Journal of Machine Learning Research 27 (2026) 1-79 Submitted 12/22; Revised 12/25; Published 1/26

Decorrelated Local Linear Estimator: Inference for Non-linear Effects in High-dimensional Additive Models

Zijian Guo ZIJGUO@ZJU.EDU.CN
Center for Data Science
Zhejiang University
Hangzhou, Zhejiang, China
Wei Yuan WY204@STAT.RUTGERS.EDU
Cunhui Zhang CZHANG@STAT.RUTGERS.EDU
Department of Statistics
Rutgers University
Piscataway, NJ 08854, USA

Editor: Mladen Kolar

Abstract

Additive models play an essential role in studying non-linear relationships. Despite many recent advances in estimation, there is a lack of methods and theories for inference in high-dimensional additive models, including confidence interval construction and hypothesis testing. Motivated by inference for non-linear treatment effects, we consider the high-dimensional additive model and make inferences for the function derivative. We propose a novel decorrelated local linear estimator and establish its asymptotic normality. The main novelty is the construction of the decorrelation weights, which is instrumental in reducing the error inherited from estimating the nuisance functions in the high-dimensional additive model. We construct the confidence interval for the function derivative and conduct the related hypothesis testing. We demonstrate our proposed method over large-scale simulation studies and apply it to identify non-linear effects in the motif regression problem. Our proposed method is implemented in the R package `DLL` available from CRAN.

Keywords: Bias correction, local polynomial, sparse model, asymptotic normality, confidence interval

1. Introduction

Additive models play an important role in modern data analysis (Buja et al., 1989; Wood, 2017; Hastie and Tibshirani, 1986). The additive model is useful as it relaxes the stringent linearity assumption imposed in the multiple linear models and generalizes the nice interpretation of linear models. In the low-dimensional setting, additive models have been carefully investigated (Buja et al., 1989; Wood, 2017; Hastie and Tibshirani, 1986; Opsomer, 2000; Horowitz and Mammen, 2004; Mammen et al., 1999, e.g.). Recently, there has been a growing interest in the high-dimensional additive model, which generalizes the high-dimensional linear regression. Much progress has been made to understand the prediction performance of various proposals, including Meier et al. (2009); Koltchinskii and Yuan (2010); Raskutti et al. (2012); Suzuki and Sugiyama (2013); Tan and Zhang (2019); Yang and Tokdar (2015); Yuan and Zhou (2016); Ravikumar et al. (2009). However, the statistical inference problem in the high-dimensional additive model is far less understood from both methodological and theoretical perspectives.

Statistical inference in high-dimensional additive models is well-motivated from causal inference with observational studies. Causal conclusions from observational studies are invalidated due to unmeasured confounders (Imbens and Rubin, 2015; Pearl, 2009, e.g.). A commonly used approach is to condition on a large number of measured covariates such that the ignorability condition holds (Belloni et al., 2017; Hernan and Robins, 2025, e.g.). This idea has been carefully investigated by utilizing high-dimensional linear models. However, the linear model imposes a stringent assumption that the exposure has a constant effect on the outcome regardless of the exposure value. Such an assumption might not be plausible for various applications. Non-linear effects have been commonly observed in scientific studies, including the return to schooling (Card, 2001), climate on crop yields (Schlenker and Roberts, 2008), and the climate change on the economic outcomes (Deschênes and Greenstone, 2012; Dell et al., 2014). The additive model relaxes the linearity assumption and better accommodates the possibly non-linear effect.

In this paper, we consider the additive model for the outcome variable $Y_i \in \mathbb{R}$,

$$Y_i = f(D_i) + g(X_i) + \epsilon_i \quad \text{with} \quad g(X_i) = \sum_{j=1}^{p} g_j(X_{i,j}), \quad \text{for} \quad 1 \leq i \leq n, \tag{1}$$

where $D_i \in \mathbb{R}$ is the variable of interest (e.g. the exposure or treatment variable), $X_i \in \mathbb{R}^p$ is the high-dimensional baseline covariates, $\mathbf{E}(\epsilon_i \mid D_i, X_i) = 0$, and $f : \mathbb{R} \to \mathbb{R}$ and $g_j : \mathbb{R} \to \mathbb{R}$ for $1 \leq j \leq p$ are unknown functions. Throughout the paper, we further assume that the treatment model follows either a sparse linear model or non-linear but additive model; see Section 2.2 and 2.6 for details. The observed data $\{Y_i, D_i, X_i\}_{1 \leq i \leq n}$ are assumed to be independently and identically distributed. For a pre-specified evaluation value $a_0 \in \mathbb{R}$, when the treatment value is changed from a_0 to $a_0 + \tau$ with $\tau \in \mathbb{R}$, the conditional mean specified in (1) changes as follows,

$$\mathbf{E}(Y_i \mid D_i = a_0 + \tau, X_i) - \mathbf{E}(Y_i \mid D_i = a_0, X_i) = f(a_0 + \tau) - f(a_0).$$

Under (1), $\left(f(a_0 + \tau) - f(a_0)\right)/\tau$ measures the rate of change at a_0. With τ approaching zero, the derivative of the function $f'(a_0)$ is an important measure of treatment effect (Belloni et al., 2015; Kozbur, 2021). We now leverage the potential outcome framework to give a causal interpretation of the derivative $f'(a_0)$. Let $Y_i(a)$ denote the potential outcome for unit i under the treatment level a. Combining the additive model (1) with the following standard causal assumptions,

1. Consistency: $Y_i = Y_i(D_i)$;

2. Unconfoundedness: $Y_i(a) \perp\!\!\!\perp D_i \mid X_i$ for all $a \in \mathbb{R}$;

we establish that

$$E\big(Y_i(a_0 + h) - Y_i(a_0) \mid X_i\big) = f(a_0 + h) - f(a_0), \tag{2}$$

which implies that the rate of change of the conditional outcome satisfies

$$f'(a_0) = \lim_{h \to 0} \frac{1}{h} \Big[E\big(Y_i(a_0 + h) - Y_i(a_0) \mid X_i\big) \Big].$$

We include a proof of (2) in Section E.1.

1.1 Results and Contributions

In the univariate setting, the local linear estimator is the state-of-the-art method to make inference for $f'(a_0)$ (Fan and Gijbels, 1996; Fan, 1993, e.g.). However, the inference problem under the high-dimensional additive model (1) is much more challenging due to the presence of the unknown high-dimensional function g. With an accurate estimator $\widehat{g}$, the estimated variables $\widehat{Y}_i = Y_i - \widehat{g}(X_i)$ for $1 \leq i \leq n$ can be used as proxies for $\{f(D_i)\}_{1\leq i\leq n}$. A natural idea is to estimate $f'(a_0)$ by applying the local linear method to $\{D_i, \widehat{Y}_i\}_{1\leq i\leq n}$ with $\{\widehat{Y}_i\}_{1\leq i\leq n}$ as the outcome variables. However, such a plug-in estimator suffers from large estimation bias due to the estimation error of $\widehat{g}$; see Table 1 in Section 4.1 for illustrations.

We propose a novel Decorrelated Local Linear (`DLL`) estimator of $f'(a_0)$. The classical local linear estimator can be expressed as a weighted average of the outcome variables, where the local linear kernel induces the weights. As the major novelty, we construct the *decorrelation weights* to mitigate the error inherited from the high-dimensional estimator $\widehat{g}$. Meanwhile, the constructed decorrelation weights ensure that the standard error of our proposed `DLL` estimator is comparable to that of the classical local linear estimator. The decorrelation weights are constructed in a non-parametric way and designed for the bias correction of the local linear estimator.

In Theorem 1, we establish the asymptotic normality of our proposed `DLL` estimator as long as the estimator $\widehat{g}$ is consistent. We further establish the asymptotic normality of our proposed estimator and show that our proposed estimator's asymptotic variance matches the optimal rate in the univariate setting (Fan and Gijbels, 1996). We construct the confidence interval for $f'(a_0)$ and test for the hypothesis $H_0 : f'(a_0) = 0$. Unlike previous approaches where debiased approaches are applied to the basis transformations of the variable of interest, which requires strong sparsity assumption in such regression, we avoid such restrictive assumptions by imposing a linear model assumption on the relationship between the variable of interest and covariates. In Section 2.6, we further discuss the possibility of relaxing this linear model to a nonlinear model assumption.

In Section 4, we demonstrate the validity of our theoretical results in moderate sample sizes, address practical issues on algorithm implementation, and provide practical recommendations. Our proposed method is implemented in the R package `DLL`, which is available from CRAN. The simulation results show that the `DLL` estimator outperforms the plug-in estimator and the ReSmoothing estimator (Gregory et al., 2021), in terms of the bias correction and coverage property. Regarding the empirical coverage and length, the confidence intervals (CIs) based on the `DLL` estimator are comparable to the oracle CIs, which are constructed with the oracle knowledge of the high-dimensional function g. We apply our proposed method to the motif regression problem (Yuan et al., 2007) and observe a highly non-linear relationship between the gene expression level and the motif scores. Our results in Section 5 demonstrate the advantage of our proposed method over the statistical inference method assuming the linear outcome model.

1.2 Literature Review and Comparison

Three recent works Gregory et al. (2021), Kozbur (2021) and Lu et al. (2020) studied the inference problems in high-dimensional additive models. Specifically, Gregory et al. (2021) proposed a two-step ReSmoothing (`RS`) estimator: in the first step, a pre-smoothing estimator was obtained; in the second step, the pre-smoothing estimator was taken as the proxy outcome, and standard univariate non-parametric technique was then applied. In Section 4.2, we compare our proposed `DLL` estimator with the `RS` estimator and observe that the `RS` estimator suffers from a large bias of estimating the function derivative for relatively small sample sizes while our proposed `DLL` estimator corrects the bias effectively. Consequently, our proposed confidence interval has better empirical coverage than that based on the `RS` estimator; see Table 4 for the detailed comparison. In addition, Lu et al. (2020) considered the confidence band construction problem under the high-dimensional sparse additive model, which is a different inference problem from the current paper.

Kozbur (2021) proposed a Post Nonparametric Double selection method to make inference for linear functionals of f in (1), including $f'(d_0)$ as a special case. Particularly, Kozbur (2021) assumed that f and g in (1) are respectively approximated by the linear combinations of K_f and K_g basis functions and further assumed that that coefficients in the K_g basis are s_0 sparse. They conducted $K_f + 1$ Lasso regressions, using Y and K_f basis functions as outcomes and K_g basis functions as high-dimensional predictors, to select a reduced dictionary for each outcome. Their confidence interval requires the size of the union of the selected dictionaries to be of a smaller order of magnitude than $K_f^{\alpha} s_0$ for some $\alpha > 0$. In our setting, when both f and components of g are twice differentiable, we may set $K_f = n^{1/5}$ and $s_0 = \|g\|_0 \cdot n^{1/5}$, with $\|g\|_0$ denoting the number of nonzero components in the additive model g. Kozbur (2021) assumed that $K_f^{1+\alpha'+\alpha/2} s_0/n^{1/3} \to 0$ for some positive α', α. Their sparsity condition is restrictive and does not cover our setting with $K_f = n^{1/5}$ and $s_0 = \|g\|_0 \cdot n^{1/5}$, and their condition on the size of the selected dictionaries is not verified under the additive model specified in (1); see Assumption 7 and the discussion after Assumption 7 in Kozbur (2021). Our proposal is completely different from Kozbur (2021), and our condition is less restrictive and more closely related to traditionally imposed smoothness conditions. In Section F.5 in the appendix, we compare our proposal with theirs in simulation studies and observe that our point estimator has a much smaller bias and our confidence interval is valid but has a much smaller length.

Inference for function derivative has been actively studied in the non-parametric modeling, including local linear estimator (Fan, 1993), regression spline (Zhou and Wolfe, 2000), kernel methods (Gasser and Müller, 1984), empirical likelihood methods (Qin and Tsao, 2005), and others cited therein. However, as discussed, the unknown high-dimensional function in the additive model (1) poses great challenges to statistical inference for the function derivative at a local point. Belloni et al. (2015) studied the inference for the function derivative in additive models without the sparsity structure. The penalty is essential to recovering the high-dimensional sparse model, which creates an additional bias to correct in the following inference step. In contrast to the results in Belloni et al. (2015), the statistical inference problem with the sparsity structure requires extra innovation in terms of both method and theory.

A recent line of active research was focused on statistical inference in high-dimensional linear regression. Debiased estimators or Neyman's Orthogonalization were proposed for inference for single regression coefficients (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014; Belloni et al., 2014; Chernozhukov et al., 2015; Farrell, 2015; Chernozhukov et al., 2018; Ning and Liu, 2017, e.g.). The linear model is a special case of the additive model, where the function derivative $f'(a_0)$ is assumed to be a constant for any $a_0 \in \mathbb{R}$. Statistical inference for the non-linear effect in the additive model is more challenging, which requires novel methods and theories to address the non-linearity. Both the rate of convergence and the sufficient conditions for confidence interval construction are different from those established in the high-dimensional linear regression. A more detailed methodological comparison is presented in Remark 4. The real data analysis in Section 5 shows that a misleading scientific conclusion might be obtained without accounting for the possible non-linear effects.

Beyond the high-dimensional linear regression, Chernozhukov et al. (2018) and Zhu et al. (2019) studied the inference procedure for the partially linear model. However, the focus is still on the inference problem for the linear component instead of the non-linear component addressed here.

Organization. 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 conduct a large-scale simulation study to demonstrate the finite-sample performance of the `DLL` estimator. In Section 5, we apply the `DLL` estimator to the motif regression problem. In Section 6, we provide conclusion and discussion.

Notations. For a sequence of random variables X_n indexed by n, we use $X_n \overset{p}{\to} X$ and $X_n \overset{d}{\to} X$ to represent that X_n converges to X in probability and in distribution, respectively. For a matrix X, we use $X_{i,j}$, X_i, and $X_{\cdot,j}$ to denote its (i,j) entry, the i-th row and j-th column, respectively; for index sets S_1 and S_2, X_{S_1,S_2} denotes the sub-matrix of X with row and column indices belonging to S_1 and S_2, respectively. 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 $\limsup_{n\to\infty} a_n/b_n = 0$ and $a_n \gg b_n$ if $b_n \ll a_n$.

2. Decorrelated Local Linear Estimator

We consider the data $\{X_i, D_i, Y_i\}_{1\leq i\leq n}$ being i.i.d. generated, where for the i-th subject, $Y_i \in \mathbb{R}$ denotes the outcome variable, $D_i \in \mathbb{R}$ denotes the variable of interest, and $X_i \in \mathbb{R}^p$ denotes the high-dimensional baseline covariates. We focus on the additive outcome model (1). Our goal is to make inference for the function derivative $f'(a_0)$ with $a_0 \in \mathbb{R}$ denoting a pre-specified evaluation value belonging to the range of D_i.

2.1 Decorrelation for the Local Linear Estimator

In the following, we review the local polynomial estimator and then describe the novel decorrelation idea designed for the local linear estimator. The local polynomial estimator (Fan, 1993; Stone, 1977; Cleveland, 1979; Fan, 1992) has been developed under the univariate non-parametric regression $Y_i = f(D_i) + \epsilon_i$, for $1 \leq i \leq n$, with $f : \mathbb{R} \to \mathbb{R}$ denoting an unknown smooth function. By the Taylor expansion

$$f(D_i) = f(a_0) + f'(a_0)(D_i - a_0) + r(D_i) \quad \text{for} \quad a_0 - h \leq D_i \leq a_0 + h, \tag{3}$$

we approximate f by a linear function in a small neighborhood of a_0 and estimate $f'(a_0)$ by fitting a linear model within this small neighborhood. For a pre-specified bandwidth $h > 0$, define the kernel

$$K_h(d) = \frac{1}{2h} \cdot \mathbf{1}\left(|d - a_0| \leq h\right). \tag{4}$$

The local linear estimator of $f'(a_0)$ has the following explicit form,

$$\frac{\sum_{i=1}^{n} W_i^0 Y_i K_h(D_i)}{\sum_{i=1}^{n} W_i^0 (D_i - a_0) K_h(D_i)}, \quad \text{with} \quad W_i^0 = (D_i - a_0) - \frac{\sum_{j=1}^{n} (D_j - a_0) K_h(D_j)}{\sum_{j=1}^{n} K_h(D_j)}. \tag{5}$$

The local polynomial estimator is the state-of-the-art method for estimating $f'(a_0)$ in the univariate and low dimensional setting. In the current paper, we focus on the uniform kernel in (4) and the local linear estimator in (5), but our following proposed method is potentially helpful for other kernels; see Remark 2 for details.

For the high-dimensional sparse additive model in (1), the existing literature (Meier et al., 2009; Koltchinskii and Yuan, 2010; Suzuki and Sugiyama, 2013; Tan and Zhang, 2019, e.g.) was focused on accurately estimating the unknown functions f and g in (1). However, there is a lack of inference methods for $f'(a_0)$. In the remaining of the current subsection, we introduce the decorrelation idea for the local linear estimator. This idea leads to our proposed decorrelated local linear estimator, which is particularly useful for the statistical inference in additive models.

We use $\widehat{g}$ to denote an initial estimator of g with its detailed construction provided in the following Section 2.4. With the estimator $\widehat{g}$, we compute $\widehat{Y}_i = Y_i - \widehat{g}(X_i)$ for $1 \leq i \leq n$, which are proxies for the oracle outcome $Y_i^{\text{ora}} = f(D_i) + \epsilon_i$. As a direct extension of the local linear estimator in (5), we replace Y_i by $\widehat{Y}_i$ and have the following plug-in estimator,

$$\widetilde{f'(a_0)} = \frac{\sum_{i=1}^{n} W_i^0 \widehat{Y}_i K_h(D_i)}{\sum_{i=1}^{n} W_i^0 (D_i - a_0) K_h(D_i)}. \tag{6}$$

Note that $\widetilde{f'(a_0)}$ is the same as the local linear estimator applied to the data $\{D_i, \widehat{Y}_i\}_{1 \leq i \leq n}$, with $\widehat{Y}_i$ as the outcome variable. The simulation results in Section 4.1 demonstrate that the plug-in estimator $\widetilde{f'(a_0)}$ suffers from a large bias due to the estimation error of $\widehat{g}$; see Table 1 for details. Consequently, the plug-in estimator is not ready for statistical inference.

To correct the bias of the plug-in estimator, we consider the estimator of the form,

$$\widehat{f'(a_0)} = \frac{\frac{1}{n} \sum_{i=1}^{n} W_i \widehat{Y}_i K_h(D_i)}{\frac{1}{n} \sum_{i=1}^{n} W_i (D_i - a_0) K_h(D_i)}, \tag{7}$$

where $\{W_i \in \mathbb{R}\}_{1 \leq i \leq n}$ are the weights to be specified. We decompose the error of the generic estimator $\widehat{f'(a_0)} - f'(a_0)$ in (7) as

$$\frac{\frac{1}{n} \sum_{i=1}^{n} W_i [f(a_0) + r(D_i) + \epsilon_i] K_h(D_i)}{\frac{1}{n} \sum_{i=1}^{n} W_i (D_i - a_0) K_h(D_i)} + \frac{\frac{1}{n} \sum_{i=1}^{n} W_i [\widehat{g}(X_i) - g(X_i)] K_h(D_i)}{\frac{1}{n} \sum_{i=1}^{n} W_i (D_i - a_0) K_h(D_i)}, \tag{8}$$

where the remainder function $r(\cdot)$ is defined in (3). The first term of (8) appears in classical univariate non-parametric regression, while the second term is the new addition due to estimating the high-dimensional function g.

In view of (8), we construct the weights $\{W_i\}_{1\leq i\leq n}$ such that the second term of (8) is significantly reduced, but ensure that the first term in (8) is of the same scale as the univariate case. To achieve this, we define the population decorrelation weights $\{W_i\}_{1\leq i\leq n}$ as

$$W_i = (D_i - a_0) - l(X_i) \quad \text{with} \quad l(X_i) := \frac{\mathbf{E}\left([D_i - a_0]K_h(D_i)|X_i\right)}{\mathbf{E}\left(K_h(D_i)|X_i\right)}. \tag{9}$$

The population decorrelation weights ensure that the estimator $\widehat{f'(a_0)}$ in (7) is nearly unbiased and asymptotically normal. In the following, we provide intuitions on how the population decorrelation weights $\{W_i\}_{1\leq i\leq n}$ reduce the second term in (8). The weights $\{W_i\}_{1\leq i\leq n}$ guarantee

$$\mathbf{E}\left[W_i K_h(D_i) \mid X_i\right] = 0.$$

If $(D_i, X_i^{\intercal}, Y_i)^{\intercal}$ is not used to construct $\widehat{g}$, then the above equation implies

$$\mathbf{E}\left[W_i(\widehat{g}(X_i) - g(X_i))K_h(D_i) \mid X_i\right] = 0. \tag{10}$$

The zero mean of the estimation error $W_i(\widehat{g}(X_i) - g(X_i))K_h(D_i)$ guarantees the second term of the decomposition (8) converges to zero at a fast rate. In Section 2.2, we propose a non-parametric estimator of the decorrelation weight W_i defined in (9).

2.2 Construction of Decorrelation Weights

Now we estimate $l(X_i)$ and W_i defined in (9). We firstly consider decoupling the relationship between D_i and X_i via a high-dimensional sparse linear model,

$$D_i = X_i^{\intercal}\gamma + \delta_i, \quad \text{for } 1 \leq i \leq n, \tag{11}$$

where γ is a sparse vector and δ_i is independent of X_i. Let $\phi(\delta)$ denote the density function of δ_i. Under the model (11), we obtain the following expression for $l(X_i)$,

$$l(X_i) = \frac{\int_{\mu_i-h}^{\mu_i+h} (\delta - \mu_i)\, \phi(\delta) d\delta}{\int_{\mu_i-h}^{\mu_i+h} \phi(\delta) d\delta} \quad \text{with} \quad \mu_i = a_0 - X_i^{\intercal}\gamma, \quad \text{for} \quad 1 \leq i \leq n. \tag{12}$$

We also write $l(X_i, \gamma)$ for $l(X_i)$ to highlight its dependence on γ. Before presenting the data-dependent estimation of $l(X_i)$, we provide two remarks.

Remark 1 (Generalizing the linear model (11)**)** The advantage of our approach is that we do not require the basis transformation of D_i to be well approximated by a sparse linear combination of the basis functions defined on the high-dimensional covariate X_i, which is generally not verified even under the treatment model (11). Without such sparsity conditions, the independence between X_i and δ_i in (11) plays a crucial role in our analysis through simplifying the expression of $l(X_i)$ as in (12). In Section 4.1, we demonstrate the robustness of our proposed method in finite samples when the independence assumption in the model (11) does not hold; see Settings 3 and 4 in Section 4.1 and Table 1 for details. In Section 2.6, We consider the relaxation of the sparse linear model (11) to a more flexible sparse additive model.

Remark 2 (Generalizing the kernel function) *We consider the following kernel as a generalization of the box kernel in* (4),

$$K_\nu(d) \coloneqq \mathbf{E}_\nu\left[K_h(d)\right] = \int \frac{1}{2h} \cdot \mathbf{1}\left(|d - a_0| \leq h\right) d\nu(h). \tag{13}$$

where ν is a probability measure of the bandwidth h and $\mathbf{E}_{h\sim\nu}$ denotes the expectation with respect to h following the probability measure ν. We then obtain a generalization of (12) *as*

$$l(X_i) = \frac{\mathbf{E}_{h\sim\nu} \int_{\mu_i-h}^{\mu_i+h} (\delta - \mu_i)\, \phi(\delta) d\delta}{\mathbf{E}_{h\sim\nu} \int_{\mu_i-h}^{\mu_i+h} \phi(\delta) d\delta}.$$

In the following, we use the expression (12) and construct a non-parametric estimator of $l(X_i)$. We will use cross fitting to create the independence required for establishing the decorrelation property in (10); see Remark 3 for details. We randomly split the index set $\{1, 2, \cdots, n\}$ into two disjoint subsets $\mathcal{I}_a$ and $\mathcal{I}_b$, with $\mathcal{I}_a \cup \mathcal{I}_b = \{1, 2, \cdots, n\}$, $|\mathcal{I}_a| = \lfloor n/2 \rfloor$, and $|\mathcal{I}_b| = n - \lfloor n/2 \rfloor$. With the data $\{Y_i, D_i, X_i\}_{i\in\mathcal{I}_a}$, we estimate γ by the Lasso estimator $\widehat{\gamma}^a \in \mathbb{R}^p$, defined as

$$\widehat{\gamma}^a = \operatorname*{arg\,min}_{\gamma\in\mathbb{R}^p} \frac{1}{2|\mathcal{I}_a|} \sum_{i\in\mathcal{I}_a} (D_i - X_i^{\intercal}\gamma)^2 + \lambda_1 \sum_{j=1}^{p} \frac{\|X_{\mathcal{I}_a,j}\|_2}{\sqrt{n_a}} |\gamma_j|, \tag{14}$$

where $\lambda_1 > 0$ is a tuning parameter and $\|X_{\mathcal{I}_a,j}\|_2 = \sqrt{\sum_{i\in\mathcal{I}_a} X_{i,j}^2}$. We estimate $\{\mu_i\}_{i\in\mathcal{I}_b}$ and $\{\delta_i\}_{i\in\mathcal{I}_b}$ by

$$\widehat{\mu}_i = a_0 - X_i^{\intercal}\widehat{\gamma}^a \quad \text{and} \quad \widehat{\delta}_i = D_i - X_i^{\intercal}\widehat{\gamma}^a \quad \text{for} \quad i \in \mathcal{I}_b.$$

For $i \in \mathcal{I}_b$, we respectively estimate $\int_{\mu_i-h}^{\mu_i+h} (\delta - \mu_i)\, \phi(\delta) d\delta$ and $\int_{\mu_i-h}^{\mu_i+h} \phi(\delta) d\delta$ by

$$\frac{1}{|\mathcal{I}_b|} \sum_{j\in\mathcal{I}_b} (\widehat{\delta}_j - \widehat{\mu}_i)\mathbf{1}(|\widehat{\delta}_j - \widehat{\mu}_i| \leq h) \quad \text{and} \quad \frac{1}{|\mathcal{I}_b|} \sum_{j\in\mathcal{I}_b} \mathbf{1}(|\widehat{\delta}_j - \widehat{\mu}_i| \leq h).$$

Then we estimate $l(X_i)$ by

$$\widehat{l}(X_i, \widehat{\gamma}^a) = \frac{\sum_{j\in\mathcal{I}_b} (\widehat{\delta}_j - \widehat{\mu}_i)\mathbf{1}(|\widehat{\delta}_j - \widehat{\mu}_i| \leq h)}{\sum_{j\in\mathcal{I}_b} \mathbf{1}(|\widehat{\delta}_j - \widehat{\mu}_i| \leq h)} \quad \text{for} \quad i \in \mathcal{I}_b. \tag{15}$$

Our above construction ensures that $\widehat{\gamma}^a$ is independent of the data points $\{D_i, X_i\}_{i\in\mathcal{I}_b}$. We can construct the estimators of $\{l(X_i)\}_{i\in\mathcal{I}_a}$ in a similar way to (15) by switching the roles of $\mathcal{I}_a$ and $\mathcal{I}_b$. In particular, we construct the estimator $\widehat{\gamma}^b \in \mathbb{R}^p$ by applying the Lasso algorithm in (14) to the data $\{Y_i, D_i, X_i\}_{i\in\mathcal{I}_b}$ and estimate $\{\mu_i, \delta_i\}_{i\in\mathcal{I}_a}$ by

$$\widehat{\mu}_i = a_0 - X_i^{\intercal}\widehat{\gamma}^b \quad \text{and} \quad \widehat{\delta}_i = D_i - X_i^{\intercal}\widehat{\gamma}^b \quad \text{for} \quad i \in \mathcal{I}_a.$$

Similarly to (15), we estimate $\{l(X_i)\}_{i\in\mathcal{I}_a}$ by

$$\widehat{l}(X_i,\widehat{\gamma}^b)=\frac{\sum_{j\in\mathcal{I}_a}(\widehat{\delta}_j-\widehat{\mu}_i)\mathbf{1}(|\widehat{\delta}_j-\widehat{\mu}_i|\leq h)}{\sum_{j\in\mathcal{I}_a}\mathbf{1}(|\widehat{\delta}_j-\widehat{\mu}_i|\leq h)}\quad\text{for}\quad i\in\mathcal{I}_a. \tag{16}$$

By the definition in (9), we construct

$$\widetilde{W}_i=(D_i-a_0)-\widehat{l}(X_i)\quad\text{with}\quad\widehat{l}(X_i)=\begin{cases}\widehat{l}(X_i,\widehat{\gamma}^b) & \text{for } i\in\mathcal{I}_a\\ \widehat{l}(X_i,\widehat{\gamma}^a) & \text{for } i\in\mathcal{I}_b\end{cases}, \tag{17}$$

where $\widehat{l}(X_i,\widehat{\gamma}^a)$ and $\widehat{l}(X_i,\widehat{\gamma}^b)$ are defined in (15) and (16), respectively. By centering $\{\widetilde{W}_i\}_{1\leq i\leq n}$, we construct the decorrelation weights as

$$\widehat{W}_i=\widetilde{W}_i-[\sum_{j=1}^n\widetilde{W}_jK_h(D_j)]/[\sum_{j=1}^nK_h(D_j)]\quad\text{for}\quad 1\leq i\leq n. \tag{18}$$

With the data $\{Y_i,D_i,X_i\}_{i\in\mathcal{I}_a}$, we construct the initial estimator $\widehat{g}^a(\cdot)$ of $g(\cdot)$ in the following equations (23) and (24); we construct the estimators $\widehat{g}^b(\cdot)$ by applying the same algorithm to the data $\{Y_i,D_i,X_i\}_{i\in\mathcal{I}_b}$. For $1\leq i\leq n$, we compute

$$\widehat{Y}_i=Y_i-\widehat{g}(X_i)\quad\text{with}\quad\widehat{g}(X_i)=\begin{cases}\widehat{g}^b(X_i) & \text{for } i\in\mathcal{I}_a\\ \widehat{g}^a(X_i) & \text{for } i\in\mathcal{I}_b\end{cases}. \tag{19}$$

By combining $\widehat{Y}_i$ defined in (19) and the decorrelation weight $\widehat{W}_i$ defined in (18), we apply the generic form (7) and propose the Decorrelated Local Linear (`DLL`) estimator as

$$\widehat{f'(a_0)}=\frac{1}{n\widehat{S}_n}\sum_{i=1}^n\widehat{W}_i\widehat{Y}_iK_h(D_i)\quad\text{where}\quad\widehat{S}_n=\frac{1}{n}\sum_{i=1}^n\widehat{W}_i(D_i-a_0)K_h(D_i). \tag{20}$$

A few remarks are in order for the proposed `DLL` estimator.

Remark 3 The construction in (17) and (19) uses the "data swapping" idea, dated at least back to Schick (1986); Klaassen (1987) and was recently developed under the name of "cross-fitting" in the double machine learning literature (Chernozhukov et al., 2018, e.g.). That is, we swap the data and the initial estimators. In our considered context of the sparse additive models, such a procedure is used to create the independence required for the proof. In theory, our particular way of cross-fitting does not lead to asymptotic efficiency loss. In practice, however, we observe slightly longer confidence intervals with cross-fitting, as higher-order errors, ignored in the asymptotic analysis, may still induce larger finite-sample errors when the nuisance model is constructed using only a subsample. To demonstrate this point, we empirically compare the finite-sample performances of our method with and without "data swapping". The results are similar in the sense of bias and empirical coverage, except that the confidence interval is about 4% to 6% longer with "data swapping". See Section F.3 in the appendix for more details and simulation results.

Remark 4 (Comparison to debiasing methods in linear models.) The debiased inference methods have been proposed in Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014) about inference for the regression coefficients in high-dimensional regression models and extended to other high-dimensional parametric models (Ning and Liu, 2017; van de Geer et al., 2014) or other inference targets in high-dimensional linear regression (Cai and Guo, 2017; Cai et al., 2021; Athey et al., 2018; Zhu and Bradic, 2018). These methods utilize that $\widehat{\delta}$ is nearly orthogonal to X and then correct the bias with a linear function of $\widehat{\delta}$. In contrast, our proposed `DLL` estimator uses a non-linear transformation of $\widehat{\delta}$ to remove the high-dimensional error. Specifically, we construct the decorrelation weights based on certain kernel estimates with $\{\widehat{\delta}_i\}_{1\leq i\leq n}$; see (15) and (16). This new decorrelation idea is designed for bias correction of the local linear estimator.

2.3 Inference for $f'(a_0)$

In Section 3, we show that $\widehat{f'(a_0)} - f'(a_0)$ is asymptotically normal if certain reasonably good estimator $\widehat{g}$ is used in our construction. Consequently, we construct the following $1-\alpha$ confidence interval for $f'(a_0)$,

$$\mathrm{CI}[f'(a_0)] = \left(\widehat{f'(a_0)} - z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}}, \widehat{f'(a_0)} + z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}}\right) \quad \text{with} \quad \widehat{\mathrm{V}} = \frac{\widehat{\sigma}^2}{n^2\widehat{S}_n^2}\sum_{i=1}^{n}\widehat{W}_i^2 K_h^2(D_i), \tag{21}$$

where $z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile of the standard normal distribution and $\widehat{\sigma}^2$ is the variance level estimator specified in Section 2.4. To test the null hypothesis $H_0 : f'(a_0) = 0$, we develop the following procedure,

$$\psi[f'(a_0)] = \mathbf{1}\left(|\widehat{f'(a_0)}| \geq z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}}\right). \tag{22}$$

2.4 Initial Estimators

We now specify the initial estimators $\widehat{g}$ and $\widehat{\sigma}^2$ used in the construction of the `DLL` estimator in (20) and the related confidence interval in (21). In the existing literature (Meier et al., 2009; Ravikumar et al., 2009; Koltchinskii and Yuan, 2010; Suzuki and Sugiyama, 2013; Tan and Zhang, 2019), different types of penalty terms are imposed to ensure that only a small number of the unknown functions f and $\{g_j\}_{1\leq j\leq p}$ are non-zero and these non-zero functions are smooth. We adopt the basis method in Meier et al. (2009); Ravikumar et al. (2009) and generate a set of basis functions to approximate the smooth functions. In particular, for a positive integer M, we use $\{\phi_{0,l}\}_{1\leq l\leq M}$ to denote a set of B-spline basis functions for f and $\{\phi_{j,l}\}_{1\leq l\leq M}$ to denote a set of B-spline basis functions for g_j for $1\leq j\leq p$. We write $\Psi_{i,0} = (\phi_{0,1}(D_i), \cdots, \phi_{0,M}(D_i)) \in \mathbb{R}^M$ and $\Psi_{i,j} = (\phi_{j,1}(X_{i,j}), \cdots, \phi_{j,M}(X_{i,j})) \in \mathbb{R}^M$ for $1\leq j\leq p$. Following Ravikumar et al. (2009), we implement the following convex optimization problem,

$$\{\widehat{\beta}_j^a\}_{0\leq j\leq p} = \mathop{\arg\min}_{\beta_j\in\mathbb{R}^M,\, 0\leq j\leq p} \frac{1}{2|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}(Y_i - \sum_{j=0}^{p}\Psi_{i,j}^{\intercal}\beta_j)^2 + \lambda\sum_{j=0}^{p}\sqrt{\beta_j^{\intercal}\left(\frac{1}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}\Psi_{i,j}\Psi_{i,j}^{\intercal}\right)\beta_j}, \tag{23}$$

where $\lambda > 0$ is a tuning parameter to be chosen. The choice of the tuning parameter $\lambda > 0$ and the number M of basis functions are discussed at the beginning of Section 4. Define

$$\widehat{g}^a(X_i) = \sum_{j=1}^{p} \Psi_{i,j}^{\intercal} \widehat{\beta}_j^a \quad \text{and} \quad \widehat{f}^a(D_i) = \Psi_{i,0}^{\intercal} \widehat{\beta}_0^a \quad \text{for} \quad i \in \mathcal{I}_b. \tag{24}$$

Similarly, we define $\{\widehat{\beta}_j^b\}_{0 \le j \le p}$ as in (23) by replacing $\mathcal{I}_a$ with $\mathcal{I}_b$ and

$$\widehat{g}^b(X_i) = \sum_{j=1}^{p} \Psi_{i,j}^{\intercal} \widehat{\beta}_j^b \quad \text{and} \quad \widehat{f}^b(D_i) = \Psi_{i,0}^{\intercal} \widehat{\beta}_0^b \quad \text{for} \quad i \in \mathcal{I}_a.$$

Then we construct

$$\widehat{g}(X_i) = \begin{cases} \widehat{g}^b(X_i) & \text{for } i \in \mathcal{I}_a \\ \widehat{g}^a(X_i) & \text{for } i \in \mathcal{I}_b \end{cases} \quad \text{and} \quad \widehat{f}(D_i) = \begin{cases} \widehat{f}^b(D_i) & \text{for } i \in \mathcal{I}_a \\ \widehat{f}^a(D_i) & \text{for } i \in \mathcal{I}_b \end{cases}, \tag{25}$$

and estimate the variance level σ^2 by

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \widehat{f}(D_i) - \widehat{g}(X_i)]^2. \tag{26}$$

In addition to the aforementioned basis method, we can also adopt the double penalization method (Tan and Zhang, 2019) by generalizing the smoothing spline; see Section A.2 in the appendix. We also discuss the construction of additive models by firstly applying the quantile transformation to the observed covariates; see Section A.3 in the appendix.

2.5 Algorithm

We summarize our proposed DLL estimator (with data swapping) in Algorithm 1 and will present the DLL estimator without data swapping in Section A.1 in the appendix. We shall discuss the tuning parameter selection at the beginning of Section 4.

2.6 Decorrelation with the Additive Treatment Model

This section generalizes the construction of decorrelation weights by considering non-linear models between D_i and X_i. Particularly, we consider the sparse additive model for D_i,

$$D_i = \sum_{j=1}^{p} \tau_j(X_{i,j}) + \delta_i, \quad \text{for } 1 \le i \le n,$$

where $\tau_j : \mathbb{R} \to \mathbb{R}$ for $1 \le j \le p$ are unknown smooth functions and δ_i is independent of X_i. Instead of applying the Lasso algorithm (14), we implement another sparse additive model as in (23),

$$\{\widehat{\gamma}_j^a\}_{1 \le j \le p} = \underset{\gamma_j \in \mathbb{R}^M,\, 1 \le j \le p}{\arg\min} \frac{1}{2|\mathcal{I}_a|} \sum_{i \in \mathcal{I}_a} (D_i - \sum_{j=1}^{p} \Psi_{i,j}^{\intercal} \gamma_j)^2 + \lambda_1 \sum_{j=1}^{p} \sqrt{\gamma_j^{\intercal} \left(\frac{1}{|\mathcal{I}_a|} \sum_{i \in \mathcal{I}_a} \Psi_{i,j} \Psi_{i,j}^{\intercal} \right) \gamma_j}$$

Algorithm 1 Decorrelated Local Linear (DLL) Estimator

Input: Data $X \in \mathbb{R}^{n\times p}, D \in \mathbb{R}^n, Y \in \mathbb{R}^n$; the evaluation point $a_0 \in \mathbb{R}$, bandwidth h, tuning parameters $\lambda, \lambda_1 > 0$, the number of basis M.
Output: Point estimator $\widehat{f'(a_0)}$ and confidence interval $\mathrm{CI}[f'(a_0)]$.

```
Implement the sparse additive model in (23) with $M \geq 1$ and $\lambda > 0$;
Construct the initial estimators $\{\widehat{g}(X_i)\}_{1\leq i\leq n}$ as in (25);
Construct the noise estimator $\widehat{\sigma}^2$ as in (26);
Compute $\widehat{Y}_i = Y_i - \widehat{g}(X_i)$ for $1 \leq i \leq n$;          ▷ Initial estimators

Implement the Lasso algorithm as in (14) with $\lambda_1 > 0$;
Construct $\{\widehat{l}(X_i)\}_{1\leq i\leq n}$ as in (15) and (16);
Construct the weights $\{\widetilde{W}_i\}_{1\leq i\leq n}$ as in (17);
Construct the centered decorrelation weights $\{\widehat{W}_i\}_{1\leq i\leq n}$ in (18);
Construct $\widehat{f'(a_0)}$ as (20) with $\{\widehat{Y}_i, \widehat{W}_i\}_{1\leq i\leq n}$ and $h > 0$;   ▷ DLL estimator

Compute the variance estimate $\widehat{\mathrm{V}}$ as in (21);
Construct $\mathrm{CI}[f'(a_0)]$ as in (21).                                  ▷ Confidence interval
```

where $\lambda_1 > 0$ is the tuning parameter to be chosen. We estimate $\{\mu_i, \delta_i\}_{i\in\mathcal{I}_b}$ by

$$\widehat{\mu}_i = a_0 - \sum_{j=1}^{p} \Psi_{i,j}^{\intercal} \widehat{\gamma}_j^a \quad \text{and} \quad \widehat{\delta}_i = D_i - \sum_{j=1}^{p} \Psi_{i,j}^{\intercal} \widehat{\gamma}_j^a \quad \text{for} \quad i \in \mathcal{I}_b.$$

By switching the data in $\mathcal{I}_a$ and $\mathcal{I}_b$, we construct $\{\widehat{\mu}_i, \widehat{\delta}_i\}_{i\in\mathcal{I}_a}$. With the estimates $\{\widehat{\mu}_i, \widehat{\delta}_i\}_{1\leq i\leq n}$, we construct the decorrelation weights in the same way as in (15) and (16). In Section 4.3, we compare the numerical performances between our proposals with the sparse linear treatment model in Algorithm 1 and the sparse additive treatment model above.

3. Theoretical Justification

3.1 Technical Conditions

Before presenting the main theorems, we present the technical conditions imposed on the outcome model (1) and the treatment model (11). Let $\pi(a_0)$ denote the probability density function of D_i evaluated at $a_0 \in \mathbb{R}$. The first condition is on the function of interest $f(\cdot)$, the regression error ϵ_i, and the bandwidth $h > 0$.

(A1) $f(\cdot)$ is twicely differentiable at a neighborhood of a_0 and $f''(\cdot)$ is continuous at a_0. The error ϵ_i in (1) satisfies $\mathbf{E}(\epsilon_i \mid D_i, X_i) = 0$, $\mathbf{E}(\epsilon_i^2 \mid D_i, X_i) = \sigma^2$, and $\mathbf{E}(\epsilon_i^{2+c} \mid D_i, X_i) \leq C$ for some positive constants $c > 0$ and $C > 0$. The bandwidth h used in (4) satisfies $nh\pi(a_0) \gg \log n$ and $nh^5\pi(a_0) \leq C$ for some positive constant $C > 0$.

Condition (A1) is standard for the analysis of the local polynomial estimator in the univariate case (Fan and Gijbels, 1996; Fan, 1993, e.g.). The smoothness condition on f

ensures a sufficiently small approximation error of f by a linear function in a neighborhood near a_0. The conditional moment conditions on ϵ_i are required to establish the asymptotic normality of our proposed `DLL` estimator. Since the expected number of observations $D_i \in [a_0 - h, a_0 + h]$ is about $2nh\pi(a_0)$, Condition (A1) requires that there are (asymptotically) infinitely many observations in the local neighborhood of a_0 with bandwidth h. For a twicely differentiable function f, the optimal choice of bandwidth for estimating $f'(a_0)$ is $h \asymp n^{-1/5}$, which satisfies $nh^5\pi(a_0) \leq C$.

The second model assumption is imposed on the treatment model (11). Recall that $\phi(\cdot)$ denotes the probability density function of the regression error $\delta_i = D_i - X_i^{\intercal}\gamma$ in (11) and $\mu_i = a_0 - X_i^{\intercal}\gamma$ for $1 \leq i \leq n$. We use $C_1(n) > 0$ and $C_2(n) > 0$ to denote some high probability upper bounds, defined as: with probability larger than $1 - \min\{n, p\}^{-c}$ for some positive constant $c > 0$,

$$\max_{1\leq i\leq n} \max_{|\delta-\mu_i|\leq r} \frac{|\phi'(\delta)|}{\phi(\mu_i)} \leq C_1(n), \quad \max_{1\leq i\leq n} \max_{|\delta-\mu_i|\leq r} \frac{|\phi''(\delta)|}{\phi(\mu_i)} \leq C_2(n), \tag{27}$$

where $r = C^* \sqrt{\|\gamma\|_0 \log p \log n/n} + h$ for some positive constant $C^* > 0$. The value $C_1(n)$ (or $C_2(n)$) defined in (27) captures the ratio of ϕ' (or ϕ'') over ϕ near μ_i. $C_1(n)$ and $C_2(n)$ are allowed to grow with n and p, but in general they grow to infinity at a slow rate; see Remark 5 for details. We now state the condition for the model (11) and $C_1(n)$ and $C_2(n)$ defined in (27).

(A2) The model (11) holds with $k := \|\gamma\|_0 \ll \min\{1, \pi(a_0)\} \cdot \frac{n}{\log p \log n}$, X_i and δ_i being Sub-gaussian, and the error δ_i being independent of X_i. The variance of δ_i is a positive constant and $\Sigma = \mathbf{E}X_iX_i^{\intercal}$ satisfies $c_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0$ for some positive constant $c_0 > 0$ and $C_0 > 0$. The density function ϕ of δ_i is upper bounded and

$$h^2 C_2(n) + (\sqrt{\|\gamma\|_0 \log p \log n/n} + h) C_1(n) \to 0, \tag{28}$$

where $C_1(n)$ and $C_2(n)$ are defined in (27).

For a constant $\pi(a_0) > 0$, $k \ll n/[\log p \log n]$ is almost the weakest sparsity condition to identify γ for the high-dimensional linear model. The conditions on $\mathrm{Var}(\delta_i)$ and the covariance matrix Σ are standard for the high-dimensional analysis. The condition (28) is mild with $C_1(n)$ and $C_2(n)$ growing at the polynomial order of $\log n$ and $h \asymp n^{-1/5}$; see Remark 5. The independence assumption between δ_i and X_i is stringent but we believe it is mainly imposed for the technical analysis. In numerical studies, we test the performance of our proposed method when the independence assumption between δ_i and X_i is violated; see Settings 3 and 4 and Table 1 in Section 4.1 for details.

Remark 5 (Growth rates of $C_1(n)$ and $C_2(n)$.) We discuss the order of magnitudes for $C_1(n)$ and $C_2(n)$ over two examples. Firstly, consider the setting that there exist positive constant $C_0 > 0$ such that $\max_{1\leq i\leq n} |\mu_i| \leq C_0$. If $\min_{|\delta|\leq C_0} \phi(\delta) \geq c$ for some positive constant $c > 0$, and $\phi(\delta)$ is twicely differentiable for $\delta \in (C_0 - r, C_0 + r)$, then $C_1(n)$ and $C_2(n)$ are of constant orders. Secondly, we consider that X_i is sub-gaussian and μ_i may be unbounded in this case. If ϕ is the Gaussian density, and $(\sqrt{\|\gamma\|_0 \log p \log n/n} + h)\sqrt{\log n} \lesssim 1$, then with probability larger than $1 - n^{-c}$ for some positive constant $c > 0$,

$$C_1(n) \lesssim \sqrt{\log n} \quad \text{and} \quad C_2(n) \lesssim \log n.$$

Finally, we require that the initial estimator $\widehat{g}$ estimates $g = \sum_{j=1}^{p} g_j$ to certain accuracy. We use $\mathrm{Err}(\widehat{g})$ to denote the estimation accuracy of $\widehat{g}$, which is defined as follows: with probability larger than $1-\min\{n,p\}^{-c}$ for some positive constant $c > 0$, the initial estimator $\widehat{g}$ defined in (19) satisfies

$$\sqrt{\mathbf{E}_{X_*}(\widehat{g}^a(X_*) - g(X_*))^2 + \mathbf{E}_{X_*}(\widehat{g}^b(X_*) - g(X_*))^2} \lesssim \mathrm{Err}(\widehat{g}), \tag{29}$$

where the expectation $\mathbf{E}_{X_*}$ is taken with respect to the independent copy X_* of $\{X_i\}_{1\leq i\leq n}$. The last assumption is on the rate of convergence of $\mathrm{Err}(\widehat{g})$.

(A3) The estimation accuracy $\mathrm{Err}(\widehat{g})$ of initial estimator $\widehat{g}$ is required to satisfy

$$\max\left\{(C_1^2(n) + C_2(n))\sqrt{h^3 k \log p \log n}, 1\right\} \cdot \frac{\mathrm{Err}(\widehat{g})}{\sqrt{\pi(a_0)}} \to 0, \tag{30}$$

where $\mathrm{Err}(\widehat{g})$ is defined in (29).

We discuss Condition (A3) by focusing on a commonly used regime with $C_1^2(n) + C_2(n) \lesssim \log n$, $\pi(a_0)$ being of a constant order, and $h \asymp n^{-1/5}$. If $k \log p (\log n)^3/n^{3/5} \leq c$ for some positive constant $c > 0$, then any consistent estimator $\widehat{g}$ with $\mathrm{Err}(\widehat{g}) \to 0$ will automatically satisfy the condition (30). Most estimators proposed in the high-dimensional sparse additive model can be shown to satisfy the assumption (A3). More discussion about (A3) can be found in Section A.4 in the appendix.

3.2 Asymptotic Normality and Inference Properties

We establish the asymptotic limiting distribution for our proposed DLL estimator.

Theorem 1 *Suppose that Conditions* (A1)*,* (A2) *and* (A3) *hold. Then our proposed estimator* $\widehat{f'(a_0)}$ *in* (20) *satisfies,*

$$\frac{1}{\sqrt{\mathrm{V}}}\left(\widehat{f'(a_0)} - f'(a_0)\right) \xrightarrow{d} N(0,1) \quad \textit{with} \ \mathrm{V} := \frac{\sigma^2}{n^2 \widehat{S}_n^2} \sum_{i=1}^{n} \widehat{W}_i^2 K_h^2(D_i), \tag{31}$$

where $\widehat{S}_n$ *is defined in* (20) *and* $nh^3\mathrm{V} \xrightarrow{p} \frac{3\sigma^2}{\pi(a_0)}$.

It is well known that the minimax–optimal rate for estimating $f'(a_0)$ over functions with two derivatives is $n^{-1/5}$ (Tsybakov, 2009). In Theorem 1 above, the proposed DLL estimator attains this optimal $n^{-1/5}$ rate when $h \asymp n^{-1/5}$ and σ and $\pi(a_0)$ are bounded away from zero and infinity. Our analysis additionally assumes that the second derivative is continuous, which yields a slightly smaller function class than the standard Hölder class with smoothness 2. Nevertheless, the optimal $n^{-1/5}$ rate continues to hold for this smaller class, as we do not impose any specific modulus of continuity on f''. Furthermore, the DLL estimator is asymptotically normal and the asymptotic variance depends on the value a_0 through the density level $\pi(a_0)$. In finite samples, we compare the variance level of our DLL estimator to that of the oracle estimator by assuming the knowledge of g; See Table 1 for details.

As a Corollary of Theorem 1, we establish the properties of our proposed confidence interval $\mathrm{CI}[f'(a_0)]$ defined in (21).

Corollary 1 *Suppose that Conditions* (A1), (A2) *and* (A3) *hold and* $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$*. For any* $\alpha \in (0, 1/2)$*, our proposed confidence interval* $\mathrm{CI}[f'(a_0)]$ *defined in* (21) *satisfies,*

$$\liminf_{n\to\infty} \mathbb{P}(f'(a_0) \in \mathrm{CI}[f'(a_0)]) = 1 - \alpha,$$

and

$$\limsup_{n\to\infty} \mathbb{P}\left(\mathbf{L}\left(\mathrm{CI}[f'(a_0)]\right) \geq (2 + \delta_0) z_{\alpha/2} \sqrt{\frac{3}{2nh^3 \cdot \pi(a_0)}}\sigma\right) = 0,$$

where $\mathbf{L}\left(\mathrm{CI}[f'(a_0)]\right)$ *denotes the length of the interval,* $\delta_0 > 0$ *is any positive constant, and* $z_{\alpha/2}$ *denotes the upper* $\alpha/2$ *quantile of the standard normal distribution.*

Beyond Conditions (A1)-(A3), the above corollary requires a consistent estimator of σ^2 such that our proposed variance estimator $\widehat{\mathrm{V}}$ consistently estimates V. In Proposition 1 in Section A.5 in the appendix, we show that our proposed estimator $\widehat{\sigma}^2$ in (26) satisfies $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$ if both $\widehat{f}$ and $\widehat{g}$ are consistent. Similarly, we can establish the validity of the proposed testing procedure $\psi[f'(a_0)]$ in (22).

Corollary 2 *Suppose that the conditions of Corollary 1 hold. If* $f'(a_0) = 0$*, then the proposed testing procedure* $\psi[f'(a_0)]$ *defined in* (22) *controls the type I error, that is,*

$$\limsup_{n\to\infty} \mathbb{P}(\psi[f'(a_0)] = 1) = \alpha.$$

3.3 Theoretical Reasoning for Decorrelation

In the following, we explain why our constructed decorrelated weight is effective. With $\Delta(X_i) = g(X_i) - \widehat{g}(X_i)$, the estimation error of the `DLL` estimator is decomposed as

$$\widehat{f'(a_0)} - f'(a_0) = \underbrace{\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n} \widehat{W}_i \epsilon_i K_h(D_i)}_{\text{Stochastic Error}} + \underbrace{\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n} \widehat{W}_i r(D_i) K_h(D_i)}_{\text{Approximation Error}} + \underbrace{\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n} \widehat{W}_i \Delta(X_i) K_h(D_i)}_{\text{High-dimensional Error}}.$$

The stochastic error represents a random component with mean zero and, after rescaling, following an asymptotic normal limiting distribution. The approximation error is the error of approximating the non-linear function f by a linear function at a local neighborhood of a_0. The high-dimensional error is due to the estimation of the unknown function g in high dimensions. Both the stochastic and approximation errors appear in the classical non-parametric regression, while the high-dimensional error is the new addition here.

The following theorem demonstrates that our proposed decorrelation method is effective in reducing the high-dimensional error.

Theorem 2 *Suppose that Condition* (A1) *and* (A2) *hold. For* $\Delta(X_i) = g(X_i) - \widehat{g}(X_i)$ *where* $\widehat{g}$ *is defined in* (19)*, then with probability larger than* $1 - \frac{1}{t} - \min\{n, p\}^{-c}$ *for some* $t > 1$*,*

$$\left|\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n} \widehat{W}_i \Delta(X_i) K_h(D_i)\right| \leq t^2 \left[1 + \sqrt{h^3 k \log p \log n}\left(C_1^2(n) + C_2(n)\right)\right] \frac{\mathrm{Err}(\widehat{g})}{\sqrt{nh^3 \pi^2(a_0)}}, \tag{32}$$

where $\mathrm{Err}(\widehat{g})$ *is defined in* (29).

Our constructed decorrelation weights are instrumental in reducing the error due to estimating g. This happens mainly due to the fact that the expectation of $W_i\Delta(X_i)K_h(D_i)$ is zero. Condition (A3) and the upper bound (32) imply

$$\frac{1}{\sqrt{\mathrm{V}}}\left|\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_i\Delta(X_i)K_h(D_i)\right| \xrightarrow{p} 0.$$

The data swapping step creates the independence between the error function Δ and the data $\{X_i, D_i, W_i\}$, which is required for the proof of (32). We believe that a more refined analysis might remove the data swapping step.

4. Simulation

We provide more details about the tuning parameter selection for Algorithm 1. For the high-dimensional sparse additive model, we compute the initial estimators $\widehat{f}$ and $\widehat{g}$ by applying the R package `SAM` (Jiang et al., 2021) and choose the tuning parameter λ and the number of basis functions M in (23) by cross validation. We construct the Lasso estimator of γ in (14) by applying the R package `glmnet` (Friedman et al., 2010) and choose the tuning parameter λ_1 by cross validation. For local linear methods, choosing a good bandwidth is essential for the finite-sample performance. There are many methods for bandwidth selection. After exploration in the simulation study, we observe that the "Rule of Thumb" method proposed in Fan and Gijbels (1996) leads to the most stable performance. This bandwidth selection method is implemented in the R package `locpol` (Cabrera, 2018) with the `thumbBw()` function. The "Rule of Thumb" is used as our default bandwidth selection method. We demonstrate the performance of the `DLL` estimator with other bandwidth selection methods in Section F.2 in the appendix. The package for our proposed method can be found at `https://github.com/zijguo/HighDim-Additive-Inference` and the codes for replicating the simulation results can be found at `https://github.com/ywwes26/DLL-Replication`.

Since we believe that the data swapping is introduced for technical analysis, we mainly report the simulation results for the `DLL` estimator without the data swapping, which is described in Section A.1 in the appendix. We compare our constructed confidence intervals with and without data swapping in Section F.3 in the appendix. Both confidence intervals attain the desired coverage level. When the sample size is relatively large, they have similar performance; for relatively small sample size, the confidence interval without data swapping can be shorter than that with data swapping.

We demonstrate the finite-sample performance of our proposed `DLL` estimator across various settings and compare it with other estimators described as follows,

- The plug-in estimator (`Plug`) is implemented in (6), where the initial estimator $\widehat{g}$ and the bandwidth h are constructed in the same way as our proposed `DLL` estimator. For implementation of the local linear estimator and the related confidence interval, we follow the output of the package `nprobust` (Calonico et al., 2019).

- The oracle estimator (`Orac`) denotes the local linear estimator applied to the data $\{D_i, Y_i^{\mathrm{ora}}\}_{1\leq i\leq n}$ with $Y_i^{\mathrm{ora}} = Y_i - g(X_i) = f(D_i) + \epsilon_i$. The oracle estimator is used as the benchmark with which to compare. For the implementation of the local linear estimator and the related confidence interval, we follow the output of the package `nprobust` (Calonico et al., 2019) and the bandwidth h are constructed in the same way as our proposed `DLL` estimator.

- The ReSmoothing (`RS`) estimator is a two-step estimator proposed in Gregory et al. (2021). First, we implement the code available at `https://github.com/gregorkb/spaddinf` and obtain a pre-smoothing estimator of f, denoted as $\widehat{f}^{\text{pre}}$; Second, we apply the local polynomial estimator to the data $\{D_i, \widehat{f}^{\text{pre}}(D_i)\}_{1\leq i\leq n}$, where $\widehat{f}^{\text{pre}}(D_i)$ is used as the outcome. We fit the local linear estimator by the package `nprobust` (Calonico et al., 2019) with the bandwidth h constructed in the same way as as our `DLL` estimator.

We generate the outcome following the model (1) and consider both exactly sparse and approximately sparse settings.

Exactly sparse. We set the first six functions as follows and $g_j = 0$ for $6 \leq j \leq p$,

$$\begin{aligned} &f(d) = 1.5\sin(d) \quad &&g_1(x) = 2\exp(-x/2) \quad &&g_2(x) = (x-1)^2 - 25/12 \\ &g_3(x) = x - 1/3 \quad &&g_4(x) = 0.75x \quad &&g_5(x) = 0.5x. \end{aligned} \tag{33}$$

More complicated relationships often exist in real life and the additive model might not be exactly sparse. We further introduce an approximately sparse setting.

Approximately sparse. We set f and $\{g_j\}_{1\leq j\leq 5}$ as in (33), generate $\{g_j\}_{6\leq j\leq 14}$ as

$$g_6(x) = 0.4x \quad g_7(x) = 0.3x \quad g_8(x) = 0.2x \quad g_9(x) = 0.1x$$

$$g_{10}(x) = 0.1\sin(2\pi x) \quad g_{11}(x) = 0.2\cos(2\pi x) \quad g_{12}(x) = 0.3\sin^2(2\pi x)$$

$$g_{13}(x) = 0.4\cos^3(2\pi x) \quad g_{14}(x) = 0.5\sin^3(2\pi x)$$

and generate $\{g_j\}_{15\leq j\leq p}$ as linear functions with $g_j(x) = x/(j-1)$.

In addition, we explore the finite-sample performance for different non-linear functions by switching the role of f and g_1 function; see the results in Section F.1 in the appendix.

4.1 Comparison with Plug-in and Oracle Estimators

In the following, we compare our proposed `DLL` estimator with the plug-in(`Plug`) and oracle(`Orac`) estimators. We consider four different settings for generating D_i and X_i, where the independence assumption between X_i and δ_i in (A2) is violated in Settings 3 and 4.

Setting 1. We generate $(D_i, X_i^{\intercal})^{\intercal}$ following the multivariate Normal distribution $N(\mu, \Sigma)$, where $\mu_j = -0.25$ for $1 \leq j \leq p+1$ and $\Sigma \in \mathbb{R}^{(p+1)\times(p+1)}$ is a toeplitz covariance matrix with $\Sigma_{jj} = 1$ for $1 \leq j \leq p+1$ and for $1 \leq j \neq l \leq p+1$,

$$\Sigma_{j,l} = 0.7\cdot\mathbf{1}(|j-l| = 1) + 0.5\cdot\mathbf{1}(|j-l| = 2) + 0.3\cdot\mathbf{1}(|j-l| = 3) + \frac{p - |j-l|}{10(p-4)}\mathbf{1}(|j-l| \geq 4)$$

For $|j-l| \geq 4$, the correlation gradually decays from 0.1 to 0.

Setting 2. $(D_i, X_i^{\intercal})^{\intercal}$ is generated in the same way as in Setting 1. With G denoting the CDF of $N(-0.25, 1)$, we generate the outcome model as:

$$Y_i = f(G(D_i)) + \sum_{j=1}^{p} g_j(G(X_{i,j})) + \epsilon_i \quad \text{for} \quad 1 \leq i \leq n$$

The main difference from Setting 1 is to apply a quantile transformation to D_i and $\{X_{i,j}\}_{1\leq j\leq p}$ before applying the additive model transformation. The goal is to make inference for $(f^*)'(a_0)$ with $f^* = f \circ G$. We generate the additive model following (33) but set $f(d) = -1.5\sin(\pi d)$, $g_1(x) = 2\exp(-x)$, $g_4(x) = x^3 - 1/2$, $g_5(x) = x/(1+x)$.

Setting 3. We generate $(D_i^0, (X_i^0)^\intercal)^\intercal$ following $N(\mu, \Sigma)$ with the same μ and Σ as in Setting 1. We define $D_i = 5(G(D_i^0) - 0.5)$ and $X_{i,j} = 5(G(X_{i,j}^0) - 0.5)$ for $1 \leq j \leq p$, with G denoting the CDF of $N(-0.25, 1)$. The marginal distributions of D_i and $X_{i,j}$ are Uniform$(-2.5, 2.5)$ and D_i is correlated with $\{X_{i,j}\}_{1\leq j\leq p}$.

Setting 4. We generate $(D_i, X_i^\intercal)^\intercal$ following a centered multivariate t distribution with the same covariance matrix Σ as in Setting 1. The freedom degree varies across $\{10, 15\}$.

We fix the dimension $p = 1500$ and vary the sample size n across {500, 1000, 1500, 2000}. The evaluation points a_0 are {-1.25, -0.5, 0.1, 0.25, 1}. For Setting 1, we generate the outcome using both exactly and approximately sparse models; for Settings 2 to 4, we only consider the exactly sparse outcome model. We generate the simulation data 500 times and then use the following metrics to compare these methods: 1. Bias, the absolute difference between the average of the 500 point estimates and the true value; 2. Root Mean Square Error (RMSE); 3. Standard Error (SE), the empirical standard deviation of the 500 point estimates; 4. Coverage, the empirical coverage out of 500 simulations; 5. Length, the average length of the constructed confidence interval (CI). In Table 1, we compare our proposed `DLL` with `Plug`, and `Orac` across four simulation settings and we take an average of the metrics across different sample sizes, evaluation functions, and evaluation points.

We summarize the results in Table 1. For the `Plug` estimator, the bias component is a dominating term in RMSE, while our `DLL` estimator is effective in bias correction. The RMSE of our proposed `DLL` estimator is similar to that of the oracle estimator, which is uniformly smaller than that of the `Plug` estimator. The coverage error is computed as the absolute difference between the empirical coverage and 95%; in most cases, the coverage error results from the undercoverage. The CIs based on the `Plug` estimator are in general undercoverage while our proposed CIs achieve the desired coverage level. Our proposed CI is of a similar length to the length of the oracle CI.

	Bias Percentage			RMSE Ratio		SE			Coverage Error			Length Ratio	
Setting	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`
1	0.128	0.407	0.090	1.152	1.248	0.293	0.286	0.269	1.05%	4.19%	0.88%	1.157	1.103
2	0.133	0.563	0.045	1.057	1.267	0.350	0.344	0.337	1.06%	7.46%	0.80%	1.045	1.010
3	0.065	0.280	0.037	1.049	1.080	0.502	0.497	0.479	0.77%	1.94%	0.81%	1.050	1.031
4	0.151	0.520	0.049	1.034	1.213	0.320	0.316	0.316	1.15%	6.72%	0.91%	1.037	0.986

Table 1: Comparison of `DLL`, plug-in (`Plug`), and oracle (`Orac`) estimators. For each setting, metrics are averaged over the total 40 combinations of $n \in \{500, 1000, 1500, 2000\}$, $f(d) \in \{\sin(d), \exp(d)\}$, and $a_0 \in \{-1.25, -0.5, 0.1, 0.25, 1\}$ with $p = 1500$. The columns indexed with "Bias Percentage" report the percentage of the bias out of RMSE; the columns indexed with "RMSE Ratio" report the ratio of RMSE to the oracle estimator's RMSE; the columns indexed with "SE" report the empirical standard error; the columns indexed with "Coverage Error" report the absolute difference between the empirical coverage and 95%; the columns indexed with "Length Ratio" report the ratio of the CI length to the length of the CI based on the oracle estimator.

In Table 2, we report the detailed simulation results for Settings 1 to 4 with $a_0 \in \{0.1, 0.25\}$ and $n \in \{500, 1000, 1500\}$, and the complete simulation results are presented in Section F.1 in the appendix. The results are consistent with the observations reported in Table 1: our proposed CI achieves the desired coverage and has a similar length to the oracle CI. In addition, the coverage improvement of our proposed CI over the `Plug` estimator can be quite substantial as our `DLL` estimator effectively corrects the bias. For Settings 3 and 4, our proposed method is still effective even if the independence assumption required in Condition (A2) is violated.

4.2 Comparison with the ReSmoothing Method

We compare `DLL` with the ReSmoothing estimator (`RS`) in Gregory et al. (2021). Gregory et al. (2021) does not directly assume the sparse linear or additive treatment model as our current paper and we adopt the simulation of Gregory et al. (2021). For dimension $(p+1) \in \{50, 150\}$ and sample size $n \in \{100, 1000\}$, generate n iid data points as model (1). Set f, g_1, g_2, and g_3 as:

$$f(d) = -\sin(2d);\; g_1(x) = x^2 - \frac{25}{12};\; g_2(x) = x;\; g_3(x) = e^{-x} - \frac{2}{5}\sinh(\frac{5}{2})$$

and $g_j(x) = 0$ for $4 \leq j \leq p$. For each $i = 1, \cdots, n$, generate $(D_i, X_i^{\intercal})^{\intercal} \in \mathbb{R}^{p+1}$ with each dimension marginally following Uniform distribution on $[-2.5, 2.5]$ and the correlation between two different dimensions $1 \leq j, j', \leq p$ of $(D_i, X_i^{\intercal})^{\intercal}$ is $r^{|j-j'|}$, where r varies in $\{0, 0.1, 0.3, 0.5\}$. The evaluation points $a_0 \in \{-1, 0.5\}$, and f, g_1, g_2, g_3, g_4 take turns to be the function of interest. In this setting, the conditional mean model of D_i given X_i is non-linear, violating our treatment model (11). However, we observe that our proposal still effectively works in this setting and we further explore the robustness of our method against non-linear treatment model in Section 4.3.

To obtain `RS` point estimator, we follow Gregory et al. (2021) and apply the local linear estimator to $\{D_i, \widehat{f}^{\text{pre}}(D_i)\}_{1 \leq i \leq n}$ with the presmoothing estimators $\{\widehat{f}^{\text{pre}}(D_i)\}_{1 \leq i \leq n}$, where $\{\widehat{f}^{\text{pre}}(D_i)\}_{1 \leq i \leq n}$, are constructed by the authors' original code available at `https://github.com/gregorkb/spaddinf` with their default choices of the tuning parameters. Since Gregory et al. (2021) did not directly provide full implementation details of the uncertainty quantification associated with estimating $f'(a_0)$ in their simulation studies, we estimate the standard error of the `RS` estimator in an oracle way. Particularly, we compute the `RS` estimator over 500 simulations and compute the corresponding sample standard error based on the 500 `RS` estimates. Such an Oracle standard error estimator can be viewed as a favorable implementation of uncertainty quantification of the `RS` estimator and cannot be implemented for real data analysis. We then construct the oracle confidence interval, denoted as `OraRS`, by assuming the asymptotic normality of the `RS` estimator. As reported in Table 3, when the sample size is small (n=100), the `OraRS` confidence interval does not achieve the desired coverage level since it has a large bias. For the large sample size (n=1000), the bias of the `OraRS` estimator is reduced and the `OraRS` confidence interval achieves the desired coverage level. In contrast, our method has a smaller bias, and our proposed CI achieves the desired coverage level for both $n = 100$ and $n = 1000$. See Section F.4 in the appendix for full simulation results.

Setting 1: approximately sparse

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
0.10	1.49	500	0.21	0.46	0.02	0.39	0.38	0.39	0.91	0.72	0.94	1.51	1.45	1.49
		1000	0.07	0.31	0.00	0.34	0.33	0.33	0.93	0.83	0.93	1.31	1.27	1.27
		1500	0.05	0.26	0.01	0.31	0.30	0.29	0.94	0.86	0.95	1.18	1.15	1.15
0.25	1.45	500	0.20	0.45	0.00	0.41	0.39	0.39	0.91	0.77	0.94	1.56	1.50	1.55
		1000	0.07	0.31	0.01	0.35	0.34	0.35	0.94	0.83	0.94	1.35	1.32	1.32
		1500	0.07	0.27	0.03	0.31	0.31	0.30	0.96	0.84	0.96	1.22	1.19	1.18

Setting 2: exactly sparse

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
0.10	0.74	500	0.19	0.29	0.03	0.38	0.37	0.39	0.92	0.86	0.96	1.51	1.44	1.53
		1000	0.15	0.25	0.06	0.32	0.31	0.32	0.93	0.88	0.95	1.30	1.24	1.28
		1500	0.11	0.20	0.03	0.31	0.30	0.31	0.94	0.88	0.93	1.17	1.13	1.15
0.25	0.94	500	0.21	0.31	0.06	0.40	0.38	0.40	0.92	0.86	0.95	1.55	1.48	1.58
		1000	0.18	0.28	0.09	0.32	0.31	0.32	0.92	0.87	0.95	1.34	1.28	1.33
		1500	0.12	0.21	0.04	0.32	0.31	0.31	0.92	0.86	0.94	1.22	1.17	1.20

Setting 3: exactly sparse

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
0.10	1.49	500	0.12	0.24	0.01	0.71	0.70	0.69	0.94	0.92	0.93	2.81	2.73	2.60
		1000	0.05	0.19	0.05	0.63	0.63	0.60	0.95	0.93	0.95	2.42	2.39	2.26
		1500	0.04	0.15	0.02	0.57	0.56	0.55	0.96	0.94	0.95	2.19	2.16	2.05
0.25	1.45	500	0.08	0.21	0.00	0.72	0.71	0.68	0.94	0.93	0.94	2.79	2.73	2.59
		1000	0.04	0.17	0.03	0.62	0.62	0.58	0.95	0.94	0.95	2.41	2.38	2.25
		1500	0.03	0.15	0.02	0.56	0.55	0.52	0.95	0.93	0.94	2.18	2.14	2.04

Setting 4: exactly sparse with df=10

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
0.10	1.49	500	0.17	0.36	0.04	0.34	0.33	0.32	0.92	0.78	0.95	1.33	1.27	1.32
		1000	0.10	0.28	0.03	0.29	0.28	0.28	0.94	0.80	0.95	1.13	1.06	1.06
		1500	0.06	0.23	0.02	0.24	0.24	0.23	0.96	0.83	0.97	1.00	0.94	0.92
0.25	1.45	500	0.18	0.38	0.05	0.35	0.35	0.35	0.89	0.76	0.94	1.34	1.28	1.34
		1000	0.10	0.29	0.01	0.29	0.28	0.28	0.93	0.78	0.93	1.14	1.07	1.08
		1500	0.06	0.24	0.01	0.24	0.24	0.24	0.96	0.83	0.96	1.01	0.95	0.94

Table 2: Comparison of `DLL`, the plug-in (`Plug`), and oracle (`Orac`) estimators for Settings 1 to 4, across different sample sizes n and evaluation points a_0 with $p = 1500$. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias", "RMSE" and "SE" report the absolute bias, the root mean square error, and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

In addition, we compare `DLL` and `RS` in Setting 1 of the current paper; our method outperforms the `RS` estimator with a much smaller bias and desired coverage; see Section F.4 in the appendix for the results.

Bias in the Setting of Gregory et al. (2021): a_0 = -1

			$f'(a_0)$		$g_1'(a_0)$		$g_2'(a_0)$		$g_3'(a_0)$		$g_4'(a_0)$	
n	p	r	DLL	OraRS	DLL	OraRS	DLL	OraRS	DLL	OraRS	DLL	OraRS
100	50	0.1	0.20	0.64	0.13	0.94	0.03	0.57	0.14	0.80	0.00	0.04
		0.5	0.14	0.61	0.13	0.91	0.05	0.94	0.13	1.06	0.01	0.11
	150	0.1	0.22	0.78	0.20	1.13	0.01	0.65	0.23	0.94	0.01	0.00
		0.5	0.30	0.61	0.08	0.96	0.22	0.91	0.01	1.24	0.08	0.10
1000	50	0.1	0.02	0.03	0.03	0.08	0.05	0.06	0.01	0.05	0.02	0.03
		0.5	0.05	0.08	0.02	0.09	0.01	0.16	0.09	0.12	0.03	0.04
	150	0.1	0.05	0.03	0.05	0.09	0.02	0.02	0.11	0.10	0.02	0.04
		0.5	0.02	0.01	0.01	0.16	0.03	0.18	0.03	0.18	0.00	0.04

Coverage in the Setting of Gregory et al. (2021): a_0 = -1

			$f'(a_0)$		$g_1'(a_0)$		$g_2'(a_0)$		$g_3'(a_0)$		$g_4'(a_0)$	
n	p	r	DLL	OraRS	DLL	OraRS	DLL	OraRS	DLL	OraRS	DLL	OraRS
100	50	0.1	0.96	0.79	0.93	0.83	0.95	0.90	0.96	0.89	0.96	0.95
		0.5	0.94	0.90	0.94	0.88	0.93	0.76	0.94	0.83	0.94	0.93
	150	0.1	0.92	0.87	0.93	0.74	0.94	0.88	0.95	0.87	0.96	0.94
		0.5	0.93	0.92	0.92	0.84	0.92	0.73	0.94	0.79	0.93	0.95
1000	50	0.1	0.96	0.95	0.97	0.95	0.96	0.94	0.97	0.95	0.96	0.95
		0.5	0.96	0.96	0.95	0.96	0.95	0.93	0.95	0.95	0.94	0.94
	150	0.1	0.94	0.95	0.97	0.95	0.95	0.94	0.96	0.93	0.94	0.95
		0.5	0.96	0.95	0.94	0.95	0.95	0.95	0.96	0.95	0.94	0.94

CI Length in the Setting of Gregory et al. (2021): a_0 = -1

			$f'(a_0)$		$g_1'(a_0)$		$g_2'(a_0)$		$g_3'(a_0)$		$g_4'(a_0)$	
n	p	r	DLL	OraRS	DLL	OraRS	DLL	OraRS	DLL	OraRS	DLL	OraRS
100	50	0.1	4.72	1.82	4.76	3.63	4.82	2.78	4.73	4.20	4.75	1.58
		0.5	4.75	2.63	4.91	4.20	4.91	2.38	4.83	4.16	4.98	2.14
	150	0.1	4.51	2.05	4.55	3.44	4.62	2.50	4.58	4.27	4.52	1.53
		0.5	4.58	2.34	4.55	3.79	4.53	2.08	4.58	4.17	4.57	2.22
1000	50	0.1	2.85	3.19	2.86	3.39	2.86	2.12	2.87	3.51	2.84	1.90
		0.5	2.86	2.99	2.86	3.60	2.87	2.29	2.88	3.55	2.86	2.06
	150	0.1	2.85	3.12	2.84	3.56	2.84	2.21	2.87	3.52	2.84	1.99
		0.5	2.87	3.03	2.89	3.50	2.89	2.55	2.89	3.81	2.88	2.01

Table 3: Comparison of bias coverage, and CI length of DLL, ReSmoothing (OraRS) in the setting of Gregory et al. (2021), across different sample sizes n, dimension of covariates p, and the correlation parameter r. The f, g_1, g_2, g_3, and g_4 represent the functions of interest to estimate their derivatives at a_0. The entries of the table represent the bias across 500 simulations.

4.3 Linear Treatment Model vs Non-linear Treatment Model

In this section, we test the performance of the generalized DLL estimator proposed in Section 2.6, which decorrelates with the sparse additive model. The estimator is referred to as DLL-S. The main difference between DLL and DLL-S is, DLL proposed in Section 2.2 uses the Lasso algorithm to fit the treatment model while DLL-S uses the sparse additive model. To generate a linear treatment model between D_i and X_i, we use Setting 1 where D_i and X_i are jointly multivariate Normal. To generate a non-linear treatment model, we generate $\{X_i\}_{1 \le i \le n}$ following the multivariate Normal distribution in Setting 1, but generate D_i as: $D_i = -0.5\exp(-X_{i,1}/2) + 0.5\sin(X_{i,2}) + 0.25X_{i,3}^2 - 0.5X_{i,4} - 0.25X_{i,5}^2 + 0.5\cos(X_{i,6}) -$

$0.25\exp(-X_{i,7}/2) + 0.25X_{i,8} + \delta_i$ with $\delta_i \sim N(0, 0.5)$. The outcome model is generated following the exactly sparse model in (33). The results are reported in Table 4.

Linear Treatment Model, exactly sparse: $f(d) = 1.5\sin(d)$

			Bias				Coverage				Length			
a_0	True	n	DLL	DLL-S	Plug	Orac	DLL	DLL-S	Plug	Orac	DLL	DLL-S	Plug	Orac
0.10	1.49	500	0.18	0.22	0.42	0.02	0.89	0.89	0.76	0.95	1.49	1.48	1.44	1.49
		1000	0.10	0.14	0.34	0.03	0.95	0.94	0.82	0.95	1.30	1.30	1.26	1.27
		1500	0.05	0.08	0.26	0.00	0.95	0.94	0.84	0.96	1.18	1.18	1.14	1.15
		2000	0.05	0.06	0.23	0.02	0.96	0.95	0.88	0.94	1.11	1.11	1.09	1.08
0.25	1.45	500	0.19	0.23	0.43	0.03	0.90	0.89	0.77	0.95	1.54	1.53	1.47	1.54
		1000	0.09	0.13	0.32	0.01	0.94	0.94	0.83	0.95	1.35	1.34	1.30	1.31
		1500	0.05	0.08	0.25	0.00	0.95	0.95	0.86	0.95	1.22	1.22	1.19	1.19
		2000	0.04	0.05	0.22	0.00	0.95	0.95	0.87	0.96	1.15	1.15	1.12	1.12

Non-linear Treatment Model, exactly sparse: $f(d) = 1.5\sin(d)$

			Bias				Coverage				Length			
a_0	True	n	DLL	DLL-S	Plug	Orac	DLL	DLL-S	Plug	Orac	DLL	DLL-S	Plug	Orac
0.10	1.49	500	0.31	0.23	0.37	0.01	0.84	0.91	0.78	0.95	1.40	1.44	1.34	1.50
		1000	0.20	0.12	0.25	0.00	0.87	0.93	0.83	0.95	1.24	1.24	1.18	1.25
		1500	0.18	0.10	0.23	0.03	0.92	0.95	0.88	0.96	1.12	1.14	1.07	1.12
		2000	0.13	0.08	0.17	0.00	0.91	0.94	0.87	0.94	1.05	1.05	1.01	1.05
0.25	1.45	500	0.33	0.23	0.39	0.06	0.80	0.87	0.74	0.94	1.34	1.36	1.28	1.42
		1000	0.19	0.12	0.24	0.01	0.90	0.92	0.85	0.95	1.17	1.18	1.12	1.19
		1500	0.16	0.10	0.21	0.02	0.92	0.95	0.87	0.95	1.06	1.07	1.02	1.06
		2000	0.13	0.08	0.18	0.00	0.91	0.95	0.87	0.95	1.00	0.99	0.95	0.99

Table 4: Comparison of `DLL`, `DLL-S`, plug-in (`Plug`), and oracle (`Orac`) for the linear and non-linear treatment models, across different sample sizes n and evaluation points a_0. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" report the absolute bias; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

In Table 4, we observe that in the linear treatment model, `DLL` and `DLL-S` have a similar performance in terms of bias correction and empirical coverage. For the highly non-linear treatment model, `DLL-S` improves the performance of `DLL` in terms of both bias and coverage. Even with a misspecified treatment model, the `DLL` estimator still corrects the bias and outperforms the `Plug` estimator. See Section F.1 in the appendix for results with additional evaluation points.

5. Real Data Analysis

The Motif Regression has important applications to biology, which studies the effect of the motif candidates' matching scores on the gene expression level (Yuan et al., 2007; Beer and Tavazoie, 2004; Conlon et al., 2003; Das et al., 2004). Motifs are the DNA sequences bound to transcription factors, which control the transcription activities, e.g., gene expressions (Yuan et al., 2007). The matching score of a motif describes the abundance of occurrence, that is, how well the motif is represented in the upstream regions of the genes. A gene's expression level can be well-predicted by the matching scores of a set of motifs (Yuan et al., 2007; Beer and Tavazoie, 2004; Conlon et al., 2003; Das et al., 2004). The data set consists of the expression values of $n = 2587$ genes and the scores of $p + 1 = 666$ motifs. For our analysis, the outcome $\{Y_i\}_{1\leq i\leq 2587}$ denote the gene expression level and $\{(D_i, X_i^\intercal)^\intercal\}_{1\leq i\leq 2587}$ are the matching scores of the 666 motifs.

We define an index subset for the motifs $\mathcal{I} = \{37, 51, 53, 73, 75, 76, 138, 199, 586, 665\} \subset \{1, \cdots, 666\}$. To demonstrate our method, we choose one index from $\mathcal{I}$ and set the corresponding motif score as the variable of interest D and the remaining 665 motif scores as the baseline covariates. We compute its sample mean and standard error for a chosen variable of interest. We choose three different evaluation points a_0: mean, mean + standard error, mean - standard error. To demonstrate our method, we compare it with the existing inference method for the high-dimensional linear model, which assumes the linear and constant effect. Specifically, we apply the `LF()` function in the R package `SIHR` (Rakshit et al., 2021) and denote the corresponding estimator as `SIHR`. We report the comparison in Figure 1.

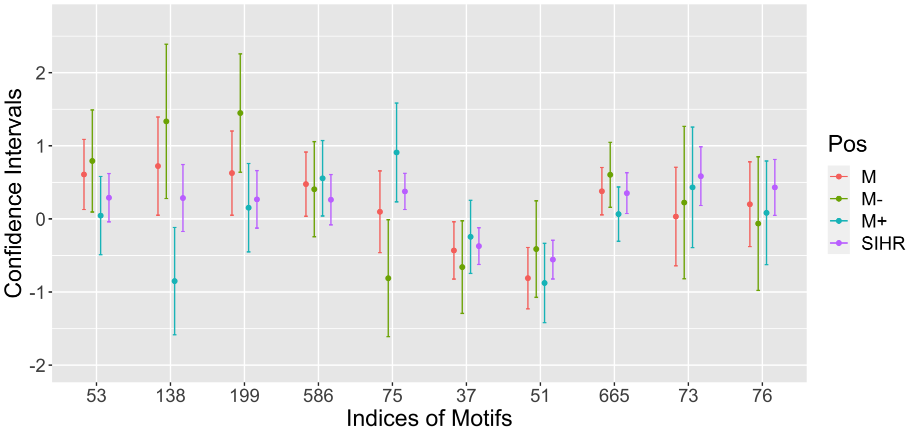

Figure 1: Confidence intervals for $f'(a_0)$ by `DLL` and `SIHR`. "M","M+" and "M-" represent the `DLL` estimator for $f'(a_0)$ with a_0 set as mean, mean + standard error and mean - standard error, respectively. `SIHR` represents inference for the constant effect in the high-dimensional linear model.

Figure 1 demonstrates several interesting observations. First of all, for the motifs with indexes 53, 138, 199, 586, we observe that these motifs do not have significant effects if we assume the model to be linear and apply the `SIHR` package. In contrast, if we assume the additive model, our proposed `DLL` estimator shows that they have non-linear effects; for example, for the motif 138, the CIs by `DLL` at M and M- are above zero, and the CI by `DLL` at M+ is below zero, while the CI by `SIHR` covers 0.

Second, for the motif 75, we observe that the `SIHR` package leads to a significant positive linear effect; however, after applying our proposed `DLL` method, the effects vary across different evaluation points: the CI at M covers 0, the CI at M- is below 0 and the CI at M+ is above 0. The above two observations indicate the relationship between gene expression level and motifs might be highly non-linear.

Lastly, for motifs with indexes 37, 51, 665, they have significant effects for both `SIHR` and `DLL` (at two evaluation points) and their effects are very similar assuming either linear or addtive model; for motifs with indexes 73 and 76, the effects are significant for linear model, but not significant for additive model.

We design a semi-real simulation study to further compare the finite-sample performance of our proposed DLL method and the SIHR method. We keep the data $\{D_i, X_i\}_{1\leq i\leq 2587}$ the same as the real data. After analyzing the original real data, we construct the noise level estimator $\widehat{\sigma}^2$ and $\widehat{f}$ and $\widehat{g}$. We simulate the synthetic response variable $Y_i^{syn} = \widehat{f}(D_i) + \widehat{g}(X_i) + \bar{\epsilon}_i$ for $1 \leq i \leq 2587$ with the i.i.d. regression error terms $\{\bar{\epsilon}_i\}_{1\leq i\leq 2587}$ following $N(0, \widehat{\sigma}^2)$. We repeat the simulation 500 times and evaluate DLL on the same three evaluation points as in the real data analysis.

We compare the results with SIHR and report the comparison in Table 5. The SIHR method, which assumes the linear outcome model, suffers from low empirical coverage for some motifs. Our proposed CIs by the DLL estimators achieve the desired coverage levels in most settings.

6. Conclusion and Discussion

We have proposed the decorrelated local linear estimator to mitigate the error caused by estimating the unknown nuisance functions in the high-dimensional additive model. We have established the asymptotic normality of our proposed estimator. We demonstrate the validity of the theoretical results in moderate sample sizes and provide practical recommendations for the algorithm implementation. Our proposed decorrelation idea is a novel and computationally efficient method designed for bias correction in non-parametric models. An interesting future research direction is extending our proposed method to accommodate other kernel functions and higher-order local polynomials. In addition to inference for $f'(a_0)$, there are other interesting statistical inference problems in the high-dimensional additive model, such as the significance test $H_0 : f = 0$. We leave these problems for future research.

Acknowledgments

We thank Dr. Lukas Meier for sharing the motif regression data. A large proportion of this research was conducted while Z. Guo was working at Rutgers University. The research of C. Zhang was supported in part by NSF Grants DMS-1721495, IIS-1741390, CCF-1934924, and DMS-2052949.

References

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

Michael A Beer and Saeed Tavazoie. Predicting gene expression from sequence. *Cell*, 117 (2):185–198, 2004.

Part I: Bias and Standard Error

	Bias				SE			
Motif	M	M-	M+	SIHR	M	M-	M+	SIHR
37	0.04	0.11	0.28	0.08	0.21	0.56	0.54	0.13
51	0.14	0.15	0.08	0.07	0.24	0.34	0.29	0.13
53	0.17	0.15	0.05	0.02	0.24	0.34	0.27	0.16
73	0.13	0.15	0.16	0.21	0.33	0.49	0.40	0.21
75	0.17	0.23	0.20	0.07	0.25	0.39	0.34	0.13
76	0.11	0.09	0.03	0.10	0.33	0.48	0.42	0.19
138	0.03	0.06	0.16	0.01	0.32	0.49	0.38	0.22
199	0.20	0.18	0.11	0.10	0.28	0.42	0.30	0.18
586	0.06	0.04	0.16	0.05	0.23	0.35	0.28	0.17
665	0.12	0.01	0.13	0.12	0.23	0.33	0.28	0.15

Part II: Coverage and Length

	Coverage				Length			
Motif	M	M-	M+	SIHR	M	M-	M+	SIHR
37	0.93	0.95	0.94	0.87	0.76	2.20	2.11	0.48
51	0.91	0.95	0.94	0.91	0.87	1.30	1.08	0.51
53	0.92	0.94	0.95	0.95	0.93	1.31	1.02	0.64
73	0.92	0.94	0.86	0.77	1.21	1.85	1.44	0.77
75	0.91	0.91	0.91	0.89	0.97	1.39	1.18	0.48
76	0.94	0.97	0.95	0.91	1.25	1.92	1.56	0.73
138	0.95	0.95	0.92	0.95	1.24	1.80	1.28	0.87
199	0.92	0.94	0.95	0.93	1.15	1.60	1.26	0.75
586	0.94	0.96	0.92	0.94	0.88	1.28	1.03	0.66
665	0.92	0.94	0.94	0.83	0.86	1.23	1.08	0.53

Table 5: Comparison of DLL and SIHR in the semi-real simulation study. The columns indexed with "M","M+" and "M-" report the performance of our proposed DLL inference methods for $f'(a_0)$, with a_0 set as mean, mean + standard error, and mean - standard error, respectively. SIHR refers to the high-dimensional inference methods assuming the linear model. The columns indexed with "Bias", and "SE" report the absolute bias, and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length. For SIHR, the bias is taken as the minimal bias across the three evaluation points and the coverage is taken as the maximal coverage across the three evaluation points.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

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.

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.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510, 1989.

Jorge Luis Ojeda Cabrera. *locpol: Kernel Local Polynomial Regression*, 2018. URL `https://CRAN.R-project.org/package=locpol`. R package version 0.7-0.

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.

Tianxi Cai, T Tony Cai, and Zijian Guo. Optimal statistical inference for individualized treatment effects in high-dimensional models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):669–719, 2021.

Sebastian Calonico, Matias D. Cattaneo, and Max H. Farrell. nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. *Journal of Statistical Software*, 91(8):1–33, 2019.

David Card. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160, 2001.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.*, 7(1): 649–688, 2015.

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.

William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

Erin M Conlon, X Shirley Liu, Jason D Lieb, and Jun S Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences*, 100(6):3339–3344, 2003.

Debopriya Das, Nilanjana Banerjee, and Michael Q Zhang. Interacting models of cooperative gene regulation. *Proceedings of the National Academy of Sciences*, 101(46):16234–16239, 2004.

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.

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.

Jianqing Fan. Design-adaptive nonparametric regression. *Journal of the American statistical Association*, 87(420):998–1004, 1992.

Jianqing Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21(1):196–216, 1993.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. CRC Press, Boca Raton, Florida, 1996.

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Theo Gasser and Hans-Georg Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11(3):171–185, 1984.

Karl Gregory, Enno Mammen, and Martin Wahl. Statistical inference in sparse high-dimensional additive models. *The Annals of Statistics*, 49(3):1514–1536, 2021.

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.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1 (3):297–318, 1986.

Tristen Hayfield and Jeffrey S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5):1–32, 2008.

M.A. Hernan and J.M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, Boca Raton, Florida, 2025.

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.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, Cambridge, UK, 2015.

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.

Haoming Jiang, Yukun Ma, Han Liu, Kathryn Roeder, Xingguo Li, and Tuo Zhao. *SAM: Sparse Additive Modelling*, 2021. URL `https://CRAN.R-project.org/package=SAM`. R package version 1.1.3.

Chris AJ Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, 15(4):1548–1562, 1987.

Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010.

Damian Kozbur. Inference in additively separable models with a high-dimensional set of conditioning variables. *Journal of Business & Economic Statistics*, 39(4):984–1000, 2021. doi: 10.1080/07350015.2020.1753524.

Junwei Lu, Mladen Kolar, and Han Liu. Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *Journal of the American Statistical Association*, 115 (532):2084–2099, 2020.

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.

Lukas Meier, Sara Van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

Jean D Opsomer. Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 73(2):166–179, 2000.

Judea Pearl. *Causality*. Cambridge university press, Cambridge, UK, 2009.

Gengsheng Qin and Min Tsao. Empirical likelihood based inference for the derivative of the nonparametric regression function. *Bernoulli*, 11(4):715–735, 2005.

Prabrisha Rakshit, T. Tony Cai, and Zijian Guo. Sihr: An r package for statistical inference in high-dimensional linear and logistic regression models. *arXiv preprint arXiv:2109.03365*, 2021.

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.

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.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151, 1986.

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, Cambridge, Massachusetts, 2008.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4): 595–620, 1977.

Taiji Suzuki and Masashi Sugiyama. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *The Annals of Statistics*, 41(3):1381–1405, 2013.

Zhiqiang Tan and Cun-Hui Zhang. Doubly penalized estimation in additive regression with high-dimensional data. *The Annals of Statistics*, 47(5):2567–2600, 2019.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation.* Springer, New York, 2009.

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.

Simon N Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, Boca Raton, Florida, 2017.

David A Wooff. Bounds on reciprocal moments with applications and developments in stein estimation and post-stratification. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(2):362–371, 1985.

Yun Yang and Surya T Tokdar. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43(2):652–674, 2015.

Ming Yuan and Ding-Xuan Zhou. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564–2593, 2016.

Yuan Yuan, Lei Guo, Lei Shen, and Jun S Liu. Predicting gene expression from sequence: A reexamination. *PLOS Computational Biology*, 3(11):1–7, 11 2007.

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.

Shanggang Zhou and Douglas A Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, 10(1):93–108, 2000.

Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113(524):1583–1600, 2018.

Ying Zhu, Zhuqing Yu, and Guang Cheng. High dimensional inference in partially linear models. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 2760–2769, Naha, Japan, 2019.

Appendix Appendix A. Additional Discussions

A.1 Algorithm without Data Swapping

In this section, we present the DLL estimator without data swapping. Different from the DLL estimator with data swapping, we use all the samples, rather than half of them, when fitting the sparse additive model and constructing the decorrelation weights. Following the steps of Algorithm 1, we make a few changes to implement the DLL estimator without data swapping.

In Step 1, implement the sparse additive model as the following optimization problem with $M \geq 1$ and $\lambda > 0$:

$$\{\widehat{\beta}_j\}_{1\leq j\leq p} = \underset{\beta_j\in\mathbb{R}^M,\, 0\leq j\leq p}{\arg\min} \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \sum_{j=0}^{p}\Psi_{i,j}^{\intercal}\beta_j)^2 + \lambda\sum_{j=0}^{p}\sqrt{\beta_j^{\intercal}\left(\frac{1}{n}\sum_{i=1}^{n}\Psi_{i,j}\Psi_{i,j}^{\intercal}\right)\beta_j}. \tag{34}$$

In Step 2, construct the initial estimator $\{\widehat{g}(X_i)\}_{1\leq i\leq n}$ as:

$$\widehat{g}(X_i) = \sum_{j=1}^{p}\Psi_{i,j}^{\intercal}\widehat{\beta}_j \quad \text{and} \quad \widehat{f}(D_i) = \Psi_{i,0}^{\intercal}\widehat{\beta}_0. \tag{35}$$

In Step 5, implement Lasso algorithm as follows with the tuning parameter $\lambda_1 > 0$:

$$\widehat{\gamma} = \underset{\gamma\in\mathbb{R}^p}{\arg\min}\frac{1}{2n}\sum_{i=1}^{n}\left(D_i - X_i^{\intercal}\gamma\right)^2 + \lambda_1\sum_{j=1}^{p}\frac{\|X_j\|_2}{\sqrt{n}}|\gamma_j|,$$

and compute

$$\widehat{\mu}_i = a_0 - X_i^{\intercal}\widehat{\gamma} \quad \text{and} \quad \widehat{\delta}_i = D_i - X_i^{\intercal}\widehat{\gamma} \quad \text{for} \quad 1\leq i\leq n.$$

In Step 6, construct $\{\widehat{l}(X_i)\}_{1\leq i\leq n}$ and the weights $\{\widetilde{W}_i\}_{1\leq i\leq n}$ as:

$$\widetilde{W}_i = (D_i - a_0) - \widehat{l}(X_i) \quad \text{with} \quad \widehat{l}(X_i,\widehat{\gamma}) = \frac{\frac{1}{n}\sum_{j=1}^{n}(\widehat{\delta}_j - \widehat{\mu}_i)\mathbf{1}(|\widehat{\delta}_j - \widehat{\mu}_i|\leq h)}{\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}(|\widehat{\delta}_j - \widehat{\mu}_i|\leq h)}.$$

The other steps are the same as Algorithm 1. We use the same parameter tuning procedures stated in Section 4. We compare our constructed confidence intervals with and without data swapping in Section F.3.

A.2 Double Penalization

In the following, we review the double penalization method (Tan and Zhang, 2019) to construct initial estimators of f and g. This can be viewed as an alternative method to the estimators in (34) and (35). To construct the penalty term, we define the complexity measure of a univariate function f as

$$\mathcal{C}(f) = \lambda_n(\|f\|_n + \rho_n\|f\|_F) \tag{36}$$

where $\lambda_n > 0$, $\rho_n > 0$, $\|f\|_n = (\sum_{i=1}^{n} f^2(D_i))^{1/2}$ denotes the function's empirical L_2 norm, and $\|f\|_F = (\int[f''(t)]^2dt)^{1/2}$ is a measure of the function's smoothness.

Specifically, with the positive tuning parameters $\rho_n > 0$ and $\lambda_n > 0$, we define the initial estimators as

$$\left\{\widehat{f}, \{\widehat{g}_j\}_{1\leq j\leq p}\right\} = \arg\min \frac{1}{n}\sum_{i=1}^{n}[Y_i - f(D_i) - \sum_{j=1}^{p} g_j(X_{ij})]^2 + \mathcal{C}(f) + \sum_{j=1}^{p}\mathcal{C}(g_j),$$

where the complexity measure $\mathcal{C}(\cdot)$ is defined in (36).

A.3 Initial Estimators with Quantile Transformation

We consider the construction of the initial estimator $\widehat{g}$ by applying the quantile transformation to all variables. Particularly, we transform D_i to $\widetilde{D}_i$, with

$$\widetilde{D}_i = \frac{\text{Ordering of } D_i}{n} \in (0, 1];$$

similarly, for $1 \leq j \leq p$, we transform $X_{i,j}$ to $\widetilde{X}_{i,j}$, with

$$\widetilde{X}_{i,j} = \frac{\text{Ordering of } X_{i,j}}{n} \in (0, 1].$$

We construct the initial estimators $\widehat{f}$ and $\widehat{g}$ by applying the sparse additive algorithm to $\{Y_i, \widetilde{D}_i, \widetilde{X}_i\}_{1\leq i\leq n}$. Except for constructing the initial estimator differently, the other steps are the same as those in Algorithm 1. This `DLL` estimator with the extra quantile transformation is refered to as `Trans` and we do not apply the data swapping for `Trans` estimator. We compare the performance of `Trans` with the regular `DLL` estimator in Section F.3; see Tables A10 and A11 for details.

A.4 Further Discussions on the Condition (A3)

In a more general setting, we may plug-in the existing convergence rate of $\mathrm{Err}(\widehat{g})$ and then the condition (30) is reduced to a simultaneous condition on $k := \|\gamma\|_0$ and $s := \|g\|_0$, where $\|g\|_0$ denotes the number of non-zero functions of $\{g_j\}_{1\leq j\leq p}$. Particularly, we follow Tan and Zhang (2019) by assuming that the individual functions $\{g_j\}_{1\leq j\leq p}$ belong to the Sobolev space $\mathcal{W}_2^m$ for $m > 1/2$ and f'' is continuous. We apply Proposition 4 and Theorem 2 in Tan and Zhang (2019) and Corollaries 4 and 5 in Guo and Zhang (2019) and establish that

$$\mathrm{Err}^2(\widehat{g}) \lesssim n^{-\frac{4}{5}} + s \cdot n^{-\frac{2m}{2m+1}} + (s+1) \cdot \log p/n.$$

Then the condition (30) is simplified as

$$k \cdot s \ll n^{\frac{2m}{2m+1}+\frac{3}{5}} \text{ and } \max\{k, s \cdot n^{\frac{1}{2m+1}}\} \ll n \quad \text{up to a polynomial order of } \log p.$$

If we set $m = 2$, then the above sparsity condition is much weaker than the one in Gregory et al. (2021), which requires $s \ll n^{\frac{3}{10}}$ and $k \ll n^{\frac{4}{15}}$ up to a polynomial order of $\log p$.

A.5 Consistent Estimators of σ^2

Similar to the definition of $\mathrm{Err}(\widehat{g})$ in (29), we use $\mathrm{Err}(\widehat{f})$ to denote the accuracy measure of $\widehat{f}$, which is defined as follows: with probability larger than $1 - \min$ for some positive constant $c > 0$,

$$\sqrt{\mathbf{E}_{D_*}(\widehat{f}^a(D_*) - f(D_*))^2 + \mathbf{E}_{D_*}(\widehat{f}^b(D_*) - f(D_*))^2} \lesssim \mathrm{Err}(\widehat{f}), \tag{37}$$

where $\widehat{f}^a$ and $\widehat{f}^b$ are defined in (25) and the expectation is taken with respect to the independent copy D_* of $\{D_i\}_{1\leq i\leq n}$.

Proposition 1 *Suppose that Condition* (A1) *holds and* $\max\{\mathrm{Err}(\widehat{f}), \mathrm{Err}(\widehat{g})\} \to 0$. *Then the estimator* $\widehat{\sigma}^2$ *defined in* (26) *satisfies* $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$.

Proposition 1 shows that our proposed $\widehat{\sigma}^2$ is consistent if both $\widehat{f}$ and $\widehat{g}$ are consistent. The proof of Proposition 1 is presented in Section E.6.

Appendix Appendix B. Notations, Events and Lemmas

We introduce some notations and events, which will be used throughout the proof. Let $q(D_i \mid X_i)$ denote the conditional distribution of D_i given X_i and ϕ denote the density function of the error $\delta_i = D_i - X_i^{\intercal}\gamma$. Since $D_i - X_i^{\intercal}\gamma$ is independent of X_i, we have

$$q(D_i = a_0 \mid X_i) = \phi(a_0 - X_i^{\intercal}\gamma) = \phi(\mu_i) \quad \text{with} \quad \mu_i = a_0 - X_i^{\intercal}\gamma.$$

We express the density function π of the random variable D_i as

$$\pi(a_0) = \mathbf{E}_{X_i}\left[q(D_i = a_0 \mid X_i)\right] = \mathbf{E}_{X_i}\left[\phi\left(a_0 - X_i^{\intercal}\gamma\right)\right].$$

Define the following events,

$$\begin{aligned}
\mathcal{A}_0 &= \left\{\max_{i\in\mathcal{I}_a}\max_{|\delta-\mu_i|\leq r}\frac{|\phi'(\delta)|}{\phi(\mu_i)} \leq C_1(n), \quad \max_{i\in\mathcal{I}_a}\max_{|\delta-\mu_i|\leq r}\frac{|\phi''(\delta)|}{\phi(\mu_i)} \leq C_2(n)\right\},\\
\mathcal{A}_1 &= \left\{\|\widehat{\gamma}^b - \gamma\| \lesssim \sqrt{\frac{k\log p}{n}}, \ \max_{i\in\mathcal{I}_a}|X_i^{\intercal}(\widehat{\gamma}^b - \gamma)| \leq C^*\sqrt{\frac{k\log p\log n}{n}}\right\},\\
\mathcal{A}_2 &= \left\{\mathbf{E}_{X_*}(\widehat{g}^b(X_*) - g(X_*))^2 \lesssim \mathrm{Err}^2(\widehat{g})\right\}.
\end{aligned} \tag{38}$$

By the definitions of $C_1(n)$ and $C_2(n)$ in (27), we have $\mathbf{P}(\mathcal{A}_0^c) \leq \min\{n,p\}^{-c}$ for some positive constant $c > 0$. Throughout the proof, we shall assume $\mathbf{P}(\mathcal{A}_0^c) \ll h\pi(a_0)$ and this will automatically hold in our considered regime.

Define $\mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2$. Theorem 7.2 in Bickel et al. (2009) implies that the Lasso estimator $\widehat{\gamma}^b$ satisfies

$$\mathbf{P}\left(\|\widehat{\gamma}^b - \gamma\| \lesssim \sqrt{\frac{k\log p}{n}}\right) \geq 1 - p^{-c},$$

for some positive constant $c > 0$. Conditioning on the data in $\mathcal{I}_b$, the random variable $X_i^\intercal(\widehat{\gamma}^b - \gamma)/\|\widehat{\gamma}^b - \gamma\|_2$ is sub-gaussian random variable, which implies

$$\mathbf{P}\left(\max_{i\in\mathcal{I}_a}\left|X_i^\intercal(\widehat{\gamma}^b - \gamma)/\|\widehat{\gamma}^b - \gamma\|_2 - \mathbf{E}\left[X_i^\intercal(\widehat{\gamma}^b - \gamma)/\|\widehat{\gamma}^b - \gamma\|_2\right| \mid \mathcal{I}_b\right] \gtrsim \sqrt{\log n} \mid \mathcal{I}_b\right) \leq n^{-c},$$

for some positive constant $c > 0$. Note that

$$\left|\mathbf{E}\left[X_i^\intercal(\widehat{\gamma}^b - \gamma)/\|\widehat{\gamma}^b - \gamma\|_2 \mid \mathcal{I}_b\right]\right| \leq \sqrt{(\widehat{\gamma} - \gamma)^\intercal\Sigma(\widehat{\gamma} - \gamma)}/\|\widehat{\gamma}^b - \gamma\|_2 \leq C.$$

The above two inequalities imply that, there exists a constant $C^* > 0$ independent of n and p such that

$$\mathbf{P}\left(\max_{i\in\mathcal{I}_a}|X_i^\intercal(\widehat{\gamma}^b - \gamma)| \leq C^*\sqrt{\frac{k\log p\log n}{n}}\right) \geq 1 - \min\{n, p\}^{-c},$$

for some positive constant $c > 0$. Together with the definition in (27) and (29), we establish

$$\mathbf{P}\left(\mathcal{A}\right) \geq 1 - \min\{n, p\}^c \text{ for some constant } c > 1. \tag{39}$$

The following lemma states the expectation of terms involved with $K_h(D_i)$, whose proof can be found in Section E.3.

Lemma 1 *Suppose that Condition* (A2) *holds, then we have*

$$\left|\frac{\mathbf{E}\left(K_h(D_i) \mid X_i\right)}{q(a_0 \mid X_i)} - 1\right| \cdot \mathbf{1}_{\mathcal{A}_0} \leq \frac{h^2}{6}C_2(n); \tag{40}$$

$$\left|\frac{\mathbf{E}\left(K_h(D_i)\right)}{\pi(a_0)} - 1\right| \lesssim h^2C_2(n) + \frac{\mathbf{P}(\mathcal{A}_0^c)}{\pi(a_0)}; \tag{41}$$

$$\left|\frac{\mathbf{E}\left[(D_i - a_0)K_h(D_i)\right]}{\pi(a_0)}\right| \leq \frac{1}{3}h^2(C_1(n) + \frac{3}{8}hC_2(n)) + \frac{\mathbf{P}(\mathcal{A}_0^c)}{\pi(a_0)}; \tag{42}$$

$$\left|\frac{\mathbf{E}\left[(D_i - a_0)^2K_h(D_i)\right]}{\frac{1}{3}h^2\pi(a_0)} - 1\right| \lesssim \frac{1}{10}h^2C_2(n) + \frac{\mathbf{P}(\mathcal{A}_0^c)}{h\pi(a_0)}; \tag{43}$$

$$\left|\frac{\mathbf{E}\left(W_i^2K_h(D_i) \mid X_i\right)}{\frac{1}{3}h^2q(a_0 \mid X_i)} - 1\right| \cdot \mathbf{1}_{\mathcal{A}_0} \lesssim h^2[C_1^2(n) + C_2(n)]; \tag{44}$$

$$\left|\frac{\mathbf{E}W_i^2K_h^2(D_i)}{\frac{1}{3}h\pi(a_0)} - 1\right| \lesssim h^2[C_1^2(n) + C_2(n)] + \frac{\mathbf{P}(\mathcal{A}_0^c)}{h\pi(a_0)}; \tag{45}$$

$$\left|\frac{\mathbf{E}W_i(D_i - a_0)K_h(D_i)}{\frac{1}{3}h^2\pi(a_0)} - 1\right| \lesssim h^2[C_1^2(n) + C_2(n)] + \frac{\mathbf{P}(\mathcal{A}_0^c)}{h\pi(a_0)}; \tag{46}$$

$$\left|\mathbf{E}W_i\frac{(D_i - a_0)^2}{2}K_h(D_i) \cdot \mathbf{1}_{\mathcal{A}_0^c}\right| \lesssim h^4\left[C_1(n) + hC_2(n)\right]\pi(a_0) + h^2\mathbf{P}(\mathcal{A}_0^c). \tag{47}$$

The following lemma is about the concentration results for terms involved with $K_h(D_i)$, whose proof can be found in Section E.4.

Lemma 2 *Suppose that Condition* (A2) *holds, then for a sufficiently large* n*, with probability* $1-\exp(-t^2)$*,*

$$c\pi(a_0)\left[1-\frac{t}{\sqrt{nh\pi(a_0)}}\right] \leq \left|\frac{1}{n}\sum_{i=1}^{n} K_h(D_i)\right| \leq C\pi(a_0)\left[1+\frac{t}{\sqrt{nh\pi(a_0)}}\right]; \tag{48}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} W_i K_h(D_i)\right| \lesssim t\sqrt{\frac{h}{n}\pi(a_0)}; \tag{49}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n}(D_i-a_0)K_h(D_i)\right| \lesssim \frac{\pi(a_0)}{3}h^2(C_1(n)+\frac{3}{8}hC_2(n))+\mathbf{P}(\mathcal{A}_0^c)+t\sqrt{\frac{h}{n}\pi(a_0)}; \tag{50}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} W_i(D_i-a_0)K_h(D_i)-\mathbf{E}W_i(D_i-a_0)K_h(D_i)\right| \leq Ch^2\pi(a_0)t\sqrt{\frac{1}{nh\pi(a_0)}}; \tag{51}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} W_i\frac{(D_i-a_0)^2}{2}K_h(D_i)\right| \lesssim h^4\left[C_1(n)+hC_2(n)\right]\pi(a_0)+h^2\mathbf{P}(\mathcal{A}_0^c)+t\sqrt{\frac{h^5}{n}\pi(a_0)}; \tag{52}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} W_i^2K_h^2(D_i)-\mathbf{E}W_i^2K_h^2(D_i)\right| \lesssim t\sqrt{\frac{h\pi(a_0)}{n}}. \tag{53}$$

Combining (48) and (49), we establish that, with probability $1-\exp(-t^2)$,

$$|\bar{\mu}_W| \lesssim t\sqrt{\frac{h}{n\pi(a_0)}}. \tag{54}$$

Appendix Appendix C. Proof of Theorem 2

Recall the definition of $\widehat{W_i}$ in (18). We define an accuracy measure of estimating the decorrelation weights as

$$\mathrm{Err}(\widehat{W}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\widehat{W_i}-(W_i-\bar{\mu}_W)]^2K_h(D_i)} \quad \text{with } \bar{\mu}_W = \frac{\frac{1}{n}\sum_{i=1}^{n} W_iK_h(D_i)}{\frac{1}{n}\sum_{i=1}^{n} K_h(D_i)}. \tag{55}$$

We first introduce the following important intermediate results. With probability larger than $1-\frac{C}{t^2}-\min\{n,p\}^{-c}$ for some $t>1$ and positive constants $C>0, c>0$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} W_i\Delta(X_i)K_h(D_i)\right| \leq Ct\sqrt{h/n}\cdot\mathrm{Err}(\widehat{g}), \tag{56}$$

$$\left|\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_i\Delta(X_i)K_h(D_i)\right| \leq t\left(t\sqrt{\frac{1}{nh^3\pi^2(a_0)}}+\frac{\mathrm{Err}(\widehat{W})}{h^2\pi(a_0)}\right)\mathrm{Err}(\widehat{g}), \tag{57}$$

and

$$\mathrm{Err}(\widehat{W}) \lesssim t\left(\sqrt{\frac{h}{n}}+\sqrt{\frac{k\log p\log n}{n}}h^2\left(C_1^2(n)+C_2(n)\right)\right). \tag{58}$$

A combination of (57) and (58) leads to the bound (32). We shall prove (56), (57), and (58) in Sections C.1, C.2, and C.3, respectively.

C.1 Proof of (56)

Recall the following notations,

- $\mathcal{I}_a$ and $\mathcal{I}_b$ are two disjoint subsets with approximately equal sample size, with $\mathcal{I}_a \cap \mathcal{I}_b$ empty and $\mathcal{I}_a \cup \mathcal{I}_b = \{1, 2, \cdots, n\}$.
- $\widehat{g}^a$ and $\widehat{g}^b$ denote the initial estimator of g based on the data $(X_i, D_i, Y_i)_{i\in\mathcal{I}_a}$ and $(X_i, D_i, Y_i)_{i\in\mathcal{I}_b}$, respectively.

The proof relies on the independence created by data swapping. Define the estimation error as $\Delta^a(X_i) = \widehat{g}^a(X_i) - g(X_i)$ and $\Delta^b(X_i) = \widehat{g}^b(X_i) - g(X_i)$. We write $\mathbf{E}(\cdot \mid \mathcal{I}_a)$,$\mathrm{Var}(\cdot \mid \mathcal{I}_a)$ and $\mathbf{P}(\cdot \mid \mathcal{I}_a)$ as the expectation, variance and probability conditioning on the sample $(X_i, D_i, Y_i)_{i\in\mathcal{I}_a}$, respectively. Similarly, we define $\mathbf{E}(\cdot \mid \mathcal{I}_b), \mathrm{Var}(\cdot \mid \mathcal{I}_b)$ and $\mathbf{P}(\cdot \mid \mathcal{I}_b)$ conditioning on $(X_i, D_i, Y_i)_{i\in\mathcal{I}_b}$. For $1 \leq i \leq n$, we have the following decomposition

$$\frac{1}{n}\sum_{i=1}^{n}W_i\Delta(X_i)K_h(D_i) = \frac{1}{n}\sum_{i\in\mathcal{I}_a}W_i\Delta^b(X_i)K_h(D_i) + \frac{1}{n}\sum_{i\in\mathcal{I}_b}W_i\Delta^a(X_i)K_h(D_i). \tag{59}$$

We will control the first term $\frac{1}{n}\sum_{i\in\mathcal{I}_a}W_i\Delta^b(X_i)K_h(D_i)$ in the following and the second term can be controlled by a similar argument. Since

$$\mathbf{P}\left(|\frac{1}{n}\sum_{i\in\mathcal{I}_a}W_i\Delta^b(X_i)K_h(D_i)| \neq |\frac{1}{n}\sum_{i\in\mathcal{I}_a}W_i\Delta^b(X_i)K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_2}|\right) \leq \min\{n,p\}^{-c},$$

it is sufficient to analyze

$$\frac{1}{n}\sum_{i=1}^{n}W_i\Delta^b(X_i)K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_2},$$

where $\mathcal{A}_0$ and $\mathcal{A}_2$ are defined in (38). The above term has two sources of randomness: the initial estimator $\Delta^b = \widehat{g}^b - g$ and the data $\{X_i, D_i\}_{i\in\mathcal{I}_a}$. Since the randomness of Δ^b is induced from the data $(X_i, D_i, Y_i)_{i\in\mathcal{I}_b}$, the estimation error Δ^b is independent of the data $\{X_i, D_i\}_{i\in\mathcal{I}_a}$.

Since W_i is constructed such that $\mathbf{E}\left[W_iK_h(D_i) \mid X_i\right] = 0$, then we have

$$\mathbf{E}\left(\frac{1}{n}\sum_{i\in\mathcal{I}_a}W_i\Delta^b(X_i)K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_2} \mid \mathcal{I}_b, \{X_i\}_{i\in\mathcal{I}_a}\right) = 0. \tag{60}$$

We control the second order moment as

$$\begin{aligned}
&\mathbf{E}\left(\left(\frac{1}{n}\sum_{i\in\mathcal{I}_a} W_i\Delta^b(X_i)K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_2}\right)^2\right)\\
&=\mathbf{E}\left[\mathbf{E}\left(\left(\frac{1}{n}\sum_{i\in\mathcal{I}_a} W_i\Delta^b(X_i)K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_2}\right)^2\mid\mathcal{I}_b,\{X_i\}_{i\in\mathcal{I}_a}\right)\right]\\
&=\frac{1}{n^2}\sum_{i\in\mathcal{I}_a}\mathbf{E}\left[\mathbf{E}\left(W_i^2K_h^2(D_i)\mid\mathcal{I}_b,\{X_i\}_{i\in\mathcal{I}_a}\right)(\Delta^b(X_i))^2\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_2}\right],
\end{aligned}$$

where the last equality follows from (60). By (44) and the definition of $\mathrm{Err}(\widehat{g})$ in (29), we establish

$$\mathbf{E}\left[\mathbf{E}\left(W_i^2K_h^2(D_i)\mid\mathcal{I}_b,\{X_i\}_{i\in\mathcal{I}_a}\right)(\Delta^b(X_i))^2\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_2}\right]\lesssim h\mathrm{Err}(\widehat{g})^2.$$

Hence we establish

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i\in\mathcal{I}_a} W_i\Delta^b(X_i)K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_2}\right|\le t\sqrt{\frac{h}{n}}\cdot\mathrm{Err}(\widehat{g})\right)\ge 1-\frac{1}{t^2}-\min\{n,p\}^{-c}.$$

By symmetry and the decomposition (59), we establish (56).

C.2 Proof of (57)

We decompose $\frac{1}{n}\sum_{i=1}^n\widehat{W}_i\Delta(X_i)K_h(D_i)$ as

$$\frac{1}{n}\sum_{i=1}^n\left(\widehat{W}_i-(W_i-\bar{\mu}_W)\right)\Delta(X_i)K_h(D_i)+\frac{1}{n}\sum_{i=1}^n W_i\Delta(X_i)K_h(D_i)-\bar{\mu}_W\cdot\frac{1}{n}\sum_{i=1}^n\Delta(X_i)K_h(D_i). \tag{61}$$

By the Cauchy-Schwarz inequality, we have

$$\left|\frac{1}{n}\sum_{i=1}^n\left(\widehat{W}_i-(W_i-\bar{\mu}_W)\right)\Delta(X_i)K_h(D_i)\right|\le\mathrm{Err}(\widehat{W})\cdot\sqrt{\frac{1}{n}\sum_{i=1}^n\Delta(X_i)^2K_h(D_i)}, \tag{62}$$

where $\mathrm{Err}(\widehat{W})$ is defined in (55). Hence, it is sufficient to control $\sqrt{\frac{1}{n}\sum_{i=1}^n\Delta(X_i)^2K_h(D_i)}$. Similar to (59), we have

$$\frac{1}{n}\sum_{i=1}^n\Delta(X_i)^2K_h(D_i)=\frac{1}{n}\sum_{i\in\mathcal{I}_a}\left|\Delta^b(X_i)\right|^2K_h(D_i)+\frac{1}{n}\sum_{i\in\mathcal{I}_b}\left|\Delta^a(X_i)\right|^2K_h(D_i), \tag{63}$$

and it is sufficient to control

$$\frac{1}{n}\sum_{i\in\mathcal{I}_a}\left|\Delta^b(X_i)\right|^2K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_2}.$$

Note that

$$\begin{aligned}&\mathbf{E}\left(\frac{1}{n}\sum_{i\in\mathcal{I}_a}\left|\Delta^b(X_i)\right|^2 K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_2}\mid\mathcal{I}_b,\{X_i\}_{i\in\mathcal{I}_a}\right)\\ =&\frac{1}{n}\sum_{i\in\mathcal{I}_a}\left|\Delta^b(X_i)\right|^2\cdot\mathbf{E}\left[K_h(D_i)\mid X_i\right]\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_2}\lesssim\mathrm{Err}^2(\widehat{g}),\end{aligned}$$

where the last inequality follows from (40) and the bounded conditional density $q(a_0 \mid X_i)$. The above moment bound implies

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\left|\Delta^b(X_i)\right|^2 K_h(D_i)\right|\leq Ct^2\mathrm{Err}^2(\widehat{g})\right)\geq 1-\frac{1}{t^2}-\min\{n,p\}^{-c}.$$

By symmetry and the decomposition (63), we have

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\Delta(X_i)^2 K_h(D_i)\right|\leq Ct^2\mathrm{Err}^2(\widehat{g})\right)\geq 1-\frac{1}{t^2}-\min\{n,p\}^{-c}. \tag{64}$$

Combined with (62), we obtain

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\left(\widehat{W}_i-(W_i-\bar{\mu}_W)\right)\Delta(X_i)K_h(D_i)\right|\lesssim t^2\mathrm{Err}(\widehat{W})\cdot\mathrm{Err}(\widehat{g})\right)\geq 1-\frac{1}{t^2}-\min\{n,p\}^{-c}. \tag{65}$$

Note that

$$\frac{1}{n}\sum_{i=1}^n\left|\Delta(X_i)\right|K_h(D_i)\leq\sqrt{\frac{1}{n}\sum_{i=1}^n K_h(D_i)}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^n\Delta(X_i)^2K_h(D_i)}.$$

Together with (48), (54), and (64), we establish

$$\mathbf{P}\left(\left|\bar{\mu}_W\cdot\frac{1}{n}\sum_{i=1}^n\Delta(X_i)K_h(D_i)\right|\lesssim t\mathrm{Err}(\widehat{g})\sqrt{h/n}\right)\geq 1-\frac{1}{t^2}-\min\{n,p\}^{-c}.$$

Together with (56), (65) and the decomposition (61), we have

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\widehat{W}_i\Delta(X_i)K_h(D_i)\right|\lesssim t\left(t\sqrt{h/n}+\mathrm{Err}(\widehat{W})\right)\mathrm{Err}(\widehat{g})\right)\geq 1-\frac{1}{t^2}-\min\{n,p\}^{-c}.$$

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

C.3 Proof of (58)

In the following, we first establish

$$\mathrm{Err}^2(\widehat{W})\lesssim\frac{1}{n}\sum_{i=1}^n\left(l(X_i,\widehat{\gamma})-l(X_i,\gamma)\right)^2K_h(D_i). \tag{66}$$

Recall that the uncentered weight $\widetilde{W}_i$ is defined in (17) and $\widehat{W}_i$ is the corresponding centered weight defined in (18). Note that

$$\left(\widehat{W}_i - (W_i - \bar{\mu}_W)\right)^2 \lesssim \left(\widetilde{W}_i - W_i\right)^2 + \left(\frac{\frac{1}{n}\sum_{i=1}^n (\widetilde{W}_i - W_i) K_h(D_i)}{\frac{1}{n}\sum_{i=1}^n K_h(D_i)}\right)^2. \tag{67}$$

By Cauchy-Schwarz inequality

$$\left(\frac{1}{n}\sum_{i=1}^n (\widetilde{W}_i - W_i) K_h(D_i)\right)^2 \leq \left(\frac{1}{n}\sum_{i=1}^n (\widetilde{W}_i - W_i)^2 K_h(D_i)\right) \cdot \left(\frac{1}{n}\sum_{i=1}^n K_h(D_i)\right),$$

we have

$$\frac{1}{n}\sum_{i=1}^n \left(\frac{\frac{1}{n}\sum_{i=1}^n (\widetilde{W}_i - W_i) K_h(D_i)}{\frac{1}{n}\sum_{i=1}^n K_h(D_i)}\right)^2 K_h(D_i) \leq \frac{1}{n}\sum_{i=1}^n (\widetilde{W}_i - W_i)^2 K_h(D_i).$$

By applying the above inequality and (67), we obtain

$$\sqrt{\frac{1}{n}\sum_{i=1}^n \left(\widehat{W}_i - (W_i - \bar{\mu}_W)\right)^2 K_h(D_i)} \lesssim \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\widetilde{W}_i - W_i\right)^2 K_h(D_i)}$$
$$+ \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\frac{\frac{1}{n}\sum_{i=1}^n (\widetilde{W}_i - W_i) K_h(D_i)}{\frac{1}{n}\sum_{i=1}^n K_h(D_i)}\right)^2 K_h(D_i)} \lesssim \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\widetilde{W}_i - W_i\right)^2 K_h(D_i)}.$$

By the definitions $W_i = (D_i - a_0) - l(X_i, \gamma)$ and $\widetilde{W}_i = (D_i - a_0) - \widehat{l}(X_i, \widehat{\gamma})$, the above inequality implies (66).

Then the proof of (58) is reduced to establishing an upper bound for

$$\frac{2}{n}\sum_{i=1}^n \left(\widehat{l}(X_i, \widehat{\gamma}) - l(X_i, \gamma)\right)^2 K_h(D_i).$$

We divide the above summation into two parts,

$$\frac{1}{n_a}\sum_{i\in\mathcal{I}_a} \left(\widehat{l}(X_i, \widehat{\gamma}^b) - l(X_i, \gamma)\right)^2 K_h(D_i) + \frac{1}{n_b}\sum_{i\in\mathcal{I}_b} \left(\widehat{l}(X_i, \widehat{\gamma}^a) - l(X_i, \gamma)\right)^2 K_h(D_i).$$

By symmetry, we focus on the first summation

$$\frac{1}{n_a}\sum_{i\in\mathcal{I}_a} \left(\widehat{l}(X_i, \widehat{\gamma}^b) - l(X_i, \gamma)\right)^2 K_h(D_i), \tag{68}$$

and adopt the notation $\widehat{\gamma}^b = \widehat{\gamma}$. We note that

$$\widehat{\delta}_j - \widehat{\mu}_i = \delta_j - \mu_i - a_{ij}, \quad \text{with} \quad a_{ij} = (X_j - X_i)^{\intercal}(\widehat{\gamma} - \gamma).$$

On the event $\mathcal{A}_1$ defined in (38), we have

$$|a_{ij}| \le C^* \sqrt{\frac{\|\gamma\|_0 \log p}{n}} \sqrt{\log n}. \tag{69}$$

To facilitate the discussion, we introduce the following notations,

$$\widehat{\mathbf{I}}_{ij} = \mathbf{1}(|\delta_j - \mu_i - a_{ij}| \le h) \quad \text{and} \quad \mathbf{I}_{ij} = \mathbf{1}(|\delta_j - \mu_i| \le h). \tag{70}$$

The estimator $\widehat{l}(X_i, \widehat{\gamma})$ defined in (16) can be written as

$$\widehat{l}(X_i, \widehat{\gamma}) = \frac{\frac{1}{n_a} \sum_{j \in \mathcal{I}_a} (\delta_j - \mu_i - a_{ij}) \widehat{\mathbf{I}}_{ij}}{\frac{1}{n_a} \sum_{j \in \mathcal{I}_a} \widehat{\mathbf{I}}_{ij}}. \tag{71}$$

The randomness of $\widehat{l}(X_i, \widehat{\gamma})$ comes from the following three parts,

- the noise $\{\delta_j\}_{j \in \mathcal{I}_a}$
- the variable μ_i with $\mu_i = a_0 - X_i^\intercal \gamma$ for the pre-fixed index i
- the estimation error a_{ij}, which depends on X_i, X_j and the initial estimator $\widehat{\gamma}$ computed on the data $\mathcal{I}_b$.

Note that $\{\delta_j\}_{j \in \mathcal{I}_a}$ is independent of the other two random sources. We shall write $\mathbf{E}_{\delta_j}$ as the expectation with respect to δ_j but condition on the other two components X_i and $\mathcal{I}_b$. We use $\mathbf{E}_{\delta_j}\left[\cdot \mid \widehat{\mathbf{I}}_{ij}\right]$ to denote the conditional expectation by only considering the randomness of δ_j. Specifically, for $j \in \mathcal{I}_a$, $\mathbf{E}_{\delta_j}$ and $\mathbf{E}_{\delta_j}\left[\cdot \mid \widehat{\mathbf{I}}_{ij}\right]$ are shorthanded for

$$\mathbf{E}_{\delta_j}[\cdot] = \mathbf{E}[\cdot \mid \{X_i\}_{i \in \mathcal{I}_a}, \mathcal{I}_b] \quad \text{and} \quad \mathbf{E}_{\delta_j}\left[\cdot \mid \widehat{\mathbf{I}}_{ij}\right] = \mathbf{E}\left[\cdot \mid \widehat{\mathbf{I}}_{ij}, \{X_i\}_{i \in \mathcal{I}_a}, \mathcal{I}_b\right].$$

We use $\mathbf{E}_{\delta|\widehat{\mathbf{I}}_{i\cdot}}$ to denote the conditional expectation of $\{\delta_j\}_{j \in \mathcal{I}_a}$ given the events $\widehat{\mathbf{I}}_{i\cdot} = \{\widehat{\mathbf{I}}_{ij}\}_{j \in \mathcal{I}_a}$, by only considering the randomness of $\{\delta_j\}_{j \in \mathcal{I}_a}$, that is

$$\mathbf{E}_{\delta|\widehat{\mathbf{I}}_{i\cdot}}[\cdot] = \mathbf{E}\left[\cdot \mid \widehat{\mathbf{I}}_{i\cdot}, \{X_i\}_{i \in \mathcal{I}_a}, \mathcal{I}_b\right].$$

We compute the following difference,

$$\begin{aligned} &\frac{\frac{1}{n_a} \sum_{j \in \mathcal{I}_a} (\delta_j - \mu_i - a_{ij}) \widehat{\mathbf{I}}_{ij}}{\frac{1}{n_a} \sum_{j \in \mathcal{I}_a} \widehat{\mathbf{I}}_{ij}} - \frac{\frac{1}{n_a} \sum_{j \in \mathcal{I}_a} \mathbf{E}_{\delta_j}\left[(\delta_j - \mu_i - a_{ij}) \widehat{\mathbf{I}}_{ij} \mid \widehat{\mathbf{I}}_{ij}\right]}{\frac{1}{n_a} \sum_{j \in \mathcal{I}_a} \widehat{\mathbf{I}}_{ij}} \\ &= \frac{\frac{1}{n_a} \sum_{j \in \mathcal{I}_a} \widehat{\mathbf{I}}_{ij} \left((\delta_j - \mu_i - a_{ij}) - \mathbf{E}_{\delta_j}\left(\delta_j - \mu_i - a_{ij} \mid \widehat{\mathbf{I}}_{ij}\right)\right)}{\frac{1}{n_a} \sum_{j \in \mathcal{I}_a} \widehat{\mathbf{I}}_{ij}}. \end{aligned} \tag{72}$$

Define

$$F_{ij}(t) = \frac{\int_{\mu_i + t a_{ij} - h}^{\mu_i + t a_{ij} + h} (\delta - \mu_i - t a_{ij}) \phi(\delta) d\delta}{\int_{\mu_i + t a_{ij} - h}^{\mu_i + t a_{ij} + h} \phi(\delta) d\delta}. \tag{73}$$

Note that

$$\mathbf{E}_{\delta_j}\left(\delta_j-\mu_i-a_{ij}\mid\widehat{\mathbf{I}}_{ij}\right)=\frac{\int_{\mu_i+a_{ij}-h}^{\mu_i+a_{ij}+h}(\delta-\mu_i-a_{ij})\phi(\delta)d\delta}{\int_{\mu_i+a_{ij}-h}^{\mu_i+a_{ij}+h}\phi(\delta)d\delta}=F_{ij}(1). \tag{74}$$

By (71), (72), and (74), we establish

$$\begin{aligned}\widehat{l}(X_i,\widehat{\gamma})-l(X_i,\gamma)&=\frac{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}(\delta_j-\mu_i-a_{ij})\widehat{\mathbf{I}}_{ij}}{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}}-\frac{\int_{\mu_i-h}^{\mu_i+h}(\delta-\mu_i)\,\phi(\delta)d\delta}{\int_{\mu_i-h}^{\mu_i+h}\phi(\delta)d\delta}\\ &=\frac{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}\left[(\delta_j-\mu_i-a_{ij})-\mathbf{E}_{\delta_j}\left(\delta_j-\mu_i-a_{ij}\mid\widehat{\mathbf{I}}_{ij}\right)\right]}{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}}+\frac{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}\left[F_{ij}(1)-F_{ij}(0)\right]}{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}},\end{aligned} \tag{75}$$

where the last component holds since $F_{ij}(0)$ defined in (73) does not depend on the index j.

The decomposition (75) and the following two inequalities lead to an upper bound for (68).

$$\frac{1}{n_a}\sum_{i\in\mathcal{I}_a}\mathbf{E}\left[\left(\frac{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}\left[(\delta_j-\mu_i-a_{ij})-\mathbf{E}_{\delta_j}\left(\delta_j-\mu_i-a_{ij}\mid\widehat{\mathbf{I}}_{ij}\right)\right]}{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}}\right)^2\frac{1}{2h}\mathbf{1}(|\delta_i-\mu_i|\le h)\right]\lesssim\frac{h}{n_a}, \tag{76}$$

and

$$\begin{aligned}&\frac{1}{n_a}\sum_{i\in\mathcal{I}_a}\mathbf{E}\left[\left(\frac{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}\left[F_{ij}(1)-F_{ij}(0)\right]}{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}}\right)^2\frac{1}{2h}\mathbf{1}(|\delta_i-\mu_i|\le h)\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_1}\right]\\ &\lesssim h^4\left(C_1^2(n)+C_2(n)\right)^2\frac{k\log p\log n}{n}.\end{aligned} \tag{77}$$

We establish (58) by combining (66), (75), (76), (77), and (39).

C.3.1 Proof of (76)

In the following proof, we fix $i\in\mathcal{I}_a$ and $\mathbf{I}_{ii}=\widehat{\mathbf{I}}_{ii}$. For $j_1,j_2\in\mathcal{I}_a$ with $j_1\neq j_2$,

$$\begin{aligned}\mathbf{E}_{\delta|\widehat{\mathbf{I}}_{i\cdot}}\Big\{&\left[(\delta_{j_1}-\mu_i-a_{ij_1})-\mathbf{E}_{\delta_{j_1}}\left(\delta_{j_1}-\mu_i-a_{ij_1}\mid\widehat{\mathbf{I}}_{ij_1}\right)\right]\cdot\\ &\left[(\delta_{j_2}-\mu_i-a_{ij_2})-\mathbf{E}_{\delta_{j_2}}\left(\delta_{j_2}-\mu_i-a_{ij_2}\mid\widehat{\mathbf{I}}_{ij_2}\right)\right]\Big\}=0,\end{aligned}$$

and

$$\mathbf{E}_{\delta_j|\widehat{\mathbf{I}}_{ij}}\left[(\delta_j-\mu_i-a_{ij})-\mathbf{E}_{\delta_j}\left(\delta_j-\mu_i-a_{ij}\mid\widehat{\mathbf{I}}_{ij}\right)\right]^2\le\mathbf{E}_{\delta_j|\widehat{\mathbf{I}}_{ij}}\left[(\delta_j-\mu_i-a_{ij})\right]^2\le h^2.$$

By the above two expressions, we obtain

$$\begin{aligned}
&\mathbf{E}_{\delta|\widehat{\mathbf{I}}_{i\cdot}}\left[\left(\frac{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}\left[(\delta_j-\mu_i-a_{ij})-\mathbf{E}_{\delta_j}\left(\delta_j-\mu_i-a_{ij}\mid\widehat{\mathbf{I}}_{ij}\right)\right]}{\frac{1}{n_a}+\frac{1}{n_a}\sum_{j\neq i}\widehat{\mathbf{I}}_{ij}}\right)^2\frac{1}{2h}\mathbf{I}_{ii}\right]\\
&=\frac{\frac{1}{n_a^2}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}\mathbf{E}_{\delta_j|\widehat{\mathbf{I}}_{ij}}\left[(\delta_j-\mu_i-a_{ij})-\mathbf{E}_{\delta_j}\left(\delta_j-\mu_i-a_{ij}\mid\widehat{\mathbf{I}}_{ij}\right)\right]^2}{\left(\frac{1}{n_a}+\frac{1}{n_a}\sum_{j\neq i}\widehat{\mathbf{I}}_{ij}\right)^2}\frac{1}{2h}\mathbf{I}_{ii} \qquad (78)\\
&\leq\frac{h}{2n_a}\frac{\mathbf{I}_{ii}}{\frac{1}{n_a}+\frac{1}{n_a}\sum_{j\neq i}\widehat{\mathbf{I}}_{ij}},
\end{aligned}$$

We further take expectation with respect to $\mathbf{I}$ (but conditioning on μ_i and a_{ij}) in (78) and obtain that

$$\mathbf{E}_\delta\left(\frac{h}{n_a}\frac{\mathbf{I}_{ii}}{\frac{1}{n_a}+\frac{1}{n_a}\sum_{j\neq i}\widehat{\mathbf{I}}_{ij}}\right)=\frac{h}{n_a}\cdot\left(\mathbf{E}_\delta\mathbf{I}_{ii}\right)\cdot\left(\mathbf{E}_\delta\frac{1}{\frac{1}{n_a}+\frac{1}{n_a}\sum_{j\neq i}\widehat{\mathbf{I}}_{ij}}\right). \tag{79}$$

Conditioning on μ_i and a_{ij}, we define the conditional probability with respect to δ_j as

$$e_{ij}=\mathbf{E}_{\delta_j}[\widehat{\mathbf{I}}_{ij}]=\int_{\mu_i+a_{ij}-h}^{\mu_i+a_{ij}+h}\phi(\delta_j)d\delta_j. \tag{80}$$

By change of variable $\tau_j=\delta_j-(\mu_i+a_{ij})$, we have

$$\begin{aligned}
\int_{\mu_i+a_{ij}-h}^{\mu_i+a_{ij}+h}\phi(\delta_j)d\delta_j&=\int_{-h}^{h}\phi(\tau_j+\mu_i+a_{ij})d\tau_j\\
&=\int_{-h}^{h}[\phi(\mu_i+a_{ij})+\tau_j\phi'(\mu_i+a_{ij})+\frac{\tau_j^2}{2}\phi''(\mu_i+a_{ij}+c\tau_j)]d\tau_j,
\end{aligned}$$

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

$$\left|\int_{\mu_i+a_{ij}-h}^{\mu_i+a_{ij}+h}\phi(\delta_j)d\delta_j-2h\phi(\mu_i+a_{ij})\right|\leq\frac{h^3}{3}\max_{|c|\leq 1}\left|\phi''(\mu_i+a_{ij}+ch)\right|,$$

and then

$$|e_{ij}-2h\phi(\mu_i)|\leq\frac{h^3}{3}\max_{|c|\leq 1}\left|\phi''(\mu_i+a_{ij}+ch)\right|+2h|a_{ij}|\max_{|c|\leq 1}\left|\phi'(\mu_i+ca_{ij})\right|.$$

Hence, we establish

$$\left|\frac{e_{ij}}{2h\phi(\mu_i)}-1\right|\leq\frac{h^2}{3}C_2(n)+|a_{ij}|C_1(n). \tag{81}$$

We note that the non-negative random variable $\frac{1}{n_a}\sum_{j\neq i}\widehat{\mathbf{I}}_{ij}$ satisfies

$$\mathbf{E}\left[\frac{1}{n_a}\sum_{j\neq i}\widehat{\mathbf{I}}_{ij}\mid\{X_i\}_{i\in\mathcal{I}_a},\mathcal{I}_b\right]=\frac{1}{n_a}\sum_{j\neq i}e_{ij},$$

and

$$\mathrm{Var}\left[\frac{1}{n_a}\sum_{j\neq i}\widehat{\mathbf{I}}_{ij} \mid \{X_i\}_{i\in\mathcal{I}_a},\mathcal{I}_b\right]=\frac{1}{n_a^2}\sum_{j\neq i}e_{ij}(1-e_{ij}).$$

Hence, we apply the equation (5) in Wooff (1985) and obtain

$$\begin{aligned}\mathbf{E}_\delta\left[\frac{1}{\frac{1}{n_a}+\frac{1}{n_a}\sum_{j\neq i}\widehat{\mathbf{I}}_{ij}}\right]&=\mathbf{E}\left[\frac{1}{\frac{1}{n_a}+\frac{1}{n_a}\sum_{j\neq i}\widehat{\mathbf{I}}_{ij}} \mid \{X_i\}_{i\in\mathcal{I}_a},\mathcal{I}_b\right]\\&\leq\frac{1}{\frac{1}{n_a}\sum_{j\neq i}e_{ij}+\frac{1}{n_a}}\left(1+\frac{\frac{1}{n_a^2}\sum_{j\neq i}e_{ij}(1-e_{ij})}{\frac{1}{n_a^2}\sum_{j\neq i}e_{ij}}\right)\qquad(82)\\&\leq\frac{2}{\frac{1}{n_a}\sum_{j\neq i}e_{ij}+\frac{1}{n_a}}.\end{aligned}$$

A combination of (79), (81) and (82) leads to

$$\mathbf{E}_\delta\left(\frac{h}{n_a}\frac{\mathbf{I}_{ii}}{\frac{1}{n_a}+\frac{1}{n_a}\sum_{j\neq i}\widehat{\mathbf{I}}_{ij}}\right)\leq\frac{h}{n_a}\frac{2e_{ii}}{\frac{1}{n_a}\sum_{j\neq i}e_{ij}+\frac{1}{n_a}}\lesssim\frac{h}{n_a}.\qquad(83)$$

By taking expectation with respect to $\{X_i\}_{i\in\mathcal{I}_a}$ and $\mathcal{I}_b$, we establish (76).

C.3.2 Proof of (77)

We calculate the expression of $F_{ij}(t)$ in (73) as

$$\begin{aligned}\frac{dF_{ij}(t)}{dt}=&\frac{a_{ij}h\left[\phi(\mu_i+ta_{ij}+h)+\phi(\mu_i+ta_{ij}-h)-\frac{1}{h}\int_{\mu_i+ta_{ij}-h}^{\mu_i+ta_{ij}+h}\phi(\delta_j)d\delta_j\right]}{\int_{\mu_i+ta_{ij}-h}^{\mu_i+ta_{ij}+h}\phi(\delta_j)d\delta_j}\\&-\frac{\int_{\mu_i+ta_{ij}-h}^{\mu_i+ta_{ij}+h}(\delta_j-\mu_i-ta_{ij})\phi(\delta_j)d\delta_j}{\left[\int_{\mu_i+ta_{ij}-h}^{\mu_i+ta_{ij}+h}\phi(\delta_j)d\delta_j\right]^2}\cdot a_{ij}\left[\phi(\mu_i+ta_{ij}+h)-\phi(\mu_i+ta_{ij}-h)\right].\end{aligned}\qquad(84)$$

Define

$$T_1(a_{ij})=\phi(\mu_i+ta_{ij}+h)+\phi(\mu_i+ta_{ij}-h)-\frac{1}{h}\int_{\mu_i+ta_{ij}-h}^{\mu_i+ta_{ij}+h}\phi(\delta_j)d\delta_j;$$

$$T_2(a_{ij})=\phi(\mu_i+ta_{ij}+h)-\phi(\mu_i+ta_{ij}-h);$$

and

$$T_3(a_{ij})=\int_{\mu_i+ta_{ij}-h}^{\mu_i+ta_{ij}+h}[\delta_j-(\mu_i+ta_{ij})]\phi(\delta_j)d\delta_j.$$

Then we can simplify the derivative of $F_{ij}(t)$ in (84) as

$$\frac{dF_{ij}(t)}{dt}=\frac{ha_{ij}T_1(a_{ij})}{e_{ij}}-\frac{a_{ij}T_2(a_{ij})}{e_{ij}}\cdot\frac{T_3(a_{ij})}{e_{ij}},\qquad(85)$$

where e_{ij} is defined in (80). To bound the above terms, we introduce the following lemma to control all of the above terms.

Lemma 3 *Suppose that $\phi(t)$ is twice differentiable for $t \in [\mu-\tau, \mu+\tau]$, there exists some positive constant $C>0$ such that*

$$\left|\phi(\mu+\tau)+\phi(\mu-\tau)-\frac{1}{\tau}\int_{\mu-\tau}^{\mu+\tau}\phi(t)dt\right| \leq C\tau^2 \cdot \max_{t\in[\mu-\tau,\mu+\tau]}\left|\phi''(t)\right| \tag{86}$$

$$\left|\int_{\mu-\tau}^{\mu+\tau}(t-\mu)\,\phi(t)dt\right| \leq C\tau^3 \max_{t\in[\mu-\tau,\mu+\tau]}\left|\phi'(t)\right| \tag{87}$$

$$|\phi(\mu+\tau)-\phi(\mu-\tau)| \leq C\tau \max_{t\in[\mu-\tau,\mu+\tau]}\left|\phi'(t)\right| \tag{88}$$

It follows from (86) that

$$|T_1(a_{ij})| \lesssim h^2 \cdot \max_{|\delta-(\mu_i+ta_{ij})|\leq h}|\phi''(\delta)| \leq h^2 \cdot \max_{|\delta-\mu_i|\leq r}|\phi''(\delta)|.$$

It follows from (87) that

$$|T_2(a_{ij})| \lesssim h \cdot \max_{|\delta-(\mu_i+ta_{ij})|\leq h}|\phi'(\delta)| \leq h \cdot \max_{|\delta-\mu_i|\leq r}|\phi'(\delta)|.$$

It follows from (88) that

$$|T_3(a_{ij})| \lesssim h^3 \cdot \max_{|\delta-(\mu_i+ta_{ij})|\leq h}|\phi'(\delta)| \leq h^3 \cdot \max_{|\delta-\mu_i|\leq r}|\phi'(\delta)|.$$

Together with (81), we establish

$$\frac{|T_1(a_{ij})|}{e_{ij}} \lesssim h \cdot \max_{|\delta-\mu_i|\leq r}\frac{|\phi''(\delta)|}{\phi(\mu_i)}, \quad \frac{|T_2(a_{ij})|}{e_{ij}} \lesssim \max_{|\delta-\mu_i|\leq r}\frac{|\phi'(\delta)|}{\phi(\mu_i)}, \quad \frac{|T_3(a_{ij})|}{e_{ij}} \lesssim h^2 \cdot \max_{|\delta-\mu_i|\leq r}\frac{|\phi'(\delta)|}{\phi(\mu_i)},$$

where $r = C^*\sqrt{\|\gamma\|_0 \log p \log n/n} + h$. Together with the expression (85) and the upper bound (69), we establish

$$\left|\frac{dF_{ij}(t)}{dt}\right| \mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_1} \lesssim \sqrt{\frac{k\log p\log n}{n}}h^2\left(C_1^2(n)+C_2(n)\right).$$

Hence, we have

$$\left(\frac{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}\left[F_{ij}(1)-F_{ij}(0)\right]}{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}}\right)^2 \mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_1} \lesssim \left(\sqrt{\frac{k\log p\log n}{n}}h^2\left(C_1^2(n)+C_2(n)\right)\right)^2,$$

and

$$\begin{aligned}
&\frac{1}{n_a}\sum_{i=1}^{n_a}\mathbf{E}\left(\frac{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}\left[F_{ij}(1)-F_{ij}(0)\right]}{\frac{1}{n_a}\sum_{j\in\mathcal{I}_a}\widehat{\mathbf{I}}_{ij}}\right)^2\frac{1}{2h}\mathbf{1}(|\delta_i-\mu_i|\leq h)\cdot\mathbf{1}_{\mathcal{A}_0\cap\mathcal{A}_1} \\
&\lesssim \left(\sqrt{\frac{k\log p\log n}{n}}h^2\left(C_1^2(n)+C_2(n)\right)\right)^2\frac{1}{n_a}\sum_{i=1}^{n_a}\mathbf{E}\frac{1}{2h}\mathbf{1}(|\delta_i-\mu_i|\leq h) \\
&= \left(\sqrt{\frac{k\log p\log n}{n}}h^2\left(C_1^2(n)+C_2(n)\right)\right)^2\frac{1}{n_a}\sum_{i=1}^{n_a}\mathbf{E}K_h(D_i).
\end{aligned} \tag{89}$$

Together with (41), we establish (77).

Appendix Appendix D. Proof of Theorem 1

We start with the following error decomposition of $\widehat{f'(a_0)} - f'(a_0)$,

$$\underbrace{\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_i\epsilon_i K_h(D_i)}_{\text{Stochastic Error}} + \underbrace{\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_i r(D_i) K_h(D_i)}_{\text{Approximation Error}} + \underbrace{\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_i\Delta(X_i) K_h(D_i)}_{\text{High-dimensional Error}}. \tag{90}$$

The high-dimensional error is controlled in Theorem 2. We shall control the stochastic error and the approximation error in Sections D.1 and D.2, respectively. We present the proof of (31) in Section D.3.

D.1 Analysis of the Stochastic Error

We shall establish the following limiting distribution,

$$\frac{1}{\sqrt{\mathrm{V}}}\frac{\sum_{i=1}^{n}\widehat{W}_i\epsilon_i K_h(D_i)}{n\widehat{S}_n} \xrightarrow{d} N(0,1), \tag{91}$$

with

$$\mathrm{V} = \frac{\sigma^2}{n^2\widehat{S}_n^2}\sum_{i=1}^{n}\widehat{W}_i^2 K_h^2(D_i) \xrightarrow{p} \frac{3\sigma^2}{nh^3\cdot\pi(a_0)}. \tag{92}$$

In the following, we shall provide proofs for both (91) and (92).

D.1.1 Proof of (92)

We decompose the error between $\frac{1}{n}\sum_{i=1}^{n}\widehat{W}_i^2 K_h^2(D_i)$ and its corresponding estimand,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{W}_i^2 K_h^2(D_i) - \frac{1}{n}\sum_{i=1}^{n}\left(W_i-\bar{\mu}_W\right)^2 K_h^2(D_i)\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n}\left[2\left(W_i-\bar{\mu}_W\right)\cdot\left(\widehat{W}_i-\left(W_i-\bar{\mu}_W\right)\right) + \left(\widehat{W}_i-\left(W_i-\bar{\mu}_W\right)\right)^2\right]K_h^2(D_i)\right| \\
&\le \frac{1}{2h}\left(2\mathrm{Err}(\widehat{W})\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\bar{\mu}_W\right)^2 K_h(D_i)} + \mathrm{Err}^2(\widehat{W})\right),
\end{aligned} \tag{93}$$

where the inequality follows from triangle inequality, $|K_h(D_i)| \le 1/(2h)$, and Cauchy-Schwarz inequality. We bound the difference between the sum of centered variables and that of uncentered variables,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\bar{\mu}_W\right)^2 K_h(D_i) - \frac{1}{n}\sum_{i=1}^{n}W_i^2 K_h(D_i)\right| \le 2\left|\bar{\mu}_W\right|\cdot\left|\frac{1}{n}\sum_{i=1}^{n}W_i K_h(D_i)\right| + 2\bar{\mu}_W^2\left|\frac{1}{n}\sum_{i=1}^{n}K_h(D_i)\right|.$$

By applying (48) and (49) in Lemma 2 and (54), we establish that, with probability larger than $1-\exp(-t^2)$ for $t \ll \sqrt{nh\pi(a_0)}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\bar{\mu}_W\right)^2 K_h(D_i) - \frac{1}{n}\sum_{i=1}^{n}W_i^2 K_h(D_i)\right| \lesssim t^2\sqrt{\frac{h}{n\pi(a_0)}}\sqrt{\frac{h}{n}\pi(a_0)} + \frac{t^2 h}{n} \lesssim \frac{ht^2}{n}. \tag{94}$$

Note that

$$\frac{1}{2h}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\bar{\mu}_W\right)^2 K_h(D_i)=\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\bar{\mu}_W\right)^2 K_h^2(D_i).$$

We apply (45), (53), and (94) and establish

$$\left|\frac{\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\bar{\mu}_W\right)^2 K_h^2(D_i)}{\frac{1}{3}h\pi(a_0)}-1\right|\lesssim h^2[C_1^2(n)+C_2(n)]+\frac{\mathbf{P}(\mathcal{A}_0^c)}{h\pi(a_0)}+\frac{t}{\sqrt{nh\pi(a_0)}}. \tag{95}$$

We shall choose $t=\sqrt{\log n}$ and establish that, with probability larger than $1-n^{-c}$,

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\bar{\mu}_W\right)^2 K_h(D_i)}{\frac{2}{3}h^2\pi(a_0)}\asymp 1.$$

Combined with (93) and (95), we establish that, with probability larger than $1-n^{-c}$,

$$\left|\frac{\frac{1}{n}\sum_{i=1}^{n}\widehat{W}_i^2 K_h^2(D_i)}{\frac{1}{3}h\pi(a_0)}-1\right|\lesssim\frac{\mathrm{Err}^2(\widehat{W})}{h^2\pi(a_0)}+\sqrt{\frac{\mathrm{Err}^2(\widehat{W})}{h^2\pi(a_0)}}+h^2[C_1^2(n)+C_2(n)]+\frac{\mathbf{P}(\mathcal{A}_0^c)}{h\pi(a_0)}+\frac{t}{\sqrt{nh\pi(a_0)}}. \tag{96}$$

For $\widehat{S}_n$ defined in (20), we approximate it by its corresponding estimand,

$$\begin{aligned}&\left|\widehat{S}_n-\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\bar{\mu}_W\right)(D_i-a_0)K_h(D_i)\right|\\&\le\mathrm{Err}(\widehat{W})\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(D_i-a_0\right)^2K_h(D_i)}\le h\cdot\mathrm{Err}(\widehat{W})\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}K_h(D_i)}.\end{aligned}\tag{97}$$

By applying (46) and (51), we establish that, with probability larger than $1-\exp(-t^2)$,

$$\left|\frac{\frac{1}{n}\sum_{i=1}^{n}W_i(D_i-a_0)K_h(D_i)}{\frac{1}{3}h^2\pi(a_0)}-1\right|\lesssim h^2[C_1^2(n)+C_2(n)]+\frac{\mathbf{P}(\mathcal{A}_0^c)}{h\pi(a_0)}+t\sqrt{\frac{1}{nh\pi(a_0)}}. \tag{98}$$

By (54) and (50) in Lemma 2, we establish that, with probability larger than $1-\exp(-t^2)$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\bar{\mu}_W(D_i-a_0)K_h(D_i)\right|\lesssim t\sqrt{\frac{h}{n\pi(a_0)}}\cdot\left(\frac{\pi(a_0)}{3}h^2(C_1(n)+hC_2(n))+\mathbf{P}(\mathcal{A}_0^c)+t\sqrt{\frac{h}{n}\pi(a_0)}\right). \tag{99}$$

Together with (97) and (98), we establish that, with probability larger than $1-\exp(-t^2)$,

$$\left|\frac{\widehat{S}_n}{\frac{1}{3}h^2\pi(a_0)}-1\right|\lesssim\frac{\mathrm{Err}(\widehat{W})}{h\sqrt{\pi(a_0)}}+\left(h^2+\sqrt{\frac{h}{n\pi(a_0)}}\right)[C_1^2(n)+C_2(n)]+\frac{\mathbf{P}(\mathcal{A}_0^c)}{h\pi(a_0)}+t\sqrt{\frac{1}{nh\pi(a_0)}}. \tag{100}$$

Under the condition that $\mathrm{Err}(\widehat{W})\ll h\sqrt{\pi(a_0)}$, $h^2C_2(n)+hC_1(n)\to 0$, $\mathbf{P}(\mathcal{A}_0^c)\ll h\pi(a_0)$, and $nh\pi(a_0)\gg\log n$, we establish (92) by combining (96) and (100).

D.1.2 Proof of (91)

Define $Z_i = \frac{1}{\sqrt{\mathrm{V}/\sigma^2}} \frac{\widehat{W}_i K_h(D_i)}{n\widehat{S}_n} \in \mathbb{R}$. We rewrite the stochastic error as follows,

$$\frac{1}{\sqrt{\mathrm{V}}} \frac{\sum_{i=1}^n \widehat{W}_i \epsilon_i K_h(D_i)}{n\widehat{S}_n} = \sum_{i=1}^n Z_i \cdot \epsilon_i/\sigma.$$

We use $\mathcal{O}$ to denote the data $\mathcal{O} = \{D_i, X_i\}_{1\leq i\leq n}$. Conditioning on $\mathcal{O}$, $Z_i \cdot \epsilon_i/\sigma$ are independent random variables with

$$\mathbf{E}\left(Z_i \cdot \epsilon_i/\sigma \mid \mathcal{O}\right) = 0$$

and

$$\sum_{i=1}^n \mathrm{Var}\left(Z_i \cdot \epsilon_i/\sigma \mid \mathcal{O}\right) = 1.$$

Define the event

$$\mathcal{G}_0 = \left\{ \left| \frac{\frac{1}{n}\sum_{i=1}^n \widehat{W}_i^2 K_h^2(D_i)}{\frac{1}{3}h\pi(a_0)} - 1 \right| \leq 1/10 \right\}.$$

The high probability inequality in (96) implies

$$\mathbf{P}(\mathcal{G}_0) \geq 1 - n^{-c}. \tag{101}$$

By applying (96) and the fact that $|\widehat{W}_i K_h(D_i)| \leq C$ for a positive constant $C > 0$, we obtain that, on the event $\mathcal{G}_0$,

$$|Z_i| = \left| \frac{\widehat{W}_i K_h(D_i)}{\sqrt{\sum_{i=1}^n \widehat{W}_i^2 K_h^2(D_i)}} \right| \lesssim \frac{1}{\sqrt{nh\pi(a_0)}}. \tag{102}$$

It is sufficient to check the Linderberg condition

$$\begin{aligned}
\sum_{i=1}^n \mathbf{E}\left[(Z_i \cdot \epsilon_i/\sigma)^2 \mathbf{1}\left(|Z_i \cdot \epsilon_i/\sigma| \geq \tau\right) \mid \mathcal{O}\right] &= \sum_{i=1}^n Z_i^2 \mathbf{E}\left[\frac{\epsilon_i^2}{\sigma^2}\mathbf{1}\left(\left|\frac{\epsilon_i}{\sigma}\right| \geq \frac{\tau}{|Z_i|}\right) \mid \mathcal{O}\right] \\
&\leq \sum_{i=1}^n Z_i^2 \mathbf{E}\left[\frac{\epsilon_i^2}{\sigma^2}\mathbf{1}\left(\left|\frac{\epsilon_i}{\sigma}\right| \gtrsim \tau\sqrt{nh\pi(a_0)}\right) \mid \mathcal{O}\right] \\
&\leq \sum_{i=1}^n Z_i^2 \max_{1\leq i\leq n} \mathbf{E}\left[\frac{\epsilon_i^2}{\sigma^2}\mathbf{1}\left(\left|\frac{\epsilon_i}{\sigma}\right| \gtrsim \tau\sqrt{nh\pi(a_0)}\right) \mid \mathcal{O}\right] \\
&\leq (\tau\sqrt{nh\pi(a_0)})^{-c},
\end{aligned}$$

where the first inequality follows from (102) and the last inequality follows from the condition that $\mathbf{E}(\epsilon_i^{2+c} \mid D_i, X_i) \leq C$ for some positive constant $c > 0$ and $C > 0$.

Then we apply the Linderberg condition and establish that

$$\sum_{i=1}^n \frac{1}{\sqrt{\mathrm{V}}} \frac{\widehat{W}_i \epsilon_i K_h(D_i)}{n\widehat{S}_n} \mid \mathcal{O} \in \mathcal{G}_0 \xrightarrow{d} N(0,1). \tag{103}$$

By (101) and (103), we have

$$\mathbf{E}\left(\mathbf{E}\left[\exp\left(it(\sum_{i=1}^{n} Z_i\epsilon_i/\sigma)\right) \mid \mathcal{O}\right]\cdot \mathbf{1}_{\mathcal{G}_0}\right) \to \exp(-t^2/2).$$

Together with

$$\left|\mathbf{E}\exp\left(it(\sum_{i=1}^{n} Z_i\epsilon_i/\sigma)\right) - \mathbf{E}\left(\mathbf{E}\left[\exp\left(it(\sum_{i=1}^{n} Z_i\epsilon_i/\sigma)\right) \mid \mathcal{O}\right]\cdot \mathbf{1}_{\mathcal{G}_0}\right)\right| \leq \mathbf{P}\left(\mathcal{G}_0^c\right),$$

we establish (91).

D.2 Analysis of the Approximation Error

In the following, we show that, with probability larger than $1 - n^{-c}$,

$$\left|\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_i\left[r(D_i) - \frac{(D_i-a_0)^2}{2}f''(a_0)\right]K_h(D_i)\right| \lesssim \max_{|d-a_0|\leq h}|f''(d)-f''(a_0)|\cdot\left(\frac{\mathrm{Err}(\widehat{W})}{\sqrt{\pi(a_0)}}+h\right), \tag{104}$$

and

$$\left|\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_i\frac{(D_i-a_0)^2}{2}f''(a_0)K_h(D_i)\right| \lesssim \frac{\mathrm{Err}(\widehat{W})}{\sqrt{\pi(a_0)}} + hc_u, \tag{105}$$

with

$$c_u = hC_1(n) + h^2C_2(n) + \sqrt{\frac{\log n}{nh\pi(a_0)}} + \frac{\mathbf{P}(\mathcal{A}_0^c)}{h\pi(a_0)}.$$

By the continuity of f'' at the point a_0, we combine (104) and (105) and establish that,

$$\left|\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_i r(D_i)K_h(D_i)\right| \lesssim \frac{\mathrm{Err}(\widehat{W})}{\sqrt{\pi(a_0)}} + \left(c_u + \max_{|d-a_0|\leq h}\left|f''(d)-f''(a_0)\right|\right)\cdot h. \tag{106}$$

D.2.1 Proof of (104)

There exists some $c\in(0,1)$ such that

$$\begin{aligned} r(D_i) &= f(D_i) - f(a_0) - (D_i-a_0)f'(a_0)\\ &= \frac{(D_i-a_0)^2}{2}f''(a_0) + \frac{(D_i-a_0)^2}{2}\left[f''(a_0+c(D_i-a_0)) - f''(a_0)\right]. \end{aligned}$$

Hence, we have

$$\frac{2}{h^2}\left|r(D_i)\mathbf{1}\left(\left|\frac{D_i-a_0}{h}\right|\leq 1\right) - \frac{(D_i-a_0)^2}{2}f''(a_0)\mathbf{1}\left(\left|\frac{D_i-a_0}{h}\right|\leq 1\right)\right| \leq \left|f''(d)-f''(a_0)\right|,$$

for some d satisfying $a_0 - h \leq d \leq a_0 + h$. The above inequality implies that

$$\frac{\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{W}_i r(D_i)K_h(D_i) - \frac{1}{n}\sum_{i=1}^{n}\widehat{W}_i\frac{(D_i-a_0)^2}{2}f''(a_0)K_h(D_i)\right|}{h^2\pi(a_0)\cdot\frac{1}{n\pi(a_0)}\sum_{i=1}^{n}\left|\widehat{W}_i\right|K_h(D_i)} \lesssim \max_{|d-a_0|\leq h}\left|f''(d)-f''(a_0)\right|. \tag{107}$$

We now control the term $\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{W}_i\right| K_h(D_i)$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left|\widehat{W}_i\right| K_h(D_i) &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{W}_i-(W_i-\bar{\mu}_W)\right| K_h(D_i)+\frac{1}{n}\sum_{i=1}^{n}\left(|W_i|+|\bar{\mu}_W|\right) K_h(D_i)\\ &\leq \operatorname{Err}(\widehat{W})\sqrt{\frac{1}{n}\sum_{i=1}^{n}K_h(D_i)}+\frac{2}{n}\sum_{i=1}^{n}|W_i|\,K_h(D_i)\\ &\leq \operatorname{Err}(\widehat{W})\sqrt{\frac{1}{n}\sum_{i=1}^{n}K_h(D_i)}+4h\left(\frac{1}{n}\sum_{i=1}^{n}K_h(D_i)\right),\end{aligned}$$

where the last inequality follows from the fact that $|W_i|\,K_h(D_i)\leq 2hK_h(D_i)$. Together with (48) with $t=\sqrt{\log n}$, we establish that, with probability larger than $1-n^{-c}$,

$$\frac{1}{n\pi(a_0)}\sum_{i=1}^{n}\left|\widehat{W}_i\right| K_h(D_i)\lesssim\frac{\operatorname{Err}(\widehat{W})}{\sqrt{\pi(a_0)}}+h. \tag{108}$$

We establish (104) by combining (107), (108), and (100).

D.2.2 Proof of (105)

By the expression $\widehat{W}_i=(W_i-\bar{\mu}_W)+\widehat{W}_i-(W_i-\bar{\mu}_W)$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\widehat{W}_i\frac{(D_i-a_0)^2}{2}f''(a_0)K_h(D_i)&=\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{W}_i-(W_i-\bar{\mu}_W)\right]\frac{(D_i-a_0)^2}{2}f''(a_0)K_h(D_i)\\ +\frac{1}{n}\sum_{i=1}^{n}W_i\frac{(D_i-a_0)^2}{2}f''(a_0)K_h(D_i)&-\bar{\mu}_W\frac{1}{n}\sum_{i=1}^{n}\frac{(D_i-a_0)^2}{2}f''(a_0)K_h(D_i).\end{aligned} \tag{109}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{W}_i-(W_i-\bar{\mu}_W)\right]\frac{(D_i-a_0)^2}{2}f''(a_0)K_h(D_i)\right|\\ &\lesssim\left|f''(a_0)\right|\operatorname{Err}(\widehat{W})\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(D_i-a_0)^4}{2}K_h(D_i)} \qquad (110)\\ &\leq\left|f''(a_0)\right|\cdot h^2\cdot\operatorname{Err}(\widehat{W})\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}K_h(D_i)},\end{aligned}$$

where the last inequality follows from the fact that $(D_i-a_0)^4K_h(D_i)\leq h^4K_h(D_i)$. In addition, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}W_i\frac{(D_i-a_0)^2}{2}f''(a_0)K_h(D_i)\right|=\left|f''(a_0)\right|\cdot\left|\frac{1}{n}\sum_{i=1}^{n}W_i\frac{(D_i-a_0)^2}{2}K_h(D_i)\right|, \tag{111}$$

and

$$\begin{aligned}\left|\bar{\mu}_W \frac{1}{n}\sum_{i=1}^{n}\frac{(D_i-a_0)^2}{2}f''(a_0)K_h(D_i)\right| &= |\bar{\mu}_W|\cdot\left|f''(a_0)\right|\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\frac{(D_i-a_0)^2}{2}K_h(D_i)\right| \\ &\leq \frac{h^2}{2}\,|\bar{\mu}_W|\cdot\left|f''(a_0)\right|\cdot\frac{1}{n}\sum_{i=1}^{n}K_h(D_i),\end{aligned} \tag{112}$$

where the last inequality follows from the fact that $(D_i-a_0)^2K_h(D_i)\leq h^2K_h(D_i)$. We now apply (48), (52), (54), the decomposition (109) with the error bounds in (110), (111), and (112). We establish that, with probability larger than $1-n^{-c}$,

$$\begin{aligned}&\frac{1}{h^2\pi(a_0)}\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{W}_i\frac{(D_i-a_0)^2}{2}f''(a_0)K_h(D_i)\right| \\ &\lesssim \frac{\mathrm{Err}(\widehat{W})}{\sqrt{\pi(a_0)}}+h\left(hC_1(n)+h^2C_2(n)+\sqrt{\frac{\log n}{nh\pi(a_0)}}+\frac{\mathbf{P}(\mathcal{A}_0^c)}{h\pi(a_0)}\right).\end{aligned}$$

Together with (100), we establish (104).

D.3 Proof of (31)

Under the conditions $\mathrm{Err}(\widehat{W})\ll\min\{\sqrt{nh^3},h\sqrt{\pi(a_0)}\}$, $h^2C_2(n)+hC_1(n)\to 0$, $\mathbf{P}(\mathcal{A}_0^c)\ll h\pi(a_0)$, and $nh^5\pi(a_0)\leq c$, we apply (91), (92), and (106) and establish

$$\frac{1}{\sqrt{\mathrm{V}}}\left(\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_i\epsilon_iK_h(D_i)+\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_ir(D_i)K_h(D_i)\right)\xrightarrow{d}N(0,1). \tag{113}$$

It follows from (100) and (32) that, with probability larger than $1-\frac{1}{t}-\min\{n,p\}^{-c}$ for some $t>1$

$$\frac{1}{\sqrt{V}}\left|\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_i\Delta(X_i)K_h(D_i)\right|\lesssim t^2\left[1+\sqrt{h^3k\log p\log n}\left(C_1^2(n)+C_2(n)\right)\right]\frac{\mathrm{Err}(\widehat{g})}{\sqrt{\pi(a_0)}}.$$

The condition (A3) implies that

$$\frac{1}{\sqrt{V}}\left|\frac{1}{n\widehat{S}_n}\sum_{i=1}^{n}\widehat{W}_i\Delta(X_i)K_h(D_i)\right|\xrightarrow{p}0.$$

Combined with (113), we establish the limiting distribution (31).

Appendix Appendix E. Proofs of Extra Lemmas

E.1 Proof of Equation (2)

For any given treatment level $a\in\mathbb{R}$, we have:

$$\begin{aligned}E(Y_i(a) \mid X_i) &= (Y_i(a) \mid D_i = a, X_i)\\ &= E(Y_i \mid D_i = 1, X_i)\\ &= f(a) + g(X_i)\end{aligned}$$

where the first equation applies the unconfoundedness condition and the second equation relies on the consistency condition. Since a is any treatment level, we apply the above equation to deduce (2).

E.2 Proof of Lemma 3

By Taylor's expansion, we have for $c_1, c_2 \in (0,1)$

$$\phi(\mu+\tau)+\phi(\mu-\tau) = \phi(\mu)+\tau\cdot\phi'(\mu)+\frac{\tau^2}{2}\cdot\phi''(\mu+c_1\tau)+\phi(\mu)-\tau\cdot\phi'(\mu)+\frac{\tau^2}{2}\cdot\phi''(\mu+c_2\tau)$$

and

$$\frac{1}{\tau}\int_{\mu-\tau}^{\mu+\tau}\phi(t)dt = \frac{1}{\tau}\int_{\mu-\tau}^{\mu+\tau}\left[\phi(\mu)+(t-\mu)\phi'(\mu)+\frac{(t-\mu)^2}{2}\phi''(\mu+c_3(t)(t-\mu))\right]dt,$$

where $c_3(t) \in (0,1)$. Hence, we have

$$\begin{aligned}\phi(\mu+\tau)+\phi(\mu-\tau)-\frac{1}{\tau}\int_{\mu-\tau}^{\mu+\tau}\phi(t)dt &= \frac{\tau^2}{2}\cdot\big(\phi''(\mu+c_1\tau)+\phi''(\mu+c_2\tau)\big)\\ &\quad+\frac{1}{\tau}\int_{\mu-\tau}^{\mu+\tau}\frac{(t-\mu)^2}{2}\phi''(\mu+c_3(t)(t-\mu))dt.\end{aligned}$$

Hence, we establish (86). Note that

$$\int_{\mu-\tau}^{\mu+\tau}(t-\mu)\,\phi(t)dt = \int_{\mu-\tau}^{\mu+\tau}\left[(t-\mu)\phi(\mu)+(t-\mu)^2\phi'(\mu+c_4(t)(t-\mu))\right]dt,$$

where $c_4(t) \in (0,1)$. Hence we establish (87). Note that

$$\phi(\mu+\tau)-\phi(\mu-\tau) = 2\tau\cdot\phi'(\mu+c_5\tau),$$

for $c_5 \in (-1,1)$. Hence we establish (88).

E.3 Proof of Lemma 1

E.3.1 Proof of (40) and (41)

We start with the expression of $\mathbf{E}\left[K_h(D_i) \mid X_i\right]$,

$$\mathbf{E}\left[K_h(D_i) \mid X_i\right] = \int_{\left|\frac{D_i-a_0}{h}\right|\leq 1}\frac{1}{2h}q(D_i \mid X_i)dD_i.$$

By setting $z = \frac{D_i - a_0}{h}$, we simplify the above expression as,

$$\int_{|z|\leq 1} \frac{1}{2} q(a_0 + hz \mid X_i) dz = \\ \frac{1}{2} \int_{|z|\leq 1} \left[q(a_0 \mid X_i) + hz q'(a_0 \mid X_i) + \frac{h^2 z^2}{2} q''(a_0 + c(z)hz \mid X_i) \right] dz \quad (114)$$

for some $c(z) \in (0,1)$. 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

$$|\mathbf{E}\left[K_h(D_i) \mid X_i\right] - q(a_0 \mid X_i)| \leq \frac{1}{6} h^2 \max_{|c|\leq 1} q''(a_0 + ch \mid X_i). \quad (115)$$

By Condition (A2), we have

$$\max_{|c|\leq 1} \left| \frac{q''(a_0 + ch \mid X_i)}{q(a_0 \mid X_i)} \right| \cdot \mathbf{1}_{\mathcal{A}_0} \leq C_2(n) \quad (116)$$

where $C_2(n)$ is defined in (27). Together with (115), we establish (40).

We apply the boundedness of $\phi(\delta)$ and establish $\mathbf{E}\left[K_h(D_i) \mid X_i\right] \leq C$ for some positive constant $C > 0$. Together with (40), we establish (41).

E.3.2 Proof of (42) and (43)

We first prove (43) by analyzing the term $\mathbf{E}\left[(D_i - a_0)^2 K_h(D_i) \mid X_i\right]$. Similar to (114), we write down the following explicit expression,

$$\mathbf{E}\left[(D_i - a_0)^2 K_h(D_i) \mid X_i\right] = \int_{\left|\frac{D_i - a_0}{h}\right| \leq 1} \left[D_i - a_0\right]^2 \frac{1}{2h} q(D_i \mid X_i) dD_i$$

By setting $z = \frac{D_i - a_0}{h}$, we have

$$\begin{aligned} \mathbf{E}\left[(D_i - a_0)^2 K_h(D_i) \mid X_i\right] &= \int_{|z|\leq 1} \frac{1}{2} h^2 z^2 q(a_0 + hz \mid X_i) dz \\ &= \int_{|z|\leq 1} \frac{1}{2} h^2 z^2 \left[q(a_0 \mid X_i) + hz q'(a_0 \mid X_i) + \frac{h^2 z^2}{2} q''(a_0 + c(z)hz \mid X_i) \right] dz. \end{aligned} \quad (117)$$

Hence, we have

$$\left| \mathbf{E}\left[(D_i - a_0)^2 K_h(D_i) \mid X_i\right] - \frac{1}{3} h^2 q(a_0 \mid X_i) \right| \leq \frac{1}{10} h^4 \max_{|c|\leq 1} q''(a_0 + ch \mid X_i). \quad (118)$$

Then we have

$$\left| \frac{\mathbf{E}\left[(D_i - a_0)^2 K_h(D_i) \mid X_i\right]}{\frac{1}{3} h^2 q(a_0 \mid X_i)} - 1 \right| \cdot \mathbf{1}_{\mathcal{A}_0} \leq \frac{1}{10} h^2 C_2(n). \quad (119)$$

Together with $(D_i - a_0)^2 K_h(D_i) \leq h$, we establish (43).

We now control (42). Similar to (114), we write down the following explicit expression,

$$\mathbf{E}\left[(D_i - a_0)K_h(D_i) \mid X_i\right] = \int_{\left|\frac{D_i-a_0}{h}\right|\leq 1} \left[D_i - a_0\right] \frac{1}{2h} q(D_i \mid X_i) dD_i.$$

Then we have

$$\begin{aligned}\mathbf{E}\left[(D_i - a_0)K_h(D_i) \mid X_i\right] &= \int_{|z|\leq 1} \frac{hz}{2} q(a_0 + hz \mid X_i) dz \\ = \int_{|z|\leq 1} \frac{hz}{2} &\left[q(a_0 \mid X_i) + hzq'(a_0 \mid X_i) + \frac{h^2z^2}{2} q''(a_0 + c(z)hz \mid X_i)\right] dz.\end{aligned}$$

Hence, we have

$$\frac{\mathbf{E}\left[(D_i - a_0)K_h(D_i) \mid X_i\right]}{q(a_0 \mid X_i)} \cdot \mathbf{1}_{\mathcal{A}_0} \leq \frac{1}{3}h^2(C_1(n) + \frac{3}{8}hC_2(n)). \tag{120}$$

Hence, together with $|(D_i - a_0)K_h(D_i)| \leq 1$, we have (42).

E.3.3 Proof of (44) and (45)

By the iterated expectation, we have

$$\begin{aligned}\mathbf{E}\left(W_i^2 K_h^2(D_i)\right) &= \mathbf{E}\left(W_i^2 K_h^2(D_i) \cdot \mathbf{1}_{\mathcal{A}_0}\right) + \mathbf{E}\left(W_i^2 K_h^2(D_i) \cdot \mathbf{1}_{\mathcal{A}_0^c}\right) \\ &= \mathbf{E}\left[\mathbf{E}\left(W_i^2 K_h^2(D_i) \mid X_i\right) \mathbf{1}_{\mathcal{A}_0}\right] + +\mathbf{E}\left(W_i^2 K_h^2(D_i) \cdot \mathbf{1}_{\mathcal{A}_0^c}\right).\end{aligned}$$

We first analyze $\mathbf{E}\left(W_i^2 K_h^2(D_i) \mid X_i\right) \mathbf{1}_{\mathcal{A}_0}$, by noting that

$$\begin{aligned}\mathbf{E}\left(W_i^2 K_h^2(D_i) \mid X_i\right) &= \frac{1}{h}\mathbf{E}\left(W_i^2 K_h(D_i) \mid X_i\right) \\ = \frac{1}{h}&\left(\mathbf{E}\left[(D_i - a_0)^2 K_h(D_i) \mid X_i\right] - \frac{\{\mathbf{E}\left[(D_i - a_0)K_h(D_i) \mid X_i\right]\}^2}{\mathbf{E}\left[K_h(D_i) \mid X_i\right]}\right),\end{aligned} \tag{121}$$

where the last equality follows from the definition of W_i.

Note that

$$\begin{aligned}&\frac{1}{\frac{1}{3}h^2 q(a_0 \mid X_i)} \cdot \frac{\{\mathbf{E}\left[(D_i - a_0)K_h(D_i) \mid X_i\right]\}^2}{\mathbf{E}\left[K_h(D_i) \mid X_i\right]} \cdot \mathbf{1}_{\mathcal{A}_0} \\ &= \frac{3}{h^2} \frac{\{\mathbf{E}\left[(D_i - a_0)K_h(D_i) \mid X_i\right]\}^2}{q^2(a_0 \mid X_i)} \frac{q(a_0 \mid X_i)}{\mathbf{E}\left[K_h(D_i) \mid X_i\right]} \cdot \mathbf{1}_{\mathcal{A}_0}\end{aligned}$$

By the above expression, (40), and (120), we establish

$$\frac{1}{\frac{1}{3}h^2 q(a_0 \mid X_i)} \cdot \frac{\{\mathbf{E}\left[(D_i - a_0)K_h(D_i) \mid X_i\right]\}^2}{\mathbf{E}\left[K_h(D_i) \mid X_i\right]} \cdot \mathbf{1}_{\mathcal{A}_0} \leq \frac{h^2}{3} \frac{\left[(C_1(n) + \frac{3}{8}hC_2(n))\right]^2}{1 - \frac{h^2}{6}C_2(n)}. \tag{122}$$

We apply (121) together with (119) and (122) and establish (44). Since $|W_i|K_h(D_i) \leq 1$, we have

$$\left|\frac{\mathbf{E}\left(W_i^2 K_h^2(D_i)\mathbf{1}_{\mathcal{A}_0^c}\right)}{\frac{1}{3}h\pi(a_0)}\right| \leq \frac{3\mathbf{P}(\mathcal{A}_0^c)}{h\pi(a_0)}. \tag{123}$$

Combining (44) and (123), we establish (45).

E.3.4 Proof of (46)

The proof of (46) is similar to that of (45). We first have the following decomposition,

$$\mathbf{E}W_i(D_i-a_0)K_h(D_i)=\mathbf{E}W_i(D_i-a_0)K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0}+\mathbf{E}W_i(D_i-a_0)K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0^c}.$$

Since $W_i(D_i-a_0)K_h(D_i)\leq h$, we have

$$\left|\frac{\mathbf{E}W_i(D_i-a_0)K_h(D_i)\mathbf{1}_{\mathcal{A}_0^c}}{\frac{2}{3}h^2\pi(a_0)}\right|\leq\frac{\mathbf{P}(\mathcal{A}_0^c)}{\frac{2}{3}h\pi(a_0)}. \tag{124}$$

Note that

$$\begin{aligned}&\mathbf{E}\left[W_i(D_i-a_0)K_h(D_i)\mid X_i\right]\mathbf{1}_{\mathcal{A}_0}\\&=\left(\mathbf{E}\left[(D_i-a_0)^2K_h(D_i)\mid X_i\right]-\frac{\{\mathbf{E}\left[(D_i-a_0)K_h(D_i)\mid X_i\right]\}^2}{\mathbf{E}\left[K_h(D_i)\mid X_i\right]}\right)\mathbf{1}_{\mathcal{A}_0}\\&=\mathbf{E}\left(W_i^2K_h(D_i)\mid X_i\right)\mathbf{1}_{\mathcal{A}_0}.\end{aligned}$$

We apply (44) and (124) to establish (46).

E.3.5 Proof of (47)

Note that

$$\mathbf{E}W_i\frac{(D_i-a_0)^2}{2}K_h(D_i)=\mathbf{E}W_i\frac{(D_i-a_0)^2}{2}K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0}+\mathbf{E}W_i\frac{(D_i-a_0)^2}{2}K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0^c}.$$

For the first term, we apply the iterated expectation and obtain

$$\mathbf{E}W_i\frac{(D_i-a_0)^2}{2}K_h(D_i)\mathbf{1}_{\mathcal{A}_0}=\mathbf{E}\left[\mathbf{E}\left(W_i\frac{(D_i-a_0)^2}{2}K_h(D_i)\mid X_i\right)\mathbf{1}_{\mathcal{A}_0}\right],$$

with

$$\begin{aligned}&\mathbf{E}\left(W_i\frac{(D_i-a_0)^2}{2}K_h(D_i)\mid X_i\right)\\=&\mathbf{E}\left(\frac{(D_i-a_0)^3}{2}K_h(D_i)\mid X_i\right)-l(X_i)\mathbf{E}\left(\frac{(D_i-a_0)^2}{2}K_h(D_i)\mid X_i\right)\\=&\mathbf{E}\left(\frac{(D_i-a_0)^3}{2}K_h(D_i)\mid X_i\right)-\frac{\mathbf{E}\left((D_i-a_0)K_h(D_i)\mid X_i\right)\mathbf{E}\left(\frac{(D_i-a_0)^2}{2}K_h(D_i)\mid X_i\right)}{\mathbf{E}\left(K_h(D_i)\mid X_i\right)}.\end{aligned}$$

Then it is sufficient to control the terms

$$\mathbf{E}\left[(D_i-a_0)^3K_h(D_i)\mid X_i\right]\mathbf{1}_{\mathcal{A}_0},$$

and

$$\frac{\mathbf{E}\left((D_i-a_0)K_h(D_i)\mid X_i\right)\mathbf{E}\left(\frac{(D_i-a_0)^2}{2}K_h(D_i)\mid X_i\right)}{\mathbf{E}\left(K_h(D_i)\mid X_i\right)}\mathbf{1}_{\mathcal{A}_0}. \tag{125}$$

Since

$$\mathbf{E}\left(\frac{(D_i-a_0)^2}{2}K_h(D_i)\mid X_i\right)\leq \frac{h^2}{2}\mathbf{E}\left(K_h(D_i)\mid X_i\right),$$

the term in (125) can be upper bounded by

$$\frac{h^2}{2}\mathbf{E}\left((D_i-a_0)K_h(D_i)\mid X_i\right).$$

It follows from (120) that

$$\frac{h^2}{2}\mathbf{E}\left((D_i-a_0)K_h(D_i)\mid X_i\right)\mathbf{1}_{\mathcal{A}_0}\lesssim q(a_0\mid X_i)\cdot h^4(C_1(n)+\frac{3}{8}hC_2(n)). \tag{126}$$

We control the term $\mathbf{E}\left[(D_i-a_0)^3K_h(D_i)\mid X_i\right]$ in the following.

$$\begin{aligned}\mathbf{E}\left[(D_i-a_0)^3K_h(D_i)\mid X_i\right]&=\int_{|z|\leq 1}\frac{1}{2}h^3z^3q(a_0+hz\mid X_i)dz\\ =\int_{|z|\leq 1}\frac{1}{2}h^3z^3&\left[q(a_0\mid X_i)+hzq'(a_0\mid X_i)+\frac{h^2z^2}{2}q''(a_0+c(z)\mid X_i)\right]dz,\end{aligned}$$

and then have

$$\left|\frac{\mathbf{E}\left[(D_i-a_0)^3K_h(D_i)\mid X_i\right]}{q(a_0\mid X_i)}\right|\cdot\mathbf{1}_{\mathcal{A}_0}\leq\frac{h^4}{5}\left[C_1(n)+\frac{5}{12}hC_2(n)\right]\pi(a_0).$$

Together with (126) and

$$\left|\mathbf{E}W_i\frac{(D_i-a_0)^2}{2}K_h(D_i)\cdot\mathbf{1}_{\mathcal{A}_0^c}\right|\leq h^2\mathbf{P}(\mathcal{A}_0^c),$$

we establish (47).

E.4 Proof of Lemma 2

The proofs rely on the Bernstein inequality (Bennett, 1962), which is restated in the following lemma.

Lemma 4 *Suppose that $\{H_i\}_{1\leq i\leq n}$ are independent zero mean random variables and $|H_i|\leq M$ almost surely. Then we have*

$$\mathbf{P}\left(\left|\sum_{i=1}^{n}H_i\right|\geq T\right)\leq 2\exp\left(-\frac{T^2/2}{\sum_{i=1}^{n}\mathbf{E}H_i^2+MT/3}\right).$$

Proof of (48)

We shall apply Lemma 4 by taking $H_i=K_h(D_i)-\mathbf{E}\left(K_h(D_i)\right)$. By (41), there exists $0<c<1/2$ such that

$$(2-c)\pi(a_0)\leq\mathbf{E}\left(K_h(D_i)\right)\leq(2+c)\pi(a_0). \tag{127}$$

Note that $|K_h(D_i) - \mathbf{E}\left(K_h(D_i)\right)| \leq 1/h$ and

$$\mathbf{E}\left(K_h^2(D_i)\right) = \mathbf{E}\left(K_h(D_i)\right)/h \leq (2+c)\pi(a_0)/h.$$

By (127) and Lemma 4 with $T = t \cdot \max\left\{\sqrt{\frac{n\pi(a_0)}{h}}, \frac{1}{h}\right\}$, we establish (48).

Proof of (49)

By the definition of W_i, the term $\frac{1}{n}\sum_{i=1}^{n} W_i K_h(D_i)$ satisfies

$$\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} W_i K_h(D_i)\right) = 0.$$

Note that $|W_i K_h(D_i)| \leq 2$. By (45), we apply Lemma 4 with $T = t \cdot \max\left\{\sqrt{nh\pi(a_0)}, 1\right\}$ and establish (49).

Proof of (50)

It follows from (42) that

$$\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(D_i - a_0)K_h(D_i)\right) = \mathbf{E}(D_i - a_0)K_h(D_i) \leq \frac{\pi(a_0)}{3}h^2(C_1(n) + \frac{3}{8}hC_2(n)) + \mathbf{P}(\mathcal{A}_0^c).$$

We apply (43) and establish

$$\mathrm{Var}\left((D_i - a_0)K_h(D_i)\right) \leq \frac{1}{h}\mathbf{E}(D_i - a_0)^2 K_h(D_i) \lesssim \frac{1}{3}h\pi(a_0).$$

Note that $|(D_i - a_0)K_h(D_i)| \leq 1$, we apply Lemma 4 with $T = t \cdot \max\left\{\sqrt{nh\pi(a_0)}, 1\right\}$ and establish (50).

Proof of (51)

It follows from (46) that

$$\left|\mathbf{E}\left(W_i(D_i - a_0)K_h(D_i)\right) - \frac{2}{3}h^2\pi(a_0)\right| \leq c\frac{2}{3}h^2\pi(a_0),$$

for some small positive constant $c \in (0,1)$. Note that

$$\mathrm{Var}\left(W_i(D_i - a_0)K_h(D_i)\right) \leq \frac{1}{h}\mathbf{E}\left(W_i^2(D_i - a_0)^2 K_h(D_i)\right) \lesssim h^3\mathbf{E}\left(K_h(D_i)\right),$$

and

$$|W_i(D_i - a_0)K_h(D_i)| \leq h.$$

We apply Lemma 4 with $T = t \cdot \max\left\{\sqrt{nh^3\pi(a_0)}, h\right\}$ and establish (51).

Proof of (52)

The term $\frac{1}{n}\sum_{i=1}^{n} W_i \frac{(D_i - a_0)^2}{2} K_h(D_i)$ satisfies

$$\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} W_i \frac{(D_i - a_0)^2}{2} K_h(D_i)\right) = \mathbf{E}\left[W_i \frac{(D_i - a_0)^2}{2} K_h(D_i)\right]$$
$$\lesssim h^4\left[C_1(n) + hC_2(n)\right] + h^2\mathbf{P}(\mathcal{A}_0^c),$$

where the last inequality follows from (47). Note that

$$\mathrm{Var}\left(W_i\frac{(D_i-a_0)^2}{2}K_h(D_i)\right)\leq\frac{1}{h}\mathbf{E}\left(W_i^2\frac{(D_i-a_0)^4}{4}K_h(D_i)\right)\leq\frac{h^5}{4}\mathbf{E}\left(K_h(D_i)\right),$$

and $\left|W_i\frac{(D_i-a_0)^2}{2}K_h(D_i)\right|\leq h^2$. We apply Lemma 4 with $T=t\cdot\max\left\{\sqrt{nh^5\pi(a_0)},h^2\right\}$ and establish (52).

Proof of (53). Note that both $\left|W_i^2K_h^2(D_i)-\mathbf{E}W_i^2K_h^2(D_i)\right|$ is upper bounded by a constant and

$$\mathrm{Var}(W_i^2K_h^2(D_i))\leq\mathbf{E}W_i^4K_h^4(D_i)\leq\mathbf{E}W_i^2K_h^2(D_i)\lesssim\frac{1}{3}h\pi(a_0),$$

where the second inequality follows from the fact $W_i^2K_h^2(D_i)\leq 1$ and the last inequality follows from (45). We apply Lemma 4 with $T=t\cdot\max\left\{\sqrt{nh\pi(a_0)},1\right\}$ and establish (53).

E.5 Proof of Lemma 3

By Taylor's expansion, we have for $c_1,c_2\in(0,1)$

$$\phi(\mu+\Delta)+\phi(\mu-\Delta)=\phi(\mu)+\Delta\cdot\phi'(\mu)+\frac{\Delta^2}{2}\cdot\phi''(\mu+c_1\Delta)+\phi(\mu)-\Delta\cdot\phi'(\mu)+\frac{\Delta^2}{2}\cdot\phi''(\mu+c_2\Delta)$$

and

$$\frac{1}{\Delta}\int_{\mu-\Delta}^{\mu+\Delta}\phi(t)dt=\frac{1}{\Delta}\int_{\mu-\Delta}^{\mu+\Delta}\left[\phi(\mu)+(t-\mu)\phi'(\mu)+\frac{(t-\mu)^2}{2}\phi''(\mu+c_3(t)(t-\mu))\right]dt$$

where $c_3(t)\in(0,1)$. Hence, we have

$$\begin{aligned}\phi(\mu+\Delta)+\phi(\mu-\Delta)-\frac{1}{\Delta}\int_{\mu-\Delta}^{\mu+\Delta}\phi(t)dt=&\frac{\Delta^2}{2}\cdot\big(\phi''(\mu+c_1\Delta)+\phi''(\mu+c_2\Delta)\big)\\&+\frac{1}{\Delta}\int_{\mu-\Delta}^{\mu+\Delta}\frac{(t-\mu)^2}{2}\phi''(\mu+c_3(t)(t-\mu))dt\end{aligned}$$

Hence, we establish (86). Note that

$$\int_{\mu-\Delta}^{\mu+\Delta}(t-\mu)\,\phi(t)dt=\int_{\mu-\Delta}^{\mu+\Delta}\left[(t-\mu)\phi(\mu)+(t-\mu)^2\phi'(\mu+c_4(t)(t-\mu))\right]dt\tag{128}$$

where $c_4(t)\in(0,1)$. Hence we establish (87). Note that

$$\phi(\mu+\Delta)-\phi(\mu-\Delta)=2\Delta\cdot\phi'(\mu+c_5\Delta),\tag{129}$$

for $c_5\in(-1,1)$. Hence we establish (88).

E.6 Proof of Proposition 1

Note that

$$
\begin{aligned}
\widehat{\sigma}^2-\sigma^2 =& \frac{1}{n}\sum_{i\in\mathcal{I}_a}\left[\left(\epsilon_i-[\widehat{f}^b(D_i)-f(D_i)]-[\widehat{g}^b(X_i)-g(X_i)]\right)^2-\sigma^2\right] \\
&+\frac{1}{n}\sum_{i\in\mathcal{I}_b}\left[\left(\epsilon_i-[\widehat{f}^a(D_i)-f(D_i)]-[\widehat{g}^a(X_i)-g(X_i)]\right)^2-\sigma^2\right].
\end{aligned}
$$

It is sufficient to show that

$$
\frac{1}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}\left[\left(\epsilon_i-[\widehat{f}^b(D_i)-f(D_i)]-[\widehat{g}^b(X_i)-g(X_i)]\right)^2-\sigma^2\right]\overset{p}{\to}0. \tag{130}
$$

For the last hand side of (130), we have the decomposition

$$
\begin{aligned}
&\frac{1}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}(\epsilon_i^2-\sigma^2)+\frac{1}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}\left([\widehat{f}^b(D_i)-f(D_i)]+[\widehat{g}^b(X_i)-g(X_i)]\right)^2 \\
&-\frac{2}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}\epsilon_i\cdot\left([\widehat{f}^b(D_i)-f(D_i)]+[\widehat{g}^b(X_i)-g(X_i)]\right).
\end{aligned} \tag{131}
$$

By the law of large numbers, we have

$$
\frac{1}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}(\epsilon_i^2-\sigma^2)\overset{p}{\to}0. \tag{132}
$$

Define the event

$$
\mathcal{A}_2'=\left\{\mathbf{E}_{X_*}(\widehat{g}^b(X_*)-g(X_*))^2\lesssim\mathrm{Err}^2(\widehat{g}),\ \ \mathbf{E}_{D_*}(\widehat{f}^b(D_*)-f(D_*))^2\lesssim\mathrm{Err}^2(\widehat{f})\right\}
$$

and by the definition of $\mathrm{Err}(\widehat{f})$ and $\mathrm{Err}(\widehat{g})$, we have

$$
\mathbf{P}(\mathcal{A}_2')\geq 1-\min\{n,p\}^{-c}. \tag{133}
$$

In the following analysis, we condition on the data in $\mathcal{I}_b$ and take the conditional expectation as

$$
\mathbf{E}\left[\frac{1}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}\left([\widehat{f}^b(D_i)-f(D_i)]+[\widehat{g}^b(X_i)-g(X_i)]\right)^2\cdot\mathbf{1}_{\mathcal{A}_2'}\mid\mathcal{I}_b\right]\lesssim\mathrm{Err}^2(\widehat{f})+\mathrm{Err}^2(\widehat{g}). \tag{134}
$$

By the Cauchy inequality, we have

$$
\begin{aligned}
&\mathbf{E}\left[\left|\frac{1}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}\epsilon_i\cdot\left([\widehat{f}^b(D_i)-f(D_i)]+[\widehat{g}^b(X_i)-g(X_i)]\right)\cdot\mathbf{1}_{\mathcal{A}_2'}\right|^2\mid\mathcal{I}_b\right] \\
&\leq\mathbf{E}\left[\left(\frac{1}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}\epsilon_i^2\right)\cdot\left(\frac{1}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}\left([\widehat{f}^b(D_i)-f(D_i)]+[\widehat{g}^b(X_i)-g(X_i)]\right)^2\cdot\mathbf{1}_{\mathcal{A}_2'}\right)\mid\mathcal{I}_b\right] \\
&\lesssim\sigma^2\cdot\left(\mathrm{Err}^2(\widehat{f})+\mathrm{Err}^2(\widehat{g})\right),
\end{aligned}
$$

where the least inequality follows from (134) and $\mathbf{E}(\epsilon_i^2 \mid D_i, X_i) = \sigma^2$. By the Markov inequality, we establish that, with probability larger than $1 - \frac{1}{t}$ for some $t > 1$,

$$\begin{aligned}
&\left|\frac{1}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}\left([\widehat{f}^b(D_i) - f(D_i)] + [\widehat{g}^b(X_i) - g(X_i)]\right)^2\right| \cdot \mathbf{1}_{\mathcal{A}_2'} \\
&+ \left|\frac{2}{|\mathcal{I}_a|}\sum_{i\in\mathcal{I}_a}\epsilon_i \cdot \left([\widehat{f}^b(D_i) - f(D_i)] + [\widehat{g}^b(X_i) - g(X_i)]\right)\right| \cdot \mathbf{1}_{\mathcal{A}_2'} \\
&\lesssim t\left(\mathrm{Err}^2(\widehat{f}) + \mathrm{Err}^2(\widehat{g})\right) + \sqrt{t\left(\mathrm{Err}^2(\widehat{f}) + \mathrm{Err}^2(\widehat{g})\right)}.
\end{aligned}$$

Combined with (132) and (133), we establish $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$ if $\max\{\mathrm{Err}(\widehat{f}), \mathrm{Err}(\widehat{g})\} \to 0$.

Appendix Appendix F. Additional Simulation Results

F.1 Setting 1-4 and Nonlinear Treatment Model

In this section, we present complete simulation results for Setting 1-4 and the nonlinear treatment model. With p fixed at 1500, The sample sizes are varied across {500,1000,1500, 2000} and a_0 is varied across {-1.25,-0.5,0.1,0.25,1}. We consider two generating models for f and g_1 as follows:

- $f(d) = 2\exp(-d/2)$ and $g_1(x) = 1.5\sin(x)$;
- $f(d) = 1.5\sin(d)$ and $g_1(x) = 2\exp(-x/2)$.

The complete results for Setting 1 are summarised in Table A1 and Table A2. Similar to the results presented in the main paper, our DLL method achieves desired coverage, and the CI length is close to the confidence interval by the oracle estimator. Besides, our DLL method outperforms the plug-in method in terms of coverage since the DLL estimator has a smaller bias. The coverage for the plug-in estimator is relatively good at the boundary {-1.25,1} since only a few samples are used with the chosen bandwidth, and the standard error for the plug-in estimator is large, leading to a wide CI.

The complete results for Setting 2 and 3 are in Table A3 and Table A4, respectively. The results for Setting 2 and Setting 3 are similar to those for Setting 1. For Setting 4, $(D_i, X_i^\intercal)^\intercal$ follows a t-distribution and we vary the degree of freedom in {10,15}. The results are reported in Tables A5 and A6. In Settings 3 and 4, we test the robustness of our proposed method to the violation of assumption in (A2). The results demonstrate that our proposed DLL method still corrects the bias of the plug-in estimator and attains the desired coverage level, with the CI length similar to the confidence interval by the oracle estimator.

The complete results of the non-linear treatment model are presented in Table A7. We see that DLL-S correct more bias than the DLL estimator, and the coverage for DLL-S improves along with this additional bias-correction. However, the bias for DLL is still smaller than the plug-in estimator, and better coverage is obtained.

Setting 1, exactly sparse: $f(d) = 1.5\sin(d)$

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	-1.00	500	0.00	0.13	0.03	0.49	0.48	0.49	0.95	0.93	0.94	2.01	1.90	1.90
		1000	0.02	0.10	0.03	0.41	0.40	0.40	0.96	0.94	0.95	1.68	1.60	1.59
		1500	0.02	0.13	0.01	0.38	0.37	0.36	0.94	0.94	0.96	1.51	1.44	1.43
		2000	0.01	0.09	0.02	0.37	0.36	0.36	0.96	0.94	0.95	1.42	1.35	1.35
-0.50	-0.56	500	0.00	0.15	0.06	0.40	0.39	0.38	0.95	0.94	0.94	1.59	1.50	1.50
		1000	0.02	0.11	0.04	0.32	0.31	0.31	0.96	0.94	0.95	1.32	1.25	1.24
		1500	0.01	0.13	0.01	0.29	0.29	0.29	0.97	0.95	0.96	1.19	1.14	1.13
		2000	0.03	0.08	0.03	0.28	0.28	0.27	0.95	0.94	0.94	1.12	1.07	1.07
0.10	0.74	500	0.17	0.30	0.06	0.42	0.42	0.41	0.92	0.85	0.92	1.61	1.53	1.53
		1000	0.08	0.21	0.04	0.32	0.31	0.32	0.95	0.91	0.95	1.34	1.28	1.27
		1500	0.07	0.19	0.04	0.30	0.30	0.30	0.95	0.88	0.95	1.21	1.16	1.15
		2000	0.07	0.18	0.05	0.29	0.28	0.28	0.96	0.88	0.95	1.13	1.09	1.08
0.25	0.94	500	0.17	0.30	0.08	0.43	0.41	0.42	0.92	0.87	0.93	1.67	1.58	1.58
		1000	0.08	0.22	0.04	0.35	0.34	0.35	0.93	0.89	0.94	1.38	1.33	1.32
		1500	0.08	0.20	0.04	0.30	0.29	0.30	0.96	0.91	0.95	1.25	1.20	1.19
		2000	0.07	0.17	0.04	0.29	0.29	0.29	0.95	0.89	0.94	1.17	1.13	1.12
1.00	0.81	500	0.08	0.18	0.00	0.61	0.58	0.58	0.93	0.90	0.93	2.33	2.19	2.19
		1000	0.00	0.12	0.04	0.48	0.47	0.47	0.96	0.94	0.95	1.94	1.83	1.83
		1500	0.03	0.14	0.01	0.44	0.43	0.42	0.95	0.93	0.95	1.75	1.66	1.66
		2000	0.01	0.10	0.02	0.40	0.39	0.39	0.95	0.93	0.95	1.63	1.55	1.55

Setting 1, approximately sparse: $f(d) = 1.5\sin(d)$

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	-1.00	500	0.02	0.08	0.01	0.47	0.47	0.50	0.96	0.94	0.94	1.89	1.78	1.92
		1000	0.04	0.06	0.01	0.39	0.38	0.39	0.96	0.96	0.96	1.62	1.56	1.61
		1500	0.02	0.06	0.01	0.35	0.34	0.36	0.98	0.95	0.96	1.47	1.41	1.44
		2000	0.03	0.04	0.01	0.34	0.34	0.34	0.95	0.95	0.95	1.37	1.31	1.34
-0.50	-0.56	500	0.03	0.08	0.02	0.38	0.37	0.38	0.95	0.95	0.95	1.48	1.41	1.50
		1000	0.03	0.07	0.02	0.33	0.32	0.32	0.94	0.93	0.94	1.28	1.23	1.26
		1500	0.03	0.06	0.03	0.28	0.28	0.28	0.96	0.95	0.96	1.16	1.11	1.14
		2000	0.02	0.06	0.01	0.28	0.27	0.26	0.96	0.93	0.96	1.08	1.04	1.06
0.10	0.74	500	0.19	0.29	0.03	0.38	0.37	0.39	0.92	0.86	0.96	1.51	1.44	1.53
		1000	0.15	0.25	0.06	0.32	0.31	0.32	0.93	0.88	0.95	1.30	1.24	1.28
		1500	0.11	0.20	0.03	0.31	0.30	0.31	0.94	0.88	0.93	1.17	1.13	1.15
		2000	0.10	0.18	0.04	0.27	0.27	0.27	0.95	0.90	0.94	1.10	1.05	1.08
0.25	0.94	500	0.21	0.31	0.06	0.40	0.38	0.40	0.92	0.86	0.95	1.55	1.48	1.58
		1000	0.18	0.28	0.09	0.32	0.31	0.32	0.92	0.87	0.95	1.34	1.28	1.33
		1500	0.12	0.21	0.04	0.32	0.31	0.31	0.92	0.86	0.94	1.22	1.17	1.20
		2000	0.12	0.20	0.06	0.28	0.27	0.28	0.95	0.87	0.95	1.13	1.09	1.11
1.00	0.81	500	0.12	0.19	0.01	0.58	0.57	0.58	0.92	0.90	0.93	2.19	2.07	2.21
		1000	0.04	0.12	0.02	0.46	0.45	0.46	0.96	0.94	0.95	1.88	1.79	1.86
		1500	0.01	0.07	0.04	0.42	0.42	0.44	0.95	0.94	0.95	1.70	1.62	1.67
		2000	0.05	0.12	0.01	0.40	0.39	0.39	0.94	0.93	0.95	1.58	1.52	1.55

Table A1: Comparison of `DLL`, plug-in (`Plug`), oracle (`Orac`) estimators in Setting 1 when $f(d) = 1.5\sin(d)$, across different sample sizes n and evaluation points a_0 with $p = 1500$. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" and "SE" report the absolute bias and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

Setting 1, exactly sparse: $f(d) = 2\exp(-d/2)$

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	-0.41	500	0.03	0.14	0.01	0.30	0.29	0.25	0.96	0.92	0.95	1.19	1.11	0.94
		1000	0.05	0.16	0.02	0.24	0.23	0.18	0.94	0.88	0.95	0.92	0.87	0.70
		1500	0.02	0.13	0.01	0.21	0.21	0.15	0.94	0.90	0.95	0.82	0.78	0.62
		2000	0.02	0.12	0.01	0.20	0.19	0.15	0.95	0.89	0.95	0.74	0.71	0.56
-0.50	-0.52	500	0.00	0.01	0.02	0.24	0.23	0.20	0.95	0.94	0.92	0.93	0.87	0.73
		1000	0.01	0.00	0.03	0.19	0.19	0.15	0.93	0.93	0.92	0.73	0.69	0.55
		1500	0.02	0.00	0.02	0.16	0.16	0.12	0.95	0.94	0.92	0.65	0.62	0.48
		2000	0.02	0.00	0.03	0.15	0.15	0.11	0.95	0.95	0.93	0.59	0.56	0.44
0.10	-0.40	500	0.01	0.13	0.01	0.25	0.24	0.20	0.94	0.90	0.94	0.94	0.89	0.75
		1000	0.01	0.10	0.00	0.19	0.18	0.15	0.95	0.90	0.94	0.74	0.70	0.56
		1500	0.01	0.09	0.00	0.16	0.16	0.12	0.96	0.92	0.95	0.66	0.63	0.49
		2000	0.02	0.07	0.01	0.15	0.15	0.12	0.96	0.92	0.94	0.60	0.57	0.45
0.25	-0.35	500	0.01	0.14	0.01	0.26	0.25	0.21	0.95	0.90	0.95	0.97	0.92	0.77
		1000	0.01	0.12	0.00	0.20	0.19	0.15	0.95	0.88	0.96	0.76	0.73	0.58
		1500	0.02	0.11	0.01	0.17	0.16	0.13	0.95	0.91	0.95	0.68	0.65	0.51
		2000	0.02	0.09	0.02	0.16	0.15	0.12	0.95	0.92	0.95	0.62	0.59	0.46
1.00	-0.15	500	0.01	0.20	0.03	0.37	0.35	0.29	0.94	0.91	0.93	1.37	1.29	1.07
		1000	0.04	0.14	0.02	0.28	0.27	0.21	0.95	0.92	0.95	1.06	1.01	0.80
		1500	0.02	0.14	0.02	0.25	0.24	0.19	0.96	0.90	0.95	0.94	0.90	0.71
		2000	0.01	0.14	0.01	0.23	0.23	0.17	0.93	0.89	0.94	0.86	0.83	0.64

Setting 1, approximately sparse: $f(d) = 2\exp(-d/2)$

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	-0.41	500	0.01	0.08	0.01	0.29	0.28	0.24	0.95	0.94	0.96	1.11	1.03	0.92
		1000	0.01	0.08	0.01	0.22	0.22	0.19	0.96	0.93	0.95	0.89	0.84	0.71
		1500	0.02	0.09	0.02	0.21	0.21	0.16	0.95	0.93	0.94	0.80	0.76	0.62
		2000	0.01	0.06	0.01	0.19	0.18	0.15	0.94	0.93	0.95	0.73	0.69	0.57
-0.50	-0.52	500	0.05	0.08	0.03	0.23	0.22	0.19	0.92	0.89	0.92	0.87	0.82	0.72
		1000	0.03	0.04	0.03	0.19	0.18	0.15	0.93	0.92	0.92	0.70	0.67	0.56
		1500	0.02	0.03	0.02	0.17	0.16	0.12	0.92	0.91	0.95	0.63	0.60	0.49
		2000	0.01	0.02	0.01	0.15	0.14	0.12	0.95	0.92	0.94	0.57	0.55	0.45
0.10	-0.40	500	0.01	0.15	0.02	0.23	0.23	0.20	0.95	0.90	0.95	0.89	0.84	0.74
		1000	0.00	0.13	0.00	0.19	0.18	0.15	0.93	0.86	0.94	0.71	0.68	0.57
		1500	0.01	0.12	0.00	0.16	0.15	0.13	0.95	0.86	0.94	0.64	0.61	0.50
		2000	0.01	0.10	0.01	0.15	0.15	0.11	0.95	0.86	0.96	0.58	0.56	0.46
0.25	-0.35	500	0.00	0.16	0.03	0.23	0.22	0.21	0.96	0.90	0.93	0.91	0.86	0.76
		1000	0.00	0.15	0.00	0.20	0.19	0.16	0.92	0.86	0.94	0.74	0.70	0.58
		1500	0.01	0.14	0.00	0.16	0.16	0.13	0.96	0.85	0.95	0.66	0.63	0.52
		2000	0.02	0.11	0.01	0.15	0.15	0.12	0.96	0.88	0.95	0.60	0.58	0.47
1.00	-0.15	500	0.01	0.23	0.02	0.35	0.33	0.30	0.94	0.88	0.92	1.28	1.22	1.07
		1000	0.02	0.17	0.02	0.29	0.27	0.23	0.94	0.89	0.93	1.03	0.98	0.82
		1500	0.01	0.16	0.01	0.23	0.23	0.17	0.93	0.89	0.95	0.92	0.88	0.72
		2000	0.00	0.15	0.01	0.22	0.21	0.18	0.95	0.87	0.94	0.84	0.81	0.66

Table A2: Comparison of `DLL`, plug-in (`Plug`), oracle (`Orac`) estimators in Setting 1 when $f(d) = 2\exp(-d/2)$, across different sample sizes n and evaluation points a_0 with $p = 1500$. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" and "SE" report the absolute bias and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

Setting 2, exactly sparse: $f(d) = 1.5\sin(d)$

			Bias			SE			Coverage			CI Length		
a_0	True	n	DLL	Plug	Orac	DLL	Plug	Orac	DLL	Plug	Orac	DLL	Plug	Orac
-1.25	0.47	500	0.12	0.38	0.06	0.48	0.47	0.48	0.94	0.87	0.94	1.92	1.84	1.89
		1000	0.04	0.26	0.02	0.41	0.41	0.42	0.95	0.90	0.94	1.64	1.58	1.58
		1500	0.03	0.21	0.02	0.39	0.39	0.38	0.96	0.90	0.95	1.49	1.44	1.44
		2000	0.03	0.19	0.00	0.34	0.34	0.33	0.94	0.92	0.94	1.39	1.36	1.35
-0.50	1.32	500	0.19	0.44	0.03	0.37	0.37	0.38	0.91	0.77	0.94	1.49	1.43	1.47
		1000	0.07	0.30	0.01	0.31	0.31	0.31	0.96	0.83	0.94	1.29	1.25	1.25
		1500	0.06	0.25	0.02	0.29	0.29	0.29	0.95	0.85	0.95	1.17	1.13	1.14
		2000	0.04	0.22	0.01	0.29	0.29	0.28	0.94	0.86	0.95	1.10	1.07	1.07
0.10	1.49	500	0.21	0.46	0.02	0.39	0.38	0.39	0.91	0.72	0.94	1.51	1.45	1.49
		1000	0.07	0.31	0.00	0.34	0.33	0.33	0.93	0.83	0.93	1.31	1.27	1.27
		1500	0.05	0.26	0.01	0.31	0.30	0.29	0.94	0.86	0.95	1.18	1.15	1.15
		2000	0.05	0.24	0.02	0.29	0.28	0.28	0.94	0.86	0.96	1.12	1.08	1.08
0.25	1.45	500	0.20	0.45	0.00	0.41	0.39	0.39	0.91	0.77	0.94	1.56	1.50	1.55
		1000	0.07	0.31	0.01	0.35	0.34	0.35	0.94	0.83	0.94	1.35	1.32	1.32
		1500	0.07	0.27	0.03	0.31	0.31	0.30	0.96	0.84	0.96	1.22	1.19	1.18
		2000	0.05	0.24	0.02	0.28	0.28	0.28	0.95	0.87	0.96	1.15	1.12	1.12
1.00	0.81	500	0.17	0.38	0.00	0.52	0.51	0.54	0.96	0.90	0.94	2.19	2.06	2.15
		1000	0.05	0.25	0.04	0.50	0.49	0.50	0.95	0.89	0.93	1.89	1.81	1.83
		1500	0.02	0.19	0.04	0.44	0.43	0.43	0.94	0.92	0.94	1.71	1.65	1.65
		2000	0.03	0.18	0.00	0.41	0.40	0.40	0.95	0.92	0.95	1.60	1.56	1.56

Setting 2, exactly sparse: $f(d) = 2\exp(-d/2)$

			Bias			SE			Coverage			CI Length		
a_0	True	n	DLL	Plug	Orac	DLL	Plug	Orac	DLL	Plug	Orac	DLL	Plug	Orac
-1.25	-1.87	500	0.04	0.23	0.03	0.44	0.43	0.41	0.95	0.90	0.94	1.70	1.62	1.57
		1000	0.01	0.16	0.01	0.37	0.36	0.35	0.95	0.92	0.94	1.43	1.39	1.34
		1500	0.01	0.16	0.02	0.34	0.33	0.32	0.94	0.91	0.93	1.30	1.27	1.24
		2000	0.00	0.13	0.00	0.30	0.29	0.29	0.96	0.95	0.96	1.21	1.18	1.15
-0.50	-1.28	500	0.04	0.27	0.01	0.37	0.36	0.32	0.93	0.83	0.94	1.34	1.29	1.24
		1000	0.01	0.19	0.01	0.30	0.29	0.27	0.94	0.89	0.94	1.13	1.09	1.06
		1500	0.04	0.20	0.03	0.27	0.27	0.26	0.94	0.85	0.94	1.02	1.00	0.97
		2000	0.01	0.13	0.01	0.23	0.22	0.22	0.97	0.93	0.96	0.95	0.92	0.91
0.10	-0.95	500	0.07	0.32	0.01	0.34	0.34	0.32	0.95	0.83	0.95	1.37	1.31	1.26
		1000	0.02	0.22	0.00	0.30	0.30	0.27	0.93	0.85	0.95	1.14	1.10	1.07
		1500	0.00	0.17	0.01	0.25	0.25	0.24	0.96	0.90	0.96	1.04	1.01	0.99
		2000	0.01	0.14	0.01	0.24	0.24	0.23	0.96	0.92	0.95	0.97	0.94	0.92
0.25	-0.88	500	0.06	0.31	0.01	0.36	0.35	0.34	0.94	0.86	0.94	1.41	1.35	1.30
		1000	0.01	0.22	0.00	0.30	0.30	0.28	0.96	0.88	0.96	1.18	1.14	1.10
		1500	0.02	0.19	0.02	0.28	0.27	0.26	0.95	0.89	0.95	1.07	1.04	1.02
		2000	0.01	0.14	0.02	0.24	0.24	0.23	0.96	0.92	0.97	1.00	0.97	0.95
1.00	-0.61	500	0.05	0.27	0.01	0.48	0.47	0.45	0.95	0.92	0.95	1.97	1.87	1.81
		1000	0.02	0.17	0.02	0.41	0.40	0.40	0.95	0.92	0.95	1.65	1.58	1.54
		1500	0.00	0.16	0.00	0.37	0.37	0.36	0.95	0.92	0.95	1.50	1.45	1.42
		2000	0.01	0.15	0.00	0.37	0.36	0.36	0.94	0.92	0.94	1.39	1.36	1.33

Table A3: Comparison of DLL, plug-in (Plug), oracle (Orac) estimators in Setting 2, across different sample sizes n and evaluation points a_0 with $p = 1500$. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" and "SE" report the absolute bias and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

Setting 3, exactly sparse: $f(d) = 1.5\sin(d)$

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	0.47	500	0.12	0.32	0.02	0.72	0.71	0.70	0.95	0.92	0.94	2.85	2.77	2.62
		1000	0.02	0.19	0.02	0.62	0.62	0.57	0.96	0.92	0.96	2.42	2.36	2.23
		1500	0.03	0.17	0.00	0.61	0.60	0.56	0.94	0.90	0.94	2.19	2.16	2.04
		2000	0.01	0.13	0.02	0.51	0.51	0.49	0.96	0.94	0.95	2.06	2.02	1.93
-0.50	1.32	500	0.10	0.24	0.02	0.77	0.76	0.70	0.93	0.91	0.94	2.82	2.75	2.60
		1000	0.00	0.14	0.01	0.62	0.60	0.55	0.96	0.95	0.96	2.41	2.38	2.26
		1500	0.04	0.17	0.04	0.57	0.57	0.52	0.95	0.95	0.96	2.19	2.18	2.06
		2000	0.02	0.11	0.01	0.48	0.48	0.45	0.98	0.96	0.96	2.05	2.04	1.94
0.10	1.49	500	0.12	0.24	0.01	0.71	0.70	0.69	0.94	0.92	0.93	2.81	2.73	2.60
		1000	0.05	0.19	0.05	0.63	0.63	0.60	0.95	0.93	0.95	2.42	2.39	2.26
		1500	0.04	0.15	0.02	0.57	0.56	0.55	0.96	0.94	0.95	2.19	2.16	2.05
		2000	0.01	0.10	0.00	0.54	0.54	0.52	0.95	0.94	0.93	2.05	2.01	1.92
0.25	1.45	500	0.08	0.21	0.00	0.72	0.71	0.68	0.94	0.93	0.94	2.79	2.73	2.59
		1000	0.04	0.17	0.03	0.62	0.62	0.58	0.95	0.94	0.95	2.41	2.38	2.25
		1500	0.03	0.15	0.02	0.56	0.55	0.52	0.95	0.93	0.94	2.18	2.14	2.04
		2000	0.02	0.09	0.01	0.50	0.50	0.49	0.96	0.95	0.95	2.05	2.02	1.92
1.00	0.81	500	0.09	0.27	0.02	0.70	0.69	0.65	0.96	0.94	0.95	2.82	2.72	2.61
		1000	0.03	0.17	0.00	0.60	0.59	0.58	0.97	0.95	0.96	2.42	2.35	2.23
		1500	0.04	0.17	0.03	0.58	0.58	0.56	0.95	0.94	0.94	2.19	2.14	2.04
		2000	0.02	0.14	0.03	0.52	0.51	0.48	0.95	0.94	0.95	2.05	2.02	1.93

Setting 3, exactly sparse: $f(d) = 2\exp(-d/2)$

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	-1.87	500	0.07	0.15	0.02	0.45	0.44	0.45	0.96	0.92	0.94	1.80	1.76	1.74
		1000	0.05	0.11	0.01	0.42	0.41	0.42	0.93	0.92	0.92	1.55	1.53	1.51
		1500	0.02	0.11	0.01	0.38	0.37	0.37	0.95	0.92	0.93	1.41	1.39	1.37
		2000	0.04	0.08	0.02	0.35	0.36	0.34	0.94	0.93	0.96	1.33	1.32	1.29
-0.50	-1.28	500	0.00	0.13	0.03	0.46	0.46	0.45	0.95	0.93	0.94	1.78	1.75	1.75
		1000	0.01	0.13	0.01	0.40	0.40	0.38	0.95	0.92	0.95	1.55	1.52	1.50
		1500	0.01	0.11	0.01	0.36	0.35	0.36	0.95	0.95	0.94	1.41	1.39	1.37
		2000	0.02	0.08	0.00	0.35	0.34	0.33	0.95	0.95	0.95	1.33	1.32	1.29
0.10	-0.95	500	0.05	0.17	0.01	0.49	0.49	0.46	0.94	0.91	0.94	1.79	1.75	1.76
		1000	0.01	0.12	0.01	0.41	0.40	0.40	0.94	0.94	0.93	1.54	1.52	1.51
		1500	0.03	0.08	0.01	0.35	0.35	0.35	0.95	0.94	0.95	1.41	1.39	1.38
		2000	0.01	0.10	0.01	0.34	0.34	0.34	0.94	0.93	0.93	1.33	1.31	1.28
0.25	-0.88	500	0.07	0.20	0.04	0.48	0.47	0.47	0.94	0.92	0.94	1.79	1.75	1.76
		1000	0.03	0.15	0.02	0.38	0.38	0.38	0.97	0.95	0.95	1.54	1.52	1.50
		1500	0.02	0.09	0.01	0.36	0.36	0.34	0.95	0.93	0.95	1.41	1.39	1.38
		2000	0.00	0.11	0.02	0.34	0.33	0.32	0.95	0.94	0.95	1.33	1.31	1.29
1.00	-0.61	500	0.04	0.14	0.03	0.47	0.46	0.46	0.94	0.91	0.95	1.78	1.73	1.77
		1000	0.02	0.12	0.01	0.41	0.40	0.39	0.94	0.92	0.95	1.54	1.50	1.50
		1500	0.01	0.11	0.02	0.37	0.37	0.36	0.95	0.93	0.95	1.41	1.38	1.37
		2000	0.00	0.10	0.02	0.36	0.35	0.34	0.94	0.92	0.94	1.33	1.31	1.29

Table A4: Comparison of `DLL`, plug-in (`Plug`), oracle (`Orac`) estimators in Setting 3, across different sample sizes n and evaluation points a_0 with $p = 1500$. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" and "SE" report the absolute bias and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

Setting 4, exactly sparse: $f(d) = 1.5\sin(d)$ and df=10

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	0.47	500	0.13	0.27	0.00	0.48	0.48	0.49	0.94	0.88	0.94	1.84	1.76	1.82
		1000	0.04	0.13	0.00	0.41	0.41	0.41	0.95	0.92	0.92	1.55	1.47	1.46
		1500	0.01	0.04	0.03	0.33	0.33	0.31	0.97	0.95	0.97	1.37	1.31	1.27
		2000	0.02	0.00	0.04	0.33	0.33	0.32	0.94	0.92	0.93	1.28	1.21	1.17
-0.50	1.32	500	0.16	0.33	0.01	0.36	0.36	0.36	0.92	0.82	0.95	1.41	1.35	1.41
		1000	0.08	0.22	0.01	0.30	0.29	0.30	0.96	0.86	0.94	1.19	1.11	1.11
		1500	0.05	0.16	0.01	0.27	0.27	0.25	0.94	0.88	0.95	1.05	0.98	0.97
		2000	0.03	0.13	0.01	0.23	0.23	0.22	0.96	0.91	0.95	0.98	0.92	0.90
0.10	1.49	500	0.17	0.36	0.04	0.34	0.33	0.32	0.92	0.78	0.95	1.33	1.27	1.32
		1000	0.10	0.28	0.03	0.29	0.28	0.28	0.94	0.80	0.95	1.13	1.06	1.06
		1500	0.06	0.23	0.02	0.24	0.24	0.23	0.96	0.83	0.97	1.00	0.94	0.92
		2000	0.04	0.21	0.01	0.23	0.23	0.23	0.95	0.83	0.93	0.93	0.88	0.85
0.25	1.45	500	0.18	0.38	0.05	0.35	0.35	0.35	0.89	0.76	0.94	1.34	1.28	1.34
		1000	0.10	0.29	0.01	0.29	0.28	0.28	0.93	0.78	0.93	1.14	1.07	1.08
		1500	0.06	0.24	0.01	0.24	0.24	0.24	0.96	0.83	0.96	1.01	0.95	0.94
		2000	0.07	0.24	0.03	0.23	0.23	0.21	0.95	0.80	0.97	0.94	0.89	0.86
1.00	0.81	500	0.12	0.30	0.04	0.42	0.41	0.43	0.96	0.86	0.94	1.64	1.54	1.64
		1000	0.09	0.26	0.01	0.34	0.33	0.33	0.96	0.89	0.94	1.38	1.31	1.31
		1500	0.05	0.22	0.01	0.31	0.31	0.30	0.94	0.86	0.95	1.23	1.16	1.14
		2000	0.05	0.22	0.00	0.28	0.28	0.27	0.97	0.87	0.93	1.15	1.08	1.05

Setting 4, exactly sparse: $f(d) = 2\exp(-d/2)$ and df=10

			Bias			SE			Coverage			CI Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	-1.87	500	0.05	0.19	0.04	0.44	0.44	0.45	0.95	0.93	0.95	1.80	1.70	1.81
		1000	0.06	0.20	0.04	0.40	0.39	0.40	0.95	0.90	0.95	1.61	1.52	1.58
		1500	0.02	0.15	0.02	0.37	0.37	0.37	0.95	0.92	0.94	1.49	1.41	1.45
		2000	0.03	0.15	0.02	0.33	0.33	0.34	0.97	0.93	0.96	1.41	1.34	1.38
-0.50	-1.28	500	0.03	0.22	0.04	0.35	0.35	0.36	0.94	0.89	0.94	1.38	1.31	1.39
		1000	0.03	0.20	0.01	0.31	0.30	0.30	0.96	0.89	0.95	1.23	1.16	1.20
		1500	0.03	0.18	0.01	0.29	0.29	0.29	0.96	0.90	0.94	1.14	1.08	1.11
		2000	0.02	0.16	0.01	0.26	0.25	0.27	0.97	0.92	0.95	1.08	1.01	1.05
0.10	-0.95	500	0.07	0.26	0.01	0.34	0.33	0.33	0.94	0.88	0.95	1.31	1.25	1.32
		1000	0.04	0.19	0.00	0.28	0.28	0.28	0.96	0.89	0.96	1.17	1.10	1.14
		1500	0.01	0.15	0.00	0.27	0.26	0.28	0.96	0.90	0.94	1.08	1.01	1.05
		2000	0.03	0.15	0.01	0.27	0.26	0.27	0.94	0.87	0.93	1.02	0.96	1.00
0.25	-0.88	500	0.08	0.26	0.01	0.33	0.32	0.32	0.93	0.86	0.96	1.32	1.26	1.33
		1000	0.02	0.17	0.01	0.30	0.29	0.30	0.96	0.90	0.94	1.18	1.11	1.15
		1500	0.01	0.14	0.00	0.26	0.26	0.27	0.97	0.91	0.96	1.09	1.03	1.06
		2000	0.01	0.14	0.01	0.25	0.25	0.26	0.96	0.91	0.95	1.04	0.97	1.01
1.00	-0.61	500	0.01	0.13	0.03	0.41	0.40	0.42	0.95	0.92	0.94	1.61	1.49	1.61
		1000	0.01	0.09	0.01	0.35	0.35	0.36	0.98	0.93	0.93	1.44	1.35	1.41
		1500	0.00	0.08	0.01	0.32	0.32	0.33	0.96	0.95	0.95	1.33	1.25	1.29
		2000	0.01	0.06	0.01	0.30	0.29	0.30	0.96	0.94	0.94	1.26	1.18	1.22

Table A5: Comparison of `DLL`, plug-in (`Plug`), oracle (`Orac`) estimators in Setting 4 with df = 10, across different sample sizes n and evaluation points a_0 with $p = 1500$. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" and "SE" report the absolute bias and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

Setting 4, exactly sparse: $f(d) = 1.5\sin(d)$ and df=15

			Bias			SE			Coverage			CI Length		
a_0	True	n	DLL	Plug	Orac	DLL	Plug	Orac	DLL	Plug	Orac	DLL	Plug	Orac
-1.25	0.47	500	0.17	0.36	0.01	0.48	0.47	0.48	0.94	0.88	0.97	1.97	1.91	1.99
		1000	0.09	0.23	0.04	0.44	0.43	0.41	0.94	0.90	0.95	1.69	1.62	1.62
		1500	0.01	0.11	0.03	0.39	0.38	0.38	0.96	0.94	0.94	1.53	1.48	1.45
		2000	0.00	0.08	0.02	0.33	0.33	0.33	0.97	0.96	0.95	1.41	1.36	1.34
-0.50	1.32	500	0.19	0.39	0.05	0.36	0.35	0.37	0.93	0.78	0.94	1.47	1.41	1.47
		1000	0.07	0.24	0.02	0.31	0.30	0.31	0.95	0.87	0.95	1.26	1.21	1.22
		1500	0.04	0.18	0.01	0.28	0.28	0.28	0.96	0.91	0.95	1.14	1.09	1.09
		2000	0.02	0.14	0.01	0.26	0.25	0.25	0.97	0.93	0.95	1.06	1.01	1.00
0.10	1.49	500	0.15	0.36	0.00	0.36	0.36	0.35	0.92	0.77	0.96	1.39	1.33	1.39
		1000	0.07	0.27	0.02	0.30	0.30	0.30	0.95	0.82	0.95	1.20	1.15	1.15
		1500	0.07	0.26	0.03	0.26	0.25	0.26	0.96	0.87	0.94	1.08	1.04	1.03
		2000	0.06	0.23	0.03	0.25	0.24	0.23	0.95	0.84	0.95	1.01	0.96	0.95
0.25	1.45	500	0.17	0.38	0.02	0.36	0.36	0.34	0.90	0.77	0.95	1.41	1.35	1.42
		1000	0.07	0.27	0.01	0.31	0.31	0.30	0.95	0.84	0.95	1.21	1.16	1.17
		1500	0.04	0.23	0.01	0.26	0.26	0.25	0.97	0.87	0.96	1.09	1.05	1.04
		2000	0.04	0.22	0.01	0.26	0.26	0.25	0.94	0.88	0.95	1.02	0.97	0.96
1.00	0.81	500	0.16	0.34	0.03	0.43	0.43	0.46	0.95	0.85	0.94	1.74	1.64	1.73
		1000	0.07	0.26	0.00	0.39	0.38	0.38	0.94	0.87	0.94	1.49	1.43	1.44
		1500	0.05	0.23	0.00	0.34	0.33	0.34	0.95	0.90	0.94	1.35	1.29	1.28
		2000	0.04	0.20	0.00	0.31	0.30	0.30	0.96	0.90	0.95	1.25	1.20	1.18

Setting 4, exactly sparse: $f(d) = 2\exp(-d/2)$ and df=15

			Bias			SE			Coverage			CI Length		
a_0	True	n	DLL	Plug	Orac	DLL	Plug	Orac	DLL	Plug	Orac	DLL	Plug	Orac
-1.25	-1.87	500	0.04	0.21	0.02	0.46	0.45	0.44	0.93	0.89	0.95	1.78	1.70	1.73
		1000	0.01	0.16	0.00	0.40	0.39	0.40	0.96	0.92	0.93	1.55	1.49	1.50
		1500	0.01	0.16	0.01	0.36	0.36	0.35	0.96	0.91	0.96	1.44	1.39	1.38
		2000	0.01	0.14	0.01	0.32	0.32	0.31	0.96	0.94	0.96	1.34	1.30	1.30
-0.50	-1.28	500	0.06	0.27	0.01	0.33	0.33	0.33	0.95	0.87	0.94	1.34	1.31	1.32
		1000	0.00	0.18	0.01	0.30	0.30	0.29	0.95	0.89	0.95	1.16	1.11	1.12
		1500	0.00	0.16	0.01	0.29	0.28	0.28	0.93	0.90	0.95	1.07	1.04	1.04
		2000	0.01	0.16	0.01	0.26	0.26	0.26	0.95	0.92	0.94	1.01	0.97	0.97
0.10	-0.95	500	0.07	0.27	0.01	0.31	0.31	0.30	0.94	0.85	0.96	1.27	1.22	1.25
		1000	0.02	0.19	0.00	0.28	0.27	0.28	0.94	0.90	0.92	1.10	1.05	1.06
		1500	0.01	0.15	0.01	0.26	0.26	0.25	0.95	0.91	0.95	1.01	0.97	0.98
		2000	0.01	0.12	0.01	0.24	0.24	0.24	0.95	0.92	0.94	0.96	0.92	0.92
0.25	-0.88	500	0.06	0.26	0.01	0.33	0.32	0.32	0.95	0.87	0.93	1.29	1.23	1.27
		1000	0.02	0.18	0.00	0.28	0.28	0.27	0.95	0.90	0.95	1.11	1.07	1.07
		1500	0.01	0.14	0.00	0.24	0.24	0.24	0.97	0.92	0.96	1.03	0.98	0.99
		2000	0.02	0.14	0.02	0.25	0.25	0.25	0.94	0.91	0.94	0.97	0.93	0.93
1.00	-0.61	500	0.00	0.15	0.05	0.42	0.41	0.41	0.93	0.92	0.92	1.59	1.49	1.56
		1000	0.01	0.13	0.00	0.32	0.31	0.32	0.97	0.95	0.96	1.38	1.31	1.33
		1500	0.01	0.09	0.01	0.30	0.30	0.30	0.97	0.95	0.96	1.27	1.21	1.22
		2000	0.01	0.08	0.00	0.29	0.29	0.29	0.95	0.93	0.94	1.19	1.14	1.15

Table A6: Comparison of DLL, plug-in (Plug), oracle (Orac) estimators in Setting 4 with df = 15, across different sample sizes n and evaluation points a_0 with $p = 1500$. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" and "SE" report the absolute bias, and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

Non-linear Treatment Model, exactly sparse: $f(d) = 1.5\sin(d)$

			Bias				Coverage				Length			
a_0	True	n	DLL	DLL-S	Plug	Orac	DLL	DLL-S	Plug	Orac	DLL	DLL-S	Plug	Orac
-1.25	0.47	500	0.29	0.12	0.33	0.01	0.94	0.95	0.90	0.90	4.36	4.27	3.41	3.60
		1000	0.21	0.13	0.22	0.06	0.94	0.96	0.91	0.93	3.55	3.61	3.31	3.51
		1500	0.06	0.11	0.07	0.15	0.94	0.96	0.91	0.92	3.24	3.24	3.10	3.25
		2000	0.18	0.12	0.19	0.01	0.96	0.95	0.93	0.94	2.98	3.02	2.84	2.95
-0.50	1.32	500	0.32	0.26	0.36	0.01	0.90	0.93	0.87	0.95	2.04	2.08	1.97	2.20
		1000	0.24	0.15	0.27	0.00	0.93	0.94	0.90	0.94	1.78	1.80	1.71	1.81
		1500	0.18	0.10	0.21	0.02	0.94	0.94	0.91	0.96	1.61	1.65	1.55	1.61
		2000	0.14	0.08	0.17	0.01	0.92	0.95	0.90	0.95	1.50	1.52	1.45	1.51
0.10	1.49	500	0.31	0.23	0.37	0.01	0.84	0.91	0.78	0.95	1.40	1.44	1.34	1.50
		1000	0.20	0.12	0.25	0.00	0.87	0.93	0.83	0.95	1.24	1.24	1.18	1.25
		1500	0.18	0.10	0.23	0.03	0.92	0.95	0.88	0.96	1.12	1.14	1.07	1.12
		2000	0.13	0.08	0.17	0.00	0.91	0.94	0.87	0.94	1.05	1.05	1.01	1.05
0.25	1.45	500	0.33	0.23	0.39	0.06	0.80	0.87	0.74	0.94	1.34	1.36	1.28	1.42
		1000	0.19	0.12	0.24	0.01	0.90	0.92	0.85	0.95	1.17	1.18	1.12	1.19
		1500	0.16	0.10	0.21	0.02	0.92	0.95	0.87	0.95	1.06	1.07	1.02	1.06
		2000	0.13	0.08	0.18	0.00	0.91	0.95	0.87	0.95	1.00	0.99	0.95	0.99
1.00	0.81	500	0.20	0.19	0.26	0.02	0.90	0.92	0.86	0.95	1.31	1.32	1.23	1.38
		1000	0.16	0.13	0.21	0.01	0.92	0.93	0.87	0.94	1.15	1.14	1.10	1.16
		1500	0.13	0.11	0.17	0.01	0.93	0.93	0.86	0.96	1.04	1.05	0.99	1.04
		2000	0.11	0.09	0.15	0.01	0.93	0.93	0.88	0.94	0.98	0.97	0.92	0.97

Non-linear Treatment Model, exactly sparse: $f(d) = 2\exp(-d/2)$

			Bias				Coverage				Length			
a_0	True	n	DLL	DLL-S	Plug	Orac	DLL	DLL-S	Plug	Orac	DLL	DLL-S	Plug	Orac
-1.25	-1.87	500	0.11	0.12	0.11	0.09	0.94	0.94	0.91	0.91	2.88	2.89	2.68	2.81
		1000	0.03	0.09	0.02	0.04	0.95	0.94	0.93	0.92	2.39	2.42	2.31	2.35
		1500	0.01	0.03	0.00	0.02	0.97	0.95	0.95	0.96	2.17	2.22	2.11	2.13
		2000	0.06	0.02	0.04	0.05	0.94	0.96	0.92	0.93	2.05	2.08	2.01	2.04
-0.50	-1.28	500	0.03	0.03	0.07	0.01	0.96	0.95	0.94	0.94	1.43	1.46	1.38	1.42
		1000	0.01	0.00	0.04	0.04	0.96	0.94	0.95	0.95	1.22	1.25	1.19	1.21
		1500	0.02	0.01	0.05	0.02	0.95	0.94	0.94	0.95	1.11	1.14	1.08	1.09
		2000	0.01	0.03	0.04	0.03	0.93	0.94	0.92	0.93	1.04	1.07	1.02	1.03
0.10	-0.95	500	0.10	0.08	0.14	0.01	0.94	0.92	0.92	0.94	1.00	1.02	0.96	1.00
		1000	0.08	0.06	0.12	0.00	0.92	0.95	0.90	0.94	0.85	0.87	0.83	0.84
		1500	0.05	0.05	0.10	0.00	0.95	0.95	0.93	0.94	0.78	0.79	0.76	0.77
		2000	0.05	0.03	0.10	0.01	0.95	0.94	0.92	0.95	0.73	0.74	0.71	0.72
0.25	-0.88	500	0.11	0.10	0.15	0.00	0.92	0.92	0.88	0.95	0.94	0.96	0.91	0.94
		1000	0.09	0.07	0.13	0.02	0.93	0.92	0.90	0.94	0.81	0.82	0.79	0.80
		1500	0.04	0.06	0.09	0.01	0.94	0.93	0.91	0.93	0.74	0.75	0.72	0.73
		2000	0.04	0.05	0.09	0.01	0.95	0.93	0.93	0.95	0.69	0.70	0.68	0.68
1.00	-0.61	500	0.11	0.16	0.16	0.01	0.92	0.88	0.86	0.95	0.93	0.93	0.90	0.93
		1000	0.07	0.12	0.12	0.01	0.93	0.90	0.90	0.95	0.80	0.79	0.77	0.78
		1500	0.08	0.07	0.13	0.01	0.92	0.93	0.87	0.95	0.73	0.73	0.71	0.71
		2000	0.06	0.07	0.11	0.00	0.92	0.93	0.87	0.94	0.68	0.68	0.67	0.67

Table A7: Comparison of DLL, DLL-S, plug-in (Plug), oracle (Orac) estimators for the non-linear treatment model, across different sample sizes n and evaluation points a_0 with $p = 1500$. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" report the absolute bias; the columns indexed with "Coverage" report the empirical coverage and the columns indexed with "Length" report the average CI length.

F.2 Other Bandwidth Selection Methods

We also investigate the performances of `DLL`, plug-in (`Plug`), and oracle (`Orac`) estimators using other bandwidth selection methods: the methods `regCVBwSelC()` implemented in Cabrera (2018) and `npregbw()` implemented in Hayfield and Racine (2008). We generate X_i, D_i as in Setting 2, and generate the outcome model as the exactly sparse model. The results are summarised in Table A8 and Table A9. We observe that using these two bandwidth selection methods might lead to a bad coverage for `DLL`, or a wide confidence interval. For the undercoverage settings for `DLL`, the oracle CI (the benchmark) does not attain the desired coverage level. This indicates that these bandwidth selections are not stable for our simulation studies. Hence, we select the bandwidth by the function `thumbBw()` in `locpol` as mentioned in the main paper.

Setting 2, exactly sparse: $f(d) = 2\exp(-d/2)$ with `regCVBwSelC()` in `locpol`

			Bias			SE			Coverage			Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	-0.41	500	0.04	0.11	0.01	0.24	0.15	0.14	0.95	0.82	0.92	0.90	0.52	0.49
		1000	0.01	0.10	0.01	0.20	0.16	0.11	0.96	0.78	0.94	0.74	0.41	0.36
		1500	0.02	0.10	0.01	0.18	0.11	0.09	0.95	0.74	0.94	0.67	0.34	0.31
		2000	0.01	0.09	0.01	0.17	0.09	0.08	0.96	0.71	0.96	0.62	0.31	0.28
-0.50	-0.52	500	0.01	0.05	0.06	0.19	0.14	0.11	0.94	0.84	0.85	0.71	0.42	0.39
		1000	0.01	0.03	0.06	0.16	0.11	0.09	0.95	0.81	0.79	0.59	0.33	0.29
		1500	0.02	0.04	0.06	0.15	0.10	0.07	0.94	0.78	0.74	0.53	0.27	0.25
		2000	0.01	0.03	0.05	0.13	0.08	0.07	0.96	0.81	0.73	0.49	0.24	0.22
0.10	-0.40	500	0.00	0.10	0.01	0.20	0.13	0.11	0.93	0.80	0.95	0.72	0.43	0.40
		1000	0.01	0.08	0.00	0.17	0.10	0.08	0.95	0.79	0.94	0.59	0.33	0.29
		1500	0.01	0.08	0.01	0.16	0.09	0.07	0.93	0.73	0.94	0.54	0.28	0.25
		2000	0.01	0.07	0.00	0.14	0.07	0.06	0.95	0.76	0.96	0.49	0.25	0.22
0.25	-0.35	500	0.01	0.11	0.01	0.19	0.13	0.12	0.94	0.82	0.94	0.74	0.44	0.41
		1000	0.03	0.09	0.02	0.16	0.10	0.09	0.95	0.80	0.94	0.61	0.34	0.30
		1500	0.02	0.08	0.01	0.15	0.10	0.07	0.95	0.73	0.95	0.55	0.29	0.26
		2000	0.02	0.08	0.01	0.14	0.08	0.07	0.95	0.76	0.94	0.51	0.25	0.23
1.00	-0.15	500	0.04	0.12	0.06	0.26	0.18	0.16	0.96	0.87	0.87	1.04	0.61	0.56
		1000	0.03	0.09	0.07	0.22	0.15	0.13	0.96	0.85	0.81	0.86	0.48	0.42
		1500	0.01	0.09	0.06	0.20	0.14	0.12	0.94	0.85	0.80	0.77	0.40	0.36
		2000	0.01	0.08	0.05	0.20	0.11	0.10	0.94	0.84	0.82	0.71	0.35	0.32

Setting 2, exactly sparse: $f(d) = 1.5\sin(d)$ with `regCVBwSelC()` in `locpol`

			Bias			SE			Coverage			Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	-1.00	500	0.00	0.06	0.11	0.27	0.24	0.22	0.94	0.85	0.76	0.97	0.73	0.73
		1000	0.03	0.04	0.11	0.20	0.19	0.18	0.96	0.86	0.74	0.77	0.58	0.57
		1500	0.02	0.04	0.08	0.18	0.18	0.16	0.95	0.87	0.78	0.70	0.54	0.53
		2000	0.03	0.02	0.09	0.16	0.16	0.16	0.94	0.83	0.72	0.63	0.48	0.47
-0.50	-0.56	500	0.10	0.00	0.20	0.20	0.19	0.19	0.89	0.86	0.55	0.77	0.58	0.58
		1000	0.10	0.02	0.18	0.17	0.16	0.15	0.87	0.80	0.51	0.61	0.46	0.45
		1500	0.09	0.02	0.16	0.16	0.16	0.15	0.87	0.77	0.53	0.55	0.43	0.42
		2000	0.09	0.03	0.16	0.13	0.14	0.14	0.85	0.74	0.54	0.50	0.38	0.37
0.10	0.74	500	0.27	0.47	0.26	0.21	0.22	0.21	0.63	0.23	0.49	0.78	0.59	0.59
		1000	0.20	0.41	0.23	0.17	0.17	0.17	0.68	0.21	0.44	0.62	0.47	0.45
		1500	0.16	0.35	0.20	0.17	0.18	0.18	0.72	0.24	0.46	0.56	0.43	0.43
		2000	0.16	0.33	0.20	0.14	0.16	0.16	0.68	0.25	0.47	0.50	0.38	0.38
0.25	0.94	500	0.30	0.51	0.29	0.23	0.23	0.23	0.59	0.24	0.44	0.80	0.61	0.61
		1000	0.22	0.44	0.27	0.18	0.19	0.20	0.65	0.20	0.42	0.64	0.48	0.47
		1500	0.18	0.37	0.22	0.19	0.20	0.19	0.64	0.24	0.44	0.58	0.45	0.44
		2000	0.16	0.36	0.23	0.14	0.18	0.18	0.67	0.27	0.46	0.52	0.40	0.39
1.00	0.81	500	0.06	0.19	0.01	0.31	0.27	0.27	0.94	0.80	0.92	1.11	0.85	0.84
		1000	0.00	0.12	0.03	0.25	0.20	0.19	0.96	0.78	0.93	0.89	0.68	0.66
		1500	0.01	0.11	0.03	0.21	0.19	0.18	0.96	0.82	0.94	0.81	0.63	0.62
		2000	0.01	0.12	0.00	0.19	0.16	0.15	0.95	0.80	0.93	0.72	0.55	0.54

Table A8: Comparison of `DLL`, plug-in (`Plug`), and oracle (`Orac`) estimators using `regCVBwSelC()` for bandwidth selection, across different sample sizes n and evaluation points a_0 with $p = 1500$. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" and "SE" report the absolute bias and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

Setting 2, exactly sparse: $f(d) = 2\exp(-d/2)$ with `npregbw()` in `np`

			Bias			SE			Coverage			Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	-0.41	500	0.02	0.12	0.01	0.88	0.89	0.71	0.94	0.85	0.90	1.89	1.76	1.43
		1000	0.01	0.08	0.02	0.56	0.56	0.41	0.94	0.87	0.92	1.44	1.36	1.12
		1500	0.01	0.08	0.02	0.46	0.46	0.43	0.96	0.84	0.94	1.32	1.27	1.00
		2000	0.02	0.12	0.00	0.47	0.47	0.45	0.95	0.88	0.93	1.10	1.06	0.91
-0.50	-0.52	500	0.02	0.02	0.02	0.63	0.62	0.62	0.88	0.83	0.74	1.48	1.41	1.16
		1000	0.03	0.02	0.03	0.49	0.50	0.41	0.91	0.86	0.81	1.14	1.08	0.89
		1500	0.01	0.02	0.03	0.49	0.49	0.32	0.88	0.86	0.76	1.05	1.02	0.78
		2000	0.01	0.00	0.04	0.36	0.36	0.33	0.88	0.86	0.82	0.86	0.84	0.71
0.10	-0.40	500	0.02	0.11	0.00	0.66	0.66	0.50	0.92	0.84	0.91	1.51	1.45	1.15
		1000	0.00	0.10	0.00	0.46	0.45	0.39	0.94	0.83	0.93	1.16	1.10	0.90
		1500	0.01	0.11	0.00	0.59	0.63	0.34	0.95	0.84	0.94	1.07	1.03	0.80
		2000	0.03	0.06	0.00	0.36	0.36	0.30	0.95	0.87	0.95	0.87	0.84	0.71
0.25	-0.35	500	0.05	0.09	0.02	0.59	0.60	0.52	0.92	0.85	0.91	1.55	1.47	1.20
		1000	0.05	0.18	0.03	0.63	0.62	0.50	0.94	0.87	0.94	1.19	1.14	0.94
		1500	0.05	0.17	0.00	0.47	0.46	0.32	0.94	0.84	0.95	1.10	1.06	0.81
		2000	0.01	0.12	0.01	0.33	0.33	0.34	0.93	0.86	0.95	0.90	0.88	0.73
1.00	-0.15	500	0.05	0.11	0.08	1.13	1.02	0.89	0.84	0.84	0.71	2.21	2.07	1.68
		1000	0.02	0.14	0.03	0.81	0.79	0.54	0.85	0.86	0.77	1.67	1.58	1.28
		1500	0.02	0.16	0.03	0.73	0.68	0.54	0.92	0.86	0.81	1.54	1.49	1.14
		2000	0.03	0.11	0.02	0.50	0.48	0.44	0.90	0.88	0.83	1.26	1.23	1.05

Setting 2, exactly sparse: $f(d) = 1.5\sin(d)$ with `npregbw()` in `np`

			Bias			SE			Coverage			Length		
a_0	True	n	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`	`DLL`	`Plug`	`Orac`
-1.25	-1.00	500	0.05	0.18	0.03	0.79	0.74	0.76	0.93	0.90	0.93	2.12	1.93	2.15
		1000	0.02	0.14	0.02	0.57	0.56	0.65	0.95	0.93	0.95	1.67	1.58	1.76
		1500	0.03	0.08	0.07	0.73	0.73	0.61	0.94	0.94	0.95	1.58	1.50	1.62
		2000	0.00	0.10	0.01	0.51	0.50	0.52	0.95	0.92	0.94	1.36	1.28	1.41
-0.50	-0.56	500	0.05	0.09	0.11	0.57	0.56	0.62	0.93	0.93	0.90	1.65	1.56	1.75
		1000	0.07	0.07	0.07	0.50	0.50	0.50	0.93	0.92	0.91	1.33	1.26	1.42
		1500	0.07	0.05	0.07	0.49	0.47	0.47	0.94	0.94	0.92	1.24	1.19	1.28
		2000	0.05	0.06	0.07	0.43	0.43	0.41	0.93	0.90	0.91	1.07	1.02	1.12
0.10	0.74	500	0.22	0.36	0.11	0.60	0.60	0.66	0.81	0.66	0.86	1.66	1.56	1.77
		1000	0.13	0.26	0.08	0.54	0.53	0.67	0.87	0.70	0.90	1.34	1.28	1.43
		1500	0.13	0.25	0.10	0.45	0.48	0.43	0.89	0.72	0.93	1.26	1.22	1.31
		2000	0.08	0.18	0.04	0.32	0.31	0.38	0.92	0.78	0.93	1.09	1.04	1.14
0.25	0.94	500	0.20	0.35	0.16	0.66	0.65	0.68	0.81	0.66	0.84	1.74	1.64	1.83
		1000	0.14	0.27	0.09	0.39	0.38	0.51	0.86	0.71	0.89	1.40	1.33	1.48
		1500	0.14	0.25	0.08	0.41	0.40	0.51	0.88	0.72	0.89	1.30	1.25	1.34
		2000	0.07	0.18	0.05	0.34	0.33	0.38	0.91	0.77	0.88	1.13	1.09	1.18
1.00	0.81	500	0.02	0.15	0.07	0.84	0.74	0.96	0.96	0.91	0.94	2.44	2.22	2.48
		1000	0.07	0.19	0.02	0.74	0.72	0.87	0.96	0.93	0.94	1.98	1.86	2.07
		1500	0.06	0.06	0.05	0.94	0.87	0.69	0.92	0.90	0.92	1.82	1.71	1.85
		2000	0.01	0.11	0.01	0.51	0.52	0.52	0.96	0.93	0.95	1.57	1.49	1.63

Table A9: Comparison of `DLL`, plug-in (`Plug`), and oracle (`Orac`) estimators using `npregbw()` for bandwidth selection, across different sample sizes n and evaluation points a_0 with $p = 1500$. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" and "SE" report the absolute bias and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

F.3 Data Swap and Quantile Transformation

In Table A10 and Table A11, we compare the DLL estimator without data swapping and the DLL with data swapping (Swap). The data is generated as in Setting 2, with the sparse additive model being exactly sparse. The CIs with and without data swapping attain the desired coverage level. When the sample size is relatively large, they have similar performance; for relatively small sample size, the confidence interval without data swapping can be shorter than that with data swapping. This happens because no data swapping uses the entire data to construct initial estimators of g and γ. When the sample size is relatively small (e.g., $n = 500$ and $p = 1500$), the DLL with data swapping might be slightly noisier than the one without data swapping.

We now investigate the performance of our proposed method with quantile transformation (Trans), which is detailed in Section A.3. We report the comparison with the Trans estimator in Tables A10 and A11. The data is generated as in Setting 2, with the sparse additive model being exactly sparse. As reported in Table A11, the method using quantile transformation leads to slightly better performance for $f(d) = 2\exp(-d/2)$: the bias is slightly smaller, and the CI is shorter. Nevertheless, the regular DLL method still attains the desired coverage.

Setting 2, exactly sparse: $f(d) = 1.5\sin(d)$

			Bias			SE			Coverage			Length		
a_0	True	n	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`
-1.25	-0.41	500	0.00	0.09	0.04	0.49	0.55	0.55	0.95	0.94	0.93	2.01	2.13	2.02
		1000	0.02	0.02	0.04	0.41	0.43	0.42	0.96	0.94	0.96	1.68	1.70	1.63
		1500	0.02	0.06	0.02	0.38	0.40	0.37	0.94	0.94	0.95	1.51	1.52	1.47
		2000	0.01	0.03	0.00	0.37	0.36	0.33	0.96	0.94	0.96	1.42	1.42	1.37
-0.50	-0.52	500	0.00	0.07	0.00	0.40	0.45	0.40	0.95	0.93	0.94	1.59	1.66	1.58
		1000	0.02	0.05	0.00	0.32	0.34	0.33	0.96	0.96	0.95	1.32	1.34	1.29
		1500	0.01	0.03	0.02	0.29	0.31	0.30	0.97	0.95	0.94	1.19	1.20	1.16
		2000	0.03	0.01	0.03	0.28	0.30	0.26	0.95	0.94	0.96	1.12	1.12	1.08
0.10	-0.40	500	0.17	0.19	0.11	0.42	0.45	0.41	0.92	0.90	0.93	1.61	1.69	1.60
		1000	0.08	0.10	0.08	0.32	0.35	0.32	0.95	0.94	0.94	1.34	1.36	1.31
		1500	0.07	0.10	0.05	0.30	0.32	0.28	0.95	0.93	0.96	1.21	1.22	1.18
		2000	0.07	0.07	0.06	0.29	0.29	0.27	0.96	0.93	0.96	1.13	1.13	1.09
0.25	-0.35	500	0.17	0.18	0.14	0.43	0.47	0.42	0.92	0.92	0.95	1.67	1.75	1.66
		1000	0.08	0.12	0.09	0.35	0.37	0.34	0.93	0.93	0.94	1.38	1.40	1.35
		1500	0.08	0.11	0.07	0.30	0.34	0.30	0.96	0.93	0.95	1.25	1.26	1.22
		2000	0.07	0.06	0.08	0.29	0.32	0.29	0.95	0.91	0.95	1.17	1.17	1.13
1.00	-0.15	500	0.08	0.10	0.03	0.61	0.64	0.59	0.93	0.93	0.95	2.33	2.45	2.33
		1000	0.00	0.09	0.02	0.48	0.50	0.48	0.96	0.94	0.94	1.94	1.97	1.89
		1500	0.03	0.03	0.02	0.44	0.45	0.45	0.95	0.95	0.94	1.75	1.76	1.70
		2000	0.01	0.00	0.00	0.40	0.44	0.40	0.95	0.94	0.97	1.63	1.64	1.58

Setting 2, approximately sparse: $f(d) = 1.5\sin(d)$

			Bias			SE			Coverage			Length		
a_0	True	n	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`
-1.25	-0.41	500	0.02	0.07	0.00	0.47	0.61	0.48	0.96	0.94	0.96	1.89	2.23	1.88
		1000	0.04	0.02	0.02	0.39	0.44	0.44	0.96	0.96	0.92	1.62	1.77	1.61
		1500	0.02	0.02	0.03	0.35	0.41	0.39	0.98	0.96	0.94	1.47	1.57	1.44
		2000	0.03	0.03	0.03	0.34	0.37	0.35	0.95	0.95	0.95	1.37	1.45	1.35
-0.50	-0.52	500	0.03	0.01	0.02	0.38	0.44	0.38	0.95	0.97	0.94	1.48	1.75	1.47
		1000	0.03	0.00	0.02	0.33	0.38	0.35	0.94	0.95	0.94	1.28	1.41	1.26
		1500	0.03	0.02	0.02	0.28	0.32	0.29	0.96	0.95	0.95	1.16	1.24	1.13
		2000	0.02	0.01	0.01	0.28	0.29	0.26	0.96	0.95	0.96	1.08	1.15	1.07
0.10	-0.40	500	0.19	0.18	0.16	0.38	0.47	0.41	0.92	0.93	0.91	1.51	1.78	1.49
		1000	0.15	0.09	0.11	0.32	0.36	0.32	0.93	0.94	0.93	1.30	1.42	1.28
		1500	0.11	0.08	0.09	0.31	0.34	0.31	0.94	0.92	0.93	1.17	1.25	1.15
		2000	0.10	0.09	0.06	0.27	0.29	0.28	0.95	0.95	0.94	1.10	1.17	1.08
0.25	-0.35	500	0.21	0.18	0.22	0.40	0.44	0.39	0.92	0.96	0.92	1.55	1.84	1.53
		1000	0.18	0.11	0.14	0.32	0.39	0.33	0.92	0.94	0.93	1.34	1.47	1.32
		1500	0.12	0.08	0.09	0.32	0.33	0.30	0.92	0.94	0.94	1.22	1.30	1.18
		2000	0.12	0.08	0.08	0.28	0.31	0.30	0.95	0.96	0.93	1.13	1.21	1.12
1.00	-0.15	500	0.12	0.01	0.07	0.58	0.65	0.57	0.92	0.95	0.94	2.19	2.58	2.17
		1000	0.04	0.08	0.01	0.46	0.54	0.48	0.96	0.93	0.94	1.88	2.06	1.85
		1500	0.01	0.01	0.02	0.42	0.48	0.43	0.95	0.95	0.95	1.70	1.82	1.66
		2000	0.05	0.02	0.03	0.40	0.46	0.41	0.94	0.93	0.94	1.58	1.68	1.57

Table A10: Comparison of `DLL`, `DLL` with data swapping (`Swap`), `DLL` with quantile transformation (`Trans`) in Setting 2 when $f(d) = 1.5\sin(d)$, across different sample sizes n and evaluation points a_0. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" and "SE" report the absolute bias and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

Setting 2, exactly sparse: $f(d) = 2\exp(-d/2)$

			Bias			SE			Coverage			Length		
a_0	True	n	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`
-1.25	-0.41	500	0.03	0.12	0.04	0.30	0.35	0.29	0.96	0.94	0.94	1.19	1.33	1.07
		1000	0.05	0.06	0.02	0.24	0.26	0.20	0.94	0.95	0.97	0.92	0.96	0.82
		1500	0.02	0.04	0.01	0.21	0.21	0.19	0.94	0.94	0.95	0.82	0.83	0.71
		2000	0.02	0.02	0.00	0.20	0.19	0.16	0.95	0.97	0.96	0.74	0.76	0.64
-0.50	-0.52	500	0.00	0.05	0.01	0.24	0.27	0.23	0.95	0.96	0.94	0.93	1.04	0.85
		1000	0.01	0.00	0.00	0.19	0.19	0.17	0.93	0.96	0.94	0.73	0.75	0.64
		1500	0.02	0.00	0.00	0.16	0.16	0.15	0.95	0.96	0.93	0.65	0.66	0.56
		2000	0.02	0.01	0.01	0.15	0.16	0.13	0.95	0.95	0.94	0.59	0.60	0.51
0.10	-0.40	500	0.01	0.03	0.02	0.25	0.27	0.22	0.94	0.95	0.96	0.94	1.06	0.86
		1000	0.01	0.01	0.02	0.19	0.20	0.18	0.95	0.95	0.96	0.74	0.77	0.65
		1500	0.01	0.01	0.00	0.16	0.17	0.15	0.96	0.97	0.95	0.66	0.67	0.57
		2000	0.02	0.01	0.01	0.15	0.16	0.14	0.96	0.96	0.94	0.60	0.61	0.52
0.25	-0.35	500	0.01	0.02	0.01	0.26	0.29	0.22	0.95	0.94	0.96	0.97	1.10	0.89
		1000	0.01	0.03	0.03	0.20	0.20	0.17	0.95	0.95	0.96	0.76	0.79	0.68
		1500	0.02	0.02	0.00	0.17	0.17	0.16	0.95	0.96	0.95	0.68	0.69	0.59
		2000	0.02	0.02	0.01	0.16	0.17	0.13	0.95	0.95	0.96	0.62	0.63	0.53
1.00	-0.15	500	0.01	0.02	0.02	0.37	0.44	0.34	0.94	0.93	0.92	1.37	1.54	1.23
		1000	0.04	0.00	0.01	0.28	0.28	0.25	0.95	0.95	0.91	1.06	1.11	0.94
		1500	0.02	0.02	0.04	0.25	0.26	0.22	0.96	0.94	0.94	0.94	0.96	0.82
		2000	0.01	0.01	0.02	0.23	0.23	0.20	0.93	0.96	0.93	0.86	0.88	0.74

Setting 2, approximately sparse: $f(d) = 2\exp(-d/2)$

			Bias			SE			Coverage			Length		
a_0	True	n	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`	`DLL`	`Swap`	`Trans`
-1.25	-0.41	500	0.01	0.05	0.01	0.29	0.36	0.28	0.95	0.94	0.96	1.11	1.34	1.03
		1000	0.01	0.03	0.00	0.22	0.25	0.21	0.96	0.95	0.94	0.89	1.00	0.79
		1500	0.02	0.03	0.01	0.21	0.23	0.18	0.95	0.95	0.94	0.80	0.87	0.69
		2000	0.01	0.04	0.01	0.19	0.20	0.17	0.94	0.95	0.94	0.73	0.79	0.62
-0.50	-0.52	500	0.05	0.01	0.04	0.23	0.27	0.22	0.92	0.94	0.93	0.87	1.05	0.81
		1000	0.03	0.00	0.02	0.19	0.20	0.17	0.93	0.95	0.92	0.70	0.79	0.62
		1500	0.02	0.01	0.02	0.17	0.18	0.14	0.92	0.94	0.95	0.63	0.68	0.54
		2000	0.01	0.00	0.02	0.15	0.16	0.13	0.95	0.96	0.92	0.57	0.62	0.49
0.10	-0.40	500	0.01	0.02	0.05	0.23	0.27	0.22	0.95	0.96	0.93	0.89	1.07	0.82
		1000	0.00	0.01	0.02	0.19	0.23	0.18	0.93	0.93	0.92	0.71	0.79	0.63
		1500	0.01	0.00	0.01	0.16	0.18	0.14	0.95	0.95	0.95	0.64	0.70	0.55
		2000	0.01	0.00	0.02	0.15	0.16	0.12	0.95	0.97	0.95	0.58	0.63	0.50
0.25	-0.35	500	0.00	0.03	0.05	0.23	0.28	0.23	0.96	0.96	0.94	0.91	1.11	0.84
		1000	0.00	0.01	0.00	0.20	0.23	0.16	0.92	0.94	0.95	0.74	0.82	0.65
		1500	0.01	0.01	0.00	0.16	0.18	0.15	0.96	0.95	0.94	0.66	0.72	0.57
		2000	0.02	0.01	0.00	0.15	0.16	0.13	0.96	0.96	0.95	0.60	0.65	0.52
1.00	-0.15	500	0.01	0.07	0.02	0.35	0.40	0.34	0.94	0.95	0.94	1.28	1.55	1.19
		1000	0.02	0.01	0.00	0.29	0.32	0.23	0.94	0.94	0.95	1.03	1.15	0.90
		1500	0.01	0.01	0.01	0.23	0.25	0.20	0.93	0.96	0.95	0.92	1.00	0.79
		2000	0.00	0.01	0.02	0.22	0.23	0.18	0.95	0.95	0.96	0.84	0.90	0.72

Table A11: Comparison of `DLL`, `DLL` with data swapping (`Swap`), `DLL` with quantile transformation (`Trans`) in Setting 2 when $f(d) = 2\exp(-d/2)$, across different sample sizes n and evaluation points a_0. The column indexed with "True" represents the true value of $f'(a_0)$. The columns indexed with "Bias" and "SE" report the absolute bias and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

F.4 Comparison with ReSmoothing Method

In the following, we provide additional details compared to the method proposed in Gregory et al. (2021). We particularly consider the following two confidence intervals for $f'(a_0)$.

(a) `RS` confidence interval: we apply the local linear estimator to $\{D_i, \widehat{f}^{\rm pre}(D_i)\}_{1\leq i\leq n}$ with the presmoothing estimators $\{\widehat{f}^{\rm pre}(D_i)\}_{1\leq i\leq n}$ as the outcome variables (Gregory et al., 2021); we construct the CI by the output of the package `nprobust` (Calonico et al., 2019). To construct the presmoothing estimators $\{\widehat{f}^{\rm pre}(D_i)\}_{1\leq i\leq n}$, we use the authors' original code available at `https://github.com/gregorkb/spaddinf` using their default choices of the tuning parameters;

(b) `OraRS` confidence interval: we estimate the standard error of the `RS` estimator by the sample standard deviation of 500 `RS` estimates and then construct the confidence interval by assuming the asymptotic normality of the `RS` estimator.

For comparison with the `RS` method, we generate the data as the simulation setting in Gregory et al. (2021) and the full simulation results are reported in A12 (bias) and A13 (Coverage). Our `DLL` method achieves the desired coverage across all sample sizes, dimensions, and correlation parameters. For a small sample size, the `RS` estimator suffers from a large bias while our proposed `DLL` effectively corrects the bias in these settings. `OraRS` does not attain the expected coverage when the sample size is small. For a large sample size, we notice that `OraRS` confidence interval achieves the desired coverage while `RS` does not, which suggests that the uncertainty quantification of `RS` is subtle and requires further investigation.

In addition, we generate the data following Setting 1 of the current paper with $p = 750$, $n \in \{500, 750, 1000\}$. We present the results with $f(d) = 1.5\sin(d), g_1(x) = 2\exp(-x/2)$ or $f(d) = 2\exp(-d/2), g_1(x) = 1.5\sin(x)$ in Table A14. Our method `DLL` has a much smaller bias than the `RS` estimator and achieves desired coverage in most settings. Additionally, the CI length for `OraRS` is large and the length of our `DLL` method is similar to the oracle confidence interval.

Bias in the Setting of Gregory et al. (2021): a_0 = -1

			$f'(a_0)$		$g_1'(a_0)$		$g_2'(a_0)$		$g_3'(a_0)$		$g_4'(a_0)$	
n	p	r	DLL	RS	DLL	RS	DLL	RS	DLL	RS	DLL	RS
100	50	0.0	0.18	0.66	0.00	0.90	0.05	0.53	0.14	0.71	0.06	0.04
		0.1	0.20	0.64	0.13	0.94	0.03	0.57	0.14	0.80	0.00	0.04
		0.3	0.22	0.66	0.40	0.95	0.18	0.71	0.15	0.90	0.01	0.03
		0.5	0.14	0.61	0.13	0.91	0.05	0.94	0.13	1.06	0.01	0.11
	150	0.0	0.34	0.70	0.07	1.03	0.04	0.61	0.16	0.84	0.08	0.00
		0.1	0.22	0.78	0.20	1.13	0.01	0.65	0.23	0.94	0.01	0.00
		0.3	0.29	0.69	0.06	1.02	0.27	0.84	0.13	1.02	0.01	0.07
		0.5	0.30	0.61	0.08	0.96	0.22	0.91	0.01	1.24	0.08	0.10
1000	50	0.0	0.01	0.04	0.01	0.01	0.01	0.01	0.02	0.04	0.01	0.01
		0.1	0.02	0.03	0.03	0.08	0.05	0.06	0.01	0.05	0.02	0.03
		0.3	0.04	0.05	0.04	0.14	0.03	0.06	0.02	0.06	0.03	0.02
		0.5	0.05	0.08	0.02	0.09	0.01	0.16	0.09	0.12	0.03	0.04
	150	0.0	0.07	0.05	0.00	0.05	0.03	0.04	0.10	0.01	0.00	0.00
		0.1	0.05	0.03	0.05	0.09	0.02	0.02	0.11	0.10	0.02	0.04
		0.3	0.01	0.03	0.03	0.05	0.04	0.10	0.01	0.11	0.03	0.02
		0.5	0.02	0.01	0.01	0.16	0.03	0.18	0.03	0.18	0.00	0.04

Bias in the Setting of Gregory et al. (2021): $a_0 = 0.5$

			$f'(a_0)$		$g_1'(a_0)$		$g_2'(a_0)$		$g_3'(a_0)$		$g_4'(a_0)$	
n	p	r	DLL	RS	DLL	RS	DLL	RS	DLL	RS	DLL	RS
100	50	0.0	0.12	0.82	0.04	0.40	0.08	0.55	0.12	0.19	0.04	0.00
		0.1	0.25	0.85	0.05	0.41	0.07	0.57	0.09	0.17	0.02	0.01
		0.3	0.27	0.83	0.04	0.28	0.09	0.70	0.11	0.28	0.01	0.01
		0.5	0.23	0.82	0.00	0.25	0.12	0.81	0.22	0.51	0.04	0.08
	150	0.0	0.16	0.93	0.09	0.52	0.11	0.63	0.17	0.20	0.08	0.00
		0.1	0.29	0.94	0.02	0.46	0.10	0.72	0.17	0.20	0.08	0.01
		0.3	0.29	0.92	0.13	0.38	0.19	0.82	0.27	0.40	0.06	0.02
		0.5	0.33	0.87	0.08	0.28	0.16	0.84	0.31	0.49	0.03	0.07
1000	50	0.0	0.01	0.01	0.00	0.05	0.01	0.03	0.02	0.02	0.05	0.03
		0.1	0.06	0.04	0.01	0.03	0.01	0.03	0.04	0.03	0.02	0.01
		0.3	0.01	0.02	0.00	0.04	0.05	0.09	0.08	0.08	0.04	0.02
		0.5	0.10	0.05	0.04	0.01	0.04	0.12	0.04	0.10	0.02	0.02
	150	0.0	0.00	0.00	0.01	0.01	0.02	0.02	0.02	0.02	0.00	0.02
		0.1	0.05	0.06	0.01	0.01	0.02	0.03	0.08	0.08	0.05	0.01
		0.3	0.07	0.07	0.05	0.03	0.03	0.01	0.09	0.13	0.02	0.01
		0.5	0.04	0.01	0.04	0.02	0.00	0.11	0.06	0.16	0.04	0.06

Table A12: Comparison of bias of DLL, ReSmoothing (RS) in the setting of Gregory et al. (2021), across different sample sizes n, dimension of covariates p, and the correlation parameter r. The f, g_1, g_2, g_3 and g_4 represent the functions of interest to estimate their derivatives at a_0. The entries of the table represents the bias across 500 simulations.

Coverage in the Setting of Gregory et al. (2021): a_0 = -1

			$f'(a_0)$			$g_1'(a_0)$			$g_2'(a_0)$			$g_3'(a_0)$			$g_4'(a_0)$		
n	p	r	DLL	RS	OraRS	DLL	RS	OraRS	DLL	RS	OraRS	DLL	RS	OraRS	DLL	RS	OraRS
100	50	0.0	0.94	0.09	0.84	0.93	0.16	0.83	0.95	0.17	0.90	0.96	0.21	0.89	0.96	0.25	0.95
		0.1	0.96	0.09	0.79	0.93	0.13	0.83	0.96	0.15	0.88	0.94	0.19	0.90	0.93	0.22	0.94
		0.3	0.94	0.07	0.85	0.95	0.11	0.89	0.95	0.09	0.87	0.95	0.17	0.87	0.94	0.21	0.92
		0.5	0.94	0.09	0.90	0.94	0.11	0.88	0.93	0.04	0.76	0.94	0.14	0.83	0.94	0.24	0.93
	150	0.0	0.94	0.06	0.43	0.94	0.11	0.82	0.95	0.12	0.91	0.94	0.17	0.87	0.95	0.23	0.94
		0.1	0.92	0.03	0.87	0.93	0.09	0.74	0.94	0.11	0.88	0.95	0.15	0.87	0.96	0.22	0.94
		0.3	0.91	0.07	0.92	0.93	0.11	0.84	0.93	0.07	0.82	0.95	0.13	0.85	0.95	0.25	0.93
		0.5	0.93	0.08	0.92	0.92	0.15	0.84	0.92	0.05	0.73	0.94	0.10	0.79	0.93	0.22	0.95
1000	50	0.0	0.96	0.01	0.95	0.95	0.06	0.95	0.97	0.00	0.94	0.96	0.07	0.95	0.94	0.01	0.95
		0.1	0.96	0.01	0.95	0.97	0.05	0.95	0.96	0.00	0.94	0.97	0.06	0.95	0.96	0.00	0.95
		0.3	0.96	0.01	0.95	0.96	0.05	0.94	0.94	0.01	0.95	0.94	0.07	0.94	0.95	0.00	0.95
		0.5	0.96	0.01	0.96	0.95	0.05	0.96	0.95	0.00	0.93	0.95	0.07	0.95	0.94	0.00	0.94
	150	0.0	0.96	0.01	0.94	0.95	0.05	0.95	0.93	0.01	0.95	0.95	0.09	0.95	0.94	0.00	0.94
		0.1	0.94	0.01	0.95	0.97	0.04	0.95	0.95	0.01	0.94	0.96	0.09	0.93	0.94	0.00	0.95
		0.3	0.95	0.02	0.94	0.96	0.04	0.95	0.95	0.01	0.95	0.96	0.06	0.94	0.95	0.00	0.94
		0.5	0.96	0.01	0.95	0.94	0.05	0.95	0.95	0.01	0.95	0.96	0.08	0.95	0.94	0.02	0.94

Coverage in the Setting of Gregory et al. (2021): a_0 = 0.5

			$f'(a_0)$			$g_1'(a_0)$			$g_2'(a_0)$			$g_3'(a_0)$			$g_4'(a_0)$		
n	p	r	DLL	RS	OraRS	DLL	RS	OraRS	DLL	RS	OraRS	DLL	RS	OraRS	DLL	RS	OraRS
100	50	0.0	0.95	0.06	0.44	0.94	0.21	0.92	0.93	0.14	0.88	0.97	0.22	0.94	0.96	0.21	0.95
		0.1	0.95	0.07	0.56	0.95	0.21	0.93	0.96	0.14	0.90	0.94	0.23	0.94	0.95	0.25	0.94
		0.3	0.93	0.07	0.81	0.94	0.22	0.94	0.95	0.11	0.85	0.95	0.23	0.94	0.96	0.18	0.92
		0.5	0.96	0.07	0.87	0.95	0.21	0.95	0.96	0.07	0.75	0.94	0.20	0.92	0.95	0.23	0.94
	150	0.0	0.92	0.02	0.30	0.94	0.21	0.92	0.95	0.13	0.87	0.97	0.25	0.95	0.95	0.23	0.94
		0.1	0.90	0.03	0.23	0.95	0.18	0.92	0.95	0.09	0.83	0.96	0.18	0.95	0.94	0.23	0.94
		0.3	0.90	0.06	0.80	0.95	0.20	0.95	0.93	0.06	0.83	0.93	0.19	0.94	0.95	0.23	0.95
		0.5	0.92	0.06	0.51	0.95	0.21	0.93	0.93	0.06	0.82	0.93	0.20	0.93	0.96	0.19	0.93
1000	50	0.0	0.96	0.03	0.94	0.94	0.03	0.94	0.95	0.00	0.95	0.95	0.03	0.95	0.96	0.01	0.94
		0.1	0.97	0.01	0.95	0.95	0.03	0.95	0.94	0.00	0.94	0.96	0.02	0.95	0.95	0.00	0.94
		0.3	0.96	0.02	0.96	0.96	0.04	0.95	0.96	0.02	0.94	0.98	0.03	0.95	0.94	0.01	0.94
		0.5	0.95	0.02	0.94	0.95	0.03	0.95	0.96	0.00	0.94	0.96	0.03	0.95	0.96	0.01	0.94
	150	0.0	0.97	0.02	0.95	0.97	0.03	0.94	0.96	0.01	0.95	0.97	0.02	0.95	0.96	0.00	0.95
		0.1	0.93	0.01	0.97	0.95	0.03	0.94	0.96	0.01	0.96	0.96	0.03	0.94	0.95	0.01	0.94
		0.3	0.95	0.01	0.95	0.95	0.03	0.94	0.96	0.01	0.95	0.97	0.04	0.94	0.95	0.01	0.94
		0.5	0.94	0.01	0.95	0.94	0.03	0.95	0.96	0.01	0.94	0.96	0.03	0.96	0.96	0.01	0.95

Table A13: Comparison of coverage of DLL, ReSmoothing (RS), and OraRS in the setting of Gregory et al. (2021), across different sample sizes n, dimension of covariates p, and the correlation parameter r. The columns indexed with "OraRS" stand for the CI centered at the RS estimator with the standard error computed based on 500 point estimates. The f, g_1, g_2, g_3 and g_4 represent the functions of interest to estimate their derivatives at a_0. The entries of the table represents the empirical coverage across 500 simulations.

Setting 1, exactly sparse: Comparison with ReSmoothing

			Bias			SE			Coverage				Length			
a_0	f	n	DLL	RS	Orac	DLL	RS	Orac	DLL	RS	OraRS	Orac	DLL	RS	OraRS	Orac
-1.0	*exp*	500	0.05	0.42	0.03	0.39	0.94	0.35	0.95	0.01	0.91	0.95	1.50	0.05	3.68	1.42
		750	0.02	0.38	0.02	0.36	0.91	0.34	0.94	0.01	0.92	0.94	1.37	0.03	3.58	1.29
		1000	0.01	0.22	0.01	0.34	0.87	0.32	0.94	0.01	0.95	0.93	1.27	0.02	3.41	1.20
	sin	500	0.20	0.91	0.01	0.42	0.83	0.41	0.94	0.01	0.80	0.96	1.69	0.04	3.24	1.68
		750	0.07	0.69	0.01	0.40	0.90	0.39	0.93	0.01	0.88	0.95	1.55	0.02	3.53	1.51
		1000	0.05	0.53	0.01	0.35	0.89	0.35	0.96	0.01	0.92	0.95	1.46	0.02	3.48	1.42
0.5	*exp*	500	0.00	0.49	0.01	0.38	1.01	0.37	0.96	0.01	0.93	0.94	1.51	0.04	3.97	1.41
		750	0.00	0.37	0.01	0.35	0.90	0.33	0.96	0.01	0.92	0.96	1.36	0.02	3.53	1.27
		1000	0.02	0.26	0.01	0.31	0.84	0.31	0.97	0.00	0.93	0.93	1.27	0.01	3.30	1.20
	sin	500	0.20	0.97	0.02	0.42	0.78	0.41	0.92	0.01	0.78	0.95	1.69	0.04	3.06	1.68
		750	0.09	0.66	0.01	0.39	0.91	0.40	0.95	0.01	0.89	0.93	1.57	0.02	3.57	1.53
		1000	0.09	0.51	0.03	0.37	0.87	0.37	0.94	0.01	0.92	0.95	1.46	0.02	3.41	1.41

Setting 1, approximately sparse: Comparison with ReSmoothing

			Bias			SE			Coverage				Length			
a_0	f	n	DLL	RS	Orac	DLL	RS	Orac	DLL	RS	OraRS	Orac	DLL	RS	OraRS	Orac
-1.0	*exp*	500	0.09	0.73	0.00	0.35	0.85	0.35	0.93	0.01	0.86	0.95	1.33	0.04	3.34	1.41
		750	0.07	0.59	0.03	0.32	0.88	0.33	0.95	0.01	0.90	0.95	1.26	0.02	3.46	1.28
		1000	0.04	0.61	0.01	0.31	0.84	0.31	0.94	0.00	0.88	0.95	1.20	0.02	3.28	1.20
	sin	500	0.26	0.80	0.02	0.39	0.46	0.46	0.86	0.00	0.65	0.91	1.46	0.02	1.81	1.67
		750	0.18	0.73	0.01	0.35	0.63	0.38	0.94	0.00	0.80	0.95	1.42	0.01	2.46	1.51
		1000	0.12	0.66	0.01	0.35	0.73	0.38	0.94	0.00	0.86	0.93	1.37	0.01	2.87	1.42
0.5	*exp*	500	0.15	0.62	0.01	0.34	0.81	0.36	0.94	0.01	0.88	0.94	1.33	0.03	3.17	1.42
		750	0.06	0.53	0.04	0.33	0.89	0.33	0.94	0.01	0.90	0.95	1.26	0.02	3.47	1.28
		1000	0.08	0.48	0.00	0.31	0.87	0.31	0.95	0.00	0.91	0.96	1.20	0.01	3.39	1.20
	sin	500	0.29	1.08	0.02	0.40	0.44	0.46	0.86	0.00	0.26	0.93	1.47	0.02	1.72	1.68
		750	0.23	0.96	0.03	0.38	0.60	0.40	0.89	0.00	0.63	0.94	1.43	0.01	2.34	1.53
		1000	0.14	0.78	0.00	0.33	0.73	0.35	0.94	0.01	0.81	0.95	1.37	0.01	2.88	1.41

Table A14: Comparison of DLL, ReSmoothing (RS), OraRS, and oracle (Orac) estimators in Setting 1 with $p = 750$, across different sample sizes n, evaluation points a_0, and function of interest f (making inference for $f'(a_0)$). The columns indexed with "Bias", and "SE" report the absolute bias, and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

F.5 Comparison with the Post Nonparametric Double Selection method

We implement the Post-nonparametric Double Selection method (PND Span) proposed in Kozbur (2021) using the authors' original code available at https://www.tandfonline.com/doi/suppl/10.1080/07350015.2020.1753524?scroll=top. Note that our method focuses on the inference of the function derivative at a single evaluation point, while the simulation of Kozbur (2021) and its publicly available code focus on the inference for average function derivative at multiple evaluation points, including our inference target as a special case. The only changes we make to their code are the data generating process and the evaluation points part, where the original code computes average of the derivatives over a set of evaluation points, and we modify it to compute the derivative point-wisely by specifying only one point in that set. After specifying the evaluation point, the inference procedure directly employs the functions in the original code.

We generate the data using the simulation setting previously considered in Gregory et al. (2021); Meier et al. (2009). $\{D_i\}_{i=1}^n$ and $\{X_{i,j}\}_{i=1}^n$ for $j = 1, 2, \cdots, p-1$ are generated as independent random variables from the Uniform(-2.5, 2.5) distribution. We consider the following non-zero functions in the outcome model and set all other functions as zero:

$$f(d) = -\sin(2d); \quad g_1(x) = x^2 - 25/12; \quad g_2(x) = x; \quad g_3(x) = e^{-x} - 2/5\sinh(5/2)$$

The dimension p is set as 150 and the sample size n varies in $\{100, 300, 500\}$. We are interested in estimating the derivative of f at evaluation points $a_0 \in \{-1, -0.25, 0.5, 0.75, 1.25\}$ and we also consider exchanging the roles of f and g_1. The results are summarised in Table A15. Our proposed DLL method has a much smaller bias than PND Span and achieves the desired coverage. Although achieving the desirede converage, the PND Span method has a much larger bias and longer confidence interval.

We also consider the data generating process as in our Setting 1 with $p = 750$ and $n \in \{500, 750, 1000\}$. The function of interest is $f(d) = 1.5\sin(d)$, and we report the results at evaluation points $a_0 \in \{-1, 0.1, 0.25, 0.5, 1\}$ in Table A16. We observe that the PND Span method suffers from large bias at most evaluation points while our proposed DLL method has a small bias. Both methods achieves the desired coverage level but our method has a much shorter confidence interval.

Comparison with `PND Span`: $f(d) = -\sin(2d)$

		Bias		SE		Coverage		Length	
a_0	n	`DLL`	`PND`	`DLL`	`PND`	`DLL`	`PND`	`DLL`	`PND`
-1	100	0.15	0.11	1.35	6.25	0.91	0.94	4.53	23.64
	300	0.05	0.20	0.96	2.55	0.95	0.95	3.76	9.92
	500	0.06	0.13	0.90	3.28	0.92	0.95	3.28	12.94
-0.25	100	0.44	0.22	1.23	22.13	0.88	0.95	4.56	38.40
	300	0.14	0.79	0.94	7.44	0.95	0.96	3.68	30.80
	500	0.09	0.29	0.87	7.62	0.95	0.95	3.29	31.00
0.5	100	0.28	0.08	1.26	7.72	0.89	0.94	4.62	29.78
	300	0.01	0.09	0.99	5.95	0.94	0.94	3.72	22.68
	500	0.03	0.42	0.84	4.40	0.95	0.95	3.28	17.64
0.75	100	0.04	0.15	1.26	9.17	0.93	0.95	4.46	36.46
	300	0.01	0.05	1.00	3.39	0.94	0.95	3.73	13.80
	500	0.01	0.05	0.83	3.77	0.95	0.94	3.30	14.46
1.25	100	0.37	0.10	1.20	4.94	0.89	0.94	4.40	19.60
	300	0.19	0.06	0.96	3.32	0.95	0.94	3.74	12.86
	500	0.10	0.09	0.85	2.28	0.94	0.95	3.29	8.62

Comparison with `PND Span`: $f(d) = d^2 - 25/12$

		Bias		SE		Coverage		Length	
a_0	n	`DLL`	`PND`	`DLL`	`PND`	`DLL`	`PND`	`DLL`	`PND`
-1	100	0.10	1.22	1.14	5.75	0.94	0.92	4.51	21.44
	300	0.04	0.05	0.96	3.30	0.93	0.95	3.57	12.44
	500	0.04	0.32	0.84	2.53	0.94	0.95	3.16	9.82
-0.25	100	0.01	1.40	1.27	47.89	0.95	0.95	4.52	43.20
	300	0.02	0.30	0.93	8.16	0.94	0.94	3.56	31.20
	500	0.03	0.06	0.85	6.36	0.94	0.92	3.18	23.60
0.5	100	0.10	0.46	1.25	7.95	0.93	0.93	4.58	28.56
	300	0.02	0.12	0.91	6.24	0.95	0.94	3.55	23.04
	500	0.04	0.45	0.83	4.66	0.94	0.94	3.12	17.26
0.75	100	0.12	0.58	1.23	8.41	0.94	0.95	4.52	33.18
	300	0.08	0.04	0.91	4.32	0.96	0.95	3.58	16.60
	500	0.01	0.15	0.78	3.48	0.96	0.93	3.14	13.76
1.25	100	0.09	0.28	1.24	4.89	0.93	0.94	4.38	18.56
	300	0.09	0.49	0.92	2.73	0.94	0.93	3.54	10.40
	500	0.02	0.32	0.80	2.00	0.96	0.94	3.16	7.02

Table A15: Comparison of `DLL`, `PND Span` (`PND`) in simulation settings of Meier et al. (2009); Gregory et al. (2021) with $p = 150$, across different sample sizes n and evaluation points a_0. The columns indexed with "Bias" and "SE" report the absolute bias and the standard error computed by 500 estimates, respectively; the columns indexed with "Coverage" report the empirical coverage level and the columns indexed with "Length" report the average CI length.

Setting 1, exactly sparse: Comparison with PND Span

		Bias			SE			Coverage			Length		
a_0	n	DLL	PND	Orac	DLL	PND	Orac	DLL	PND	Orac	DLL	PND	Orac
-1.0	500	0.14	0.66	0.01	0.43	1.99	0.42	0.92	0.94	0.94	1.72	7.66	1.68
	750	0.05	0.82	0.03	0.41	2.34	0.39	0.95	0.87	0.96	1.65	7.72	1.53
	1000	0.05	0.74	0.01	0.39	2.15	0.36	0.95	0.94	0.97	1.57	8.40	1.43
0.1	500	0.15	0.49	0.01	0.39	5.48	0.38	0.93	0.98	0.94	1.54	23.38	1.52
	750	0.05	0.21	0.02	0.37	6.81	0.35	0.95	0.94	0.96	1.47	25.10	1.37
	1000	0.10	0.06	0.03	0.35	7.27	0.32	0.96	0.96	0.94	1.40	27.38	1.28
0.25	500	0.15	0.10	0.00	0.40	4.06	0.39	0.92	0.97	0.94	1.58	17.14	1.56
	750	0.07	0.36	0.00	0.38	5.84	0.36	0.95	0.94	0.94	1.52	21.52	1.42
	1000	0.05	0.59	0.02	0.36	3.75	0.33	0.93	0.95	0.94	1.44	15.44	1.32
0.5	500	0.16	0.62	0.02	0.47	3.65	0.48	0.91	0.96	0.93	1.74	13.68	1.70
	750	0.07	0.13	0.01	0.42	2.71	0.39	0.94	0.96	0.95	1.64	10.94	1.53
	1000	0.07	0.15	0.02	0.40	2.91	0.37	0.94	0.95	0.95	1.57	11.82	1.44
1	500	0.11	1.37	0.03	0.57	2.87	0.58	0.95	0.91	0.92	2.26	10.64	2.19
	750	0.08	0.78	0.00	0.52	3.10	0.51	0.97	0.94	0.94	2.12	11.08	1.97
	1000	0.06	0.45	0.01	0.51	3.46	0.49	0.94	0.91	0.94	2.02	12.26	1.86

Table A16: Comparison of DLL, PND Span (PND), and oracle (Orac) estimators for Settings 1 with $p = 750$, across different sample sizes n and evaluation points a_0. The columns indexed with “Bias” and “SE” report the absolute bias and the standard error computed by 500 estimates, respectively; the columns indexed with “Coverage” report the empirical coverage level and the columns indexed with “Length” report the average CI length.