Journal of Machine Learning Research 21 (2020) 1-45 Submitted 3/19; Revised 4/20; Published 4/20

Smoothed Nonparametric Derivative Estimation using Weighted Difference Quotients

Yu Liu YULIU@IASTATE.EDU
Department of Computer Science
Iowa State University
Ames, IA 50011, USA

Kris De Brabanter KBRABANT@IASTATE.EDU
Department of Statistics
Department of Industrial Manufacturing & Systems Engineering
Iowa State University
Ames, IA 50011, USA

Editor: Garvesh Raskutti

Abstract

Derivatives play an important role in bandwidth selection methods (e.g., plug-ins), data analysis and bias-corrected confidence intervals. Therefore, obtaining accurate derivative information is crucial. Although many derivative estimation methods exist, the majority require a fixed design assumption. In this paper, we propose an effective and fully data-driven framework to estimate the first and second order derivative in random design. We establish the asymptotic properties of the proposed derivative estimator, and also propose a fast selection method for the tuning parameters. The performance and flexibility of the method is illustrated via an extensive simulation study.

Keywords: derivative estimation, asymptotic properties, random design

1. Introduction

The next section describes previous methods and the current state-of-the-art for nonparametric derivative estimation. Also, we summarize the main differences between derivative estimation in the equispaced and random design for our type of estimator and give a brief overview of local polynomial regression.

1.1. Previous work and current state-of-the-art

Since the mid sixties nonparametric density and regression estimation have become a popular and well studied area in statistics. These methods have provided researchers with more flexibility to analyze data without relying on parametric assumptions. Although the literature of nonparametric regression estimators is vast, see e.g., Fan and Gijbels (1996), Györfi et al. (2006) and Tsybakov (2008), derivative estimation also plays an important role in different research areas and applications such as exploration of the structure of data (detecting jump discontinuities (Gijbels and Goderniaux, 2005), revealing important features from curve estimation (Chaudhuri and Marron, 1999), analyzing significant trends (Rondonotti et al., 2007)), comparing regression curves (Park and Kang, 2008), bias-corrected confidence intervals (Eubank and Speckman, 1993; Xia, 1998), analyzing human growth data (Müller, 2012; Ramsay and Silverman, 2007) and neural network pruning (Hassibi and Stork, 1993).

Our proposed methodology provides a data-driven way to estimate derivatives nonparametrically without having to estimate the regression model first. This is especially important when the regression function is difficult to estimate. Although a myriad of papers are published regarding derivative estimation in the mid nineties, many open problems still remain. Ramsay (1998) noted that typically one sees derivatives go wild at the extremes, and the higher the order of the derivative, the wilder the behavior. Further, problems arise in the smoothing parameter or bandwidth selection processes where cross-validation (CV) and generalized CV can be poor guides (Härdle, 1990). Based on Rice (1986), Charnigo et al. (2011) proposed a generalized C_p criterion to determine the smoothing/tuning parameters for derivative estimation for the equispaced design case.

Nonparametric derivative estimation methods can be categorized in three groups: local polynomial regression, regression/smoothing splines, and difference quotients (Müller et al., 1987). Due to the tremendous and well established work done in the field of local polynomial smoothing, the research activity regarding to nonparametric regression and derivative estimation seems to be somewhat stalled. In local polynomial regression, the derivative can be estimated by the coefficient of the q-th order derivative of the local polynomial regression fitted at point x, i.e. the local slope. Theoretical properties are studied in Fan and Gijbels (1996) and Delecroix and Rosa (1996). The bandwidth choice for the derivative estimator (based on a factor rule) is discussed in Fan and Gijbels (1996). Stone (1985) showed that derivative estimation with splines can achieve the optimal L_2 rate of convergence under mild assumptions. Further asymptotic properties are obtained by Zhou and Wolfe (2000) in the random design setting. However, the smoothing parameter selection problem remained unanswered. Wahba and Wang (1990) noticed that this was particularly difficult for smoothing splines since the smoothing parameter depends on the order of the derivative.

Difference quotient based derivative estimators (Müller et al., 1987; Härdle, 1990) produce a noisy data set which can be smoothed by any nonparametric regression estimator. Smoothing turns out to be quite difficult in practice due to difference quotient's large variance which is $O(n^2)$, where n is the sample size. Therefore, the main goal is to significantly reduce the variance at the cost of a slight bias increase. To obtain such a variance reduction, Iserles (2009) proposed a variance-reducing linear combination of k symmetric difference quotients in the field of numerical mathematics where k is considered to be a tuning parameter. Difference quotients are certainly not new (Müller et al., 1987; Charnigo et al., 2011; De Brabanter et al., 2013), but all results were obtained under the equispaced design assumption. Extending these estimators to the random design setting is possible, however they are no longer consistent for derivatives of order higher than two. This is due to the accumulation of errors associated with the design which will cause higher order terms to blow up. Such an effect is not present when considering equispaced design. Wang and Lin (2015) proposed a sequence of approximate linear regression representations in which the derivative is the intercept term. Although their results are very appealing, they rely on rather stringent assumptions on the regression function. These assumptions are relaxed in Dai et al. (2016) where a linear combination of the dependent variables, depending on two tuning parameters, are used to obtain derivatives. The variance reducing weights are obtained by solving a constraint optimization problem for which the authors derived a closed form solution. They further showed that the symmetric form used in Charnigo et al. (2011) and De Brabanter et al. (2013) reduces the order of estimation bias without increasing the estimation variance in the interior. They proposed an asymmetric estimator for the derivatives at the boundaries. All results from Wang and Lin (2015) and Dai et al. (2016) assume an equispaced design and both authors do not mention the extension to the random design setting.

In this paper we extend the difference quotient based estimator to the random design to estimate the first order derivative and propose a new consistent estimator for second order derivatives. This framework is flexible so it can be used to extend other difference based estimators in fixed design to the random design. An initial idea of this paper is given in the conference paper of Liu and De Brabanter (2018). Since it is not straightforward to propose an asymptotically consistent estimator for the general case, we will first provide a framework to estimate the first and second order derivative in the uniform random design and then generalize it to arbitrary distributions. Because this method produces a new data set containing correlated errors, we use the local polynomial regression estimator with bandwidth selection method of De Brabanter et al. (2018) to smooth the noisy derivatives and derive the asymptotic properties of the smoothed derivative estimators.

The paper is organized as follows. We discuss the main theoretical differences between equispaced and random design for this type of estimator and give a short description of local polynomial regression. Section 2 illustrates the first order derivative estimation based on variance reducing weighted difference quotients. Bias, variance and pointwise consistency are established. In addition, bandwidth selection and behavior at the boundary for noisy derivative estimators are also described. Finally, we discuss how to smooth the data with correlated noise and study the asymptotic properties of the smoothed derivatives. Section 3 extends the framework to second order derivatives. In section 4, we conduct Monte Carlo experiments to compare the proposed methodology with smoothing splines and local polynomial regression. Finally, Section 5 states the conclusions and future work.

1.2. Equispaced design vs. random design

Consider the data (X_1, Y_1),...,(X_n, Y_n) which form an independent and identically distributed (i.i.d.) sample from a population (X, Y), where $X_i \in \mathcal{X} = [a, b] \subseteq \mathbb{R}$ and $Y \in \mathbb{R}$ for all $i = 1, \ldots, n$. In the equispaced design case, the response variables are assumed to satisfy

$$Y_i = m(x_i) + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where $x_1, \ldots, x_n$ are nonrandom numbers and $x_{i+1} - x_i = (b - a)/(n - 1)$ is constant for all i. In this setting, the regression function is given by $m(x) = \mathbf{E}[Y]$ and we assume that $\mathbf{E}[e] = 0$ and $\mathbf{Var}[e] = \sigma_e^2 < \infty$. In contrast to the equispaced design, the X are random variables in random design and are generated from an unknown density and distribution f and F respectively. Consider the following model

$$Y_i = m(X_i) + e_i, \quad i = 1, \ldots, n, \tag{2}$$

where the regression function is given by $m(X) = \mathbf{E}[Y|X = x]$ and assume that $\mathbf{E}[e] = 0$, $\mathbf{Var}[e] = \sigma_e^2 < \infty$, X and e are independent. The derivative estimators discussed in Charnigo et al. (2011) and De Brabanter et al. (2013) use the symmetric property $x_{i+j} - x_i = x_i - x_{i-j}$ since they both assumed equispaced design. However, in the random design this property no longer holds which introduces extra estimation error. In addition, it is fairly complicated to obtain an asymptotic expression for the difference $X_{i+j} - X_i$ when the X's are generated from an unknown distribution, leading to theoretical difficulties in obtaining asymptotic properties of the derivative estimator.

1.3. Local Polynomial Regression

The local polynomial regression estimator in an arbitrary point x is given by minimizing the following weighted least squares problem (Fan and Gijbels, 1996)

$$\min_{\beta_j \in \mathbb{R}} \sum_{i=1}^{n} \Big\{Y_i - \sum_{j=0}^{p} \beta_j \, (X_i - x)^j\Big\}^2 K_h(X_i - x), \tag{3}$$

where β_j are the solutions to the weighted least squares problem, K is a symmetric probability density function with $K_h(\cdot) = K(\cdot/h)/h$. Note that $\hat{m}^{(q)}(x) = q!\hat{\beta}_q$ is an estimator for the q-th order derivative $m^{(q)}(x)$, $q = 0, 1, \dots, p$. In matrix notation the solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \, \mathbf{W} \, \mathbf{X})^{-1} \, \mathbf{X}^T \, \mathbf{W} \, \mathbf{y},$$

where $\mathbf{y} = (Y_1, \dots, Y_n)^T$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ and

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^p \end{pmatrix},$$

with $\mathbf{W} = \mathrm{diag}\{K_h(X_i - x)\}$ a $n \times n$ diagonal matrix of weights based on the kernel function and the bandwidth h.

2. First order derivative estimation

Müller et al. (1987) introduced the first order difference quotients to produce noisy derivative data. However, all their results are obtained for the equispaced design setting. In case of random design, their estimator for the first order (noisy) derivative at design point X_i is denoted by $\hat{q}_i^{(1)}$ and is

$$\hat{q}^{(1)}(X_i) = \hat{q}_i^{(1)} = \frac{Y_i - Y_{i-1}}{X_i - X_{i-1}}. \tag{4}$$

Although quite appealing and intuitive, this estimator has major drawbacks, i.e. (i) a large variance and (ii) difficulties in studying its asymptotic properties in random design. The variance is $O(n^2)$ and $O_p\{(X_i - X_{i-1})^{-2}\}$ for the equispaced and random design respectively. In the latter case, it is obvious that this can be very large when the distance between two neighboring X is small. Consequently, reducing variance in these type of estimators is paramount and can be accomplished by means of a variance-reducing linear combination of symmetric difference quotients. Second, in order to discuss the asymptotic properties of this different quotient, we need to obtain an asymptotic expression for the difference $X_i - X_{i-1}$ which is not trivial in the random design setting. However, in a special case, i.e. $X = U \sim \mathcal{U}(0,1)$ and arranging the random variables in order of magnitude according to U (order statistics), the asymptotic properties of the first order quotient (4) can be obtained. In what follows, $\mathcal{U}(0,1)$ denotes the uniform distribution between 0 and 1. For the sake of simplicity, we will first discuss a special case, i.e. $U = X \sim \mathcal{U}(0,1)$, before we formulate the estimator for arbitrary distributions.

2.1. Approach based on order statistics

Consider n bivariate data forming an i.i.d sample from a population (U, Y) and further assume $U \sim \mathcal{U}(0,1)$. Arrange the bivariate data (U, Y) in order of magnitude according to U, i.e. $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$ where $U_{(i)}$, $i = 1, \ldots, n$ is the i-th order statistic. In order to avoid ties and hence division by zero we also require $U_{(i)} \neq U_{(j)}$ for $i \neq j$. The first order difference quotient (4) is

$$\hat{q}^{(1)}(U_{(i)}) = \hat{q}_i^{(1)} = \frac{Y_i - Y_{i-1}}{U_{(i)} - U_{(i-1)}}. \tag{5}$$

The difference $U_{(i)} - U_{(i-1)}$ is the difference of uniform order statistics and it is well-known that (David and Nagaraja, 1970, p. 14)

$$U_{(s)} - U_{(r)} \sim \text{Beta}(s - r, n - s + r + 1) \quad \text{for } s > r.$$

This result immediately leads to Lemma 1.

Lemma 1 *Let* $U \overset{i.i.d.}{\sim} \mathcal{U}(0,1)$. *Arrange the random variables in order of magnitude* $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$. *Then, for* $i > j$

$$U_{(i+j)} - U_{(i-j)} = \frac{2j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right),$$

$$U_{(i+j)} - U_{(i)} = \frac{j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right)$$

and

$$U_{(i)} - U_{(i-j)} = \frac{j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right).$$

Proof: see Appendix A. ■

Consider the model

$$Y_i = r(U_{(i)}) + e_i, \tag{6}$$

where $r(u) = \mathbf{E}[Y|U = u]$ is the regression function and assume $\mathbf{E}[e] = 0$, $\mathbf{Var}[e] = \sigma_e^2 < \infty$, U and e are independent. Assume r is twice continuously differentiable on $[0, 1]$. A Taylor expansion of $r(U_{(i\pm j)})$ in a neighborhood of $U_{(i)}$ gives

$$r(U_{(i\pm j)}) = r(U_{(i)}) + r^{(1)}(U_{(i)})(U_{(i\pm j)} - U_{(i)}) + O_p\left(\frac{j^2}{n^2}\right). \tag{7}$$

Using Lemma 1 for $j = 1$ and (7) yields

$$\mathbf{E}[\hat{q}_i^{(1)}|U_{(i-1)}, U_{(i)}] = \mathbf{E}\left[\frac{Y_i - Y_{i-1}}{U_{(i)} - U_{(i-1)}}|U_{(i-1)}, U_{(i)}\right] = r^{(1)}(\xi_i)$$

for $\xi_i \in [U_{(i-1)}, U_{(i)}]$ and

$$\mathbf{Var}\big[\hat{q}_i^{(1)}|U_{(i-1)}, U_{(i)}\big] = \mathbf{Var}\left[\frac{Y_i - Y_{i-1}}{U_{(i)} - U_{(i-1)}}|U_{(i-1)}, U_{(i)}\right] = \frac{2\sigma_e^2}{(U_{(i)} - U_{(i-1)})^2} = O_p(n^2).$$

It is immediately clear that this estimator is asymptotically unbiased. However, the variance of this estimator can be arbitrary large and hence it will be difficult to estimate the smoothed derivative function. A possible way to reduce the variance is described in Iserles (2009) and used in Charnigo et al. (2011) and De Brabanter et al. (2013) which involves a combination of symmetric difference quotients around the i-th point. Our proposed derivative estimator for random design involving uniform order statistics is

$$\hat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right), \tag{8}$$

where the weights $w_{i,1}, \ldots, w_{i,k}$ sum up to one. Note that (8) is valid for $k+1 \le i \le n-k$ and hence $k \le (n-1)/2$. For the boundary regions, i.e. $2 \le i \le k$ and $n-k+1 \le i \le n-1$, the estimator (8) needs to be modified and is discussed in Section 2.3.2. The estimator (8) does not provide results for $\hat{Y}_1^{(1)}$ and $\hat{Y}_n^{(1)}$. One can ignore these two points from consideration or have them coincide with $\hat{Y}_2^{(1)}$ and $\hat{Y}_{n-1}^{(1)}$ (see Charnigo et al. (2011)).

The following proposition states the optimal weights $w_{i,j}$, optimal in the sense of minimizing the variance of the estimator (8).

Proposition 1 *For $k+1 \le i \le n-k$ and under model* (6)*, the weights $w_{i,j}$ that minimize the variance of* (8)*, satisfying $\sum_{j=1}^{k} w_{i,j} = 1$, are given by*

$$w_{i,j} = \frac{(U_{(i+j)} - U_{(i-j)})^2}{\sum_{l=1}^{k}(U_{(i+l)} - U_{(i-l)})^2}, \quad j = 1, \ldots, k. \tag{9}$$

Proof: see Appendix B. ■

For fixed i, the j-th weight (9) is proportional to the inverse variance of the difference quotient $\frac{Y_{i+j}-Y_{i-j}}{U_{(i+j)}-U_{(i-j)}}$ in (8). At first sight, these weights seem to be different than the weights obtained by Charnigo et al. (2011) and De Brabanter et al. (2013) for the equispaced design case. Plugging in the difference $u_{i+j} - u_{i-j} = 2j(b-a)/(n-1)$ for equispaced design on $[a, b]$ yields

$$w_{i,j} = \frac{(u_{i+j} - u_{i-j})^2}{\sum_{l=1}^{k}(u_{i+l} - u_{i-l})^2} = \frac{\frac{4j^2}{(n-1)^2}}{\frac{4}{(n-1)^2}\sum_{l=1}^{k} l^2} = \frac{6j^2}{k(k+1)(2k+1)}.$$

These are exactly the weights obtained in Charnigo et al. (2011) & De Brabanter et al. (2013). This shows that the weights for equispaced design are a special case of the weights in Proposition 1. However, one parameter still remains unknown, i.e. k, the number of symmetric difference quotients (around i). Theorem 1 (asymptotic conditional bias and variance) provides valuable insights how to choose k.

2.2. Asymptotic properties of the first order derivative estimator

The following theorems establish the asymptotic conditional bias and variance of our proposed estimator (8) for the interior points, i.e. $k+1 \leq i \leq n-k$. In what follows we denote $\mathbb{U} = (U_{(i-j)}, \ldots, U_{(i+j)})$ for $i > j$, $i + j \leq n$ and $j = 1, \ldots, k$.

Theorem 1 *Under model* (6) *and assume r is twice continuously differentiable on* $[0,1]$ *and $k \to \infty$ as $n \to \infty$. Then, for uniform random design on* $[0,1]$ *and the weights in Proposition 1, the conditional (absolute) bias and conditional variance of* (8) *are*

$$\left|\text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]\right| \leq \sup_{u\in[0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} + o_p(n^{-1}k)$$

and

$$\mathbf{Var}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] = \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} + o_p(n^2k^{-3})$$

uniformly for $k+1 \leq i \leq n-k$.

Proof: see Appendix C. ■

From Theorem 1, the pointwise consistency of (8) immediately follows.

Corollary 1 *Under the assumptions of Theorem 1, $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$ and $n^2k^{-3} \to 0$. Then, for $\sigma_e^2 < \infty$ and the weights given in Proposition 1, we have for any $\varepsilon > 0$*

$$\mathbf{P}(|\hat{Y}_i^{(1)} - r^{(1)}(U_{(i)})| \geq \varepsilon) \to 0$$

for $k+1 \leq i \leq n-k$.

Proof: see Appendix D. ■

According to Theorem 1 and Corollary 1, the conditional bias and conditional variance of (8) tend to zero and k is at least $O(n^{2/3})$ but slower than $O(n)$. Next, we develop a rule-of-thumb tuning method for k such that $k = O(n^{4/5})$ and the fastest possible rate at which $\mathbf{E}[(\hat{Y}_i^{(1)} - r^{(1)}(U_{(i)}))^2|\mathbb{U}] \to 0$ (L_2 rate of convergence) is $O_p(n^{-2/5})$. Using Jensen's inequality, similar results can be shown for the L_1 rate of convergence, i.e.

$$\mathbf{E}\big[|\hat{Y}_i^{(1)} - r^{(1)}(U_{(i)})| \mid \mathbb{U}\big] \leq \left|\text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]\right| + \sqrt{\mathbf{Var}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]} = O_p(n^{-1/5}).$$

From Theorem 1, it is clear that the parameter k in (8) controls the bias-variance trade-off. Based on Theorem 1, we choose k that minimizes the asymptotic upper bound of the conditional mean integrated squared error (MISE). The result is given in Corollary 2.

Corollary 2 *Under the assumptions of Theorem 1 and denote $\mathcal{B} = \sup_{u\in[0,1]} |r^{(2)}(u)|$, then k that minimizes the asymptotic upper bound of the conditional MISE is*

$$k_{\text{opt}} = \underset{k\in\mathbb{N}^+\setminus\{0\}}{\arg\min} \left\{ \mathcal{B}^2 \frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2} + \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} \right\}.$$

Proof: see Appendix E. ■

Corollary 2 provides a fast and easy parameter tuning method in practice, however some unknown quantities still need to be estimated. The error variance can be estimated by Hall's $\sqrt{n}$-consistent estimator (Hall et al., 1990)

$$\hat{\sigma}_e^2 = \frac{1}{n-2}\sum_{i=1}^{n-2}(0.809Y_i - 0.5Y_{i+1} - 0.309Y_{i+2})^2.$$

The second unknown quantity $\mathcal{B}$ can be (roughly) estimated with a local polynomial regression estimator of order $p = 3$. The performance of our proposed model is not so sensitive to the accuracy of $\mathcal{B}$, thus a rough estimate of the second order derivative is sufficient. By plugging in these two estimators for σ_e^2 and $\mathcal{B}$ in Corollary 2, the optimal value k_{opt} can be obtained for example by a grid search over the integer set $[1, \lfloor \frac{n-1}{2} \rfloor]$ where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. As an alternative, any root solving algorithm can also be used.

Remark 1 *By setting the derivative of the expression in Corollary 2 to zero, we cannot obtain a closed form for k_{opt}. However, by only retaining the higher order terms, we can obtain a fairly accurate estimate for k_{opt}. Given estimates for σ_e^2 and $\mathcal{B}$ we have*

$$\hat{k}_{\text{opt}} = \lfloor 2^{4/5}\hat{\sigma}_e^{2/5}\widehat{\mathcal{B}}^{-2/5}n^{4/5} \rfloor.$$

2.3. Exact bias expression and boundary correction

In this section, we further investigate the bias and propose a simple but effect boundary correction to reduce the variance by adding a small amount of bias.

2.3.1. Asymptotic order of the conditional bias and continuous differentiability of the regression function

In Theorem 1, we bounded the conditional bias above. From a theoretical point of view, it is helpful to derive an exact expression for the conditional bias and discuss its dependence on the continuous differentiability of the regression function r. It also allows us to compare with the bias in fixed design and explain the extra bias due to the asymmetric differences $U_{(i+j)} - U_{(i)} \neq U_{(i)} - U_{(i-j)}$ in random design. Assume the first $q+1$ derivatives of r exist on $[0, 1]$. A Taylor series of $r(U_{(i\pm j)})$ in a neighborhood of U_i and using Lemma 1 yields

$$\begin{aligned} r(U_{(i+j)}) &= r(U_{(i)}) + \sum_{l=1}^{q}\frac{1}{l!}(U_{(i+j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p(U_{(i+j)} - U_{(i)})^{q+1} \\ &= r(U_{(i)}) + \sum_{l=1}^{q}\frac{1}{l!}(U_{(i+j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p\{(j/n)^{q+1}\} \end{aligned}$$

and

$$\begin{aligned} r(U_{(i-j)}) &= r(U_{(i)}) + \sum_{l=1}^{q}\frac{1}{l!}(U_{(i-j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p(U_{(i-j)} - U_{(i)})^{q+1} \\ &= r(U_{(i)}) + \sum_{l=1}^{q}\frac{1}{l!}(U_{(i-j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p\{(j/n)^{q+1}\}. \end{aligned}$$

Using Lemma 1, assume $k \to \infty$ as $n \to \infty$, and for the weights in Proposition 1 we obtain the asymptotic order of the exact conditional bias for different values of q

$$\text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] = \begin{cases} O_p\left(\frac{k}{n}\right) & , \quad q = 1. \\ O_p\left(\max\left\{\frac{k^{\frac{1}{2}}}{n}, \frac{k^2}{n^2}\right\}\right), & q \geq 2. \end{cases}$$

The proof is given in Appendix F. For $q = 1$ (i.e. r is twice continuously differentiable), the leading order of exact conditional bias is the same as that of the bias upperbound given in Theorem 1. For $q = 2$, r is three times continuously differentiable on $[0, 1]$, the exact bias achieves smaller order than $O_p(k/n)$. Unfortunately, adding additional assumptions on the differentiability of r, i.e. $q > 2$, will no longer improve the asymptotic rate of the bias. This can be seen as follows: for $q \geq 2$, the bias is

$$\text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] = \frac{\sum_{j=1}^k (U_{(i+j)} - U_{(i-j)})\big[\sum_{l=2}^q \frac{r^{(l)}(U_{(i)})\{(U_{(i+j)}-U_{(i)})^l-(U_{(i-j)}-U_{(i)})^l\}}{l!} + O_p\{(j/n)^{q+1}\}\big]}{\sum_{p=1}^k (U_{(i+p)} - U_{(i-p)})^2}.$$

This can be split into two terms: odd and even with $l \geq 2$

$$\text{bias}_{\text{odd}}[\hat{Y}_i^{(1)}|\mathbb{U}] = O_p\left(\frac{k^2}{n^2}\right) \qquad \text{and} \qquad \text{bias}_{\text{even}}[\hat{Y}_i^{(1)}|\mathbb{U}] = O_p\left(\frac{k^{\frac{1}{2}}}{n}\right).$$

resulting in

$$\begin{aligned} \text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= \text{bias}_{\text{odd}}[\hat{Y}_i^{(1)}|\mathbb{U}] + \text{bias}_{\text{even}}[\hat{Y}_i^{(1)}|\mathbb{U}] \\ &= O_p\left\{\max\left(\frac{k^2}{n^2}, \frac{k^{\frac{1}{2}}}{n}\right)\right\}. \end{aligned}$$

In fixed design, $\text{bias}_{\text{even}} = 0$ due to symmetry: $u_{(i+j)} - u_{(i)} = u_{(i)} - u_{(i-j)}$. Unfortunately, in the random design, we cannot remove $\text{bias}_{\text{even}}$. It is this fact that will lead to the inconsistency of third and higher order derivatives if these estimators are defined in a fully recursive way as in Charnigo et al. (2011).

2.3.2. Boundary Correction

We discussed the proposed estimator at the interior points and in this section we provide a simple but effective boundary correction. Points with index $i < k + 1$ and $i > n - k$ are points located at the left and right boundary respectively. Since there are not enough k pairs of neighbors at the boundary, we use a weighted linear combination of $k(i)$ pairs of points U_i instead, where $k(i) = i - 1$ for the left boundary and $k(i) = n - i$ for the right boundary. This is the approach of Charnigo et al. (2011) and De Brabanter et al. (2013). The first order derivative estimator at the boundary is obtained by replacing k with $k(i)$ in (8) and weights in Proposition 1. From Section 2.3.1, we know that if r is three times continuously differentiable on $[0, 1]$ the asymptotic order of the conditional bias at the boundary is $O_p\{\max(\frac{k(i)^2}{n^2}, \frac{k(i)^{1/2}}{n})\}$, which is smaller than for the interior points.

However, the asymptotic order of the conditional variance is $O_p\left\{\frac{3\sigma_e^2(n+1)^2}{k(i)(k(i)+1)(2k(i)+1)}\right\}$ and attains $O_p(n^2)$, as i is close to either 2 or $n-1$.

In order to reduce the variance at the boundary we propose the following modification to (8). For points at the left boundary, $i < k+1$, consider the estimator

$$\hat{Y}_i^{(1)} \quad = \quad \sum_{j=1}^{k(i)} w_{i,j} \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right) + \sum_{j=k(i)+1}^{k} w_{i,j} \cdot \left(\frac{Y_{i+j} - Y_i}{U_{(i+j)} - U_{(i)}}\right) \tag{10}$$

with

$$w_{i,j} = \begin{cases} \dfrac{(U_{(i+j)} - U_{(i-j)})^2}{\sum_{l=1}^{k(i)}(U_{(i+l)} - U_{(i-l)})^2 + \sum_{l=k(i)+1}^{k}(U_{(i+l)} - U_{(i)})^2}, & 1 \le j \le k(i); \\ \dfrac{(U_{(i+j)} - U_{(i)})^2}{\sum_{l=1}^{k(i)}(U_{(i+l)} - U_{(i-l)})^2 + \sum_{l=k(i)+1}^{k}(U_{(i+l)} - U_{(i)})^2}, & k(i) < j \le k. \end{cases}$$

This modification leads to

$$\text{bias}[\hat{Y}_i^{(1)}|\mathbb{U}] = O_p\left\{\max\left(\frac{k(i)^{7/2}}{k^3 n}, \frac{k(i)^5}{k^3 n^2}, \frac{k-k(i)}{n}\right)\right\}$$

and

$$\mathbf{Var}[\hat{Y}_i^{(1)}|\mathbb{U}] = O_p\left\{\max\left(\frac{n^2}{k^3}, \frac{n^2(k-k(i))^2}{k^4}\right)\right\}.$$

The proof is given in Appendix G. The bias$[\hat{Y}_i^{(1)}|\mathbb{U}] \to 0$ when $n \to \infty$ indicating that (10) is still asymptotically unbiased at the boundary. Worst case scenario, the variance is of the order $O_p(n^2/k^2)$ which is smaller than $O_p(n^2)$. A similar result can be obtained for the right boundary.

2.4. Smoothing the noisy derivatives

Noisy first order derivative estimators (8) and (10) have two problems: (i) derivative estimators contain the noise coming from the unknown errors $e_i, i = 1, \ldots, n$ in model (6) and (ii) derivative estimators can only be evaluated at the design points $U_{(i)}, i = 1, \ldots, n$. Hence some type of smoothing will be needed to remove the noise and evaluate the derivative in an arbitrary point. The first order derivative estimator (8) can be written as

$$\hat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{r(U_{(i+j)}) - r(U_{(i-j)})}{U_{(i+j)} - U_{(i-j)}}\right) + \sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{e_{i+j} - e_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right) \tag{11}$$

where the second term $\tilde{e}_i = \sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{e_{i+j}-e_{i-j}}{U_{(i+j)}-U_{(i-j)}}\right)$ is the new error and $w_{i,j}$ are given in Proposition 1. We have $Y_i, i = 1, \ldots n$ are independent and $e_i, i = 1, \ldots n$ are independent. It is clear that $\tilde{e}_i, i = 2, \ldots, n-1$ are correlated and the generated derivatives $\hat{Y}_i^{(1)}, i =$

$2, \dots, n-1$ are also correlated. In order to obtain a smoothed version of the derivative, we regard the (correlated) data $(U, \hat{Y}^{(1)})$ as being generated from the model

$$\hat{Y}^{(1)}(U) = r^{(1)}(U) + \tilde{e}.$$

Since the i.i.d. assumption of the errors is no longer valid for the above model, bandwidth selection for any nonparametric smoothing method becomes increasingly difficult (Opsomer et al., 2001; De Brabanter et al., 2018). In this paper we use the idea of De Brabanter et al. (2018) by using a kernel K such that $K(0) = 0$. By using such a kernel, De Brabanter et al. (2018) have shown that under mild assumptions, the effect of the correlation on the bandwidth selection process is removed without any prior knowledge about the correlation structure.

For interior points $k+1 \leq i \leq n-k$, all $\hat{Y}_i^{(1)}$ are asymptotic consistent estimators. Without loss of generality, we show the properties of the smoothed derivative estimator in the interior. The local polynomial estimator at an arbitrary point u_0 is

$$\hat{r}^{(1)}(u_0) = \boldsymbol{\epsilon}_1^T \hat{\boldsymbol{\beta}} = \boldsymbol{\epsilon}_1^T \mathbf{S}_n^{-1} \mathbf{U}_\mathrm{u}{}^T \mathbf{W}_\mathrm{u} \hat{\mathbf{Y}}^{(1)} \tag{12}$$

where $\boldsymbol{\epsilon}_1 = (1, 0, \dots, 0)^T$ is a unit vector with 1 in the first position and $\hat{r}^{(q+1)}(u_0) = q! \hat{\beta}_q$. $\hat{\mathbf{Y}}^{(1)} = (\hat{Y}_{k+1}^{(1)}, \dots, \hat{Y}_{n-k}^{(1)})$, $\mathbf{W}_\mathrm{u}$ is the diagonal matrix of weights, i.e. $\mathrm{diag}\{K_h(U_{(i)} - u_0)\}$ with kernel K, bandwidth h and $K_h(\cdot) = K(\cdot/h)/h$, $\mathbf{S}_n = \mathbf{U}_\mathrm{u}{}^T \mathbf{W}_\mathrm{u} \mathbf{U}_\mathrm{u}$, and

$$\mathbf{U}_\mathrm{u} = \begin{pmatrix} 1 & (U_{(k+1)} - u_0) & \cdots & (U_{(k+1)} - u_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (U_{(n-k)} - u_0) & \cdots & (U_{(n-k)} - u_0)^p \end{pmatrix}.$$

The term $\tilde{e}_i = \sum_{j=1}^k w_{i,j} \cdot \left(\frac{e_{i+j} - e_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right)$ in (11) satisfies $\mathbf{E}[\tilde{e}_i | U] = 0$ and $\mathbf{Cov}(\tilde{e}_i, \tilde{e}_j | U_{(i)}, U_{(j)}) = \sigma_{\tilde{e}}^2 \rho_n(U_{(i)} - U_{(j)})$ for $i \neq j$ with $\sigma_{\tilde{e}}^2 < \infty$ and ρ_n is a stationary correlation function satisfying $\rho_n(0) = 1, \rho_n(u) = \rho_n(-u)$ and $|\rho_n(u)| \leq 1$ for all u. The subscript n allows the correlation function ρ_n to shrink as $n \to \infty$ (De Brabanter et al., 2018). In what follows, we denote $\widetilde{\mathbb{U}} = (U_{(1)}, \dots, U_{(n)})$. Under the following assumptions:

- Assumption 1. $h \to 0$ and $nh \to \infty$ as $n \to \infty$;
- Assumption 2. There exists a constant $K_{\max}$ such that $|K(x)| < K_{\max}$, and $K(x) \geq 0$ for all x;
- Assumption 3. K is symmetric and Lipschitz continuous at 0;
- Assumption 4. $\lim_{|u| \to \infty} |u|^l K(u) < \infty$ for $l = 0, \dots, p$;
- Assumption 5. The correlation function ρ_n is an element of a sequence $\{\rho_n\}$ with the following properties for all n: there exists constants ρ_{max} and ρ_c such that $n \int |\rho_n(x)| dx < \rho_{max}$ and $\lim_{n \to \infty} n \int \rho_n(x) dx = \rho_c$; and for any sequence $\epsilon_n > 0$ satisfying $n\epsilon_n \to \infty$,
$$n \int_{|x| \geq \epsilon_n} |\rho_n(x)| dx \to 0, n \to \infty,$$

Assumption 5 requires the correlation to be short range dependent (Opsomer et al., 2001; Francisco-Fernández et al., 2004). This is a not uncommon assumption in the area of spatial statistics (Cressie, 1993). Two correlation functions satisfying Assumption 5 are

$$\rho_n(x) = \exp(-\alpha\, n|x|) \quad \text{and} \quad \rho_n(x) = \frac{1}{1+\alpha\, n^2 x^2}, \quad \alpha > 0. \tag{13}$$

Via semi-variograms and autocorrelation plots, we can verify the claim that the $\tilde{e}_i$'s support the hypothesis of short-range dependency (and hence Assumption 5 holds). For brevity, we have not included this in the current paper but the interested reader can contact the second author to obtain these results.

Next, we derive the conditional bias and variance of the smoothed derivative $\hat{r}^{(1)}(u_0)$ (for random uniform design on $[0,1]$) by applying Theorem 1 in De Brabanter et al. (2018).

Theorem 2 *Assume $r^{(p+1)}(\cdot), p \geq 1$ be continuous in a neighbourhood of u_0. Under Assumptions $1-7$, $k \to \infty$ as $n \to \infty$, $\sigma_e^2 < \infty$ and the weights given in Proposition 1, the conditional bias and conditional variance of* (12) *for p odd is*

$$\begin{aligned}
\text{bias}\big[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\big] &\leq \boldsymbol{\epsilon}_1^T\, \mathbf{S}^{-1}\left[\frac{c_p}{(p+1)!} r^{(p+2)}(u_0) h^{p+1} + \mathcal{B}\frac{3k(k+1)}{4(n+1)(2k+1)}\tilde{c}_p\right]\{1+o_p(1)\} \\
&= \left[\left(\int t^{p+1} K_0^*(t)dt\right)\frac{1}{(p+1)!} r^{(p+2)}(u_0) h^{p+1}\right. \\
&+ \left.\mathcal{B}\frac{3k(k+1)}{4(n+1)(2k+1)}\left(\int K_0^*(t)dt\right)\right]\{1+o_p(1)\}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{Var}\big[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\big] &= \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}\frac{1+\rho_c}{h(n-2k)}\boldsymbol{\epsilon}_1^T\, \mathbf{S}^{-1}\, \mathbf{S}^*\, \mathbf{S}^{-1}\, \boldsymbol{\epsilon}_1\{1+o_p(1)\} \\
&= \int K_0^{*2}(t)dt \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}\frac{1+\rho_c}{h(n-2k)}\{1+o_p(1)\}
\end{aligned}$$

where $\mathcal{B} = \sup_{u\in[0,1]}|r^{(2)}(u)|$, $\mathbf{S} = (\mu_{i+j})_{0\leq i,j\leq p}$ with $\mu_j = \int u^j K(u)du$, $\mathbf{S}^* = (\nu_{i+j})_{0\leq i,j\leq p}$ with $\nu_j = \int u^j K^2(u)du$, $c_p = (\mu_{p+1},\ldots,\mu_{2p+1})^T$, $\tilde{c}_p = (\mu_0,\mu_1,\ldots,\mu_p)^T$, $\boldsymbol{\epsilon}_1 = (1,0,\ldots,0)^T$, and the equivalent kernel $K_0^*(t) = \boldsymbol{\epsilon}_1^T S^{-1}(1,t,\ldots,t^p)^T K(t)$.

Proof: see Appendix H. ■

The asymptotic upper bound of the conditional MISE is minimized for $h = O(n^{-\frac{2}{5p+6}})$ and $k = O(n^{\frac{3p+4}{5p+6}})$. The corresponding L_2 rate of convergence is $O_p(n^{-\frac{4p+4}{5p+6}})$. In this paper, we will not use the variance-bias trade-off in Theorem 2 to select the bandwidth h and the parameter k simultaneously, since it requires estimating ρ_c, which is not straightforward. To have an easy and efficient tuning method at the cost of a slower rate of convergence, we use Corollary 2 to select k then select bandwidth h as follows. First, use kernel $\overline{K}(u) = (2/\sqrt{\pi})u^2\exp(-u^2)$ to obtain the bandwidth h_b by minimizing the residual sum of squares (RSS) of interior points $(U_{(i)}, \hat{Y}_i^{(1)})$ with $k+1 \leq i \leq n-k$, i.e.

$\text{RSS}(h_b) = (n-2k)^{-1}\sum_{i=k+1}^{n-k}(\hat{r}^{(1)}(U_{(i)}) - \hat{Y}_i^{(1)})^2$. RSS does not contain the boundary points, since noisy derivatives $\hat{Y}_i^{(1)}$ at the boundary have larger variance. Second, as bimodal kernels introduce extra error in the estimation due to their non-optimality we overcome this issue by using $\hat{h}_b$ as a pilot bandwidth and relate it to a bandwidth $\hat{h}$ of a more optimal (unimodal) kernel, say the Gaussian kernel. As shown in De Brabanter et al. (2018), this can be achieved without any extra smoothing step. For local cubic regression, the relation between the bimodal and unimodal bandwidth is

$$\hat{h} = 1.01431\hat{h}_b$$

when using $\overline{K}(u) = (2/\sqrt{\pi})u^2\exp(-u^2)$ and $K(u) = (1/\sqrt{2\pi})\exp(-u^2/2)$ as bimodal and unimodal kernel respectively.

From Theorem 2, the pointwise consistency of (12) for p odd immediately follows.

Corollary 3 *Under the assumptions of Theorem 2, $h \to 0$ and $nh \to \infty$ as $n \to \infty$, $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$ and $nk^{-3}h^{-1} \to 0$. Then, for $\sigma_e^2 < \infty$ and the weights given in Proposition 1, we have for any $\varepsilon > 0$*

$$\mathbf{P}(|\hat{r}^{(1)}(u_0) - r^{(1)}(u_0)| \geq \varepsilon) \to 0$$

Proof: see Appendix I. ■

For $k = O(n^{4/5})$ in Corollary 2 and $h = O(n^{-1/(2p+3)})$ for p odd, then by Corollary 1 in De Brabanter et al. (2018), we have

$$\mathbf{E}[(\hat{r}^{(1)}(u_0) - r^{(1)}(u_0))^2|\widetilde{\mathbb{U}}] = O_p(n^{-2/5})$$

which matches the convergence rate obtained by Stone (1985) for $p = 2$. Using Jensen's inequality, the L_1 rate of convergence is $O_p(n^{-1/5})$.

2.5. Generalizing results for first order derivatives to arbitrary distributions

It is possible to find a closed form expression for the distribution of the differences $X_{(i+j)} - X_{(i-j)}$ with $X \overset{i.i.d}{\sim} F$ where F is unknown and continuous (David and Nagaraja, 1970) such that the density function $f(x) = F'(x)$ and let f be bounded away from zero. Since this result is quite unattractive from a theoretical point of view, we advocate the use of the probability integral transform (PIT) (Casella and Berger, 2002)

$$F(X) \sim U(0,1). \tag{14}$$

By using the probability integral transform we know that the new data set $(F(X_{(1)}), Y_1), \ldots, (F(X_{(n)}), Y_n)$ has the same distribution as $(U_{(1)}, Y_1), \ldots, (U_{(n)}, Y_n)$. This leads to the original setting of uniform order statistics discussed earlier. The final step is to transform back to the original space. In order for this step to work, we need the existence of a density f. Since $m(X) = r(F(X))$ and by the chain rule

$$\frac{dm(X)}{dX} = \frac{dr(U)}{dU}\frac{dU}{dX} = f(X)\frac{dr(U)}{dU}, \tag{15}$$

yielding $m^{(1)}(X) = f(X)r^{(1)}(U)$ which is the smoothed version of the first order derivative in the original space. In practice, the distribution F and density f need to be estimated yielding $\widehat{m}^{(1)}(X) = \widehat{f}(X)\widehat{r}^{(1)}(U)$. In this paper we use the kernel density estimator (Rosenblatt, 1956; Parzen, 1962) to estimate the density f and distribution F with plug-in bandwidth (Wand and Jones, 1994).

Remark 2 *One of the anonymous referees noted that in real data however, it could be that the density f of the data is multi-modal with regions of low dimension in between the modes. Consequently, our assumption that the density f is bounded away from zero does no longer hold and our theoretical results will no longer be valid. A possible remedy for this problem would be to consider a ridge parameter approach similar to approach in density deconvolution (Meister, 2009). Although beyond the scope of this paper, we believe it is an interesting idea for further research.*

3. Higher Order Derivatives

In practice, first and second order derivatives are widely used. However, higher order derivatives become progressively more difficult to estimate, i.e. they suffer from higher bias and variance and consequently slower rate of convergence. In this section, we construct an efficient estimator for the second order derivative and discuss its asymptotic properties. A similar procedure can be applied to estimate derivatives with order higher than two.

3.1. Asymptotic Results for Second Order Noisy Derivatives Under Standard Uniform Distribution

As before, assume $U \sim \mathcal{U}(0,1)$, and (U, Y) are sorted according to ascending order of U. We define the second order noisy derivative estimator as

$$\hat{Y}_i^{(2)} = 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{Y_{i+j+k_1} - Y_{i+j}}{U_{(i+j+k_1)} - U_{(i+j)}} - \frac{Y_{i-j-k_1} - Y_{i-j}}{U_{(i-j-k_1)} - U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}} \tag{16}$$

where both k_1 and k_2 are positive integers and the weights $w_{i,j,2}$ sum up to one. The subscript 2 is used to indicate the weight for the second order derivative. Let ${}^{+}\hat{Y}_{i+j}^{(1)} = \frac{Y_{i+j+k_1} - Y_{i+j}}{U_{(i+j+k_1)} - U_{(i+j)}}$, which is an asymptotically conditional unbiased estimator of $r^{(1)}(U_{(i+j)})$. $\mathbf{Var}[{}^{+}\hat{Y}_{i+j}^{(1)}|\widetilde{\mathbb{U}}] = O_p(\frac{\sigma_e^2 n^2}{k_1^2})$, where the parameter k_1 controls the variance of ${}^{+}\hat{Y}_{i+j}^{(1)}$. The left superscript "+" indicates the estimator only uses data on the right hand side of Y_{i+j}. Similarly ${}^{-}\hat{Y}_{i-j}^{(1)} = \frac{Y_{i-j-k_1} - Y_{i-j}}{U_{(i-j-k_1)} - U_{(i-j)}}$ is an asymptotically conditional unbiased estimator of $r^{(1)}(U_{(i-j)})$ and the conditional variance is $O_p(\frac{\sigma_e^2 n^2}{k_1^2})$. The left superscript "−" indicates the estimator only uses data on the left hand side of Y_{i-j}. Hence the estimator (16) can be rewritten as

$$\hat{Y}_i^{(2)} = \sum_{j=1}^{k_2} w_{i,j,2} \frac{{}^{+}\hat{Y}_{i+j}^{(1)} - {}^{-}\hat{Y}_{i-j}^{(1)}}{C_{i,j,k_1}}$$

where $C_{i,j,k_1} = (U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)})/2$. By defining the estimator in this way, the variance of $\hat{Y}_i^{(2)}$ is reduced by decreasing the correlation between different quotients. Assume r is three times continuously differentiable on the compact interval $[0,1]$, $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$. Applying Lemma 1 gives

$$\begin{aligned}
\mathbf{E}[{}^{+}\hat{Y}_{i+j}^{(1)} - {}^{-}\hat{Y}_{i-j}^{(1)}|\widetilde{\mathbb{U}}] &= \frac{r(U_{(i+j+k_1)}) - r(U_{(i+j)})}{U_{(i+j+k_1)} - U_{(i+j)}} - \frac{r(U_{(i-j-k_1)}) - r(U_{(i-j)})}{U_{(i-j-k_1)} - U_{(i-j)}} \\
&= r^{(1)}(U_{(i+j)}) + \frac{1}{2} r^{(2)}(U_{(i+j)})(U_{(i+j+k_1)} - U_{(i+j)})\{1 + o_p(1)\} \\
&- r^{(1)}(U_{(i-j)}) - \frac{1}{2} r^{(2)}(U_{(i-j)})(U_{(i-j-k_1)} - U_{(i-j)})\{1 + o_p(1)\} \\
&= \frac{1}{2} r^{(2)}(U_{(i)})(U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)})\{1 + o_p(1)\}
\end{aligned}$$

where C_{i,j,k_1} is chosen such that each individual quotient $\frac{{}^{+}\hat{Y}_{i+j}^{(1)} - {}^{-}\hat{Y}_{i-j}^{(1)}}{C_{i,j,k_1}}$, $j = 1, \ldots, k_2$ is an asymptotic unbiased estimator of the second order derivative $r^{(2)}(U_{(i)})$.

The exact weight is selected to be proportional to the inverse of the conditional variance of each quotient $\frac{{}^{+}\hat{Y}_{i+j}^{(1)} - {}^{-}\hat{Y}_{i-j}^{(1)}}{C_{i,j,k_1}}$

$$\tilde{w}_{i,j,2} = \frac{1/\,\mathbf{Var}\left[\frac{\left(\frac{Y_{i+j+k_1} - Y_{i+j}}{U_{(i+j+k_1)} - U_{(i+j)}} - \frac{Y_{i-j-k_1} - Y_{i-j}}{U_{(i-j-k_1)} - U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}} |\widetilde{\mathbb{U}}\right]}{\sum_{j=1}^{k_2} 1/\,\mathbf{Var}\left[\frac{\left(\frac{Y_{i+j+k_1} - Y_{i+j}}{U_{(i+j+k_1)} - U_{(i+j)}} - \frac{Y_{i-j-k_1} - Y_{i-j}}{U_{(i-j-k_1)} - U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}} |\widetilde{\mathbb{U}}\right]}$$

By Lemma 1, the leading order of the weight $\tilde{w}_{i,j,2}$ is

$$w_{i,j,2} = \frac{(2j + k_1)^2}{\sum_{j=1}^{k_2} (2j + k_1)^2} \tag{17}$$

such that $\tilde{w}_{i,j,2} = w_{i,j,2}\{1 + o_p(1)\}$ for $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$. Similar to the first order noisy derivative, boundary issues arise in (16) when $i < \sum_{j=1}^{2} k_j + 1$ or $i > n - \sum_{j=1}^{2} k_j$. Theorem 3 states the asymptotic conditional bias and variance of (16) using the weights (17). It is difficult to get the exact asymptotic expression for the conditional bias and variance of the noisy second order derivative estimator. Therefore, we provide a suitable upperbound.

Theorem 3 *Under model* (6) *and assume r is three times continuously differentiable on* $[0,1]$, *$k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$. Then, for the weights* (17), *the conditional (absolute) bias and the conditional variance of* (16) *are bounded above*

$$\left|\text{bias}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]\right| \le \frac{\sup_{u \in [0,1]} |r^{(3)}(u)|}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3} k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3} k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \{1 + o_p(1)\}$$

and

$$\mathbf{Var}[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}] \leq \frac{4(n+1)^4\sigma_e^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1+o_p(1)\}$$

uniformly for $\sum_{j=1}^{2} k_j + 1 \leq i \leq n - \sum_{j=1}^{2} k_j$.

Proof: see Appendix J.

From Theorem 3 the pointwise consistency easily follows

Corollary 4 *Under the assumptions of Theorem 3 and for the weight sequence defined in* (17)*,* $k_1 \to \infty$ *and* $k_2 \to \infty$ *as* $n \to \infty$ *such that* $n^{-1}k_1 \to 0$*,* $n^{-1}k_2 \to 0$*,* $n^4k_1^{-2}k_2^{-3} \to 0$ *and* $n^4k_1^{-4}k_2^{-1} \to 0$*, it follows for any* $\epsilon > 0$

$$\mathbf{P}\left[\left|\hat{Y}_i^{(2)} - r^{(2)}(U_{(i)})\right| > \epsilon\right] \to 0.$$

Proof: see Appendix K.

Assuming the order of k_1 is the same as the order of k_2, then according to Theorem 3 and Corollary 4, the conditional bias and conditional variance of (16) tends to zero as $k_1 \to \infty$ and $k_2 \to \infty$ asymptotically faster than $n^{4/5}$ but slower than n. It is easy to show that the fastest possible rate at which $\mathbf{E}[(\hat{Y}_i^{(2)} - r^{(2)}(U_i))^2| \mid \widetilde{\mathbb{U}}] \to 0$ (L_2 rate of convergence) is $O_p(n^{-2/7})$ and the fastest rate is attained for $k_1 = O(n^{6/7})$ and $k_2 = O(n^{6/7})$. Using Jensen's inequality, similar results can be shown for the L_1 rate of convergence, i.e.

$$\mathbf{E}\left[\left|\hat{Y}_i^{(2)} - r^{(2)}(U_{(i)})\right| \mid \widetilde{\mathbb{U}}\right] \leq \left|\mathrm{bias}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]\right| + \sqrt{\mathbf{Var}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]} = O_p(n^{-1/7}).$$

3.2. Optimal Tuning parameter selection for k_1 and k_2

As for the first order derivative with one tuning parameter, the second order derivative has two, which control the bias-variance trade-off. Based on the asymptotic upperbounds of the bias and variance in Theorem 3, we choose k_1 and k_2 as follows.

Corollary 5 *Under the assumptions of Theorem 3 and denote* $\mathcal{B}_2 = \sup_{u\in[0,1]} |r^{(3)}(u)|$*, then* k_1 *and* k_2 *that minimize the asymptotic upper bound of the conditional MISE are*

$$\begin{aligned}(k_1,k_2)_{\mathrm{opt}} &= \underset{k_1,k_2\in\mathbb{N}^+\setminus\{0\}}{\arg\min}\left\{\left(\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2\right. \\ &+ \left.\frac{4(n+1)^4\sigma_e^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\right\}\end{aligned}$$

Proof: see Appendix L. ∎

The second unknown quantity $\mathcal{B}_2$ can be (roughly) estimated with a local polynomial regression estimator of order $p = 4$. By plugging in two estimators for σ_e^2 (Hall et al., 1990) and $\mathcal{B}_2$ in Corollary 5, the optimal value pair $(k_1, k_2)_{\mathrm{opt}}$ can be obtained using a grid search (or any other optimization method) over a Cartesian product set.

3.3. Exact Bias

Although we use the upper bound of the absolute conditional bias to tune the parameters in the estimator of second order derivatives, the exact conditional bias for noisy second order derivative estimation (16) is also important. It provides a clear comparision with the second order derivative estimator used in fixed design (Charnigo et al., 2011; De Brabanter et al., 2013) and illustrates why we can not use a similar framework in random design.

Adapting the fixed design framework from Charnigo et al. (2011) and De Brabanter et al. (2013) to random design under the standard uniform distribution for the q-th order derivative

$$\hat{Y}_i^{(q)} = \sum_{i=1}^{k_q} w_{i,j} \frac{\hat{Y}_{i+j}^{(q-1)} - \hat{Y}_{i-j}^{(q-1)}}{U_{i+j} - U_{i-j}}, \quad q = 1, 2, \ldots \tag{18}$$

where $k_1, k_2, \ldots, k_q$ are tuning parameters. Due to the asymmetry $U_{(i+j)} - U_{(i)} \neq U_{(i)} - U_{(i-j)}$ in random design, extra bias will be introduced in the first order noisy derivative estimator $\hat{Y}_i^{(1)}, i = 2, \ldots, n-1$. Using the recursive relation in (18), the extra bias will accumulate as q increases in random design. The estimator (18) is no longer a consistent estimator when $q > 2$. The exact bias of the proposed second order derivative estimator in (16) is smaller and is given by

$$\text{bias}\big[\hat{Y}_i^{(2)} | \widetilde{\mathbb{U}}\big] = O_p\left(\max\left\{\frac{k_1^{\frac{1}{2}}}{n}, \frac{k_2^{\frac{1}{2}}}{n}, \frac{k_1^2}{n^2}, \frac{k_2^2}{n^2}\right\}\right) \tag{19}$$

The proof is given in Appendix M. The boundary issue still arises for the second order derivative estimator since there are not enough k_2 pairs of ${}^{+}\hat{Y}_{i+j}^{(1)}$ and ${}^{-}\hat{Y}_{i-j}^{(1)}$ at the boundary. Similar to Section 2.3.2, at the boundary $i < 1 + k_1 + k_2$ and $i > n - k_1 - k_2$, $k_1(i)$ and $k_2(i)$ are the maximum number of available quotients in the first and second empirical derivatives.

3.4. Smoothing the noisy second order derivatives

The second order derivative estimator (16) can be written as

$$\begin{aligned} \hat{Y}_i^{(2)} &= 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{r(U_{(i+j+k_1)}) - r(U_{(i+j)})}{U_{(i+j+k_1)} - U_{(i+j)}} - \frac{r(U_{(i-j-k_1)}) - r(U_{(i-j)})}{U_{(i-j-k_1)} - U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}} \\ &+ 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{e_{i+j+k_1} - e_{i+j}}{U_{(i+j+k_1)} - U_{(i+j)}} - \frac{e_{i-j-k_1} - e_{i-j}}{U_{(i-j-k_1)} - U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}} \end{aligned} \tag{20}$$

where the second term is the new error term and is denoted as $\acute{e}_i$. It is clear that for $\acute{e}_i, i = 3, \ldots, n-2$ the i.i.d. assumption is no longer valid. Similar to Section 2.4, we apply a kernel K such that $K(0) = 0$ to remove the effects of correlation on the bandwidth selection process (De Brabanter et al., 2018).

Without loss of the generality, we show the properties of the smoothed estimator for the interior points $\sum_{j=1}^{2} k_j + 1 \leq i \leq n - \sum_{j=1}^{2} k_j$. For an arbitrary point u_0

$$\hat{r}^{(2)}(u_0) = \boldsymbol{\epsilon}_1^T \hat{\boldsymbol{\beta}} = \boldsymbol{\epsilon}_1^T \, \mathbf{S}_n^{-1} \, \mathbf{U}_{\mathrm{u}}{}^T \, \mathbf{W}_{\mathrm{u}} \, \hat{\mathbf{Y}}^{(2)} \tag{21}$$

where $\boldsymbol{\epsilon}_1 = (1, 0, \ldots, 0)^T$ is a unit vector with 1 in the first position. $\hat{\mathbf{Y}}^{(2)} = (\hat{Y}^{(2)}_{k_1+k_2+1}, \ldots, \hat{Y}^{(2)}_{n-k_1-k_2})$, $\mathbf{W}_{\mathrm{u}}$ is the diagonal matrix of weights, i.e. $\mathrm{diag}\{K_h(U_{(i)} - u_0)\}$ with kernel K, bandwidth h and $K_h(\cdot) = K(\cdot/h)/h$, $\mathbf{S}_n = \mathbf{U}_{\mathrm{u}}{}^T\, \mathbf{W}_{\mathrm{u}}\, \mathbf{U}_{\mathrm{u}}$, and

$$\mathbf{U}_{\mathrm{u}} = \begin{pmatrix} 1 & (U_{(k_1+k_2+1)} - u_0) & \cdots & (U_{(k_1+k_2+1)} - u_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (U_{(n-k_1-k_2)} - u_0) & \cdots & (U_{(n-k_1-k_2)} - u_0)^p \end{pmatrix}$$

with $\acute{e}_i$ in (20) satisfying $\mathbf{E}[\acute{e}_i|U] = 0$ and $\mathbf{Cov}(\acute{e}_i, \acute{e}_j|U_{(i)}, U_{(j)}) = \sigma^2_{\acute{e}}\rho'_n(U_{(i)} - U_{(j)})$ for $i \neq j$ with $\sigma^2_{\acute{e}} < \infty$ and ρ'_n is a stationary correlation function satisfying $\rho'_n(0) = 1, \rho'_n(u) = \rho'_n(-u)$ and $|\rho'_n(u)| \leq 1$ for all u. Applying Theorem 1 in De Brabanter et al. (2018) yields the following theorem.

Theorem 4 *Let $r^{(p+1)}(\cdot), p \geq 2$ be continuous in a neighbourhood of u_0. Under the Assumptions $1-5$ and $k_1 \to \infty$, $k_2 \to \infty$ as $n \to \infty$. For $\sigma^2_e < \infty$ and the weights given in* (17)*, the conditional bias and conditional variance of* (21) *for p odd are bounded above*

$$\begin{aligned}
\mathrm{bias}\big[\hat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\big] \;&\leq\; \boldsymbol{\epsilon}_1^T\, \mathbf{S}^{-1} \Bigg[\frac{\mathcal{B}_2}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j}\tilde{c}_p \\
&+\; \frac{c_p}{(p+1)!} r^{(p+3)}(u_0) h^{p+1}\Bigg]\{1 + o_p(1)\} \\
&=\; \Bigg[\left(\int K_0^*(t)dt\right) \frac{\mathcal{B}_2}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \\
&+\; \left(\int t^{p+1} K_0^*(t)dt\right) \frac{1}{(p+1)!} r^{(p+3)}(u_0) h^{p+1}\Bigg]\{1 + o_p(1)\}
\end{aligned}$$

where $\mathcal{B}_2 = \sup_{u\in[0,1]} |r^{(3)}(u)|$ and

$$\begin{aligned}
\mathbf{Var}\big[\hat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\big] \;&\leq\; \frac{4(n+1)^4\sigma_e^2}{k_1^2 \sum_{j=1}^{k_2}(2j + k_1)^2} \frac{1 + \rho'_c}{h(n - 2k_1 - 2k_2)} \boldsymbol{\epsilon}_1^T\, \mathbf{S}^{-1}\, \mathbf{S}^*\, \mathbf{S}^{-1}\, \boldsymbol{\epsilon}_1 \{1 + o_p(1)\} \\
&=\; \frac{4(n+1)^4\sigma_e^2}{k_1^2 \sum_{j=1}^{k_2}(2j + k_1)^2} \frac{1 + \rho'_c}{h(n - 2k_1 - 2k_2)} \left(\int K_0^{*2}(t)dt\right) \{1 + o_p(1)\}.
\end{aligned}$$

Proof: see Appendix N. ■

If the order of k_1 is the same as the order of k_2, the asymptotic upper bound of the conditional MISE is minimized at $h = O(n^{-\frac{2}{7p+8}})$, $k_1 = O(n^{\frac{5p+6}{7p+8}})$ and $k_2 = O(n^{\frac{5p+6}{7p+8}})$ and L_2 rates of convergence is $O_p(n^{-\frac{4p+4}{7p+8}})$. The way to select the bandwidth h is the same as for the first order smoothed derivative estimator. We use Corollary 5 to select k_1 and k_2, and then select bandwidth h by minimizing the RSS in order to avoid estimating ρ'_c. The proposed estimator with a two step parameter tuning is still asymptotic consistent. From Theorem 4, the pointwise consistency of (21) for p odd immediately follows.

Corollary 6 *Under the assumptions of Theorem 4, $h \to 0$ and $nh \to \infty$ as $n \to \infty$, $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ such that $n^{-1}k_1 \to 0$, $n^{-1}k_2 \to 0$, $n^3k_1^{-2}k_2^{-3}h^{-1} \to 0$ and $n^3k_1^{-4}k_2^{-1}h^{-1} \to 0$. Then, for $\sigma_e^2 < \infty$ and the weights given in (17), we have for any $\varepsilon > 0$*

$$\mathbf{P}(|\hat{r}^{(2)}(u_0) - r^{(2)}(u_0)| \geq \varepsilon) \to 0$$

Proof: Analogous to the proof in Appendix H. ■.

Assume k_1 and k_2 have the same order, then for $k_1 = O(n^{6/7})$ and $k_2 = O(n^{6/7})$ in Corollary 5, and $h = O(n^{-1/(2p+3)})$ for p odd from Corollary 1 in De Brabanter et al. (2018), the L_2 rates of convergence is

$$\mathbf{E}[(\hat{r}^{(2)}(u_0) - r^{(2)}(u_0))^2|\widetilde{\mathbb{U}}] = O_p(n^{-2/7})$$

which matches the convergence rate obtained by Stone (1985) for $p = 3$. Using Jensen's inequality, the L_1 rate of convergence is $O_p(n^{-1/7})$.

3.5. Generalizing Noisy Second Order Derivative to arbitrary distributions

As before, we use the Probability Integral Transform (PIT) as in (14) to transform the random variables X to U. Assume the second order derivative of $F(X)$ exists, taking the derivative on both sides of $m(X) = r(F(X))$ with respect to X

$$\frac{d^2m}{dX^2} = \frac{d}{dX}\left(\frac{dr}{dU}\frac{dU}{dX}\right) = \frac{d}{dX}\left(f(X)r^{(1)}(U)\right) = f^{(1)}(X)r^{(1)}(U) + f(X)r^{(2)}(U) \tag{22}$$

leading to $m^{(2)}(X) = f^{(1)}(X)r^{(1)}(U) + f(X)r^{(2)}(U)$, where $f^{(1)}(X) = \frac{df(X)}{dX}$. The derivative of the density can be estimated via the kernel density derivative estimator

$$\hat{f}'(x) = \frac{1}{nh^2}\sum_{i=1}^{n} L'\left(\frac{x - X_i}{h}\right)$$

assuming the kernel L satisfies the necessary differentiability conditions (e.g. Gaussian kernel) and the bandwidth $h > 0$. An automated procedure, including bandwidth selection, is available in the R package `kedd` (Arsalane, 2015).

4. Simulation Study

In Theorem 2 and Theorem 4, $\hat{r}^{(1)}(\cdot)$ and $\hat{r}^{(2)}(\cdot)$ are based on noisy derivative data for interior points. In the simulation, we include the noisy derivative data at the boundary to obtain the local polynomial regression estimator for the final smoothed derivatives.

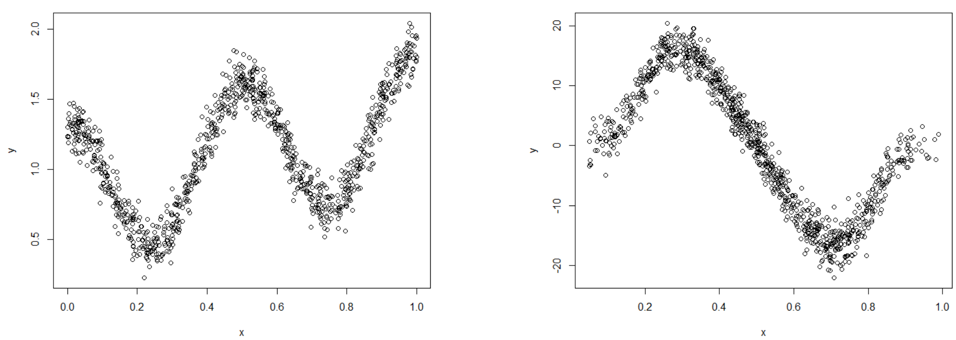
4.1. First Order Derivative Estimation

Consider the following two functions

$$m(X) = \cos^2(2\pi X) + \log(4/3 + X) \quad \text{for} \quad X \sim \mathcal{U}(0, 1) \tag{23}$$

$$m(X) = 50e^{-8(1-2X)^4}(1 - 2X) \quad \text{for} \quad X \sim \text{beta}(2, 2). \tag{24}$$

In all simulations, we estimate the density f and distribution F using kernel methods (R package `ks` (Duong, 2018)). The tuning parameter k is selected based on Corollary 5 over a positive integer set $\{1, 2, \ldots, 499\}$. We use local cubic regression ($p = 3$) with bimodal kernel to initially smooth the data. Bandwidths h were selected from the set $\{0.04, 0.045, \ldots, 0.1\}$ for both (23) and (24) and corrected for a unimodal Gaussian kernel. The sample size for both models is $n = 1000$ with $e \sim N(0, 0.1^2)$ and $e \sim N(0, 2^2)$ for (23) and (24) respectively. Figure 1 shows the raw data (X, Y) for both model (23) and model (24). Figure 2 and 3 show the first order noisy derivative (blue dots), the true first order derivative (full line) and smoothed first order derivative (dashed line) for both model (23) and model (24) separately.



(a) (b)

Figure 1: Raw data generated according to (a) model (23) and (b) model (24).

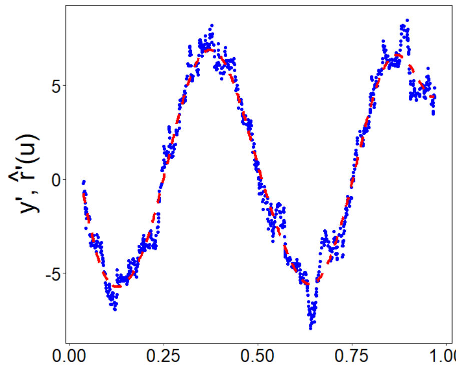
Next, we compare the proposed methodology with several popular methods for nonparametric derivative estimation, i.e. the local slope of the local polynomial regression with $p = 2$, $p = 3$ (R package `locpol` (Ojeda, 2012)) and penalized smoothing splines (R package **`pspline`** (Ramsey and Ripley, 2017)). The order of the local polynomial is recommended to be $p = 2$ since p minus the order of the derivative is odd (Fan and Gijbels, 1996). In case of penalized smoothing splines, cubic splines were used. For the Monte Carlo study, we constructed data sets of size $n = 700$ and generated the function

$$m(X) = \sqrt{X(1-X)}\sin((2.1\pi)/(X+0.05)) \quad \text{for} \quad X \sim \mathcal{U}(0.25, 1) \tag{25}$$

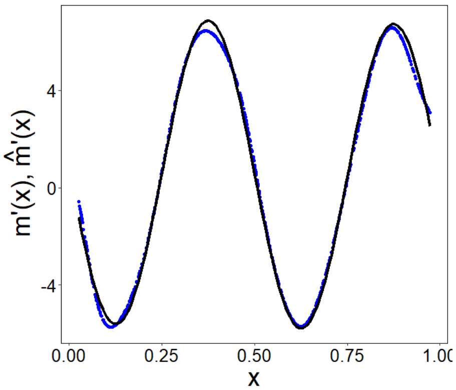
100 times according to model (2) with $e \sim N(0, 0.2^2)$. Bandwidths were selected from the set $\{0.03, 0.035, \ldots, 0.07\}$ and corrected for a unimodal Gaussian kernel. In order to remove the effect of boundary issues on the performance for all three methods, we use the adjusted mean absolute error as a performance measure defined as

$$\text{MAEadjusted} = \frac{1}{650}\sum_{i=26}^{675} |\hat{m}'_n(X_i) - m'(X_i)|.$$

Figure 4 shows the raw data in one random run in Monte Carlo study and its estimated first order derivatives using the proposed estimator, local quadratic polynomial regression

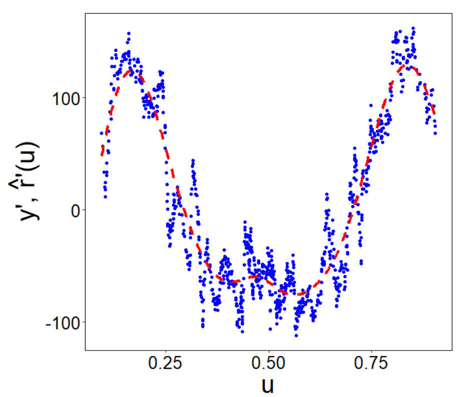


(a)

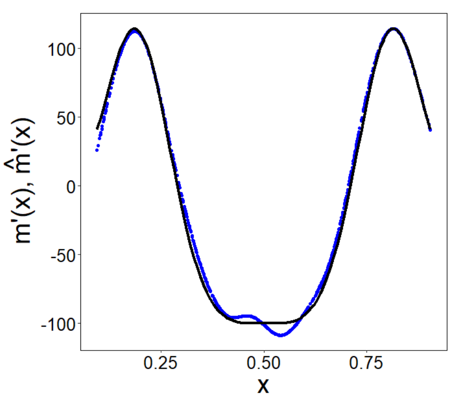


(b) Back transform according to $\widehat{m}^{(1)}(X) = \widehat{f}(X)\widehat{r}^{(1)}(U)$

Figure 2: First order noisy derivative (dots) of model (23) based on $k = 26$. Smoothed derivative based on local cubic regression (dashed line) and true derivative (full line). (a) First step of the smoothing process for arbitrary distributions using the probability integral transform; (b) True first order derivative (full line) and the proposed smoothed derivative of $m(X)$ (dashed line) in the original space. Boundary points are not shown for visual purposes.



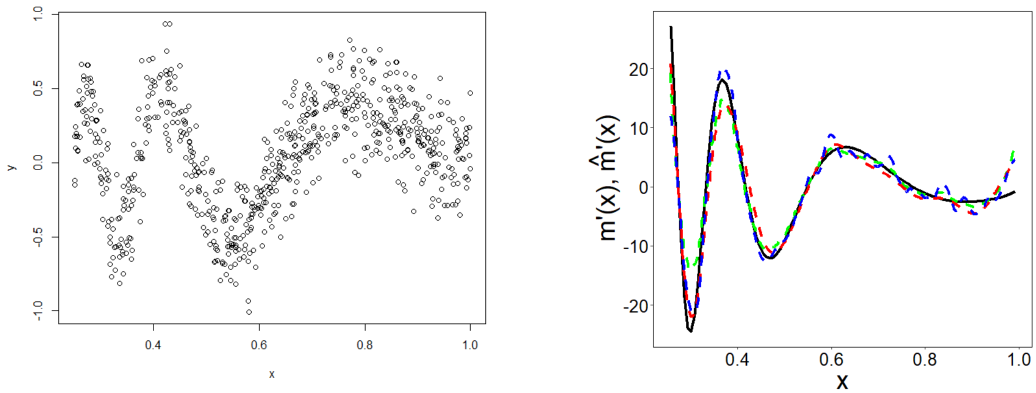
(a)



(b) Back transform according to $\widehat{m}^{(1)}(X) = \widehat{f}(X)\widehat{r}^{(1)}(U)$

Figure 3: First order noisy derivative (dots) of model (24) based on $k = 22$, smoothed derivative based on local cubic regression (dashed line) and true derivative (full line). (a) First step of the smoothing process for arbitrary distributions using the probability integral transform; (b) True first order derivative (full line) and the proposed smoothed derivative of $m(X)$ (dashed line) in the original space. Boundary points are not shown for visual purposes.

and cubic penalized smoothing splines. The variance in this model is large which increases the difficulty in estimation and the proposed estimator is slightly better than local poly-



(a) (b)

Figure 4: One random run for model (25). (a) Raw data. (b) True first order derivatives (full solid line) with estimated first order derivatives using three different method: the proposed estimator with $k = 8$ (red dash line), the local polynomial estimator with $p = 2$ (green dash line) and cubic penalized smoothing (blue dash line).

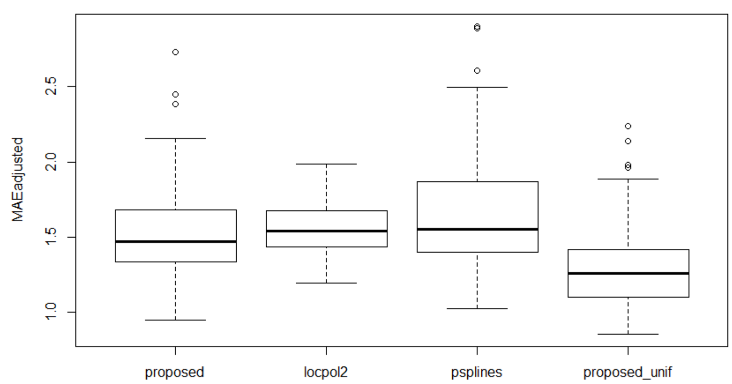


Figure 5: Result of the Monte Carlo study for the proposed methodology, local polynomial regression and penalized smoothing splines for first order derivative estimation.

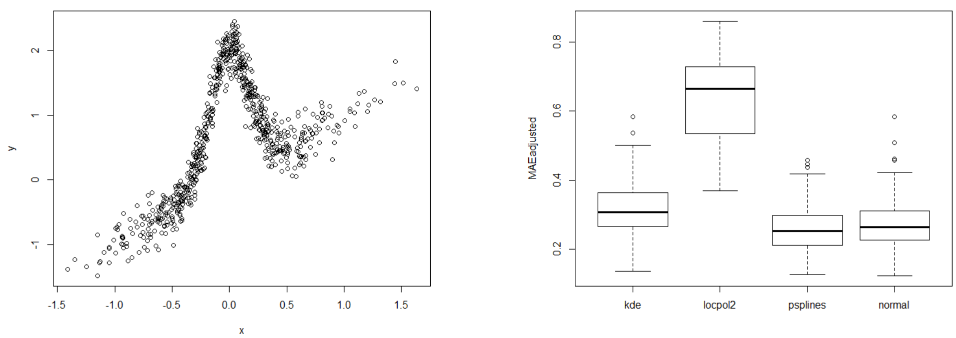
nomial estimator in peaks and dips and smoother then penalized smoothing on the right part, but overall three methods have equal performance in estimating first order derivative of this regression function.

The first three boxplots in Figure 5 represent the performance of those three methods in 100 repetitions. The proposed model has a similar performance as the local polynomial regression with $p = 2$ (locpol2) and cubic penalized smoothing splines (psplines). To illustrate the loss of accuracy due to estimation of the density f and distribution F we use the true density and distribution to compute the derivative in the fourth boxplot.

We also run a Monte Carlo simulation for a non-uniform distribution

$$m(X) = X + 2\exp(-16X^2) \quad \text{for} \quad X \sim \mathcal{N}(0, 0.5^2) \tag{26}$$

100 times according to model (2) with $e \sim N(0, 0.2^2)$. Bandwidths were selected from the set $\{0.04, 0.045, \ldots, 0.08\}$ and corrected for a unimodal Gaussian kernel. As before, we compare the proposed model with local polynomial regression with $p = 2$ (locpol2) and cubic penalized smoothing splines (psplines) in Figure 6. Proposed model and cubic penalized smoothing splines are better than local polynomial regression with $p = 2$ in this case. The cubic penalized smoothing splines is slightly better than the proposed model using kernel density estimator and have similar performance with the proposed model if using true density.



(a) (b)

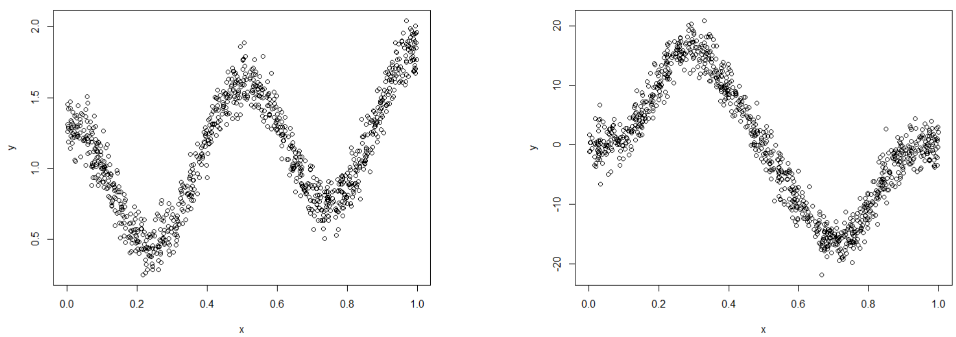
Figure 6: One random run for model (26). (a) Raw data. (b) Result of the Monte Carlo study for the proposed methodology using kernel density estimator(kde), local polynomial regression(locpol2), penalized smoothing splines(psplines) and proposed methodology with true normal density for first order derivative estimation.

4.2. Second Order Derivative Estimation

Similar to the first order derivative, the tuning parameters k_1 and k_2 could be determined by minimizing the criterion in Corollary 5 through grid search over a product set. We use local cubic regression ($p = 3$) with a kernel K such that $K(0) = 0$ to smooth the noisy second order derivatives. The bandwidth obtained with the kernel K such that $K(0) = 0$ is then corrected for a unimodal kernel. The second order derivative estimation for any distribution is given in (22). In the simulation, we only show the performance of the proposed second order derivative estimator under the assumption that $X \sim \mathcal{U}[0, 1]$. For model (24), we change the assumption on the distribution of X as follows

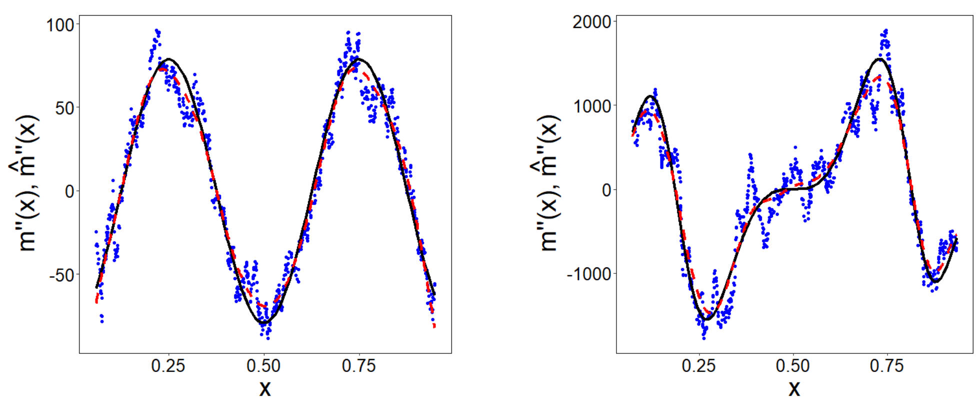
$$m(X) = 50e^{-8(1-2X)^4}(1 - 2X) \quad \text{for} \qquad X \sim \mathcal{U}(0, 1). \tag{27}$$

Figure 7 shows the raw data (X, Y) for both models (23) and (27). The sample size is taken to be $n = 1000$ for both functions with $e \sim N(0, 0.1^2)$ and $e \sim N(0, 2^2)$ for (23) and (24) respectively. In this simulation, we choose the gird search space of (k_1, k_2) to be $\{1, 2, \ldots, 100\} \otimes \{1, 2, \ldots, 100\}$ for all models. Bandwidths h are selected from the set $\{0.05, 0.055, \ldots, 0.1\}$ for both functions (23) and (27). The results for second order derivative estimation of function (23) and (27) are shown in Figure 8. For visual purposes the boundary points have been removed. Figure 8 shows the second order noisy derivative (blue dots), the true first order derivative (full line) and smoothed first order derivative (red dashed line) for both models (23) and (27) respectively.



(a) model (23) (b) model (27)

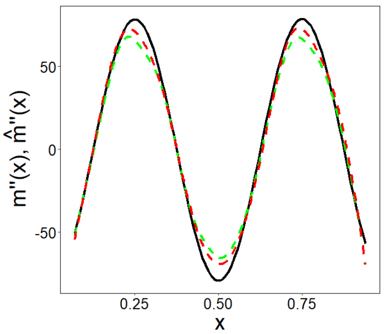
Figure 7: Raw Data for both models



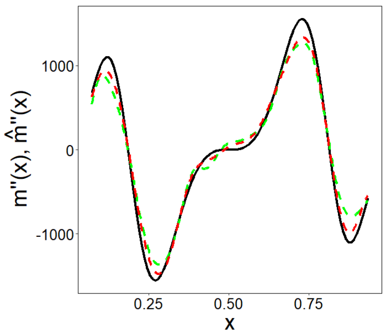
(a) model (23) (b) model (27)

Figure 8: Second order derivatives smoothed by $p = 3$ local polynomial regression using a kernel K such that $K(0) = 0$ (red dashed line) on the noisy second order derivative data(blue dots) and true derivative function (full line). (a) Second order derivative of model (23), with $k_1 = 42$ and $k_2 = 23$; (b) Second order derivative of model (27) with $k_1 = 44$ and $k_2 = 24$. Boundary points are not shown for visual purposes.

To compare the proposed smoothed second order derivative estimator with the cubic local polynomial estimator, we show both estimators for model (23) and model (27) in



(a)



(b)

Figure 9: Second order derivatives smoothed by $p = 3$ local polynomial regression using a kernel K such that $K(0) = 0$ (red dashed line) on the noisy second order derivative data, the local polynomial estimator with $p = 3$ (green dash line) and true derivative function (full line). (a) Second order derivative of model (23) with $k_1 = 42$ and $k_2 = 23$ (b) Second order derivative of model (27) with $k_1 = 44$ and $k_2 = 24$. Boundary points are not shown for visual purposes.

Figure 9. It is clear that the proposed second order derivative estimator slightly outperforms the local polynomial ($p = 3$) estimates. For the Monte Carlo study, we construct data sets of size $n = 700$ for the function

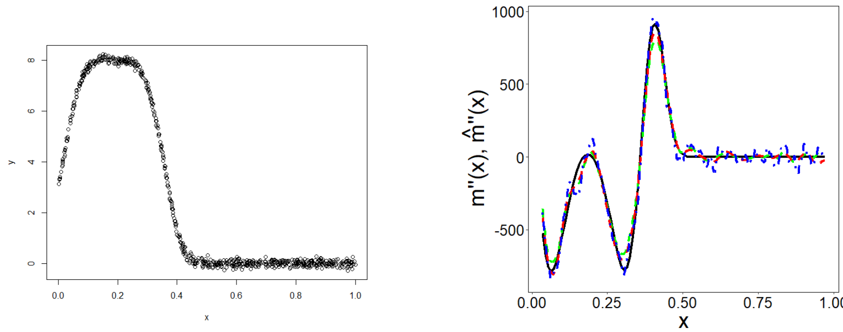
$$m(x) = 8e^{-(1-5x)^3(1-7x)} \quad \text{for} \quad X \sim \mathcal{U}(0,1) \tag{28}$$

100 times according to model (2) with $e \sim N(0, 0.1^2)$. As a measure of performance, we define the adjusted mean absolute error as

$$\text{MAEadjusted} = \frac{1}{640} \sum_{i=31}^{670} |\hat{m}_n^{(2)}(X_i) - m^{(2)}(X_i)|$$

to ignore the boundary effects in the simulation result. Bandwidths are selected from interval $\{0.03, 0.035, \ldots, 0.1\}$.

Similar to the first order derivative, we compare the proposed methodology with local polynomial regression (R package `locpol` (Ojeda, 2012)) and penalized smoothing splines (R package `stat` (Ramsey and Ripley, 2017)). The order of the local polynomial was taken to be $p = 3$ since p minus the order of the derivative is odd (Fan and Gijbels, 1996). In case of penalized smoothing splines, cubic splines were used. Figure 10 shows the raw data in one random run in Monte Carlo study and its estimated second order derivatives using three methods. Those three methods have equal performance in estimating second order derivative for raw data in Figure 10. The results for Monte Carlo study is shown in Figure 11. The proposed estimator has a slightly better performance compared to local cubic polynomial estimates and penalized cubic splines.



(a) (b)

Figure 10: One random run for model (28). (a) Raw data. (b) True second order derivatives (full solid line) with estimated second order derivatives using three different method: the proposed estimator with $k_1 = 15$, $k_2 = 8$ (red dash line), the local polynomial estimator with $p = 3$ (green dash line) and cubic penalized smoothing (blue dash line).

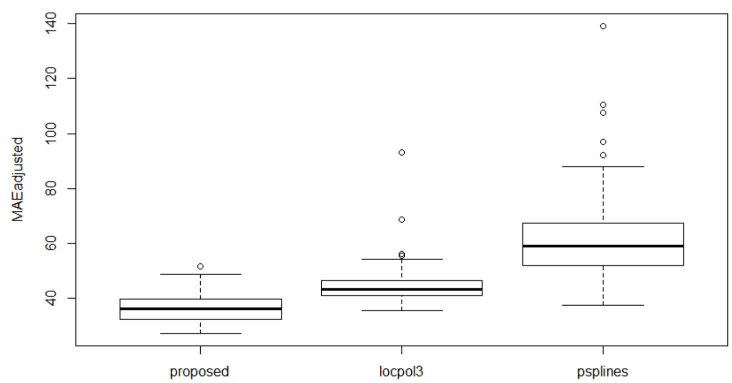
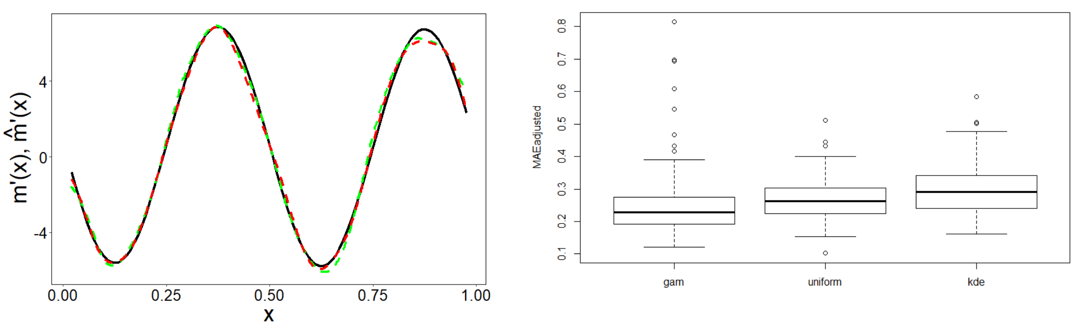


Figure 11: Result of the Monte Carlo study for the proposed methodology, local polynomial regression and penalized smoothing splines for second order derivative estimation.

One anonymous referee suggested to smooth the data first by means of adaptive splines followed by taking discrete derivatives i.e., using first or second order differencing for the first or second order derivatives respectively. The main idea for the first order derivative estimator is given by

$$\frac{\widehat{m}(X_i) - \widehat{m}(X_{i-1})}{X_i - X_{i-1}} \approx \widehat{m}'(\xi_i) \tag{29}$$

with $\xi_i \in [X_{i-1}, X_i]$ and $\widehat{m}$ the adaptive spline estimator (denoted as gam). This approach can have promising results (see Figure 12), but it does not immediately allow to evaluate the derivatives in an arbitrary point. We conducted a Monte Carlo simulation for model 23 with global k. In Figure 12(b), the median of (29) is slightly lower than the proposed model (kde) and the one assuming the true underlying distribution is known (uniform). Further, the mean for the three methods is 0.267 (gam), 0.268 (uniform) and 0.302 (kde) and the variances are 0.016, 0.004, 0.006 respectively. Based on this simulation we can state that the adaptive spline estimator and the proposed method have a similar performance on model (23).



(a) (b)

Figure 12: One random run for model (22). (a) True first order derivatives (full) with estimated first order derivatives using an adaptive spline estimator in (29) (green dashed) and the proposed estimator (red dashed); (b) Monte Carlo result based on 100 runs for (29) (gam), assuming the true underlying distribution is known (uniform) and the proposed method (kde).

5. Conclusions

We proposed a method for derivative estimation in random design and discussed the asymptotic properties of the proposed estimators. The proposed methodology estimates derivatives nonparametrically without having to estimate the regression function. Asymptotic bias and variance are derived, L_1 and L_2 rates of convergence are established. Our analysis showed that estimating higher order derivatives becomes increasingly more difficult and slower rates of convergence are to be expected. Further, we provide a rule-of-thumb to choose the parameter(s) for the first and second order noisy derivatives. Finally, since the independence assumption of the newly created data set does no longer hold, we use a simple but effective smoothing methodology based on kernels K such that $K(0) = 0$ combined with the flexibility of local polynomial regression. Additionally, we discussed the property of the smoothed noisy derivative estimates.

One drawback of the proposed framework is that the proposed first and second order derivative estimator requires the estimation of the density f and distribution F. A first topic of further research interest is to adapt the proposed framework directly for arbitrary distributions without transformation. Second, finding an efficient way to tune h and k simultaneously would greatly benefit the rate of convergence of the proposed methodology. A potential lead could be found in the use of empirical semi-variograms.

Appendix A. Proof of Lemma 1

Following David and Nagaraja (1970, p. 14) we have

$$U_{(i+j)} - U_{(i-j)} \sim \text{Beta}(2j, n+1-2j).$$

It immediately follows that

$$\begin{aligned} U_{(i+j)} - U_{(i-j)} &= \mathbf{E}\{U_{(i+j)} - U_{(i-j)}\} + O_p\left[\sqrt{\mathbf{Var}\{U_{(i+j)} - U_{(i-j)}\}}\right] \\ &= \frac{2j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right) \end{aligned}$$

Similarly, according to the property of uniform order statistics we have

$$U_{(i+j)} - U_{(i)} \sim \text{Beta}(j, n+1-j)$$

and

$$\begin{aligned} U_{(i+j)} - U_{(i)} &= \mathbf{E}\{U_{(i+j)} - U_{(i)}\} + O_p\left[\sqrt{\mathbf{Var}\{U_{(i+j)} - U_{(i)}\}}\right] \\ &= \frac{j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right). \end{aligned}$$

The proof of the third part of the lemma is analogous to the proof above and is therefore omitted.

Appendix B. Proof of Proposition 1

$$\begin{aligned} \mathbf{Var}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= \mathbf{Var}\left[\sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right)|\mathbb{U}\right] \\ &= \left(1 - \sum_{j=2}^{k} w_{i,j}\right)^2 \mathbf{Var}\left[\frac{Y_{i+1} - Y_{i-1}}{U_{(i+1)} - U_{(i-1)}}|\mathbb{U}\right] + \sum_{j=2}^{k} w_{i,j}^2\, \mathbf{Var}\left[\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}|\mathbb{U}\right] \\ &= \left(1 - \sum_{j=2}^{k} w_{i,j}\right)^2 \frac{2\sigma_e^2}{(U_{(i+1)} - U_{(i-1)})^2} + \frac{2\sigma_e^2}{(U_{(i+j)} - U_{(i-j)})^2} \sum_{j=2}^{k} w_{i,j}^2. \end{aligned}$$

Setting the partial derivatives to zero yields

$$w_{i,j} = w_{i,1} \frac{(U_{(i+j)} - U_{(i-j)})^2}{(U_{(i+1)} - U_{(i-1)})^2}.$$

Using the fact that $\sum_{j=1}^{k} w_{i,j} = 1$ results in

$$\sum_{j=1}^{k} w_{i,j} = w_{i,1} \sum_{j=1}^{k} \frac{(U_{(i+j)} - U_{(i-j)})^2}{(U_{(i+1)} - U_{(i-1)})^2} = 1.$$

Consequently, this gives

$$w_{i,j}\frac{(U_{(i+1)}-U_{(i-1)})^2}{(U_{(i+j)}-U_{(i-j)})^2}\sum_{j=1}^{k}\frac{(U_{(i+j)}-U_{(i-j)})^2}{(U_{(i+1)}-U_{(i-1)})^2}=1$$

proving the proposition.

Appendix C. Proof of Theorem 1

Since r is twice continuously differentiable on $[0,1]$, the following Taylor expansions are valid for $r(U_{(i+j)})$ and $r(U_{(i-j)})$ in a neighborhood of $U_{(i)}$:

$$r(U_{(i+j)})=r(U_{(i)})+(U_{(i+j)}-U_{(i)})r'(U_{(i)})+\frac{(U_{(i+j)}-U_{(i)})^2}{2}r^{(2)}(\zeta_{i,i+j})$$

and

$$r(U_{(i-j)})=r(U_{(i)})+(U_{(i-j)}-U_{(i)})r'(U_{(i)})+\frac{(U_{(i-j)}-U_{(i)})^2}{2}r^{(2)}(\zeta_{i-j,i}),$$

where $\zeta_{i,i+j}\in]U_{(i)},U_{(i+j)}[$ and $\zeta_{i-j,i}\in]U_{(i-j)},U_{(i)}[$. Using Lemma 1 and Proposition 1, the absolute conditional bias is bounded above by

$$\begin{aligned}
\left|\text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]\right| &= \left|\mathbf{E}\left[\sum_{j=1}^{k}w_{i,j}\cdot\left(\frac{Y_{i+j}-Y_{i-j}}{U_{(i+j)}-U_{(i-j)}}\right)|\mathbb{U}\right]-r'(U_i)\right|\\
&= \frac{1}{2}\left|\sum_{j=1}^{k}w_{i,j}\frac{(U_{(i+j)}-U_{(i)})^2r^{(2)}(\zeta_{i,i+j})-(U_{(i-j)}-U_{(i)})^2r^{(2)}(\zeta_{i-j,i})}{U_{(i+j)}-U_{(i-j)}}\right|\\
&= \frac{1}{2}\left|\frac{\sum_{j=1}^{k}(U_{(i+j)}-U_{(i-j)})\big\{(U_{(i+j)}-U_{(i)})^2r^{(2)}(\zeta_{i,i+j})-(U_{(i-j)}-U_{(i)})^2r^{(2)}(\zeta_{i-j,i})\big\}}{\sum_{l=1}^{k}(U_{(i+l)}-U_{(i-l)})^2}\right|\\
&\leq \frac{1}{2}\sup_{u\in[0,1]}|r^{(2)}(u)|\frac{\sum_{j=1}^{k}(U_{(i+j)}-U_{(i-j)})\big\{(U_{(i+j)}-U_{(i)})^2+(U_{(i-j)}-U_{(i)})^2\big\}}{\sum_{l=1}^{k}(U_{(i+l)}-U_{(i-l)})^2}\\
&= \frac{1}{2}\sup_{u\in[0,1]}|r^{(2)}(u)|\frac{\frac{k^2(k+1)^2}{(n+1)^3}\big\{1+O_p(\frac{1}{\sqrt{k}})\big\}}{\frac{2k(k+1)(2k+1)}{3(n+1)^2}\big\{1+O_p(\frac{1}{\sqrt{k}})\big\}}\\
&= \sup_{u\in[0,1]}|r^{(2)}(u)|\frac{3k(k+1)}{4(n+1)(2k+1)}\left\{1+O_p\left(\frac{1}{\sqrt{k}}\right)\right\}.
\end{aligned}$$

Then for $k\to\infty$ as $n\to\infty$

$$\left|\text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]\right|\leq\sup_{u\in[0,1]}|r^{(2)}(u)|\frac{3k(k+1)}{4(n+1)(2k+1)}\left\{1+o_p(1)\right\}$$

Using Proposition 1, the conditional variance yields

$$\begin{aligned}
\mathbf{Var}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= \mathbf{Var}\left[\sum_{j=1}^{k} w_{i,j}\cdot\left(\frac{Y_{i+j}-Y_{i-j}}{U_{(i+j)}-U_{(i-j)}}\right)|\mathbb{U}\right] \\
&= 2\sigma_e^2\frac{\sum_{j=1}^{k}(U_{(i+j)}-U_{(i-j)})^2}{\left(\sum_{l=1}^{k}(U_{(i+l)}-U_{(i-l)})^2\right)^2} \\
&= 2\sigma_e^2\frac{1}{\sum_{l=1}^{k}\left(U_{(i+l)}-U_{(i-l)}\right)^2} \\
&= 2\sigma_e^2\frac{1}{\frac{2k(k+1)(2k+1)}{3(n+1)^2}\left\{1+o_p(1)\right\}} \\
&= \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}\left\{1+o_p(1)\right\},
\end{aligned}$$

provided that $k\to\infty$ as $n\to\infty$. Both results hold uniformly for $k+1\le i\le n-k$.

Appendix D. Proof of Corollary 1

Under the conditions $k\to\infty$ as $n\to\infty$ such that $n^{-1}k\to 0$ and $n^2k^{-3}\to 0$, Theorem 1 states that the upperbound of conditional bias and conditional variance go to zero. Consequently, we have that

$$\lim_{n\to\infty}\mathrm{MSE}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]=0.$$

According to Chebyshev's inequality the proof is complete.

Appendix E. Proof of Corollary 2

From the bias-variance decomposition of the mean squared error (MSE), it follows that

$$\mathrm{MSE}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big]\le\mathcal{B}^2\frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2}+\frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}+o_p(n^{-2}k^2+n^2k^{-3}),$$

with $\mathcal{B}=\sup_{u\in[0,1]}|r^{(2)}(u)|$. Since $U\sim\mathcal{U}(0,1)$, the conditional mean integrated squared error (MISE) which measures the average global error is:

$$\begin{aligned}
\mathrm{MISE}\big[\hat{Y}^{(1)}|\mathbb{U}\big] &= \mathbf{E}\int_0^1\big[\hat{Y}^{(1)}(U)-r^{(1)}(U)|\mathbb{U}\big]^2dU \\
&= \int_0^1\mathbf{E}\big[\hat{Y}^{(1)}(U)-r^{(1)}(U)|\mathbb{U}\big]^2dU \\
&\le B^2\frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2}+\frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}+o_p(n^{-2}k^2+n^2k^{-3})
\end{aligned}$$

with $\hat{Y}^{(1)}(U)$ represents the first order derivative estimator at design point U. Denote the asymptotic conditional MISE (AMISE) by

$$\mathrm{AMISE}\big[\hat{Y}^{(1)}|\mathbb{U}\big]\le\mathcal{B}^2\frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2}+\frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}.$$

Appendix F. Proof of Exact Bias

Assume the $q+1$ derivatives of r exist on $[0,1]$, according to lemma 1, the following Taylor expansions are valid for $r(U_{(i+j)})$ and $r(U_{(i-j)})$ in a neighborhood of $U_{(i)}$

$$\begin{aligned} r(U_{(i+j)}) &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i+j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p(U_{(i+j)} - U_{(i)})^{q+1} \\ &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i+j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p\{(j/n)^{q+1}\} \end{aligned}$$

and

$$\begin{aligned} r(U_{(i-j)}) &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i-j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p(U_{(i-j)} - U_{(i)})^{q+1} \\ &= r(U_{(i)}) + \sum_{l=1}^{q} \frac{1}{l!}(U_{(i-j)} - U_{(i)})^l r^{(l)}(U_{(i)}) + O_p\{(j/n)^{q+1}\}. \end{aligned}$$

Taking expectations and for $\sum_{j=1}^{k} w_{i,j} = 1$

$$\begin{aligned} \mathbf{E}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= \sum_{j=1}^{k} w_{i,j} \frac{r(U_{(i+j)}) - r(U_{(i-j)})}{U_{(i+j)} - U_{(i-j)}} \\ &= \frac{\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})\Big[\sum_{l=1}^{q} \frac{r^{(l)}(U_{(i)})}{l!}\big\{(U_{(i+j)} - U_{(i)})^l - (U_{(i-j)} - U_{(i)})^l\big\} + O_p\{(j/n)^{q+1}\}\Big]}{\sum_{p=1}^{k}(U_{(i+p)} - U_{(i-p)})^2} \end{aligned}$$

For $q = 1$, the second order derivative of r exists on $[0,1]$

$$\begin{aligned} \text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= \frac{r^{(1)}(U_{(i)})\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})^2 + O_p\left(k^4/n^3\right)}{\sum_{p=1}^{k}(U_{(i+p)} - U_{(i-p)})^2} - r^{(1)}(U_{(i)}) \\ &= O_p\left(\frac{k}{n}\right) \end{aligned}$$

For $q = 2$, the third order derivative of r exists on $[0,1]$

$$\begin{aligned} \text{bias}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= \frac{r^{(2)}(U_{(i)})\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})\big\{(U_{(i+j)} - U_{(i)})^2 - (U_{(i-j)} - U_{(i)})^2\big\} + O_p\left(k^5/n^4\right)}{2\sum_{p=1}^{k}(U_{(i+p)} - U_{(i-p)})^2} \\ &= \frac{O_p(k^{\frac{7}{2}}/n^3) + O_p\left(k^5/n^4\right)}{O_p(k^3/n^2)} \\ &= O_p\left(\max\left\{\frac{k^{\frac{1}{2}}}{n}, \frac{k^2}{n^2}\right\}\right) \end{aligned}$$

The bias can be split into two terms, $\text{bias}_{\text{even}} = O_p\left(\frac{k^{\frac{1}{2}}}{n}\right)$ and $\text{bias}_{\text{odd}} = O_p\left(\frac{k^2}{n^2}\right)$. $\text{bias}_{\text{even}}$ indicates the bias from the even order terms in the Taylor expansion of $r(U_{(i\pm j)})$ and bias_{odd} for the odd order terms respectively.
For $q > 2$, we have

$$\begin{aligned}\mathbf{E}\big[\hat{Y}_i^{(1)}|\mathbb{U}\big] &= r^{(1)}(U_{(i)}) \\ &+ \frac{\sum_{j=1}^{k}(U_{(i+j)}-U_{(i-j)})\Big[\sum_{l=2}^{q}\frac{r^{(l)}(U_{(i)})}{l!}\big\{(U_{(i+j)}-U_{(i)})^l-(U_{(i-j)}-U_{(i)})^l\big\}+O_p\{(j/n)^{q+1}\}\Big]}{\sum_{p=1}^{k}(U_{(i+p)}-U_{(i-p)})^2}.\end{aligned}$$

Splitting the second term in $\text{bias}_{\text{even}}$ and $\text{bias}_{\text{even}}$ yields

$$\begin{aligned}\text{bias}_{\text{odd}}[\hat{Y}_i^{(1)}|\mathbb{U}] &\stackrel{\text{def}}{=} \frac{\sum_{j=1}^{k}(U_{(i+j)}-U_{(i-j)})\left[\sum_{l=3,5,\ldots}^{2\lceil q/2\rceil-1}\frac{r^{(l)}(U_{(i)})}{l!}\left((U_{(i+j)}-U_{(i)})^l-(U_{(i-j)}-U_{(i)})^l\right)\right]}{\sum_{p=1}^{k}(U_{(i+p)}-U_{(i-p)})^2} \\ &= \frac{\sum_{j=1}^{k}(U_{(i+j)}-U_{(i-j)})\left[\frac{r^{(3)}(U_i)}{6}\left((U_{(i+j)}-U_{(i)})^3-(U_{(i-j)}-U_{(i)})^3\right)\right]}{\sum_{p=1}^{k}(U_{(i+p)}-U_{(i-p)})^2}\big\{1+o_p(1)\big\} \\ &= \left(\frac{r^{(3)}(U_{(i)})(3k^2+3k-1)}{30(n+1)^2}+O_p\left(\frac{k^{\frac{3}{2}}}{n^2}\right)\right)\big\{1+o_p(1)\big\} \\ &= O_p\left\{\frac{k^2}{n^2}\right\}\end{aligned}$$

$$\begin{aligned}\text{bias}_{\text{even}}[\hat{Y}_i^{(1)}|\mathbb{U}] &\stackrel{\text{def}}{=} \frac{\sum_{j=1}^{k}(U_{(i+j)}-U_{(i-j)})\left[\sum_{l=2,4,\ldots}^{2\lfloor q/2\rfloor}\frac{r^{(l)}(U_{(i)})}{l!}\left((U_{(i+j)}-U_{(i)})^l-(U_{(i-j)}-U_{(i)})^l\right)\right]}{\sum_{p=1}^{k}(U_{(i+p)}-U_{(i-p)})^2} \\ &= \frac{\sum_{j=1}^{k}(U_{(i+j)}-U_{(i-j)})\left[\frac{r^{(2)}(U_{(i)})}{2}\left((U_{(i+j)}-U_{(i)})^2-(U_{(i-j)}-U_{(i)})^2\right)\right]}{\sum_{p=1}^{k}(U_{(i+p)}-U_{(i-p)})^2}\big\{1+o_p(1)\big\} \\ &= O_p\left\{\frac{k^{\frac{1}{2}}}{n}\right\}\end{aligned}$$

Appendix G. Bias and Variance at the Left Boundary

Assume that r is three times continuously differentiable on $[0,1]$. At the left boundary $i < k+1$, we have

$$
\begin{aligned}
\text{bias}[\hat{Y}_i^{(1)}|\mathbb{U}] &= \sum_{j=1}^{k(i)} w_{i,j} \cdot \left(\frac{\frac{1}{2}\left[(U_{(i+j)} - U_{(i)})^2 - \frac{1}{2}(U_{(i-j)} - U_{(i)})^2\right] r^{(2)}(U_{(i)})}{U_{(i+j)} - U_{(i-j)}} \right) \\
&+ \sum_{j=1}^{k(i)} w_{i,j} \cdot \left(\frac{O_p\big(j^3/n^3\big)}{U_{(i+j)} - U_{(i-j)}} \right) \\
&+ \sum_{j=k(i)+1}^{k} w_{i,j} \cdot \left(\frac{1}{2}(U_{(i+j)} - U_{(i)}) r^{(2)}(U_{(i)}) \right) \{1 + o_p(1)\} \\
&= O_p\left\{ \max\left(\frac{k(i)^{7/2}}{k^3 n}, \frac{k(i)^5}{k^3 n^2}, \frac{k - k(i)}{n} \right) \right\}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{Var}[\hat{Y}_i^{(1)}|\mathbb{U}] &= \mathbf{Var}\left[\sum_{j=1}^{k(i)} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right) |\mathbb{U} \right] + \mathbf{Var}\left[\sum_{j=k(i)+1}^{k} w_{i,j} \left(\frac{Y_{i+j} - Y_i}{U_{(i+j)} - U_{(i)}} \right) |\mathbb{U} \right] \\
&= 2\sigma_e^2 \sum_{j=1}^{k(i)} \left(\frac{w_{i,j}}{U_{(i+j)} - U_{(i-j)}} \right)^2 + \sigma_e^2 \sum_{j=k(i)+1}^{k} \left(\frac{w_{i,j}}{U_{(i+j)} - U_{(i)}} \right)^2 \\
&+ \sigma_e^2 \left[\sum_{j=k(i)+1}^{k} \left(\frac{w_{i,j}}{U_{(i+j)} - U_{(i)}} \right) \right]^2 \\
&= O_p\left\{ \max\left(\frac{n^2}{k^3}, \frac{n^2(k - k(i))^2}{k^4} \right) \right\}.
\end{aligned}
$$

Appendix H. Proof of Theorem 2

Part I (conditional bias)

$$
\begin{aligned}
\text{bias}\big[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\big] &= \mathbf{E}\big[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\big] - r^{(1)}(u_0) \\
&= \boldsymbol{\epsilon}_1^T \, \mathbf{S}_{n-2k}^{-1} \, \mathbf{U}_{\mathrm{u}}{}^T \, \mathbf{W}_{\mathrm{u}} \, \mathbf{E}[\hat{\mathbf{Y}}^{(1)}|\widetilde{\mathbb{U}}] - r^{(1)}(u_0) \\
&= \boldsymbol{\epsilon}_1^T \, \mathbf{S}_{n-2k}^{-1} \, \mathbf{U}_{\mathrm{u}}{}^T \, \mathbf{W}_{\mathrm{u}} \left(\begin{bmatrix} r^{(1)}(U_{(k+1)}) \\ \vdots \\ r^{(1)}(U_{(n-k)}) \end{bmatrix} + \begin{bmatrix} \text{bias}[\hat{Y}_{k+1}^{(1)}|\mathbb{U}] \\ \vdots \\ \text{bias}[\hat{Y}_{n-k}^{(1)}|\mathbb{U}] \end{bmatrix} \right) - r^{(1)}(u_0)
\end{aligned}
$$

For p odd, (see Theorem 3.1 in Fan and Gijbels (1996)), the first term is

$$\begin{aligned}\boldsymbol{\epsilon}_1^T \,\mathbf{S}_{n-2k}^{-1}\,\mathbf{U}_{\mathrm{u}}{}^T\,\mathbf{W}_{\mathrm{u}}\begin{bmatrix} r^{(1)}(U_{(k+1)}) \\ \vdots \\ r^{(1)}(U_{(n-k)})\end{bmatrix} - r^{(1)}(u_0) &= \boldsymbol{\epsilon}_1^T\,\mathbf{S}^{-1}\,c_p\beta_{p+1}h^{p+1} + o_p(h^{p+1}) \\ &= \boldsymbol{\epsilon}_1^T\,\mathbf{S}^{-1}\,\frac{c_p}{(p+1)!}r^{(p+2)}(u_0)h^{p+1} + o_p(h^{p+1})\end{aligned} \tag{30}$$

where $c_p = (\mu_{p+1}, \ldots, \mu_{2p+1})^T$ with $\mu_j = \int u^j K(u)du$, and $\mathbf{S} = (\mu_{i+j})_{0\le i,j\le p}$.
Based on Theorem 1, for $k \to \infty$ as $n \to \infty$ the second term is

$$\boldsymbol{\epsilon}_1^T \,\mathbf{S}_{n-2k}^{-1}\,\mathbf{U}_{\mathrm{u}}{}^T\,\mathbf{W}_{\mathrm{u}}\begin{bmatrix} \mathrm{bias}[\hat{Y}_{k+1}^{(1)}|\mathbb{U}] \\ \vdots \\ \mathrm{bias}[\hat{Y}_{n-k}^{(1)}|\mathbb{U}]\end{bmatrix} \le \boldsymbol{\epsilon}_1^T \,\mathbf{S}_{n-2k}^{-1}\,\mathbf{U}_{\mathrm{u}}{}^T\,\mathbf{W}_{\mathrm{u}}\begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix}\sup_{u\in[0,1]}|r^{(2)}(u)|\frac{3k(k+1)}{4(n+1)(2k+1)}\{1+o_p(1)\}. \tag{31}$$

Ignore orders statistics among $U_{(k+1)}, \ldots, U_{(n-k)}$, they can be treated i.i.d samples. Let $S_{n-2k,l} = \sum_{m=k+1}^{n-k}(U_{(m)} - u_0)^l K_h(U_{(m)} - u_0)$, then for $l = 0, 1, \ldots, p$

$$S_{n-2k,l} = \mathbf{E}[S_{n-2k,l}] + O_p\left\{\sqrt{\mathbf{Var}[S_{n-2k,l}]}\right\}$$

For $h \to 0$ and $nh \to \infty$ as $n \to \infty$ we have

$$\begin{aligned}\mathbf{E}[S_{n-2k,l}] &= (n-2k)\,\mathbf{E}[(U-u_0)^l K_h(U-u_0)] \\ &= \frac{(n-2k)}{h}\int K\left(\frac{u-u_0}{h}\right)(u-u_0)^l f(u)du \\ &= (n-2k)h^l\int K(x)x^l f(u_0+xh)dx \\ &= (n-2k)h^l f(u_0)\left[\int x^l K(x)dx + o_p(1)\right] \\ &= (n-2k)h^l f(u_0)\mu_l\{1+o_p(1)\}\end{aligned}$$

and similarly

$$\begin{aligned}O_p\left\{\sqrt{\mathbf{Var}[S_{n-2k,l}]}\right\} &= O_p\left(\sqrt{(n-2k)\,\mathbf{E}[(U-u_0)^{2l}K_h^2(U-u_0)]}\right) \\ &= O_p\left(\sqrt{(n-2k)\int (u-u_0)^{2l}K_h^2(u-u_0)f(u)du}\right) \\ &= O_p\left(\sqrt{(n-2k)h^{2l-1}f(u_0)\int x^{2l}K^2(x)dx}\right) \\ &= O_p\left(\sqrt{(n-2k)h^{2l-1}}\right)\end{aligned}$$

Thus for $h \to 0$ and $nh \to \infty$ as $n \to \infty$, $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$:

$$\begin{aligned} S_{n-2k,l} &= (n-2k)h^l f(u_0)\mu_l \left\{1 + o_p(1) + O_p\left(\sqrt{1/(h(n-2k))}\right)\right\} \\ &= (n-2k)h^l f(u_0)\mu_l \{1 + o_p(1)\} \end{aligned} \tag{32}$$

and

$$\begin{aligned} \mathbf{S}_{n-2k} &= \mathbf{U}_{\mathrm{u}}{}^T \mathbf{W}_{\mathrm{u}} \mathbf{U}_{\mathrm{u}} = \begin{bmatrix} S_{n-2k,0} & S_{n-2k,1} & \cdots & S_{n-2k,p} \\ S_{n-2k,1} & S_{n-2k,1} & \cdots & S_{n-2k,p+1} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n-2k,p} & S_{n-2k,p+1} & \cdots & S_{n-2k,2p} \end{bmatrix} \\ &= (n-2k)f(u_0)H\,\mathbf{S}\,H\{1+o_p(1)\} \end{aligned} \tag{33}$$

where $H = diag\{1, h, \ldots, h^p\}$. Next,

$$\begin{aligned} \mathbf{U}_{\mathrm{u}}{}^T \mathbf{W}_{\mathrm{u}} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} &= \begin{bmatrix} \sum_{m=k+1}^{n-k} K_h(U_{(m)} - u_0) \\ \sum_{m=k+1}^{n-k} (U_{(m)} - u) K_h(U_{(m)} - u_0) \\ \vdots \\ \sum_{m=k+1}^{n-k} (U_{(m)} - u)^p K_h(U_{(m)} - u_0) \end{bmatrix} = \begin{bmatrix} S_{n-2k,0} \\ S_{n-2k,1} \\ \vdots \\ S_{n-2k,p} \end{bmatrix} \\ &= (n-2k)f(u_0)H\tilde{c}_p\{1+o_p(1)\} \end{aligned} \tag{34}$$

where $\tilde{c}_p = (\mu_0, \mu_1, \ldots, \mu_p)^T$. Plugging (33),(34) into (31) gives

$$\begin{aligned} \boldsymbol{\epsilon}_1^T \mathbf{S}_{n-2k}^{-1} \mathbf{U}_{\mathrm{u}}{}^T \mathbf{W}_{\mathrm{u}} \begin{bmatrix} \text{bias}[\hat{Y}_{k+1}^{(1)}|\mathbb{U}] \\ \vdots \\ \text{bias}[\hat{Y}_{n-k}^{(1)}|\mathbb{U}] \end{bmatrix} &\leq \sup_{u\in[0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} \boldsymbol{\epsilon}_1^T H^{-1} \mathbf{S}^{-1} \tilde{c}_p\{1+o_p(1)\} \\ &= \sup_{u\in[0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} \boldsymbol{\epsilon}_1^T \mathbf{S}^{-1} \tilde{c}_p\{1+o_p(1)\}. \end{aligned} \tag{35}$$

Based on (30) and (35), we have

$$\begin{aligned} \text{bias}\big[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\big] &= \boldsymbol{\epsilon}_1^T \mathbf{S}_{n-2k}^{-1} \mathbf{U}_{\mathrm{u}}{}^T \mathbf{W}_{\mathrm{u}} \begin{bmatrix} r^{(1)}(U_{(k+1)}) \\ \vdots \\ r^{(1)}(U_{(n-k)}) \end{bmatrix} - r^{(1)}(u_0) + \boldsymbol{\epsilon}_1^T \mathbf{S}_{n-2k}^{-1} \mathbf{U}_{\mathrm{u}}{}^T \mathbf{W}_{\mathrm{u}} \begin{bmatrix} \text{bias}[\hat{Y}_{k+1}^{(1)}|\mathbb{U}] \\ \vdots \\ \text{bias}[\hat{Y}_{n-k}^{(1)}|\mathbb{U}] \end{bmatrix} \\ &\leq \boldsymbol{\epsilon}_1^T \mathbf{S}^{-1} \left[\frac{c_p}{(p+1)!} r^{(p+2)}(u_0)h^{p+1} + \sup_{u\in[0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} \tilde{c}_p\right]\{1+o_p(1)\}. \end{aligned} \tag{36}$$

Part II (conditional variance)
Provided that $k \to \infty$ as $n \to \infty$, consider the conditional variance in Theorem 1

$$\mathbf{Var}[\hat{Y}_i^{(1)}|\widetilde{\mathbb{U}}] = \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}\{1+o_p(1)\}$$

and by Theorem 1 in De Brabanter et al. (2018)

$$\begin{aligned}\mathbf{Var}\big[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\big] &= \boldsymbol{\epsilon}_1^T\,\mathbf{S}_{n-2k}^{-1}(\mathbf{U}_{\mathrm{u}}{}^T\,\mathbf{W}_{\mathrm{u}}\,\mathbf{Var}[\hat{\mathbf{Y}}^{(1)}|\widetilde{\mathbb{U}}]\,\mathbf{W}_{\mathrm{u}}\,\mathbf{U}_{\mathrm{u}})\,\mathbf{S}_{n-2k}^{-1}\,\boldsymbol{\epsilon}_1 \\ &= \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}\frac{1+f(u_0)\rho_c}{h(n-2k)f(u_0)}\boldsymbol{\epsilon}_1^T\,\mathbf{S}^{-1}\,\mathbf{S}^*\,\mathbf{S}^{-1}\,\boldsymbol{\epsilon}_1\{1+o_p(1)\} \quad (37)\end{aligned}$$

with $\lim_{n\to\infty} n\int \rho_n(x)dx = \rho_c$ and $\mathbf{S}^* = (\nu_{i+j})_{0\le i,j\le p}$ with $\nu_j = \int w^j K^2(u)du$.
For p odd, (see Theorem 3.1 in Fan and Gijbels (1996)):

$$\begin{aligned}\int K_0^*(t)dt &= \boldsymbol{\epsilon}_1^T S^{-1}\left(\int K(t)dt, \int tK(t)dt, \ldots, \int t^p K(t)dt\right)^T \\ &= \boldsymbol{\epsilon}_1^T S^{-1}\tilde{c}_p \qquad (38)\end{aligned}$$

Similarly, we obtain

$$\int t^{p+1}K_0^*(t)dt = \boldsymbol{\epsilon}_1^T S^{-1}c_p, \quad \int K_0^{*2}(t)dt = \boldsymbol{\epsilon}_1^T S^{-1}S^*S^{-1}\boldsymbol{\epsilon}_1 \qquad (39)$$

plugging (38) and (39) into (36) and (37) provides the second part of Theorem 2.

Appendix I. Proof of Corollary 3

For $h\to 0$, $nh\to\infty$ and $k\to\infty$ as $n\to\infty$ such that $n^{-1}k\to 0$ and $nk^{-3}h^{-1}\to 0$, then theorem 2 states that the upperbound of the conditional bias and conditional variance go to zero. Consequently, we have that

$$\lim_{n\to\infty} \mathrm{MSE}\big[\hat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\big] = 0.$$

According to Chebyshev's inequality the proof is complete.

Appendix J. Proof of Theorem 3

The proof for the asymptotic properties of the second order derivatives is similar to that of the first order derivatives. Since r is three times continuously differentiable on the compact interval $[0,1]$, the following Taylor expansions are valid for $r(U_{(i+j+k_1)})$ and $r(U_{(i-j-k_1)})$ in a neighborhood of $U_{(i+j)}$ and $U_{(i-j)}$ respectively

$$r(U_{(i+j+k_1)}) = r(U_{(i+j)}) + \sum_{q=1}^{2}\frac{1}{q!}(U_{(i+j+k_1)}-U_{(i+j)})^q r^{(q)}(U_{(i+j)}) + \frac{(U_{(i+j+k_1)}-U_{(i+j)})^3}{6}r^{(3)}(\zeta_{i+j,i+j+k_1})$$

and

$$r(U_{(i-j-k_1)}) = r(U_{(i-j)}) + \sum_{q=1}^{2}\frac{1}{q!}(U_{(i-j-k_1)}-U_{(i-j)})^q r^{(q)}(U_{(i-j)}) + \frac{(U_{(i-j-k_1)}-U_{(i-j)})^3}{6}r^{(3)}(\zeta_{i-j-k_1,i-j}),$$

where $\zeta_{i+j,i+j+k_1} \in]U_{(i+j)}, U_{(i+j+k_1)}[$ and $\zeta_{i-j-k_1,i-j} \in]U_{(i-j-k_1)}, U_{(i-j)}[$.

Similarly, the following Taylor expansions are also valid for $r^{(1)}(U_{(i+j)})$ and $r^{(1)}(U_{(i-j)})$ in a neighborhood of $U_{(i)}$:

$$r^{(1)}(U_{(i+j)}) = r^{(1)}(U_{(i)}) + (U_{(i+j)} - U_{(i)})r^{(2)}(U_{(i)}) + \frac{(U_{(i+j)} - U_{(i)})^2}{2} r^{(3)}(\zeta_{i,i+j})$$

and

$$r^{(1)}(U_{(i-j)}) = r^{(1)}(U_{(i)}) + (U_{(i-j)} - U_{(i)})r^{(2)}(U_{(i)}) + \frac{(U_{(i-j)} - U_{(i)})^2}{2} r^{(3)}(\zeta_{i-j,i}),$$

where $\zeta_{i,i+j} \in]U_{(i)}, U_{(i+j)}[$ and $\zeta_{i-j,i} \in]U_{(i-j)}, U_{(i)}[$.

$$r^{(2)}(U_{(i+j)}) = r^{(2)}(U_{(i)}) + (U_{(i+j)} - U_{(i)})r^{(3)}(\zeta^{'}_{i,i+j})$$

and

$$r^{(2)}(U_{(i-j)}) = r^{(2)}(U_{(i)}) + (U_{(i-j)} - U_{(i)})r^{(3)}(\zeta^{'}_{i-j,i}),$$

where $\zeta^{'}_{i,i+j} \in]U_{(i)}, U_{(i+j)}[$ and $\zeta^{'}_{i-j,i} \in]U_{(i-j)}, U_{(i)}[$. Since $\sum_{j=1}^{k_2} w_{i,j,2} = 1$, the absolute conditional bias is

$$\begin{aligned}
\Big|\text{bias}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big]\Big| &= \left|\mathbf{E}\left[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right] - r^{(2)}(U_{(i)})\right| \\
&= \left|2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{r(U_{(i+j+k_1)})-r(U_{(i+j)})}{U_{(i+j+k_1)}-U_{(i+j)}} - \frac{r(U_{(i-j-k_1)})-r(U_{(i-j)})}{U_{(i-j-k_1)}-U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}} - r^{(2)}(U_{(i)})\right| \\
&= \left|2\sum_{j=1}^{k_2} w_{i,j,2}\left\{\frac{\left(r^{(1)}(U_{(i+j)}) - r^{(1)}(U_{(i-j)}) + \frac{1}{2}r^{(2)}(U_{(i+j)})(U_{(i+j+k_1)} - U_{(i+j)})\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}}\right.\right. \\
&+ \frac{-\frac{1}{2}r^{(2)}(U_{(i-j)})(U_{(i-j-k_1)} - U_{(i-j)}) + \frac{1}{6}r^{(3)}(\zeta_{i+j,i+j+k_1})(U_{(i+j+k_1)} - U_{(i+j)})^2}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}} \\
&+ \left.\left.\frac{-\frac{1}{6}r^{(3)}(\zeta_{i-j-k_1,i-j})(U_{(i-j-k_1)} - U_{(i-j)})^2}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}}\right\} - r^{(2)}(U_{(i)})\right|
\end{aligned}$$

$$
\begin{aligned}
= & \Bigg|2\sum_{j=1}^{k_2} w_{i,j,2}\Bigg\{\frac{\frac{1}{2}r^{(3)}(\zeta_{i,i+j})(U_{(i+j)}-U_{(i)})^2-\frac{1}{2}r^{(3)}(\zeta_{i-j,i})(U_{(i-j)}-U_{(i)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
+ & \frac{\frac{1}{2}r^{(3)}(\zeta'_{i,i+j})(U_{(i+j)}-U_{(i)})(U_{(i+j+k_1)}-U_{(i+j)})}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
- & \frac{\frac{1}{2}r^{(3)}(\zeta'_{i-j,i})(U_{(i-j)}-U_{(i)})(U_{(i-j-k_1)}-U_{(i-j)})}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
+ & \frac{\frac{1}{6}r^{(3)}(\zeta_{i+j,i+j+k_1})(U_{(i+j+k_1)}-U_{(i+j)})^2-\frac{1}{6}r^{(3)}(\zeta_{i-j-k_1,i-j})(U_{(i-j-k_1)}-U_{(i-j)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\Bigg\}\Bigg| \\
\leq & \sup_{u\in[0,1]}|r^{(3)}(u)|\Bigg(\sum_{j=1}^{k_2} w_{i,j,2}\frac{(U_{(i+j)}-U_{(i)})^2+(U_{(i-j)}-U_{(i)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
+ & \sum_{j=1}^{k_2} w_{i,j,2}\frac{(U_{(i+j)}-U_{(i)})(U_{(i+j+k_1)}-U_{(i+j)})+(U_{(i-j)}-U_{(i)})(U_{(i-j-k_1)}-U_{(i-j)})}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
+ & \sum_{j=1}^{k_2} w_{i,j,2}\frac{\frac{1}{3}(U_{(i+j+k_1)}-U_{(i+j)})^2+\frac{1}{3}(U_{(i-j-k_1)}-U_{(i-j)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\Bigg)
\end{aligned}
$$

Using Lemma 1, the weights in Equation (17), $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ gives

$$
\Big|\text{bias}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big]\Big| \quad \leq \quad \frac{\sup_{u\in[0,1]}|r^{(3)}(u)|}{n+1}\frac{2\sum_{j=1}^{k_2} j^3+3k_1\sum_{j=1}^{k_2} j^2+\frac{5}{3}k_1^2\sum_{j=1}^{k_2} j+\frac{1}{3}k_1^3k_2}{4\sum_{j=1}^{k_2} j^2+k_1^2k_2+4k_1\sum_{j=1}^{k_2} j}\{1+o_p(1)\}
$$

Using the weights in Equation (17) and Lemma 1, the conditional variance is

$$
\begin{aligned}
&\mathbf{Var}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big] = \mathbf{Cov}\Bigg[2\sum_{j=1}^{k_2} w_{i,j,2}\frac{\left(\frac{Y_{i+j+k_1}-Y_{i+j}}{U_{(i+j+k_1)}-U_{(i+j)}}-\frac{Y_{i-j-k_1}-Y_{i-j}}{U_{(i-j-k_1)}-U_{(i-j)}}\right)}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}, \\
&2\sum_{l=1}^{k_2} w_{i,l,2}\frac{\left(\frac{Y_{i+l+k_1}-Y_{i+l}}{U_{(i+l+k_1)}-U_{(i+l)}}-\frac{Y_{i-l-k_1}-Y_{i-l}}{U_{(i-l-k_1)}-U_{(i-l)}}\right)}{U_{(i+l+k_1)}+U_{(i+l)}-U_{(i-l-k_1)}-U_{(i-l)}}|\widetilde{\mathbb{U}}\Bigg] \\
&= 4\sum_{j=1}^{k_2}\sum_{l=1}^{k_2}\frac{w_{i,j,2}w_{i,l,2}}{(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)})(U_{(i+l+k_1)}+U_{(i+l)}-U_{(i-l-k_1)}-U_{(i-l)})} \\
&\Bigg\{\frac{\mathbf{Cov}\big[Y_{i+j+k_1}-Y_{i+j},Y_{i+l+k_1}-Y_{i+l}\big]}{(U_{(i+j+k_1)}-U_{(i+j)})(U_{(i+l+k_1)}-U_{(i+l)})}-\frac{\mathbf{Cov}\big[Y_{i+j+k_1}-Y_{i+j},Y_{i-l-k_1}-Y_{i-l}\big]}{(U_{(i+j+k_1)}-U_{(i+j)})(U_{(i-l-k_1)}-U_{(i-l)})} \\
&-\frac{\mathbf{Cov}\big[Y_{i-j-k_1}-Y_{i-j},Y_{i+l+k_1}-Y_{i+l}\big]}{(U_{(i-j-k_1)}-U_{(i-j)})(U_{(i+l+k_1)}-U_{(i+l)})}+\frac{\mathbf{Cov}\big[Y_{i-j-k_1}-Y_{i-j},Y_{i-l-k_1}-Y_{i-l}\big]}{(U_{(i-j-k_1)}-U_{(i-j)})(U_{(i-l-k_1)}-U_{(i-l)})}\Bigg\}
\end{aligned}
$$

in which $\mathbf{Cov}\big[Y_{i+j+k_1}-Y_{i+j},Y_{i+l+k_1}-Y_{i+l}\big] = \mathbf{Cov}\big[Y_{i+j+k_1},Y_{i+l+k_1}\big]-\mathbf{Cov}\big[Y_{i+j},Y_{i+l+k_1}\big]-\mathbf{Cov}\big[Y_{i+j+k_1},Y_{i+l}\big]+\mathbf{Cov}\big[Y_{i+j},Y_{i+l}\big]$. When $j = l$, the first and the fourth covariance are not zero, when $j = l + k_1$ the second covariance is not zero, and when $j + k_1 = l$, the third covariance is not zero. The other three covariance terms can be obtained in a similar way. Thus,

$$
\begin{aligned}
&\mathbf{Var}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big] \\
&= 4\sigma^2 \sum_{j=1}^{k_2} \frac{w_{i,j,2}^2}{(U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)})^2} \left(\frac{2}{(U_{(i+j+k_1)} - U_{(i+j)})^2} + \frac{2}{(U_{(i-j-k_1)} - U_{(i-j)})^2} \right) \\
&-4\sigma^2 \sum_{j=1}^{k_2-k_1} \frac{w_{i,j,2} w_{i,j+k_1,2}}{(U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)})(U_{(i+j+2k_1)} + U_{(i+j+k_1)} - U_{(i-j-2k_1)} - U_{(i-j-k_1)})} \\
&\left(\frac{1}{(U_{(i+j+k_1)} - U_{(i+j)})(U_{(i+j+2k_1)} - U_{(i+j+k_1)})} + \frac{1}{(U_{(i-j-k_1)} - U_{(i-j)})(U_{(i-j-2k_1)} - U_{(i-j-k_1)})} \right) \\
&-4\sigma^2 \sum_{j=1+k_1}^{k_2} \frac{w_{i,j,2} w_{i,j-k_1,2}}{(U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)})(U_{(i+j)} + U_{(i+j-k_1)} - U_{(i-j)} - U_{(i-j+k_1)})} \\
&\left(\frac{1}{(U_{(i+j+k_1)} - U_{(i+j)})(U_{(i+j)} - U_{(i+j-k_1)})} + \frac{1}{(U_{(i-j-k_1)} - U_{(i-j)})(U_{(i-j)} - U_{(i-j+k_1)})} \right) \\
&\leq 4\sigma^2 \sum_{j=1}^{k_2} \frac{w_{i,j,2}^2}{(U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)})^2} \left(\frac{2}{(U_{(i+j+k_1)} - U_{(i+j)})^2} + \frac{2}{(U_{(i-j-k_1)} - U_{(i-j)})^2} \right) \\
&= \frac{4(n+1)^4\sigma^2}{k_1^2 \sum_{j=1}^{k_2} (2j + k_1)^2} \{1 + o_p(1)\}
\end{aligned}
$$

provided that $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$. Both results hold uniformly for $\sum_{j=1}^{2} k_j + 1 \leq i \leq n - \sum_{j=1}^{2} k_j$.

Appendix K. Proof of Corollary 4

For $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ and from Theorem 3

$$
\begin{aligned}
\left|\text{bias}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big]\right| &\leq \frac{\sup_{u\in[0,1]} |r^{(3)}(u)|}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \{1 + o_p(1)\} \\
&= O_p\left(\max\left\{ \frac{k_1}{n}, \frac{k_2}{n} \right\} \right)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{Var}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big] &\leq \frac{4(n+1)^4\sigma^2}{k_1^2 \sum_{j=1}^{k_2} (2j + k_1)^2} \{1 + o_p(1)\} \\
&= O_p\left(\max\left\{ \frac{n^4}{k_1^2 k_2^3}, \frac{n^4}{k_1^4 k_2} \right\} \right)
\end{aligned}
$$

Under the conditions $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ such that $n^{-1}k_1 \to 0$, $n^{-1}k_2 \to 0$, $n^4 k_1^{-2} k_2^{-3} \to 0$ and $n^4 k_1^{-4} k_2^{-1} \to 0$, Theorem 3 states that the conditional bias and conditional variance go to zero. Consequently, we have that

$$\lim_{n\to\infty} \text{MSE}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big] = \lim_{n\to\infty} \Big(\text{bias}^2\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big] + \mathbf{Var}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big]\Big) = 0.$$

According to Chebyshev's inequality the proof is complete.

Appendix L. Proof of Corollary 5

From the bias-variance decomposition of the mean squared error (MSE), it follows that

$$\begin{aligned}
\text{MSE}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big] &\leq \left(\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2 \{1 + o_p(1)\} \\
&+ \frac{4(n+1)^4\sigma^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1 + o_p(1)\}
\end{aligned}$$

with $\mathcal{B}_2 = \sup_{u\in[0,1]} |r^{(3)}(u)|$. Since $U \sim \mathcal{U}(0,1)$, the mean integrated squared error (MISE) which measures the average global error is

$$\begin{aligned}
\text{MISE}\big[\hat{Y}^{(2)}|\widetilde{\mathbb{U}}\big] &= \mathbf{E}\int_0^1 \big[\hat{Y}^{(2)}(U) - r^{(2)}(U)|\mathbb{U}\big]^2 dU \\
&= \int_0^1 \mathbf{E}\big[\hat{Y}^{(2)}(U) - r^{(2)}(U)|\mathbb{U}\big]^2 dU \\
&\leq \left(\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2 \{1 + o_p(1)\} \\
&+ \frac{4(n+1)^4\sigma^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1 + o_p(1)\}
\end{aligned}$$

where $\hat{Y}^{(2)}(U)$ represents the second order noisy derivative estimator at the design point U. Denote the asymptotic conditional MISE (AMISE) by

$$\text{AMISE}\big[\hat{Y}^{(2)}|\widetilde{\mathbb{U}}\big] \leq \left(\frac{B_2}{n+1}\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2 + \frac{4(n+1)^4\sigma^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}.$$

Appendix M. Proof of Exact Bias for the Second Order Derivative

Assume the fourth order derivative of r exist on [0, 1]; using Lemma 1 and weights in Equation (17), the exact bias of (16) is

$$
\begin{aligned}
\text{bias}\big[\hat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\big] &= 2\sum_{j=1}^{k_2} w_{i,j,2}\frac{\left(\frac{r(U_{(i+j+k_1)})-r(U_{(i+j)})}{U_{(i+j+k_1)}-U_{(i+j)}} - \frac{r(U_{(i-j-k_1)})-r(U_{(i-j)})}{U_{(i-j-k_1)}-U_{(i-j)}}\right)}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} - r^{(2)}(U_{(i)}) \\
&= 2\sum_{j=1}^{k_2} w_{i,j,2}\Bigg\{\frac{\left(r'(U_{(i+j)})-r'(U_{(i-j)})+\frac{1}{2}r^{(2)}(U_{(i+j)})(U_{(i+j+k_1)}-U_{(i+j)})\right)}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
&+ \frac{-\frac{1}{2}r^{(2)}(U_{(i-j)})(U_{(i-j-k_1)}-U_{(i-j)})+\frac{1}{6}r^{(3)}(U_{(i+j)})(U_{(i+j+k_1)}-U_{(i+j)})^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
&+ \frac{-\frac{1}{6}r^{(3)}(U_{(i-j)})(U_{(i-j-k_1)}-U_{(i-j)})^2+O_p(\frac{k_1^3}{n^3})}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\Bigg\} - r^{(2)}(U_{(i)}) \\
&= 2\sum_{j=1}^{k_2} w_{i,j,2}\Bigg\{\frac{\frac{1}{2}r^{(3)}(U_{(i)})[(U_{(i+j)}-U_{(i)})^2-(U_{(i-j)}-U_{(i)})^2]}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
&+ \frac{\frac{1}{2}r^{(3)}(U_{(i)})[(U_{(i+j)}-U_{(i)})(U_{(i+j+k_1)}-U_{(i+j)})-(U_{(i-j)}-U_{(i)})(U_{(i-j-k_1)}-U_{(i-j)})]}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \\
&+ \frac{\frac{1}{6}r^{(3)}(U_{(i)})[(U_{(i+j+k_1)}-U_{(i+j)})^2-(U_{(i-j-k_1)}-U_{(i-j)})^2]+O_p(\frac{k_1^3}{n^3})+O_p(\frac{j^3}{n^3})}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}\Bigg\} \\
&= 2\sum_{j=1}^{k_2} w_{i,j,2}\frac{O_p(\frac{j^{\frac{3}{2}}}{n^2})+O_p(\frac{k_1^{\frac{3}{2}}}{n^2})+O_p(\frac{k_1^3}{n^3})+O_p(\frac{j^3}{n^3})}{\frac{4j+2k_1}{n+1}\left\{1+O_p(\max\{j^{-1/2},k_1^{-1/2}\})\right\}} \\
&= O_p\left(\max\left\{\frac{k_1^{\frac{1}{2}}}{n},\frac{k_2^{\frac{1}{2}}}{n},\frac{k_1^2}{n^2},\frac{k_2^2}{n^2}\right\}\right)
\end{aligned}
$$

Appendix N. Proof of Theorem 4

The proof is analogous to the proof of Theorem 2 in Appendix H. Denote $k' = k_1 + k_2$, following the proof of Theorem 3.1 in Fan and Gijbels (1996) and based on Theorem 3, for $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ we have

$$
\begin{aligned}
&\boldsymbol{\epsilon}_1^T\,\mathbf{S}_{n-2k'}^{-1}\,\mathbf{U}_{\mathrm{u}}{}^T\,\mathbf{W}_{\mathrm{u}}\begin{bmatrix}\text{bias}[\hat{Y}_{k'+1}^{(2)}|\widetilde{\mathbb{U}}] \\ \vdots \\ \text{bias}[\hat{Y}_{n-k'}^{(2)}|\widetilde{\mathbb{U}}]\end{bmatrix} \\
\leq\ &\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2} j^3+3k_1\sum_{j=1}^{k_2} j^2+\frac{5}{3}k_1^2\sum_{j=1}^{k_2} j+\frac{1}{3}k_1^3k_2}{4\sum_{j=1}^{k_2} j^2+k_1^2k_2+4k_1\sum_{j=1}^{k_2} j}\boldsymbol{\epsilon}_1^T\,\mathbf{S}^{-1}\,\tilde{c}_p\{1+o_p(1)\}
\end{aligned}
$$

and for p odd,

$$\begin{aligned}
\boldsymbol{\epsilon}_1^T\,\mathbf{S}_{n-2k'}^{-1}\,\mathbf{U}_{\mathrm{u}}{}^T\,\mathbf{W}_{\mathrm{u}}\begin{bmatrix} r^{(2)}(U_{(k'+1)}) \\ \vdots \\ r^{(2)}(U_{(n-k')}) \end{bmatrix} - r^{(2)}(u_0) &= \boldsymbol{\epsilon}_1^T\,\mathbf{S}^{-1}\,c_p\beta_{p+1}h^{p+1} + o_p(h^{p+1}) \\
&= \boldsymbol{\epsilon}_1^T\,\mathbf{S}^{-1}\,\frac{c_p}{(p+1)!}r^{(p+3)}(u_0)h^{p+1} + o_p(h^{p+1}).
\end{aligned}$$

Combining the two above expressions yields

$$\begin{aligned}
\mathrm{bias}\big[\hat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\big] &= \mathbf{E}\big[\hat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\big] - r^{(2)}(u_0) \\
&= \boldsymbol{\epsilon}_1^T\,\mathbf{S}_{n-2k'}^{-1}\,\mathbf{U}_{\mathrm{u}}{}^T\,\mathbf{W}_{\mathrm{u}}\,\mathbf{E}[\hat{\mathbf{Y}}^{(2)}|\widetilde{\mathbb{U}}] - r^{(2)}(u_0) \\
&= \boldsymbol{\epsilon}_1^T\,\mathbf{S}_{n-2k'}^{-1}\,\mathbf{U}_{\mathrm{u}}{}^T\,\mathbf{W}_{\mathrm{u}}\left(\begin{bmatrix} r^{(2)}(U_{(k'+1)}) \\ \vdots \\ r^{(2)}(U_{(n-k')}) \end{bmatrix} + \begin{bmatrix} \mathrm{bias}[\hat{Y}^{(2)}_{k'+1}|\mathbb{U}] \\ \vdots \\ \mathrm{bias}[\hat{Y}^{(2)}_{n-k'}|\mathbb{U}] \end{bmatrix}\right) - r^{(2)}(u_0) \\
&\le \boldsymbol{\epsilon}_1^T\,\mathbf{S}^{-1}\Bigg[\tilde{c}_p\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j} \\
&+ \frac{c_p}{(p+1)!}r^{(p+3)}(u_0)h^{p+1}\Bigg]\{1+o_p(1)\}.
\end{aligned}$$

Plugging (38) and (39) into this conditional bias gives the second term of the conditional bias of Theorem 4.

According to Theorem 1 in De Brabanter et al. (2018)

$$\begin{aligned}
\mathbf{Var}\big[\hat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\big] &= \boldsymbol{\epsilon}_1^T\,\mathbf{S}_{n-2k'}^{-1}(\mathbf{U}_{\mathrm{u}}{}^T\,\mathbf{W}_{\mathrm{u}}\,\mathbf{Var}[\hat{\mathbf{Y}}^{(2)}|\widetilde{\mathbb{U}}]\,\mathbf{W}_{\mathrm{u}}\,\mathbf{U}_{\mathrm{u}})\,\mathbf{S}_{n-2k'}^{-1}\,\boldsymbol{\epsilon}_1 \\
&\le \frac{4(n+1)^4\sigma_e^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\frac{1+f(u_0)\rho_c'}{h(n-2k')f(u_0)}\boldsymbol{\epsilon}_1^T\,\mathbf{S}^{-1}\,\mathbf{S}^*\,\mathbf{S}^{-1}\,\boldsymbol{\epsilon}_1\{1+o_p(1)\}
\end{aligned}$$

with $\lim_{n\to\infty} n\int\rho_n(x)dx = \rho_c'$ and $\mathbf{S}^* = (\nu_{i+j})_{0\le i,j\le p}$ with $\nu_j = \int u^j K^2(u)du$. Plugging (38) and (39) into this conditional variance gives the second term of the conditional variance of Theorem 4.

References

C.G. Arsalane. *Package 'kedd'*. `https://cran.r-project.org/web/packages/kedd/kedd.pdf`, October 2015.

G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002. ISBN 9780534243128. URL `https://books.google.com/books?id=0x_vAAAAMAAJ`.

R. Charnigo, B. Hall, and C. Srinivasan. A generalized c p criterion for derivative estimation. *Technometrics*, 53(3):238–253, 2011.

P. Chaudhuri and J.S. Marron. Sizer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823, 1999.

N.A.C. Cressie. *Statistics for Spatial Data, 2nd Ed.* Wiley, New York, 1993.

W. Dai, T. Tong, and M.G. Genton. Optimal estimation of derivatives in nonparametric regression. *Journal of Machine Learning Research*, 17(164):1–25, 2016.

H.A. David and H.N. Nagaraja. *Order statistics*. Wiley Online Library, 1970.

K. De Brabanter, J. De Brabanter, B. De Moor, and I. Gijbels. Derivative estimation with local polynomial fitting. *The Journal of Machine Learning Research*, 14(1):281–301, 2013.

K. De Brabanter, C. Fan, I. Gijbels, and J. Opsomer. Local polynomial regression with correlated errors in random design and unknown correlation structure. *Biometrika*, 105 (3):681–690, 2018.

M Delecroix and A.C. Rosa. Nonparametric estimation of a regression function and its derivatives under an ergodic hypothesis. *Journal of Nonparametric Statistics*, 6(4):367–382, 1996.

T Duong. Kernel smoothing. `https://cran.r-project.org/web/packages/ks/index.html`, 2018. [Online; accessed 22-April-2018].

R.L. Eubank and P.L. Speckman. Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, 88(424):1287–1301, 1993.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press, 1996.

M. Francisco-Fernández, J. Opsomer, and J.M. Vilar-Fernández. Plug-in bandwidth selector for local polynomial regression estimator with correlated errors. *Nonparametric Statistics*, 16(1-2):127–151, 2004.

I. Gijbels and A-C Goderniaux. Data-driven discontinuity detection in derivatives of a regression function. *Communications in Statistics-Theory and Methods*, 33(4):851–871, 2005.

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

P. Hall, J.W. Kay, and D.M. Titterinton. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528, 1990.

W. Härdle. *Applied nonparametric regression*. Cambridge university press, 1990.

B. Hassibi and D.G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in neural information processing systems*, pages 164–171, 1993.

A. Iserles. *A first course in the numerical analysis of differential equations*. Cambridge university press, 2009.

Y. Liu and K. De Brabanter. Derivative estimation in random design. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3449–3458. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/7604-derivative-estimation-in-random-design.pdf`.

A. Meister. *Deconvolution Problems in Nonparametric Statistics*. Springer, 2009.

H-G Müller. *Nonparametric regression analysis of longitudinal data*, volume 46. Springer Science & Business Media, 2012.

H-G Müller, U. Stadtmüller, and T. Schmitt. Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, 74(4):743–749, 1987.

C.J.L. Ojeda. locpol: Kernel local polynomial regression. `https://cran.r-project.org/web/packages/locpol/index.html`, 2012. [Online; accessed 22-April-2018].

J. Opsomer, Y. Wang, and Y. Yang. Nonparametric regression with correlated errors. *Statistical Science*, pages 134–153, 2001.

C. Park and K-H. Kang. Sizer analysis for the comparison of regression curves. *Computational Statistics & Data Analysis*, 52(8):3954–3970, 2008.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

J. Ramsay. derivative estimation. *StatLib -S-News*, 1998.

J.O. Ramsay and B.W. Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2007.

J. Ramsey and B. Ripley. pspline: Penalized smoothing splines. `https://cran.r-project.org/web/packages/pspline/index.html`, 2017. [Online; accessed 22-April-2018].

J.A. Rice. Bandwidth choice for differentiation. *Journal of Multivariate Analysis*, 19(2): 251–264, 1986.

V. Rondonotti, J.S. Marron, and C. Park. Sizer for time series: a new approach to the analysis of trends. *Electronic Journal of Statistics*, 1:268–289, 2007.

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956.

C.J. Stone. Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705, 1985.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.

G. Wahba and Y. Wang. When is the optimal regularization parameter insensitive to the choice of the loss function? *Communications in Statistics-Theory and Methods*, 19(5): 1685–1700, 1990.

M.P. Wand and M.C. Jones. *Kernel smoothing*. Crc Press, 1994.

W. Wang and L. Lin. Derivative estimation based on difference sequence via locally weighted least squares regression. *Journal of Machine Learning Research*, 16:2617–2641, 2015.

Y. Xia. Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):797–811, 1998.

S. Zhou and D.A. Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, pages 93–108, 2000.