Smoothed Nonparametric Derivative Estimation Using Weighted Difference Quotients by Liu and De Brabanter (2020)

Yikun Zhang

Department of Statistics, University of Washington, Seattle, WA, 98195

June 16, 2023

Abstract

This report discusses a nonparametric derivative estimation method for the random design proposed by the paper (Liu and De Brabanter, 2020). We examine their data-driven derivative estimation framework, which combines weighted difference quotients with local polynomial regression. In addition to scrutinizing the asymptotic properties of the proposed derivative estimators and the selection proposals of tuning parameters in the paper, we also fill the theoretical gaps by establishing the consistency results of the final proposed derivative estimators. Finally, we reproduce all of their simulation studies through `R` with some extensions that compare the proposed derivative estimators with other classical methods in terms of both estimation accuracy and computational efficiency.

1 Introduction

Assume that we observe an independent and identically distributed (i.i.d.) sample $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R} \times \mathbb{R}$ from the following model:

$$Y = m(X) + e, \tag{1}$$

where $m(x) = \mathbb{E}(Y|X = x)$ is an unknown regression function and X is a covariate with unknown density f and cumulative distribution function (CDF) F on $[a, b] \subset \mathbb{R}$. Further, it is assumed that the noise variable e is independent of X with $\mathbb{E}(e) = 0, \mathrm{Var}(e) = \sigma_e^2 < \infty$. The left panel of Figure 1 gives an example of the simulated observations from model (1).

Many applications of interest focus not only on estimating the regression function m that well-approximates the observed data but also its derivatives

$$m^{(1)}(x) = \lim_{\Delta \to 0} \frac{m(x + \Delta) - m(x)}{\Delta} \tag{2}$$

within the support $[a, b]$, given that $m^{(1)}(x)$ reveals the changing trend and local curvature information of the function m. For instance, derivatives of consumption in labor economics quantify how the marginal propensity to consume (Haavelmo, 1947) would change with respect to incomes, savings, and other factors among a specific population (Dang, 2021, 2022). In biomechanics, studying

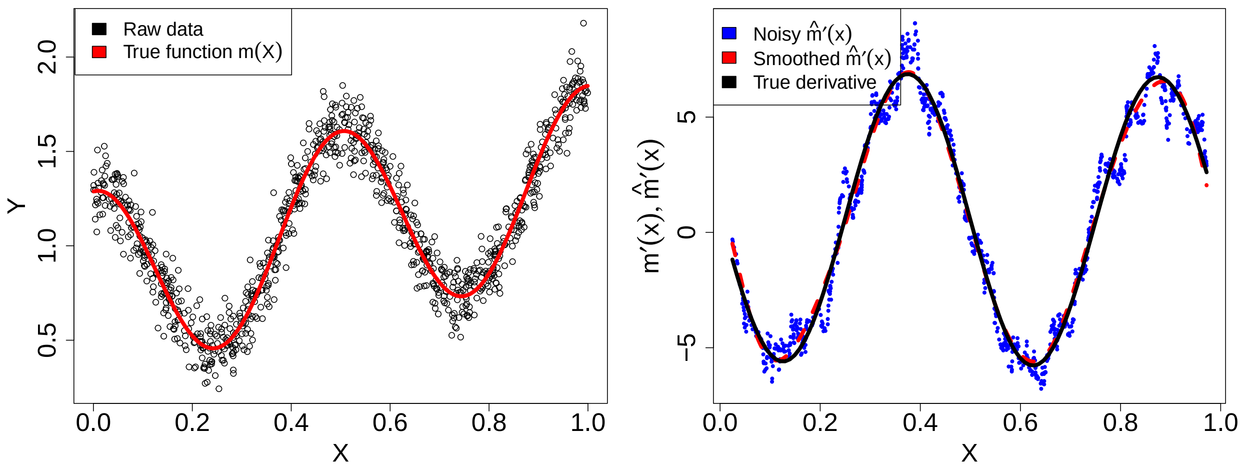


Figure 1: Simulated data $\{(X_i, Y_i)\}_{i=1}^{1000}$ from model (1) with the first-order noisy derivatives and proposed smoothed derivative estimates. The left panel plots the raw data with $m(X) = \cos^2(2\pi X) + \log(4/3 + X), X \sim \text{Unif}(0,1)$ and $e \sim N(0, 0.1^2)$. The right panel shows the first-order noisy derivatives, the proposed smoothed derivative estimator with $k = 26$, and the true derivative $m^{(1)}(X)$. (This figure is extended from Figure 1(a) and Figure 2(b) in the paper.)

the derivatives from displacement data facilitates the kinematic analysis of different body segments during human movements (Woltring, 1985). Within the fields of statistics, derivative estimation appears in the exploration of curve structures (Chaudhuri and Marron, 1999; Gijbels and Goderniaux, 2005), trend analysis in time series (Rondonotti et al., 2007), comparisons of regression curves (Park and Kang, 2008), investigation of human growth data (Müller, 1988; Ramsay and Silverman, 2002), and bias corrections of regression estimators for conducting valid inference (Eubank and Speckman, 1993; Xia, 1998; Calonico et al., 2018; Cheng and Chen, 2019).

The main challenge of the derivative estimation problem is a lack of specific data for the derivatives of $m(x)$, in that only the data from model (1) are given. Such an unavailability of derivative data also makes the parameter tuning and model selection more difficult in the context of derivative estimation. One straightforward proposal for estimating the derivatives is to derive an estimator $\widehat{m}(x)$ of the regression function and take its derivatives. Nevertheless, the performance of such a derivative estimator relies heavily on how well the original regression function is estimated, and the estimation errors accumulate as the orders of derivatives increase (De Brabanter et al., 2013).

Our discussed paper (Liu and De Brabanter, 2020), which is an extended version of Liu and De Brabanter (2018), tackles the above challenges by proposing a data-driven method for estimating derivatives directly from the observed data $\{(X_i, Y_i)\}_{i=1}^{n}$; see Section 2. It extends the framework

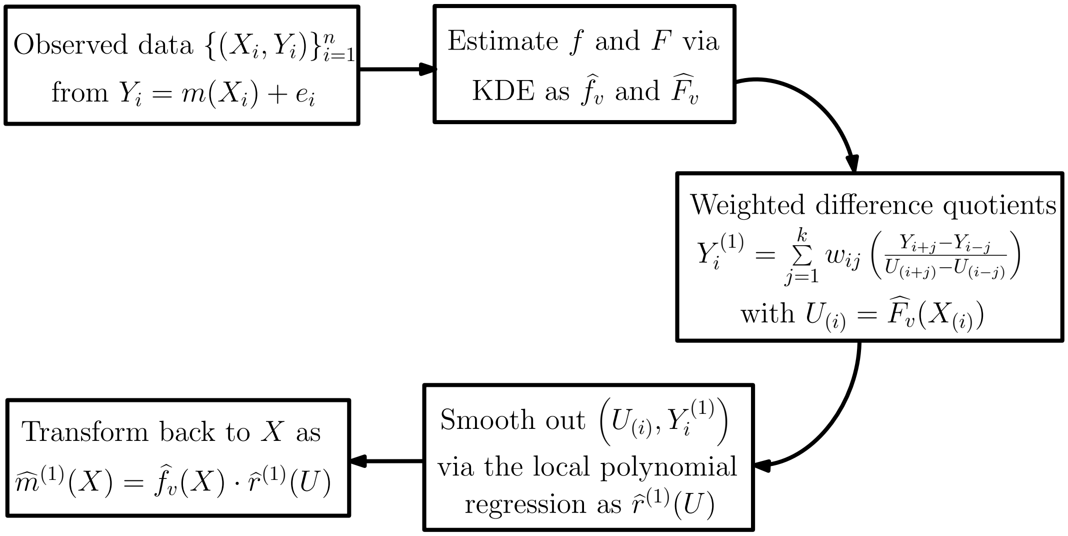


Figure 2: Summary of the derivative estimation framework in the paper.

proposed by De Brabanter et al. (2013) from the equispaced design, where $X_i = a + \frac{(i-1)(b-a)}{n-1}, i = 1, ..., n$, to the random design as model (1). In particular, a set of empirical derivatives is constructed through weighted difference quotients, and the local polynomial regression for correlated errors (De Brabanter et al., 2018) is utilized to smooth out these noisy derivatives; see Figure 2 for a brief methodological summary and the right panel of Figure 1 for an example. One crucial difference between the equispaced and random designs is that a basic assumption called the symmetric property $x_{i+j} - x_i = x_i - x_{i-j}$ no longer holds for the random design (1). Overcoming this difficulty and deriving asymptotic properties become the main contributions of the paper; see Section 3. Given that the paper only presents the asymptotic properties under the uniformly distributed covariate on $[0, 1]$, we derive the pointwise and uniform rates of convergence for the proposed derivative estimators when the unknown distribution of X is estimated by the kernel density estimator (KDE; Rosenblatt 1956; Parzen 1962; Chen 2017) as an extension; see Section 4. While the authors of the paper did not make any code publicly available, we reproduce all of their simulation studies and supplement a real-world application in Section 5 and Appendix A. The reproducible code and a new `Python` implementation are available at `https://github.com/zhangyk8/NonDeriDQ`.

1.1 Other Related Literature

In the regime of derivative estimation, parametric methods are rarely used because it is difficult to propose a valid parametric family that explains the data. The only related literature on parametric derivative estimation appears to be in signal processing (Belkić and Belkić, 2018), where a complicated form of the fast Padé transform was applied. Even when it is common to estimate the variogram via parametric models in spatial statistics, Gorsich and Genton (2000) still advocated for nonparametric derivative estimation in order to assist the variogram model selection. We briefly review the major nonparametric derivative estimation methods as follows, where the optimal rate of convergence for a nonparametric derivative estimator was studied by Stone (1980, 1982).

- **Regression/Smoothing splines:** The spline regression (de Boor, 1968) approximates the regression function $m(x)$ by a linear basis expansion as $f(x) = \sum_{j=1}^{M} \beta_j g_j(x)$, where $\{g_j : \mathbb{R} \to \mathbb{R}\}_{j=1}^{M}$ are polynomial transformations of x and $\boldsymbol{\beta} = (\beta_1, ..., \beta_M)^T \in \mathbb{R}^M$ is obtained from the least-square solution under some knot constraints or using B-splines; see Chapter 5 of Hastie et al. (2009). The L_2 rate of convergence for derivative estimators based on regression splines was derived in Stone (1985). Other asymptotic properties, including bias, variance, and normality, of the derivatives of regression splines for estimating the derivatives of $m(x)$ were studied by Zhou and Wolfe (2000). As for smoothing splines, one will search for the solution that minimizes the penalized residual sum of squares $\sum_{i=1}^{n} [Y_i - f(X_i)]^2 + \lambda \int [f''(t)]^2 \, dt$ among all functions $f(x)$ with twice continuous derivatives, which can be shown to be a natural cubic spline with knots at $\{X_i\}_{i=1}^{n}$ and $\widehat{f}(x) = \sum_{i=1}^{n} \widehat{\beta}_j g_j(x)$ can be obtained by the usual penalized least-square solution (Hastie et al., 2009). Estimating the derivatives $m^{(q)}(x), q = 1, 2, ...$ via the derivatives of smoothing splines may not be ideal, since the smoothing parameter depends on the order of the derivative (Wahba and Wang, 1990). In the case of semiparametric penalized splines, Jarrow et al. (2004) indeed noticed that more smoothing is required for derivative estimation than the smoothing parameter selected by generalized cross-validation.

- **Gasser-Müller derivative estimator:** There were also some research works about kernel-based derivative estimation methods. Given the i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^{n}$ from model (1), one particular example method is based on the Gasser-Müller regression estimator (Gasser and Müller, 1979) as $\widehat{m}_{h,GM}(x) = \frac{1}{h} \sum_{i=1}^{n} Y_i \cdot \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h}\right) du$, where $s_i = \frac{X_{(i)} + X_{(i+1)}}{2}$ for $i = 1, ..., n$ with $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$, $K : \mathbb{R} \to [0, \infty)$ is the kernel function, and $h > 0$ is the bandwidth parameter. Gasser and Müller (1984) considered using the q-th order derivative of $\widehat{m}_h(x)$ as $\widehat{m}_{h,GM}^{(q)}(x) = \frac{1}{h^{q+1}} \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K^{(q)}\left(\frac{x-u}{h}\right) du$ to estimate the true derivative $m^{(q)}(x)$ of the regression. A robust variant of the Gasser-Müller derivative estimator was also discussed in Härdle and Gasser (1985).

- **Nadaraya-Watson derivative estimator:** Another well-known kernel-based derivative estimator stems from Nadaraya-Watson regression estimator (Nadaraya, 1964; Watson, 1964). Instead of using the derivative of Nadaraya-Watson regression estimator, Mack and Müller (1989) proposed a simpler variant as $\widehat{m}_{h,NW}^{(q)}(x) = \frac{1}{nh^{q+1}} \sum_{i=1}^{n} \frac{Y_i \cdot K^{(q)}\left(\frac{x-X_i}{h}\right)}{\widehat{f}_v(X_i)}$, where $\widehat{f}_v$ is the KDE for the density f of X in model (1) with the bandwidth parameter $v > 0$.

Some uniform consistency properties of these kernel-based derivative estimators were also analyzed by Delecroix and Rosa (1996). For their bandwidth selection, Rice (1986); Müller et al. (1987) proposed a generalized cross-validation criterion that utilizes the difference quotients introduced below and a factor method that relies on a careful choice of kernel functions.

• **Local polynomial regression:** Local polynomial regression (Fan and Gijbels, 1996) generalizes Nadaraya-Watson estimator and leads to an intuitive estimate for the q-th order derivative $m^{(q)}(x)$. The idea is from Taylor's theorem (Rudin et al., 1976), where under smoothness conditions, the regression function $m(x_0)$ can be locally approximated by a polynomial of order $p > q$ as $m(x_0) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(x_0 - x)^j \equiv \sum_{j=0}^{p} \beta_j(x)\,(x_0 - x)^j$. The coefficients $\widehat{\boldsymbol{\beta}}(x) = \left(\widehat{\beta}_0(x), ..., \widehat{\beta}_p(x)\right)^T \in \mathbb{R}^{p+1}$ of the fitted polynomial at point $x \in \mathbb{R}$ can be obtained by solving the following weighted least-square problem as:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}(x) &= \underset{\boldsymbol{\beta}(x)\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{p} \beta_j(x)\cdot(X_i - x)^j\right]^2 K\left(\frac{X_i - x}{h}\right) \\ &= \underset{\boldsymbol{\beta}(x)\in\mathbb{R}^{p+1}}{\arg\min} \left\{\left[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}(x)\right]^T \boldsymbol{W} \left[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}(x)\right]\right\}, \end{aligned} \tag{3}$$

where $K : \mathbb{R} \to [0, \infty)$ is the kernel function, $h > 0$ is the smoothing bandwidth parameter, and

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^p \end{pmatrix}, \quad \boldsymbol{W} = \begin{pmatrix} K\left(\frac{X_1 - x}{h}\right) & & \\ & \ddots & \\ & & K\left(\frac{X_n - x}{h}\right) \end{pmatrix}.$$

Thus, $\widehat{\boldsymbol{\beta}}(x) = \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$. Moreover, $\widehat{m}^{(q)}(x) = q! \cdot \widehat{\beta}_q(x)$ is a natural estimator for the derivatives $m^{(q)}(x), q = 0, 1, ..., p$.

• **Difference quotient based methods:** It is natural from the definition (2) to apply the (first-order) difference quotients $\widehat{q}_i^{(1)} = \frac{Y_i - Y_{i-1}}{X_{(i)} - X_{(i-1)}}, i = 2, ..., n$ (also called Newton's quotients; Lang 1968) to estimating the first-order derivative (Müller et al., 1987; Härdle, 1990; Charnigo et al., 2011), where $X_{(1)} \leq \cdots \leq X_{(n)}$ are order statistics of $\{X_i\}_{i=1}^n$ and $Y_i, i = 1, ..., n$ are also reordered according to $\{X_{(i)}\}_{i=1}^n$. However, the variances of difference quotients are (stochastically) proportional to n^2 under some smoothness conditions on m and noises with nonzero variances as in model (1); see Section 2.1 in De Brabanter et al. (2013) and our Remark 6 in Appendix B.8. In order to reduce the variance, Iserles (2009) considered aggregating several symmetric difference quotients by linear combinations as:

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{X_{(i+j)} - X_{(i-j)}} \right) \quad \text{with} \quad \sum_{j=1}^{k} w_{i,j} = 1 \quad \text{for } i = k+1, ..., n-k, \tag{4}$$

where $k \leq \frac{n-1}{2}$ is a tuning parameter and the weights $w_{i,j}, j = 1, ..., k$ for each i are chosen by minimizing the variance $\mathrm{Var}(\widehat{Y}_i^{(1)})$. The idea of weighted difference quotients has been employed in derivative estimation by De Brabanter et al. (2013); Wang and Lin (2015), and Dai et al. (2016) further generalizes this idea by formulating a constrained optimization problem for obtaining the weights. All these works focus on the equispaced design, and it is unclear how to extend their methods to the random design.

2 Derivative Estimation via Weighted Difference Quotients

In this section, we present the methodology of estimating first and second order derivatives via weighted difference quotients (4) and local polynomial smoothing in the discussed paper.

2.1 Probability Integral Transform to the Uniform Distribution

Recall model (1) that generates our i.i.d. data $\{(X_i, Y_i)\}_{i=1}^n$, where the covariate X has density f and CDF F. It is well-known (Casella and Berger, 2002) that $F(X_i), i = 1, ..., n$ follows the uniform distribution $\mathrm{Unif}[0,1]$. Thus, it suffices to estimate the derivatives of a transformed regression function $r(U) = m(F^{-1}(U))$ and refer back to the derivatives of the original regression function $m(X) = r(F(X))$ through the chain rules as:

$$\begin{aligned} m^{(1)}(X) &= \frac{dm(X)}{dX} = \frac{dr(U)}{dU} \cdot \frac{dU}{dX} = f(X) \cdot r^{(1)}(U), \\ m^{(2)}(X) &= \frac{d^2 m(X)}{dX^2} = \frac{d}{dX} \left(f(X) \cdot \frac{dr(U)}{dU} \right) = f^{(1)}(X) \cdot r^{(1)}(U) + [f(X)]^2 \, r^{(2)}(U). \end{aligned} \tag{5}$$

While $F(X), f(X), f^{(1)}(X)$ are unknown in practice, they can be estimated by the KDE as:

$$\widehat{f}_v(X) = \frac{1}{nv} \sum_{i=1}^{n} K_{\mathrm{kde}} \left(\frac{X - X_i}{v} \right), \quad \widehat{F}_v(X) = \frac{1}{nv} \sum_{i=1}^{n} \int_{-\infty}^{X} K_{\mathrm{kde}} \left(\frac{u - X_i}{v} \right) du, \tag{6}$$

and $\widehat{f}_v^{(1)}(X) = \frac{1}{nv^2} \sum_{i=1}^{n} K_{\mathrm{kde}}^{(1)} \left(\frac{X - X_i}{v} \right)$, where $K_{\mathrm{kde}} : \mathbb{R} \to [0, \infty)$ is the kernel function and $v > 0$ is the bandwidth parameter. In the paper, the Gaussian kernel $K_{\mathrm{kde}}(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2} \right)$ is applied, and the bandwidth v is selected by the two-stage plug-in method (Sheather and Jones, 1991); see also Section 3.6 in Wand and Jones (1994). Practically, these quantities can be obtained from the `R` functions `kde`, `kcde`, `kdde` with default parameters in the `R` package `ks` (Duong, 2022).

2.2 First-Order Noisy Derivative Estimator

Given that the observed covariates $\{X_i\}_{i=1}^n$ can be transformed into (approximately) uniformly distributed random variables on $[0,1]$ as shown in Section 2.1, we consider the i.i.d. data $\{(U_i, Y_i)\}_{i=1}^n$ with $U_i \sim \text{Unif}[0,1], i = 1, ..., n$ in the sequel. Furthermore, we order the data $\{(U_i, Y_i)\}_{i=1}^n$ according to the magnitude of $U_i, i = 1, ..., n$ so that the model (1) becomes

$$Y_i = r(U_{(i)}) + e_i, \quad i = 1, ..., n, \tag{7}$$

where $r(u) = \mathbb{E}[Y|U = u] = m(F^{-1}(u))$ has the same role as the regression function $m(x)$ in (1) and $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$ are order statistics. Based on (4), the proposed first-order derivative estimator for the random design at $u = U_{(i)}$ is defined as:

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right) \quad \text{ for } \quad k+1 \leq i \leq n-k, \tag{8}$$

where $k \leq \frac{n-1}{2}$ is the tuning parameter and the weights are given by $w_{i,j} = \frac{\left(U_{(i+j)} - U_{(i-j)}\right)^2}{\sum_{\ell=1}^{k}\left(U_{(i+\ell)} - U_{(i-\ell)}\right)^2}$ for $j = 1, ..., k$ that minimize the variance of (8); see Proposition 10 in Appendix B.8. Notice that the j-th weight $w_{i,j}$ is proportional to the reciprocal variance of the difference quotient $\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}$ and incorporates the equispaced design on $[a, b]$ satisfying $U_{(i+j)} - U_{(i-j)} = \frac{2j(b-a)}{(n-1)}$ for $j = 1, ..., k$ in Charnigo et al. (2011); De Brabanter et al. (2013) as a special case.

• **Boundary Correction:** The proposed estimator (8) is only valid at $U_{(i)}$ for $k+1 \leq i \leq n-k$. Within the left and right boundary regions $2 \leq i \leq k$ and $n-k+1 \leq i \leq n-1$, one may consider using $k(i)$ weighted difference quotients in (8) instead, where $k(i) = i - 1$ for the left boundary and $k(i) = n - i$ for the right boundary. However, the asymptotic variance of $\widehat{Y}_i^{(1)}$ is $O_P\left(\frac{3\sigma_e^2(n+1)^2}{k(i)(k(i)+1)(2k(i)+1)}\right)$ and will become $O_P(n^2)$ when i is close to 2 and $n - 1$; see Theorem 1 in Section 3.1. To reduce the variance of $\widehat{Y}_i^{(1)}$ within the boundary regions, the paper proposes a boundary corrected estimator as:

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k(i)} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right) + \sum_{j=k(i)+1}^{k} w_{i,j} \left[\left(\frac{Y_{i+j} - Y_i}{U_{(i+j)} - U_{(i)}} \right) \mathbb{1}_{\{2 \leq i \leq k\}} + \left(\frac{Y_i - Y_{i-j}}{U_{(i)} - U_{(i-j)}} \right) \mathbb{1}_{\{n-k < i < n\}} \right], \tag{9}$$

where

$$w_{i,j} = \begin{cases} \frac{\left(U_{(i+j)} - U_{(i-j)}\right)^2}{\sum_{\ell=1}^{k(i)}\left(U_{(i+\ell)} - U_{(i-\ell)}\right)^2 + \sum_{\ell=k(i)+1}^{k}\left[\left(U_{(i+\ell)} - U_{(i)}\right)^2 \mathbb{1}_{\{2 \leq i \leq k\}} + \left(U_{(i)} - U_{(i-\ell)}\right)^2 \mathbb{1}_{\{n-k < i < n\}}\right]}, & 1 \leq j \leq k(i), \\ \frac{\left(U_{(i+j)} - U_{(i)}\right)^2 \mathbb{1}_{\{2 \leq i \leq k\}} + \left(U_{(i)} - U_{(i-j)}\right)^2 \mathbb{1}_{\{n-k < i < n\}}}{\sum_{\ell=1}^{k(i)}\left(U_{(i+\ell)} - U_{(i-\ell)}\right)^2 + \sum_{\ell=k(i)+1}^{k}\left[\left(U_{(i+\ell)} - U_{(i)}\right)^2 \mathbb{1}_{\{2 \leq i \leq k\}} + \left(U_{(i)} - U_{(i-\ell)}\right)^2 \mathbb{1}_{\{n-k < i < n\}}\right]}, & k(i) < j \leq k. \end{cases}$$

We also take $\widehat{Y}_1^{(1)} = \widehat{Y}_2^{(1)}$ and $\widehat{Y}_n^{(1)} = \widehat{Y}_{n-1}^{(1)}$. In the worst-case scenario, the variance of $\widehat{Y}_i^{(1)}$ given by (9) reduces to $O_P\left(\frac{n^2}{k^2}\right)$ and its bias is still of the order $O_P\left(\frac{k}{n}\right)$.

2.3 Second-Order Noisy Derivative Estimator

Under model (7), the proposed second-order derivative estimator is defined as:

$$\widehat{Y}_i^{(2)} = 2\sum_{j=1}^{k_2} w_{ij,2} \cdot \frac{\left(\frac{Y_{i+j+k_1}-Y_{i+j}}{U_{(i+j+k_1)}-U_{(i+j)}} - \frac{Y_{i-j-k_1}-Y_{i-j}}{U_{(i-j-k_1)}-U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}} \quad \text{for} \quad k_1+k_2+1 \leq i \leq n-k_1-k_2, \tag{10}$$

where k_1, k_2 are tuning parameters and $\sum_{j=1}^{k_2} w_{ij,2} = 1$. As in the first-order derivative estimator (8), the exact j-th weight $\widetilde{w}_{ij,2}$ will be proportional to the reciprocal variance of the j-th weighted difference terms conditional on $\{U_{(i)}\}_{i=1}^n$ as:

$$\widetilde{w}_{ij,2} \propto \frac{1}{\mathrm{Var}\left[\frac{\left(\frac{Y_{i+j+k_1}-Y_{i+j}}{U_{(i+j+k_1)}-U_{(i+j)}} - \frac{Y_{i-j-k_1}-Y_{i-j}}{U_{(i-j-k_1)}-U_{(i-j)}}\right)}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \middle| U_{(i)}, i=1,...,n\right]} \quad \text{for} \quad j=1,...,k_2.$$

To simplify the estimation procedure, the paper considers taking the asymptotic dominating order of $\widetilde{w}_{ij,2}$ as the actual weight $w_{ij,2} = \frac{(2j+k_1)^2}{\sum_{j=1}^{k_2}(2j+k_1)^2}$ in (10), in the sense that $\widetilde{w}_{ij,2} = w_{ij,2}\left\{1+o_P(1)\right\}$ as $k_1, k_2 \to \infty$ by Lemma 9 in Appendix B.8. In addition, given that the boundary correction estimator (9) is too complicated to implement for the second-order derivative estimator (10), the paper only utilizes the maximum numbers $k_1(i), k_2(i)$ of the available first and second order difference quotients to construct $\widehat{Y}_i^{(2)}$ within the boundary regions $i \leq k_1+k_2$ and $i > n-k_1-k_2$.

2.4 Smoothing the Noisy Derivatives Through Local Polynomial Regression

The above empirical/noisy derivative estimators (8), (9), and (10) are only defined at the design points $\{U_{(i)}\}_{i=1}^n$ and would contain noises from the error terms $e_i, i=1,...,n$ in (7). To extrapolate beyond the design points and reduce noises, the paper considers applying the local polynomial regression to the interior noisy derivative data $\{(U_{(i)}, \widehat{Y}_i^{(1)})\}_{i=k+1}^{n-k}$ for the first order and $\{(U_{(i)}, \widehat{Y}_i^{(2)})\}_{i=k_1+k_2+1}^{n-k_1-k_2}$ for the second order. Specifically, in the case of smoothing the first-order derivative estimator (8), we recall from Section 1.1 that the local polynomial estimator at point $u_0 \in [0,1]$ for estimating the derivative $r^{(1)}(u_0)$ in model (7) is given by

$$\widehat{r}^{(1)}(u_0) = \boldsymbol{\epsilon}_1^T \widehat{\boldsymbol{\beta}}(u_0) = \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{u_0}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \widehat{\boldsymbol{Y}}^{(1)}, \tag{11}$$

where $\boldsymbol{\epsilon}_1 = (1, 0, ..., 0)^T \in \mathbb{R}^{p+1}$, $\widehat{\boldsymbol{Y}}^{(1)} = \left(\widehat{Y}_{k+1}^{(1)}, ..., \widehat{Y}_{n-k}^{(1)}\right)^T \in \mathbb{R}^{n-2k}$, $\boldsymbol{S}_{u_0} = \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \boldsymbol{U}_{u_0} \equiv \boldsymbol{S}_{n-2k}$, and

$$\boldsymbol{U}_{u_0} = \begin{pmatrix} 1 & (U_{(k+1)} - u_0) & \cdots & (U_{(k+1)} - u_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (U_{(n-k)} - u_0) & \cdots & (U_{(n-k)} - u_0)^p \end{pmatrix}, \quad \boldsymbol{W}_{u_0} = \begin{pmatrix} K\left(\frac{U_{(k+1)} - u_0}{h}\right) & & \\ & \ddots & \\ & & K\left(\frac{U_{(n-k)} - u_0}{h}\right) \end{pmatrix}.$$

To smooth out the second-order derivative estimator (10), we similarly define the local polynomial estimator

$$\widehat{r}^{(2)}(u_0) = \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{u_0}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \widehat{\boldsymbol{Y}}^{(2)}, \tag{12}$$

where $\widehat{\boldsymbol{Y}}^{(2)} = \left(\widehat{Y}_{k_1+k_2+1}^{(2)}, ..., \widehat{Y}_{n-k_1-k_2}^{(2)}\right)$, while $\boldsymbol{U}_{u_0}, \boldsymbol{W}_{u_0}$ and $\boldsymbol{S}_{u_0} = \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \boldsymbol{U}_{u_0} \equiv \boldsymbol{S}_{n-2k_1-2k_2}$ are defined through the data $\{(U_{(i)}, \widehat{Y}_i^{(2)})\}_{i=k_1+k_2+1}^{n-k_1-k_2}$.

One caveat in applying the local polynomial regression (11) is that $\left\{\widehat{Y}_i^{(1)}\right\}_{i=k+1}^{n-k}$ are no longer independent even when we condition on $\{U_{(i)}\}_{i=1}^n$. To inspect this fact, one can rewrite the first-order noisy derivative estimator (8) as:

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

where we denote the second term by $\sum_{j=1}^k w_{i,j} \left(\frac{e_{i+j} - e_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right)$ as the new error terms for $k+1 \leq i \leq n-k$. The first term in the above equation is an approximation of $r^{(1)}(U_{(i)})$ with its absolute bias bounded by $O\left(\frac{k}{n}\right) \to 0$ as $n \to \infty$; see Theorem 1 below. Hence, $\{(U_{(i)}, \widehat{Y}_i^{(1)})\}_{i=k+1}^{n-k}$ can be regarded as an (ordered) random sample from the model with correlated errors $\widetilde{e}_i, i = k+1, ..., n-k$ as:

$$\widehat{Y}_i^{(1)} = r^{(1)}(U_{(i)}) + \widetilde{e}_i, \tag{13}$$

where $\mathbb{E}\left[\widetilde{e}_i | U_i\right] = 0$ and $\mathrm{Cov}\left(\widetilde{e}_i, \widetilde{e}_j | U_{(i)}, U_{(j)}\right) = \sigma_{\widetilde{e}}^2 \cdot \rho_n(U_{(i)} - U_{(j)})$ with $\sigma_{\widetilde{e}}^2 < \infty$ and ρ_n being a stationary correlation function with $\rho_n(0) = 1, \rho_n(u) = \rho_n(-u)$ and $|\rho_n(u)| \leq 1$ for all $u \in \mathbb{R}$. Such correlated error structures complicate the bandwidth selection for the local polynomial smoothing (11) and deteriorate the performance of the final derivative estimator (Opsomer et al., 2001; De Brabanter et al., 2018). To resolve this issue, the paper adopts a two-step procedure proposed by De Brabanter et al. (2018) to select the final bandwidth $\widehat{h}$ in (11) as follows.

1. We fit a local polynomial regression (11) using a bimodal kernel $\bar{K} : \mathbb{R} \to [0, \infty)$ with $\bar{K}(0) = 0$ and compute a pilot bandwidth $\widehat{h}_b$ by minimizing the residual sum of squares (RSS) as:

$$\widehat{h}_b = \arg\min_{h_b > 0} \mathrm{RSS}(h_b) = \arg\min_{h_b > 0} \left\{\frac{1}{n-2k} \sum_{i=k+1}^{n-k} \left(\widehat{r}^{(1)}(U_{(i)}) - \widehat{Y}_i^{(1)}\right)^2\right\}, \tag{14}$$

given the tuning parameter k is chosen a priori as Corollary 2 in Section 3.1. In the paper, the bimodal Gaussian kernel $\bar{K}(u) = \frac{2u^2}{\sqrt{\pi}} \exp(-u^2)$ is applied.

2. To handle the extra mean squared error caused by the non-optimality of kernels with $\bar{K}(0) = 0$, we consider the bandwidth correction as:

$$\widehat{h} = \left\{ \frac{\int \left(K_p^\star(t)\right)^2 dt \left[\int t^{p+1} \bar{K}_p^\star(t) dt\right]^2}{\int \left(\bar{K}_p^\star(t)\right)^2 dt \left[\int t^{p+1} K_p^\star(t) dt\right]^2} \right\}^{\frac{1}{2p+2}} \widehat{h}_b = 1.01431 \widehat{h}_b,$$

where $K_p^\star(u), \bar{K}_p^\star(u)$ are equivalent kernels defined by $\bar{K}(u)$ and $K(u)$ (see Section 3.2.2 in Fan and Gijbels 1996), and the last equality follows by using $\bar{K}(u) = \frac{2u^2}{\sqrt{\pi}} \exp(-u^2)$ and $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$ under the local cubic regression with $p = 3$.

The final smoothed derivative estimator (11) is computed with the unimodal kernel K and selected bandwidth $\widehat{h}$. The above bandwidth selection procedure also applies to the local polynomial smoothing of the second-order noisy derivative estimator (10).

Remark 1. The paper does not address the derivative estimation of order higher than two, because it is unwieldy to generalize the proposed framework and its asymptotic properties. More importantly, the bias of weighted difference quotient estimators accumulate as the derivative order increases under the random design setting. As a result, the proposed framework is inadequate to estimate the higher-order derivatives of the regression function $m(x)$.

3 Asymptotic Properties of the Proposed Derivative Estimators

In this section, we study the asymptotic conditional bias and variance of the proposed derivative estimators and their smoothed counterparts by local polynomial regression under model (7). These asymptotic results suggest some practical guidelines for choosing the tuning parameters k or k_1, k_2.

3.1 First-Order Derivative Estimation

Theorem 1 (Theorem 1 in Liu and De Brabanter 2020)**.** *Assume that r is twice continuously differentiable on $[0, 1]$ under model (7). Then, the conditional bias and variance of the first-order noisy derivative estimator (8) given $\mathbb{U} = \left(U_{(i-j)}, ..., U_{(i+j)}\right)$ for $i > j$ and $i + j \leq n$ are*

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

$$\mathrm{Var}\left[\widehat{Y}_i^{(1)}|\mathbb{U}\right] = \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} + o_P\left(\frac{n^2}{k^3}\right)$$

uniformly for $k+1 \le i \le n-k$ when $k \to \infty$ as $n \to \infty$. Further, if we assume that r is $q+1$ times continuously differentiable on $[0,1]$ for $q \ge 1$, then the asymptotic order of the exact conditional bias is given by

$$\mathrm{Bias}\left[\widehat{Y}_i^{(1)}|\mathbb{U}\right] = \begin{cases} O_P\left(\frac{k}{n}\right), & q=1, \\ O_P\left(\max\left\{\frac{k^{\frac{1}{2}}}{n}, \frac{k^2}{n^2}\right\}\right), & q \ge 2. \end{cases}$$

The proof of Theorem 1 can be found in Appendix B.1. It reveals the following three corollaries:

- The conditional bias and variance of $\widehat{Y}_i^{(1)}$ will tend to 0 if the tuning parameter k tends to infinity in a rate faster than $O\left(n^{\frac{2}{3}}\right)$ but slower than $O(n)$.
- $\widehat{Y}_i^{(1)}$ is a (pointwise) consistent estimator of $r^{(1)}(U_{(i)})$ as $\frac{k}{n} \to 0$ and $\frac{n^2}{k^3} \to 0$.
- The fastest L_2 rate of convergence for $\mathbb{E}\left[\left(\widehat{Y}_i^{(1)} - r^{(1)}(U_{(i)})\right)^2 |\mathbb{U}\right]$ is $O_P\left(n^{-\frac{2}{5}}\right)$ when $k = O\left(n^{\frac{4}{5}}\right)$, which is also the optimal rate of convergence among twice differentiable functions (Stone, 1982).

Theorem 1 also provides us with a practical guideline for selecting the tuning parameter k in (8). In principle, we will minimize the asymptotic conditional mean integrated squared error (MISE).

Corollary 2 (Corollary 2 in Liu and De Brabanter 2020)**.** *Let $\mathcal{B} = \sup_{u\in[0,1]}\left|r^{(2)}(u)\right|$. Under the assumptions of Theorem 1, the tuning parameter k that minimizes the asymptotic upper bound of the conditional MISE is given by*

$$k_{opt} = \underset{k=1,2,\dots,\lfloor\frac{n-1}{2}\rfloor}{\arg\min}\left[\mathcal{B}^2\frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2} + \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}\right].$$

The proof of Corollary 2 is given in Appendix B.2. To apply Corollary 2 in practice, the unknown quantity $\mathcal{B} = \sup_{u\in[0,1]}\left|r^{(2)}(u)\right|$ can be approximated by the second-order local slope of a local polynomial regression of order $p=3$ fitted to the data $\{(U_{(i)}, Y_i)\}_{i=1}^n$, while the noise variance σ_e^2 can be estimated by Hall's $\sqrt{n}$-consistent estimator with the optimal second-order difference sequence (Hall et al., 1990) as $\widehat{\sigma}_e^2 = \frac{1}{n-2}\sum_{i=1}^{n-2}(0.809Y_i - 0.5Y_{i+1} - 0.309Y_{i+2})^2$. Then, the optimal value k_{opt} can be obtained by searching over the integer set within $\left[1, \frac{n-1}{2}\right]$.

Now, we study the asymptotic pointwise conditional bias and variance of the smoothed derivative estimator (11). Recall from Section 2.4 that the first-order noisy derivative data $\left\{\left(U_{(i)}, \widehat{Y}_i^{(1)}\right)\right\}_{i=k+1}^{n-k}$ obtained from (8) can be viewed as observations from the additive noise model (13), where the errors $\widetilde{e}_i, i = k+1, ..., n-k$ satisfy $\mathbb{E}\left[\widetilde{e}_i|U_i\right] = 0$ and $\mathrm{Cov}\left(\widetilde{e}_i, \widetilde{e}_j|U_{(i)}, U_{(j)}\right) = \sigma_{\widetilde{e}}^2 \cdot \rho_n(U_{(i)} - U_{(j)})$ with $\sigma_{\widetilde{e}}^2 < \infty$ and ρ_n being a stationary correlation function with $\rho_n(0) = 1, \rho_n(u) = \rho_n(-u)$ and $|\rho_n(u)| \le 1$ for all $u \in \mathbb{R}$. We make the following assumptions.

Assumption 1. The kernel function $K : \mathbb{R} \to [0, \infty)$ is bounded, symmetric, and Lipschitz continuous at 0. Furthermore, it satisfies $\lim_{|u|\to\infty} |u|^\ell K(u) < \infty$ for $\ell = 0, ..., p$.

Assumption 2. The correlation function ρ_n is an element of a sequence $\{\rho_n\}_{n=1}^\infty$ with the following properties for all $n \geq 1$: there exist constants $\rho_{\max}, \rho_c > 0$ such that $n \int |\rho_n(x)| dx < \rho_{\max}$ and $\lim_{n\to\infty} n \int \rho_n(x) dx = \rho_c$. In addition, for any sequence $\epsilon_n > 0$ with $n\epsilon_n \to \infty$, it holds that $n \int_{|x|\geq\epsilon_n} |\rho_n(x)| dx \to 0$ as $n \to \infty$.

Assumption 1 is a standard and mild condition for kernel functions (Wasserman, 2006). Assumption 2 requires the correlation to be short-range dependent (Opsomer et al., 2001) and can be satisfies by $\widetilde{e}_i, i = k+1, ..., n-k$ in model (13) when k is small. Other correlation functions satisfying Assumption 2 includes $\rho_n(x) = \exp(-\alpha n|x|)$ and $\rho_n(x) = \frac{1}{1+\alpha n^2 x^2}$ for $\alpha > 0$. The asymptotic pointwise conditional bias and variance of $\widehat{r}^{(1)}(u_0)$ in (11) are given in the following theorem.

Theorem 3 (Theorem 2 in Liu and De Brabanter 2020)**.** *Assume that $r(\cdot)$ under model (7) is $(p+2)$ times continuously differentiable in a neighborhood of u_0. Under Assumptions 1 and 2, the conditional bias and variance of (11) with $u_0 \in [0,1]$ for p odd are*

$$\begin{aligned}
\mathrm{Bias}\left[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\right] &\leq \left[\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} c_p \cdot \frac{r^{(p+2)}(u_0)}{(p+1)!} \cdot h^{p+1} + \left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right| \widetilde{c}_p \cdot \frac{3k(k+1)\mathcal{B}}{4(n+1)(2k+1)}\right] [1 + o_P(1)] \\
&= \left[\left(\int t^{p+1} K_0^\star(t) dt\right) \frac{r^{(p+2)}(u_0)}{(p+1)!} \cdot h^{p+1} + \left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right| \widetilde{c}_p \cdot \frac{3k(k+1)\mathcal{B}}{4(n+1)(2k+1)}\right] [1 + o_P(1)], \\
\mathrm{Var}\left[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\right] &= \frac{3\sigma_e^2 (n+1)^2 (1+\rho_c)}{k(k+1)(2k+1)(n-2k)h} \cdot \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} \boldsymbol{S}^* \boldsymbol{S}^{-1} \boldsymbol{\epsilon}_1 [1 + o_P(1)] \\
&= \left(\int K_0^\star(t)^2 dt\right) \frac{3\sigma_e^2 (n+1)^2 (1+\rho_c)}{k(k+1)(2k+1)(n-2k)h} [1 + o_P(1)]
\end{aligned}$$

as $h \to 0, nh \to \infty, k \to \infty$ with $n \to \infty$, where $\widetilde{\mathbb{U}} = \left(U_{(1)}, ..., U_{(n)}\right)$, $\mathcal{B} = \sup_{u\in[0,1]} \left|r^{(2)}(u)\right|$, $\boldsymbol{S} = (\mu_{i+j-2})_{1\leq i,j\leq p+1}$ with $\mu_j = \int u^j K(u) du$, $\boldsymbol{S}^* = (\nu_{i+j-2})_{1\leq i,j\leq p+1}$ with $\nu_j = \int u^j K(u)^2 du$, $c_p = (\mu_{p+1}, ..., \mu_{2p+1})^T$, $\widetilde{c}_p = (\widetilde{\mu}_0, ..., \widetilde{\mu}_p)^T$ with $\widetilde{\mu}_j = \int |u|^j K(u) du$, $\boldsymbol{\epsilon}_1 = (1, 0, ..., 0)^T \in \mathbb{R}^{p+1}$, $\left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right|$ means elementwise absolute values of $\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}$, and the equivalent kernel $K_0^\star(t) = \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} (1, t, ..., t^p)^T K(t)$.

The proof of Theorem 3 is in Appendix B.3. It implies that the optimal L_2 rate of convergence is $O_P\left(n^{-\frac{4p+4}{5p+6}}\right)$, which is attained when $h = O\left(n^{-\frac{2}{5p+6}}\right)$ and $k = O\left(n^{\frac{3p+4}{5p+6}}\right)$; see Remark 3 in Appendix B.3. Unlike Corollary 2, it is complicated to leverage the bias-variance trade-off in Theorem 3 to select the optimal bandwidth h and tuning parameter k simultaneously, since there are too many unknown quantities that need estimating. Instead, the bandwidth h is selected through a two-step procedure by minimizing the residual sum of squares under a bimodal kernel $\bar{K}$ with an additional correction; recall the details in Section 2.4. The rationale behind this procedure is due to the asymptotic equivalence between minimizing $\mathrm{RSS}(h) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left(\widehat{r}^{(1)}(U_{(i)}) - \widehat{Y}_i^{(1)}\right)^2$ and the sample squared error $\mathrm{SSE}(h) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left(\widehat{r}^{(1)}(U_{(i)}) - r^{(1)}(U_{(i)})\right)^2$ as follows.

Lemma 4 (Theorem 2 in De Brabanter et al. 2018)**.** *Under the assumptions in Theorem 3 with a local polynomial regression estimator $\widehat{r}^{(1)}$ given by (11), if $n^\delta \int |\rho_n(t)|dt < \rho_\delta$ for $\delta > 1$, p is odd, and $h \in \mathcal{H}_n$ with $\mathcal{H}_n = \left[c_1 n^{-\frac{1}{2p+3}}, c_2 n^{-\frac{1}{2p+3}}\right]$ for some constants $0 < c_1 < c_2 < \infty$, then*

$$\mathrm{RSS}(h) = \mathrm{SSE}(h) + \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\widetilde{e}_i^2 - \frac{2\sigma_{\widetilde{e}}^2 \cdot K(0)\cdot \left(\boldsymbol{S}^{-1}\right)_{11}\cdot(1+\rho_c)}{nh} + o_P\left(n^{-\frac{2p+2}{2p+3}}\right),$$

recalling that the domain of $r^{(1)}$ is $[0,1]$ and $\left(\boldsymbol{S}^{-1}\right)_{11}$ is the first element in the first row of $\boldsymbol{S}^{-1}$.

According to Lemma 4, the term related to the correlation structures of errors in model (13) will be removed if we employ a bimodal kernel $\bar{K}$ with $\bar{K}(0) = 0$. Hence, $\frac{\mathrm{RSS}(h)}{\mathrm{SSE}(h)} = 1 + o_P(1)$ as $n \to \infty$, and we can select the optimal bandwidth by minimizing $\mathrm{RSS}(h)$ without any prior knowledge about the correlation structures of errors.

3.2 Second-Order Derivative Estimation

Theorem 5 (Theorem 3 in Liu and De Brabanter 2020)**.** *Assume that r is three times continuously differentiable on $[0,1]$ under model (7). Then, under the weight $w_{ij,2} = \frac{(2j+k_1)^2}{\sum_{j=1}^{k_2}(2j+k_1)^2}$, the conditional bias and variance of the second-order noisy derivative estimator (10) given $\widetilde{\mathbb{U}} = \left(U_{(1)}, ..., U_{(n)}\right)$ are bounded by*

$$\left|\mathrm{Bias}\left[\widehat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]\right| \le \frac{\sup_{u\in[0,1]}\left|r^{(3)}(u)\right|}{n+1}\left(\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)[1 + o_P(1)],$$
$$\mathrm{Var}\left[\widehat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right] \le \frac{4(n+1)^4\sigma_e^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}[1 + o_P(1)]$$

uniformly for $k_1 + k_2 + 1 \le i \le n - k_1 - k_2$ when $k_1, k_2 \to \infty$ as $n \to \infty$.

The proof of Theorem 5 is given in Appendix B.4. This theorem implies the following corollaries:

- Assuming that k_1, k_2 have the same asymptotic order with respect to n. The conditional bias and variance of $\widehat{Y}_i^{(2)}$ tend to 0 if $k_1, k_2 \to \infty$ in a rate faster than $O\left(n^{\frac{4}{5}}\right)$ but slower than $O(n)$.
- $\widehat{Y}_i^{(2)}$ is a (pointwise) consistent estimator of $r^{(2)}(U_{(i)})$ if $\frac{k_1}{n} \to 0$, $\frac{k_2}{n} \to 0$, $\frac{n^4}{k_1^2 k_2^3} \to 0$, $\frac{n^4}{k_1^4 k_2} \to 0$ as $n \to \infty$; see Remark 4 in Appendix B.4.

- The fastest possible L_2 rate of convergence for $\mathbb{E}\left[\left(\widehat{Y}_i^{(2)} - r^{(2)}(U_{(i)})\right)^2 |\widetilde{\mathbb{U}}\right] \to 0$ is $O_P\left(n^{-\frac{2}{7}}\right)$ when $k_1, k_2 = O\left(n^{\frac{6}{7}}\right)$.

Analogous to the first-order derivative estimation through (8), Theorem 5 also sheds light on the bias-variance rationale of selecting the tuning parameters k_1, k_2 for the second-order noisy derivative estimator (10).

Corollary 6 (Corollary 5 in Liu and De Brabanter 2020)**.** *Let $\mathcal{B}_2 = \sup_{u\in[0,1]} \left|r^{(3)}(u)\right|$. Under the assumptions of Theorem 5, the tuning parameters k_1 and k_2 that minimize the asymptotic upper bound of the conditional MISE are*

$$(k_1, k_2)_{opt} = \underset{k_1,k_2=1,2,\ldots}{\arg\min} \left[\frac{\mathcal{B}_2^2}{(n+1)^2}\left(\frac{2\sum\limits_{j=1}^{k_2} j^3 + 3k_1\sum\limits_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum\limits_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2 + \frac{4(n+1)^4\sigma_e^2}{k_1^2\sum\limits_{j=1}^{k_2}(2j+k_1)^2}\right].$$

The proof of Corollary 6 is given in Appendix B.5. The unknown quantity $\mathcal{B}_2$ can be approximated by the local polynomial regression estimator with $p = 4$, while σ_e^2 is again estimated by Hall's $\sqrt{n}$-consistent estimator described after Corollary 2. Then, the optimal pair $(k_1, k_2)_{\text{opt}}$ can be obtained by grid-searching over a Cartesian product set $\left\{1, 2, ..., \lfloor\frac{n-1}{2}\rfloor\right\} \otimes \left\{1, 2, ..., \lfloor\frac{n-1}{2}\rfloor\right\}$.

Finally, we study the asymptotic pointwise conditional bias and variance of the smoothed second-order derivative estimator (12). As in (13), the data $\{(U_{(i)}, \widehat{Y}_i^{(2)})\}_{i=k_1+k_2+1}^{n-k_1-k_2}$ can also be viewed as an (ordered) random sample from the model $\widehat{Y}_i^{(2)} = r^{(2)}(U_{(i)}) + \acute{e}_i$, in which the error terms $\acute{e}_i, i = k_1+k_2+1, ..., n-k_1-k_2$ are correlated with $\mathbb{E}\left(\acute{e}_i|U_{(i)}\right) = 0, \operatorname{Cov}\left(\acute{e}_i, \acute{e}_j|U_{(i)}, U_{(j)}\right) = \sigma_{\acute{e}}^2 \cdot \acute{\rho}_n(U_{(i)} - U_{(j)})$ for $i \neq j$, and $\acute{\rho}_n$ is a stationary correlation function with $\acute{\rho}_n(0) = 1$, $\acute{\rho}_n(u) = \acute{\rho}_n(-u)$, and $|\acute{\rho}_n(u)| \leq 1$ for all $u \in \mathbb{R}$. Analogous to Theorem 3, we have the following theorem for the asymptotic upper bound of the conditional bias and variance of $\widehat{r}^{(2)}(u_0)$ in (12).

Theorem 7 (Theorem 4 in Liu and De Brabanter 2020)**.** *Assume that $r(\cdot)$ under model (7) is $(p+3)$ times continuously differentiable in a neighborhood of u_0. Under Assumptions 1 and 2 on $\acute{\rho}_n$, the conditional bias and variance of (12) with $u_0 \in [0,1]$ for p odd are*

$$\begin{aligned}
\operatorname{Bias}\left[\widehat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\right] \leq & \left[\left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right| \widetilde{c}_p \cdot \left(\frac{\mathcal{B}_2}{n+1}\right)\left(\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)\right. \\
& \left. + \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} c_p \cdot \frac{r^{(p+3)}(u_0)}{(p+1)!} \cdot h^{p+1}\right] [1 + o_P(1)] \\
\operatorname{Var}\left[\widehat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\right] \leq & \frac{4(n+1)^4\sigma_e^2(1+\acute{\rho}_c)}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2(n-2k_1-2k_2)h} \cdot \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\boldsymbol{S}^*\boldsymbol{S}^{-1}\boldsymbol{\epsilon}_1 [1 + o_P(1)]
\end{aligned}$$

$$= \frac{4(n+1)^4\sigma_e^2(1+\acute{\rho}_c)}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2(n-2k_1-2k_2)h}\left(\int K^\star(t)^2dt\right)[1+o_P(1)]$$

when $h \to 0$, $nh \to \infty$, $k_1, k_2 \to \infty$ *as* $n \to \infty$, *where* $\mathcal{B}_2 = \sup_{u\in[0,1]}\left|r^{(3)}(u)\right|$, $\widetilde{\mathbb{U}} = \left(U_{(1)}, ..., U_{(n)}\right)$, $\boldsymbol{S} = (\mu_{i+j-2})_{1\le i,j\le p+1}$ *with* $\mu_j = \int u^j K(u)du$, $\boldsymbol{S}^* = (\nu_{i+j-2})_{1\le i,j\le p+1}$ *with* $\nu_j = \int u^j K(u)^2 du$, $c_p = (\mu_{p+1}, ..., \mu_{2p+1})^T$, $\widetilde{c}_p = (\widetilde{\mu}_0, ..., \widetilde{\mu}_p)^T$ *with* $\widetilde{\mu}_j = \int |u|^j K(u)du$, $\boldsymbol{\epsilon}_1 = (1,0,...,0)^T \in \mathbb{R}^{p+1}$, $\left|\boldsymbol{\epsilon}_1^T\boldsymbol{S}^{-1}\right|$ *means elementwise absolute values of* $\boldsymbol{\epsilon}_1^T\boldsymbol{S}^{-1}$, *and the equivalent kernel* $K_0^\star(t) = \boldsymbol{\epsilon}_1^T\boldsymbol{S}^{-1}\left(1,t,...,t^p\right)^T K(t)$.

The proof of Theorem 7 is in Appendix B.6. When the tuning parameters k_1, k_2 have the same asymptotic order as k, the optimal upper bound for the conditional MISE is $O_P\left(n^{-\frac{4p+4}{7p+8}}\right)$, which is attained when $k = O\left(n^{\frac{5p+6}{7p+8}}\right)$ and $h = O\left(n^{-\frac{2}{7p+8}}\right)$; see Remark 5 in Appendix B.6. Analogous to the first-order smoothed derivative estimator, we leverage the two-step procedure to select the bandwidth h after choosing k_1, k_2 via Corollary 6.

4 Extensions

All the asymptotic properties in the discussed paper (Liu and De Brabanter, 2020) are presented after the probability integral transform in Section 2.1, in which the covariate $U = F(X)$ is Unif$[0,1]$ distributed; see also Section 3 above. Under an arbitrary distribution of X, the paper suggests using KDE (6) to estimate its distribution without any theoretical justification. Thus, it is still unclear how the asymptotic rate of convergence for the final derivative estimators

$$\widehat{m}^{(1)}(x) = \widehat{f}_v(x)\cdot\widehat{r}^{(1)}(u) \quad \text{and} \quad \widehat{m}^{(2)}(x) = \widehat{f}_v^{(1)}(x)\cdot\widehat{r}^{(1)}(u) + \left[\widehat{f}_v(x)\right]^2\widehat{r}^{(2)}(u) \tag{15}$$

would be when the unknown distribution of X is estimated by the KDE (6). Here, we leverage the convergence theories for KDE (Giné and Guillou, 2002; Einmahl and Mason, 2005; Chacón et al., 2011) and local polynomial regression (Francisco-Fernández et al., 2003) to derive the pointwise and uniform rates of convergence for the final derivative estimators (15).

Assumption 3. The kernel function for KDE $K_{\text{kde}} : \mathbb{R} \to [0,\infty)$ is bounded, symmetric, and differentiable (almost everywhere) with $\int u^2 K_{\text{kde}}(u)du < \infty$ and $\int K_{\text{kde}}^{(\alpha)}(u)^2 du < \infty$ for $\alpha = 0, 1$.

Assumption 4. Let $\mathcal{K} = \left\{y \mapsto K_{\text{kde}}^{(\alpha)}\left(\frac{x-y}{v}\right) : x \in \mathbb{R}, v > 0, \alpha = 0, 1\right\}$. We assume that $\mathcal{K}$ is a bounded VC (subgraph) class of measurable functions on $\mathbb{R}$, *i.e.*, there exist absolute constants $A, \nu > 0$ such that for any $\epsilon \in (0,1)$, $\sup_Q N\left(\mathcal{K}, L_2(Q), \epsilon\, ||F||_{L_2(Q)}\right) \le \left(\frac{A}{\epsilon}\right)^\nu$, where $N\left(\mathcal{K}, L_2(Q), \epsilon\right)$ is the ϵ-covering number of the normed space $\left(\mathcal{K}, ||\cdot||_{L_2(Q)}\right)$, Q is any probability measure on $\mathbb{R}$, and F is an envelope function of $\mathcal{K}$. Here, the norm $||F||_{L_2(Q)}$ is defined as $\left[\int_{\mathbb{R}} |F(x)|^2 dQ(x)\right]^{\frac{1}{2}}$.

Assumption 5. The stationary correlation functions ρ_n and $\acute{\rho}_n$ of the error terms $\widetilde{e}_i, \acute{e}_i$ in the first and second order noisy derivative estimators come from a first-order autoregressive process with $\mathbb{E}\left(|\widetilde{e}_i|^\delta\right) < \infty, \mathbb{E}\left(|\acute{e}_i|^\delta\right) < \infty$ and are α-mixing with mixing coefficients $\alpha(k)$ such that $\sum_{k=1}^{\infty} k \cdot \alpha(k)^{1-\frac{2}{\delta}}$ for some $\delta > 2$. Moreover, define the sequence $M_n = \left(n \log n (\log\log n)^{1+\gamma}\right)^{\frac{1}{\delta}}$ for some $0 < \gamma < 1$. Then, the bandwidth $h = h_n$ satisfies that $\gamma_n = \left(\frac{nM_n^2}{h_n^3 \log n}\right)^{\frac{1}{2}} \to \infty$ and $b_n = \left(\frac{nh_n}{M_n^2 \log n}\right)^{\frac{1}{2}} \to \infty$ as $n \to \infty$. Finally, the α-mixing sequence $\alpha(k)$ satisfies $\sum_{n=1}^{\infty} \frac{n\gamma_n}{b_n} \left(\frac{nM_n^2}{h_n \log n}\right)^{\frac{1}{2}} \alpha(b_n) < \infty$.

Assumptions 3 and 4 are not stringent and can be satisfied by the Gaussian kernel and other compactly supported kernel functions due to Lemma 22 in Nolan and Pollard (1987). In particular, Assumption 4 was assumed by Giné and Guillou (2002); Einmahl and Mason (2005) to control the complexity of the kernel and establish the uniform consistency of KDE. Assumption 5 regularizes the correlation structures of the error terms from the noisy derivative estimates. It is imposed by Francisco-Fernández et al. (2003) to determine an appropriate truncation sequence to obtain a precise block size for the Bernstein's block technique. This regularity condition also appeared in Masry (1996) when the author studied the local polynomial regression for time series data.

Theorem 8. *Assume that $m(\cdot)$ under model (1) is $(p+3)$ times continuously differetiable within $[a,b]$, and the density f of X is at least three times continuously differentiable with $\inf_{x\in[a,b]} f(x) > c > 0$ for some constant c. Then,*

- **Pointwise consistency:** *under Assumptions 1, 2, and 3, the derivative estimators in (15) for $q = 1, 2$ and any fixed $x \in [a,b]$ satisfy*

$$\left|\widehat{m}^{(q)}(x) - m^{(q)}(x)\right| = O\left(h^{p+1}\right) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n^{2q-1}}{k^{2q+1}h}}\right) + O(v^2) + O_P\left(\sqrt{\frac{1}{nv^{2q-1}}}\right)$$

  when $h \to 0, \frac{k}{n} \to 0, \frac{n^{2q-1}}{k^{2q+1}h} \to 0, v \to 0, nv^{2q-1} \to \infty$ as $n \to \infty$.

- **Uniform consistency:** *under Assumptions 1, 2, 3, 4, and 5, when $h \to 0, \frac{k}{n} \to 0, \frac{n^{2q-1}\log n}{k^{2q+1}h} \to 0, v \to 0, \frac{nv^{2q-1}}{\log n} \to \infty$ as $n \to \infty$, we have that*

$$\sup_{x\in[a,b]} \left|\widehat{m}^{(q)}(x) - m^{(q)}(x)\right| = O\left(h^{p+1}\right) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n^{2q-1}\log n}{k^{2q+1}h}}\right) + O(v^2) + O_P\left(\sqrt{\frac{\log n}{nv^{2q-1}}}\right).$$

The proof of Theorem 8 is in Appendix B.7. Compared with Theorems 3 and 7, Theorem 8 suggests that the correct rates of convergence for the final derivative estimators (15) have extra additive terms that depend on the bandwidth v in KDE (6). However, if we select the bandwidth to be $v = O\left(n^{-\frac{1}{2q+3}}\right)$ for $q = 1, 2$ that minimizes the MISE, then these rates of convergence in Theorem 8 will be dominated by the optimal rates of the proposed derivative estimators based on the minimization of conditional MISEs in Theorems 3 and 7 and can be ignored as in the paper.

5 Simulation Studies and Real-World Applications

In this section, we compare the proposed derivative estimators via weighted difference quotients and local polynomial smoothing described in Section 2 with some well-studied nonparametric derivative estimation methods on some simulated data. We also apply the proposed first-order derivative estimator to analyzing Washington state-level COVID-19 case rates in Appendix A.4. The methods that we compare include but is not limited to those that have been discussed in the paper as:

1. **Penalized smoothing splines:** Recall from Section 1.1 that taking the derivatives of penalized smoothing splines is another classical nonparametric method for estimating the derivatives $m^{(q)}(x)$. This method is implemented in `R` package `pspline` (Ramsey and Ripley, 2022).

2. **Local polynomial regression:** As introduced in (3), the q-order local slope $\widehat{m}^{(q)}(x) = q! \cdot \widehat{\beta}_q(x)$ with $q \leq p$ of a local polynomial regression with degree p is a natural estimator of $m^{(q)}(x)$. This method is implemented in `R` package `locpol` (Ojeda Cabrera, 2022).

3. **Gasser-Müller derivative estimator:** We implement the Gasser-Müller derivative estimator introduced in Section 1.1 with the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$. The default bandwidth parameter is selected via the optimal cross-validated bandwidth for the local polynomial regression with $p = 0$ in `R` function `regCVBwSelC` in the package `locpol`.

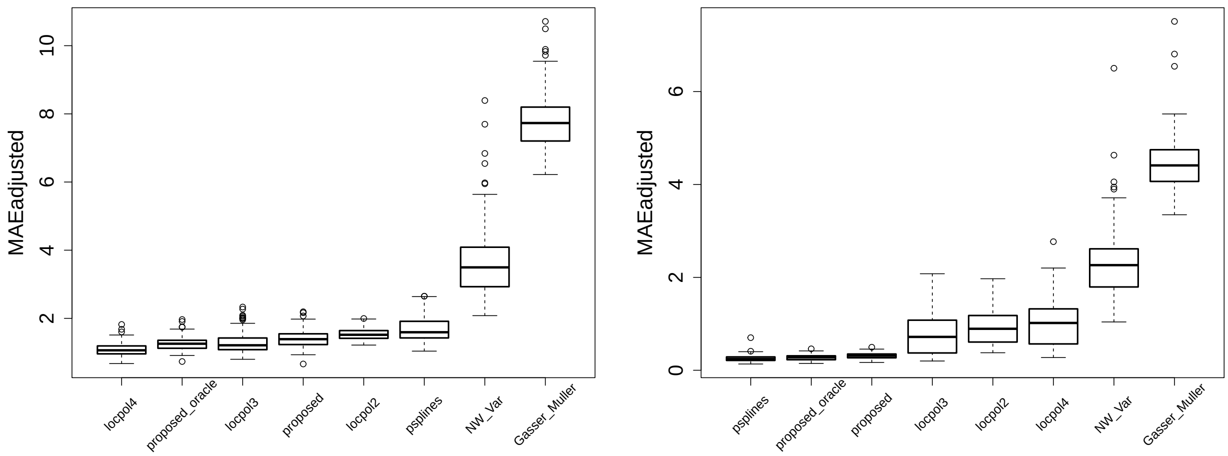
4. **Nadaraya-Watson derivative estimator:** We implement the Nadaraya-Watson derivative estimator described in Section 1.1, where we apply the Gaussian kernel and select the bandwidths v for KDE and h for the derivative estimator $\widehat{m}_{h,NW}^{(q)}$ using the two-stage plug-in method (Sheather and Jones, 1991) and the optimal cross-validated bandwidth for the local polynomial regression with $p = 0$ in `R` function `regCVBwSelC` in the package `locpol`, respectively.

5.1 Simulation Studies on the First-Order Derivative Estimation

As in the discussed paper, we simulate data sets of size $n = 700$ from model (1) with the function

$$m(X) = \sqrt{X(1-X)} \cdot \sin\left(\frac{2.1\pi}{X + 0.05}\right) \quad \text{with} \quad X \sim \text{Unif}(0.25, 1) \quad \text{and} \quad e \sim N(0, 0.2^2) \tag{16}$$

for 100 times. The tuning parameter k is selected via Corollary 2 over the positive integer set $\left\{1, 2, ..., \lfloor \frac{n-1}{2} \rfloor\right\}$ unless otherwise stated. The bandwidth h of the proposed estimator (11) is ini-



(a) Results for simulation study (16).

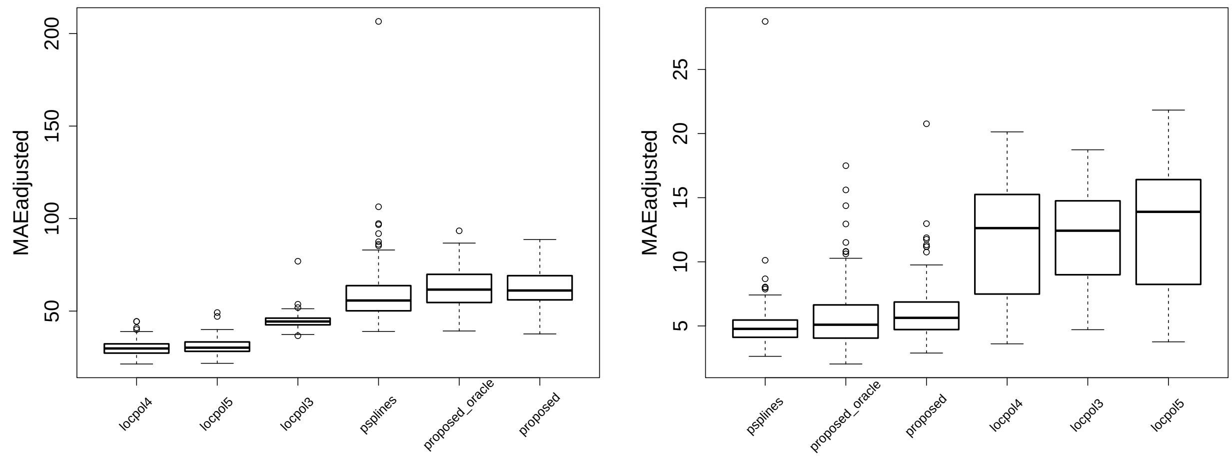
(b) Results for simulation study (17).

Figure 3: Comparative boxplots of the proposed first-order derivative estimator and all the methods that we compare under the Monte Carlo simulation studies (16) and (17). We order the boxplots in each panel according to the increasing order of the average values of their adjusted MAEs. (This figure is extended from Figure 5 and Figure 6(b) in the paper.)

tially selected from the set $\{0.03, 0.035, ..., 0.07\}$ through a local cubic regression with the bimodal Gaussian kernel and corrected for the unimodal Gaussian kernel as in Section 2.4. The paper also considered another Monte Carlo simulation with a nonuniform distribution of X under model (1) as:

$$m(X) = X + 2\exp\left(-16X^2\right) \quad \text{with} \quad X \sim N(0, 0.5^2) \quad \text{and} \quad e \sim N(0, 0.2^2), \tag{17}$$

where the sample size is again $n = 700$ and the data generating process is repeated for 100 times. The initial bandwidth in this case is selected from the set $\{0.04, 0.045, ..., 0.08\}$ and corrected for a unimodal Gaussian kernel as well. We adopt the adjusted mean absolute error $\text{MAEadj} = \frac{1}{650}\sum_{i=26}^{675}\left|\widehat{m}^{(1)}(X_{(i)}) - m^{(1)}(X_{(i)})\right|$ without boundary points from the paper as an evaluation metric. Based on our offline experiments, this performance measure yields some similar comparative results to the scenarios when we use a more robust median absolute error metric. Figure 3 demonstrates that even with oracle knowledge about the distribution of covariate X, the proposed first-order derivative estimator is outperformed by local polynomial regression with $p = 4$ or penalized smoothing cubic splines in the above simulation settings (16) and (17) from the discussed paper. The reason why we can discover the cases when the proposed first-order derivative estimation method is outperformed by classical methods is that our experiments are more comprehensive than those in the paper.



(a) Results for simulation study (18).

(b) Results for simulation study (17).

Figure 4: Comparative boxplots of the proposed second-order derivative estimator and the methods that we compare under the Monte Carlo simulation studies (18) and (17). We order the boxplots in each panel according to the increasing order of the average values of their adjusted MAEs. (This figure is extended from Figure 11 and Figure 6(b) in the paper.)

5.2 Simulation Studies on the Second-Order Derivative Estimation

As in the discussed paper, we simulate data sets of size $n = 700$ from model (1) with the function

$$m(X) = 8e^{-(1-5X)^3(1-7X)} \quad \text{with} \quad X \sim \text{Unif}(0,1) \quad \text{and} \quad e \sim N(0, 0.1^2) \tag{18}$$

for 100 times. The tuning parameters k_1, k_2 are selected via Corollary 6 over the product integer set $\{1, ..., 100\} \otimes \{1, ..., 100\}$. The initial bandwidth h of the proposed estimator (12) is selected from the set $\{0.03, 0.035, ..., 0.1\}$ through a local cubic regression with the bimodal Gaussian kernel and then corrected for the unimodal Gaussian kernel as in Section 2.4. Given that the paper did not consider the nonuniform distributional design of X, we make an extension here by reusing the Monte Carlo simulation study (17) but estimating the second-order derivative of m. The performance measure is adopted from the paper as another adjusted mean absolute error MAEadj $= \frac{1}{640}\sum_{i=31}^{670} \left|\widehat{m}^{(2)}(X_{(i)}) - m^{(2)}(X_{(i)})\right|$. Figure 4 presents the comparative boxplots of various second-order derivative estimation methods under these two simulation studies, where we remove the results of Gasser-Müller and Nadaraya-Watson derivative estimators due to their inferior performances. Different from what the paper claimed, the proposed second-order derivative again behaves worse than the local polynomial regression of some certain order and penalized smoothing cubic splines.

6 Discussion

The discussed paper (Liu and De Brabanter, 2020) proposes a data-driven method for derivative estimation under the random design by combining the weighted difference quotients with local polynomial regression and studies the asymptotic properties of the proposed derivative estimators. While the proposed derivative estimation framework is not novel given the previous works (De Brabanter et al., 2013, 2018), the theoretical analysis under the random design in the paper has its own merit, especially with our complementary results (Theorem 8). Nevertheless, our reproducing and extensive simulation studies demonstrate that the proposed first and second-order derivative estimators are outperformed by other classical derivative estimation methods under the simulation settings of the paper. Furthermore, we record the elapsed time for each derivative estimation method in the Monte Carlo simulation study (16) in Figure 5, where the proposed method is less computationally efficient than other methods due to its time required to select tuning parameters.

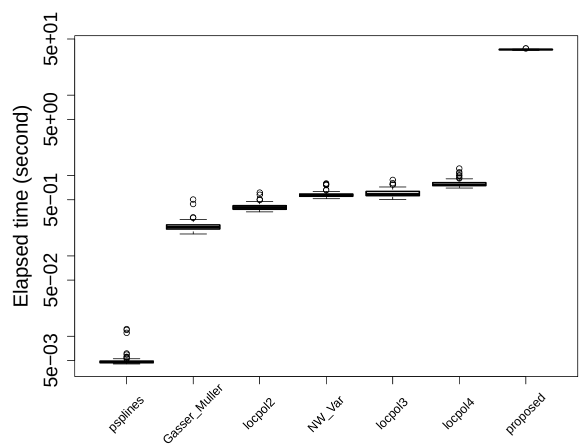


Figure 5: Time comparisons of different first-order derivative estimation methods.

One promising direction for improving the accuracy of the proposed method is to smooth the observed sample $\{(X_i, Y_i)\}_{i=1}^n$ first by penalized smoothing splines (see Simulation 8 in Appendix A.3) or local random forests (Dang, 2021, 2022) before taking weighted difference quotients as noisy derivative estimators. Studying the asymptotic properties of such a pre-smoothed derivative estimator will be of research interest. In addition, while Dang (2021) already considered generalizing the proposed method to estimating (partial) derivatives of a multivariate regression function under the random design, a strong independence assumption between covariates in its multivariate probability integral transformation step was imposed. In general, when the covariates are dependent, it is not true that the CDF of the covariate vector $\boldsymbol{X} \in \mathbb{R}^d$ is uniformly distributed on $[0, 1]$. Nor is it possible to reconstruct the distribution of $\boldsymbol{X}$ through its joint CDF (Genest and Rivest, 2001). To tackle this multivariate derivative estimation problem, one may resort to the associated copula (Nelsen, 2007) in order to model the dependence structure between $\boldsymbol{X}$ and characterize its entire distribution.

Acknowledgement

I would like to thank Professor Thomas Richardson for his detailed and insightful comments on this report and my progress towards the preliminary exam throughout the entire quarter. I also acknowledge Professor Marina Meilă, Professor Yen-Chi Chen, Professor Alexander Giessing, Steven Wilkins-Reeves, Apara Venkat, and members of the Geometric Data Analysis reading group at the University of Washington for their comments and suggestions to my presentation.

References

D. Belkić and K. Belkić. Validation of reconstructed component spectra from non-parametric derivative envelopes: comparison with component lineshapes from parametric derivative estimations with the solved quantification problem. *Journal of Mathematical Chemistry*, 56:2537–2578, 2018.

S. Calonico, M. D. Cattaneo, and M. H. Farrell. On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113(522): 767–779, 2018.

G. Casella and R. Berger. *Statistical Inference*. Duxbury advanced series. Thomson Learning, 2nd edition, 2002.

J. E. Chacón, T. Duong, and M. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, pages 807–840, 2011.

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.

Y.-C. Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

G. Cheng and Y.-C. Chen. Nonparametric inference via bootstrapping the debiased estimator. *Electronic Journal of Statistics*, 13(1):2194 – 2256, 2019.

W. Dai, T. Tong, and M. G. Genton. Optimal estimation of derivatives in nonparametric regression. *Journal of Machine Learning Research*, 17(1):5700–5724, 2016.

J. Dang. Smoothed nonparametric derivative estimation using random forest based weighted difference quotients. 2021. URL https://economics.ucr.edu/wp-content/uploads/2021/10/JustinDangJMP.pdf.

J. Dang. *Machine Learning Estimation of Nonparametric Econometric Models and Marginal Effects*. PhD thesis, University of California, Riverside, 2022.

C. de Boor. On uniform approximation by splines. *Journal of Approximation Theory*, 1(2):219–235, 1968.

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

K. De Brabanter, F. Cao, 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. 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. *ks: Kernel Smoothing*, 2022. URL https://CRAN.R-project.org/package=ks. R package version 1.14.0 [Online; accessed 3-April-2023].

U. Einmahl and D. M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005.

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*, volume 66. Chapman & Hall/CRC, 1996.

M. Francisco-Fernández, J. M. Vilar-Fernández, and J. A. Vilar-Fernández. On the uniform strong consistency of local polynomial regression under dependence conditions. *Communications in Statistics-Theory and Methods*, 32(12):2415–2440, 2003.

T. Gasser and H.-G. Müller. Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation: Proceedings of a Workshop Held in Heidelberg, April 2–4, 1979*, pages 23–68. Springer, 1979.

T. Gasser and H.-G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian journal of statistics*, pages 171–185, 1984.

C. Genest and L.-P. Rivest. On the multivariate probability integral transformation. *Statistics & probability letters*, 53(4):391–399, 2001.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511 – 1545, 2014.

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.

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907–921, 2002.

D. J. Gorsich and M. G. Genton. Variogram model selection via nonparametric derivative estimation. *Mathematical geology*, 32:249–270, 2000.

T. Haavelmo. Methods of measuring the marginal propensity to consume. *Journal of the American Statistical Association*, 42(237):105–122, 1947.

P. Hall. Large sample optimality of least squares cross-validation in density estimation. *The Annals of Statistics*, pages 1156–1174, 1983.

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

W. Härdle. *Applied nonparametric regression*. Number 19. Cambridge university press, 1990.

W. Härdle and T. Gasser. On robust kernel estimation of derivatives of regression functions. *Scandinavian journal of statistics*, pages 233–240, 1985.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, second edition, 2009.

A. Iserles. *A first course in the numerical analysis of differential equations*. Number 44. Cambridge university press, 2009.

R. Jarrow, D. Ruppert, and Y. Yu. Estimating the interest rate term structure of corporate debt with a semiparametric penalized spline model. *Journal of the American Statistical Association*, 99(465):57–66, 2004.

S. Lang. *Analysis I*. Addison-Wesley series in mathematics. Addison-Wesley Publishing Company, 1968.

Y. Liu and K. De Brabanter. Derivative estimation in random design. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Y. Liu and K. De Brabanter. Smoothed nonparametric derivative estimation using weighted difference quotients. *Journal of Machine Learning Research*, 21(1):2438–2482, 2020.

Y. Mack and H.-G. Müller. Derivative estimation in nonparametric regression with random predictor variable. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 59–72, 1989.

E. Masry. Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and their Applications*, 65(1):81–101, 1996.

H.-G. Müller. *Nonparametric regression analysis of longitudinal data*. Springer-Verlag, 1988.

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

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

R. B. Nelsen. *An introduction to copulas*. Springer science & business media, 2007.

D. Nolan and D. Pollard. U-processes: Rates of convergence. *The Annals of Statistics*, 15(2): 780–799, 1987.

J. L. Ojeda Cabrera. *locpol: Kernel Local Polynomial Regression*, 2022. URL https://CRAN.R-project.org/package=locpol. R package version 0.8.0 [Online; accessed 3-April-2023].

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 and B. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer Series in Statistics. Springer New York, 2002.

J. Ramsey and B. Ripley. *pspline: Penalized Smoothing Splines*, 2022. URL https://CRAN.R-project.org/package=pspline. R package version 1.0-19 [Online; accessed 3-April-2023].

A. Reinhart, L. Brooks, M. Jahja, A. Rumack, J. Tang, S. Agrawal, W. Al Saeed, T. Arnold, A. Basu, J. Bien, et al. An open repository of real-time covid-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51):e2111452118, 2021.

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(none):268 – 289, 2007.

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

W. Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3): 683–690, 1991.

B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, pages 1348–1360, 1980.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

C. J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2): 689–705, 1985.

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. W. Wang and L. Lin. Derivative estimation based on difference sequence via locally weighted least squares regression. *Journal of Machine Learning Research*, 16(1):2617–2641, 2015.

L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

H. J. Woltring. On optimal smoothing and derivative estimation from noisy displacement data in biomechanics. *Human Movement Science*, 4(3):229–245, 1985.

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.

Summary of the Appendix:

- Appendix A: We reproduce all the simulation studies and figures in the discussed paper (Liu and De Brabanter, 2020) and showcase a real-world application of the proposed first-order derivative estimation methods to Washington state-level COVID-19 case rates.
- Appendix B: We provide all the proofs of theorems and theoretical results in the main report.

A Other Reproducing Simulation Studies

This section presents all the simulation studies and figures in the paper (Liu and De Brabanter, 2020) that have yet been discussed in Section 5 of the main report. Since the authors of the discussed paper did not make any code publicly available, I programmed the proposed derivative estimation methods as well as all the simulation studies by myself. Although the smoothed estimators $\widehat{r}^{(1)}$ and $\widehat{r}^{(2)}$ are constructed with noisy derivative data for interior points in Theorem 3 and Theorem 7, the paper also includes the noisy derivative data at the boundary to obtain these smoothed derivative estimators by the local polynomial regression in its simulation studies. *I noticed that only including the interior noisy derivative data for the local polynomial regression as suggested by Theorem 3 and Theorem 7 makes our reproducing figures look more similar to the original ones in the paper, and the proposed derivative estimation methods have lower estimation errors. Thus, we will use this strategy for our reproducing simulation studies in this report.* In addition, it is worth mentioning that exactly reproducing some figures without any knowledge about the random seeds for simulations is almost impossible, so there may be some tiny discrepancies between the original figures in the paper and the ones that I reproduce as follows.

A.1 Simulation Studies on the First-Order Derivative Estimation

For all the simulation studies in this section, the density f and CDF F of the covariate X are estimated by the `R` functions `kde` and `kcde` with default parameters in the `R` package `ks` (Duong, 2022). The tuning parameter k is selected via Corollary 2 over the positive integer set $\left\{1, 2, ..., \lfloor \frac{n-1}{2} \rfloor\right\}$ unless otherwise stated. We apply the local cubic regression ($p = 3$) with the bimodal kernel $\bar{K}(u) = \frac{2u^2}{\sqrt{\pi}} \exp\left(-u^2\right)$ to initially select the bandwidth parameter from a set and correct it for a unimodal Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$ as described in Section 2.4.

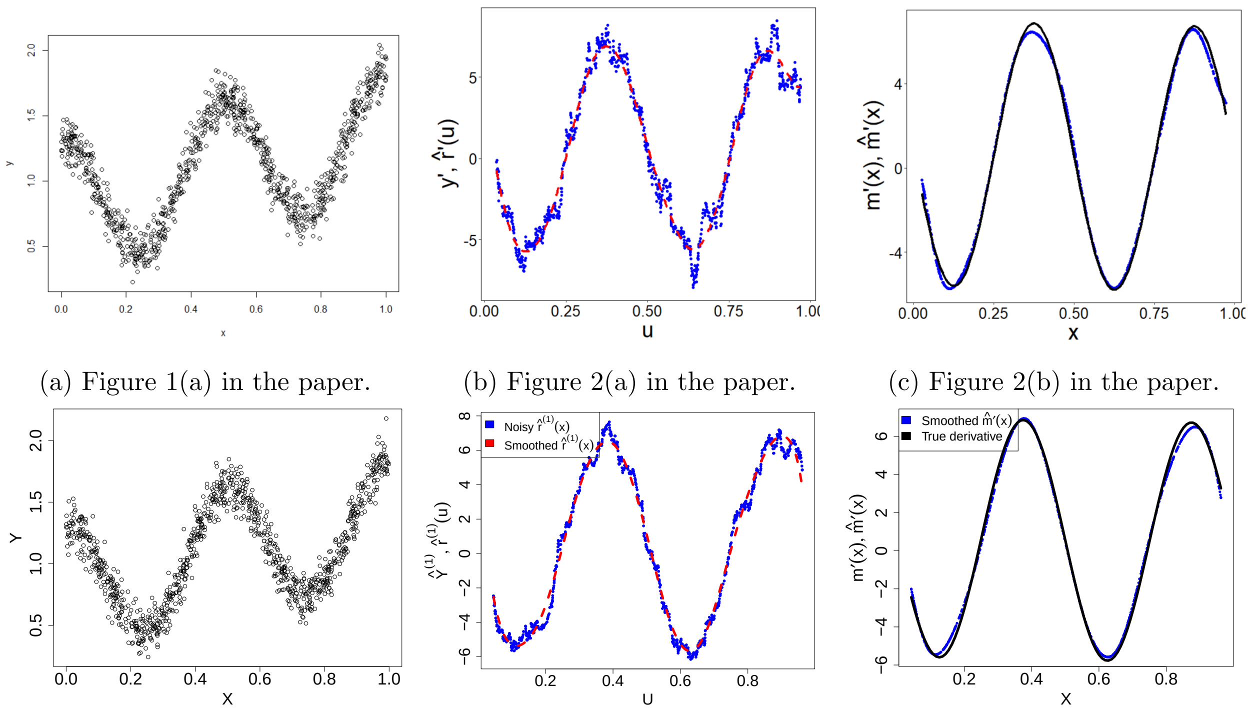
- **Simulation 1:** The first simulation study in the paper (Liu and De Brabanter, 2020) has been presented in Figure 1 of the main report, in which the simulated observations $\{(X_i, Y_i)\}_{i=1}^n$ with $n = 1000$ is sampled from model (1) with

$$m(X) = \cos^2(2\pi X) + \log(4/3 + X) \quad \text{for} \quad X \sim \text{Unif}(0,1) \tag{19}$$

and $e \sim N(0, 0.1^2)$, whose true first-order derivative is $m^{(1)}(X) = -2\pi \sin(4\pi X) + \frac{3}{3X+4}$. The initial bandwidth for the proposed derivative estimator is selected from the set $\{0.04, 0.045, ..., 0.1\}$. We compare our reproducing figures with the original ones in the paper in Figure 6 as well.

- **Simulation 2:** The second simulation study in the paper (Liu and De Brabanter, 2020) considers estimating the first-order derivative when the true distribution of X is no longer uniform on $[0, 1]$. In this case, the simulated observations $\{(X_i, Y_i)\}_{i=1}^n$ with $n = 1000$ is sampled from model (1) with

$$m(X) = 50e^{-8(1-2X)^4}(1 - 2X) \quad \text{for} \quad X \sim \text{Beta}(2,2) \tag{20}$$



(a) Figure 1(a) in the paper.

(b) Figure 2(a) in the paper.

(c) Figure 2(b) in the paper.

(d) Simulated observations.

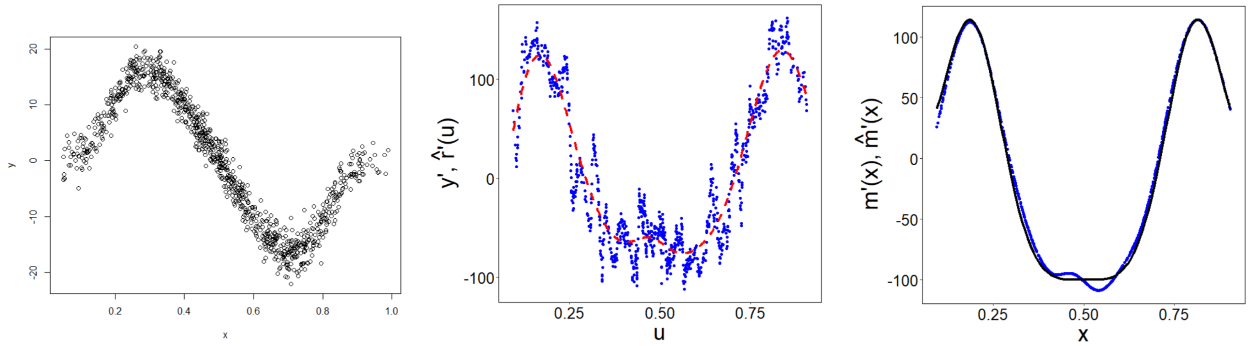
(e) First-order noisy derivatives with chosen $k = 37$ and the smoothed estimates on the transformed space $[0, 1]$.

(f) The proposed smoothed derivative estimates back-transformed to the original space with the true derivative.

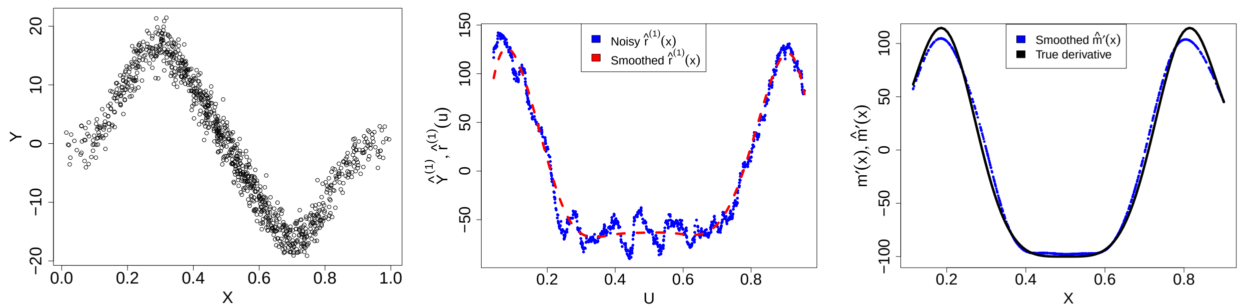
Figure 6: **Reproducing Figure 1(a) and Figure 2 in the paper:** Simulated data $\{(X_i, Y_i)\}_{i=1}^{1000}$ from model (1) under (19) with the first-order noisy derivatives and the proposed smoothed derivative estimates. The first row contains figures in the original paper, while the second row presents our reproduced figures.

and $e \sim N(0, 2^2)$, whose true first-order derivative is $m^{(1)}(X) = 100\left[32(1-2X)^4 - 1\right] e^{-8(1-2X)^4}$. The initial bandwidth for the proposed derivative estimator is selected from the set $\{0.04, 0.045, ..., 0.1\}$. We compare our reproducing figures with the original ones in the paper in Figure 7.

• **Simulation 3:** The third simulation study in the paper (Liu and De Brabanter, 2020) is a Monte Carlo repeated simulation study described in (16) of the main report, where the authors only compared the proposed first-order derivative estimator with the local slope of the local polynomial regression of order $p = 2$ that is implemented in `R` package `locpol` (Ojeda Cabrera, 2022) and the first-order derivative of the penalized smoothing cubic spline that is implemented in `R` package `pspline` (Ramsey and Ripley, 2022). The tuning parameter k is chosen via Corollary 2 and the initial bandwidth is selected from the set $\{0.03, 0.035, ..., 0.07\}$ for each Monte Carlo simulated data



(a) Figure 1(b) in the paper. (b) Figure 3(a) in the paper. (c) Figure 3(b) in the paper.



(d) Simulated observations. (e) First-order noisy derivatives with chosen $k = 39$ and the smoothed estimates on the transformed space $[0, 1]$. (f) The proposed smoothed derivative estimates back-transformed to the original space with the true derivative.

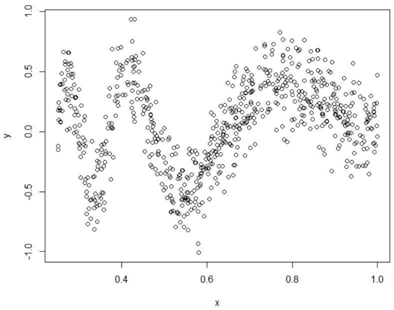
Figure 7: **Reproducing Figure 1(b) and Figure 3 in the paper:** Simulated data $\{(X_i, Y_i)\}_{i=1}^{1000}$ from model (1) under (20) with the first-order noisy derivatives and the proposed smoothed derivative estimates. The first row contains figures in the original paper, while the second row presents our reproduced figures.

set. In order to alleviate the boundary effects, the paper used an adjusted mean absolute error

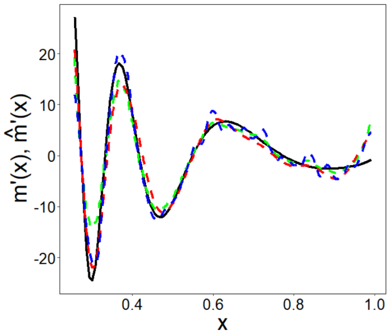
$$\text{MAEadj} = \frac{1}{650} \sum_{i=26}^{675} \left| \widehat{m}^{(1)}(X_{(i)}) - m^{(1)}(X_{(i)}) \right|$$

as an evaluation metric. We compare our reproducing figures with the original ones in the paper in Figure 8, in which we also consider plugging in the true density f and CDF F in the probability integral transform step of the proposed estimator to illuminate the loss of accuracy due to the estimation of f and F in Figure 8(c,f).

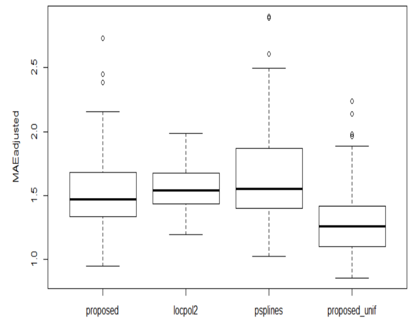
- **Simulation 4:** The fourth simulation study in the paper (Liu and De Brabanter, 2020) is another Monte Carlo repeated simulation study described in (17) of the main report, where the covariate X is no longer uniformly distributed. Again, the authors only compared the proposed



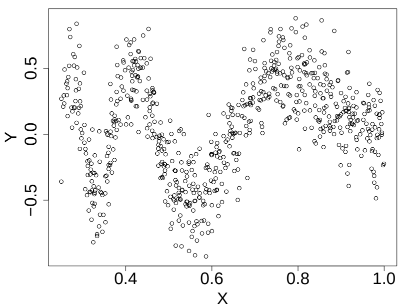
(a) Figure 4(a) in the paper.



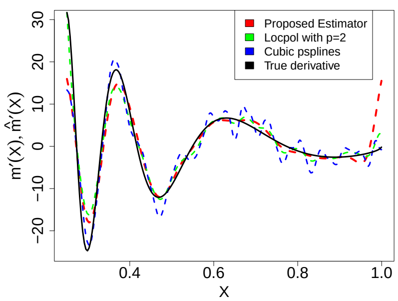
(b) Figure 4(b) in the paper.



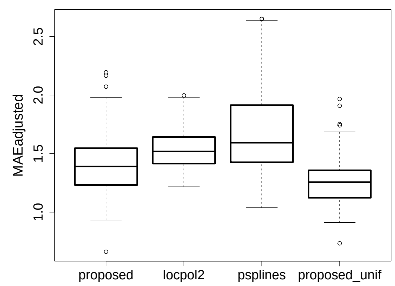
(c) Figure 5 in the paper.



(d) Simulated observations from one random run of model (16).



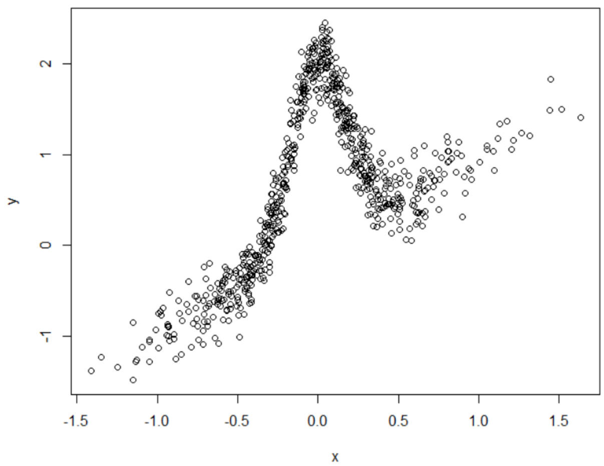
(e) Proposed derivative estimator with chosen $k = 10$, "locpol2", "psplines", and the true derivative.



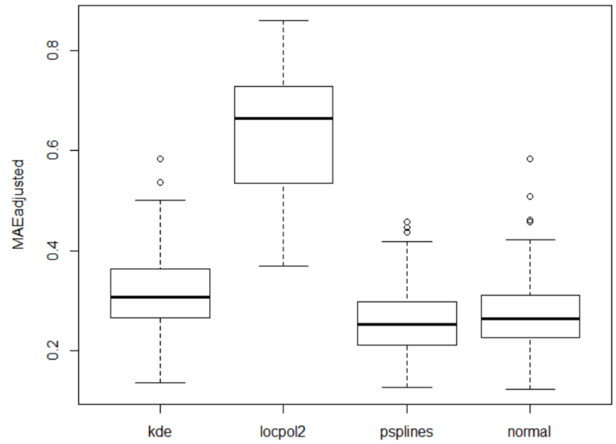
(f) Adjusted mean absolute errors under 100 Monte Carlo repeated experiments.

Figure 8: **Reproducing Figures 4 and 5 in the paper:** Monte Carlo comparative studies from model (1) under (16) for the proposed first-order derivative estimator ("proposed"), the proposed first-order derivative estimator under the oracle distribution of X ("proposed_unif"), local polynomial regression estimator with $p = 2$ ("locpol2"), and penalized smoothing cubic spline estimator ("psplines"). The first row contains figures in the original paper, while the second row presents our reproduced figures.

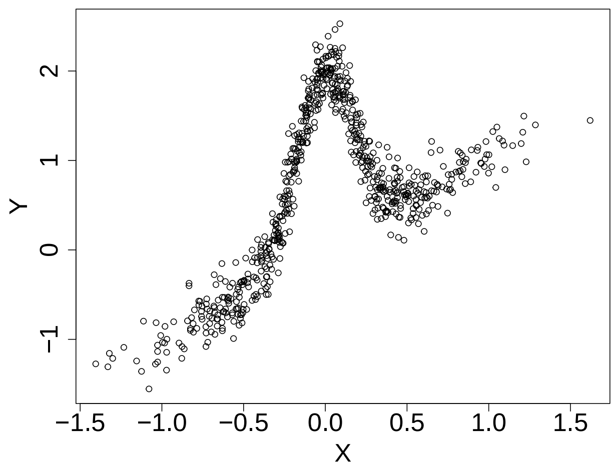
first-order derivative estimator with the local slope of the local polynomial regression of order $p = 2$ and the first-order derivative of the penalized smoothing cubic spline with respect to the adjusted mean absolute error. The initial bandwidth for the proposed derivative estimator is selected from the set $\{0.04, 0.045, ..., 0.08\}$. We compare our reproducing figures with the original ones in the paper in Figure 9, in which we also consider plugging in the true density f and CDF F in the probability integral transform step of the proposed estimator to illuminate the loss of accuracy due to the estimation of f and F in Figure 9(b,d).



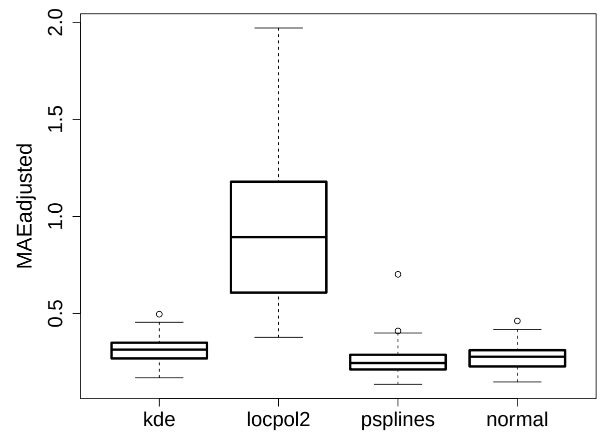
(a) Figure 6(a) in the paper.



(b) Figure 6(b) in the paper.



(c) Simulation observations from one random run of model (17).



(d) Adjusted mean absolute errors under 100 Monte Carlo repeated experiments.

Figure 9: **Reproducing Figure 6 in the paper:** Monte Carlo comparative studies from model (1) under (17) for the proposed first-order derivative estimator with KDE for the distribution of X ("kde"), the proposed first-order derivative estimator under the oracle distribution of X ("normal"), local polynomial regression estimator with $p = 2$ ("locpol2"), and and penalized smoothing cubic spline estimator ("psplines"). The first row contains figures in the original paper, while the second row presents our reproduced figures.

A.2 Simulation Studies on the Second-Order Derivative Estimation

Similar to the first-order derivative estimation, we estimate the density f, its derivative $f^{(1)}$, and CDF F of the covariate X through the `R` functions `kde`, `kdde`, and `kcde` with default parameters in the `R` package `ks` (Duong, 2022). The tuning parameters k_1, k_2 are selected via Corollary 6 over a product set of positive integers $\{1,2,...,100\} \otimes \{1,2,...,100\}$ unless otherwise stated. We apply the local cubic regression ($p=3$) with the bimodal kernel $\bar{K}(u) = \frac{2u^2}{\sqrt{\pi}} \exp\left(-u^2\right)$ to initially select the bandwidth parameter from a set and correct it for a unimodal Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$ as described in Section 2.4.

• **Simulation 5:** The fifth simulation study in the paper (Liu and De Brabanter, 2020) reuses the data-generating mechanism of (19) so that the true second-order derivative is $m^{(2)}(X) = -8\pi^2 \cos(4\pi X) - \frac{9}{(3X+4)^2}$. The initial bandwidth for the proposed derivative estimator is selected from the set $\{0.05, 0.055, ..., 0.1\}$. We compare our reproducing figures with the original ones in the paper in Figure 10.

• **Simulation 6:** The sixth simulation study in the paper (Liu and De Brabanter, 2020) generates $n = 1000$ observations $\{(X_i, Y_i)\}_{i=1}^n$ from model (1) with the function

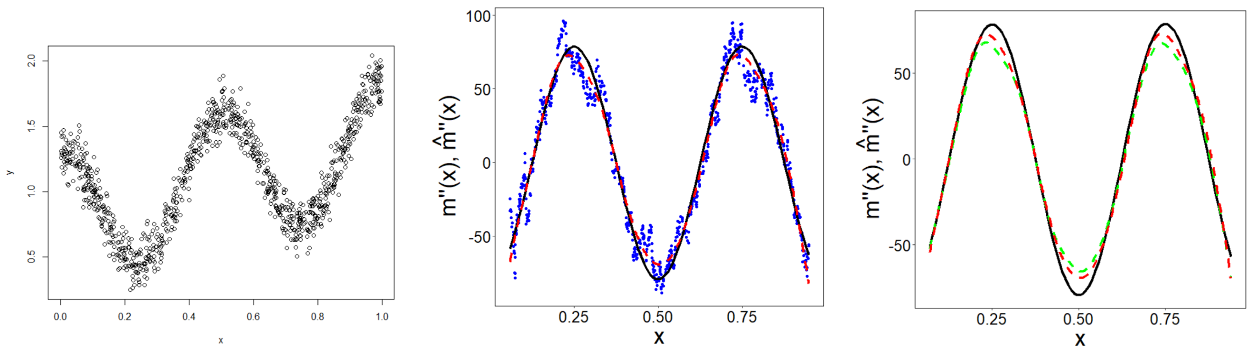
$$m(X) = 50e^{-8(1-2X)^4}(1-2X) \quad \text{for} \quad X \sim \text{Unif}[0,1] \quad \text{and} \quad e \sim N(0, 2^2), \tag{21}$$

whose true second-order derivative is $m^{(2)}(X) = 6400(1-2X)^3\left[32(1-2X)^3 - 5\right]e^{-8(1-2X)^4}$. The initial bandwidth for the proposed derivative estimator is again selected from the set $\{0.05, 0.055, ..., 0.1\}$. We compare our reproducing figures with the original ones in the paper in Figure 11.

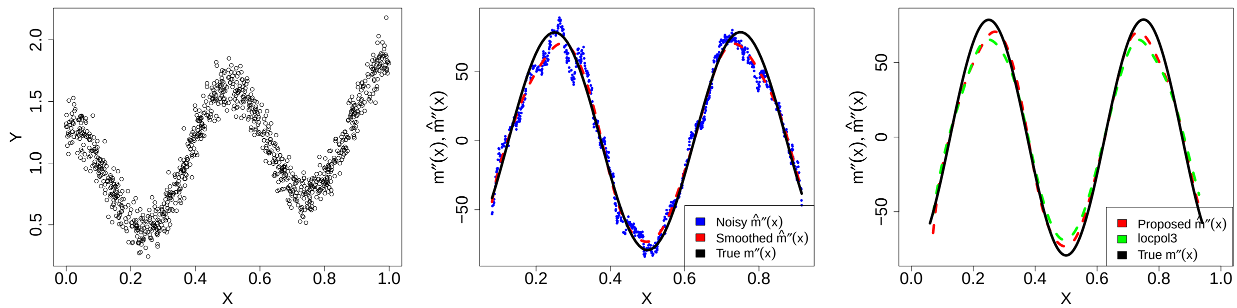
• **Simulation 7:** The seventh simulation study in the paper (Liu and De Brabanter, 2020) is a Monte Carlo repeated simulation study described in (18), where the authors only compared the proposed second-order derivative estimator with the local slope of the local polynomial regression of order $p = 3$ implemented in `R` package `locpol` (Ojeda Cabrera, 2022) and the second-order derivative of the penalized smoothing cubic spline implemented in `R` package `pspline` (Ramsey and Ripley, 2022). The tuning parameters k_1, k_2 are chosen via Corollary 6 and the initial bandwidth is selected from the set $\{0.03, 0.035, ..., 0.1\}$ for each Monte Carlo simulated data set. In order to alleviate the boundary effects, the paper again used an adjusted mean absolute error

$$\text{MAEadj} = \frac{1}{640} \sum_{i=31}^{670} \left|\widehat{m}^{(2)}(X_{(i)}) - m^{(2)}(X_{(i)})\right|$$

as an evaluation metric. We compare our reproducing figures with the original ones in the paper in Figure 12. Notice that there are some discrepancies between Figure 12(c) in the original paper and Figure 12(f) of our reproducing figure. We explored other choices of the tuning parameters, tried to include the noisy derivative data at the boundary, and did several rounds of code review for the proposed second-order derivative estimation method. None of these attempts help reproduce the original boxplot to the exact extent. We suspect that there might be some minor programming tricks in the proposed second-order derivative estimation method that may result in these discrepancies.



(a) Figure 7(a) in the paper. (b) Figure 8(a) in the paper. (c) Figure 9(a) in the paper.



(d) Simulated observations.

(e) Second-order noisy derivatives with chosen $k_1 = 32, k_2 = 46$, the smoothed estimates, and the true derivative.

(f) The proposed smoothed derivative estimates and the derivative of the local polynomial regression with $p = 3$.

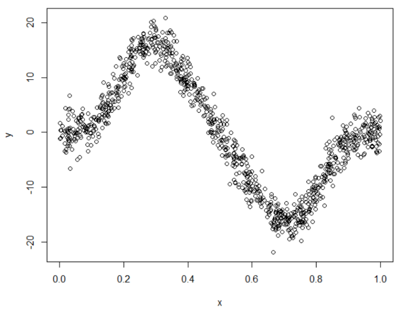
Figure 10: **Reproducing Figure 7(a), Figure 8(a), and Figure 9(a) in the paper:** Simulated data $\{(X_i, Y_i)\}_{i=1}^{1000}$ from model (1) under (19) with the second-order noisy derivatives, the proposed smoothed derivative estimates, and their comparisons with the local polynomial regression estimator with $p = 3$. The first row contains figures in the original paper, while the second row presents our reproduced figures.

A.3 Potential Approach for Improving the Estimation Accuracy

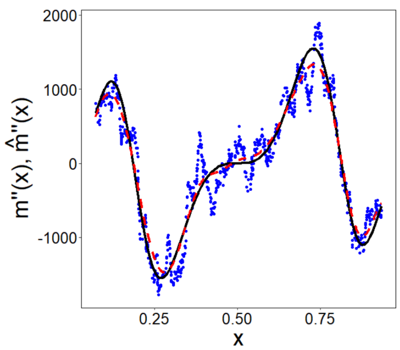
• **Simulation 8:** One reviewer of the paper suggested that it is possible to smooth out the data $\{(X_i, Y_i)\}_{i=1}^n$ by penalized smoothing splines[1] before taking the noisy derivatives. Specifically, we fit a penalized smoothing cubic splines $\widehat{m}$ on the data $\{(X_i, Y_i)\}_{i=1}^n$ and compute the difference quotients as:

$$\frac{\widehat{m}(X_{(i)}) - \widehat{m}(X_{(i-1)})}{X_{(i)} - X_{(i-1)}} \approx \widehat{m}^{(1)}(\xi) \tag{22}$$

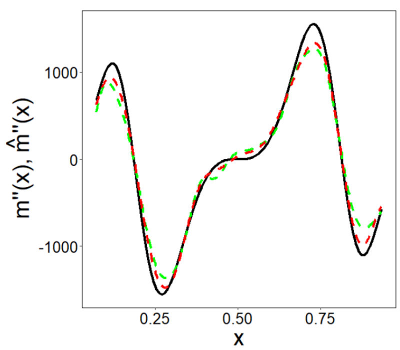
[1] The paper claimed that they smoothed out the data via the adaptive splines. However, based on our reproducing work, it is more realistic that the author was smoothing out the data via penalized smoothing cubic splines.



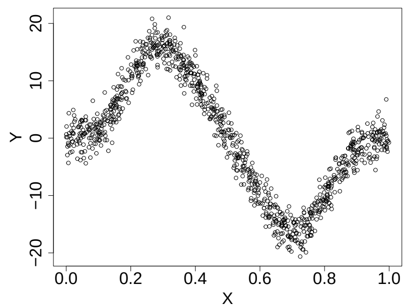
(a) Figure 7(b) in the paper.



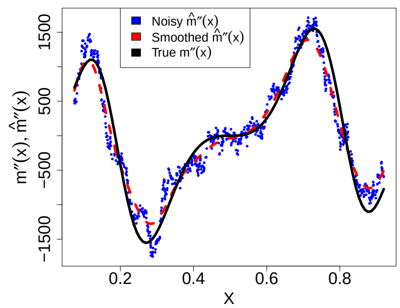
(b) Figure 8(b) in the paper.



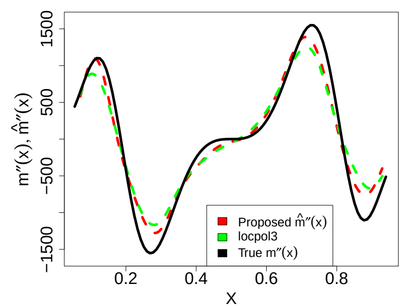
(c) Figure 9(b) in the paper.



(d) Simulated observations.



(e) Second-order noisy derivatives with chosen $k_1 = 30, k_2 = 42$, the smoothed estimates, and the true derivative.



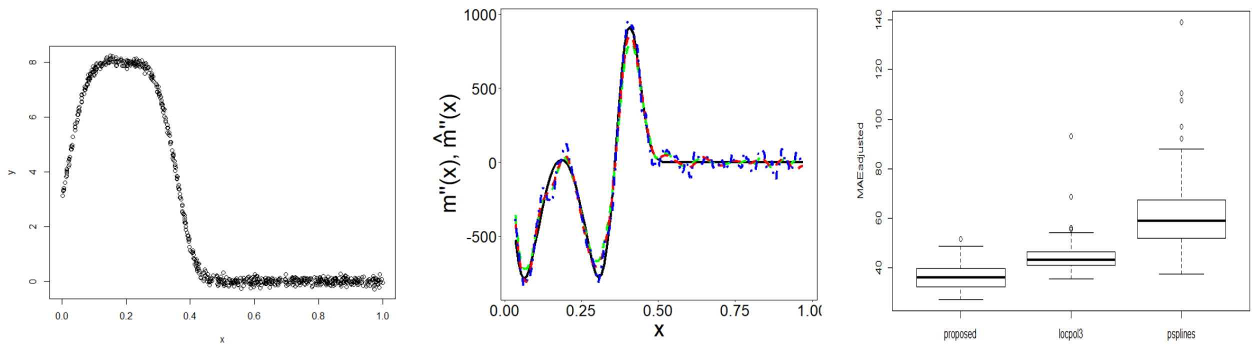
(f) The proposed smoothed derivative estimates and the derivative of the local polynomial regression with $p = 3$.

Figure 11: **Reproducing Figure 7(b), Figure 8(b), and Figure 9(b) in the paper:** Simulated data $\{(X_i, Y_i)\}_{i=1}^{1000}$ from model (1) under (21) with the second-order noisy derivatives, the proposed smoothed derivative estimates, and their comparisons with the local polynomial regression estimator with $p = 3$. The first row contains figures in the original paper, while the second row presents our reproduced figures.

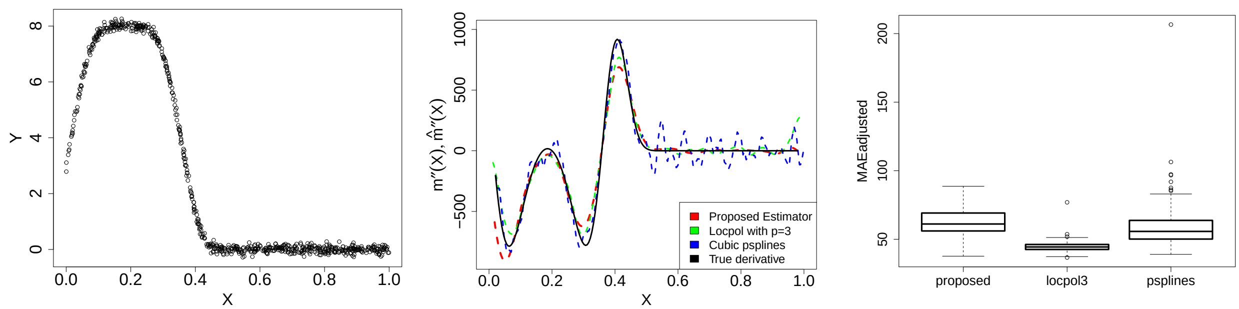
with $\xi \in \left[X_{(i-1)}, X_{(i)}\right]$. We conduct a Monte Carlo simulation study from model (1) under (19) and compare our results with the ones in the paper in Figure 13.

A.4 Case Study: Washington State-Level COVID-19 Case Rates

Although the discussed paper did not incorporate any real-world application of their proposed methods, we found a potentially interesting real-world application of the derivative estimation on an academic website `https://cmu-delphi.github.io/covidcast/modeltoolsR/articles/estimate-deriv.html`. Here, we consider estimating the Washington state-level COVID-19 case



(a) Figure 10(a) in the paper. (b) Figure 10(b) in the paper. (c) Figure 11 in the paper.



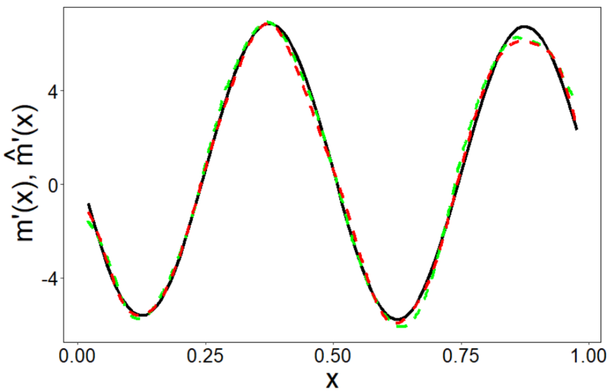
(d) Simulated observations from one random run of (18).

(e) Second-order noisy derivatives with chosen $k_1 = 30, k_2 = 42$, the smoothed estimates, and the true derivative.

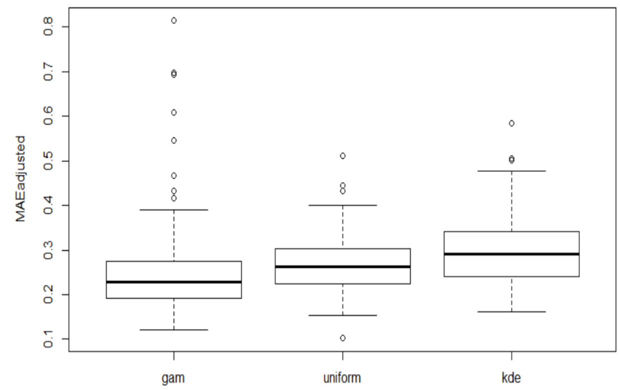
(f) The proposed smoothed derivative estimates and the derivative of the local polynomial regression with $p = 3$.

Figure 12: **Reproducing Figures 10 and 11 in the paper:** Monte Carlo comparative studies from model (1) under (18) for the proposed second-order derivative estimator, the local polynomial regression estimator with $p = 3$ ("locpol3"), and and penalized smoothing cubic spline estimator ("psplines"). The first row contains figures in the original paper, while the second row presents our reproduced figures.

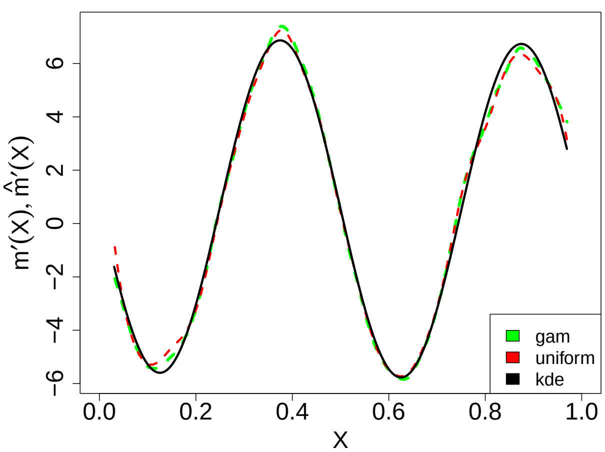
rates at the early stage of the pandemic according to the USAFacts data stored in `R` package `covidcast` (Reinhart et al., 2021). We restrict the studied dates to the range from "2020-05-01" to "2020-12-15" at the Washington State, which capture both the initial surge of COVID cases and the first spike during the winter of 2020; see Figure 14(a). We apply the proposed first-order derivative estimator, local polynomial regression of order $p = 2$, and penalized smoothing cubic splines to estimating the case rates within this selected period. The tuning parameter k is selected via Corollary 2 over the positive integer set $\left\{1, 2, ..., \lfloor \frac{n-1}{2} \rfloor\right\}$, where n is the number of dates in this context. The initial bandwidth for the proposed derivative estimator is selected from the



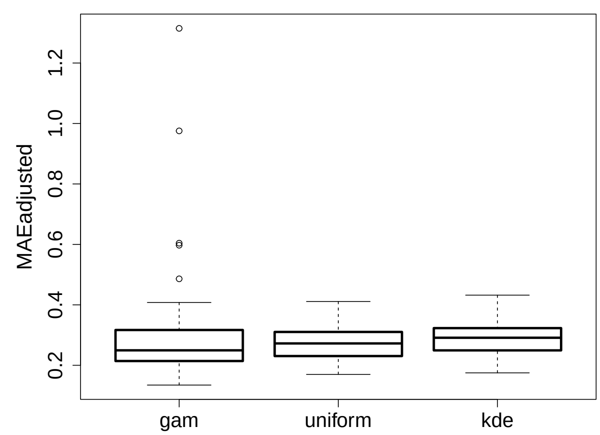
(a) Figure 12(a) in the paper.



(b) Figure 12(b) in the paper.



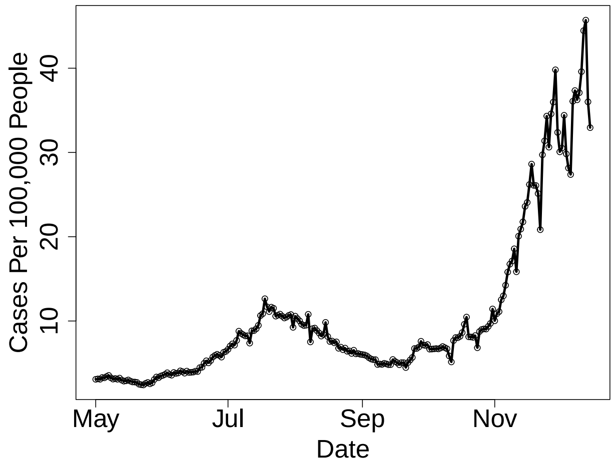
(c) Comparative results from one random run of model (19).



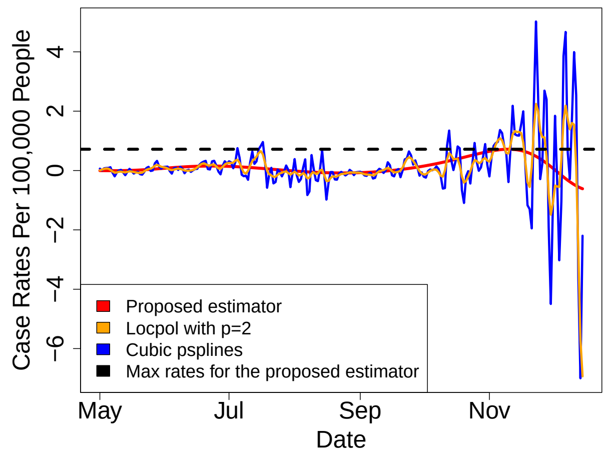
(d) Adjusted mean absolute errors under 100 Monte Carlo repeated experiments.

Figure 13: **Reproducing Figure 12 in the paper:** Monte Carlo comparative studies from model (1) under (19) for the proposed first-order derivative estimator with KDE for the distribution of X ("kde"), the proposed first-order derivative estimator under the oracle distribution of X ("uniform"), and the pre-smoothing approach described in (22) ("gam"). The first row contains figures in the original paper, while the second row presents our reproduced figures.

set $\{0.001, 0.002, ..., 0.2\}$. Figure 14 displays the estimated COVID-19 case rates by these three derivative estimation methods, where the proposed derivative estimator with data-driven tuning parameters produce the smoothest estimates. Moreover, the estimated change rates of COVID-19 new cases at the Washington state by the proposed estimator never exceed 71.78%. Compared with the actual reported cases, however, it is obvious that the increasing rates of COVID-19 cases in November 2020 should be much higher than this number. To some extent, it suggests that the



(a) Reported COVID-19 cases.



(b) Estimated case rates for different methods.

Figure 14: Reported COVID-19 cases at the Washington state between "2020-05-01" and "2020-12-15" as well as the estimated case rates by the proposed first-order derivative estimator ("Proposed estimator"), local polynomial regression of order $p = 2$ ("Locpol with $p = 2$"), and penalized smoothing cubic splines ("Cubic psplines").

proposed derivative estimators are not quite applicable to the analysis of COVID-19 case rates compared with the other two methods, especially because these rates change rapidly across time and we need more sensible derivative estimators to capture this rapid changing trend for disease control and political decision making.

B Proofs

This section supplements the proofs of theorems and theoretical results in the main report. Different from the discussed paper (Liu and De Brabanter, 2020), I will give a short summary for each proof, fill in more details for the proofs, and rewrite/correct some arguments according to our own understanding. In addition, all the remarks after the proofs are inspired by or extended from the discussed paper.

B.1 Proof of Theorem 1

Theorem 1 (Theorem 1 in Liu and De Brabanter 2020). *Assume that r is twice continuously differentiable on $[0, 1]$ under model (7). Then, the conditional bias and variance of the first-order* *noisy derivative estimator* ([8](#)) *given* $\mathbb{U} = \left(U_{(i-j)}, ..., U_{(i+j)}\right)$ *for* $i > j$ *and* $i + j \leq n$ *are*

$$\left|\text{Bias}\left[\widehat{Y}_i^{(1)} | \mathbb{U}\right]\right| \leq \sup_{u\in[0,1]} \left|r^{(2)}(u)\right| \frac{3k(k+1)}{4(n+1)(2k+1)} + o_P\left(\frac{k}{n}\right),$$
$$\text{Var}\left[\widehat{Y}_i^{(1)} | \mathbb{U}\right] = \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} + o_P\left(\frac{n^2}{k^3}\right)$$

uniformly for $k+1 \leq i \leq n-k$ *when* $k \to \infty$ *as* $n \to \infty$*. Further, if we assume that* r *is* $q+1$ *times continuously differentiable on* $[0,1]$ *for* $q \geq 1$*, then the asymptotic order of the exact conditional bias is given by*

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

Proof of [Theorem 1](#). **Summary of the Proof:** The proof follows by applying Taylor's theorem over r in model ([7](#)) and using Lemma [9](#) in the calculations of $\left|\text{Bias}\left[\widehat{Y}_i^{(1)} | \mathbb{U}\right]\right|$ and $\text{Var}\left[\widehat{Y}_i^{(1)} | \mathbb{U}\right]$.

Since r is twice continuously differentiable on $[0,1]$, we have the following Taylor's expansions of $r(U_{(i+j)})$ and $r(U_{(i-j)})$ in the neighborhood of $U_{(i)}$ as:

$$\begin{aligned} r(U_{(i+j)}) &= r(U_{(i)}) + \left(U_{(i+j)} - U_{(i)}\right) r^{(1)}(U_{(i)}) + \frac{\left(U_{(i+j)} - U_{(i)}\right)^2}{2} \cdot r^{(2)}(\zeta_{i,i+j}), \\ r(U_{(i-j)}) &= r(U_{(i)}) + \left(U_{(i-j)} - U_{(i)}\right) r^{(1)}(U_{(i)}) + \frac{\left(U_{(i-j)} - U_{(i)}\right)^2}{2} \cdot r^{(2)}(\zeta_{i-j,i}), \end{aligned} \tag{23}$$

where $\zeta_{i,i+j} \in \left[U_{(i)}, U_{(i+j)}\right]$ and $\zeta_{i-j,i} \in \left[U_{(i-j)}, U_{(i)}\right]$. The absolute conditional bias is bounded above by

$$\begin{aligned} &\left|\text{Bias}\left[\widehat{Y}_i^{(1)} | \mathbb{U}\right]\right| \\ &= \left|\mathbb{E}\left[\sum_{j=1}^{k} w_{i,j}\left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right) \Big| \mathbb{U}\right] - r^{(1)}(U_{(i)})\right| \\ &\overset{\text{(i)}}{=} \frac{1}{2}\left|\sum_{j=1}^{k} w_{i,j}\left[\frac{(U_{(i+j)} - U_{(i)})^2 r^{(2)}(\zeta_{i,i+j}) - (U_{(i-j)} - U_{(i)})^2 r^{(2)}(\zeta_{i-j,i})}{U_{(i+j)} - U_{(i-j)}}\right]\right| \\ &\overset{\text{(ii)}}{=} \frac{1}{2}\left|\frac{\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})\left[(U_{(i+j)} - U_{(i)})^2 \cdot r^{(2)}(\zeta_{i,i+j}) - (U_{(i-j)} - U_{(i)})^2 \cdot r^{(2)}(\zeta_{i-j,i})\right]}{\sum_{\ell=1}^{k}\left(U_{(i+\ell)} - U_{(i-\ell)}\right)^2}\right| \\ &\leq \frac{1}{2}\sup_{u\in[0,1]}\left|r^{(2)}(u)\right| \cdot \frac{\sum_{j=1}^{k}(U_{(i+j)} - U_{(i-j)})\left[(U_{(i+j)} - U_{(i)})^2 + (U_{(i-j)} - U_{(i)})^2\right]}{\sum_{\ell=1}^{k}\left(U_{(i+\ell)} - U_{(i-\ell)}\right)^2} \\ &\overset{\text{(iii)}}{=} \frac{1}{2}\sup_{u\in[0,1]}\left|r^{(2)}(u)\right| \cdot \frac{\frac{k^2(k+1)^2}{(n+1)^3}\left[1 + O_P\left(\frac{1}{\sqrt{k}}\right)\right]}{\frac{2k(k+1)(2k+1)}{3(n+1)^2}\left[1 + O_P\left(\frac{1}{\sqrt{k}}\right)\right]} \end{aligned}$$

$$= \sup_{u\in[0,1]} \left|r^{(2)}(u)\right| \frac{3k(k+1)}{4(n+1)(2k+1)} \left[1 + O_P\left(\frac{1}{\sqrt{k}}\right)\right],$$

where (i) follows from (23), (ii) is due to Proposition 10, and (iii) is based on Lemma 9 and the following calculations as:

$$\begin{aligned}
\sum_{\ell=1}^{k} \left(U_{(i+\ell)} - U_{(i-\ell)}\right)^2 &= \sum_{\ell=1}^{k} \left[\frac{2\ell}{n+1} + O_P\left(\sqrt{\frac{\ell}{n^2}}\right)\right]^2 \\
&= \frac{4}{(n+1)^2} \cdot \frac{k(k+1)(2k+1)}{6} + \frac{4}{n+1} \sum_{\ell=1}^{k} \ell \cdot O_P\left(\sqrt{\frac{\ell}{n^2}}\right) + \sum_{\ell=1}^{k} O_P\left(\frac{\ell}{n^2}\right) \\
&= \frac{2k(k+1)(2k+1)}{3(n+1)^2} \left[1 + O_P\left(\sqrt{\frac{1}{k}}\right)\right]
\end{aligned} \tag{24}$$

and

$$\begin{aligned}
&\sum_{j=1}^{k} (U_{(i+j)} - U_{(i-j)}) \left[(U_{(i+j)} - U_{(i)})^2 + (U_{(i-j)} - U_{(i)})^2\right] \\
&= \sum_{j=1}^{k} \left[\frac{2j}{n+1} + O_P\left(\sqrt{\frac{j}{n^2}}\right)\right] \left\{\left[\frac{j}{n+1} + O_P\left(\sqrt{\frac{j}{n^2}}\right)\right]^2 + \left[\frac{j}{n+1} + O_P\left(\sqrt{\frac{j}{n^2}}\right)\right]^2\right\} \\
&= \sum_{j=1}^{k} \frac{4j^3}{(n+1)^3} \left[1 + O_P\left(\sqrt{\frac{1}{k}}\right)\right] \\
&= \frac{k^2(k+1)^2}{(n+1)^3} \left[1 + O_P\left(\sqrt{\frac{1}{k}}\right)\right].
\end{aligned}$$

Thus, for $k \to \infty$ as $n \to \infty$,

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

Similarly, by Proposition 10, the conditional variance is

$$\begin{aligned}
\text{Var}\left[\widehat{Y}_i^{(1)} | \mathbb{U}\right] &= \text{Var}\left[\sum_{j=1}^{k} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right) \Big| \mathbb{U}\right] \\
&= 2\sigma_e^2 \cdot \frac{\sum_{j=1}^{k} \left(U_{(i+j)} - U_{(i-j)}\right)^2}{\left(\sum_{\ell=1}^{n} (U_{(i+\ell)} - U_{(i-\ell)})^2\right)^2} \\
&= \frac{2\sigma_e^2}{\sum_{\ell=1}^{n} (U_{(i+\ell)} - U_{(i-\ell)})^2} \\
&\overset{\text{(iv)}}{=} \frac{2\sigma_e^2}{\frac{2k(k+1)(2k+1)}{3(n+1)^2} \left[1 + O_P\left(\sqrt{\frac{1}{k}}\right)\right]}
\end{aligned}$$

$$= \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} [1 + o_P(1)]$$

when $k \to \infty$ as $n \to \infty$, where we leverage our calculation (24) in (iv). In addition, the above results hold uniformly for $k+1 \le i \le n-k$.

Now, if r is $q+1$ times continuously differentiable on $[0,1]$ for $q \ge 1$, then applying Taylor's theorem and Lemma 9 to $r(U_{(i+j)})$ and $r(U_{(i-j)})$ in a neighborhood of $U_{(i)}$ yields that

$$\begin{aligned} r(U_{(i+j)}) &= r(U_{(i)}) + \sum_{\ell=1}^{q+1} \frac{r^{(\ell)}(U_{(i)})}{\ell!} \left(U_{(i+j)} - U_{(i)}\right)^\ell + O\left(|U_{(i+j)} - U_{(i)}|^{q+2}\right) \\ &= r(U_{(i)}) + \sum_{\ell=1}^{q+1} \frac{r^{(\ell)}(U_{(i)})}{\ell!} \left(U_{(i+j)} - U_{(i)}\right)^\ell + O_P\left(\left(\frac{j}{n}\right)^{q+2}\right) \end{aligned}$$

and

$$\begin{aligned} r(U_{(i-j)}) &= r(U_{(i)}) + \sum_{\ell=1}^{q+1} \frac{r^{(\ell)}(U_{(i)})}{\ell!} \left(U_{(i-j)} - U_{(i)}\right)^\ell + O\left(|U_{(i-j)} - U_{(i)}|^{q+2}\right) \\ &= r(U_{(i)}) + \sum_{\ell=1}^{q+1} \frac{r^{(\ell)}(U_{(i)})}{\ell!} \left(U_{(i-j)} - U_{(i)}\right)^\ell + O_P\left(\left(\frac{j}{n}\right)^{q+2}\right). \end{aligned}$$

For $q=1$, $r^{(2)}$ exists and is continuous on $[0,1]$, so the conditional bias becomes

$$\begin{aligned} \text{Bias}\left[\widehat{Y}_i^{(1)} | \mathbb{U}\right] &= \frac{r^{(1)}(U_{(i)}) \sum_{j=1}^k \left(U_{(i+j)} - U_{(i-j)}\right)^2 + O_P\left(\frac{k^4}{n^3}\right)}{\sum_{\ell=1}^k \left(U_{(i+\ell)} - U_{(i-\ell)}\right)^2} - r^{(1)}(U_{(i)}) \\ &= O_P\left(\frac{k}{n}\right), \end{aligned}$$

where we can calculate that $\sum_{j=1}^k O_P\left(\left(\frac{j}{n}\right)^{q+2}\right) = O_P\left(\frac{k^{q+3}}{n^{q+2}}\right)$.
For $q=2$, $r^{(3)}$ exists on $[0,1]$ and the conditional bias becomes

$$\begin{aligned} \text{Bias}\left[\widehat{Y}_i^{(1)} | \mathbb{U}\right] &= \frac{r^{(2)}(U_{(i)}) \sum_{j=1}^k \left(U_{(i+j)} - U_{(i-j)}\right) \left[\left(U_{(i+j)} - U_{(i)}\right)^2 - \left(U_{(i-j)} - U_{(i)}\right)^2\right] + O_P\left(\frac{k^5}{n^4}\right)}{2\sum_{\ell=1}^k \left(U_{(i+\ell)} - U_{(i-\ell)}\right)^2} \\ &= \frac{O_P\left(\frac{k^{\frac{7}{2}}}{n^3}\right) + O_P\left(\frac{k^5}{n^4}\right)}{O_P\left(\frac{k^3}{n^2}\right)} \\ &= O_P\left(\max\left\{\frac{k^{\frac{1}{2}}}{n}, \frac{k^2}{n^2}\right\}\right). \end{aligned}$$

For $q > 2$, we can split the conditional bias into even order terms in the Taylor's expansion of $r(U_{(i\pm j)})$ and odd order terms respectively as:

$$
\begin{aligned}
\text{Bias}\left[\widehat{Y}_i^{(1)}\middle|\mathbb{U}\right] &= \frac{\sum_{j=1}^{k}\left(U_{(i+j)}-U_{(i-j)}\right)\left[\sum\limits_{\ell=3,5,\ldots,2\lceil q/2\rceil-1}\frac{r^{(\ell)}(U_{(i)})}{\ell!}\left(\left(U_{(i+j)}-U_{(i)}\right)^{\ell}-\left(U_{(i-j)}-U_{(i)}\right)^{\ell}\right)\right]}{\sum_{p=1}^{k}\left(U_{(i+p)}-U_{(i-p)}\right)^2} \\
&+\frac{\sum_{j=1}^{k}\left(U_{(i+j)}-U_{(i-j)}\right)\left[\sum\limits_{\ell=2,4,\ldots,2\lceil q/2\rceil}\frac{r^{(\ell)}(U_{(i)})}{\ell!}\left(\left(U_{(i+j)}-U_{(i)}\right)^{\ell}-\left(U_{(i-j)}-U_{(i)}\right)^{\ell}\right)\right]}{\sum_{p=1}^{k}\left(U_{(i+p)}-U_{(i-p)}\right)^2} \\
&= \frac{\sum_{j=1}^{k}\left(U_{(i+j)}-U_{(i-j)}\right)\left[\frac{r^{(3)}(U_{(i)})}{6}\left(\left(U_{(i+j)}-U_{(i)}\right)^{3}-\left(U_{(i-j)}-U_{(i)}\right)^{3}\right)\right]}{\sum_{p=1}^{k}\left(U_{(i+p)}-U_{(i-p)}\right)^2}[1+o_P(1)] \\
&+\frac{\sum_{j=1}^{k}\left(U_{(i+j)}-U_{(i-j)}\right)\left[\frac{r^{(2)}(U_{(i)})}{2}\left(\left(U_{(i+j)}-U_{(i)}\right)^{3}-\left(U_{(i-j)}-U_{(i)}\right)^{3}\right)\right]}{\sum_{p=1}^{k}\left(U_{(i+p)}-U_{(i-p)}\right)^2}[1+o_P(1)] \\
&= \frac{\sum_{j=1}^{k}\left[\frac{2j}{n+1}+O_P\left(\sqrt{\frac{j}{n^2}}\right)\right]\frac{r^{(3)}(U_{(i)})}{6}\left[\frac{j^3}{(n+1)^3}+O_P\left(\frac{j^{\frac{5}{2}}}{n^3}\right)\right]}{O_P\left(\frac{k^3}{n^2}\right)} \\
&+\frac{\sum_{j=1}^{k}\left[\frac{2j}{n+1}+O_P\left(\sqrt{\frac{j}{n^2}}\right)\right]\frac{r^{(2)}(U_{(i)})}{2}\cdot O_P\left(\frac{j^{\frac{3}{2}}}{n}\right)}{O_P\left(\frac{k^3}{n^2}\right)} \\
&= O_P\left(\frac{k^2}{n^2}\right)+O_P\left(\frac{k^{\frac{3}{2}}}{n^2}\right)+O_P\left(\frac{k^{\frac{1}{2}}}{n}\right) \\
&= O_P\left(\max\left\{\frac{k^{\frac{1}{2}}}{n},\frac{k^2}{n^2}\right\}\right).
\end{aligned}
$$

The proof is thus completed. □

B.2 Proof of Corollary 2

Corollary 2 (Corollary 2 in Liu and De Brabanter 2020)**.** *Let $\mathcal{B} = \sup_{u\in[0,1]}\left|r^{(2)}(u)\right|$. Under the assumptions of Theorem 1, the tuning parameter k that minimizes the asymptotic upper bound of the conditional MISE is given by*

$$
k_{opt} = \underset{k=1,2,\ldots,\lfloor\frac{n-1}{2}\rfloor}{\arg\min}\left[\mathcal{B}^2\frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2}+\frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}\right].
$$

Proof of Corollary 2. **Summary of the Proof:** The proof follows directly from the definition of the conditional mean integrated squared error (MISE) and the bias-variance decomposition in

Theorem 1.

By the bias-variance decomposition in Theorem 1 of the conditional mean squared error, we have that

$$\mathbb{E}\left[\left(\widehat{Y}^{(1)}(U) - r^{(1)}(U)\right)^2 \middle| \mathbb{U}\right] \leq \mathcal{B}^2 \frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2} + \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} + o_P\left(\frac{k^2}{n^2} + \frac{n^2}{k^3}\right).$$

Since $U \sim \mathrm{Unif}[0,1]$, the conditional mean integrated squared error (MISE) is given by

$$\begin{aligned}
\mathrm{MISE}\left[\widehat{Y}^{(1)}|\mathbb{U}\right] &= \mathbb{E}\left\{\int_0^1 \left[\widehat{Y}^{(1)}(U) - r^{(1)}(U)\right]^2 dU \,\middle|\, \mathbb{U}\right\} \\
&= \int_0^1 \mathbb{E}\left[\left(\widehat{Y}^{(1)}(U) - r^{(1)}(U)\right)^2 \middle| \mathbb{U}\right] dU \\
&\leq \mathcal{B}^2 \frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2} + \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} + o_P\left(\frac{k^2}{n^2} + \frac{n^2}{k^3}\right),
\end{aligned}$$

where $\widehat{Y}^{(1)}(U)$ denotes the first-order derivative estimator at design point U and the first two terms comprise the upper bound of the (asymptotic) conditional MISE. $\square$

Remark 2. Selecting the optimal tuning parameter as the minimizer of the asymptotic MISE (*i.e.*, through the bias-variance trade-off; Wasserman 2006) is a common approach in Statistics. For instance, all those well-known bandwidth selection methods in kernel density estimation, such as Silverman's rule of thumb (Pages 45 and 47 in Silverman 1986), least-square cross validation (Hall, 1983), and plug-in method (Section 3.6 in Wand and Jones 1994), are based on the rationale of minimizing the (asymptotic) MISE in different ways. In our context of choosing the number k of weighted difference quotients in (8), we can also solve for k_{opt} by minimizing the leading terms in Corollary 2 as:

$$\begin{aligned}
k_{\mathrm{opt}} &= \underset{k=1,2,\ldots}{\arg\min} \left[\mathcal{B}^2 \frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2} + \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}\right] \\
&\approx \underset{k=1,2,\ldots}{\arg\min} \left[\mathcal{B}^2 \frac{9k^2}{64n^2} + \frac{3\sigma_e^2 n^2}{2k^3}\right] \\
&= \lfloor 2\sigma_e^{\frac{2}{5}} \mathcal{B}^{-\frac{2}{5}} n^{\frac{4}{5}} \rfloor,
\end{aligned}$$

where the unknown quantities $\mathcal{B}$ and σ_e^2 can be estimated according to our suggestions in Section 3.1.

B.3 Proof of Theorem 3

Theorem 3 (Theorem 2 in Liu and De Brabanter 2020)**.** *Assume that $r(\cdot)$ under model (7) is $(p+2)$ times continuously differentiable in a neighborhood of u_0. Under Assumptions 1 and 2, the* *conditional bias and variance of* (11) *with* $u_0 \in [0,1]$ *for* p *odd are*

$$\begin{aligned}
\text{Bias}\left[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\right] &\leq \left[\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} c_p \cdot \frac{r^{(p+2)}(u_0)}{(p+1)!} \cdot h^{p+1} + \left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right| \widetilde{c}_p \cdot \frac{3k(k+1)\mathcal{B}}{4(n+1)(2k+1)}\right] [1+o_P(1)] \\
&= \left[\left(\int t^{p+1} K_0^\star(t) dt\right) \frac{r^{(p+2)}(u_0)}{(p+1)!} \cdot h^{p+1} + \left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right| \widetilde{c}_p \cdot \frac{3k(k+1)\mathcal{B}}{4(n+1)(2k+1)}\right] [1+o_P(1)], \\
\text{Var}\left[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\right] &= \frac{3\sigma_e^2(n+1)^2(1+\rho_c)}{k(k+1)(2k+1)(n-2k)h} \cdot \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} \boldsymbol{S}^* \boldsymbol{S}^{-1} \boldsymbol{\epsilon}_1 [1+o_P(1)] \\
&= \left(\int K_0^\star(t)^2 dt\right) \frac{3\sigma_e^2(n+1)^2(1+\rho_c)}{k(k+1)(2k+1)(n-2k)h} [1+o_P(1)]
\end{aligned}$$

as $h \to 0, nh \to \infty, k \to \infty$ *with* $n \to \infty$, *where* $\widetilde{\mathbb{U}} = \left(U_{(1)}, ..., U_{(n)}\right)$, $\mathcal{B} = \sup_{u \in [0,1]} \left|r^{(2)}(u)\right|$, $\boldsymbol{S} = (\mu_{i+j-2})_{1 \leq i,j \leq p+1}$ *with* $\mu_j = \int u^j K(u) du$, $\boldsymbol{S}^* = (\nu_{i+j-2})_{1 \leq i,j \leq p+1}$ *with* $\nu_j = \int u^j K(u)^2 du$, $c_p = (\mu_{p+1}, ..., \mu_{2p+1})^T$, $\widetilde{c}_p = (\widetilde{\mu}_0, ..., \widetilde{\mu}_p)^T$ *with* $\widetilde{\mu}_j = \int |u|^j K(u) du$, $\boldsymbol{\epsilon}_1 = (1, 0, ..., 0)^T \in \mathbb{R}^{p+1}$, $\left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right|$ *means elementwise absolute values of* $\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}$, *and the equivalent kernel* $K_0^\star(t) = \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} (1, t, ..., t^p)^T K(t)$.

Proof of Theorem 3. **Summary of the Proof:** The proof follows from the arguments of Theorem 3.1 in Fan and Gijbels (1996) (see its Section 3.7) and the results of Theorem 1 in De Brabanter et al. (2018). In particular, we derive the asymptotic expression for each entry of the matrix $\boldsymbol{S}_{n-2k} \equiv \boldsymbol{S}_{u_0}$ in (11) and utilize the identity

$$(A + hB)^{-1} = A^{-1} - hA^{-1}BA^{-1} + O(h^2)$$

to handle the asymptotic expression of $\boldsymbol{S}_{n-2k}^{-1}$. Notice that there is an incorrect inequality (Eq.(31) on Page 34 of the discussed paper) in the original proof of this theorem in Liu and De Brabanter (2020). I fix this minor mistake in the following argument by slightly changing the statement of Theorem 3 here.

- **Conditional variance:** Recall from Theorem 1 that when $k \to \infty$ as $n \to \infty$,

$$\text{Var}\left[\widehat{Y}_i^{(1)}|\widetilde{\mathbb{U}}\right] = \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} [1+o_P(1)].$$

By Theorem 1 in De Brabanter et al. (2018) and the definition of $\widehat{r}^{(1)}(u_0)$ in (11), we have that

$$\begin{aligned}
\text{Var}\left[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}\right] &= \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k}^{-1} \left(\boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \cdot \text{Var}\left[\widehat{\boldsymbol{Y}}^{(1)}|\widetilde{\mathbb{U}}\right] \cdot \boldsymbol{W}_{u_0} \boldsymbol{U}_{u_0}\right) \boldsymbol{S}_{n-2k}^{-1} \boldsymbol{\epsilon}_1 \\
&= \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} \cdot \frac{1+f(u_0)\rho_c}{h(n-2k)f(u_0)} \cdot \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} \boldsymbol{S}^* \boldsymbol{S}^{-1} \boldsymbol{\epsilon}_1 [1+o_P(1)] \\
&= \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} \cdot \frac{1+\rho_c}{h(n-2k)} \cdot \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} \boldsymbol{S}^* \boldsymbol{S}^{-1} \boldsymbol{\epsilon}_1 [1+o_P(1)]
\end{aligned}$$

under Assumptions 1 and 2, where $\boldsymbol{S}^* = (\nu_{i+j-2})_{1\le i,j\le p+1}$ with $\nu_j = \int u^j K(u)^2 du$ and $f(u_0) = 1$ for any $u_0 \in [0,1]$. By the definition of the equivalent kernel (see also Section 3.2.2 in Fan and Gijbels 1996), one can also write

$$\begin{aligned}
\mathrm{Var}\left[\widehat{r}^{(1)}(u_0)\middle|\widetilde{\mathbb{U}}\right] &= \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} \cdot \frac{1+\rho_c}{h(n-2k)} \cdot \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\boldsymbol{S}^*\boldsymbol{S}^{-1}\boldsymbol{\epsilon}_1 \left[1+o_P(1)\right] \\
&= \left(\int K_0^\star(t)^2 dt\right) \frac{3\sigma_e^2(n+1)^2(1+\rho_c)}{k(k+1)(2k+1)(n-2k)h} \left[1+o_P(1)\right],
\end{aligned}$$

given that $\int K_0^\star(t)^2 dt = \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\boldsymbol{S}^*\boldsymbol{S}^{-1}\boldsymbol{\epsilon}_1$.

- **Conditional bias:** Notice from (11) that

$$\begin{aligned}
\mathrm{Bias}\left[\widehat{r}^{(1)}(u_0)\middle|\widetilde{\mathbb{U}}\right] &= \mathbb{E}\left[\widehat{r}^{(1)}(u_0)\middle|\widetilde{\mathbb{U}}\right] - r^{(1)}(u_0) \\
&= \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \cdot \mathbb{E}\left[\widehat{\boldsymbol{Y}}^{(1)}\middle|\widetilde{\mathbb{U}}\right] - r^{(1)}(u_0) \\
&= \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \left(\begin{bmatrix} r^{(1)}(U_{(k+1)}) \\ \vdots \\ r^{(1)}(U_{(n-k)}) \end{bmatrix} + \begin{bmatrix} \mathrm{Bias}\left[\widehat{Y}_{k+1}^{(1)}\middle|\widetilde{\mathbb{U}}\right] \\ \vdots \\ \mathrm{Bias}\left[\widehat{Y}_{n-k}^{(1)}\middle|\widetilde{\mathbb{U}}\right] \end{bmatrix} \right) - r^{(1)}(u_0) \\
&= \underbrace{\boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \begin{bmatrix} r^{(1)}(U_{(k+1)}) \\ \vdots \\ r^{(1)}(U_{(n-k)}) \end{bmatrix} - r^{(1)}(u_0)}_{\text{Term I}} + \underbrace{\boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \begin{bmatrix} \mathrm{Bias}\left[\widehat{Y}_{k+1}^{(1)}\middle|\widetilde{\mathbb{U}}\right] \\ \vdots \\ \mathrm{Bias}\left[\widehat{Y}_{n-k}^{(1)}\middle|\widetilde{\mathbb{U}}\right] \end{bmatrix}}_{\text{Term II}}.
\end{aligned} \tag{25}$$

By direct calculations,

$$\boldsymbol{S}_{n-2k} = \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \boldsymbol{U}_{u_0} = \begin{bmatrix} S_{n-2k,0} & S_{n-2k,1} & \cdots & S_{n-2k,p} \\ S_{n-2k,1} & S_{n-2k,2} & \cdots & S_{n-2k,p+1} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n-2k,p} & S_{n-2k,p+1} & \cdots & S_{n-2k,2p} \end{bmatrix},$$

where

$$S_{n-2k,\ell} = \sum_{m=k+1}^{n-k} \left(U_{(m)} - u_0\right)^\ell K\left(\frac{U_{(m)} - u_0}{h}\right) = \mathbb{E}\left[S_{n-2k,\ell}\right] + O_P\left(\sqrt{\mathrm{Var}\left[S_{n-2k,\ell}\right]}\right)$$

for $\ell = 0,1,...,2p$. More importantly, $U_{(k+1)},...,U_{(n-k)}$ can be regarded as an i.i.d. sample when we sum over $k+1,...,n-k$ as in $S_{n-2k,\ell}$. Thus, when $h \to 0$ and $nh \to \infty$ as $n \to \infty$, we have that

$$\mathbb{E}\left[S_{n-2k,\ell}\right] = (n-2k) \cdot \mathbb{E}\left[(U-u_0)^\ell K\left(\frac{U-u_0}{h}\right)\right]$$

$$
\begin{aligned}
&= (n-2k)\int K\left(\frac{u-u_0}{h}\right)(u-u_0)^\ell f(u)du \\
&= (n-2k)h^{\ell+1}\int K(x)x^\ell f(u_0+hx)dx \\
&= (n-2k)h^{\ell+1}f(u_0)\left[\int x^\ell K(x)dx + O(h)\right] \\
&= (n-2k)h^{\ell+1}f(u_0)\mu^\ell\left[1+O(h)\right]
\end{aligned}
$$

and

$$
\begin{aligned}
O_P\left(\sqrt{\mathrm{Var}\left[S_{n-2k,\ell}\right]}\right) &= O_P\left(\sqrt{(n-2k)\cdot\mathbb{E}\left[(U-u_0)^{2\ell}\cdot K\left(\frac{U-u_0}{h}\right)^2\right]}\right) \\
&= O_P\left(\sqrt{(n-2k)\int(u-u_0)^{2\ell}K\left(\frac{u-u_0}{h}\right)^2 f(u)du}\right) \\
&= O_P\left(\sqrt{(n-2k)h^{2\ell+1}f(u_0)\int x^{2\ell}K^2(x)dx}\right) \\
&= O_P\left(\sqrt{(n-2k)h^{2\ell+1}}\right).
\end{aligned}
$$

These results imply that, for $h \to 0$, $nh \to \infty$, and $k \to \infty$ with $\frac{k}{n} \to 0$ as $n \to \infty$,

$$
S_{n-2k,\ell} = (n-2k)h^{\ell+1}f(u_0)\mu_\ell\left[1+O(h)+O_P\left(\sqrt{\frac{1}{(n-2k)h}}\right)\right]
$$

so that

$$
\boldsymbol{S}_{n-2k} = \begin{bmatrix} S_{n-2k,0} & S_{n-2k,1} & \cdots & S_{n-2k,p} \\ S_{n-2k,1} & S_{n-2k,2} & \cdots & S_{n-2k,p+1} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n-2k,p} & S_{n-2k,p+1} & \cdots & S_{n-2k,2p} \end{bmatrix} = (n-2k)hf(u_0)\boldsymbol{HSH}\left[1+o_P(1)\right],
$$

where $\boldsymbol{H} = \mathtt{Diag}\left(1,h,...,h^p\right)$ and $\boldsymbol{S} = \left(\mu_{i+j-2}\right)_{1\le i,j\le p+1}$ with $\mu_j = \int u^j K(u)du$.

Term I in (25)*:* When p is odd, we can directly leverage the results of Theorem 3.1 in Fan and Gijbels (1996) to obtain that

$$
\begin{aligned}
\text{Term I} &= \boldsymbol{\epsilon}_1^T\boldsymbol{S}_{n-2k}^{-1}\boldsymbol{U}_{u_0}^T\boldsymbol{W}_{u_0}\begin{bmatrix} r^{(1)}(U_{(k+1)}) \\ \vdots \\ r^{(1)}(U_{(n-k)}) \end{bmatrix} - r^{(1)}(u_0) \\
&= \boldsymbol{\epsilon}_1^T\boldsymbol{S}_{n-2k}^{-1}\boldsymbol{U}_{u_0}^T\boldsymbol{W}_{u_0}\left(\begin{bmatrix} r^{(1)}(U_{(k+1)}) \\ \vdots \\ r^{(1)}(U_{(n-k)}) \end{bmatrix} - \boldsymbol{U}_{u_0}\boldsymbol{\beta}_{u_0}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \begin{bmatrix} r^{(1)}(U_{(k+1)}) - \sum_{j=0}^{p} \left(U_{(k+1)} - u_0\right)^j \cdot \frac{r^{(j+1)}(u_0)}{j!} \\ \vdots \\ r^{(1)}(U_{(n-k)}) - \sum_{j=0}^{p} \left(U_{(n-k)} - u_0\right)^j \cdot \frac{r^{(j+1)}(u_0)}{j!} \end{bmatrix} \\
&\overset{\text{(i)}}{=} \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \begin{bmatrix} \frac{r^{(p+2)}(u_0)}{(p+1)!}(U_{(k+1)} - u_0)^{p+1} + o\left((U_{(k+1)} - u_0)^{p+1}\right) \\ \vdots \\ \frac{r^{(p+2)}(u_0)}{(p+1)!}(U_{(n-k)} - u_0)^{p+1} + o\left((U_{(n-k)} - u_0)^{p+1}\right) \end{bmatrix} \\
&\overset{\text{(ii)}}{=} \boldsymbol{\epsilon}_1^T \left\{(n-2k)hf(u_0)\boldsymbol{HSH}\left[1 + O(h)\right]\right\}^{-1} \cdot \frac{r^{(p+2)}(u_0)}{(p+1)!} \cdot (n-2k)h^{p+2}\boldsymbol{H}c_p\left[1 + O(h)\right] \\
&\overset{\text{(iii)}}{=} \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} c_p \cdot \frac{r^{(p+2)}(u_0)}{(p+1)!} \cdot h^{p+1} + o_P\left(h^{p+1}\right),
\end{aligned}
$$

where $\boldsymbol{\beta}_{u_0} = \left(r^{(1)}(u_0), ..., \frac{r^{(p+1)}(u_0)}{p!}\right)^T \in \mathbb{R}^{p+1}$ and $c_p = (\mu_{p+1}, ..., \mu_{2p+1})^T \in \mathbb{R}^{p+1}$. Here, we use the Taylor's expansion of $r^{(1)}$ around u_0 in equality (i), apply the asymptotic expressions for $\boldsymbol{S}_{n-2k}$ and $\boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0}$ in equality (ii), and utilize the identity $(A + hB)^{-1} = A^{-1} - hA^{-1}BA^{-1} + O(h^2)$ in equality (iii).

Term II: According to Theorem 1, we know that, for $k \to \infty$ as $n \to \infty$,

$$
\begin{aligned}
\text{Term II} &= \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \begin{bmatrix} \text{Bias}\left[\widehat{Y}_{k+1}^{(1)} | \widetilde{\mathbb{U}}\right] \\ \vdots \\ \text{Bias}\left[\widehat{Y}_{n-k}^{(1)} | \widetilde{\mathbb{U}}\right] \end{bmatrix} \\
&= \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k}^{-1} \begin{bmatrix} \sum_{m=k+1}^{n-k} \text{Bias}\left[\widehat{Y}_{m}^{(1)} | \widetilde{\mathbb{U}}\right] K\left(\frac{U_{(m)} - u_0}{h}\right) \\ \sum_{m=k+1}^{n-k} \text{Bias}\left[\widehat{Y}_{m}^{(1)} | \widetilde{\mathbb{U}}\right] \left(U_{(m)} - u_0\right) K\left(\frac{U_{(m)} - u_0}{h}\right) \\ \vdots \\ \sum_{m=k+1}^{n-k} \text{Bias}\left[\widehat{Y}_{m}^{(1)} | \widetilde{\mathbb{U}}\right] \left(U_{(m)} - u_0\right)^p K\left(\frac{U_{(m)} - u_0}{h}\right) \end{bmatrix} \\
&\le \left|\boldsymbol{\epsilon}_1^T \left\{(n-2k)hf(u_0)\boldsymbol{HSH}\left[1 + O(h)\right]\right\}^{-1}\right| (n-2k)h\boldsymbol{H}\widetilde{c}_p\left[1 + O(h)\right] \cdot \frac{3k(k+1)\mathcal{B}}{4(n+1)(2k+1)} \\
&= \left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right| \widetilde{c}_p \cdot \frac{3k(k+1)\mathcal{B}}{4(n+1)(2k+1)} \cdot \left[1 + o_P(1)\right]
\end{aligned}
$$

where $\mathcal{B} = \sup_{u \in [0,1]} \left|r^{(2)}(u)\right|$ and $\widetilde{c}_p = (\widetilde{\mu}_0, ..., \widetilde{\mu}_p)^T$ with $\widetilde{\mu}_j = \int |u|^j K(u) du$.

Combining *Term I* and *Term II* with (25) yields that

$$
\text{Bias}\left[\widehat{r}^{(1)}(u_0) | \widetilde{\mathbb{U}}\right] = \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \begin{bmatrix} r^{(1)}(U_{(k+1)}) \\ \vdots \\ r^{(1)}(U_{(n-k)}) \end{bmatrix} - r^{(1)}(u_0) + \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \begin{bmatrix} \text{Bias}\left[\widehat{Y}_{k+1}^{(1)} | \widetilde{\mathbb{U}}\right] \\ \vdots \\ \text{Bias}\left[\widehat{Y}_{n-k}^{(1)} | \widetilde{\mathbb{U}}\right] \end{bmatrix}
$$

$$\le \left[\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} c_p \cdot \frac{r^{(p+2)}(u_0)}{(p+1)!} \cdot h^{p+1} + \left| \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} \right| \widetilde{c}_p \cdot \frac{3k(k+1)\mathcal{B}}{4(n+1)(2k+1)} \right] [1 + o_P(1)]$$
$$= \left[\left(\int t^{p+1} K_0^\star(t) dt \right) \frac{r^{(p+2)}(u_0)}{(p+1)!} \cdot h^{p+1} + \left| \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} \right| \widetilde{c}_p \cdot \frac{3k(k+1)\mathcal{B}}{4(n+1)(2k+1)} \right] [1 + o_P(1)],$$

where $\int t^{p+1} K_0^\star(t) dt = \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} c_p$. The results follow. □

Remark 3. Under the assumptions of Theorem 3, the conditional mean integrated squared error (MISE) of the local polynomial regression estimator (11) is upper bounded by

$$\begin{aligned}
&\text{MISE}\left[\widehat{r}^{(1)} \middle| \widetilde{\mathbb{U}} \right] \\
&= \mathbb{E}\left\{ \int_0^1 \left[\widehat{r}^{(1)}(u_0) - r^{(1)}(u_0) \right]^2 du_0 \middle| \widetilde{\mathbb{U}} \right\} \\
&= \int_0^1 \mathbb{E}\left[\left(\widehat{r}^{(1)}(u_0) - r^{(1)}(u_0) \right)^2 \middle| \widetilde{\mathbb{U}} \right] du_0 \\
&\le \left[\left| \int t^{p+1} K_0^\star(t) dt \right| \frac{\sup_{u \in [0,1]} \left| r^{(p+2)}(u) \right|}{(p+1)!} \cdot h^{p+1} + \left| \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} \right| \widetilde{c}_p \cdot \frac{3k(k+1)\mathcal{B}}{4(n+1)(2k+1)} \right]^2 [1 + o_P(1)] \\
&\quad + \left(\int K_0^\star(t)^2 dt \right) \frac{3\sigma_e^2 (n+1)^2 (1+\rho_c)}{k(k+1)(2k+1)(n-2k)h} [1 + o_P(1)] \\
&= O_P\left(h^{2p+2} \right) + O_P\left(\frac{k h^{p+1}}{n} \right) + O_P\left(\frac{k^2}{n^2} \right) + O_P\left(\frac{n}{k^3 h} \right),
\end{aligned}$$

given that k is always of smaller order than $O(n)$. By taking the partial derivatives with respect to h and k and setting them to 0, we obtain a system of equations

$$\begin{cases} h^{2p+1} + \frac{k h^p}{n} - \frac{n}{k^3 h^2} \asymp 0, \\ \frac{h^{p+1}}{n} + \frac{k}{n^2} - \frac{n}{k^4 h} \asymp 0, \end{cases}$$

where we introduce the asymptotic equivalence symbol "$\asymp$" to get rid of all the constant factors. Solving this system of equations gives us that $k = O\left(n^{\frac{3p+4}{5p+6}} \right)$ and $h = O\left(n^{-\frac{2}{5p+6}} \right)$, which leading to an optimal rate of convergence for the upper bound of MISE $\left[\widehat{r}^{(1)} \middle| \widetilde{\mathbb{U}} \right]$ as $O_P\left(n^{-\frac{4p+4}{5p+6}} \right)$.

B.4 Proof of Theorem 5

Theorem 5 (Theorem 3 in Liu and De Brabanter 2020)**.** *Assume that r is three times continuously differentiable on $[0,1]$ under model (7). Then, under the weight $w_{ij,2} = \frac{(2j+k_1)^2}{\sum_{j=1}^{k_2}(2j+k_1)^2}$, the conditional bias and variance of the second-order noisy derivative estimator (10) given $\widetilde{\mathbb{U}} = \left(U_{(1)}, ..., U_{(n)} \right)$ are bounded by*

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

$$\mathrm{Var}\left[\widehat{Y}_i^{(2)}\big|\widetilde{\mathbb{U}}\right] \leq \frac{4(n+1)^4\sigma_e^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\cdot[1+o_P(1)]$$

uniformly for $k_1+k_2+1\leq i\leq n-k_1-k_2$ *when* $k_1,k_2\to\infty$ *as* $n\to\infty$.

Proof of Theorem 5. **Summary of the Proof:** The proof is similar to our arguments in Appendix B.1 for the proof of Theorem 1. In particular, we apply Taylor's theorem to the function r and utilize Lemma 9 to handle the asymptotic terms in $\left|\mathrm{Bias}\left[\widehat{Y}_i^{(2)}\big|\widetilde{\mathbb{U}}\right]\right|$ and $\mathrm{Var}\left[\widehat{Y}_i^{(2)}\big|\widetilde{\mathbb{U}}\right]$.

Since r is three times continuously differentiable on $[0,1]$, we have the following Taylor's expansions of $r(U_{(i+j+k_1)})$ and $r(U_{(i-j-k_1)})$ in the neighborhoods of $U_{(i+j)}$ and $U_{(i-j)}$ respectively as:

$$\begin{aligned}
r(U_{(i+j+k_1)}) &= \sum_{q=0}^{2}\frac{\left(U_{(i+j+k_1)}-U_{(i+j)}\right)^q}{q!}\cdot r^{(q)}(U_{(i+j)}) + \frac{\left(U_{(i+j+k_1)}-U_{(i+j)}\right)^3}{6}\cdot r^{(3)}(\zeta_{i+j,i+j+k_1}),\\
r(U_{(i-j-k_1)}) &= \sum_{q=0}^{2}\frac{\left(U_{(i-j-k_1)}-U_{(i-j)}\right)^q}{q!}\cdot r^{(q)}(U_{(i-j)}) + \frac{\left(U_{(i-j-k_1)}-U_{(i-j)}\right)^3}{6}\cdot r^{(3)}(\zeta_{i-j-k_1,i-j}),
\end{aligned} \tag{26}$$

where $\zeta_{i+j,i+j+k_1}\in\left[U_{(i+j)},U_{(i+j+k_1)}\right]$ and $\zeta_{i-j-k_1,i-j}\in\left[U_{(i-j-k_1)},U_{(i-j)}\right]$. In addition, the following Taylor's expansions of $r^{(1)}(U_{(i+j)})$ and $r^{(1)}(U_{(i-j)})$ are also valid in a neighborhood of $U_{(i)}$ as:

$$\begin{aligned}
r^{(1)}(U_{(i+j)}) &= r^{(1)}(U_{(i)}) + \left(U_{(i+j)}-U_{(i)}\right)r^{(2)}(U_{(i)}) + \frac{\left(U_{(i+j)}-U_{(i)}\right)^2}{2}\cdot r^{(3)}(\zeta_{i,i+j}),\\
r^{(1)}(U_{(i-j)}) &= r^{(1)}(U_{(i)}) + \left(U_{(i-j)}-U_{(i)}\right)r^{(2)}(U_{(i)}) + \frac{\left(U_{(i-j)}-U_{(i)}\right)^2}{2}\cdot r^{(3)}(\zeta_{i-j,i}),
\end{aligned} \tag{27}$$

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

$$\begin{aligned}
r^{(2)}(U_{(i+j)}) &= r^{(2)}(U_{(i)}) + \left(U_{(i+j)}-U_{(i)}\right)\cdot r^{(3)}(\zeta'_{i,i+j}),\\
r^{(2)}(U_{(i-j)}) &= r^{(2)}(U_{(i)}) + \left(U_{(i-j)}-U_{(i)}\right)\cdot r^{(3)}(\zeta'_{i-j,i}),
\end{aligned} \tag{28}$$

where $\zeta'_{i,i+j}\in\left[U_{(i)},U_{(i+j)}\right]$ and $\zeta'_{i-j,i}\in\left[U_{(i-j)},U_{(i)}\right]$.

Conditional bias: Given the above Taylor's expansions and the property that $\sum_{j=1}^{k_2}w_{ij,2}=1$, we can upper bound the absolute conditional bias as:

$$\begin{aligned}
&\left|\mathrm{Bias}\left[\widehat{Y}_i^{(2)}\big|\widetilde{\mathbb{U}}\right]\right|\\
&=\left|\mathbb{E}\left[\widehat{Y}_i^{(2)}\big|\widetilde{\mathbb{U}}\right]-r^{(2)}(U_{(i)})\right|
\end{aligned}$$

$$
\begin{aligned}
&= \left| 2\sum_{j=1}^{k_2} w_{ij,2} \cdot \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| \\
&\overset{\text{(i)}}{=} \left| 2\sum_{j=1}^{k_2} w_{ij,2} \left[\frac{r^{(1)}(U_{(i+j)}) - r^{(1)}(U_{(i-j)}) + \frac{r^{(2)}(U_{(i+j)})}{2}\left(U_{(i+j+k_1)}-U_{(i+j)}\right) - \frac{r^{(2)}(U_{(i-j)})}{2}\left(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. \\
&\quad \left.\left. + \frac{r^{(3)}(\zeta_{i+j,i+j+k_1})\left(U_{(i+j+k_1)}-U_{(i+j)}\right)^2 - r^{(3)}(\zeta_{i-j-k_1,i-j})\left(U_{(i-j-k_1)}-U_{(i-j)}\right)^2}{6\left(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}\right)} \right] - r^{(2)}(U_{(i)}) \right| \\
&\overset{\text{(ii)}}{=} \left| 2\sum_{j=1}^{k_2} w_{ij,2} \left[\frac{r^{(3)}(\zeta_{i,i+j})\left(U_{(i+j)}-U_{(i)}\right)^2 - r^{(3)}(\zeta_{i-j,i})\left(U_{(i-j)}-U_{(i)}\right)^2}{2\left(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}\right)} \right.\right. \\
&\quad + \frac{r^{(3)}(\zeta'_{i,i+j})\left(U_{(i+j)}-U_{(i)}\right)\left(U_{(i+j+k_1)}-U_{(i+j)}\right) - r^{(3)}(\zeta'_{i-j,i})\left(U_{(i-j)}-U_{(i)}\right)\left(U_{(i-j-k_1)}-U_{(i-j)}\right)}{2\left(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}\right)} \\
&\quad \left.\left. + \frac{r^{(3)}(\zeta_{i+j,i+j+k_1})\left(U_{(i+j+k_1)}-U_{(i+j)}\right)^2 - r^{(3)}(\zeta_{i-j-k_1,i-j})\left(U_{(i-j-k_1)}-U_{(i-j)}\right)^2}{6\left(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}\right)} \right] \right| \\
&\le \sup_{u\in[0,1]} \left| r^{(3)}(u) \right| \left(\sum_{j=1}^{k_2} w_{ij,2} \left[\frac{\left(U_{(i+j)}-U_{(i)}\right)^2 + \left(U_{(i-j)}-U_{(i)}\right)^2}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}} \right.\right. \\
&\quad + \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)}} \\
&\quad \left.\left. + \frac{\left(U_{(i+j+k_1)}-U_{(i+j)}\right)^2 + \left(U_{(i-j-k_1)}-U_{(i-j)}\right)^2}{3\left(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}\right)} \right] \right),
\end{aligned}
$$

where we plug in (26) to obtain (i) as well as use both (27) and (28) with $\sum_{j=1}^{k_2} w_{ij,2} = 1$ to derive (ii). By Lemma 9 with the weight $w_{ij,2} = \frac{(2j+k_1)^2}{\sum_{j=1}^{k_2}(2j+k_1)^2}$, we have that

$$
\begin{aligned}
\left|\mathrm{Bias}\left[\widehat{Y}_i^{(2)} \middle| \widetilde{\mathbb{U}}\right]\right| &\le \sup_{u\in[0,1]} \left|r^{(3)}(u)\right| \sum_{j=1}^{k_2} \left[\frac{(2j+k_1)^2}{\sum_{\ell=1}^{k_2}(2\ell+k_1)^2} \cdot \frac{2\left(\frac{j}{n+1}\right)^2 + \frac{2jk_1}{(n+1)^2} + \frac{2k_1^2}{3(n+1)^2}}{\frac{2(j+k_1)}{n+1} + \frac{2j}{n+1}} \right] [1 + o_P(1)] \\
&= \frac{\sup_{u\in[0,1]}\left|r^{(3)}(u)\right|}{n+1} \left(\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j} \right) [1 + o_P(1)].
\end{aligned}
$$

Conditional variance: By some direct calculations, we have that

$$
\begin{aligned}
&\mathrm{Var}\left[\widehat{Y}_i^{(2)} \middle| \widetilde{\mathbb{U}}\right] \\
&= \mathrm{Cov}\left[2\sum_{j=1}^{k_2} w_{ij,2} \cdot \frac{\left(\frac{Y_{i+j+k_1}-Y_{i+j}}{U_{(i+j+k_1)}-U_{(i+j)}} - \frac{Y_{i-j-k_1}-Y_{i-j}}{U_{(i-j-k_1)}-U_{(i-j)}}\right)}{U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}}, \right. \\
&\qquad \left. 2\sum_{\ell=1}^{k_2} w_{i\ell,2} \cdot \frac{\left(\frac{Y_{i+\ell+k_1}-Y_{i+\ell}}{U_{(i+\ell+k_1)}-U_{(i+\ell)}} - \frac{Y_{i-\ell-k_1}-Y_{i-\ell}}{U_{(i-\ell-k_1)}-U_{(i-\ell)}}\right)}{U_{(i+\ell+k_1)}+U_{(i+\ell)}-U_{(i-\ell-k_1)}-U_{(i-\ell)}} \middle| \widetilde{\mathbb{U}} \right] \\
&= 4\sum_{j=1}^{k_2}\sum_{\ell=1}^{k_2} \frac{w_{ij,2} w_{i\ell,2}}{\left(U_{(i+j+k_1)}+U_{(i+j)}-U_{(i-j-k_1)}-U_{(i-j)}\right)\left(U_{(i+\ell+k_1)}+U_{(i+\ell)}-U_{(i-\ell-k_1)}-U_{(i-\ell)}\right)}
\end{aligned}
$$

$$
\times \left\{ \frac{\mathrm{Cov}\left[Y_{i+j+k_1} - Y_{i+j},\, Y_{i+\ell+k_1} - Y_{i+\ell}\right]}{\left(U_{(i+j+k_1)} - U_{(i+j)}\right)\left(U_{(i+\ell+k_1)} - U_{(i+\ell)}\right)} - \frac{\mathrm{Cov}\left[Y_{i+j+k_1} - Y_{i+j},\, Y_{i+\ell+k_1} - Y_{i+\ell}\right]}{\left(U_{(i+j+k_1)} - U_{(i+j)}\right)\left(U_{(i-\ell-k_1)} - U_{(i-\ell)}\right)} \right.
$$

$$
\left. - \frac{\mathrm{Cov}\left[Y_{i-j-k_1} - Y_{i-j},\, Y_{i+\ell+k_1} - Y_{i+\ell}\right]}{\left(U_{(i-j-k_1)} - U_{(i-j)}\right)\left(U_{(i+\ell+k_1)} - U_{(i+\ell)}\right)} + \frac{\mathrm{Cov}\left[Y_{i-j-k_1} - Y_{i-j},\, Y_{i-\ell-k_1} - Y_{i-\ell}\right]}{\left(U_{(i-j-k_1)} - U_{(i-j)}\right)\left(U_{(i-\ell-k_1)} - U_{(i-\ell)}\right)} \right\}.
$$

Notice that

$$
\begin{aligned}
&\mathrm{Cov}\left[Y_{i+j+k_1} - Y_{i+j},\, Y_{i+\ell+k_1} - Y_{i+\ell}\right] \\
&= \mathrm{Cov}\left[Y_{i+j+k_1}, Y_{i+\ell+k_1}\right] - \mathrm{Cov}\left[Y_{i+j}, Y_{i+\ell+k_1}\right] - \mathrm{Cov}\left[Y_{i+j+k_1}, Y_{i+\ell}\right] + \mathrm{Cov}\left[Y_{i+j}, Y_{i+\ell}\right].
\end{aligned}
$$

When $j = \ell$, the first and fourth covariance are not zero; when $j = \ell + k_1$, the second covariance is not zero; and when $j + k_1 = \ell$, the third covariance is not zero. The other three covariance terms in $\mathrm{Var}\left[\widehat{Y}_i^{(2)} | \widetilde{\mathbb{U}}\right]$ can be derived in a similar way. Thus,

$$
\begin{aligned}
&\mathrm{Var}\left[\widehat{Y}_i^{(2)} | \widetilde{\mathbb{U}}\right] \\
&= 4\sigma_e^2 \sum_{j=1}^{k_2} \frac{w_{ij,2}^2}{\left(U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}\right)^2} \left[\frac{2}{\left(U_{(i+j+k_1)} - U_{(i+j)}\right)^2} + \frac{2}{\left(U_{(i-j-k_1)} - U_{(i-j)}\right)^2} \right] \\
&\quad - 4\sigma_e^2 \sum_{j=1}^{k_2-k_1} \frac{w_{ij,2} \cdot w_{i(j+k_1),2}}{\left(U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}\right)\left(U_{(i+j+2k_1)} + U_{(i+j+k_1)} - U_{(i-j-2k_1)} - U_{(i-j-k_1)}\right)} \\
&\qquad \times \left[\frac{1}{\left(U_{(i+j+k_1)} - U_{(i+j)}\right)\left(U_{(i+j+2k_1)} - U_{(i+j+k_1)}\right)} + \frac{1}{\left(U_{(i-j-k_1)} - U_{(i-j)}\right)\left(U_{(i-j-2k_1)} - U_{(i-j-k_1)}\right)} \right] \\
&\quad - 4\sigma_e^2 \sum_{j=1+k_1}^{k_2} \frac{w_{ij,2} \cdot w_{i(j-k_1),2}}{\left(U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}\right)\left(U_{(i+j)} + U_{(i+j-k_1)} - U_{(i-j)} - U_{(i-j+k_1)}\right)} \\
&\qquad \times \left[\frac{1}{\left(U_{(i+j+k_1)} - U_{(i+j)}\right)\left(U_{(i+j+2k_1)} - U_{(i+j+k_1)}\right)} + \frac{1}{\left(U_{(i-j-k_1)} - U_{(i-j)}\right)\left(U_{(i-j-2k_1)} - U_{(i-j-k_1)}\right)} \right] \\
&\leq 4\sigma_e^2 \sum_{j=1}^{k_2} \frac{w_{ij,2}^2}{\left(U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}\right)^2} \left[\frac{2}{\left(U_{(i+j+k_1)} - U_{(i+j)}\right)^2} + \frac{2}{\left(U_{(i-j-k_1)} - U_{(i-j)}\right)^2} \right] \\
&= \frac{4(n+1)^4 \sigma_e^2}{k_1^2 \sum_{j=1}^{k_2} (2j + k_1)^2} \left[1 + o_P(1)\right]
\end{aligned}
$$

when $k_1, k_2 \to \infty$ as $n \to \infty$, where the last equality follows from Lemma 9. Both results hold uniformly for $k_1 + k_2 + 1 \leq i \leq n - k_1 - k_2$, and the proof is thus completed. □

Remark 4. According to Theorem 5,

$$
\left| \mathrm{Bias}\left[\widehat{Y}_i^{(2)} | \widetilde{\mathbb{U}}\right] \right| \leq \frac{\sup_{u \in [0,1]} \left| r^{(3)}(u) \right|}{n+1} \left(\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3} k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3} k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \right) \left[1 + o_P(1)\right]
$$

$$
= O_P\left(\max\left\{\frac{k_1}{n}, \frac{k_2}{n}\right\}\right)
$$

and

$$
\begin{aligned}
\mathrm{Var}\left[\widehat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right] &\leq \frac{4(n+1)^4\sigma_e^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\left[1+o_P(1)\right] \\
&= O_P\left(\max\left\{\frac{n^4}{k_1^2k_2^3}, \frac{n^4}{k_1^4k_2}\right\}\right)
\end{aligned}
$$

when $k_1, k_2 \to \infty$ as $n \to \infty$. When $k_1, k_2 \asymp k$ are of the same order with respect to n, we know that the conditional mean square error of $\widehat{Y}_i^{(2)}$ satisfies that

$$
\begin{aligned}
\mathbb{E}\left[\left(\widehat{Y}_i^{(2)} - r^{(2)}(U_{(i)})\right)^2|\widetilde{\mathbb{U}}\right] &= \mathrm{Bias}\left[\widehat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right]^2 + \mathrm{Var}\left[\widehat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right] \\
&\leq O_P\left(\frac{k^2}{n^2}\right) + O_P\left(\frac{n^4}{k^5}\right),
\end{aligned}
$$

which is minimized when $k = O_P\left(n^{\frac{6}{7}}\right)$.

B.5 Proof of Corollary 6

Corollary 6 (Corollary 5 in Liu and De Brabanter 2020)**.** *Let $\mathcal{B}_2 = \sup_{u\in[0,1]}\left|r^{(3)}(u)\right|$. Under the assumptions of Theorem 5, the tuning parameters k_1 and k_2 that minimize the asymptotic upper bound of the conditional MISE are*

$$
(k_1,k_2)_{opt} = \underset{k_1,k_2=1,2,\ldots}{\arg\min}\left[\frac{\mathcal{B}_2^2}{(n+1)^2}\left(\frac{2\sum\limits_{j=1}^{k_2} j^3 + 3k_1\sum\limits_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum\limits_{j=1}^{k_2} j + \frac{1}{3}k_1^3k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2 + \frac{4(n+1)^4\sigma_e^2}{k_1^2\sum\limits_{j=1}^{k_2}(2j+k_1)^2}\right].
$$

Proof of Corollary 6. **Summary of the Proof:** The proof follows directly from the definition of the conditional MISE and the conditional bias-variance decomposition in Theorem 5.

From Remark 4, we know that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\widehat{Y}^{(2)}(U) - r^{(2)}(U)\right)^2\middle|\widetilde{\mathbb{U}}\right] \\
&\leq \frac{\mathcal{B}_2^2}{(n+1)^2}\left(\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2\left[1+o_P(1)\right] \\
&\quad + \frac{4(n+1)^4\sigma_e^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\left[1+o_P(1)\right].
\end{aligned}
$$

Since $U \sim \text{Unif}[0,1]$, the conditional MISE of $\widehat{Y}^{(2)}$ is given by

$$\begin{aligned}
\text{MISE}\left[\widehat{Y}^{(2)} | \mathbb{U}\right] &= \mathbb{E}\left\{\int_0^1 \left[\widehat{Y}^{(2)}(U) - r^{(1)}(U)\right]^2 dU \,\middle|\, \widetilde{\mathbb{U}}\right\} \\
&= \int_0^1 \mathbb{E}\left[\left(\widehat{Y}^{(2)}(U) - r^{(2)}(U)\right)^2 \middle| \widetilde{\mathbb{U}}\right] dU \\
&\leq \frac{\mathcal{B}_2^2}{(n+1)^2}\left(\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2 [1 + o_P(1)] \\
&\quad + \frac{4(n+1)^4\sigma_e^2}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\left[1 + o_P(1)\right].
\end{aligned}$$

This implies the asymptotic upper bound of the conditional MISE delineated in the statement of the corollary. The result follows. □

B.6 Proof of Theorem 7

Theorem 7 (Theorem 4 in Liu and De Brabanter 2020)**.** *Assume that $r(\cdot)$ under model (7) is $(p+3)$ times continuously differentiable in a neighborhood of u_0. Under Assumptions 1 and 2 on $\acute{\rho}_n$, the conditional bias and variance of (12) with $u_0 \in [0,1]$ for p odd are*

$$\begin{aligned}
\text{Bias}\left[\widehat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\right] &\leq \Bigg[\left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right| \widetilde{c}_p \left(\frac{\mathcal{B}_2}{n+1}\right)\left(\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right) \\
&\quad + \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} c_p \cdot \frac{r^{(p+3)}(u_0)}{(p+1)!} \cdot h^{p+1}\Bigg] [1 + o_P(1)] \\
\text{Var}\left[\widehat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\right] &\leq \frac{4(n+1)^4\sigma_e^2(1+\acute{\rho}_c)}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2(n-2k_1-2k_2)h} \cdot \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\boldsymbol{S}^*\boldsymbol{S}^{-1}\boldsymbol{\epsilon}_1 \left[1 + o_P(1)\right] \\
&= \frac{4(n+1)^4\sigma_e^2(1+\acute{\rho}_c)}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2(n-2k_1-2k_2)h}\left(\int K^\star(t)^2 dt\right)[1 + o_P(1)]
\end{aligned}$$

when $h \to 0$, $nh \to \infty$, $k_1, k_2 \to \infty$ as $n \to \infty$, where $\mathcal{B}_2 = \sup_{u\in[0,1]}\left|r^{(3)}(u)\right|$, $\widetilde{\mathbb{U}} = \left(U_{(1)}, ..., U_{(n)}\right)$, $\boldsymbol{S} = (\mu_{i+j-2})_{1\leq i,j\leq p+1}$ with $\mu_j = \int u^j K(u)du$, $\boldsymbol{S}^* = (\nu_{i+j-2})_{1\leq i,j\leq p+1}$ with $\nu_j = \int u^j K(u)^2 du$, $c_p = (\mu_{p+1}, ..., \mu_{2p+1})^T$, $\widetilde{c}_p = (\widetilde{\mu}_0, ..., \widetilde{\mu}_p)^T$ with $\widetilde{\mu}_j = \int |u|^j K(u)du$, $\boldsymbol{\epsilon}_1 = (1, 0, ..., 0)^T \in \mathbb{R}^{p+1}$, $\left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right|$ means elementwise absolute values of $\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}$, and the equivalent kernel $K_0^\star(t) = \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\left(1, t, ..., t^p\right)^T K(t)$.

Proof of Theorem 7. **Summary of the Proof:** The proof is almost identical to the one of Theorem 3 in Appendix B.3, where we will utilize the results of Theorem 3.1 in Fan and Gijbels (1996) and Theorem 1 in De Brabanter et al. (2018) to bound the conditional bias and variance of $\widehat{r}^{(2)}(u_0)$.

• **Conditional variance:** Let $k' = k_1 + k_2$. By Theorem 5 here and Theorem 1 in De Brabanter et al. (2018), we have that

$$\begin{aligned}
\mathrm{Var}\left[\widehat{Y}_i^{(2)}|\widetilde{\mathbb{U}}\right] &= \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k'}^{-1} \left(\boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \cdot \mathrm{Var}\left[\widehat{\boldsymbol{Y}}^{(2)}|\widetilde{\mathbb{U}}\right] \cdot \boldsymbol{W}_{u_0} \boldsymbol{U}_{u_0} \right) \boldsymbol{S}_{n-2k'}^{-1} \boldsymbol{\epsilon}_1 \\
&\leq \frac{4(n+1)^4 \sigma_e^2 (1+\acute{\rho}_c)}{k_1^2 \sum_{j=1}^{k_2} (2j+k_1)^2 (n-2k_1-2k_2) h} \cdot \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} \boldsymbol{S}^* \boldsymbol{S}^{-1} \boldsymbol{\epsilon}_1 \left[1 + o_P(1)\right],
\end{aligned}$$

where $\lim_{n\to\infty} n \int \rho_n(x) dx = \acute{\rho}_c$, $\boldsymbol{S} = (\mu_{i+j-2})_{1\leq i,j\leq p+1}$ with $\mu_j = \int u^j K(u) du$, and $\boldsymbol{S}^* = (\nu_{i+j-2})_{1\leq i,j\leq p+1}$ with $\nu_j = \int u^j K(u)^2 du$. The second term of the conditional variance follows from the definition of the equivalent kernel.

• **Conditional bias:** Using our arguments in the proof of Theorem 3 in Appendix B.3, we obtain that

$$\boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k'}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \begin{bmatrix} r^{(2)}(U_{(k'+1)}) \\ \vdots \\ r^{(2)}(U_{(n-k')}) \end{bmatrix} - r^{(2)}(u_0) = \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} c_p \cdot \frac{r^{(p+3)}(u_0)}{(p+1)!} \cdot h^{p+1} + o_P\left(h^{p+1}\right)$$

and

$$\begin{aligned}
&\boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k'}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \begin{bmatrix} \mathrm{Bias}\left[\widehat{Y}_{k'+1}^{(2)}|\widetilde{\mathbb{U}}\right] \\ \vdots \\ \mathrm{Bias}\left[\widehat{Y}_{n-k'}^{(2)}|\widetilde{\mathbb{U}}\right] \end{bmatrix} \\
&\leq \left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right| \widetilde{c}_p \cdot \left(\frac{\mathcal{B}_2}{n+1}\right) \left(\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3} k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3} k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \right).
\end{aligned}$$

Combining these two expressions yields that

$$\begin{aligned}
\mathrm{Bias}\left[\widehat{r}^{(2)}(u_0)|\widetilde{\mathbb{U}}\right] &= \boldsymbol{\epsilon}_1^T \boldsymbol{S}_{n-2k'}^{-1} \boldsymbol{U}_{u_0}^T \boldsymbol{W}_{u_0} \left(\begin{bmatrix} r^{(2)}(U_{(k'+1)}) \\ \vdots \\ r^{(2)}(U_{(n-k')}) \end{bmatrix} + \begin{bmatrix} \mathrm{Bias}\left[\widehat{Y}_{k'+1}^{(2)}|\widetilde{\mathbb{U}}\right] \\ \vdots \\ \mathrm{Bias}\left[\widehat{Y}_{n-k'}^{(2)}|\widetilde{\mathbb{U}}\right] \end{bmatrix} \right) - r^{(2)}(u_0) \\
&\leq \Bigg[\left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right| \widetilde{c}_p \left(\frac{\mathcal{B}_2}{n+1}\right) \left(\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3} k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3} k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \right) \\
&\quad + \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} c_p \cdot \frac{r^{(p+3)}(u_0)}{(p+1)!} \cdot h^{p+1} \Bigg] \left[1 + o_P(1)\right],
\end{aligned}$$

where $\mathcal{B}_2 = \sup_{u\in[0,1]} \left|r^{(3)}(u)\right|$. The results follow. $\square$

Remark 5. Assume that the assumptions of Theorem 7 hold and $k_1, k_2 \asymp k$ have the same asymptotic order. Then, the conditional mean integrated squared error (MISE) of the local polynomial regression estimator (12) is upper bounded by

$$
\begin{aligned}
&\text{MISE}\left[\widehat{r}^{(2)}|\widetilde{\mathbb{U}}\right] \\
&= \mathbb{E}\left\{\int_0^1 \left[\widehat{r}^{(2)}(u_0) - r^{(2)}(u_0)\right]^2 du_0 \,\middle|\, \widetilde{\mathbb{U}}\right\} \\
&= \int_0^1 \mathbb{E}\left[\left(\widehat{r}^{(2)}(u_0) - r^{(2)}(u_0)\right)^2 \middle| \widetilde{\mathbb{U}}\right] du_0 \\
&\leq \left[\left|\boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1}\right| \widetilde{c}_p \cdot \left(\frac{\mathcal{B}_2}{n+1}\right)\left(\frac{2\sum_{j=1}^k j^3 + 3k\sum_{j=1}^k j^2 + \frac{5}{3}k^2\sum_{j=1}^k j + \frac{1}{3}k^4}{4\sum_{j=1}^k j^2 + k^3 + 4k\sum_{j=1}^k j}\right)\right. \\
&\quad \left. + \boldsymbol{\epsilon}_1^T \boldsymbol{S}^{-1} c_p \cdot \frac{r^{(p+3)}(u_0)}{(p+1)!} \cdot h^{p+1}\right]^2 [1 + o_P(1)] \\
&\quad + \frac{4(n+1)^4 \sigma_e^2 (1 + \acute{\rho}_c)}{k^2 \sum_{j=1}^k (2j+k)^2 (n-4k) h}\left(\int K^\star(t)^2 dt\right)[1 + o_P(1)] \\
&= O_P\left(h^{2p+2}\right) + O_P\left(\frac{kh^{p+1}}{n}\right) + O_P\left(\frac{k^2}{n^2}\right) + O_P\left(\frac{n^3}{k^5 h}\right),
\end{aligned}
$$

given that k is always of smaller order than $O(n)$. By taking the partial derivatives with respect to h and k and setting them to 0, we obtain a system of equations

$$
\begin{cases}
h^{2p+1} + \frac{kh^p}{n} - \frac{n^3}{k^5 h^2} \asymp 0, \\
\frac{h^{p+1}}{n} + \frac{k}{n^2} - \frac{n^3}{k^6 h} \asymp 0,
\end{cases}
$$

where we introduce the asymptotic equivalence symbol "$\asymp$" to get rid of all the constant factors. Solving this system of equations gives us that $k = O\left(n^{\frac{5p+6}{7p+8}}\right)$ and $h = O\left(n^{-\frac{2}{7p+8}}\right)$, which leading to an optimal rate of convergence for the upper bound of MISE $\left[\widehat{r}^{(2)}|\widetilde{\mathbb{U}}\right]$ as $O_P\left(n^{-\frac{4p+4}{7p+8}}\right)$.

B.7 Proof of Theorem 8

Theorem 8. *Assume that $m(\cdot)$ under model (1) is $(p+3)$ times continuously differetiable within $[a,b]$, and the density f of X is at least three times continuously differentiable with $\inf_{x\in[a,b]} f(x) > c > 0$ for some constant c. Then,*

- **Pointwise consistency:** *under Assumptions 1, 2, and 3, the derivative estimators in (15) for $q = 1, 2$ and any fixed $x \in [a,b]$ satisfy*

$$
\left|\widehat{m}^{(q)}(x) - m^{(q)}(x)\right| = O\left(h^{p+1}\right) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n^{2q-1}}{k^{2q+1}h}}\right) + O(v^2) + O_P\left(\sqrt{\frac{1}{nv^{2q-1}}}\right)
$$

when $h \to 0, \frac{k}{n} \to 0, \frac{n^{2q-1}}{k^{2q+1}h} \to 0, v \to 0, nv^{2q-1} \to \infty$ as $n \to \infty$.

- **Uniform consistency:** *under Assumptions 1, 2, 3, 4, and 5, when* $h \to 0$, $\frac{k}{n} \to 0$, $\frac{n^{2q-1}\log n}{k^{2q+1}h} \to 0, v \to 0, \frac{nv^{2q-1}}{\log n} \to \infty$ *as* $n \to \infty$, *we have that*

$$\sup_{x\in[a,b]} \left|\widehat{m}^{(q)}(x) - m^{(q)}(x)\right| = O\left(h^{p+1}\right) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n^{2q-1}\log n}{k^{2q+1}h}}\right) + O(v^2) + O_P\left(\sqrt{\frac{\log n}{nv^{2q-1}}}\right).$$

Proof of Theorem 8. The proof follows from standard consistency results for KDE (see, *e.g.*, Section 2.1 in Chen 2017) and Corollary 1 in Francisco-Fernández et al. (2003). In particular, under our kernel assumptions (*i.e.*, Assumptions 3 and 4), one can use the techniques in Giné and Guillou (2002); Einmahl and Mason (2005); Chacón et al. (2011) to show that

$$\widehat{f}_v^{(\alpha)}(x) - f^{(\alpha)}(x) = O(v^2) + O_P\left(\sqrt{\frac{1}{nv^{1+2\alpha}}}\right),$$

$$\sup_{x\in[a,b]} \left|\widehat{f}_v^{(\alpha)}(x) - f^{(\alpha)}(x)\right| = O(v^2) + O_P\left(\sqrt{\frac{\log n}{nv^{1+2\alpha}}}\right)$$

for $\alpha = 0, 1$; see also Section 5 in Genovese et al. (2014). According to (15), we have that

$$\begin{aligned}
&\left|\widehat{m}^{(1)}(x) - m^{(1)}(x)\right| \\
&= \left|\widehat{f}_v(x)\cdot \widehat{r}^{(1)}(u) - f(x)\cdot r^{(1)}(u)\right| \\
&\le \left|\widehat{f}_v(x)\right| \cdot \left|\widehat{r}^{(1)}(u) - r^{(1)}(u)\right| + \left|r^{(1)}(u)\right| \cdot \left|\widehat{f}_v(x) - f(x)\right| \\
&\overset{\text{(i)}}{\le} \sup_{x\in[a,b]} \left|\widehat{f}_v(x)\right| \left[O(h^{p+1}) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n}{hk^3}}\right)\right] + \sup_{u\in[0,1]} \left|r^{(1)}(u)\right| \left[O(v^2) + O_P\left(\sqrt{\frac{1}{nv}}\right)\right] \\
&= O(h^{p+1}) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n}{hk^3}}\right) + O(v^2) + O_P\left(\sqrt{\frac{1}{nv}}\right),
\end{aligned}$$

where $u = F(x)$ and we use the results from Theorem 3 to obtain inequality (i). Notice also that $\sup_{x\in[a,b]} \left|\widehat{f}_v(x)\right| < \infty$ and $\sup_{u\in[0,1]} \left|r^{(1)}(u)\right| < \infty$ by the differentiability assumptions on K_{kde} and m. The uniform consistency of $\widehat{m}^{(1)}(x)$ follows similarly as:

$$\begin{aligned}
&\sup_{x\in[a,b]} \left|\widehat{m}^{(1)}(x) - m^{(1)}(x)\right| \\
&= \sup_{x\in[a,b],u\in[0,1]} \left|\widehat{f}_v(x)\cdot \widehat{r}^{(1)}(u) - f(x)\cdot r^{(1)}(u)\right| \\
&\le \sup_{x\in[a,b]} \left|\widehat{f}_v(x)\right| \cdot \sup_{u\in[0,1]} \left|\widehat{r}^{(1)}(u) - r^{(1)}(u)\right| + \sup_{u\in[0,1]} \left|r^{(1)}(u)\right| \cdot \sup_{x\in[a,b]} \left|\widehat{f}_v(x) - f(x)\right| \\
&\overset{\text{(ii)}}{\le} \sup_{x\in[a,b]} \left|\widehat{f}_v(x)\right| \left[O(h^{p+1}) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n\log n}{hk^3}}\right)\right] + \sup_{u\in[0,1]} \left|r^{(1)}(u)\right| \left[O(v^2) + O_P\left(\sqrt{\frac{\log n}{nv}}\right)\right] \\
&= O(h^{p+1}) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n\log n}{hk^3}}\right) + O(v^2) + O_P\left(\sqrt{\frac{\log n}{nv}}\right),
\end{aligned}$$

where we apply Corollary 1 in Francisco-Fernández et al. (2003) and the above rate of convergence for KDE to derive inequality (ii).

As for $\widehat{m}^{(2)}(x)$, we derive analogously from (15) as:

$$
\begin{aligned}
&\left|\widehat{m}^{(2)}(x) - m^{(2)}(x)\right| \\
&= \left|\widehat{f}_v^{(1)}(x) \cdot \widehat{r}^{(1)}(u) + \left[\widehat{f}_v(x)\right]^2 \widehat{r}^{(2)}(u) - f^{(1)}(x) \cdot r^{(1)}(u) - [f(x)]^2 r^{(2)}(u)\right| \\
&\leq \left|\widehat{f}_v^{(1)}(x)\right| \cdot \left|\widehat{r}^{(1)}(u) - r^{(1)}(u)\right| + \left|r^{(1)}(u)\right| \cdot \left|\widehat{f}_v^{(1)}(x) - f_v^{(1)}(x)\right| + \left[\widehat{f}_v(x)\right]^2 \left|\widehat{r}^{(2)}(u) - r^{(2)}(u)\right| \\
&\quad + \left|r^{(2)}(u)\right| \left|\left[\widehat{f}_v(x)\right]^2 - [f(x)]^2\right| \\
&\overset{\text{(iii)}}{\leq} \sup_{x\in[a,b]} \left|\widehat{f}_v^{(1)}(x)\right| \left[O(h^{p+1}) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n}{hk^3}}\right)\right] + \sup_{u\in[0,1]} \left|r^{(1)}(u)\right| \left[O(v^2) + O_P\left(\sqrt{\frac{1}{nv^3}}\right)\right] \\
&\quad + \sup_{x\in[a,b]} \left[\widehat{f}_v(x)\right]^2 \left[O(h^{p+1}) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n^3}{hk^5}}\right)\right] \\
&\quad + \sup_{u\in[0,1]} \left|r^{(1)}(u)\right| \sup_{x\in[a,b]} \left|\widehat{f}_v(x) + f(x)\right| \left[O(v^2) + O_P\left(\sqrt{\frac{1}{nv}}\right)\right] \\
&= O(h^{p+1}) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n}{k^5 h}}\right) + O(v^2) + O_P\left(\sqrt{\frac{1}{nv^3}}\right),
\end{aligned}
$$

where $u = F(x)$ and we use the results from Theorem 7 to obtain inequality (iii). Finally, the uniform consistency of $\widehat{m}^{(2)}(x)$ follows similarly as:

$$
\begin{aligned}
&\sup_{x\in[a,b]} \left|\widehat{m}^{(2)}(x) - m^{(2)}(x)\right| \\
&= \sup_{x\in[a,b],u\in[0,1]} \left|\widehat{f}_v^{(1)}(x) \cdot \widehat{r}^{(1)}(u) + \left[\widehat{f}_v(x)\right]^2 \widehat{r}^{(2)}(u) - f^{(1)}(x) \cdot r^{(1)}(u) - [f(x)]^2 r^{(2)}(u)\right| \\
&\leq \sup_{x\in[a,b]} \left|\widehat{f}_v^{(1)}(x)\right| \cdot \sup_{u\in[0,1]} \left|\widehat{r}^{(1)}(u) - r^{(1)}(u)\right| + \sup_{u\in[0,1]} \left|r^{(1)}(u)\right| \cdot \sup_{x\in[a,b]} \left|\widehat{f}_v^{(1)}(x) - f_v^{(1)}(x)\right| \\
&\quad + \sup_{x\in[a,b]} \left[\widehat{f}_v(x)\right]^2 \sup_{u\in[0,1]} \left|\widehat{r}^{(2)}(u) - r^{(2)}(u)\right| + \sup_{u\in[0,1]} \left|r^{(2)}(u)\right| \sup_{x\in[a,b]} \left|\left[\widehat{f}_v(x)\right]^2 - [f(x)]^2\right| \\
&\overset{\text{(iv)}}{\leq} \sup_{x\in[a,b]} \left|\widehat{f}_v^{(1)}(x)\right| \left[O(h^{p+1}) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n\log n}{hk^3}}\right)\right] + \sup_{u\in[0,1]} \left|r^{(1)}(u)\right| \left[O(v^2) + O_P\left(\sqrt{\frac{\log n}{nv^3}}\right)\right] \\
&\quad + \sup_{x\in[a,b]} \left[\widehat{f}_v(x)\right]^2 \left[O(h^{p+1}) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n^3\log n}{hk^5}}\right)\right] \\
&\quad + \sup_{u\in[0,1]} \left|r^{(1)}(u)\right| \sup_{x\in[a,b]} \left|\widehat{f}_v(x) + f(x)\right| \left[O(v^2) + O_P\left(\sqrt{\frac{\log n}{nv}}\right)\right] \\
&= O(h^{p+1}) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n\log n}{k^5 h}}\right) + O(v^2) + O_P\left(\sqrt{\frac{\log n}{nv^3}}\right),
\end{aligned}
$$

where we use Corollary 1 in Francisco-Fernández et al. (2003) and the uniform rates of convergence for $\widehat{f}_v, \widehat{f}_v^{(1)}$ again to derive inequality (iv). The proof is thus completed. $\square$

B.8 Auxiliary Results

Lemma 9. *Let $\{U_i\}_{i=1}^n$ be i.i.d. observations from* Unif$[0,1]$. *Then, the order statistics $U_{(1)} \leq \cdots \leq U_{(n)}$ satisfy*

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

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

and

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

for $i > j$.

Proof of Lemma 9. **Summary of the Proof:** The proof utilizes the result that

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

for $1 \leq i < j \leq n$ and Chebyshev's inequality.

To establish this result, we rename $Z = U_{(j)}, W = U_{(i)}$ and their joint density is given by (Theorem 5.4.6 in Casella and Berger 2002)

$$f_{W,Z}(w,z) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \cdot w^{i-1}(z-w)^{j-i-1}(1-z)^{n-j}, \quad 0 < w < z < 1.$$

Given that the distribution of $Z - W$ are of interest, we apply the change of variables

$$\begin{cases} V_1 = Z - W, \\ V_2 = Z + W, \end{cases} \iff \begin{cases} Z = \frac{V_1+V_2}{2}, \\ W = \frac{V_2-V_1}{2}. \end{cases}$$

The joint density of (V_1, V_2) becomes

$$f_{V_1,V_2}(v_1,v_2) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \cdot \frac{(v_2-v_1)^{i-1}}{2^{i-1}} \cdot v_1^{j-i-1}\left[1 - \left(\frac{v_1+v_2}{2}\right)\right]^{n-j} \frac{1}{2},$$

where $0 < v_1 < v_2 < 2 - v_1$. Hence, the (marginal) density of V_1 is

$$f_{V_1}(v_1) = \frac{n! \cdot v_1^{j-i-1}}{2(i-1)!(j-i-1)!(n-j)!} \int_{v_1}^{2-v_1} \frac{(v_2-v_1)^{i-1}(2-v_1-v_2)^{n-j}}{2^{n-j+i-1}} dv_2$$

$$= \frac{n! \cdot v_1^{j-i-1}}{2(i-1)!(j-i-1)!(n-j)!} \cdot 2(1-v_1)^{n-j+i} \int_0^1 x^{i-1}(1-x)^{n-j} dx \quad \text{by } x = \frac{v_2 - v_1}{2 - 2v_1}$$
$$= \frac{n!}{(n-j+i)!(j-i-1)!} \cdot v_1^{j-i-1}(1-v_1)^{n-j+i}$$

with $0 < v_1 < 1$, which is the density of Beta$(j-i, n-j+i+1)$.

Therefore,

$$U_{(i+j)} - U_{(i-j)} \sim \text{Beta}\left(2j, n-2j+1\right), \quad U_{(i+j)} - U_{(i)} \text{ or } U_{(i)} - U_{(i-j)} \sim \text{Beta}\left(j, n+1-j\right),$$

and by Chebyshev's inequality,

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

and

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

The same asymptotic property holds for $U_{(i)} - U_{(i-j)}$ given that it has the same distribution as $U_{(i+j)} - U_{(i)} \sim \text{Beta}\left(j, n+1-j\right)$. □

Remark 6. Under model (7) and the assumption that r is twice continuously differentiable on $[0,1]$, we use the Taylor's expansion with Lemma 9 that

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

Therefore, the first-order difference quotients $\widehat{q}_i^{(1)} = \frac{Y_i - Y_{i-1}}{U_{(i)} - U_{(i-1)}}, i = 1, ..., n$ satisfy that

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

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

$$\text{Var}\left[\widehat{q}_i^{(1)} \middle| U_{(i-1)}, U_{(i)}\right] = \text{Var}\left[\frac{Y_i - Y_{i-1}}{U_{(i)} - U_{(i-1)}} \middle| U_{(i-1)}, U_{(i)}\right] = \frac{2\sigma_e^2}{\left(U_{(i)} - U_{(i-1)}\right)^2} = O_P\left(n^2\right).$$

We make two remarks on the above results. First, the first-order difference quotients $\widehat{q}_i^{(1)} = \frac{Y_i - Y_{i-1}}{U_{(i)} - U_{(i-1)}}, i = 2, ..., n$ are asymptotically unbiased estimators of $r^{(1)}(U_{(i)}), i = 2, ..., n$. Second, the variances of $\widehat{q}_i^{(1)}, i = 2, ..., n$ tend to infinity as the sample size n increases. This result for the random design resembles the derivation in Section 2.1 in De Brabanter et al. (2013) for the equispaced design. It emphasizes the necessity of aggregating several symmetric difference quotients as in (8) to reduce the variance of the first-order noisy derivative estimator.

Proposition 10 (Proposition 1 in Liu and De Brabanter 2020)**.** *For $k + 1 \leq i \leq n - k$ and under model (7), the weights $w_{i,j}, j = 1, ..., k$ with $\sum_{j=1}^k w_{i,j} = 1$ that minimize the variance of (8) are given by*

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

Proof of Proposition 10. **Summary of the Proof:** The proof follows from a direct minimization of the variance of $\widehat{Y}_i^{(1)}$ conditional on $\mathbb{U} = \left(U_{(i-j)}, ..., U_{(i+j)}\right)$ for $i > j$, $i + j \leq n$, and $j = 1, ..., k$.

Recall from model (7) and (8) that $Y_i = r(U_{(i)}) + e_i$ with $\text{Var}(e_i) = \sigma_e^2$ and

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

where we use the (conditional) independence between $Y_{i+j} - Y_{i-j}$ for different $j = 1, ..., k$ given $\mathbb{U}$ in equality (i) and the (conditional) independence between Y_{i+j} and Y_{i-j} for $j = 1, ..., k$ given $\mathbb{U}$ in equality (ii). Under the constraint $\sum_{j=1}^k w_{i,j} = 1$, we compute the partial derivatives of the Lagrangian function $\mathcal{L}(w_{i,1}, ..., w_{i,k}, \lambda) = \sum_{j=1}^k w_{i,j}^2 \cdot \frac{2\sigma_e^2}{\left(U_{(i+j)} - U_{(i-j)}\right)^2} + \lambda\left(\sum_{j=1}^k w_{i,j} - 1\right)$ and set them to 0 as:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_{i,j}} &= 2w_{i,j} \cdot \frac{2\sigma_e^2}{\left(U_{(i+j)} - U_{(i-j)}\right)^2} + \lambda = 0, \quad j = 1, ..., k, \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= \sum_{j=1}^k w_{i,j} - 1 = 0.
\end{aligned}$$

Solving the above system of equations yields that

$$\lambda = -\frac{4\sigma_e^2}{\sum_{j=1}^k \left(U_{(i+j)} - U_{(i-j)}\right)^2} \quad \text{and} \quad w_{i,j} = \frac{\left(U_{(i+j)} - U_{(i-j)}\right)^2}{\sum_{\ell=1}^k \left(U_{(i+\ell)} - U_{(i-\ell)}\right)^2}, \quad j = 1, ..., k.$$

Finally, computing the Hessian matrix of $\mathcal{L}$ leads to a positive definite matrix, so the above weights minimize the variance $\mathrm{Var}\left[\widehat{Y}_i^{(1)} \middle| \mathbb{U}\right]$. $\square$

Remark 7. From the proof of Proposition 10, we note that $\mathrm{Var}\left[\frac{Y_{i+j}-Y_{i-j}}{U_{(i+j)}-U_{(i-j)}} \middle| \mathbb{U}\right] = \frac{2\sigma_e^2}{\left(U_{(i+j)}-U_{(i-j)}\right)^2}$, and the j-th weight $w_{i,j}$ is thus proportional to the reciprocal variance of the difference quotient $\frac{Y_{i+j}-Y_{i-j}}{U_{(i+j)}-U_{(i-j)}}$.