Journal of Machine Learning Research 17 (2016) 1-25 Submitted 12/15; Revised 4/16; Published 8/16

Optimal Estimation of Derivatives in Nonparametric Regression

Wenlin Dai WENLIN.DAI@KAUST.EDU.SA
CEMSE Division
King Abdullah University of Science and Technology
Saudi Arabia

Tiejun Tong TONGT@HKBU.EDU.HK
Department of Mathematics
Hong Kong Baptist University
Hong Kong

Marc G. Genton MARC.GENTON@KAUST.EDU.SA
CEMSE Division
King Abdullah University of Science and Technology
Saudi Arabia

Editor: Xiaotong Shen

Abstract

We propose a simple framework for estimating derivatives without fitting the regression function in nonparametric regression. Unlike most existing methods that use the symmetric difference quotients, our method is constructed as a linear combination of observations. It is hence very flexible and applicable to both interior and boundary points, including most existing methods as special cases of ours. Within this framework, we define the variance-minimizing estimators for any order derivative of the regression function with a fixed bias-reduction level. For the equidistant design, we derive the asymptotic variance and bias of these estimators. We also show that our new method will, for the first time, achieve the asymptotically optimal convergence rate for difference-based estimators. Finally, we provide an effective criterion for selection of tuning parameters and demonstrate the usefulness of the proposed method through extensive simulation studies of the first- and second-order derivative estimators.

Keywords: Linear combination, Nonparametric derivative estimation, Nonparametric regression, Optimal sequence, Taylor expansion

1. Introduction

Consider the following nonparametric regression model:

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

where x_i are the design points satisfying $0 \le x_1 < \cdots < x_n \le 1$, $m(x)$ is the regression function, Y_i are the observations, and ε_i are independent and identically distributed random errors with $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2 < \infty$. Estimation of $m(x)$ is an important problem in nonparametric regression and has received sustained attention in the literature. Such

methods include, for example, kernel smoothing (Härdle, 1990), spline smoothing (Wahba, 1990), and local polynomial regression (Fan and Gijbels, 1996). It has been noted that the estimation of the first- or higher-order derivatives of $m(x)$ is also important for practical implementations including, but not limited to, the modeling of human growth data (Ramsay and Silverman, 2002), kidney function for a lupus nephritis patient (Ramsay, 2006), and Raman spectra of bulk materials (Charnigo et al., 2011). Derivative estimation is also needed in nonparametric regression to construct confidence intervals for regression functions (Eubank and Speckman, 1993), to select kernel bandwidths (Ruppert et al., 1995), and to compare regression curves (Park and Kang, 2008).

Most existing methods for pth-order derivative estimation can be expressed as a weighted average of the responses,

$$\hat{m}^{(p)}(x) = \sum_{i=1}^{n} w_i(x) Y_i,$$

where $w_i(x)$ are weights assigned to each observation Y_i. These estimators can be separated into two classes by their ability to directly or indirectly assess the weights, $w_i(x)$. In the indirect methods, the regression function is initially estimated as $\hat{m}(x) = \sum_{i=1}^{n} c_i(x) Y_i$ by the aforementioned nonparametric smoothing techniques, where $c_i(x)$ are smooth functions. Then, $w_i(x)$ are estimated as $\mathrm{d}^p c_i(x)/\mathrm{d}x^p$ (Gasser and Müller, 1984; Müller et al., 1987; Fan and Gijbels, 1995; Zhou and Wolfe, 2000; Boente and Rodriguez, 2006; Cao, 2014). We note, however, that the optimal bandwidths may differ for estimating the regression function and for estimating the derivatives, respectively. That is, a good estimate of the regression function may not guarantee the generation of good estimates of the derivatives.

Direct methods lead to the second class, which estimate the derivatives directly without fitting the regression function. The two key steps for such methods are constructing point-wise estimates for the derivatives of each design point and determining the amount of smoothing or the bandwidth. To select the bandwidth, one may refer to some classical methods in Müller et al. (1987), Härdle (1990), Fan and Gijbels (1996), Opsomer et al. (2001), Lahiri (2003), and Kim et al. (2009), among others. In contrast, little attention has been paid to the improvement of the point-wise estimation of the derivatives. One simple point-wise estimator for derivatives uses difference quotients. This method is, however, very noisy. For example, the variance of the first-order difference quotient $(Y_i - Y_{i-1})/(x_i - x_{i-1})$ is of order $O(n^2)$. Charnigo et al. (2011) proposed a variance-reducing linear combination of symmetric difference quotients, called *empirical derivatives*, and applied it to their generalized C_p criterion for tuning parameter selection. De Brabanter et al. (2013) established the L_1 and L_2 convergence rates for the empirical derivatives. Specifically, they defined the empirical derivatives as

$$Y_i^{(L)} = \sum_{j=1}^{k_L} w_{j,L} \left(\frac{Y_{i+j}^{(L-1)} - Y_{i-j}^{(L-1)}}{x_{i+j} - x_{i-j}} \right), \quad L = 1, \ldots p,$$

where $Y_i^{(L)}$ denotes the estimated Lth-order derivative at x_i, $Y_i^{(0)} = Y_i$ and $w_{j,L}$ are the associated weights. When $L = 1$, $w_{j,1}$ are chosen as the optimal weights that minimize the estimation variance. For $L \geq 2$, $w_{j,L}$ are determined intuitively instead of by optimizing the estimation variance. As a consequence, their higher-order empirical derivatives may not

be optimally defined. Another attempt was made recently by Wang and Lin (2015). They estimated the derivative as the intercept of a linear regression model through the weighted least squares method. They further showed that their proposed estimators achieve better control of the estimation bias, which makes them superior to empirical derivatives when the signal-to-noise ratio is large. Finally, it is noteworthy that their method only applies to equidistant designs and hence the practical applications are somewhat limited.

In this paper, we propose a simple framework for estimating derivatives in model (1) without fitting the regression function. Our method does not rely on symmetric difference quotients; hence, it is more flexible than existing methods. Within this framework, we define the variance-minimizing estimators for any order derivative of $m(x)$ with a fixed bias-reduction level. For the equidistant design, we derive the asymptotic variance and bias of these estimators. We also show that the proposed estimators perform well on both interior and boundary points and, more importantly, that they achieve the optimal convergence rate for the mean squared error (MSE).

The rest of this paper is organized as follows. In Section 2, we propose a new framework for first-order derivative estimation and show that most existing estimators are special cases of ours. We also investigate the theoretical properties of the proposed estimator, including the optimal sequence, the asymptotic variance and bias, the point-wise consistency, and the boundary behavior. In Section 3, we extend the proposed method to higher-order derivative estimation and provide an effective criterion for the selection of tuning parameters. We then report extensive simulation studies in Section 4 that validate the proposed method. We conclude the paper with a discussion in Section 5. Technical proofs of the theoretical results are given in the Appendix.

2. First-order derivative estimation

In this section, we propose a new framework for estimating derivatives in nonparametric regression. Within this framework, we define the optimal estimator for the first-order derivative by minimizing the estimation variance. Theoretical results including the asymptotic variance and bias, and point-wise consistency are derived for the proposed optimal estimators under the equidistant design. We also investigate the performance of the estimators on the boundaries.

2.1 New framework

Recall that most existing methods are weighted average of symmetric difference quotients, which limits their implementation to some extent. All these estimators can be expressed as a linear combination of the observations for fixed design points. To proceed, we define

$$DY_i = \sum_{k=0}^{r} d_k Y_{i+k}, \quad 1 \le i \le n - r,$$

where $(d_0, \ldots, d_r)$ is a sequence of real numbers, and r is referred to as the order of the sequence. Assuming that $m(x)$ is a smooth enough function, we have the following Taylor

expansion at x_{i+l} for each $m(x_{i+k})$,

$$m(x_{i+k}) = m(x_{i+l}) + \sum_{j=1}^{\infty} \frac{(x_{i+k} - x_{i+l})^j}{j!} m^{(j)}(x_{i+l}), \quad 0 \le l \le r.$$

Note that x_{i+l} can be any design point within $[x_i, x_{i+r}]$, which frees our method from the symmetric form restriction. If we further assume that x_i are equidistant, then $x_i = i/n, i = 1, \ldots, n$. Define $C_{j,l} = \sum_{k=0}^{r} d_k (k-l)^j / (n^j j!)$, $j = 0, 1, \ldots$ and $l = 0, \ldots, r$. The expectation of DY_i can be expressed as

$$\mathrm{E}(DY_i) = \sum_{j=0}^{\infty} C_{j,l} m^{(j)}(x_{i+l}), \quad 1 \le i \le n - r. \tag{2}$$

To estimate the first-order derivative at x_{i+l} with DY_i, we let $C_{0,l} = 0$ and $C_{1,l} = 1$ so that

$$\mathrm{E}(DY_i) = m'(x_{i+l}) + \sum_{j=2}^{\infty} C_{j,l} m^{(j)}(x_{i+l}),$$

where the second term on the right side is the estimation bias. When the regression function is oscillating around x_{i+l}, we can alter our model by controlling the estimation bias at a higher level. Specifically, if we let

$$C_{1,l} = 1 \text{ and } C_{j,l} = 0, \ 0 \le j \ne 1 \le q - 1, \tag{3}$$

then

$$\mathrm{E}(DY_i) = m'(x_{i+l}) + \sum_{j=q}^{\infty} C_{j,l} m^{(j)}(x_{i+l}).$$

When $q = 2$, condition (3) reduces to $C_{1,l} = 1$ and $C_{0,l} = 0$. When $q \ge 3$, condition (3) eliminates the estimation bias up to order $q - 1$.

2.2 Theoretical results

If we use a sequence with an order $r \ge q$, an infinite number of choices satisfying (3) is available. Among them, we choose the one(s) minimizing the estimation variance, $\mathrm{var}(DY_i) = \sigma^2 \sum_{k=0}^{r} d_k^2$, which leads to the following optimization problem,

$$(d_0, \ldots, d_r)_{1,q} = \underset{(d_0,\ldots,d_r) \in \mathbb{R}^{r+1}}{\operatorname{argmin}} \sum_{k=0}^{r} d_k^2, \quad \text{s.t.} \quad \text{condition (3) holds.}$$

We denote this variance-minimizing sequence as $(d_0, \ldots, d_r)_{1,q}$. For simplicity of notation, the dependence of d_k on l is suppressed. In addition, we introduce the following notation:

$I_i^{(l)} = \sum_{k=0}^{r} (k-l)^i, \quad l = 0, \ldots, r \quad \text{and} \quad i = 0, 1, \ldots;$

$U^{(l)}$ denotes a $q \times q$ matrix with $u_{ij}^{(l)} = I_{i+j-2}^{(l)}$;

$V^{(l)} = (U^{(l)})^{-1}$ is the inverse matrix of $U^{(l)}$.

Then, we present the theoretical results for $(d_0, \ldots, d_r)_{1,q}$ in the following proposition.

Proposition 1 *Assume that model (1) holds with equidistant design and $m(x)$ has a finite qth-order derivative on $[0,1]$. For $1 \le i \le n-r$ and $0 \le l \le r$, the unique variance-minimizing sequence is*

$$(d_k)_{1,q} = n \sum_{j=0}^{q-1} V^{(l)}_{(j+1,2)} (k-l)^j, \quad k = 0, \ldots, r,$$

for estimating $m'(x_{i+l})$ with an order of accuracy up to $m^{(q)}(x_{i+l})$, $q \ge 2$. Here, $V^{(l)}_{(i,j)}$ denotes the element in the ith row and the jth column of the matrix $V^{(l)}$.
Proof: see Appendix A.

When q is fixed, the optimal sequence depends only on l, which makes it quite convenient for practical implementation. When r is even and $l = r/2$, we get the symmetric form used in De Brabanter et al. (2013) and Wang and Lin (2015). For this case, it is easy to verify that $d_k = -d_{r-k}$, which eliminates all the even-order derivatives in (2). The sequence is derived for the equidistant design on $[0,1]$. To extend the result to equidistant designs on an arbitrary interval, $[a,b] \subset \mathbb{R}$, we can simply use $d_k/(b-a)$ instead. We treat the DY_i built on $(d_0, \ldots, d_r)_{1,q}$ as the estimator for the first-order derivative with a bias-reduction level of q, denoted by $\hat{m}'_q(x_{i+l})$.

Theorem 1 *Assume that model (1) holds with equidistant design, $m(x)$ has a finite qth-order derivative on $[0,1]$ and $r = o(n)$, $r \to \infty$. For $1 \le i \le n-r$ and $0 \le l \le r$, we have*

$$\begin{aligned} \operatorname{var}[\hat{m}'_q(x_{i+l})] &= n^2 V^{(l)}_{(2,2)} \sigma^2 = O\left(\frac{n^2}{r^3}\right), \\ \operatorname{bias}[\hat{m}'_q(x_{i+l})] &= \frac{1}{q! n^{q-1}} \sum_{j=0}^{q-1} V^{(l)}_{(j+1,2)} I^{(l)}_{j+q} m^{(q)}(x_{i+l}) + o\left(\frac{r^{q-1}}{n^{q-1}}\right). \end{aligned}$$

Proof: see Appendix B.

For a larger q, the order of estimation bias is indeed reduced as expected, and the estimation variance surprisingly retains the same order at the same time. Assuming $r = n^\lambda$ and $2/3 < \lambda < 1$, we can establish the point-wise consistency of our estimator, $\hat{m}'_q(x_{i+l}) \xrightarrow{P} m'(x_{i+l})$, where "$\xrightarrow{P}$" means convergence in probability.

Corollary 1 *Assume that the conditions in Theorem 1 hold. When r is even and $l = r/2$,*

$$\hat{m}'_{2v}(x_{i+r/2}) = \hat{m}'_{2v+1}(x_{i+r/2}), \quad v = 1, 2, \ldots, \left[\frac{q-1}{2}\right],$$

where $[x]$ denotes the greatest integer less than or equal to x.

This means that, when we employ a symmetric form for our estimator, the optimal sequence is the same for $q = 2v$ and $q = 2v+1$. In other words, the symmetric form further reduces the order of estimation bias without any increase in the estimation variance. Hence, it is natural to use the symmetric form (r is even and $l = r/2$) for the interior points, $\{x_i : 1 + r/2 \leq i \leq n - r/2\}$, when the design points are equidistant. Also, we can show that the two existing estimators for the first-order derivative (De Brabanter et al., 2013; Wang and Lin, 2015) are special cases of our method.

When $q = 2$ or $q = 3$, we get the same sequence as

$$(d_k)_{1,2} = (d_k)_{1,3} = \frac{6n(2k-r)}{r(r+1)(r+2)}, \quad k = 0, \ldots, r.$$

This results in the empirical estimator in De Brabanter et al. (2013), denoted by $\hat{m}'_{\text{emp}}$. Assuming the regression function has a finite third-order derivative on $[0, 1]$, the estimation variance and bias are respectively

$$\text{var}[\hat{m}'_2(x_{i+r/2})] = \frac{12n^2\sigma^2}{r(r+1)(r+2)} \quad \text{and} \quad \text{bias}[\hat{m}'_2(x_{i+r/2})] = \frac{r^2}{40n^2} m^{(3)}(x_{i+r/2}) + o\left(\frac{r^2}{n^2}\right).$$

When $q = 4$ or $q = 5$, we get the same sequence as

$$(d_k)_{1,4} = (d_k)_{1,5} = \frac{n\left[I_6^{(r/2)}(k - \frac{r}{2}) - I_4^{(r/2)}(k - \frac{r}{2})^3\right]}{I_2^{(r/2)} I_6^{(r/2)} - {I_4^{(r/2)}}^2}, \quad k = 0, \ldots, r.$$

This results in the least squares estimator in Wang and Lin (2015), denoted by $\hat{m}'_{\text{lse}}$. Within our framework, it is clear that the least squares estimator can be regarded as a bias-reduction modification of the empirical estimator.

Figure 1 presents an example of $m'_q(x_i)$ with different levels of control for the estimation bias ($q = 3$, 5 and 7). We follow the regression function $m(x) = \sqrt{x(1-x)} \cdot \sin\{2.1\pi/(x + 0.05)\}$ for model (1) from De Brabanter et al. (2013) and Wang and Lin (2015). Five hundred design points are equidistant on $[0.25, 1]$ and the random errors are generated from a Gaussian distribution, $N(0, 0.1^2)$. Sequence orders are chosen as $\{50, 100\}$. We observe that the estimation curves are smoother for smaller q, and the bias in oscillating areas decreases significantly for larger q. These results are consistent with our theoretical results. With various levels of bias control, we may achieve a better compromise in the trade-off between the estimation variance and bias.

2.3 Behavior on the boundaries

If we use a sequence with order r, then the boundary region will be $\{x_i : 1 \leq i \leq [r/2]$ or $n - [r/2] + 1 \leq i \leq n.\}$. Within our framework, we have two types of estimators for estimating derivatives for the boundary area. One choice is to use a sequence with smaller order, so that we can still use the symmetric estimator as suggested for the interior points. This solution is also suggested by both De Brabanter et al. (2013) and Wang and Lin (2015). The other is to hold the sequence order while using an asymmetric form of the estimator instead.

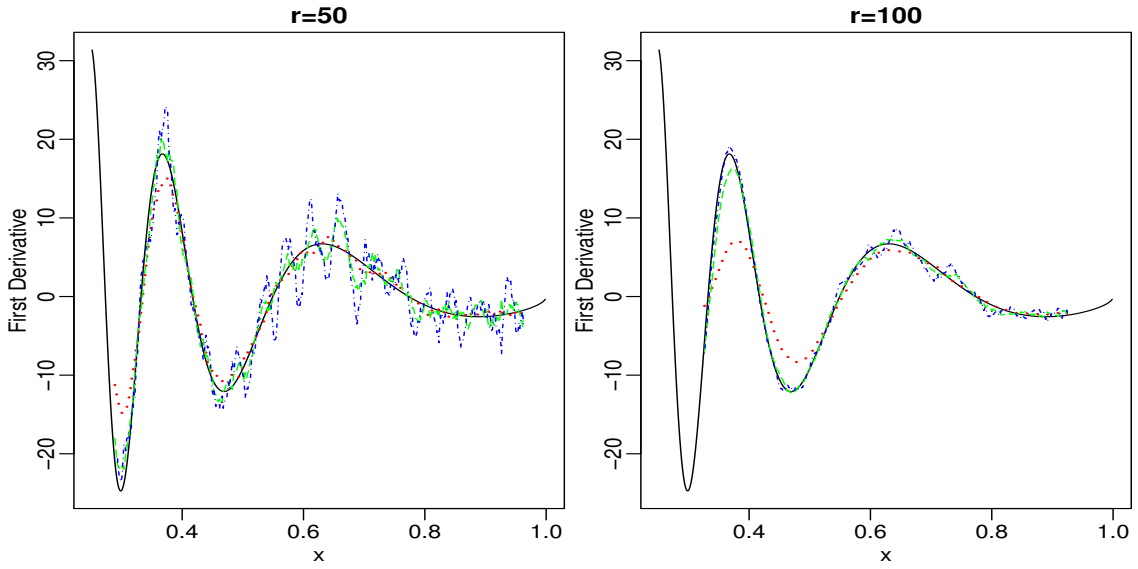

Figure 1: First-order derivative estimates with different levels of bias reduction. Red lines (dotted): $q = 3$; green lines (long dash): $q = 5$; blue lines (dot dash): $q = 7$ and black lines (solid): the true first-order derivative.

For the symmetric estimator, we can choose an even-order t satisfying $1 \leq t/2 \leq \min(i-1, n-i, [r/2])$. By Theorem 1, we have

$$\text{var}[\hat{m}_q'(x_i)] = O\left(\frac{n^2}{t^3}\right) \quad \text{and} \quad \text{bias}[\hat{m}_q'(x_i)] = O\left(\frac{t^{q-1}}{n^{q-1}}\right),$$

for $2 \leq i \leq [r/2]$ or $n - [r/2] + 1 \leq i \leq n - 1$. The closer x_i locates to the endpoints, the smaller the largest order of the chosen sequence, which means that the information we can incorporate into the estimator becomes very limited. As a consequence, the estimation variance will eventually reach an order of $O(n^2)$, which is rather noisy.

The asymmetric estimator does not require the estimated point to be located at the middle of the interval. We can still use a relatively large sequence order to include as much information as included in the interior points. The theoretical results were provided in Theorem 1:

$$\text{var}[\hat{m}_q'(x_i)] = O\left(\frac{n^2}{r^3}\right) \quad \text{and} \quad \text{bias}[\hat{m}_q'(x_i)] = O\left(\frac{r^{q-1}}{n^{q-1}}\right).$$

With a proper choice of r, we can still get a consistent estimate for the derivatives at the boundary region. Another advantage of this asymmetric form is that it is applicable to all the boundary points including x_1 and x_n, which can never be handled by the symmetric-form estimators.

It is noteworthy that Wang and Lin (2015) also proposed left-side and right-side weighted least squares estimators for the boundary points. Their estimators are, however, two special

cases of our asymmetric estimator with $q = 2$ and $l = 0$ (right-side) or $l = r$ (left-side). The estimation bias for $\hat{m}_2'(x_{i+l})$ is

$$\text{bias}[\hat{m}_2'(x_{i+l})] = \frac{r-2l}{2n} m''(x_{i+l}) + o\left(\frac{r}{n}\right).$$

To minimize the estimation bias on these boundary points, we recommend the following criterion:

$$\hat{m}_2'(x_i) = \begin{cases} DY_1 & 1 \le i \le [r/2], \\ DY_{n-r} & n - [r/2] + 1 \le i \le n. \end{cases}$$

Then, the smallest absolute estimation bias can be simply derived as

$$|\mathrm{E}[\hat{m}_2'(x_i)] - m'(x_i)| = \frac{r - 2\min(i-1, n-i)}{2n} |m''(x_i)| + o\left(\frac{r}{n}\right).$$

In summary, the asymmetric estimator generates a smaller variance, while its estimation bias is of a higher order. Consequently, the sequence order should be selected to achieve the best trade-off between the estimation variance and bias. In view of this, we recommend using the asymmetric estimator when the regression function is flat at the boundary region or when σ^2 is large; otherwise, the symmetric form should be employed.

3. Higher-order derivative estimation

In this section, we extend our method and propose higher-order derivative estimators for model (1). We further demonstrate that the new estimators possess the optimal estimation variance, which is not achieved by the two aforementioned methods (De Brabanter et al., 2013; Wang and Lin, 2015). Our new estimators also achieve the optimal convergence rate for MSE.

3.1 Theoretical results

To define an estimator for $m^{(p)}(x_{i+l})$ with a bias-reduction level up to $m^{(q)}(x_{i+l})$, we construct the new conditions on the coefficients as

$$C_{p,l} = 1 \text{ and } C_{j,l} = 0,\ 0 \le j \ne p \le q-1. \tag{4}$$

Then, the optimal sequence can be derived as the solution(s) of the following optimization problem:

$$(d_0, \ldots, d_r)_{p,q} = \underset{(d_0,\ldots,d_r)\in\mathbb{R}^{r+1}}{\text{argmin}} \sum_{k=0}^{r} d_k^2, \quad \text{s.t.} \quad \text{condition (4) holds.}$$

We present the result for $(d_0, \ldots, d_r)_{p,q}$ in the following proposition.

Proposition 2 *Assume that model (1) holds with equidistant design and $m(x)$ has a finite qth-order derivative on $[0,1]$. For $1 \le i \le n-r$ and $0 \le l \le r$, the unique variance minimizing sequence is*

$$(d_k)_{p,q} = p! n^p \sum_{j=0}^{q-1} V^{(l)}_{(j+1,p+1)} (k-l)^j, \quad k = 0, \ldots, r,$$

for estimating $m^{(p)}(x_{i+l})$ with an order of accuracy up to $m^{(q)}(x_{i+l})$, $q \geq p+1$.
Proof: see Appendix C.

To extend the result to equidistant designs on an arbitrary interval, $[a,b] \subset \mathbb{R}$, we can simply use $(d_k)_{p,q}/(b-a)^p$ instead. We treat the DY_i built on $(d_0,\ldots,d_r)_{p,q}$ as the estimator for the pth-order derivative with a bias-reduction level up to $m^{(q)}(x_{i+l})$, denoted as $\hat{m}_q^{(p)}(x_{i+l})$.

Theorem 2 *Assume that model (1) holds with equidistant design, $m(x)$ has a finite qth-order derivative on $[0,1]$ and $r = o(n)$, $r \to \infty$. For $1 \leq i \leq n-r$ and $0 \leq l \leq r$, we have*

$$\begin{aligned}
\operatorname{var}[\hat{m}_q^{(p)}(x_{i+l})] &= (p!)^2 n^{2p} V^{(l)}_{(p+1,p+1)} \sigma^2 = O\left(\frac{n^{2p}}{r^{2p+1}}\right), \\
\operatorname{bias}[\hat{m}_q^{(p)}(x_{i+l})] &= \frac{p!}{q! n^{q-p}} \sum_{j=0}^{q-1} V^{(l)}_{(j+1,p+1)} I^{(l)}_{j+q} m^{(q)}(x_{i+l}) + o\left(\frac{r^{q-p}}{n^{q-p}}\right).
\end{aligned}$$

Proof: see Appendix D.

For a fixed p and an increasing q, we can reduce the estimation bias to a lower order while keeping the order of variance unchanged. Whenever we keep the difference between q and p constant, the convergence rate of the bias is preserved for different p. When r is an even number and $l = r/2$, we can derive that $(d_k)_{p,q} = (-1)^p (d_{n-k})_{p,q}$. Consequently in this case, the optimal sequence remains the same when we increase q from $p+2\nu-1$ to $p+2\nu$, $\nu = 1,2,\ldots$, which means $\hat{m}^{(p)}_{p+2\nu-1}(x_{i+r/2}) = \hat{m}^{(p)}_{p+2\nu}(x_{i+r/2})$. Hence, for this kind of estimator, the symmetric form is also the most favorable choice for the interior points.

The optimal MSE of our estimator is of order $O(n^{-2(q-p)/(2q+1)})$, which achieves the asymptotically optimal rate established by Stone (1980). For comparison, we note that the optimal MSE of the empirical estimator in De Brabanter et al. (2013) is of order $O(n^{-4/(2p+5)})$, that is, their estimator is of the optimal order only when $q = p+2$. While for the least squares estimator in Wang and Lin (2015), they provided asymptotic results only for the first- and second-order derivative estimator. Their optimal MSE is of order $O(n^{-8/11})$ for $p=1$ and $O(n^{-8/13})$ for $p=2$, which corresponds with two special cases, i.e., when $(p,q) = (1,5)$ or $(2,6)$. From this point of view, our method has greatly improved the literature in derivative estimation and it achieves the optimal rate of MSE for any (p,q) from Theorem 2.

As mentioned at the beginning of this section, the newly defined estimator is optimal for the estimation variance, which is superior to existing estimators. In what follows, we illustrate this advantage in detail with the second-order derivative estimator, which is usually of greatest interest after the first-order derivative in practice. A similar analysis can be made for other higher-order derivative estimators. For the estimator without bias-control, e.g. $\hat{m}_4''$, we derive the following results:

$$\begin{aligned}
\operatorname{var}[\hat{m}_4''(x_{i+r/2})] &= \frac{4n^4\sigma^2 I_0^{(r/2)}}{I_0^{(r/2)} I_4^{(r/2)} - I_2^{(r/2)^2}}, \\
\operatorname{bias}[\hat{m}_4''(x_{i+r/2})] &= \frac{r^2}{14n^2} m^{(4)}(x_{i+r/2}) + o\left(\frac{r^2}{n^2}\right).
\end{aligned}$$

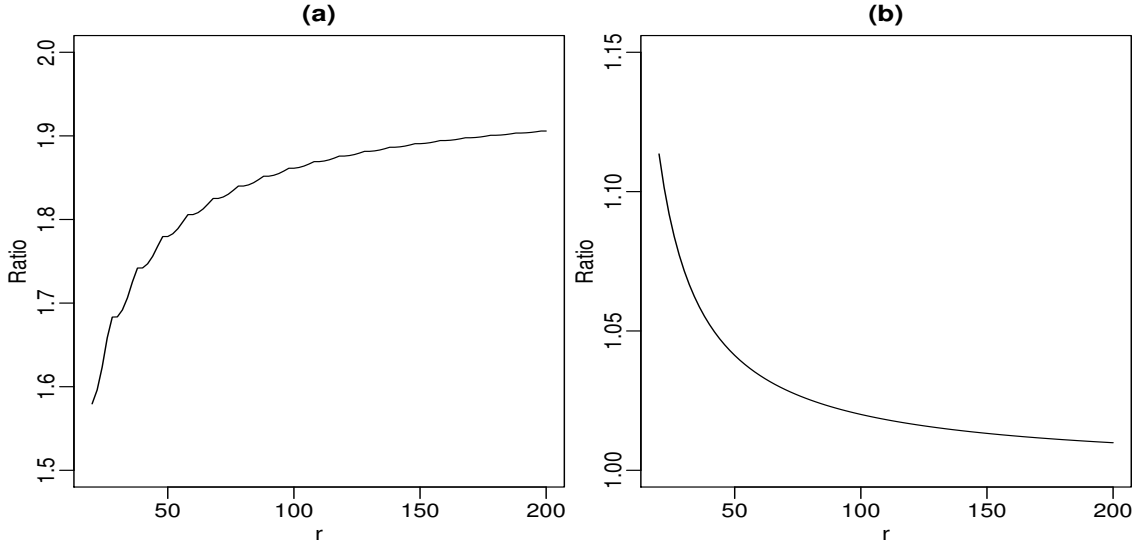

Figure 2: The ratio of estimation variance is plotted against the sequence order, r. Setting: $n = 500$ and r is chosen as an even integer ranging from 20 to 200. (a), $\mathrm{var}(\hat{m}''_{\mathrm{emp}})/\mathrm{var}(\hat{m}''_4)$; (b), $\mathrm{var}(\hat{m}''_{\mathrm{lse}})/\mathrm{var}(\hat{m}''_6)$.

The corresponding method is $\hat{m}''_{\mathrm{emp}}$ in De Brabanter et al. (2013) with regard to the accurate level. Instead of minimizing the estimation variance, they intuitively choose the weight sequences for higher-order derivative estimators, which makes it quite difficult to derive analytical asymptotic results. Hence, we make a finite sample comparison of the variance of the two estimators. We set $n = 500$ and calculate the corresponding sequences for $\hat{m}''_4$ with an even order r ranging from 20 to 200 and $l = r/2$. For $\hat{m}''_{\mathrm{emp}}$, we choose (k_1, k_2), which achieves the smallest estimation variance from $\{(k_1, k_2) : k_1 \leq k_2,\ k_1 + k_2 = r/2\}$. We do not need a specified form of the regression function, since it is not related with the estimation variance. We illustrate the ratio, $\mathrm{var}(\hat{m}''_{\mathrm{emp}})/\mathrm{var}(\hat{m}''_4)$, in the left panel of Figure 2. Obviously, the new estimator improves the estimation variance significantly, which results in a smaller MSE for smooth regression functions.

A similar comparison is carried out between $\hat{m}''_{\mathrm{lse}}$ and $\hat{m}''_6$ under the same settings, and the ratio of $\mathrm{var}(\hat{m}''_{\mathrm{lse}})/\mathrm{var}(\hat{m}''_6)$ is presented in the right panel of Figure 2. Wang and Lin (2015) built a linear model with correlated regressors but employed the weighted least squares regression, rather than the generalized least squares technique, to derive the estimator. It can be shown that our method is equivalent with the generalized least squares estimator for their model. As expected, we find that our proposed estimator performs slightly better in terms of the finite sample than the least squares estimator. In addition their asymptotic variances and biases are equivalent for the first-order term. For the boundary points, our second-order estimator also maintains the same advantages over the existing estimators as discussed in Section 2.2 for the first-order estimators.

3.2 Tuning parameter selection

As shown in Figure 1, the order, r, and the bias-reduction level, q, are both critical to the proposed estimators. For practical implementation, (r, q) should be chosen to achieve a better trade-off between the estimation variance and bias.

By Theorem 2, the approximated MSE of $\hat{m}_q^{(p)}(x_{i+l})$ is

$$\mathrm{MSE}[\hat{m}_q^{(p)}(x_{i+l})] \simeq (p!)^2 n^{2p} V_{(p+1,p+1)}^{(l)} \sigma^2 + \left[\frac{p!}{q! n^{q-p}} \sum_{j=0}^{q-1} V_{(j+1,p+1)}^{(l)} I_{j+q}^{(l)} m^{(q)}(x_{i+l}) \right]^2.$$

We define the averaged mean squared error (AMSE) as a measure of the goodness of fit for all the design points,

$$\mathrm{AMSE}(\hat{m}_q^{(p)}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{MSE}[\hat{m}_q^{(p)}(x_i)].$$

A uniform sequence is preferred for the estimate at most points (all the interior points for example) over different sequences for each design point. Hence, we can choose the parameters (r, q) minimizing the AMSE. To achieve this, we replace the unknown quantities, σ^2 and $m^{(q)}(x_{i+l})$, with their consistent estimates. The error variance can be estimated by the method in Tong and Wang (2005) and Tong et al. (2013) and $m^{(q)}(x_i)$ can be estimated by the local polynomial regression of order $q + 2$. For the high-order derivatives at the boundary points, we recommend replacing the AMSE for all the points with the following adjusted form:

$$\mathrm{AMSE}_{\mathrm{adj}}(\hat{m}_q^{(p)}) = \frac{1}{n-r} \sum_{i=1+r/2}^{n-r/2} \mathrm{MSE}[\hat{m}_q^{(p)}(x_i)] \simeq B_1 \sigma^2 + \frac{B_2}{n-r} \sum_{i=1+r/2}^{n-r/2} [m^{(q)}(x_i)]^2, \qquad (5)$$

where $B_1 = (p!)^2 n^{2p} V_{(p+1,p+1)}^{(r/2)}$ and $B_2 = \left[\frac{p!}{q! n^{q-p}} \sum_{j=0}^{q-1} V_{(j+1,p+1)}^{(r/2)} I_{j+q}^{(r/2)} \right]^2$. Given all the parameters for a specific problem, B_1 and B_2 are available quantities. The adjusted AMSE includes only derivatives at the interior points that share the identical difference sequence for an even r and $l = r/2$. Another advantage is that we only need $V^{(r/2)}$ and $I_{j+q}^{(r/2)}$ instead of $V^{(l)}$ and $I_{j+q}^{(l)}$ for $l = 0, \ldots, r$, which greatly reduces the computation time.

For the tuning parameter space of the sequence order, we recommend $r \in O = \{2i : 1 \leq i \leq k_0\}$, where $k_0 < [n/4]$, to keep a symmetric form ($l = r/2$) for the interior points and to make sure that the number of boundary points will be less than that of the interior points. For the bias-reduction level of $\hat{m}_q^{(p)}$, we consider $q \in Q = \{p + 2\nu : \nu = 1, 2, \ldots, \nu_0\}$, where $p + 2\nu_0$ is the highest level chosen by users. Only even differences are considered for $q - p$, since $\hat{m}_{p+2\nu_0-1}^{(p)} = \hat{m}_{p+2\nu_0}^{(p)}$ when we use the recommended symmetric form.

4. Simulation study

In this section, we conduct simulation studies to assess the finite sample performance of the proposed estimators, $\hat{m}_q^{(p)}$, and make comparisons with the empirical estimator, $\hat{m}_{\mathrm{emp}}^{(p)}$,

in De Brabanter et al. (2013) and the least squares estimator, $\hat{m}_{\text{lse}}^{(p)}$, in Wang and Lin (2015). We apply the three methods to both interior (Int) and boundary (Bd) areas, where Int $= \{x_i : k_0 + 1 \leq i \leq n - k_0\}$ and Bd $= \{x_i : 1 \leq i \leq k_0 \text{ or } n - k_0 + 1 \leq i \leq n\}$. Throughout the simulation, we set $k_0 = [n/10]$, which means that we treat ten percent of the design points on both sides of the interval as boundary points. We also tried some other proportions and the results were similar. For the interior part, we keep the symmetric form for $\hat{m}_q^{(p)}$ by setting r as an even number and $l = r/2$, as suggested in the theoretical results. For the boundary part, we apply the following criterion for the proposed estimators:

$$\hat{m}_q^{(p)}(x_i) = \begin{cases} DY_1 & 1 \leq i \leq [r/2], \\ DY_{n-r} & n - [r/2] + 1 \leq i \leq n. \end{cases}$$

The modified version of $\hat{m}_{\text{emp}}^{(p)}$ in De Brabanter et al. (2013) and the one-side weighted least squares estimators in Wang and Lin (2015) are investigated for the empirical and least squares estimators, respectively on the boundary points. We consider estimators for both first- and second-order derivatives, which are of most interest in practice. Similar to De Brabanter et al. (2013) and Wang and Lin (2015), the mean absolute error (MAE) is used as a measure of estimation accuracy. It is defined as follows:

$$\text{MAE} = \frac{1}{\#\mathbb{A}} \sum_{x_i \in \mathbb{A}} |\hat{m}^{(p)}(x_i) - m^{(p)}(x_i)|,$$

where $\mathbb{A} =$ Int or Bd and $\#\mathbb{A}$ denotes the number of elements in set $\mathbb{A}$.

We consider the following regression function,

$$m(x) = 5\sin(w\pi x),$$

with $\omega = 1,\ 2,\ 4$ corresponding to different levels of oscillations. The $n = 100$ and 500 sample sizes are investigated. We set the design points as $x_i = i/n$ and generate the random errors, ε_i, independently from $N(0, \sigma^2)$. For each regression function, we consider $\sigma = 0.1,\ 0.5$ and 2 to capture the small, moderate and large variances, respectively. In total, we have 18 combinations of simulation settings. Following the definitions of Int and Bd, we select the sequence order r from $\mathbb{O} = \{2i : 1 \leq i \leq k_0\}$. We choose the bias-reduction level, q, from $\mathbb{Q} = \{p+2, p+4, p+6\}$, with $q = p+2$ and $q = p+4$ corresponding to $\hat{m}_{\text{emp}}^{(p)}$ and $\hat{m}_{\text{lse}}^{(p)}$, respectively, and $q = p+6$ as an even higher level. We denote by $\hat{m}_{\text{opt}}^{(p)}$ the estimator with the selected tuning parameters. For $\hat{m}_{\text{emp}}^{(p)}$ and $\hat{m}_{\text{lse}}^{(p)}$, the parameter k is chosen from $\{i : 1 \leq i \leq k_0\}$. We investigate two scenarios (for the tuning parameters selection criterion): *oracle* and *plug-in* (see below). For each run of the simulation, we compute the MAE of the estimators at both Int and Bd and repeat the procedure 1000 times for each setting. The simulation results for $w = 2$ are reported as box-plot figures. Other results are provided in the supplementary materials.

Oracle parameters

Oracle parameters are selected by assuming that we know the true regression (derivative) function, the purpose of which is to illustrate the possible best performance of each

estimator. Specifically for $\hat{m}_q^{(p)}$, the pair of tuning parameters is chosen as

$$(r, q)_{\text{opt}} = \underset{\text{r}\in\mathbb{O},\text{q}\in\mathbb{Q}}{\text{argmin}} \left(\text{MAE}(\hat{m}_q^{(p)}) \right).$$

The bandwidths of $\hat{m}_{\text{emp}}^{(p)}$ and $\hat{m}_{\text{lse}}^{(p)}$ are selected through a similar procedure.

For the first-order derivative, we investigate $\hat{m}'_{\text{opt}}$, $\hat{m}'_{\text{emp}}$ and $\hat{m}'_{\text{lse}}$ and report the simulation results in Figure 3. On the interior points, $\hat{m}'_{\text{opt}}$ always possesses the same MAE as the smaller one of $\hat{m}'_{\text{emp}}$ and $\hat{m}'_{\text{lse}}$, due to the fact that $\hat{m}'_{\text{emp}}$ and $\hat{m}'_{\text{lse}}$ are two special cases of $\hat{m}'_q$ in this area. On the boundary points, $\hat{m}'_{\text{opt}}$ is uniformly better than the other two methods. To further explore the reason for the boundary behavior, we use an example from De Brabanter et al. (2013) and Wang and Lin (2015). The fitted results for the three estimators are illustrated in Figure 4, where the red points represent the boundary parts. The empirical estimator suffers a lot from the increasing variance when the estimated points get close to the endpoints of the interval. The least squares estimator simply estimates the boundary parts by shifting the estimates of the interior points nearby, which results in very serious estimation bias. Our estimator fits the boundary points very well, resulting from the flexibility brought by the parameter l, the relative location of the estimated point within the interval $[x_i, x_{i+r}]$.

For the second-order derivative, we include another two estimators, $\hat{m}''_4$ and $\hat{m}''_6$, which have the same bias-reduction level as $\hat{m}''_{\text{emp}}$ and $\hat{m}''_{\text{lse}}$, respectively. The sequence order, r, of the two additional estimators is optimally chosen by minimizing MAE as well. The simulation results are presented in Figure 5. The relationships between $\hat{m}''_{\text{opt}}$, $\hat{m}''_{\text{emp}}$ and $\hat{m}''_{\text{lse}}$ remain the same as those observed for the first-order derivative. We also observe that $\text{MAE}(\hat{m}''_4)$ is significantly smaller than $\text{MAE}(\hat{m}''_{\text{emp}})$ and that $\text{MAE}(\hat{m}''_6)$ is almost the same as $\text{MAE}(\hat{m}''_{\text{lse}})$, consistent with our theoretical results in Section 3.

Plug-in parameters

Plug-in parameters are chosen via minimizing the adjusted AMSE in (5) after replacing all the unknown quantities with their consistent estimates. In this simulation, we estimate σ^2 using Tong and Wang's (2005) method with the recommended bandwidth $[n^{1/3}]$. Here, $\hat{m}^{(q)}(x_i)$ $(1 + k_0 \leq i \leq n - k_0)$ are calculated with the function *locpol* in the R package locpol (Ojeda Cabrera, 2012) with the parameter $deg = q + 2$. The bandwidths of $\hat{m}_{\text{emp}}^{(p)}$ and $\hat{m}_{\text{lse}}^{(p)}$ are selected accordingly.

We report the simulation results together with those for the oracle parameters in Figures 6 and 7. From the comparison, we observe that the plug-in parameters lead to quite similar results with those for the oracle parameters, especially on the interior points. Since the tuning parameters are selected based on AMSE of derivative estimates for the interior points, the performance on the boundary is not consistent. Nevertheless, the mutual relationship of the three estimators remains the same for most cases on both interior and boundary points. Overall, the proposed plug-in method is quite effective for choosing the optimal tuning parameters.

In summary, we have demonstrated the superiority of the proposed estimators over the existing estimators through extensive simulation studies. We have further provided an effective criterion for selection of the tuning parameters for the newly defined estimator.

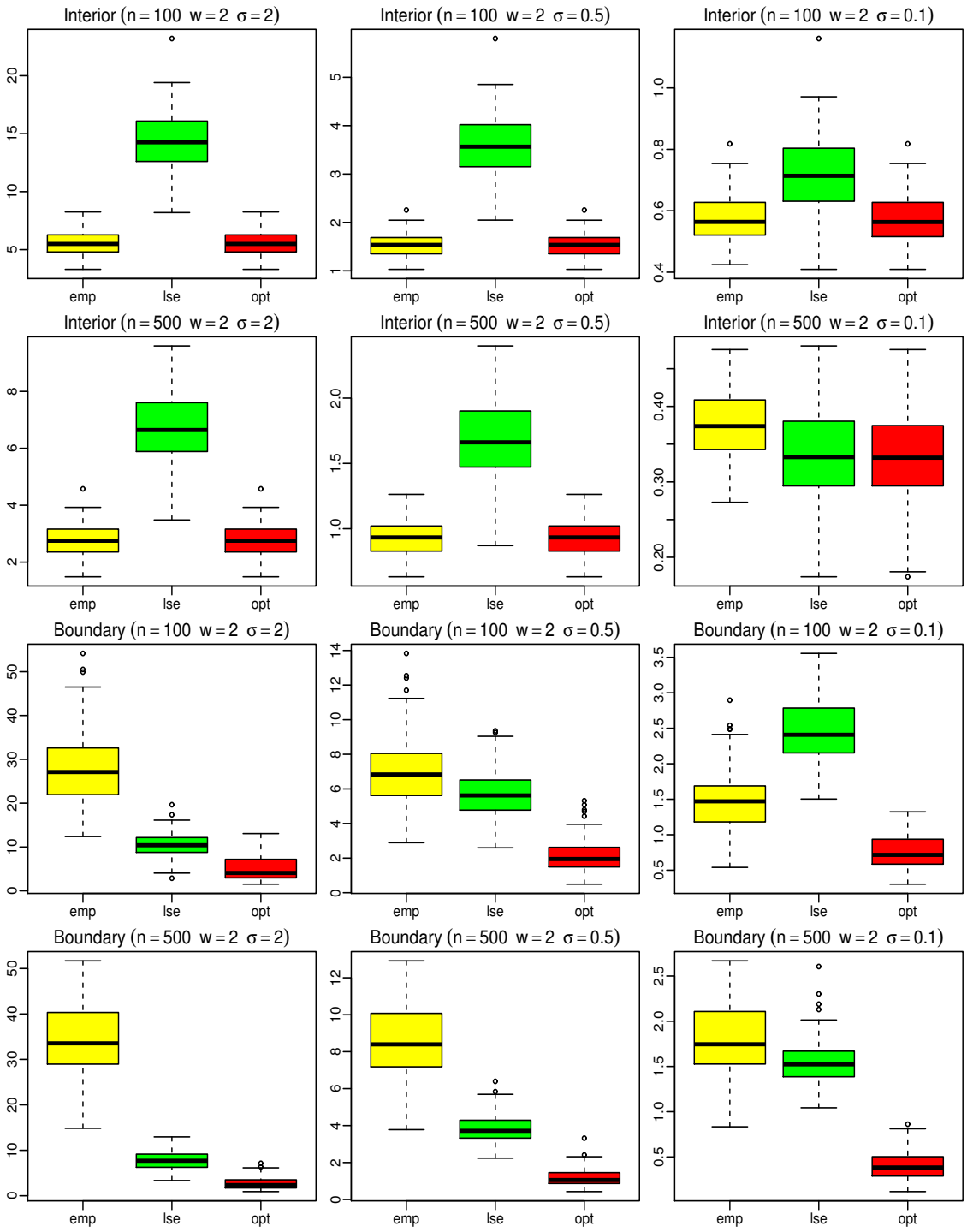

Figure 3: Mean absolute errors of three first-order derivative estimators on both interior (2 top rows) and boundary (2 bottom rows) points for various settings. $\hat{m}'_{\text{emp}}$, yellow box; $\hat{m}'_{\text{lse}}$, green box; $\hat{m}'_{\text{opt}}$, red box. $m(x) = 5\sin(2\pi x)$ and $\varepsilon \sim N(0, \sigma^2)$.

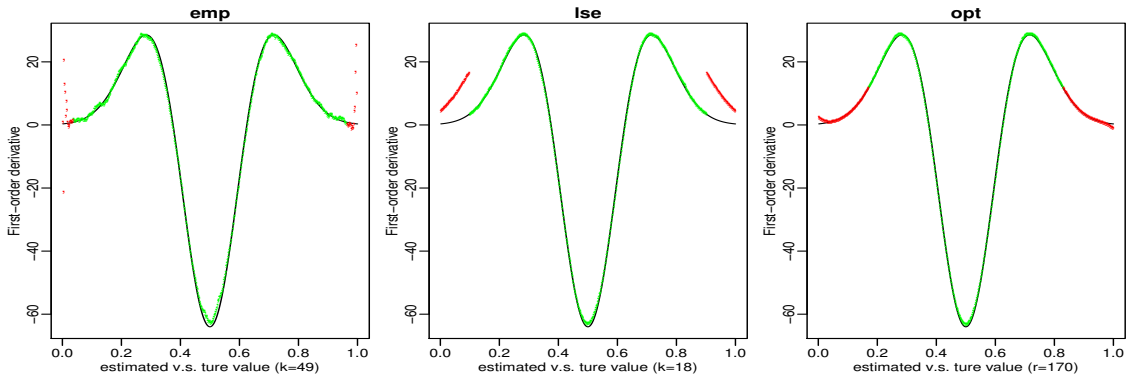

Figure 4: The fitted point-wise derivatives by the three estimators using oracle tuning parameters. $m(x) = 32e^{-8(1-2x)^2}(1-2x)$, ε_i are independent random errors from $N(0, 0.1^2)$ and $n = 500$. Interior points: green points. Boundary points: red points.

5. Conclusion

We proposed a new framework for estimating derivatives without fitting the regression function. Unlike most existing methods using the symmetric difference quotients, our method is constructed as a linear combination of the observations. It is hence very flexible and applicable to both interior and boundary points. We obtained the variance-minimizing estimators for the first- and higher-order derivatives with a fixed bias-reduction level. Under the equidistant design, we derived some theoretical results for the proposed estimators including the optimal sequence, asymptotic variance and bias, point-wise consistency, and boundary behavior. We illustrated that the order of the estimation bias can be reduced while the order of variance remains unchanged. We showed that our method achieves the optimal convergence rate for the MSE. Furthermore, we provided an effective selection procedure for the tuning parameters of the proposed estimators. Simulation studies for the first- and second-order derivative estimators demonstrated the superiority of our proposed method.

The method can be readily extended to unequally spaced designs. In this case, the symmetric form is no longer valid and the choice of l also deserves further consideration. To estimate the point-wise derivatives for unequally spaced designs, we can first find the r nearest neighbors of the estimated point and construct the variance-minimizing estimator with the linear combination of the $r+1$ points, say $x_i < \cdots < x_{i+l} < \cdots < x_{i+r}$. Assuming that $m(x)$ is smooth enough and that x_{i+l} is the estimated point, we have the expectation of DY_i as

$$\mathrm{E}(DY_i) \quad = \quad m(x_{i+l}) \sum_{k=0}^{r} d_k + \sum_{j=1}^{\infty} m^{(j)}(x_{i+l}) \sum_{k=0}^{r} d_k (x_{i+k} - x_{i+l})^j / j!.$$

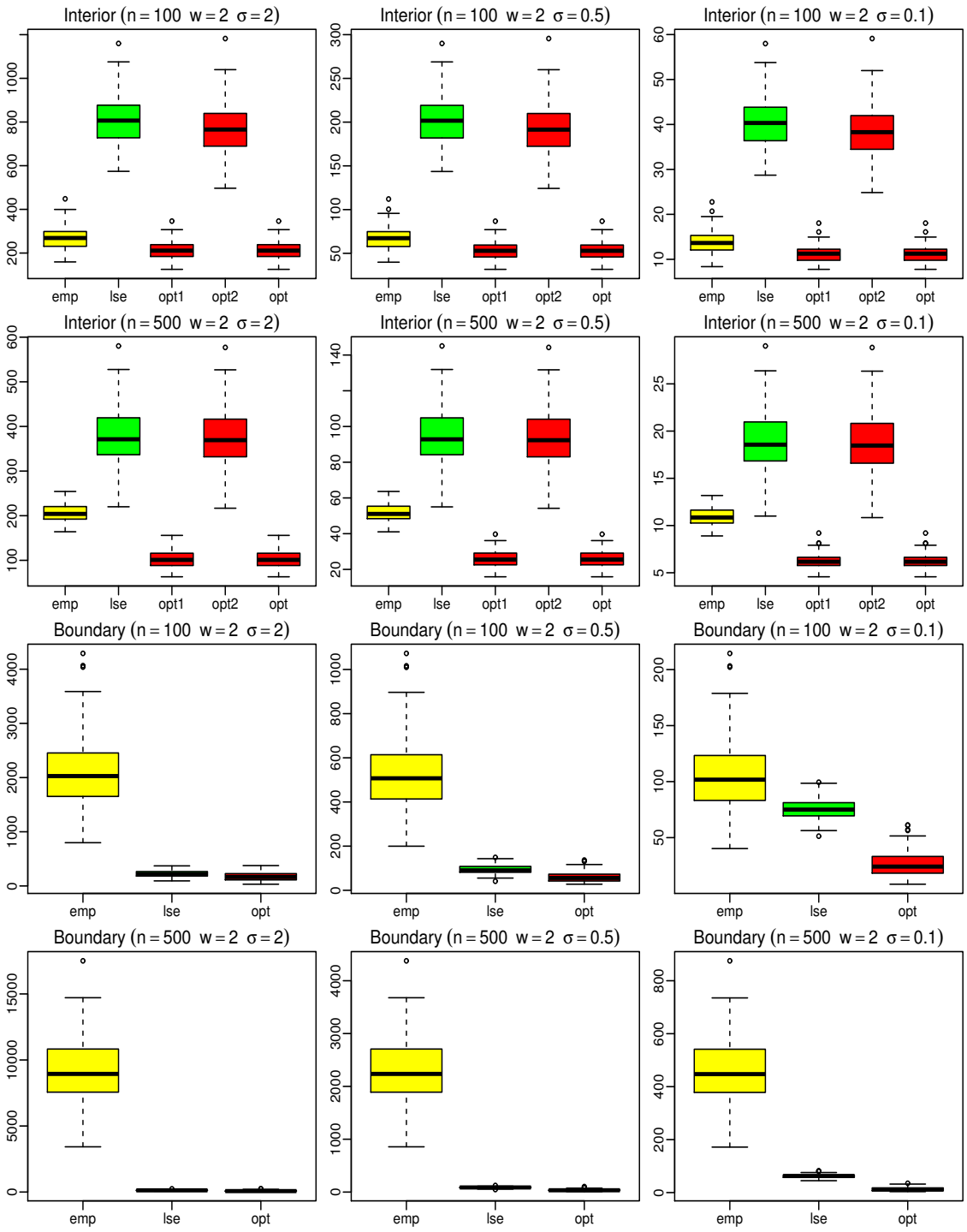

Figure 5: Mean absolute errors of three second-order derivative estimators on both interior (2 top rows) and boundary (2 bottom rows) points for various settings. $\hat{m}'_{\text{emp}}$, yellow box; $\hat{m}'_{\text{lse}}$, green box; $\hat{m}''_q$, red box. opt1: $\hat{m}''_4$; opt2: $\hat{m}''_6$; opt: $\hat{m}''_{\text{opt}}$. $m(x) = 5\sin(2\pi x)$ and $\varepsilon \sim N(0, \sigma^2)$.

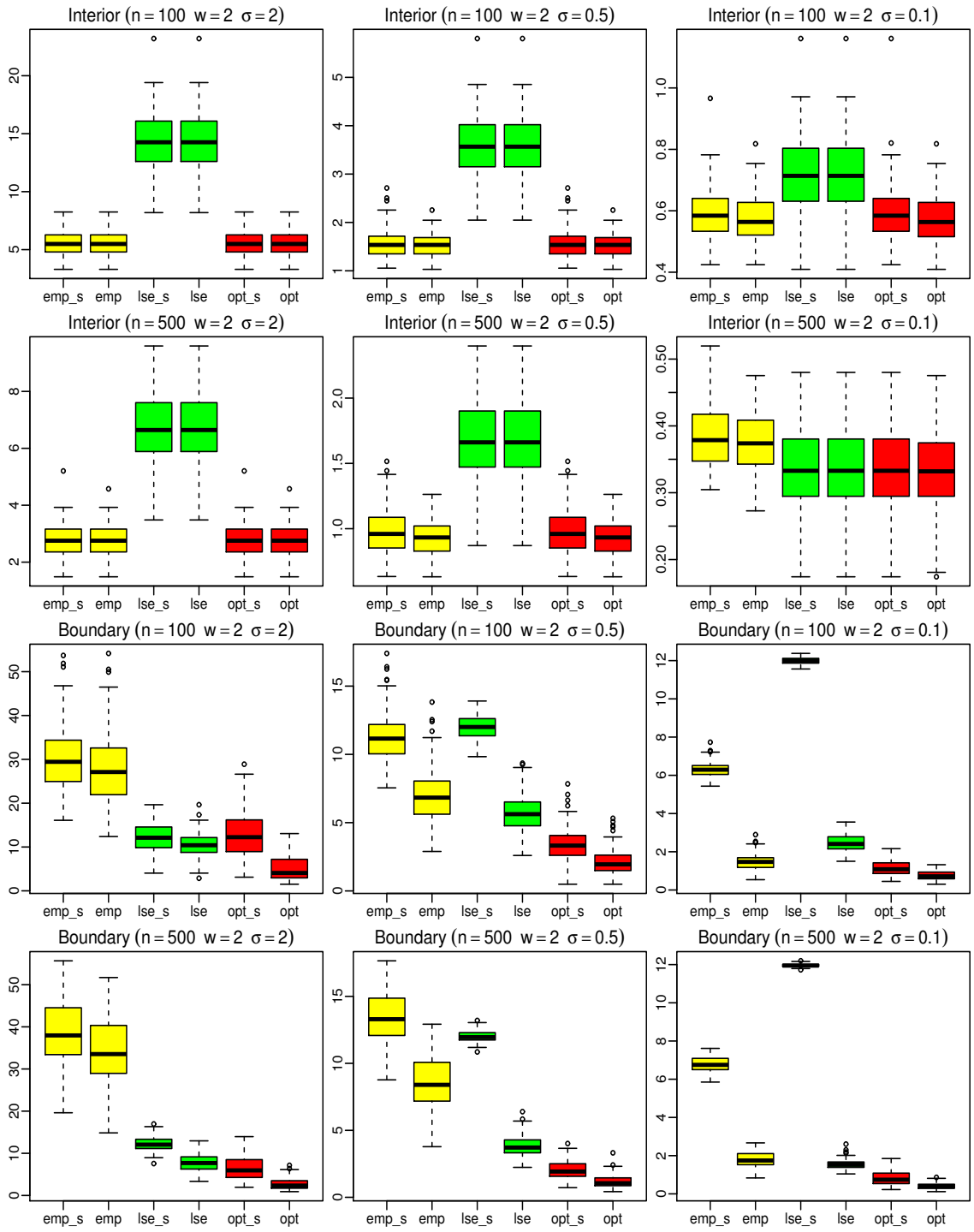

Figure 6: Comparison of the mean absolute errors on both interior (2 top rows) and boundary (2 bottom rows) points between the first-order derivative estimators with oracle tuning parameters and those with plug-in tuning parameters. $\hat{m}'_{\text{emp}}$, yellow box; $\hat{m}'_{\text{lse}}$, green box; $\hat{m}'_{\text{opt}}$, red box. "_s" denotes estimators using plug-in parameters. $m(x) = 5\sin(2\pi x)$ and $\varepsilon \sim N(0, \sigma^2)$.

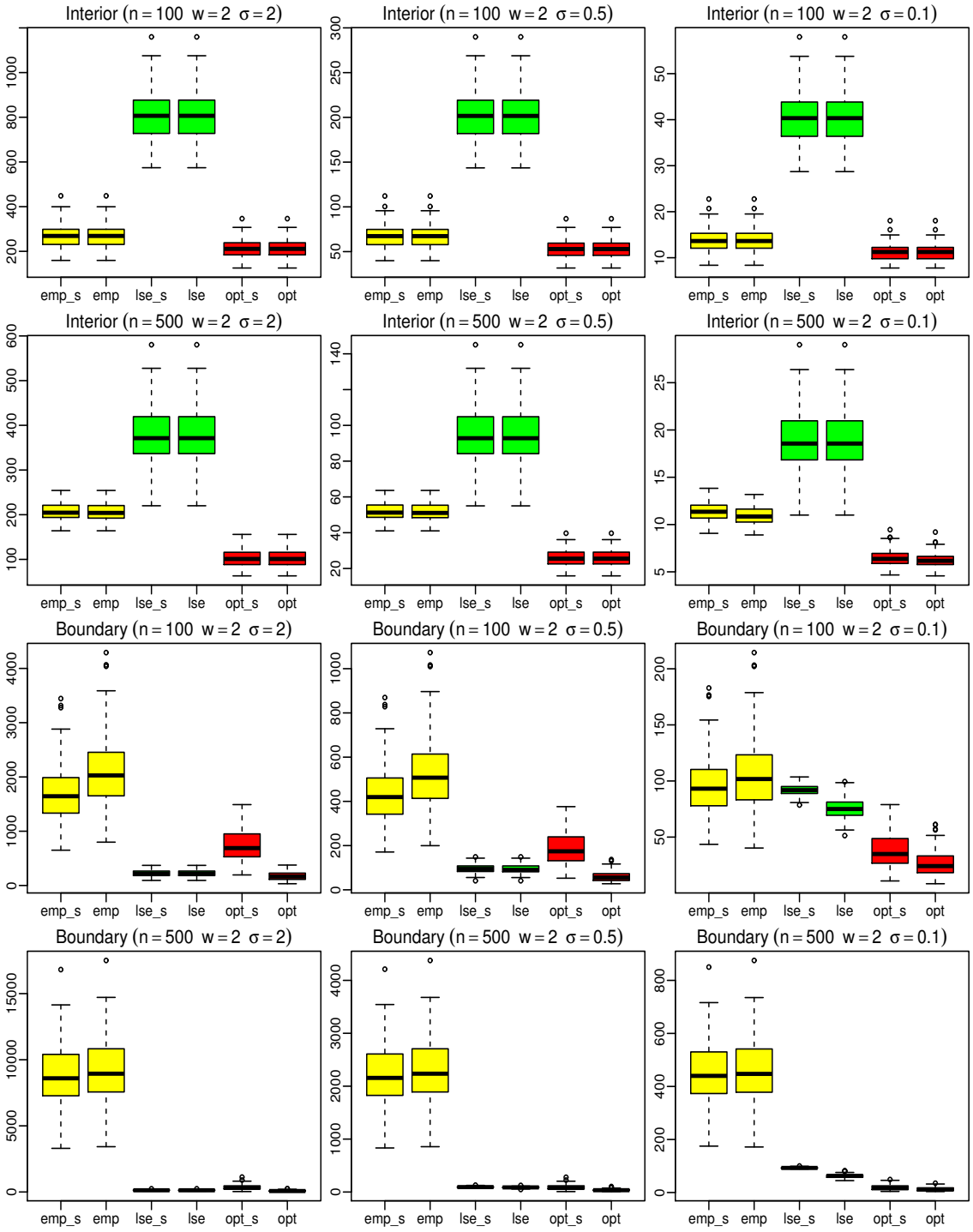

Figure 7: Comparison of the mean absolute errors on both interior (2 top rows) and boundary (2 bottom rows) points between the second-order derivative estimators with oracle tuning parameters and those with plug-in tuning parameters. $\hat{m}''_{\text{emp}}$, yellow box; $\hat{m}''_{\text{lse}}$, green box; $\hat{m}''_{\text{opt}}$, red box. "_s" denotes estimators using plug-in parameters. $m(x) = 5\sin(2\pi x)$ and $\varepsilon \sim N(0, \sigma^2)$.

The optimal sequence for estimating $m^{(p)}(x_{i+l})$ with a bias-reduction level q can be decided by solving the following optimization problem:

$$(d_0,\ldots,d_r)_{p,q} = \underset{(d_0,\ldots,d_r)\in\mathbb{R}^{r+1}}{\operatorname{argmin}} \sum_{k=0}^{r} d_k^2,$$
$$\text{s.t. } \sum_{k=0}^{r} d_k \frac{(x_{i+k}-x_{i+l})^j}{j!} = 0,\ j=0,\ldots,p-1,p+1,\ldots,q-1,\ \sum_{k=0}^{r} d_k \frac{(x_{i+k}-x_{i+l})^p}{p!} = 1.$$

The optimal difference sequences are adaptively chosen for each estimated point and they are no longer identical for all the interior design points. As a result, the parameter selection becomes more challenging and we leave this for future research. Finally, other models worthy of investigation include, for example, random design models (De Brabanter and Liu, 2015) and multivariate models (Charnigo et al., 2015; Charnigo and Srinivasan, 2015).

Acknowledgments

The authors thank the editor, the associate editor and the two referees for their constructive comments that led to a substantial improvement of the paper. The work of Wenlin Dai and Marc G. Genton was supported by King Abdullah University of Science and Technology (KAUST). Tiejun Tong's research was supported in part by Hong Kong Baptist University FRG grants FRG1/14-15/044, FRG2/15-16/038, FRG2/15-16/019 and FRG2/14-15/084.

Appendix A. Proof of Proposition 1

To find the optimal sequence for estimating the first-order derivative with qth-order accuracy, we solve the following optimization problem:

$$(d_0,\ldots,d_r)_{1,q} = \underset{(d_0,\ldots,d_r)\in\mathbb{R}^{r+1}}{\operatorname{argmin}} \sum_{k=0}^{r} d_k^2, \quad \text{s.t.} \quad \text{condition (3) holds.}$$

It is easy to check that condition (3) is equivalent to

$$\sum_{k=0}^{r} d_k(k-l) = n \text{ and } \sum_{k=0}^{r} d_k(k-l)^j = 0,\ 0 \le j \ne 1 \le q-1.$$

To apply the Lagrange multipliers method to find the optimal sequence, we transform the above problem in the following unconstrained optimization problem:

$$f(d_0,\ldots,d_r,\lambda_0,\ldots,\lambda_{q-1}) = \sum_{k=0}^{r} d_k^2 + \lambda_0 C_0 + \sum_{j=2}^{q-1}\lambda_j \sum_{k=0}^{r} d_k(k-l)^j + \lambda_1\left[\sum_{k=0}^{r} d_k(k-l) - n\right].$$

Taking the partial derivative of f with respect to each parameter and setting it to zero, we have

$$\begin{aligned}
\frac{\partial f}{\partial d_k} &= 2d_k + \lambda_0 + \sum_{j=1}^{q-1} \lambda_j (k-l)^j = 0, \quad k = 0, \ldots, r, \qquad (6)\\
\frac{\partial f}{\partial \lambda_j} &= \sum_{k=0}^{r} d_k (k-l)^j = 0, \quad 0 \le j \ne 1 \le q-1,\\
\frac{\partial f}{\partial \lambda_1} &= \sum_{k=0}^{r} d_k (k-l) = n.
\end{aligned}$$

We further make the following transformation:

$$\begin{aligned}
\sum_{k=0}^{r} (k-l)^i \frac{\partial f}{\partial d_k} &= 2\sum_{k=0}^{r} d_k (k-l)^i + \lambda_0 \sum_{k=0}^{r} (k-l)^i + \sum_{j=1}^{q-1} \lambda_j \sum_{k=0}^{r} (k-l)^{i+j}\\
&= I_i^{(l)} \lambda_0 + \sum_{j=1}^{q-1} I_{i+j}^{(l)} \lambda_j = 0, \quad 0 \le i \ne 1 \le q-1.\\
\sum_{k=0}^{r} (k-l) \frac{\partial f}{\partial d_k} &= 2\sum_{k=0}^{r} d_k (k-l) + \lambda_0 \sum_{k=0}^{r} (k-l) + \sum_{j=1}^{q-1} \lambda_j \sum_{k=0}^{r} (k-l)^{1+j}\\
&= 2n + I_1^{(l)} \lambda_0 + \sum_{j=1}^{q-1} I_{1+j}^{(l)} \lambda_j = 0,
\end{aligned}$$

where $I_i^{(l)} = \sum_{k=0}^{r} (k-l)^i$ for $i = 1, 2, \ldots$.

These results can be expressed as a matrix equation,

$$U^{(l)} (\lambda_0, \ldots, \lambda_{q-1})' = -2n\epsilon_2,$$

where $U^{(l)}$ is a $q \times q$ matrix with $u_{ij}^{(l)} = I_{i+j-2}^{(l)}$ and ϵ_2 is a $q \times 1$ vector with the second element equal to 1 and the others equal to zero. Noting that $U^{(l)}$ is an invertible matrix, we have

$$(\lambda_0, \ldots, \lambda_{q-1})' = -2nV_{(.,2)}^{(l)},$$

where $V^{(l)} = (U^{(l)})^{-1}$ and $V_{(.,2)}^{(l)}$ denotes the second column of $V^{(l)}$. This leads to $\lambda_j = -2p!n^p V_{(j+1,2)}^{(l)}$ for $j = 0, \ldots, q-1$. Combining this result with (6), we get

$$(d_k)_{1,q} = n \sum_{j=0}^{q-1} V_{(j+1,2)}^{(l)} (k-l)^j, \quad k = 0, \ldots, r.$$

This completes the proof of Proposition 1. □

Appendix B. Proof of Theorem 1

We can easily derive that

$$\begin{aligned}
\mathrm{var}[\hat{m}_q'(x_{i+l})] &= \sigma^2 \sum_{k=0}^{r} d_k^2 = \sigma^2 \sum_{k=0}^{r} d_k \left[n \sum_{j=0}^{q-1} V_{(j+1,2)}^{(l)} (k-l)^j \right] \\
&= \sigma^2 n \sum_{j=0}^{q-1} V_{(j+1,2)}^{(l)} \sum_{k=0}^{r} d_k (k-l)^j = \sigma^2 n V_{(2,2)}^{(l)} \sum_{k=0}^{r} d_k (k-l) \\
&= \sigma^2 n^2 V_{(2,2)}^{(l)}, \\
\mathrm{bias}[\hat{m}_q'(x_{i+l})] &= C_{q,l} m^{(q)}(x_{i+l}) + o(r^{q-1}/n^{q-1}),
\end{aligned}$$

where

$$\begin{aligned}
C_{q,l} &= \sum_{k=0}^{r} d_k \frac{(k-l)^q}{n^q q!} = \frac{n}{q! n^q} \sum_{k=0}^{r} \left[\sum_{j=0}^{q-1} V_{(j+1,2)}^{(l)} (k-l)^j \right] (k-l)^q \\
&= \frac{1}{q! n^{q-1}} \sum_{j=0}^{q-1} V_{(j+1,2)}^{(l)} I_{j+q}^{(l)} = O(r^{q-1}/n^{q-1}).
\end{aligned}$$

This completes the proof of Theorem 1. □

Appendix C. Proof of Proposition 2

To find the optimal sequence for estimating the pth-order derivative with qth-order accuracy, we solve the following optimization problem:

$$(d_0, \ldots, d_r)_{p,q} = \underset{(d_0,\ldots,d_r)\in\mathbb{R}^{r+1}}{\operatorname{argmin}} \sum_{k=0}^{r} d_k^2, \quad \text{s.t.} \quad \text{condition (4) holds.}$$

It is easy to check that condition (4) is equivalent to

$$\sum_{k=0}^{r} d_k (k-l)^p = p! n^p \text{ and } \sum_{k=0}^{r} d_k (k-l)^j = 0,\ 0 \le j \ne p \le q-1.$$

To apply the Lagrange multipliers method to find the optimal sequence, we transform the above problem in the following unconstrained optimization problem:

$$\begin{aligned}
f(d_0, \ldots, d_r, \lambda_0, \ldots, \lambda_{q-1}) &= \sum_{k=0}^{r} d_k^2 + \lambda_0 C_0 + \left(\sum_{j=1}^{p-1} + \sum_{j=p+1}^{q-1} \right) \lambda_j \sum_{k=0}^{r} d_k (k-l)^j \\
&\quad + \lambda_p \left[\sum_{k=0}^{r} d_k (k-l)^p - p! n^p \right].
\end{aligned}$$

Taking the partial derivative of f with respect to each parameter and setting it to zero, we have

$$\begin{aligned}
\frac{\partial f}{\partial d_k} &= 2d_k + \lambda_0 + \sum_{j=1}^{q-1} \lambda_j (k-l)^j = 0, \quad k = 0, \ldots, r, \qquad (7)\\
\frac{\partial f}{\partial \lambda_j} &= \sum_{k=0}^{r} d_k (k-l)^j = 0, \quad 0 \le j \ne p \le q-1,\\
\frac{\partial f}{\partial \lambda_p} &= \sum_{k=0}^{r} d_k (k-l)^p = p! n^p.
\end{aligned}$$

We further make the following transformation:

$$\begin{aligned}
\sum_{k=0}^{r} (k-l)^i \frac{\partial f}{\partial d_k} &= 2\sum_{k=0}^{r} d_k (k-l)^i + \lambda_0 \sum_{k=0}^{r} (k-l)^i + \sum_{j=1}^{q-1} \lambda_j \sum_{k=0}^{r} (k-l)^{i+j}\\
&= I_i^{(l)} \lambda_0 + \sum_{j=1}^{q-1} I_{i+j}^{(l)} \lambda_j = 0, \quad 0 \le i \ne p \le q-1.\\
\sum_{k=0}^{r} (k-l)^p \frac{\partial f}{\partial d_k} &= 2\sum_{k=0}^{r} d_k (k-l)^p + \lambda_0 \sum_{k=0}^{r} (k-l)^p + \sum_{j=1}^{q-1} \lambda_j \sum_{k=0}^{r} (k-l)^{p+j}\\
&= 2p! n^p + I_p^{(l)} \lambda_0 + \sum_{j=1}^{q-1} I_{p+j}^{(l)} \lambda_j = 0,
\end{aligned}$$

where $I_i^{(l)} = \sum_{k=0}^{r} (k-l)^i$ for $i = 1, 2, \ldots$.

These results can be expressed as a matrix equation,

$$U^{(l)} (\lambda_0, \ldots, \lambda_{q-1})' = -2p! n^p \epsilon_{p+1},$$

where $U^{(l)}$ is a $q \times q$ matrix with $u_{ij}^{(l)} = I_{i+j-2}^{(l)}$ and ϵ_{p+1} is a $q \times 1$ vector with the $(p+1)$th element equal to 1 and the others equal to zero. Noting that $U^{(l)}$ is an invertible matrix, we have

$$(\lambda_0, \ldots, \lambda_{q-1})' = -2p! n^p V_{(.,p+1)}^{(l)},$$

where $V^{(l)} = (U^{(l)})^{-1}$ and $V_{(.,p+1)}^{(l)}$ denotes the $(p+1)$th column of $V^{(l)}$. This leads to $\lambda_j = -2p! n^p V_{(j+1,p+1)}^{(l)}$ for $j = 0, \ldots, q-1$. Combining this result with (7), we get

$$(d_k)_{p,q} = p! n^p \sum_{j=0}^{q-1} V_{(j+1,p+1)}^{(l)} (k-l)^j, \quad k = 0, \ldots, r.$$

This completes the proof of Proposition 2. □

Appendix D. Proof of Theorem 2

We can easily derive that

$$\begin{aligned}
\mathrm{var}[\hat{m}_q^{(p)}(x_{i+l})] &= \sigma^2 \sum_{k=0}^{r} d_k^2 = \sigma^2 \sum_{k=0}^{r} d_k \left[p! n^p \sum_{j=0}^{q-1} V_{(j+1,p+1)}^{(l)} (k-l)^j \right] \\
&= \sigma^2 p! n^p \sum_{j=0}^{q-1} V_{(j+1,p+1)}^{(l)} \sum_{k=0}^{r} d_k (k-l)^j = \sigma^2 p! n^p V_{(p+1,p+1)}^{(l)} \sum_{k=0}^{r} d_k (k-l)^p \\
&= \sigma^2 (p!)^2 n^{2p} V_{(p+1,p+1)}^{(l)}, \\
\mathrm{bias}[\hat{m}_q^{(p)}(x_{i+l})] &= C_{q,l} m^{(q)}(x_{i+l}) + o(r^{q-p}/n^{q-p}),
\end{aligned}$$

where

$$\begin{aligned}
C_{q,l} &= \sum_{k=0}^{r} d_k \frac{(k-l)^q}{n^q q!} = \frac{p! n^p}{q! n^q} \sum_{k=0}^{r} \left[\sum_{j=0}^{q-1} V_{(j+1,p+1)}^{(l)} (k-l)^j \right] (k-l)^q \\
&= \frac{p!}{q! n^{q-p}} \sum_{j=0}^{q-1} V_{(j+1,p+1)}^{(l)} I_{j+q}^{(l)} = O(r^{q-p}/n^{q-p}).
\end{aligned}$$

This completes the proof of Theorem 2. □

References

Graciela Boente and Daniela Rodriguez. Robust estimators of high order derivatives of regression functions. *Statistics and Probability Letters*, 76(13):1335–1344, 2006.

Guanqun Cao. Simultaneous confidence bands for derivatives of dependent functional data. *Electronic Journal of Statistics*, 8(2):2639–2663, 2014.

Richard Charnigo and Cidambi Srinivasan. A multivariate generalized Cp and surface estimation. *Biostatistics*, 16(2):311–325, 2015.

Richard Charnigo, Benjamin Hall, and Cidambi Srinivasan. A generalized Cp criterion for derivative estimation. *Technometrics*, 53(3):238–253, 2011.

Richard Charnigo, Limin Feng, and Cidambi Srinivasan. Nonparametric and semiparametric compound estimation in multiple covariates. *Journal of Multivariate Analysis*, 141: 179–196, 2015.

Kris De Brabanter and Yu Liu. *Smoothed nonparametric derivative estimation based on weighted difference sequences,* Stochastic Models, Statistics and Their Applications, A. Steland and E. Rafajłowicz and K. Szajowski (Eds.), Chapter 4 (31-38). Springer, 2015.

Kris De Brabanter, Jos De Brabanter, Bart De Moor, and Irène Gijbels. Derivative estimation with local polynomial fitting. *Journal of Machine Learning Research*, 14(1):281–301, 2013.

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

Jianqing Fan and Irène Gijbels. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society Series B*, 57:371–394, 1995.

Jianqing Fan and Irène Gijbels. *Local Polynomial Modelling and Its Applications*. CRC Press, 1996.

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

Wolfgang Härdle. *Applied Nonparametric Regression*. Cambridge University Press, 1990.

Tae Y. Kim, Byeong U. Park, Myung S. Moon, and Chiho Kim. Using bimodal kernel for inference in nonparametric regression with correlated errors. *Journal of Multivariate Analysis*, 100:1487–1497, 2009.

Soumendra Nath Lahiri. *Resampling Methods for Dependent Data*. Springer, 2003.

Hans G. Müller, Ulrich Stadtmüller, and Thomas Schmitt. Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, 74:743–749, 1987.

Jorge L. Ojeda Cabrera. locpol: Kernel local polynomial regression. URL `http://mirrors.ustc.edu.cn/CRAN/web/packages/locpol/index.html`, 2012.

Jean Opsomer, Yuedong Wang, and Yuhong Yang. Nonparametric regression with correlated errors. *Statistical Science*, 16:134–153, 2001.

Cheolwoo Park and Kee H. Kang. SiZer analysis for the comparison of regression curves. *Computational Statistics and Data Analysis*, 52(8):3954–3970, 2008.

James O. Ramsay. *Functional Data Analysis*. Wiley, 2006.

James O. Ramsay and Bernard W Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer, 2002.

David Ruppert, Simon J. Sheather, and Matthew P. Wand. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432): 1257–1270, 1995.

Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8:1348–1360, 1980.

Tiejun Tong and Yuedong Wang. Estimating residual variance in nonparametric regression using least squares. *Biometrika*, 92:821–830, 2005.

Tiejun Tong, Yanyuan Ma, and Yuedong Wang. Optimal variance estimation without estimating the mean function. *Bernoulli*, 19(5A):1839–1854, 2013.

Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

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

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