Smoothed Nonparametric Derivative Estimation using Random Forest Based Weighted Difference Quotients

Justin Dang*

October 27, 2021

Abstract

Derivatives play an important role in economics; they help determine partial marginal effects (e.g. returns to schooling) and check the curvature or concavity of functions (e.g. production function). These play key roles in economic policy evaluations and predictions. Therefore, it is essential to estimate derivatives precisely and robustly. In this paper, we propose a derivative estimator that is completely data driven to estimate the first and second derivatives. The estimator uses difference quotients, based on a variant of random forests, that are smoothed. Incorporating random forests in estimating derivatives help add more interpretability of forest-based models to explain the relationships between variables via marginal effects. Asymptotic properties of the estimator are established, and the performance of the estimator is addressed in both simulation and in an empirical application of evaluating the concavity of a power plant production function.

Keywords: derivative estimation, econometrics, random forests

*Department of Economics, University of California, Riverside. Email: jdang015@ucr.edu

1 Introduction

Derivative estimation plays a major role in economics. Derivatives help provide interpretability of the relationships between an independent variable and a dependent variable, for example, changing X by one unit, how much will be the change on Y, holding all else fixed? This is known as the partial marginal effect (first derivative) of X on Y. An economic example of the need of derivatives include estimating the marginal propensity to consume (MPC), the proportion of extra income that is spent on consumption. In this case, consumption, as a function of income, savings, and other determinants, is estimated. To determine the MPC, the partial derivative of this estimated equation is taken. Another economic example is estimating the effect of schooling or experience on earnings. In labor economics, the relationship between education, experience, and earnings is widely studied. It is often the case that people with more education and experience have higher earnings than those with less education and experience. To quantify how much more a person would earn with more education or experience, the derivative of an earnings regression is estimated to determine the returns to schooling or experience. Clearly, estimating derivatives are vital in understanding and solving economic problems.

After estimating a linear parametric model, to find the partial effect of a given variable, the derivative of the estimated regression equation is usually taken.[1] However, for certain data based nonparametric or machine learning models, there may not be an analytical form of the estimated regression equation and as a result, the partial effect may not be estimated. As an example, the derivative of the machine learning based random forest estimator of the regression function does not have an explicit analytical form nor it is smooth. Forest based regression models produce step-wise regression functions that are not differentiable, which is a huge hindrance in determining the partial marginal effect. Machine learning models

[1] In the classical case of Ordinary Least Squares, the derivative of the estimated linear regression function is a constant. However, a constant marginal effect may be too restrictive and the derivative may vary across the space of X. To deal with this, instead of specifying a parametric form of the regression function, resulting in a parametric form of the derivative, both the regression function and the derivative are estimated in a data driven way using nonparametric methods in this paper.

are very flexible and can estimate any function well; however, some of these models are considered to be "black boxes." To alleviate this issue, we estimate derivatives of a foreset based model, which allows for a better understanding of the underlying relationship between the data and can be more interpretable in terms of how a change in a variable will change the outcome variable.

In econometrics, the partial marginal effect can be estimated by taking the derivative of the estimated regression function. However, such a marginal effect estimation obtained from the linear or non-linear parametric regression model is well known to be biased and inconsistent unless it is from a correctly specified model, which is rare. When the model is from a data based nonparametric kernel regression, the estimates of derivatives are obtained by the local polynomial regression method (Fan and Gijbels, 1996), which may not be very smooth. Also, see the estimation of derivatives from spline regression (Zhou and Wolfe (2000) and Ma et al. (2019)). Recently, in the data based machine learning regression area, some interest in estimating derivatives has emerged. This includes, for example, a paper by Fonseca et al. (2018) where a machine learning model using logistic function, called boosted smooth transition regression trees (BooST) is introduced to estimate derivatives. However, if a logistic specification is not correct then the proposed estimator of derivatives may become statistically inconsistent.

Some other methods have also developed, in which data based difference quotients (DQ) have been utilized to estimate derivatives, where the derivative is determined by taking the ratio of differences between two data points. For example, Wang and Lin (2015) use symmetric DQ to run a locally weighted linear regression where the estimate of the derivative is just the intercept term under an equispaced design, where data points are equally spaced along the support of the independent variable. Also, see Iserles (2008) and Brabanter et al. (2013) papers for the estimation of derivatives using DQ method under the equispaced design. Liu and De Brabanter (2018) and Liu and Brabanter (2020) then apply this DQ method under the random design, where the regressor X is no longer restricted to be equally spaced

apart. Such a procedure is denoted as DQSmooth. The results based on DQSmooth produce noisy estimates of the derivative, and local polynomial regression is then used to smooth the derivatives in attempt to reduce the variance. However, the derivative estimation procedure proposed by Liu and Brabanter (2020) has some shortcomings. First, their estimator uses observed data on the dependent variable Y, whereas we propose to consider estimated values of Y, in order to further reduce the variance of the derivative estimator, based on a variant of the machine learning estimator, like random forest. Second, their method only considers the scalar X variable whereas we extend the estimator and its derivatives to include multivariate X, which is commonly used in econometrics.

Random forests procedure (Breiman, 2001) is a popular method for estimating nonparametric regression estimation for predictions. However, a drawback of random forests is that they are unable to capture any smoothness in the estimated regression surface. This is most likely why derivative estimation based on random forests is limited, which is due to the nonsmooth-like nature of the estimator. In an effort to address the issue of the inability to fit smooth signals, Friedberg et al. (2018) use random forests as an adaptive kernel method, where random forests are incorporated in a local polynomial regression framework with a ridge penalty, denoted as Local Linear Forests. Since Friedberg et al. (2018) consider a local linear regression, the derivative can be estimated by the coefficient of the first order derivative of the local linear regression. However, Friedberg et al. (2018) do not focus on estimating the derivative in their paper and instead only estimate the regression function. Instead of using local linear regression, any degree polynomial may be desired, especially if higher order derivatives need to be estimated. This procedure will be denoted as Local Random Forests (LRF).[2] However, we show in simulation that the derivative obtained by LRF are not only very noisy but they also tend to zero, and thus, LRF is a poor estimator of the derivative.

Under above scenarios, in this paper, we propose to estimate derivatives using a proce-

[2]Local Random Forest is simply an extension to Local Linear Forests (Friedberg et al., 2018), where instead of local linear regression, local polynomial regression of any degree can be used.

dure, denoted as DQSmoothLRF, where difference quotients are first obtained using predictions from LRF. Then, they are smoothed through a local polynomial kernel regression. The proposed derivative estimator contributes to both the nonparametric derivative estimation literature and random forest literature by providing more interpretability of forest based models to explain the relationships between variables via marginal effects estimation. It is shown that the proposed derivative estimator improves substantially relative to LRF considered by Friedberg et al. (2018).

The rest of the paper is as follows: section 2 goes through the procedure for estimating the first derivative and its properties, section 3 discusses second derivative estimation and its properties, section 4 extends the estimator to the multivariate case, section 5 briefly discusses the LRF estimator, section 6 displays the simulation results in comparison to the benchmark estimators, section 7 walks through an emprical example of evaluating the concavity of a Chilean power plant production function, and section 8 concludes the paper.

2 First Order Derivative Estimation

Consider the bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$, which are independently and identically distributed sampled from a population (X, Y), where $X \in \mathbb{R}$ and $Y \in \mathbb{R}$. Under the random design, X is a random variable generated from some unkown density f and distribution F. Consider the model

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

where $m(X) = \mathbb{E}[Y|X = x]$ is the conditional mean function. Assume that $\mathbb{E}[e|X = x] = 0$ and $\text{Var}[e|X = x] = \sigma_e^2 < \infty$. Now, consider a special case of X when X is standard uniformly distributed. That is, let $X = U \sim \mathcal{U}(0, 1)$, where $\mathcal{U}(0, 1)$ is the uniform distribution between 0 and 1. Now, let $U \sim \mathcal{U}(0, 1)$ and consider the same model but in the special case where $X \sim \mathcal{U}(0, 1)$

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

where $r(u) = \mathbb{E}[Y|U = u]$, $\mathbb{E}[e|U = u] = 0$, and $\text{Var}(e|U = u) = \sigma_e^2 < \infty$. Assume that the bivariate data (U, Y) is ordered with respect to U.

2.1 Weighted Difference Quotients

Using the weighted combination of symmetric difference quotients around the ith point from Iserles (2008), the noisy derivative estimator proposed by Liu and Brabanter (2020) under a random design is

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k} \omega_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{U_{i+j} - U_{i-j}} \right), \tag{3}$$

for $k+1 \leq i \leq n-k$ and hence $k \leq (n-1)/2$, where $\omega_{i,j}$ sum to one for $j = 1, \ldots, k$ and $\widehat{Y}_i^{(1)}$ is an estimator of the first derivative. The term noisy is used to differentiate this derivative estimator and the smoothed derivative estimator later in this section. By minimizing the variance of Eq. (3), Liu and Brabanter (2020) show that the optimal weights are

$$\omega_{i,j} = \frac{(U_{i+j} - U_{i-j})^2}{\sum_{l=1}^{k} (U_{i+l} - U_{i-l})^2}, \quad j = 1, \ldots, k, \tag{4}$$

where the weights sum to one across $j = 1, \ldots, k$. This estimator is only for the interior points, so Eq. (3) as well as the weights need to be modified for the boundaries, with observation $1 < i < k + 1$ being in the left boundary and $n - k < i < n$ being in the right boundary. Note that this estimator will not produce estimates for $i = 1, n$, and these two observations can be ignored.

For this paper, instead of using observed values of Y for Y_{i+j} and Y_{i-j}, we propose using estimated values from a Local Random Forest (LRF) model, a machine learning (ML) model. First, we estimate Eq. (2) by LRF, and produce its fitted values as

$$\widehat{Y}_{i,LRF} = \widehat{r}(U_i), \quad i = 1, \ldots, n. \tag{5}$$

In the papers previously mentioned about derivative estimation, all estimate the derivative

via the difference quotient using observed values of Y. By estimating the relationship between U and Y first, this allows us to try to pick up the signal from the data before taking the derivative, extend the procedure to the multivariate case, and provide more interpretability of forest based models. Therefore, the proposed noisy derivative estimator is

$$\widehat{Y}^{(1)}_{i,LRF} = \sum_{j=1}^{k} w_{i,j} \frac{\widehat{r}(U_{i+j}) - \widehat{r}(U_{i-j})}{U_{i+j} - U_{i-j}}, \tag{6}$$

where $\widehat{Y}^{(1)}_{i,LRF}$ denotes the derivative estimator based on LRF predictions, $\widehat{r}(\cdot)$. By minimizing the variance of Eq. (6), the optimal weights are obtained in Proposition 1. Let $\mathbb{U} = (U_{i-j}, \ldots, U_{i+j})$ for $i > j,\ i + j \leq n$, and $j = 1, \ldots, k$.

Proposition 1. *Under the model in Eq. (2) and for interior data points, $k + 1 \leq i \leq n - k$, minimizing the variance of Eq. (6) subject to $\sum_{j=1}^{k} w_{i,j} = 1$ gives the optimal weights (minimum conditional variance) as*

$$w_{i,j} = \frac{(U_{i+j} - U_{i-j})^2/(\sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j})}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2/(\sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j})}, \quad j = 1, \ldots, k, \tag{7}$$

where $\sigma^2_{\widehat{r},i\pm j} \equiv \mathrm{Var}[\widehat{r}(U_{i\pm j})|\mathbb{U}]$, the variance of LRF estimator, $\widehat{r}(U_{i\pm j})$, for observation $i \pm j$.
Proof: see Appendix A.

We can see that the optimal weights in Eq. (7) are similar to that of Eq. (4), however these weights depend on the variance of the LRF estimator, $\sigma^2_{\widehat{r},i\pm j}$. If we make the assumption that the variance of the estimator is the same for $i > j,\ i + j \leq n$, and $j = 1, \ldots, k$, then the optimal weights are the same as Eq. (4) in Liu and Brabanter (2020).

2.2 Asymptotic Properties of the Noisy Derivative Estimator

First, notice that the difference $U_i - U_j$ is simply the difference of uniform order statistics, where

$$U_i - U_j \sim \mathrm{Beta}(i - j, n - i + j + 1) \quad \text{for } i > j \tag{8}$$

(David and Nagaraja, 1970).

Lemma 1. *Define* $U \overset{iid}{\sim} \mathcal{U}(0,1)$ *and sort the random variables in order of magnitude such that* $U_1 < \cdots < U_n$*. Then,*

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

Proof: see Liu and Brabanter (2020).

This result, combined with Proposition 1 leads to the bias and variance of the first order derivative estimator.

Theorem 1. *Under the model in Eq. (2) and assume* r *is twice continuously differentiable on* $[0,1]$*,* $k \to \infty$ *as* $n \to \infty$*, and under the assumptions of Theorem 1 in Friedberg et al. (2018), with* $\omega \leq 0.2$ *and subsamples of size* s *with* $s = n^\beta$*, for*

$$\beta_{min} := 1 - \left(1 + \frac{d}{1.56\pi} \frac{\log(\omega^{-1})}{\log((1-\omega)^{-1})}\right) < \beta < 1,$$

$\mathrm{Var}[\hat{r}(\cdot)] = O(n^{-(1-\beta)})$*, where* ω *is the minimum fraction of parent observations into each child node,* π *is the minimum probability that a variable is split, and* d *is the number of regressors. Then, from the optimal weights obtained in Proposition 1, under the uniform random design on the interval* $[0,1]$*, the conditional absolute bias and conditional variance of the proposed derivative estimator in Eq. (6) for the interior data points* $k+1 \leq i \leq n-k$

are

$$\left|\text{bias}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}]\right| \leq \sup_{u\in[0,1]}\left|r^{(2)}(u)\right| \frac{3k(k+1)}{4(n+1)(2k+1)} + o_p\left(n^{-1}k\right) \tag{9}$$

$$\text{Var}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] \leq \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)} + o_p(n^{(1+\beta)}k^{-3}). \tag{10}$$

Proof: see Appendix B.

Then, from Theorem 1, the pointwise consistency of the derivative estimator can be obtained.

Corollary 1. *Under the assumptions of Theorem 1,* $k \to \infty$ *as* $n \to \infty$ *such that* $n^{(1+\beta)}k^{-3} \to 0$ *and* $n^{-1}k \to 0$*. Then, using the weights in Proposition 1, for any* $\varepsilon > 0$ *and for* $k+1 \leq i \leq n-k$*,*

$$P\left(\left|\widehat{Y}^{(1)}_{i,LRF} - r^{(1)}(U_i)\right| \geq \varepsilon\right) \to 0 \tag{11}$$

Proof: see Appendix C.

The parameter k, the number of symmetric differences around the ith data point, depicts the bias-variance tradeoff; larger k increases bias but decreases variance. Therefore, k is chosen by minimizing the asymptotic upper bound of the coniditonal mean integrated squared error (MISE).

Corollary 2. *Under the assumptions of Theorem 1, the optimal* k *that is chosen by minimizing the asymptotic upper bound of the conditional MISE is*

$$k_{opt} = \underset{k\in\mathbb{N}^+\setminus\{0\}}{\arg\min}\left\{\left(\mathcal{B}\frac{3k(k+1)}{4(n+1)(2k+1)}\right)^2 + \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)}\right\}, \tag{12}$$

where $\mathcal{B} \equiv \sup_{u\in[0,1]}\left|r^{(2)}(u)\right|$.

Proof: See Appendix D.

To find the optimal number of symmetric difference quotients, $\mathcal{B}$ needs to be estimated,

which can be done by a local cubic polynomial regression. A grid search for k or any optimization solver can then be used to find the optimal value of k.

Remark 1. *Taking the first order condition of Eq. (12), we will not get an analytical solution for k_{opt}. However, if we retain the higher order terms, we can get an approximation for k_{opt}*

$$\hat{k}_{opt} = \lfloor 2^{4/5} \widehat{\mathcal{B}}^{-2/5} n^{(3+\beta)/5} \rfloor \tag{13}$$

given an estimate for $\mathcal{B}$.

2.3 Boundary Correction

So far, we have only discussed points in the interior. In order to reduce the variance, slight modifications to the estimator is needed at the boundaries. Similar to the boundary corrections made in Charnigo et al. (2011), Brabanter et al. (2013), and Liu and Brabanter (2020), the observations lying at the left boundary with index $1 < i < k+1$, the modified weighted difference estimator is

$$\widehat{Y}^{(1)}_{i,LRF} = \sum_{j=1}^{k(i)} w_{i,j} \left(\frac{\widehat{r}(U_{i+j}) - \widehat{r}(U_{i-j})}{U_{i+j} - U_{i-j}} \right) + \sum_{j=k(i)+1}^{k} w_{i,j} \left(\frac{\widehat{r}(U_{i+j}) - \widehat{r}(U_i)}{U_{i+j} - U_i} \right) \tag{14}$$

with weights

$$w_{i,j} = \begin{cases} \frac{(U_{i+j}-U_{i-j})^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})}{\sum_{l=1}^{k(i)}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})+\sum_{l=k(i)+1}^{k}(U_{i+l}-U_i)^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i})}, & 1 \le j \le k(i) \\ \frac{(U_{i+j}-U_i)^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i})}{\sum_{l=1}^{k(i)}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})+\sum_{l=k(i)+1}^{k}(U_{i+l}-U_i)^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i})}, & k(i) < j \le k \end{cases} \tag{15}$$

for $k(i) = i - 1$. Here, the weights are similar to those from the interior in that the weights are standardized by the inverse variances of the LRF estimator. The observations at the right boundary can be estimated in a similar fashion for $k(i) = n - i$.

2.4 Smoothing Weighted Difference Quotients

The first order derivative estimators Eq. (6) and Eq. (14) depend on the variance of the LRF estimator; forest based estimators are very noisy and therefore may affect the variance of the derivative estimator. Another issue of the weighted difference quotient estimator is that it cannot be evaluated at any arbitrary test point, and can only be evaluated for points in the training sample. Therefore, Liu and De Brabanter (2018) and Liu and Brabanter (2020) propose smoothing the estimator via local polynomial regression.

First, observe that the first order derivative estimator in Eq. (6), along with their modified boundary correction estimators, will create a new variable for observations $i = 2, \ldots, n-1$. Then, consider the new model,

$$\widehat{Y}_{LRF}^{(1)} = r^{(1)}(U) + \tilde{e}, \tag{16}$$

where the first derivative estimator, $\widehat{Y}_{LRF}^{(1)}$, is some unknown first derivative function, $r^{(1)}(\cdot)$, of U, with error, $\tilde{e}$. By construction of the first derivative estimator, the errors will be correlated and assume that $\mathbb{E}[\tilde{e}|U] = 0$, $\text{Cov}[\tilde{e}_i, \tilde{e}_j|U_i, U_j] = \sigma_{\tilde{e}}^2 \rho_n(U_i - U_j)$ for $i \neq j$, and $\sigma_{\tilde{e}}^2 < \infty$. The correlation function, $\rho_n(\cdot)$ goes to zero as $n \to \infty$ and must satisfy $\rho_n(0) = 1$, $\rho_n(u) = \rho_n(-u)$, and $-1 \leq \rho_n(u) \leq 1$, assumed in Liu and Brabanter (2020) and Brabanter et al. (2018). The goal is then to enhance the noisy derivative estimator $\widehat{Y}_{LRF}^{(1)}$ by nonparametric smoothing. However, the correlation in the errors will affect the bandwidth selection for any nonparametric smoothing and to counteract these effects, Brabanter et al. (2018) propose using a bimodal kernel K such that $K(0) = 0$ and showed that under mild assumptions, using such a kernel will remove any effects of the correlation on the bandwidth selection without the need to estimate the correlation structure. We will denote the bimodal kernel as $\bar{K}$:

$$\bar{K}(u) = (2/\sqrt{\pi})u^2 \exp(-u^2). \tag{17}$$

Next, we fit a local polynomial regression of $\widehat{Y}_{LRF}^{(1)}$ on U. The local polynomial regression

estimator of degree p for a given test observation u_0 is provided by minimizing

$$\min_{\beta_L \in \mathbb{R}} \sum_{i=1}^{n} \left\{ \widehat{Y}^{(1)}_{i,LRF} - \sum_{L=0}^{p} \beta_L \left(U_i - u_0\right)^L \right\}^2 K\left(\frac{U_i - u_0}{h}\right) \tag{18}$$

where β_L are the solutions to the weighted least squares problem and $K(\cdot)$ is a kernel function. The Lth order derivative $r^{(L)}(u_0)$ for $L = 0, 1, \ldots, p$ is estimated by $\widehat{r}^{(L)}(u_0) = L!\widehat{\beta}_L$. In matrix notation, the solution is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{U}^\top \mathbf{W} \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{W} \widehat{\mathbf{y}}^{(1)}, \tag{19}$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^\top$ is the $(p+1) \times 1$ vector of solutions to the minimization problem, $\mathbf{W} = \text{diag}(K((U_i - u_0)/h))$ is the $(n-2) \times (n-2)$ diagonal matrix of weights based on a specified kernel function, $\widehat{\mathbf{y}}^{(1)} = (\widehat{Y}^{(1)}_{2,LRF}, \ldots, \widehat{Y}^{(1)}_{n-1,LRF})$ is the $(n-2) \times 1$ vector of the first derivative estimators, and

$$\mathbf{U} = \begin{pmatrix} 1 & (U_2 - u_0) & \cdots & (U_2 - u_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (U_{n-1} - u_0) & \cdots & (U_{n-1} - u_0)^p \end{pmatrix},$$

the $(n-2) \times (p+1)$ centered regression matrix. Therefore, the smoothed first order derivative estimator is

$$\widehat{r}^{(1)}(u_0) = \boldsymbol{\epsilon}_1^\top \widehat{\boldsymbol{\beta}} = \boldsymbol{\epsilon}_1^\top (\mathbf{U}^\top \mathbf{W} \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{W} \widehat{\mathbf{y}}^{(1)} \tag{20}$$

where $\boldsymbol{\epsilon}_i$ is a column vector that picks out the ith element.

2.5 Asymptotic Properties of the Smoothed Derivative Estimator

Next, we discuss some asymptotic results of the final smoothed derivative estimator in Eq. (20). The following theorem states the upper bound of the conditional bias and variance of $\widehat{r}^{(1)}(\cdot)$.

Theorem 2. *Under the assumptions in Theorem 2 of Liu and Brabanter (2020) and in Theorem 1, $k \to \infty$ as $n \to \infty$, and the weights given in Proposition 1, the conditional bias and variance of Eq. (20) for p odd is*

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

and

$$\begin{aligned}\text{Var}[\widehat{r}^1(u_0)|\tilde{\mathbb{U}}] &\leq \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)} \frac{1+\rho_c}{h(n-2k)} \boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1} \mathbf{S}^* \mathbf{S}^{-1} \boldsymbol{\epsilon}_1 \{1 + o_p(1)\}, \\ &= \int K_0^{*2}(t) dt \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)} \frac{1+\rho_c}{h(n-2k)} \{1 + o_p(1)\}\end{aligned}$$

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

Proof: see Appendix E.

With smoothing using local polynomial regression, the bandwidth h needs to be estimated. Following Liu and Brabanter (2020), we find k and h as follows: k is found by optimizing AMISE in Corollary 2 and the bandwidth h is then estimated by using the bimodal kernel $\bar{K}$ in Eq. (17), denoted as $\widehat{h}_b$ by cross validation. Then, $\widehat{h}_b$ can be used as a pilot bandwidth which can be related to the bandwidth, $\widehat{h}$, of the usual unimodal kernel, such as the gaussian kernel,

$$K(u) = 1/\sqrt{2\pi} \exp(-u^2/2). \tag{21}$$

Brabanter et al. (2013) show the relationship between the bimodal and unimodal bandwidth,

$$\widehat{h} = 1.01431\widehat{h}_b, \tag{22}$$

for local cubic regression using a gaussian kernel.[3] Therefore, after fitting a local cubic regression of $\widehat{Y}^{(1)}_{LRF}$ on U with bimodal kernel, $\bar{K}(\cdot)$, and bandwidth, $\widehat{h}_b$, we refit a local cubic regression with unimodal kernel, $K(\cdot)$ and bandwidth, $\hat{h}$, defined in Eq. (22).

From Theorem 2, for p odd, the pointwise consistency follows.

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

$$P(|\widehat{r}^{(1)}(u_0) - r^{(1)}(u_0)| \geq \epsilon) \to 0. \tag{23}$$

Proof: See Appendix F.

2.6 Generalizing to Arbitrary Distributions

In general, X may not follow a standard uniform distribution. To generalize X with unknown distribution F, Liu and Brabanter (2020) suggest using probability integral transform (PIT)

$$F(X) \sim \mathcal{U}(0, 1). \tag{24}$$

By using the PIT, we know that the transformed data is $(F(X_1), Y_1), \ldots, (F(X_n), Y_n)$ which has the same distribution as $(U_1, Y_1), \ldots, (U_n, Y_n)$. Then, the same procedure can be used under the transformed data set. Notice, however, that the derivatives, $\widehat{r}^{(1)}(U_i)$ are in the

[3]For pth degree local regression and for different kernel functions, please see Brabanter et al. (2018).

transformed space. Now, to get the derivative in the original space, the chain rule is used

$$\frac{dm(X)}{dX} = \frac{dr(U)}{dU}\frac{dU}{dX} = f(X)\frac{dr(U)}{dU} \tag{25}$$

where $m(X) = r(F(X))$. Since the distribution and density, F and f, are unknown, kernel estimators can be used to estimate the distribution and density, with plug-in bandwidths. As a result, the full smoothed first order derivative estimator in the original space is

$$\widehat{m}^{(1)}(X) = \widehat{f}(X)\widehat{r}^{(1)}(U) \tag{26}$$

To summarize, the full algorithm is described in Algorithm 1.

Note that this procedure gives the pointwise derivatives of Y with respect to X, which may vary across the space of X. To summarize the overall derivative, we can take the sample average of the derivative estimates to estimate the average derivative, that is

$$\widehat{m}^{(1)}_{avg} = \frac{1}{n'}\sum_{i=1}^{n'}\widehat{m}^{(1)}(X_i) \tag{27}$$

This result may be useful in estimating the global partial effect of a variable or even in comparison to the partial effect given by an OLS coefficient.

Algorithm 1 Smoothed Nonparametric Derivative Estimation using Weighted Difference Quotients based on LRF

procedure DQSmoothLRF(training independent variable X, training dependent variable Y, number of symmetric difference quotients k, degree of local polynomial $d = 3$, grid of bandwidths to be considered h_{grid}, tunable hyperparameters of LRF model θ)

1: $U \leftarrow \widehat{F}(X)$
   ▷ kernel cumulative distribution function estimation
2: $\widehat{r}(U) \leftarrow$ regress Y on U by a LRF model with tunable hyperparameters θ
3: $\widehat{Y}_{LRF}^{(1)} \leftarrow$ difference quotient in Eq. (6)
4: $\widehat{h}_b \leftarrow$ bandwidth selection from local polynomial regression of $\widehat{Y}_{LRF}^{(1)}$ on U
   ▷ with degree deg, kernel $\bar{K}$ in Eq. (17), and searched over bandwidths h_{grid}
5: $\widehat{h} \leftarrow 1.01431\widehat{h}_b$ (for local cubic regression)
   ▷ bandwidth relationship between bimodal and unimodal gaussian kernels
6: $\widehat{r}^{(1)}(U) \leftarrow$ local polynomial regression
   ▷ with degree deg, unimodal gaussian kernel K, and bandwidth $\widehat{h}$
7: $\widehat{m}^{(1)}(X) \leftarrow \widehat{f}(X)\widehat{r}^{(1)}(U)$
   ▷ back transform into original space as in Eq. (26) after kernel density estimation of X

3 Second Order Derivatives

At times, economists are interested in the second derivative, which help depict the curvature of the function and how the slope changes due to a change in the independent variable. For example, with second derivatives, we can determine whether the earnings function is concave or convex or see if a production function exhibits diminishing marginal returns. Again, assume U is standard uniformly distributed and that the data (U, Y) is sorted with respect to U. The second order weighted difference quotient proposed by Liu and Brabanter (2020) is

$$\widehat{Y}_i^{(2)} = 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{Y_{i+j+k_1}-Y_{i+j}}{U_{i+j+k_1}-U_{i+j}} - \frac{Y_{i-j-k_1}-Y_{i-j}}{U_{i-j-k_1}-U_{i-j}}\right)}{U_{i+j+k_1} - U_{i+j} - U_{i-j-k_1} - U_{i-j}}, \tag{28}$$

with weights

$$w_{i,j,2} = \frac{(2j+k_1)^2}{\sum_{j=1}^{k_2}(2j+k_1)^2}, \tag{29}$$

where k_1 and k_2 are positive integers that represent the number of first and second order difference quotients about observation i such that the weights $w_{i,j,2}$ sum to one across j. Note that the estimator is for observations in the interior for $k_1 + k_2 + 1 \leq i \leq n - k_1 - k_2$. Also note that the Liu and Brabanter (2020) choose the second order derivative weights to be proportional to the inverse of the conditional variance of each quotient. Similar to the first order derivative, we replace Y with estimates of Y using the LRF estimator, denoted by $\widehat{r}$. Define ${}^{+}\widehat{Y}^{(1)}_{i+j,LRF} = \frac{\widehat{r}(U_{i+j+k_1}) - \widehat{r}(U_{i+j})}{U_{i+j+k_1} - U_{i+j}}$ and ${}^{-}\widehat{Y}^{(1)}_{i-j,LRF} = \frac{\widehat{r}(U_{i-j-k_1}) - \widehat{r}(U_{i-j})}{U_{i-j-k_1} - U_{i-j}}$. This leads to the following proposition.

Proposition 2. *Under the assumptions of Theorem 1 and $r(\cdot)$ is three times differentiable, the estimator for the second derivative based on symmetric difference quotients is*

$$\widehat{Y}^{(2)}_{i,LRF} = \sum_{j=1}^{k_2} w_{i,j,2} \frac{{}^{+}\widehat{Y}^{(1)}_{i+j,LRF} - {}^{-}\widehat{Y}^{(1)}_{i-j,LRF}}{C_{i,j,k_1}}, \tag{30}$$

where $C_{i,j,k_1} = (U_{i+j+k_1} - U_{i+j} - U_{i-j-k_1} - U_{i-j})/2$ is chosen such that each individual quotient $\frac{{}^{+}\widehat{Y}^{(1)}_{i+j,ML} - {}^{-}\widehat{Y}^{(1)}_{i-j,ML}}{C_{i,j,k_1}}$ is an asymptotically unbiased estimator of the second order derivative $r^{(2)}(\cdot)$ for $j = 1, \ldots, k_2$ and where each weight is selected to be proportional to the inverse of the conditional variance of each quotient:

$$w_{i,j} = \frac{\frac{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})^2}{{}^{+}V_{i+j} + {}^{-}V_{i-j}}}{\sum_{j=1}^{k_2} \frac{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})^2}{{}^{+}V_{i+j} + {}^{-}V_{i-j}}}, \tag{31}$$

with ${}^{+}V_{i+j} \equiv \frac{\sigma^2_{\widehat{r},i+j+k_1} + \sigma^2_{\widehat{r},i+j}}{(U_{i+j+k_1} - U_{i+j})^2}$ and ${}^{-}V_{i-j} \equiv \frac{\sigma^2_{\widehat{r},i-j-k_1} + \sigma^2_{\widehat{r},i-j}}{(U_{i-j-k_1} - U_{i-j})^2}$.

Proof: See Appendix appendix G.

Applying Lemma 1, and using the leading order of the weight, the weight can be approximated by

$$w_{i,j,2} = \frac{(2j + k_1)^2 / (\sigma^2_{\widehat{r},i+j+k_1} + \sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j-k_1} + \sigma^2_{\widehat{r},i-j})}{\sum_{j=1}^{k_2} (2j + k_1)^2 / (\sigma^2_{\widehat{r},i+j+k_1} + \sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j-k_1} + \sigma^2_{\widehat{r},i-j})}. \tag{32}$$

Notice that if the variances are approximately the same for the $2 \cdot (k_1 + k_2)$ observations around i, the the proposed weights will be equal to that of Liu and Brabanter (2020) in Eq. (29).

Theorem 3. *Assume the model in Eq. (2), r is three times continuously differentiable on $[0,1]$, $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$, and the assumptions of Theorem 1 in Friedberg et al. (2018), with $\mathrm{Var}[\hat{r}(\cdot)] = O(n^{-(1-\beta)})$. Then, from the optimal weights obtained in equation (32), under the uniform random design on the interval $[0,1]$, the conditional absolute bias and conditional variance of the proposed second derivative estimator in Eq. (30) for the interior data points $k_1 + k_2 + 1 \leq i \leq n - k_1 - k_2$ are*

$$\left| \mathrm{Bias}[\hat{Y}^{(2)}_{i,LRF} | \tilde{\mathbb{U}}] \right| \leq \frac{\sup_{u \in [0,1]} |r^{(3)}(u)|}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \{1 + o_p(1)\} \tag{33}$$

$$\mathrm{Var}[\widehat{Y}^{(2)}_{i,LRF} | \mathbb{U}] \leq \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2 \sum_{j=1}^{k_2} (2j + k_1)^2} \{1 + o_p(1)\}. \tag{34}$$

Proof: see Appendix H.

From Theorem 3, the pointwise consistency of $\hat{Y}^{(2)}_{i,LRF}$ can be obtained.

Corollary 4. *Under the assumptions of Theorem 3, $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ such that $n^{-1}k_1 \to 0$, $n^{-1}k_2 \to 0$, $n^{3+\beta}k_1^{-2}k_2^{-3} \to 0$, and $n^{3+\beta}k_1^{-4}k_2^{-1} \to 0$. Then, using the weights in equation (32), for any $\varepsilon > 0$ and for $k_1 + k_2 + 1 \leq i \leq n - k_1 - k_2$,*

$$P\left(\left| \widehat{Y}^{(2)}_{i,LRF} - r^{(2)}(U_i) \right| \geq \varepsilon \right) \to 0 \tag{35}$$

Proof: see Appendix I.

From Theorem 3, the number of difference quotients k_1 and k_2 play a role in the bias variance tradeoff. Similar to the symmetric difference quotients for the first derivative, the higher k_1 and k_2 are, the higher the bias but the lower the variance and vice versa. The

following corollary chooses the numbers of symmetric difference quotients for the second derivative considering the bias variance tradeoff.

Corollary 5. *Under the assumptions of Theorem 3, the optimal k_1 and k_2 that is chosen by minimizing the asymptotic upper bound of the conditional MISE is*

$$(k_1,k_2)_{opt} = \underset{k\in\mathbb{N}^+\setminus\{0\}}{\arg\min} \left\{ \left(\frac{\mathcal{B}_2}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j} \right)^2 + \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2} \right\} \tag{36}$$

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

Proof: See Appendix J.

The quantity $\mathcal{B}_2$ can be estimated by a local polynomial regression of order $p = 4$ to obtain an estimate of the third derivative of r. The optimal value pair $(k_1, k_2)_{opt}$ can be obtained using a grid search or any optimization method. Points on the left boundary, $i < k_1 + k_2 + 1$, and points on the right boundary, $i > n - k_1 - k_2$, need to be adjusted and can be done in a similar analysis for the noisy first derivative estimator.

In order to smooth the second order weighted difference quotients, a local polynomial regression of the second order derivative estimates on U. First, rewriting Eq. (28) as

$$\begin{aligned} \widehat{Y}_i^{(2)} = 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{r(U_{i+j+k_1}) - r(U_{i+j})}{U_{i+j+k_1} - U_{i+j}} - \frac{r(U_{i-j-k_1}) - r(U_{i-j})}{U_{i-j-k_1} - U_{i-j}} \right)}{U_{i+j+k_1} - U_{i+j} - U_{i-j-k_1} - U_{i-j}} \\ + 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{e_{i+j+k_1} - e_{i+j}}{U_{i+j+k_1} - U_{i+j}} - \frac{e_{i-j-k_1} - e_{i-j}}{U_{i-j-k_1} - U_{i-j}} \right)}{U_{i+j+k_1} - U_{i+j} - U_{i-j-k_1} - U_{i-j}}, \end{aligned} \tag{37}$$

where the first term is denoted as the true second derivative, $r^{(2)}(U)$ and the last term is the new error term, $\acute{e}$,

$$\widehat{Y}_i^{(2)} = r^{(2)}(U) + \acute{e}. \tag{38}$$

As in the case for the first derivative, a bimodal kernel, K such that $K(0) = 0$, is used to

counteract the effect of the correlated errors on bandwidth selection. Instead of using the second order weighted difference quotient in Eq. (28), we propose using Eq. (30) with weights Eq. (32) for the estimate of the second derivative,

$$\widehat{Y}^{(2)}_{i,LRF} = r^{(2)}(U) + \acute{e}. \tag{39}$$

Now, a local polynomial regression is estimated for the model Eq. (39), where $\widehat{Y}^{(1)}_{i,LRF}$ is replaced by $\widehat{Y}^{(2)}_{i,LRF}$ in Eq. (18) is minimized. Following Liu and Brabanter (2020), the error term $\acute{e}$ satisfies $\mathbb{E}[\acute{e}|U] = 0$ and $\text{Cov}[\acute{e}_i, \acute{e}_j|U_i, U_j] = \sigma^2_{\acute{e}}\rho'_n(U_i - U_j)$. Then, the solution is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{U}^\top \mathbf{W}\mathbf{U})^{-1}\mathbf{U}^\top \mathbf{W}\widehat{\mathbf{y}}^{(\mathbf{2})}, \tag{40}$$

where $\mathbf{U}$ is the centered regression matrix with the first column being ones, $\mathbf{W}$ is the diagonal matrix of kernel weights, and $\widehat{\mathbf{y}}^{(2)}$ is the vector of second order weighted difference quotients in Eq. (30). Therefore, the smoothed second order derivative estimator is given by

$$\widehat{r}^{(2)}(u_0) = \boldsymbol{\epsilon}_1^\top \widehat{\boldsymbol{\beta}} = \boldsymbol{\epsilon}_1^\top (\mathbf{U}^\top \mathbf{W}\mathbf{U})^{-1}\mathbf{U}^\top \mathbf{W}\widehat{\mathbf{y}}^{(\mathbf{2})} \tag{41}$$

To generalize to any unknown distribution for X, we again use PIT to transform the variables X to U. Since, $m(X) = r(F(X))$, the second derivative of m with respect to X is

$$\frac{d^2m}{dX^2} = \left(\frac{dr}{dU}\frac{dU}{dX}\right) = \frac{d}{dX}\left(f(X)r^{(1)}(U)\right) = f^{(1)}(X)r^{(1)}(U) + f(X)r^{(2)}(U) \tag{42}$$

4 Extension to the Multivariate Case

It is very rare to have models with univariate X in economics. In this section we try to extend the procedure to the multivariate case. Suppose we have the model

$$Y_i = m(X_{i,1}, \ldots, X_{i,d}) + e_i, \quad i = 1, \ldots, n, \tag{43}$$

where d is the number of independent variables. For now, suppose all of the regressors are standard uniformly distributed. Then, consider the regression of the form

$$Y_i = r(U_{i,1}, \ldots, U_{i,d}) + e_i. \tag{44}$$

Then, we follow the same steps as before. First, we estimate $r(\cdot)$ by LRF. Now, the first partial derivative with respect to the sth variable is given by the weighted difference quotient

$$\widehat{Y}^{(1)}_{i,s,LRF} = \sum_{j=1}^{k_s} w_{i,j} \frac{\widehat{r}(U_{i+j,s}, \bar{U}) - \widehat{r}(U_{i-j,s}, \bar{U})}{U_{i+j,s} - U_{i-j,s}}, \tag{45}$$

where $\bar{U}$ contains the other $d-1$ variables evaluated at their medians. Note that we order the data with respect to the sth variable and that variables can have different number of symmetric difference quotients denoted by k_s. We can then use this estimate of the first partial derivative as the dependent variable in a local polynomial regression on $U_1, \ldots, U_d$,

$$\widehat{Y}^{(1)}_{s,LRF} = r^{(1)}_s(U_1, \ldots, U_d) + \tilde{e}, \tag{46}$$

where the subscript s denotes the derivative with respect to the sth variable. Since the errors are correlated, we use the multivariate bimodal kernel,

$$\bar{K}(\mathbf{u}) = (2d/\pi^{\frac{d}{2}})||\mathbf{u}||^2 \exp(-||\mathbf{u}||^2). \tag{47}$$

Then, we can correct the bandwidth using the relationship between bimodal and unimodal kernel functions as in Brabanter et al. (2013) for each bandwidth to get an estimate of the partial derivative, $\widehat{r}^{(1)}_s(U_s, \bar{U})$.

Now, suppose that X_m for $m = 1 \ldots, d$ are not all standard uniformly distributed. First, we can estimate the joint CDF function of all independent variables, $F_{X_1,\ldots,X_d}(x_1, \ldots, x_d)$. To obtain the marginal CDFs of a specified variable, we take the limits in the arguments of

the joint CDF of the other variables, $F_{X_s}(x_s) = F_{X_1,\ldots,X_d}(+\infty,\ldots,x_s,\ldots,+\infty)$. Then, we have for each regressor,

$$F_{X_m}(X_m) \sim \mathcal{U}(0,1), \quad m = 1,\ldots,d. \tag{48}$$

Therefore, the new data $(F_{X_1}(X_{1,1}),\ldots,F_{X_d}(X_{1,d}),Y_1),\ldots,(F_{X_1}(X_{n,1}),\ldots,F_{X_d}(X_{n,d}),Y_n)$, has the same distribution as $(U_{1,1},\ldots,U_{1,d},Y_1),\ldots,(U_{n,1},\ldots,U_{n,d},Y_n)$.

$$\begin{aligned} Y_i &= r(F_{X_1}(X_{i,1}),\ldots,F_{X_d}(X_{i,d})) + e_i \\ &= r(U_{i,1},\ldots,U_{i,d}) + e_i, \end{aligned} \tag{49}$$

which is the case where we have all standard uniform variables as regressors. To get derivatives in the original space,

$$\frac{\partial m(X_s,\bar{X})}{\partial X_s} = \frac{\partial r(U_s,\bar{U})}{\partial U_s}\frac{\partial U_s}{\partial X_s} = f_{X_s}(X_s)\frac{\partial r(U_s,\bar{U})}{\partial X_s} \tag{50}$$

Using a multivariate kernel density estimator for $f_{X_1,\ldots,X_d}$, and marginalizing for the $s-th$ variable, the final smooth partial derivative with respect to the sth variable is

$$\widehat{m}_s^{(1)}(X_s,\bar{X}) = \widehat{f}_{X_s}(X_s)\widehat{r}_s^{(1)}(U_s,\bar{U}), \tag{51}$$

where $\widehat{f}_{X_s}(X_s) = \sum_{x_1}\cdots\sum_{x_{s-1}},\sum_{x_{s+1}}\cdots\sum_{x_d}\widehat{f}_{X_1,\ldots,X_d}(x_1,\ldots,x_s,\ldots,x_d)$ Note that $\bar{X}$ and $\bar{U}$ mean we are holding the other variables fixed (at their medians or some other fixed constant). So, we evaluate the derivatives holding the other variables fixed at their medians in the X space and transform these fixed values in the U space when fitting a regression function and smoothing the derivatives. The second derivative in the multivariate case can be estimated in a similar fashion.

$$\widehat{m}_s^{(2)}(X_s,\bar{X}) = \widehat{f}_{X_s}^{(1)}(X_s)\widehat{r}^{(1)}(U_s,\bar{U}) + \widehat{f}_{X_s}(X_s)\widehat{r}^{(2)}(U_s,\bar{U}) \tag{52}$$

5 Local Random Forest

This section briefly discusses the Local (Linear) Random Forest (LRF) estimator. In what follows, assume the i.i.d data $(\mathbf{x}_i, y_i)$, for $i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and assume the model in Eq. (1), either for the univariate case ($d = 1$) or the multivariate case ($d > 1$). LRF uses a local polynomial regression with $p = 1$ (local linear) with forest based weights instead of kernel based weights (Friedberg et al., 2018). The objective function includes a weighted quadratic loss with L_2 regularization,

$$\arg\min_{\delta_q} \sum_{i=1}^{n} \left(y_i - \sum_{q=0}^{p} \delta_q^T (\mathbf{x}_i - \mathbf{x}_0)^q \right)^2 a_i(\mathbf{x}_i, \mathbf{x}_0) + \lambda ||\delta_1||^2 \tag{53}$$

where λ is the regularization strength parameter and $a_i(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbf{1}\{x_i \in L_b(\mathbf{x}_0)\}}{|L_b(\mathbf{x}_0)|}$ is the weight function. Here, B is the number of bootstrap replications, $\mathbf{1}\{\cdot\}$ is the indicator function, and $L_b(\mathbf{x}_0)$ denotes the the leaf of the b-th bootstrapped tree that contains the testing point $\mathbf{x}_0$. δ_q, for $q = 0, 1$ denotes the conditional mean function and its derivative evaluated at $\mathbf{x}_0$. The minimization problem is the same as in Eq. (18), except for the weight function and the regularization term. Using similar notation as before, the minimization problem can be expressed in matrix form as

$$\arg\min_{\boldsymbol{\delta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\delta})^\top \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\delta}) + \lambda \boldsymbol{\delta}^\top \mathbf{J} \boldsymbol{\delta}, \tag{54}$$

where $\mathbf{J}$ is the identity matrix with the first diagonal element as zero, $\mathbf{X}$ is the centered regression matrix with the first column being ones, $\mathbf{W}$ is the diagonal matrix of weights $a(\cdot)$, $\mathbf{y}$ is the vector of the dependent variable, and $\boldsymbol{\delta}$ is the gradient vector with the first element being the estimate for the mean regression function. Then, the solution is

$$\widehat{\boldsymbol{\delta}} = (\mathbf{X}^\top \mathbf{W} \mathbf{X} + \lambda \mathbf{J})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y} \tag{55}$$

and the resulting prediction function evaluated at a test point $\mathbf{x}_0$ is

$$\widehat{m}(\mathbf{x}_0) = \boldsymbol{\epsilon}_1^\top \widehat{\boldsymbol{\delta}}. \tag{56}$$

To find λ, a random forest model is trained first and the weights $a_i(\cdot)$ are obtained from the forest. Then, the regularization parameter λ can be found by cross validation. LRF can then be used in step 3 of Algorithm 1, with transformed regressor, U.

As a reference, the first derivative can be obtained from LRF.

$$\widehat{m}^{(1)}(\mathbf{x}_0) = \boldsymbol{\epsilon}_2^\top \widehat{\boldsymbol{\delta}} \tag{57}$$

However, as stated in section 1, the derivative based on LRF will be noisy and tend to zero due to the ridge penalty. Even when there is no ridge penaly, $\lambda = 0$, the derivative will still have high variance. This is due to the nature of the high variance in random forests. In the following section, we will show that there is a huge improvement when the derivative based on LRF is compared to the proposed estimator, DQSmoothLRF.

6 Simulations

This section shows derivative estimation based on the proposed estimator, DQSmoothLRF as well as other estimators. To compare results, we consider the following data generating process (DGP)

$$m(X) = \cos(2\pi X)^2 \quad \text{for } X \sim \text{beta}(2, 2), \tag{58}$$

which is from Liu and De Brabanter (2018) with sample size $n = 700$ and $e \sim N(0, 0.2^2)$. Since we know the true function, the analytical expression of the derivative is

$$m^{(1)}(X) = -4\pi \cos(2\pi X) \sin(2\pi X). \tag{59}$$

For all simulations, we estimate the density f and distribution F by the `R` package `ks` (Duong, 2020). The parameter k, the number of symmetric difference quotients, is estimated by Corollary 2. For the weights $w_{i,j}$ based on the variances of the LRF estimator in Eq. (7), we assume that the variances are constant for the k points round i, so that the weights collapse to Eq. (4). Therefore, plots and results show the derivative estimators based on the weights obtained in Eq. (4). We do this so that results under the proposed model can be easily compared to the benchmark model, the model proposed by Liu and Brabanter (2020), and since the simulated errors are homoskedastic, this assumption may be reasonable.

For this DGP, we also consider estimating the second derivative. Given the true function for m, the second derivative is

$$m^{(2)}(X) = 8\pi^2(\sin(2\pi X)^2 - \cos(2\pi X)^2). \tag{60}$$

Then, we follow the procedure outlined in section 3, where the numbers of symmetric difference quotients for the first and second derivative, k_1 and k_2, are estimated by Corollary 5. Similar to the first derivative, for the weights $w_{i,j,2}$ in Eq. (32), we assume that the variances are constant for the k_1 and k_2 points round i, so that the weights do not depend on the variances of the LRF estimator.

We show estimates of the first and second derivatives considering four different models, (1) DQSmooth, the benchmark model of the derivative based on difference quotients proposed by Liu and Brabanter (2020), (2) DQSmoothLRF, the proposed estimator based on LRF estimates of Y, (3) LocCubic, a local cubic regression, a common regression technique to estimate derivatives in the nonparametric literature, and (4), LRF, the model proposed by Friedberg et al. (2018), a benchmark model of the derivative based on random forests. For the last estimator, LRF, the original paper by Friedberg et al. (2018) focuses on estimating the conditional mean function, not its derivatives. The paper also considers only a local linear approach. For these simulations, we consider a local cubic with $\lambda = 0$ for the LRF

estimator. The reason for zero ridge penalty, is that since the ridge parameter, λ, penalizes the curvature of the function, it will force the derivative estimates toward zero, although the conditional mean function may not be flat. All local polynomial regressions are estimated using `locpol` package (Cabrera, 2018). When estimating LRF based models, we use the `grf` package (Tibshirani et al., 2020). To assess the models, we use mean squared error (MSE) and mean absolute error (MAE), defined as

$$\text{MSE} = \frac{1}{m}\sum_{j=1}^{m}\left(\widehat{m}^{(1)}(X_j) - m^{(1)}(X_j)\right)^2 \tag{61}$$

$$\text{MAE} = \frac{1}{m}\sum_{j=1}^{m}\left|\widehat{m}^{(1)}(X_j) - m^{(1)}(X_j)\right|. \tag{62}$$

We evaluate all models at 500 evenly spaced points from 0.05 to 0.95, where $m = 500$. Models for the second derivative is evaluated in a similar fashion.

Results for the first derivative are plotted in Figure 1. The grey curves depict one estimated first derivative function for a simulation run, where all simulations are plotted. The solid colored curves represent the average of all simulations at each of the 500 evenly spaced points of X from 0.05 to 0.95 and the black curve represents the true derivative we wish to estimate. As seen from the figure, all models on average seem to estimate the true first derivative accurately. However, the difference is notable due to the variance of the models, which is depicted by how far away on average the grey curves are from their solid curves. By first glance, both DQsmooth and DQSmoothLRF seem to have lower variance and are more accurate in the sense of lower variability. In addition, DQSmoothLRF also appears to have slightly smaller variability in estimating the first derivative compared to DQSmooth. This is a direct result of using estimated values of Y by a LRF, where the signal is picked up from the noise, instead of using raw values of Y, which is the case for DQSmooth. The LocCubic and LRF models seem to have larger variability than the DQSmooth models. Notice the extremely large variability in the LRF estimator for the first derivative; this justifies the need to enhance derivative estimates based on LRF.

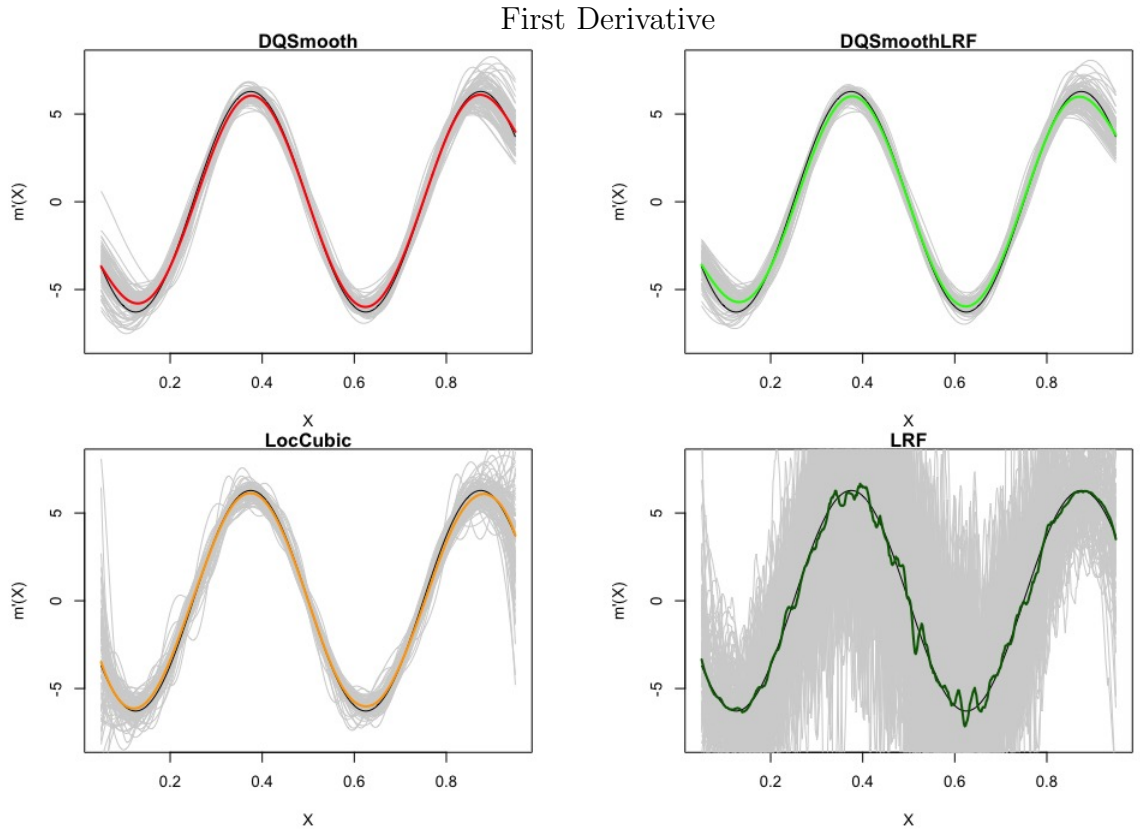

Figure 1: *Each plot shows the estimates for the first derivative for DQSmooth (Liu and Brabanter, 2020), DQSmoothLRF (the proposed estimator), LocCubic, and LRF estimators. Note that the first derivative estimator based on LRF is for* $\lambda = 0$*. The grey curves in each plot depict estimates of the first derivative for one simulation, where 100 simulations are plotted. The solid colored curves represent the mean predicted values across all simulations. The solid black curve represents the true first derivative. All estimators are evaluated at 500 evenly spaced points from 0.05 to 0.95.*

Results for the second derivative are plotted in Figure 2. The colored curves are analagous to those from the first derivative plot. Here, it appears that that most estimators roughly get the overall shape of the second derivative. Similar to the first derivative, the first noticeable difference is the variance of each estimator, where both DQSmooth and DQSmoothLRF estimators have the smallest variance. Although LocCubic seems to have smaller bias, the variance is significantly larger than the DQSmooth procedures. In the last case, LRF has extremely large variance, and there is much improvement on estimating the second derivative based on DQSmoothLRF compared to LRF alone.

Results for the simulations are shown in Table 1, where the bias, variance, MSE, and

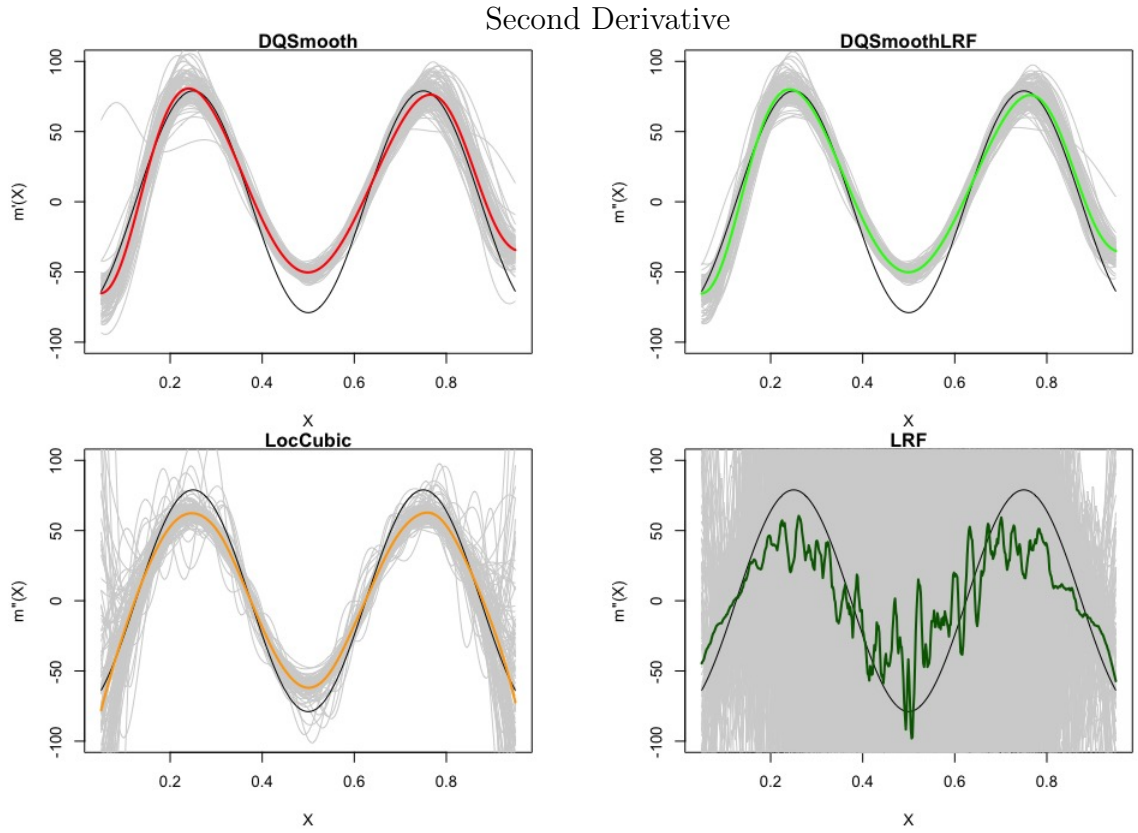

Figure 2: *Each plot shows the estimates for the second derivative for DQSmooth (Liu and Brabanter, 2020), DQSmoothLRF (the proposed estimator), LocCubic, and LRF estimators. Note that the second derivative estimator based on LRF is for $\lambda = 0$. The grey curves in each plot depict estimates of the second derivative for one simulation, where 100 simulations are plotted. The solid colored curves represent the mean predicted values across all simulations. The solid black curve represents the true second derivative. All estimators are evaluated at 500 evenly spaced points from 0.05 to 0.95.*

MAE are reported for both the first and second derivative across the four models under consideration. First, the mean bias from DQSmooth and DQSmoothLRF estimators are roughly the same indicating that the bias of derivative estimates for DQSmoothLRF is similar to that of the bias for DQSmooth. However, improvement on DQSmooth is shown through the variance of the DQSmoothLRF estimator, with a 20% and 32% reduction in the variance relative to the variance of DQSmooth. Although the LocCubic model has lower absolute mean bias compared to all models, the variance is over double of those of the DQSmooth models. The LRF estimator performs the worst and we can see a significant improvement in estimates for the first derivative and second derivative estimator when using DQSmoothLRF.

Lastly, DQSmooth performs the best in terms of both MSE and MAE compared to all models for both the first derivative and second derivative estimations. Overall, in these simulations, we have shown that DQSmoothLRF outperforms all other estimators in terms of variance, MSE, and MAE.

Model Assessment

First Derivative	Bias	Variance	MSE	MAE
DQSmooth	0.0168	0.2618	0.3194	0.4221
DQSmoothLRF	0.0121	0.2085	0.2943	0.4180
LocCubic	-0.0045	0.4437	0.4789	0.4565
LRF	-0.0209	12.9508	13.1182	2.7767
Second Derivative	**Bias**	**Variance**	**MSE**	**MAE**
DQSmooth	5.7984	86.0250	259.8858	12.6552
DQSmoothLRF	5.4323	58.7255	231.5355	12.1516
LocCubic	-1.9564	631.8972	747.2560	15.8320
LRF	-4.1653	18,042.3700	18,892.7700	96.5819

Table 1: *The top and bottom panel show the bias, variance, MSE, and MAE for the first and second derivative respectively, comparing the four models, DQSmooth (Liu and Brabanter, 2020), DQSmoothLRF (the proposed estimator), LocCubic, and LRF. All estimates are averaged across all simulations. Note that the results based on LRF is for* $\lambda = 0$*. All models are evaluated at 500 evenly spaced points from 0.05 to 0.95.*

7 Empirical Application: Convex Technology

A conventional assumption of a production possibility set or technology is convexity (Kerstens and de Woestyne, 2021). One way to check the convexity assumption is to evaluate the curvature of the production function, known as the criterion of quasi-concavity. Sauer (2006) shows that the the law of diminishing marginal productivity in at least one input and quasi-concavity are violated, indicating nonconvexities in agricultural technology. Kerstens and de Woestyne (2021) test the convexity assumption and show that cost functions determined by convex technology are heavily downward biased compared to those determined by non-convex technology. In this paper, we will check the curvature and quasi-concavity criterion

of the production function by estimating its second derivative by the proposed method.

We will use Chilean hydro-electric power generation plants (Atkinson and Dorfman, 2009). To avoid any technical change over time, we single out the year 1997 for 16 power plants with monthly data, providing 188 observations.[4] The data contain one output, electricity generated (q), prices, and quantities of three inputs: capital (k), water (w), and labor (l).[5]

First, the production function can be modeled as an unknown function of the three inputs.

$$q = m(k, w, l) + e. \tag{63}$$

We wish to evaluate the curvature of the production function to test quasi-concavity by estimating its second derivative in each of the three arguments. First, we transform the inputs as in section 4 and can rewrite the model as

$$q = r(U_k, U_w, U_l) + e_i, \tag{64}$$

where U_k, U_w, and U_l are the uniformly transformed variables of capital, water, and labor, respectively. Figure 3 depicts the first (left three plots) and second (right three plots) derivatives of the production function in Eq. (63). The first derivatives with respect to capital $\widehat{m}^{(1)}(k, \bar{w}, \bar{l})$, water $\widehat{m}^{(1)}(\bar{k}, w, \bar{l})$, and labor $\widehat{m}^{(1)}(\bar{k}, \bar{w}, l)$ are evaluated at 200 evenly spaced points across the sample space of each input while holding the other inputs fixed at their medians. The second derivatives are evaluated analogously.

From Figure 3, for capital, the first derivative is positive when capital is small, indicating that electricity output is increasing in capital. The second derivative is positive around this support, indicating that the slope of the production function with respect to capital is increasing, which implies that this part of the production function is convex. As a result the

[4]The data can be found from the Journal of Applied Econometrics Data Archive. Note that there are four missing observations for the 16 power plants during the year 1997.

[5]Further details about the data can be found in Atkinson and Dorfman (2009).

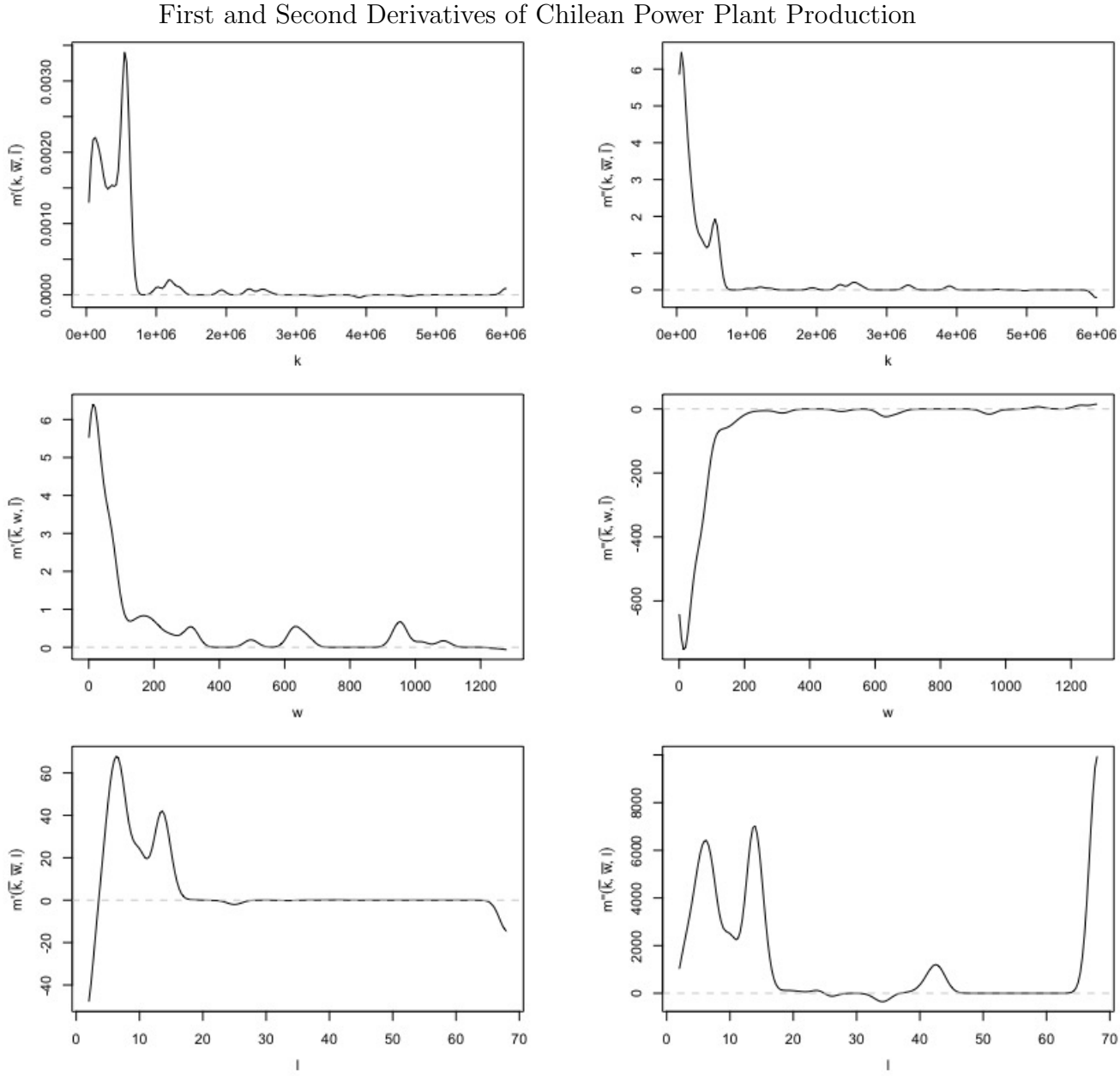

Figure 3: *The three figures on the left and right are the first and second derivatives, respectively, estimated by DQSmoothLRF. All estimates are evaluated across 200 evenly spaced points for each of the three inputs, while holding the other two inputs fixed at their medians.*

level set will be nonconvex, violating the quasi-concavity criterion and convex technology assumption. The bottom plots of Figure 3, referring to the input labor, seem to also violate these assumptions, where the production function is convex in parts of the support for labor. The middle plots show that production is increasing in the input for water and the second

derivative is negative in the start of the support for water. Therefore, production appears to be quasi-concave in water. Overall, electricity production is convex in at least two of its inputs and does not produce convex technology, which would further confirm the nonconvex results obtained in Kerstens and de Woestyne (2021).

Table 2 shows the estimated average first and second derivatives of the production function with respect to the input variables, capital, water, and labor. Average derivatives for OLS are also reported as reference, where a naive linear regression function is estimated. The first derivatives, partial marginal effects, will be constant, and as a result, the second derivative will be zero. All inputs have a positive average partial effect and the production function is increasing in the three inputs. However, when evaluating the first derivative, the average partial effect is underestimated compared to the average effect estimated by DQSmoothLRF for all of the inputs. The second derivative with respect to water is negative on average, implying that the estimated production function is concave in the input water. However, the other two inputs have a positive average second derivative, implying that the estimated production function is convex in the inputs capital and labor. Furthermore, the results from Table 2 regarding the positive second derivatives with respect to capital and labor further justify that the electricity production function breaks the convex technology assumption.

Average Derivatives of Chilean Power Plant Production

	First Derivative		Second Derivative	
	OLS	DQSmoothLRF	OLS	DQSmoothLRF
Capital	0.00001	0.0002	0	0.3264
Water	0.2176	0.5296	0	-30.0765
Labor	0.2815	4.5731	0	975.8699

Table 2: *Average partial first and second derivatives with respect to each of the inputs, capital, water, and labor, are reported. The partial derivatives are evaluated at 200 evenly spaced points across the support of each regressor, where each derivative is held constant for other factors. The reported results are the average of the evaluated 200 derivative points.*

8 Conclusion

Overall, derivatives can help economists find the partial marginal effect (first derivative) of a variable or check convexity assumptions by evaluating the curvature of a production function (second derivative). In this paper, we propose a method, DQSmoothLRF, that smooths random forest based difference quotients to estimate first and second derivatives. We improve on the original method in Liu and Brabanter (2020) by using estimated values of the dependent variable from LRF in forming difference quotients, instead of using the dependent variable itself, in hopes of reducing variance and by including multiple variables in the model, instead of the simple univariate case. Improvement is also made in comparision to derivatives estimated by LRF in Friedberg et al. (2018), where derivatives of the regression equation were not even focused on and in providing better interpretation for forest based models by evaluating derivatives derived from random forests. We have shown in simulation that the proposed estimator outperforms the benchmark ones as well as a popular method of estimating derivatives in economics, local polynomial regression; a reduction in variance, MSE, and MAE are all evident when using the proposed estimator. Lastly, we provide an empirical example using Chilean hydro-electric power generation plants data to assess the curvature of the production function in order to check the convex technology assumption assumed in the literature, in which we found that this assumption is broken for this dataset.

References

Atkinson, S. E. and Dorfman, J. H. "Feasible estimation of firm-specific allocative inefficiency through bayesian numerical methods." *Journal of Applied Econometrics*, 24(4):675–697, 2009. ISSN 08837252, 10991255.

Brabanter, K., Cao, F., Gijbels, I., and Opsomer, J. "Local polynomial regression with correlated errors in random design and unknown correlation structure." *Biometrika*, 105:681–690, 2018. doi:10.1093/biomet/asy025.

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

Breiman, L. "Random forests." *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi:10.1023/A:1010933404324.

Cabrera, J. L. O. *locpol: Kernel Local Polynomial Regression*, 2018. R package version 0.7-0.

Charnigo, R., Hall, B., and Srinivasan, C. "A generalized cp criterion for derivative estimation." *Technometrics*, 53(3):238–253, 2011. doi:10.1198/TECH.2011.09147.

David, H. and Nagaraja, H. *Order Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 1970.

Duong, T. *ks: Kernel Smoothing*, 2020. R package version 1.11.7.

Fan, J. and Gijbels, I. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1996. ISBN 9780412983214.

Fonseca, Y., Medeiros, M., Vasconcelos, G., and Veiga, A. "BooST: Boosting Smooth Trees for Partial Effect Estimation in Nonlinear Regressions." Papers 1808.03698, arXiv.org, 2018.

Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. "Local linear forests." *ArXiv*, abs/1807.11408, 2018.

Iserles, A. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2 edition, 2008. doi:10.1017/CBO9780511995569.

Kerstens, K. and de Woestyne, I. V. "Cost functions are nonconvex in the outputs when the technology is nonconvex: convexification is not harmless." *Annals of Operations Research*, 305:81–106, 2021.

Liu, Y. and Brabanter, K. D. "Smoothed nonparametric derivative estimation using weighted difference quotients." *Journal of Machine Learning Research*, 21(65):1–45, 2020.

Liu, Y. and De Brabanter, K. "Derivative estimation in random design." In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, "Advances in Neural Information Processing Systems 31," pages 3445–3454. Curran Associates, Inc., 2018.

Ma, S., Racine, J., and Ullah, A. "Nonparametric Estimation of Marginal Effects in Regression-spline Random Effects Models." Working Papers 201920, University of California at Riverside, Department of Economics, 2019.

Sauer, J. "Economic theory and econometric practice: Parametric efficiency analysis." *Empirical Economics*, 31(4):1061–1087, 2006. doi:10.1007/s00181-006-0068-3.

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Hirshberg, D., Miner, L., Sverdrup, E., Wager, S., and Wright, M. *grf: Generalized Random Forest*, 2020. R package version 1.2.0.

Wang, W. and Lin, L. "Derivative estimation based on difference sequence via locally weighted least squares regression." *Journal of Machine Learning Research*, 16(81):2617–2641, 2015.

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

A Proof of Proposition 1

$$\begin{aligned}
\mathrm{Var}[\widehat{Y}^{(1)}_{i,ML}|\mathbb{U}] &= \mathrm{Var}\left[\sum_{j=1}^{k} w_{i,j}\frac{\widehat{r}(U_{i+j})-\widehat{r}(U_{i-j})}{U_{i+j}-U_{i-j}}\middle|\mathbb{U}\right] \\
&= \left(1-\sum_{j=2}^{k} w_{i,j}\right)^2 \frac{\mathrm{Var}[\widehat{r}(U_{i+1})|\mathbb{U}]+\mathrm{Var}[\widehat{r}(U_{i-1})|\mathbb{U}]}{(U_{i+1}-U_{i-1})^2} \\
&\quad + \sum_{j=2}^{k} w_{i,j}^2 \frac{\mathrm{Var}[\widehat{r}(U_{i+j})|\mathbb{U}]+\mathrm{Var}[\widehat{r}(U_{i-j})|\mathbb{U}]}{(U_{i+j}-U_{i-j})^2} \\
&= \left(1-\sum_{j=2}^{k} w_{i,j}\right)^2 \frac{\sigma^2_{\widehat{r},i+1}+\sigma^2_{\widehat{r},i-1}}{(U_{i+1}-U_{i-1})^2} + \sum_{j=2}^{k} w_{i,j}^2 \frac{\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j}}{(U_{i+j}-U_{i-j})^2}.
\end{aligned}$$

For all $j = 1, \ldots, k$, take the partial derivative with respect to $w_{i,j}$ and set it to zero results in

$$w_{i,j} = w_{i,1}\frac{(U_{i+j}-U_{i-j})^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})}{(U_{i+1}-U_{i-1})^2/(\sigma^2_{\widehat{r},i+1}+\sigma^2_{\widehat{r},i-1})},$$

which shows the relationship between $w_{i,1}$ and $w_{i,j}$. Since $\sum_{j=1}^{k} w_{i,j} = 1$,

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

Substituting for $w_{i,1}$ gives

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

proving the proposition.

B Proof of Theorem 1

Consider the taylor expansions for $\widehat{r}(U_{i+j})$ and $\widehat{r}(U_{i-j})$ in the neighborhood of U_i

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

where $\zeta_{i,i+j} \in]U_i, U_{i+j}[$ and $\zeta_{i-j,i} \in]U_{i-j}, U_i[$. Then, using Lemma 1 and Proposition 1, the absolute conditional bias is

$$\begin{aligned}
\left|\text{Bias}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}]\right| &= \left|\mathbb{E}\left[\sum_{j=1}^{k} w_{i,j}\frac{\widehat{r}(U_{i+j}) - \widehat{r}(U_{i-j})}{U_{i+j} - U_{i-j}}\middle|\mathbb{U}\right] - r^{(1)}(U_i)\right| \\
&= \left|\sum_{j=1}^{k} w_{i,j}\frac{r(U_{i+j}) + \text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}] - r(U_{i-j}) - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i+j} - U_{i-j}} - r^{(1)}(U_i)\right| \\
&= \left|\sum_{j=1}^{k} w_{i,j}\left[\frac{r(U_{i+j}) - r(U_{i-j})}{U_{i+j} - U_{i-j}} + \frac{\text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i+j} - U_{i-j}}\right] - r^{(1)}(U_i)\right| \\
&= \left|\sum_{j=1}^{k} w_{i,j}\left[\frac{r(U_i) + r^{(1)}(U_i)(U_{i+j} - U_i) + \frac{1}{2}r^{(2)}(\zeta_{i,i+j})(U_{i+j} - U_i)^2}{U_{i+j} - U_{i-j}}\right.\right. \\
&\qquad - \frac{r(U_i) + r^{(1)}(U_i)(U_{i-j} - U_i) + \frac{1}{2}r^{(2)}(\zeta_{i-j,i})(U_{i-j} - U_i)^2}{U_{i+j} - U_{i-j}} \\
&\qquad \left.\left. + \frac{\text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i+j} - U_{i-j}}\right] - r^{(1)}(U_i)\right| \\
&= \left|\frac{1}{2}\sum_{j=1}^{k} w_{i,j}\left[\frac{r^{(2)}(\zeta_{i,i+j})(U_{i+j} - U_i)^2 - r^{(2)}(\zeta_{i-j,i})(U_{i-j} - U_i)^2}{U_{i+j} - U_{i-j}}\right]\right. \\
&\qquad \left. + \sum_{j=1}^{k} w_{i,j}\left[\frac{\text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i+j} - U_{i-j}}\right]\right|,
\end{aligned}$$

where $\zeta_{i,i+j} \in]U_i, U_{i+j}[$ and $\zeta_{i-j,i} \in]U_{i-j}, U_i[$. Now, substitute the ith weight and use $\sigma^2_{\widehat{r},i} =$

$O(n^{-(1-\beta)})$ for all i.

$$
\begin{aligned}
w_{i,j} &= \frac{(U_{i+j} - U_{i-j})^2/(\sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j})}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2/(\sigma^2_{\widehat{r},i+j} + \sigma^2_{\widehat{r},i-j})} \\
&= \frac{(U_{i+j} - U_{i-j})^2}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2}
\end{aligned}
$$

$$
\begin{aligned}
\left| \text{Bias}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] \right| &= \left| \frac{\frac{1}{2}\sum_{j=1}^{k}(U_{i+j} - U_{i-j})(r^{(2)}(\zeta_{i,i+j})(U_{i+j} - U_i)^2 - r^{(2)}(\zeta_{i-j,i})(U_{i-j} - U_i)^2)}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2} \right. \\
&\qquad \left. + \frac{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})(\text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}])}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2} \right| \\
&= \frac{1}{2} \left| \frac{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})(r^{(2)}(\zeta_{i,i+j})(U_{i+j} - U_i)^2 - r^{(2)}(\zeta_{i-j,i})(U_{i-j} - U_i)^2)}{\sum_{j=1}^{k}(U_{i+j} - U_{i-j})^2} \right|
\end{aligned}
$$

The last equality holds since we have subsamples of size s with $s = n^{\beta}$, which allows the errors of the forests to be variance-dominated. Then, the absolute conditional bias is bounded by the same bound provided by Liu and Brabanter (2020) for $k \to \infty$ as $n \to \infty$.

From Proposition 1 the conditional variance is

$$\begin{aligned}
\mathrm{Var}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] &= \mathrm{Var}\left[\sum_{j=1}^{k} w_{i,j}\frac{\widehat{r}(U_{i+j})-\widehat{r}(U_{i-j})}{U_{i+j}-U_{i-j}}\middle|\mathbb{U}\right] \\
&= \mathrm{Var}\left[\sum_{j=1}^{k}\left\{\frac{(U_{i+j}-U_{i-j})^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})}{\sum_{l=1}^{k}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})}\frac{\widehat{r}(U_{i+j})-\widehat{r}(U_{i-j})}{U_{i+j}-U_{i-j}}\right\}\middle|\mathbb{U}\right] \\
&= \mathrm{Var}\left[\frac{\sum_{j=1}^{k}[(U_{i+j}-U_{i-j})/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})][\widehat{r}(U_{i+j})-\widehat{r}(U_{i-j})]}{\sum_{l=1}^{k}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})}\middle|\mathbb{U}\right] \\
&= \frac{\sum_{j=1}^{k}[(U_{i+j}-U_{i-j})^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})^2]\,\mathrm{Var}[\widehat{r}(U_{i+j})-\widehat{r}(U_{i-j})|\mathbb{U}]}{\left(\sum_{l=1}^{k}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})\right)^2} \\
&= \frac{\sum_{j=1}^{k}[(U_{i+j}-U_{i-j})^2/(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})^2][(\sigma^2_{\widehat{r},i+j}+\sigma^2_{\widehat{r},i-j})]}{\left(\sum_{l=1}^{k}(U_{i+l}-U_{i-l})^2/(\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l})\right)^2} \\
&= \sum_{l=1}^{k}\frac{\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l}}{(U_{i+l}-U_{i-l})^2}.
\end{aligned}$$

Then, using Lemma 1 and $\sigma^2_{\widehat{r},i} = O(n^{-(1-\beta)})$ for all i, the conditional variance is bounded above by

$$\begin{aligned}
\mathrm{Var}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] &= \sum_{l=1}^{k}\frac{\sigma^2_{\widehat{r},i+l}+\sigma^2_{\widehat{r},i-l}}{(U_{i+l}-U_{i-l})^2} \\
&\leq 2n^{-(1-\beta)}\sum_{l=1}^{k}\frac{1}{(U_{i+l}-U_{i-l})^2} \\
&= 2n^{-(1-\beta)}\frac{1}{\frac{2k(k+1)(2k+1)}{3(n+1)^2}\{1+o_p(1)\}} \\
&= \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)}\{1+o_p(1)\}
\end{aligned}$$

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

C Proof of Corollary 1

As $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$ and $n^{(1+\beta)}k^{-3}$ from Theorem 1, the upperbound

of the conditional bias and conditional variance tend to zero. Therefore,

$$\lim_{n\to\infty} \text{MSE}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] = 0.$$

Use Chebyshev's inequality to complete the proof.

D Proof of Corollary 2

Using the bias-variance decomposition of means squared error (MSE), the MSE is bounded above by

$$\text{MSE}[\widehat{Y}^{(1)}_{i,LRF}|\mathbb{U}] \leq \left(\mathcal{B}\frac{3k(k+1)}{4(n+1)(2k+1)}\right)^2 + \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)}.$$

Then, the conditional mean integrated squared error (MISE), is

$$\begin{aligned}
\text{MISE}[\widehat{Y}^{(1)}_{ML}|\mathbb{U}] &= \mathbb{E}\int_0^1 \left(\widehat{Y}^{(1)}_{LRF}(U) - r^{(1)}(U)|\mathbb{U}\right)^2 dU \\
&= \int_0^1 \mathbb{E}\left(\widehat{Y}^{(1)}_{LRF}(U) - r^{(1)}(U)|\mathbb{U}\right)^2 dU \\
&\leq \left(\mathcal{B}\frac{3k(k+1)}{4(n+1)(2k+1)}\right)^2 + \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)} + o_p(n^{-2}k^2 + n^{(1+\beta)}k^{-3})
\end{aligned}$$

Therefore, the asymptotic conditional MISE (AMISE) is

$$\text{AMISE}[\widehat{Y}^{(1)}_{LRF}|\mathbb{U}] \leq \left(\mathcal{B}\frac{3k(k+1)}{4(n+1)(2k+1)}\right)^2 + \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)}.$$

E Proof of Theorem 2

Conditional bias:

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

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

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

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

$$\begin{aligned}
\boldsymbol{\epsilon}_1^\top \mathbf{S}_{n-2k}^{-1}\mathbf{U}^\top\mathbf{W}\begin{bmatrix} \text{Bias}[\widehat{Y}^{(1)}_{k+1,LRF}|\mathbb{U}] \\ \vdots \\ \text{Bias}[\widehat{Y}^{(1)}_{n-k,LRF}|\mathbb{U}] \end{bmatrix} &\le \boldsymbol{\epsilon}_1^\top \mathbf{S}_{n-2k}^{-1}\mathbf{U}^\top\mathbf{W}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}\sup_{u\in[0,1]}|r^{(2)}(u)|\frac{3k(k+1)}{4(n+1)(2k+1)}\{1+o_p(1)\} \\
&\le \sup_{u\in[0,1]}|r^{(2)}(u)|\frac{3k(k+1)}{4(n+1)(2k+1)}\boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1}\tilde{c}_p\{1+o_p(1)\},
\end{aligned}$$

and since from Liu and Brabanter (2020), it is shown that

$$\begin{aligned}\mathbf{S}_{n-2k} &= \mathbf{U}^\top \mathbf{W}\mathbf{U} \\ &= (n-2k) f(u_0) H \mathbf{S} H \{1 + o_p(1)\},\end{aligned}$$

where $H = \text{diag}\{1, h, \ldots, h^p\}$, and

$$\mathbf{U}^\top \mathbf{W} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = (n-2k) f(u_0) H \tilde{c}_p \{1 + o_p(1)\} 1,$$

where $\tilde{c}_p = (\mu_0, \mu_1, \ldots, \mu_p)^\top$. Finally, the conditional bias of the smoothed derivative estimator is bounded by

$$\text{Bias}[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}] \leq \boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1} \left[\frac{c_p}{(p+1)!} r^{(p+2)}(u_0) h^{p+1} + \sup_{u \in [0,1]} |r^{(2)}(u)| \frac{3k(k+1)}{4(n+1)(2k+1)} \right] \{1 + o_p(1)\}$$

Conditional Variance:

When $k \to \infty$ as $n \to \infty$, the conditional variance from Theorem 1 is

$$\text{Var}[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}] = \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)} \{1 + o_p(1)\}$$

and from Theorem 1 of Brabanter et al. (2018),

$$\begin{aligned}\text{Var}[\widehat{r}^{(1)}(u_0)|\widetilde{\mathbb{U}}] &= \boldsymbol{\epsilon}_1^\top \mathbf{S}_{n-2k}^{-1} (\mathbf{U}^\top \mathbf{W} \, \text{Var}[\widehat{\mathbf{Y}}^{(1)}|\widetilde{\mathbb{U}}] \mathbf{W}\mathbf{U}) \mathbf{S}_{n-2k}^{-1} \boldsymbol{\epsilon}_1 \\ &\leq \frac{3n^{-(1-\beta)}(n+1)^2}{k(k+1)(2k+1)} \frac{1 + f(u_0)\rho_c}{h(n-2k)f(u_0)} \boldsymbol{\epsilon}_1^\top \mathbf{S}^{-1} \mathbf{S}^* \mathbf{S}^{-1} \boldsymbol{\epsilon}_1 \{1 + o_p(1)\},\end{aligned}$$

where $\lim_{n\to\infty} n \int \rho_n(x) dx = \rho_c$ and $\mathbf{S}^* = (\nu_{i+j})_{0 \leq i,j \leq p}$ with $\nu_j = \int u^j K^2(u) du$. For p odd,

from Theorem 3.1 of Fan and Gijbels (1996),

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

Similarly,

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

F Proof of Corollary 3

For $\to 0$, $nh \to \infty$ and $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$ and $n^\beta k^{-3}h^{-1} \to 0$, then from Theorem 2, the upperbound of the conditional bias and conditional variance go to zero. Then,

$$\lim_{n\to\infty} \text{MSE}[\widehat{r}^{(1)}(u_0)|\tilde{\mathbb{U}}] = 0$$

and use Chebyshev's inequality to complete the proof.

G Proof of Proposition 2

Under the assumptions of Proposition 2 and using Lemma 1,

$$\begin{aligned}\mathbb{E}[{}^+\widehat{Y}^{(1)}_{i+j,LRF} -^- \widehat{Y}^{(1)}_{i-j,LRF}|\mathbb{U}] &= \frac{r(U_{i+j+k_1}) + \text{Bias}[\widehat{r}(U_{i+j+k_1})|\mathbb{U}] - r(U_{i+j}) - \text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}]}{U_{i+j+k_1} - U_{i+j}} \\ &\quad - \frac{r(U_{i-j-k_1}) + \text{Bias}[\widehat{r}(U_{i-j-k_1})|\mathbb{U}] - r(U_{i-j}) - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i-j-k_1} - U_{i-j}} \\ &= r^{(1)}(U_{i+j}) + \frac{1}{2}r^{(2)}(U_{i+j})(U_{i+j+k_1} - U_{i+j})(1 + o_p(1)) \\ &\quad - r^{(1)}(U_{i-j}) + \frac{1}{2}r^{(2)}(U_{i-j})(U_{i-j-k_1} - U_{i-j})(1 + o_p(1)) \\ &\quad + \frac{\text{Bias}[\widehat{r}(U_{i+j+k_1})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}]}{U_{i+j+k_1} - U_{i+j}}\end{aligned}$$

$$
\begin{aligned}
&- \frac{\text{Bias}[\widehat{r}(U_{i-j-k_1})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i-j-k_1} - U_{i-j}} \\
&= \frac{1}{2} r^{(2)}(U_i)(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})(1 + o_p(1)) \\
&\quad + \frac{\text{Bias}[\widehat{r}(U_{i+j+k_1})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i+j})|\mathbb{U}]}{U_{i+j+k_1} - U_{i+j}} \\
&\quad - \frac{\text{Bias}[\widehat{r}(U_{i-j-k_1})|\mathbb{U}] - \text{Bias}[\widehat{r}(U_{i-j})|\mathbb{U}]}{U_{i-j-k_1} - U_{i-j}} \\
&= \frac{1}{2} r^{(2)}(U_i)(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})(1 + o_p(1)),
\end{aligned}
$$

where the subsample rate β is chosen such that $\beta_{min} < \beta < 1$, the LRF estimator is asymptotically unbiased. The weight for observation i is selected to be proportional to the inverse of the conditional variance of each quotient,

$$
\begin{aligned}
w_{i,j,2} &= \frac{1/\operatorname{Var}\left[\frac{+\widehat{Y}^{(1)}_{i+j,LRF} - -\widehat{Y}^{(1)}_{i-j,LRF}}{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})} \middle| \mathbb{U}\right]}{\sum_{j=1}^{k_2} 1/\operatorname{Var}\left[\frac{+\widehat{Y}^{(1)}_{i+j,LRF} - -\widehat{Y}^{(1)}_{i-j,LRF}}{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})} \middle| \mathbb{U}\right]} \\
&= \frac{1/\left[\frac{\frac{\sigma^2_{\widehat{r},i+j+k_1} + \sigma^2_{\widehat{r},i+j}}{(U_{i+j+k_1} + U_{i+j})^2} + \frac{\sigma^2_{\widehat{r},i-j-k_1} + \sigma^2_{\widehat{r},i-j}}{(U_{i-j-k_1} + U_{i-j})^2}}{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})^2}\right]}{\sum_{j=1}^{k_2} 1/\left[\frac{\frac{\sigma^2_{\widehat{r},i+j+k_1} + \sigma^2_{\widehat{r},i+j}}{(U_{i+j+k_1} + U_{i+j})^2} + \frac{\sigma^2_{\widehat{r},i-j-k_1} + \sigma^2_{\widehat{r},i-j}}{(U_{i-j-k_1} + U_{i-j})^2}}{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})^2}\right]},
\end{aligned}
$$

where $\sigma^2_{\widehat{r}}$ is the variance of the LRF estimator.

H Proof of Theorem 3

Since r is three times continuously differentiable on the interval $[0, 1]$, consider the Taylor

expansions for $r(U_{i+j+k_1})$ and $r(U_{i-j-k_1})$ in the neighborhood of U_{i+j} and U_{i-j} respectively

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

where $\zeta_{i+j,i+j+k_1} \in]U_{i+j}, U_{i+j+k_1}[$ and $\zeta_{i-j-k_1,i-j} \in]U_{i-j-k_1}, U_{i-j}[$. Also, consider the Taylor expansions for $r^{(1)}(U_{i+j})$, $r^{(1)}(U_{i-j})$, $r^{(2)}(U_{i+j})$, and $r^{(2)}(U_{i-j})$ in the neighborhood of U_i

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

where $\zeta_{i,i+j} \in]U_i, U_{i+j}[$, $\zeta_{i-j,i} \in]U_{i-j}, U_i[$, $\zeta'_{i,i+j} \in]U_i, U_{i+j}[$, and $\zeta'_{i-j,i} \in]U_{i-j}, U_i[$. The absolute conditional bias of $\hat{Y}^{(2)}_{i,LRF}$ is

$$\begin{aligned}
\left| \text{Bias}[\hat{Y}^{(2)}_{i,LRF} | \tilde{\mathbb{U}}] \right| &= \left| \mathbb{E}[\hat{Y}^{(2)}_{i,LRF}] - r^{(2)}(U_i) \right| \\
&= \left| \mathbb{E}\left[2 \sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{\widehat{r}(U_{i+j+k_1}) - \widehat{r}(U_{i+j})}{U_{i+j+k_1} - U_{i+j}} - \frac{\widehat{r}(U_{i-j-k_1}) - \widehat{r}(U_{i-j})}{U_{i-j-k_1} - U_{i-j}} \right)}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \right] - r^{(2)}(U_i) \right| \\
&= \left| 2 \sum_{j=1}^{k_2} \left\{ w_{i,j,2} \frac{\left(\frac{r(U_{i+j+k_1}) - r(U_{i+j})}{U_{i+j+k_1} - U_{i+j}} - \frac{r(U_{i-j-k_1}) - r(U_{i-j})}{U_{i-j-k_1} - U_{i-j}} \right)}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \right. \right. \\
&\quad \left. \left. + \frac{\left(\frac{\text{Bias}[\widehat{r}(U_{i+j+k_1})] - \text{Bias}[\widehat{r}(U_{i+j})]}{U_{i+j+k_1} - U_{i+j}} - \frac{\text{Bias}[\widehat{r}(U_{i-j-k_1})] - \text{Bias}[\widehat{r}(U_{i-j})]}{U_{i-j-k_1} - U_{i-j}} \right)}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} \right\} - r^{(2)}(U_i) \right| \\
&= \left| 2 \sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{r(U_{i+j+k_1}) - r(U_{i+j})}{U_{i+j+k_1} - U_{i+j}} - \frac{r(U_{i-j-k_1}) - r(U_{i-j})}{U_{i-j-k_1} - U_{i-j}} \right)}{U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j}} - r^{(2)}(U_i) \right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{u\in[0,1]} |r^{(3)}(u)| \left(\sum_{j=1}^{k_2} w_{i,j,2} \frac{(U_{i+j}-U_i)^2 + (U_{i-j}-U_i)^2}{U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j}} \right. \\
&\quad + \sum_{j=1}^{k_2} w_{i,j,2} \frac{(U_{i+j}-U_i)(U_{i+j+k_1}-U_{i+j}) + (U_{i-j}-U_i)(U_{i-j-k_1}-U_{i-j})}{U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j}} \\
&\quad \left. + \sum_{j=1}^{k_2} w_{i,j,2} \frac{\frac{1}{3}(U_{i+j+k_1}-U_{i+j})^2 + \frac{1}{3}(U_{i-j-k_1}-U_{i-j})^2}{U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j}} \right)
\end{aligned}
$$

where the last equality holds for the subsample rate β such that $\beta_{min} < \beta < 1$. Then, using the weights in Eq. (32), Lemma 1, and $\sigma_{\widehat{r}} = O(n^{-(1-\beta)})$, where $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ gives

$$
\left| \text{Bias}[\hat{Y}^{(2)}_{i,LRF}|\tilde{\mathbb{U}}] \right| \leq \frac{\sup_{u\in[0,1]} |r^{(3)}(u)|}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \{1 + o_p(1)\}
$$

The variance of the second derivative estimator, $\hat{Y}^{(2)}_{i,LRF}$, is

$$
\begin{aligned}
\text{Var}[\hat{Y}^{(2)}_{i,LRF}|\tilde{\mathbb{U}}] = \text{Cov} \Bigg[& 2\sum_{j=1}^{k_2} w_{i,j,2} \frac{\left(\frac{\widehat{r}(U_{i+j+k_1})-\widehat{r}(U_{i+j})}{U_{i+j+k_1}-U_{i+j}} - \frac{\widehat{r}(U_{i-j-k_1})-\widehat{r}(U_{i-j})}{U_{i-j-k_1}-U_{i-j}} \right)}{U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j}}, \\
& 2\sum_{l=1}^{k_2} w_{i,l,2} \frac{\left(\frac{\widehat{r}(U_{i+l+k_1})-\widehat{r}(U_{i+l})}{U_{i+l+k_1}-U_{i+l}} - \frac{\widehat{r}(U_{i-l-k_1})-\widehat{r}(U_{i-l})}{U_{i-l-k_1}-U_{i-l}} \right)}{U_{i+l+k_1}+U_{i+l}-U_{i-l-k_1}-U_{i-l}} \Bigg] \\
= 4 & \sum_{j=1}^{k_2}\sum_{l=1}^{k_2} \frac{w_{i,j,2}w_{i,l,2}}{(U_{i+j+k_1}+U_{i+j}-U_{i-j-k_1}-U_{i-j})(U_{i+l+k_1}+U_{i+l}-U_{i-l-k_1}-U_{i-l})} \\
& \Bigg\{ \frac{\text{Cov}[\widehat{r}(U_{i+j+k_1})-\widehat{r}(U_{i+j}), \widehat{r}(U_{i+l+k_1})-\widehat{r}(U_{i+l})]}{(U_{i+j+k_1})-U_{i+j})((U_{i+l+k_1})-U_{i+l})} \\
& - \frac{\text{Cov}[\widehat{r}(U_{i+j+k_1})-\widehat{r}(U_{i+j}), \widehat{r}(U_{i-l-k_1})-\widehat{r}(U_{i-l})]}{(U_{i+j+k_1})-U_{i+j})((U_{i-l-k_1})-U_{i-l})} \\
& - \frac{\text{Cov}[\widehat{r}(U_{i-j-k_1})-\widehat{r}(U_{i-j}), \widehat{r}(U_{i+l+k_1})-\widehat{r}(U_{i+l})]}{(U_{i-j-k_1})-U_{i-j})((U_{i+l+k_1})-U_{i+l})} \\
& + \frac{\text{Cov}[\widehat{r}(U_{i-j-k_1})-\widehat{r}(U_{i-j}), \widehat{r}(U_{i-l-k_1})-\widehat{r}(U_{i-l})]}{(U_{i-j-k_1})-U_{i-j})((U_{i-l-k_1})-U_{i-l})} \Bigg\},
\end{aligned}
$$

where the first covariance is

$$\begin{aligned}\mathrm{Cov}[\widehat{r}(U_{i+j+k_1}) - \widehat{r}(U_{i+j}), \widehat{r}(U_{i+l+k_1}) - \widehat{r}(U_{i+l})] &= \mathrm{Cov}[\widehat{r}(U_{i+j+k_1}), \widehat{r}(U_{i+l+k_1})] \\ &\quad - \mathrm{Cov}[\widehat{r}(U_{i+j}), \widehat{r}(U_{i+l+k_1})] \\ &\quad - \mathrm{Cov}[\widehat{r}(U_{i+j+k_1}), \widehat{r}(U_{i+l})] \\ &\quad + \mathrm{Cov}[\widehat{r}(U_{i+j}), \widehat{r}(U_{i+l})]\end{aligned}$$

The first and fourth covariances are $\sigma^2_{\widehat{r},i+j+k_1}$ and $\sigma^2_{\widehat{r},i+j}$ respectively when $j = l$, the second covariance is $\sigma^2_{\widehat{r},i+l+k_1}$ when $j = l + k_1$, and the third covariance is $\sigma^2_{\widehat{r},i+l}$ when $j + k_1 = l$. The other covariances can be obtained in a similar fashion. Now, using the weights in Eq. (32), $\sigma^2_{\widehat{r}} = O(n^{-(1-\beta)})$, and Lemma 1, where $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$,

$$\begin{aligned}\mathrm{Var}[\hat{Y}^{(2)}_{i,LRF}|\tilde{\mathbb{U}}] &= \frac{4}{n^{1-\beta}} \sum_{j=1}^{k_2} \frac{w_{i,j,2}}{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})^2} \left(\frac{2}{(U_{i+j+k_1} - U_{i+j})^2} + \frac{2}{(U_{i-j-k_1} - U_{i-j})^2} \right) \\ &\quad - \frac{4}{n^{1-\beta}} \sum_{j=1}^{k_2-k_1} \frac{w_{i,j,2} w_{i,j+k_1,2}}{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})(U_{i+j+2k_1} + U_{i+j+k_1} - U_{i-j-2k_1} - U_{i-j-k_1})} \\ &\quad \left(\frac{1}{(U_{i+j+k_1} - U_{i+j})(U_{i+j+2k_1} - U_{i+j+k_1})} + \frac{1}{(U_{i-j-k_1} - U_{i-j})(U_{i-j-2k_1} - U_{i-j-k_1})} \right) \\ &\quad - \frac{4}{n^{1-\beta}} \sum_{j=1+k_1}^{k_2} \frac{w_{i,j,2} w_{i,j-k_1,2}}{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})(U_{i+j} + U_{i+j-k_1} - U_{i-j} - U_{i-j+k_1})} \\ &\quad \left(\frac{1}{(U_{i+j+k_1} - U_{i+j})(U_{i+j} - U_{i+j-k_1})} + \frac{1}{(U_{i-j-k_1} - U_{i-j})(U_{i-j} - U_{i-j+k_1})} \right) \\ &\leq \frac{4}{n^{1-\beta}} \sum_{j=1}^{k_2} \frac{w_{i,j,2}}{(U_{i+j+k_1} + U_{i+j} - U_{i-j-k_1} - U_{i-j})^2} \left(\frac{2}{(U_{i+j+k_1} - U_{i+j})^2} + \frac{2}{(U_{i-j-k_1} - U_{i-j})^2} \right) \\ &= \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2 \sum_{j=1}^{k_2} (2j + k_1)^2} \{1 + o_p(1)\}.\end{aligned}$$

I Proof of Corollary 4

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

$$\left| \text{Bias}[\hat{Y}^{(2)}_{i,LRF}|\tilde{\mathbb{U}}] \right| \leq \frac{\sup_{u\in[0,1]} |r^{(3)}(u)|}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \{1 + o_p(1)\}$$
$$= O_p\left(\max\left\{ \frac{k_1}{n}, \frac{k_2}{n} \right\} \right)$$

and

$$\text{Var}[\widehat{Y}^{(2)}_{i,LRF}|\mathbb{U}] \leq \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2 \sum_{j=1}^{k_2} (2j + k_1)^2} \{1 + o_p(1)\}$$
$$= O_p\left(\max\left\{ \frac{n^{3+\beta}}{k_1^2 k_2^3}, \frac{n^{3+\beta}}{k_1^4 k_2} \right\} \right)$$

Then, for $k_1 \to \infty$ and $k_2 \to \infty$ as $n \to \infty$ such that $n^{-1}k_1 \to 0$, $n^{-1}k_2 \to 0$, $n^{3+\beta}k_1^{-2}k_2^{-3} \to 0$, and $n^{3+\beta}k_1^{-4}k_2^{-1} \to 0$, the conditional bias and conditional variance tend to zero. Therefore,

$$\lim_{n\to\infty} \text{MSE}[\widehat{Y}^{(2)}_i|\tilde{\mathbb{U}}] = 0.$$

Use Chebyshev's inequality to complete the proof.

J Proof of Corollary 5

Using the bias-variance decomposition of means squared error (MSE), the MSE is bounded above by

$$\text{MSE}[\widehat{Y}^{(1)}_{i,LRF}|\tilde{\mathbb{U}}] \leq \left(\frac{\mathcal{B}_2}{n+1} \frac{2\sum_{j=1}^{k_2} j^3 + 3k_1 \sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2 \sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1 \sum_{j=1}^{k_2} j} \right)^2 \{1 + o_p(1)\}$$
$$+ \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2 \sum_{j=1}^{k_2} (2j + k_1)^2} \{1 + o_p(1)\}$$

Then, the conditional mean integrated squared error (MISE), is

$$\begin{aligned}
\text{MISE}[\widehat{Y}^{(2)}_{ML}|\tilde{\mathbb{U}}] &= \mathbb{E}\int_0^1 \left(\widehat{Y}^{(2)}_{LRF}(U) - r^{(2)}(U)|\mathbb{U}\right)^2 dU \\
&= \int_0^1 \mathbb{E}\left(\widehat{Y}^{(2)}_{LRF}(U) - r^{(2)}(U)|\mathbb{U}\right)^2 dU \\
&\leq \left(\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2 \{1 + o_p(1)\} \\
&\quad + \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}\{1 + o_p(1)\}
\end{aligned}$$

Therefore, the asymptotic conditional MISE (AMISE) is

$$\text{AMISE}[\widehat{Y}^{(2)}_{LRF}|\tilde{\mathbb{U}}] \leq \left(\frac{\mathcal{B}_2}{n+1}\frac{2\sum_{j=1}^{k_2} j^3 + 3k_1\sum_{j=1}^{k_2} j^2 + \frac{5}{3}k_1^2\sum_{j=1}^{k_2} j + \frac{1}{3}k_1^3 k_2}{4\sum_{j=1}^{k_2} j^2 + k_1^2 k_2 + 4k_1\sum_{j=1}^{k_2} j}\right)^2 + \frac{4n^{-(1-\beta)}(n+1)^4}{k_1^2\sum_{j=1}^{k_2}(2j+k_1)^2}.$$

K More Simulation Studies

The table below shows simulations, under the same DGP in section 3, for two other candidate estimators that are not discussed in the paper. The first is SmoothLRF, where the a LRF is trained on the data, and the fitted values are obtained, $\widehat{y}_{LRF}$. Then a local polynomial regression is used on the new dataset $(x, \widehat{y}_{LRF})$ and the first derivative can be obtained from the second element of the gradient vector. The second is DoubleLocCubic where a local cubic regression is trained on the data and the first derivative is obtained, denoted by $\widehat{y}^{(1)}_{LocCubic}$. Then another local cubic regression is used on the new dataset, $(x, \widehat{y}^{(1)}_{LocCubic})$ and the first derivative can be otbained from the first element of the gradient vector.

For the proposed estimator, DQSmoothLRF, and local cubic regression, LocCub, the results are the same as those in section 6 and are here for comparison. Overall, DQSmoothLRF still outperforms the other estimators, where SmoothLRF does worse in reducing variance, MSE, and MAE. However, the derivative estimated by SmoothLRF does significantly better than the derivative estimated by LRF. DoubleLocCubic does improve upon LocCub, show-

ing a reduction in variance, MSE, and MAE. Overall, DQSmoothLRF seems to be robust in estimating the first derivative, even to other candidate estimators.

Simulations of Other Candidate Estimators

	Bias	Variance	MSE	MAE
DQSmoothLRF	0.0121	0.2085	0.2943	0.4180
SmoothLRF	0.0011	0.3457	0.3699	0.4395
LocCubic	-0.0045	0.4437	0.4789	0.4565
DoubleLocCubic	-0.0046	0.4312	0.4667	0.4503

Table 3: *The table shows bias, variance, MSE, and MAE for the first derivative, comparing four models, DQSmoothLRF (the proposed estimator), LocCubic, and DoubleLocCubic. All estimates are averaged across all simulations. All models are evaluated at 500 evenly spaced points from 0.05 to 0.95.*