Asymptotically First-Order Optimal Bandwidth Selection for Differences of Nonparametric Estimators*

Yoichi Arai†
National Graduate Institute for Policy Studies

Hidehiko Ichimura‡
Faculty of Economics University of Tokyo

This version: February 22, 2012

Preliminary Draft

PLEASE DO NOT CIRCULATE WITHOUT AUTHORS PERMISSION

Abstract

We consider the problem of bandwidth selection to estimate the difference of two functions at points. Bandwidth selection procedures based on the asymptotic approximation of the mean squared error (AMSE) criterion are investigated. We present the simultaneous selection method of two distinct bandwidths among a class that encompass the existing methods as special cases. We show that the asymptotic behavior of the AMSE reveals dichotomous characteristics depending on the sign of the product of the second derivatives for the underlying functions and that the optimal bandwidths that minimize the AMSE does not exist when the sign of the product is positive. We introduce the modified version of the AMSE (MMSE) that exploits the higher-order bias terms as the nonvanishing penalty. We also propose "the asymptotically first-order optimal bandwidths" that are well-defined regardless of the sign of the product. The feasible version of the proposed bandwidths automatically adjust to dichotomous situation and are shown to be asymptotically as good as the asymptotically first-order optimal bandwidths. Our approach is general enough to cover the estimation problems of the difference of densities at points and the difference of regression functions at interior points as well as boundary points. We provide a detailed treatment on the application of the sharp regression discontinuity design and the modest simulation study show the promising results.

Key words: Bandwidth selection, kernel density estimation, regression discontinuity design, local linear regression, local polynomial regression
JEL Classification: C13, C14, C21, C31

1 Introduction

Many nonparametric estimators of an unknown function, such as Lebesgue density function or regression function, are proposed to reduce the misspecification concerns. Given a particular nonparametric

*Earlier versions of this paper has been presented at the Japanese Economic Association Spring Meeting in 2011 and the North American Winter Meeting of the Econometric Society in 2012. We are especially grateful to Jack Porter and Yoshihiko Nishiyama for many helpful comments.

†Arai's research was supported by Grant-in-Aid for Scientific Research (No. 23330070).

‡Ichimura's research was supported by Grant-in-Aid for Scientific Research (No. 22243020).

estimator, it is well recognized that choosing an appropriate smoothing parameter is a key implementation issue and various methods have been proposed. Among vast developments in nonparametric estimation methods, those in program evaluation highlight the need to estimate the difference of two functions at points rather than an unknown function itself. Examples include applications of the average treatment effect (ATE), the local average treatment effect (LATE) and the regression discontinuity design (RDD). For example, the average treatment effect is a parameter of interest in the sharp RDD and it is defined by the difference between the right hand limit of one function at a point and the left hand limit of another function at the same point.

The current standard approach in this context is to choose two bandwidths individually by ad hoc methods such as plug-in or cross-validation methods that are proposed to estimate a single function. One notable exception is the bandwidth selection procedure proposed by by Imbens and Kalyanaraman (2012). It is developed for the RDD estimator to choose a single bandwidth to estimate two functions at a point with paying attention to both functions. In this paper we propose the simultaneous selection method of two distinct bandwidths among a class that encompass the existing methods as special cases. The class we consider is the one that can be obtained by minimizing the asymptotic mean square error (AMSE) criterion. As Imbens and Kalyanaraman (2012) emphasize in the context of the RDD, the problem considered in this paper is how to choose local bandwidths rather than global bandwidths. Hence the bandwidths selection based on either the asymptotic mean "integrated" squared errors or cross-validation criterion cannot be never optimal since the parameter of interest in this paper is the difference of two functions at points rather than two functions on their entire supports.

To illustrate the problem and our approach, we start with the problem of nonparametric estimation for the difference of densities evaluated at two distinct points since density estimation problems are the simplest yet have all the essential features we explore later. First we show that the bandwidths that minimize the AMSE exhibits dichotomous characteristics depending on the sign of the product of the second derivatives for the the density functions at two distinct points. When the sign is negative, the bandwidths that minimizes the AMSE is well-defined and the order of both bandwidths is $n^{-1/5}$ that is the same as the order for the existing methods although they differ from ones based on ad hoc method where n stands for the number of observations. On the other hand, when the sign of the product is positive, the trade-off between bias and variance that is a key component of the optimal bandwidth selection turns out to break down and the AMSE can be made arbitrarily small without cost of increasing bias. This happens because there exists a linear combination of the bandwidths where the first-order bias terms cancel each other, and leads to large values of the bandwidths.

We take the view that the underlying functions are smooth (at least four times continuously differentiable) and consider the AMSE with the second-order bias term as the penalty that punishes large bandwidths. It turns out, however, that there exists a combination of bandwidths that can make the square of the bias term arbitrarily small, leading to large values of the bandwidths again. This analysis indicates that introducing higher-order bias terms only does not help to avoid disappearance of the trade-off. It is inevitable to make some modification on the AMSE to introduce the penalty. This finding also points out necessity for alternative optimality criterion. Hence, we propose a modified version of the AMSE (abbreviated by MMSE) to bring in a nonvanishing trade-off between bias and variance. It contains the sum of the squared first- and the squared second-order bias terms as the bias component rather than the square of their sum. We then propose "the asymptotically first-order optimal bandwidths" that are well-defined regardless of the sign of the product with the dichotomous behaviors being still present. The asymptotically first-order optimal bandwidths are of the order of $n^{-1/5}$ when the sign of the product is negative and $n^{-1/9}$ when the sign of the product is negative. Then it is shown that the bandwidths that minimizes the MMSE is asymptotically equivalent to the asymptotically first-order optimal bandwidths.

Neither the bandwidths based on the MMSE nor the asymptotically first-order optimal bandwidths are practicable since they depend on unknown quantities including the sign of the product of the second derivatives. We propose a feasible version of the asymptotically first-order optimal

bandwidths that is based on the plug-in principle. It is shown that the proposed method automatically adjusts to each of dichotomous situation without knowledge on the sign. We also show that the feasible bandwidths are asymptotically as good as the asymptotically first-order optimal bandwidths.

The proposed methods are generalized to the problem to estimate the difference of regression functions at interior points. The nonparametric regression estimators we consider are local linear estimators since, as shown in Fan (1992), they have desirable properties in terms of asymptotic efficiency, design adaptive property as well as boundary effects. First, we derive the AMSE with the second-order bias term and it can be shown that the essential features of the AMSE is exactly the same as for the estimation problem of the difference of densities. That is, it exhibits the dichotomous characteristics now depending on the sign of the product of the second derivatives for regression functions. Then we propose the MMSE and the feasible version of the bandwidths based on it. The proposed bandwidths are proved to possess all the desirable asymptotic properties as in the case of densities.

We extend the proposed method further to estimate the difference of regression functions at boundary points. Among various possibilities, we consider a problem to estimate the treatment effect in the sharp RDD. Again, we employ the local linear estimators especially because of its automatic boundary carpentry. First, we present the AMSE with the second-order bias term that now involve the lower-order term in the bandwidths comparing two previous cases because of the boundary kernel function. Consequently, the MMSE involve the lower-order term and the asymptotically first-order optimal bandwidths are of the order of $n^{-1/7}$ when the sign of the product is positive although they are the same order when the sign of the product is negative. Despite the differences, the essential features remain the same and the feasible version of the bandwidths can be proposed in the same manner and it can be shown that they are asymptotically as good as the asymptotically first-order optimal bandwidths.

After theoretical developments, we conduct a small simulation study to investigate the finite sample properties of the proposed method. We concentrate on the case of the RDD since this would be one of the most empirically relevant cases. Our limited experiment shows the promising performance of the proposed. We also provide discussion about the two additional cases where the second derivative vanishes at one of the two points. The leading bias term at the point the second derivative vanishes is of lower order than the other point but nonetheless if the signs are the same, the bias term can be cancelled and if the signs differ the bias term cannot be cancelled.

The remainder of the paper is organized as follows. In Section 2, all the essential features of our approach is presented through the estimation problem of the difference of densities at points. We generalizes the proposed method to the estimation problem of regression functions at interior points as well as boundary points in Section 3. In Section 4, we demonstrate the finite sample behavior of our approach via a small simulation study. Section 5 concludes and briefly discusses the excluded cases.

2 Nonparametric Estimation for Difference of Densities

2.1 Asymptotic Approximation of the MSE for the Difference of Kernel Density Estimators

We illustrate the nature of the problem by considering the simple case of estimating the difference of the density at two points, i.e.

$$f(x_1) - f(x_2),$$

for $x_1 \neq x_2$, where f is a Lebesgue density. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a univariate distribution with the Lebesgue density f. The kernel density estimator for f is given by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

where K is a kernel function and h a bandwidth. Then $f(x_1) - f(x_2)$ is estimated by

$$\hat{f}_{h_1}(x_1) - \hat{f}_{h_2}(x_2). \tag{2}$$

where h_1 and h_2 are bandwidths to estimate the density f at x_1 and x_2, respectively. We use the same kernel function K to estimate both $\hat{f}_{h_1}(x_1)$ and $\hat{f}_{h_2}(x_2)$ for simplicity. This simplification might be innocuous since it is well-known that the choice of kernel functions has little effect on efficiency. See, e.g. Silverman (1986, Section 3.3) for more discussion on the choice of a kernel function. It is straightforward exercise to generalize the results provided in this section to the case with multiple kernel functions if it is necessary. In fact, it is generalized to employ two different kernel functions when we consider the difference of nonparametric regression estimators at a boundary point since it becomes imperative for some applications such as the RDD estimator. This point will become clear in the next section after introducing precise definitions. In contrast to the utilization of a single kernel function, we consider the multiple bandwidths, h_1 and h_2 since there is no simple justification to use a single bandwidth. In fact it would be natural to use two different bandwidth because generally the optimal bandwidths at two points differ. Before we move on, we note that the difference of a single function at two points is a parameter of interest here rather than the difference of two functions at points. Because of the nature for nonparametric estimation methods, it is straightforward to extend the method proposed in this section to cover the latter case. This issue will be mentioned after we present a main result.

In this paper, we propose a bandwidth selection method based on the mean squared error (MSE). In the standard context of kernel density estimation, numerous methods are proposed to choose a bandwidth. See, e.g. Silverman (1986, Section 3.4) for introduction on these methods. One of the most popular and frequently used method is to choose a bandwidth based on the MSE. The asymptotic approximation for the mean "integrated" square error for entire support of the density is often minimized to choose a bandwidth. However, when interests lie in the values of the density at two points rather than those on the entire support, it is natural to choose bandwidths to minimize the MSE determined at the points of interests. The mean squared error for the difference of the two density estimators is defined by

$$MSE_n(h_1, h_2) = E\left\{\left[\left(\hat{f}_{h_1}(x_1) - \hat{f}_{h_2}(x_2)\right) - (f(x_1) - f(x_2))\right]^2\right\}$$

where the expectation is taken with respect to f. A standard approach to implement the minimization is to obtain the asymptotic approximation of the above MSE (AMSE) ignoring higher order terms and to find the bandwidths to minimize it. To do so, we make the following assumptions. The integral sign $\int$ refers to an integral over the range $(-\infty, \infty)$ unless stated otherwise.

ASSUMPTION 2.1 *1. $K(\cdot) : \mathbb{R} \to \mathbb{R}$ is a symmetric second-order kernel function that is continuous with compact support, i.e. K satisfies the followings:*

$$\int K(u)du = 1, \quad \int uK(u)du = 0, \quad \int u^2K(u)du \neq 0.$$

ASSUMPTION 2.2 *The density f is bounded and twice continuously differentiable in a neighborhood of x_1 and x_2. Also assume $f(x_1) > 0$ and $f(x_2) > 0$.*

ASSUMPTION 2.3 *The positive sequence of bandwidths is such that $h \to 0$ and $nh \to \infty$ as $n \to \infty$.*

Assumptions 2.1, 2.2, and 2.3 are standard in the literature of kernel density estimation. Under Assumptions 2.1, 2.2, and 2.3, the usual calculation shows that

$$MSE_n(h_1, h_2) = \left[\frac{\mu_2}{2}\left\{f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2\right\}\right]^2 + \frac{\nu_0}{n}\left\{\frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2}\right\} + o\left(h_1^4 + h_2^4 + \frac{1}{nh_1} + \frac{1}{nh_2}\right)$$

where

$$\mu_j = \int u^j K(u)du, \quad \nu_j = \int u^j K^2(u)du, \tag{3}$$

and $f^{(r)}(\cdot)$ stand for the rth derivative of $f(\cdot)$. See, e.g., Prakasa Rao (1983, Section 2.1). This suggests that we choose the bandwidths to minimize the following AMSE:

$$AMSE_n(h_1, h_2) = \left\{ \frac{\mu_2}{2} \left[f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2 \right] \right\}^2 + \frac{\nu_0}{n} \left[\frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2} \right]. \tag{4}$$

However, this procedure may fail. To see this, let h_1, $h_2 \in H$ where $H = (0, \infty)$ and consider a case where $f^{(2)}(x_1)f^{(2)}(x_2) > 0$. Now choose $h_2 = [f^{(2)}(x_1)/f^{(2)}(x_2)]^{1/2}h_1$. Then we have

$$AMSE_n(h_1, h_2) = \frac{\nu_0}{nh_1} \left\{ f(x_1) + f(x_2) \left[\frac{f^{(2)}(x_2)}{f^{(2)}(x_1)} \right]^{1/2} \right\}.$$

This implies that the bias can be removed completely from the AMSE by choosing the specific ratio of bandwidths and the AMSE can be arbitrarily small by choosing sufficiently large h_1. Restricting H to be a compact set $[c_\ell, c_u]$ does not improve the situation much. In this case, h_1 is always chosen to be equal to c_u irrespective of the data structure.

One reason of this nonstandard behavior is that the AMSE ignores higher order terms. If non-removable higher order terms for bias are present, we might be able to hope that they would punish choosing large values for bandwidths. In the following, we incorporate the higher-order bias terms into the AMSE when densities are smooth. We make the following assumption.

ASSUMPTION 2.4 *The density f is bounded and four times continuously differentiable in a neighborhood of x_1 and x_2. Also assume $f(x_1) > 0$ and $f(x_2) > 0$.*

In the literature of kernel density estimation, it is common to employ higher-order kernel functions when the density is four times differentiable since it is known to deliver bias reduction. See, e.g., Silverman (1986, Section 3.6). However, we have two reasons to maintain Assumption 2.1, i.e., utilization of the second order kernel functions. First, as it will be shown, we can achieve the same bias reduction without employing higher order kernel functions. Second, we end up with the exactly same problem of taking large bandwidths otherwise even if we take higher order bias terms into account. For example, the same issue arises when we remove the first-order bias term by higher order kernel functions as we may be able to cancel the second-order bias term by an appropriate choice of the bandwidths. The next lemma shows the asymptotic property of the MSE under the smoothness condition.

LEMMA 2.1 *Suppose Assumptions 2.1, 2.3 and 2.4 hold. Then it follows that*

$$\begin{aligned} MSE_n(h_1, h_2) = &\left\{ \frac{\mu_2}{2} \left[f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2 \right] + \frac{\mu_4}{4!} \left[f^{(4)}(x_1)h_1^4 - f^{(4)}(x_2)h_2^4 \right] \right\}^2 \\ &+ \frac{\nu_0}{n} \left[\frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2} \right] + o\left(h_1^8 + h_2^8 + \frac{1}{nh_1} + \frac{1}{nh_2} \right). \end{aligned} \tag{5}$$

This result suggests to consider the following AMSE:

$$\begin{aligned} AMSE_n(h_1, h_2) = &\left[\frac{\mu_2}{2} \left\{ f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2 \right\} + \frac{\mu_4}{4!} \left\{ f^{(4)}(x_1)h_1^4 - f^{(4)}(x_2)h_2^4 \right\} \right]^2 \\ &+ \frac{\nu_0}{n} \left\{ \frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2} \right\} \end{aligned} \tag{6}$$

We call this the AMSE with the second-order bias term. In the following, we consider the problem of bandwidths selection to minimize the AMSE with the second-order bias term.

2.2 Asymptotically First-Order Optimal Bandwidth

We saw that the optimal bandwidth that minimizes the AMSE does not exist when the sign of the product of the second derivatives is positive since the AMSE can be made arbitrarily small. Here we show that a similar phenomenon arises even when we consider the AMSE with the second-order bias term. Hence we propose the modified version of the AMSE (MMSE) to introduce a non-vanishing trade-off between bias and variance. The bandwidths chosen to minimize the MMSE are called "the asymptotically first-order optimal bandwidths."

First, we focus on the bias term in the AMSE with the second-order bias term. We show that the order of the bias term can be made $O(h_1^{2k})$ by choosing $h_2^2 = C(h_1, k)h_1^2$ and $C(h_1, k) = C_0 + C_1h_1^2 + C_2h_1^4 + C_3h_1^6 + \ldots + C_kh_1^{2k}$ for some constants C_0, C_1, …, C_k when the sign of the product of the second derivatives are positive. Given that bandwidths are necessarily positive, we must have $C_0 > 0$ though we allow C_1, C_2, …, C_k to be negative. We note that the following discussion might not hold for small n since positivity of bandwidths is possibly violated for small n but must hold eventually for sufficiently large n.

To gain insight, consider choosing $C(h_1, 1) = C_0 + C_1h_1^2$ where $C_0 = f^{(2)}(x_1)/f^{(2)}(x_2)$. In this case we have for the first- and the second-order bias terms

$$\begin{aligned}
&\frac{\mu_2}{2}\left[f^{(2)}(x_1) - C(h_1,1)f^{(2)}(x_2)\right]h_1^2 + \frac{\mu_4}{4!}\left[f^{(4)}(x_1) - C(h_1,1)^2 f^{(4)}(x_2)\right]h_1^4 \\
&\quad = -\frac{\mu_2}{2}C_1 f^{(2)}(x_2)h_1^4 + \frac{\mu_4}{4!}\left[f^{(4)}(x_1) - \left(C_0 + C_1h_1^2\right)^2 f^{(4)}(x_2)\right]h_1^4 \\
&\quad = \left\{-\frac{\mu_2}{2}C_1 f^{(2)}(x_2) + \frac{\mu_4}{4!}\left[f^{(4)}(x_1) - C_0^2 f^{(4)}(x_2)\right]\right\}h_1^4 - \frac{\mu_4}{12}C_0C_1 f^{(4)}(x_2)h_1^6 - \frac{\mu_4}{4!}C_1^2 f^{(4)}(x_2)h_1^8 \\
&\quad = \left\{-\frac{\mu_2}{2}C_1 f^{(2)}(x_2) + \frac{\mu_4}{4!}\left[f^{(4)}(x_1) - C_0^2 f^{(4)}(x_2)\right]\right\}h_1^4 + O(h_1^6).
\end{aligned}$$

By choosing $C_1 = \mu_4\left[f^{(4)}(x_1) - C_0^2 f^{(4)}(x_2)\right] / \left[12\mu_2 f^{(2)}(x_2)\right]$ the order of bias can be made $O(h_1^6)$.

Next consider $C(h_1, 2) = C_0 + C_1h_1^2 + C_2h_1^4$ where C_0 and C_1 are determined as above. In this case

$$\begin{aligned}
&\frac{\mu_2}{2}\left[f^{(2)}(x_1) - C(h_1,2)f^{(2)}(x_2)\right]h_1^2 + \frac{\mu_4}{4!}\left[f^{(4)}(x_1) - C(h_1,2)^2 f^{(4)}(x_2)\right]h_1^4 \\
&\quad = -\frac{\mu_2}{2}C_1 f^{(2)}(x_2)h_4 - \frac{\mu_2}{2}C_2 f^{(2)}(x_2)h_1^6 + \frac{\mu_4}{4!}\left[f^{(4)}(x_1) - \left(C_0 + C_1h_1^2 + C_2h_1^4\right)^2 f^{(4)}(x_2)\right]h_1^4 \\
&\quad = -\frac{\mu_2}{2}C_1 f^{(2)}(x_2)h_4 - \frac{\mu_2}{2}C_2 f^{(2)}(x_2)h_1^6 \\
&\qquad + \frac{\mu_4}{4!}\left\{f^{(4)}(x_1) - f^{(4)}(x_2)\left[C_0^2 + 2C_0C_1h_1^2 + \left(C_1^2 + 2C_0C_2\right)h_1^4 + 2C_1C_2h_1^6 + C_2^2h_1^8\right]\right\}h_1^4 \\
&\quad = -\left\{\frac{\mu_2}{2}C_2 f^{(2)}(x_2) + \frac{\mu_4}{12}C_0C_1 f^{(4)}(x_2)\right\}h_1^6 \\
&\qquad - \frac{\mu_4}{4!}\left(C_1^2 + 2C_0C_2\right)f^{(4)}(x_2)h_1^8 - \frac{\mu_4}{12}C_1C_2 f^{(4)}(x_2)h_1^{10} - \frac{\mu_4}{4!}C_2^2 f^{(4)}(x_2)h_1^{12} \\
&\quad = -\left\{\frac{\mu_2}{2}C_2 f^{(2)}(x_2) + \frac{\mu_4}{12}C_0C_1 f^{(4)}(x_2)\right\}h_1^6 + O(h_1^8)
\end{aligned}$$

so that by choosing $C_2 = -\mu_4 C_0 C_1 f^{(4)}(x_2)/[6\mu_2 f^{(2)}(x_2)]$, the order of bias term can be made $O(h_1^8)$. The same argument can be formulated analogously to the higher order. This implies that the order of bias can be made arbitrarily small and the bandwidths can be chosen to converge arbitrarily slowly to 0 in a polynomial rate.

This analysis indicates that introducing higher-order bias terms does not help to avoid the trade-off disappearing and it is inevitable to make some modification on the AMSE to introduce the penalty. This finding also points out necessity for alternative optimality criterion. First, we propose a modified version of the AMSE, abbreviated as "MMSE," to bring in a nonvanishing trade-off

between bias and variance. The MMSE is defined by

$$MMSE_n(h_1,h_2) = \left\{\frac{\mu_2}{2}\left[f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2\right]\right\}^2 + \left\{\frac{\mu_4}{4!}\left[f^{(4)}(x_1)h_1^4 - f^{(4)}(x_2)h_2^4\right]\right\}^2 + \frac{\nu_0}{n}\left[\frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2}\right]. \quad (7)$$

The difference between the MMSE in (7) and the AMSE with the second order bias terms in (**??**) is that the bias component is represented as the sum of the squared first- and the squared second-order bias terms. A key characteristic of the MMSE is that the first- and the second order bias terms do not disappear simultaneously even when the sign of the product of the second derivatives are positive unless the second derivatives and the fourth derivatives have a particular relationship $f^{(2)}(x_2)^2 f^{(4)}(x_1) = f^{(2)}(x_1)^2 f^{(4)}(x_2)$. Thus either term naturally works as the penalty for large bandwidths in each situation.

Next, we propose an alternative optimality criterion called the "the asymptotically first-order optimality" and we see that the bandwidths that minimizes the MMSE is asymptotically equivalent to the the asymptotically first-order optimal bandwidths. They are obtained by the following procedure:

DEFINITION 2.1 *(Asymptotically First-Order Optimal Bandwidths for Difference of Densities)*

(i) *When $f^{(2)}(x_1)f^{(2)}(x_2) < 0$, choose the bandwidths that minimize*

$$AMSE_{1n}(h_1,h_2) \stackrel{def}{=} \left\{\frac{\mu_2}{2}\left[f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2\right]\right\}^2 + \frac{\nu_0}{n}\left[\frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2}\right]. \quad (8)$$

(ii) *When $f^{(2)}(x_1)f^{(2)}(x_2) > 0$ and $f^{(2)}(x_2)^2 f^{(4)}(x_1) \neq f^{(2)}(x_1)^2 f^{(4)}(x_2)$, choose the bandwidths that minimize*

$$AMSE_{2n}(h_1,h_2) \stackrel{def}{=} \left\{\frac{\mu_4}{4!}\left[f^{(4)}(x_1)h_1^4 - f^{(4)}(x_2)h_2^4\right]\right\}^2 + \frac{\nu_0}{n}\left[\frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2}\right] \quad (9)$$

subject to the restriction $f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2 = 0$

Before we discuss the idea of the first-order optimal bandwidths, we show that the problems minimizing $AMSE_{1n}(h_1,h_2)$ in (8) and $AMSE_{2n}(h_1,h_2)$ in (9) are well-defined.

LEMMA 2.2 *Suppose that the same conditions as Lemma 2.1 hold for each case. Then it follows*

(i) *when $f^{(2)}(x_1)f^{(2)}(x_2) < 0$, $AMSE_{1n}(h_1,h_2)$ has a unique minimum with respect to h_1 and h_2 and unique minimizers, denoted by h_1^* and h_2^*, are given by*

$$h_1^* = \theta^* n^{-1/5}, \quad h_2^* = \lambda^* h_1^* \quad (10)$$

where

$$\theta^* = \left\{\frac{\nu_0 f(x_1)}{\mu_2^2 f^{(2)}(x_1)\left[f^{(2)}(x_1) - {\lambda^*}^2 f^{(2)}(x_2)\right]}\right\}^{1/5}, \quad \lambda^* = \left[-\frac{f(x_2)f^{(2)}(x_1)}{f(x_1)f^{(2)}(x_2)}\right]^{1/3}$$

(ii) *when $f^{(2)}(x_1)f^{(2)}(x_2) > 0$ and $f^{(2)}(x_2)^2 f^{(4)}(x_1) \neq f^{(2)}(x_1)^2 f^{(4)}(x_2)$, $AMSE_{2n}(h_1,h_2)$ subject to the restriction $f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2 = 0$ has a unique minimum with respect to h_1 and h_2 and unique minimizers, denoted by h_1^{**} and h_2^{**}, are given by*

$$h_1^{**} = \theta^{**} n^{-1/9}, \quad h_2^{**} = \lambda^{**} h_1^{**} \quad (11)$$

where

$$\theta^{**} = \left\{\frac{72\nu_0[f(x_1) + f(x_2)/\lambda^{**}]}{\mu_4^2\left[f^{(4)}(x_1) - {\lambda^{**}}^4 f^{(4)}(x_2)\right]^2}\right\}^{1/9}, \quad \lambda^{**} = \left[\frac{f^{(2)}(x_1)}{f^{(2)}(x_2)}\right]^{1/2}$$

Lemma 2.2 shows that the minimization problems that characterize the asymptotically first-order optimal bandwidths are well-defined and the resulting AMSEs are

$$AMSE_{1n}(h_1^*, h_2^*) = n^{-4/5}\left\{\left(\frac{\mu_2\theta^{*2}}{2}\right)^2\left[f^{(2)}(x_1) - \lambda^{*2}f^{(2)}(x_2)\right]^2 + \frac{\nu_0}{\theta^*}\left[f(x_1) + \frac{f(x_2)}{\lambda^*}\right]\right\}, \quad (12)$$

$$AMSE_{2n}(h_1^{**}, h_2^{**}) = n^{-8/9}\left\{\left(\frac{\mu_4\theta^{**4}}{4!}\right)^2\left[f^{(4)}(x_1) - \lambda^{**4}f^{(4)}(x_2)\right]^2 + \frac{\nu_0}{\theta^{**}}\left[f(x_1) + \frac{f(x_2)}{\lambda^{**}}\right]\right\}. \quad (13)$$

The fact that two bandwidths are of the same order in each case is a consequence of our assumption that they are both at least four times continuously differentiable Again, it is interesting to note that the order of the optimal bandwidths exhibits dichotomous behavior depending on the sign of the product of the second derivatives.

Now we discuss the idea behind the asymptotically first-order optimal bandwidths. First, we consider a case where $f^{(2)}(x_1)f^{(2)}(x_2) < 0$. Observe that the $AMSE_{1n}$ defined in (8) is the same as the standard AMSE without the second-order bias term considered in (4). We explain why this is reasonable. In the AMSE with the second-order bias term, the first-order bias term always dominates the second-order term in the order of n unless it disappears. Since it cannot disappear when $f^{(2)}(x_1)f^{(2)}(x_2) < 0$, the second-order bias term can be ignored asymptotically, leading to focusing on the $AMSE_{1n}$. This case is similar to the existing bandwidths selection methods of either choosing two bandwidths separately or selecting a single bandwidth to estimate two functions, in the sense that the order of the bandwidths are $n^{-1/5}$ although the constants in front of $n^{-1/5}$ are different to reflect the simultaneous selection. Second, we take a look at an atypical case where $f^{(2)}(x_1)f^{(2)}(x_2) > 0$. The $AMSE_{2n}$ is a version of the AMSE with the second-order bias term that does not contain the first-order bias term. The definition of the asymptotically first-order optimal bandwidths for this case says that the relationship of the bandwidths are chosen to set the first-order bias term to be exactly zero in the AMSE with the second-order bias term and then we select the bandwidths that minimize the remaining part subject to the relationship. This delivers the bias reduction without increasing variance and explain why we do not have to employ higher-order kernel functions even for the cases where the fourth derivative of $f(\cdot)$ exists. Advantages not to employ higher-order kernel functions are that we do not have to worry about having negative values for density estimates and increased variability for density estimates.

We call the proposed bandwidths the asymptotically first-order optimal bandwidths since the optimal value of the square of the first-order bias term is the same in each situation as that based on the bandwidths minimizing the AMSE with the second-order bias term. The discussion given at the beginning of this subsection shows the possibility of further bias reduction through the appropriate choice of the bandwidths rather than setting the first-order bias term to be exactly zero. This is obviously more desirable and leads to the asymptotically higher-order optimal bandwidth theoretically if we knew the sign of the product of the second derivatives. In reality, however, not only the sign of the product but also other quantities that depend on the underlying density function are unknown and the feasible bandwidth choice procedure must be developed. In doing so, we focus on the asymptotically first-order optimal bandwidths for two reasons. First, the first-order optimal bandwidths dominates the existing methods theoretically. Second, we can establish a feasible automatic bandwidth selection method based on the idea of the asymptotically first-order optimal bandwidths without knowledge on the underlying densities. This point will be elaborated in the next subsection.

After the discussion, the relationship between the asymptotically first-order optimal bandwidths and the bandwidths that minimizes the MMSE would be clear. To be concrete, note that the

following relationships hold for the $MMSE$, the $AMSE_{1n}$ and the $AMSE_{2n}$:

$$MMSE_n(h_1^*, h_2^*) = AMSE_{1n}(h_1^*, h_2^*) + \left[\frac{\mu_4 \theta^{*4}}{4!}\left\{f^{(4)}(x_1) - \lambda^{*4} f^{(4)}(x_2)\right\}\right]^2 n^{-8/5}, \tag{14}$$

$$MMSE_n(h_1^{**}, h_2^{**}) = AMSE_{2n}(h_1^{**}, h_2^{**}) \tag{15}$$

by the definitions of $MMSE_n$ in (7), $AMSE_{1n}$ in (8) and $AMSE_{2n}$ in (9). When the sign is positive, the bandwidths that minimizes the MMSE is exactly the same as the asymptotically first-order optimal bandwidths by (14). When the sign is negative, they are asymptotically equivalent since they are different by asymptotically negligible amount by (15).

Before we move on, we give some remarks for several cases that are not covered by the asymptotically first-order optimal bandwidths. Definition 2.1 does not cover a case where the first- and the second-order bias terms vanish simultaneously, i.e. a case where $f^{(2)}(x_2)^2 f^{(4)}(x_1) = f^{(2)}(x_1)^2 f^{(4)}(x_2)$. Definition 2.1 can be generalized to cover the excluded case in a straightforward manner if we are willing to assume the existence of the sixth derivative of f and $f^{(4)}(x_2)^3 f^{(6)}(x_1)^2 \neq f^{(4)}(x_1)^3 f^{(6)}(x_2)^2$. This case correspond to the situation where the first and the second order bias terms can be removed simultaneously by an appropriate choice of bandwidths and the third order bias term works as a penalty. When f is infinitely times continuously differentiable, an excluded case becomes $f^{(2j)}(x_2)^{j+1} f^{(2(j+1))}(x_1)^j = f^{(2j)}(x_1)^{j+1} f^{(2(j+1))}(x_2)^j$ for all integer j. Another excluded case by Definition 2.1 is when $f^{(2)}(x_1) f^{(2)}(x_2) = 0$. However, when both $f^{(2)}(x_1) = 0$ and $f^{(2)}(x_2) = 0$ hold, a simple extension of Lemma 2.2 is possible to deal with the situation as long as we can assume the existence of the sixth order derivatives. The generalization is parallel to Definition 2.1 (i) and (ii) with replacing $f^{(2)}(x_1)$, $f^{(2)}(x_2)$, $f^{(4)}(x_1)$, $f^{(4)}(x_2)$ and other parameters with $f^{(4)}(x_1)$, $f^{(4)}(x_2)$, $f^{(6)}(x_1)$, $f^{(6)}(x_2)$ and corresponding ones. When only one of $f^{(2)}(x_1) = 0$ and $f^{(2)}(x_2) = 0$ is equal to 0, the situations become complicated. An insight to the difficulty will be provided in Section 5.

2.3 Feasible Automatic Bandwidth Choice

Next, we propose a practicable procedure to choose bandwidths that are as good as the asymptotically first-order optimal bandwidth at least in large sample. The approach proposed in the last section to minimize the MMSE is obviously infeasible since it depends on unknown quantities such as $f(\cdot)$, $f^{(2)}(\cdot)$ and $f^{(4)}(\cdot)$. It is also important to note that we do not know which one of the AMSEs given in Definition 2.1 to minimize without the knowledge on the relationships between the second order derivatives of the underlying densities. If we knew the sign for the product of the second order derivatives, we would be able to use a direct plug-in method by estimating values of either the second or the fourth order derivatives and the density at the points of interest. Hence one possibility might be a pretesting method. That is we first test whether the sign is positive or negative, and then we use the plug-in method after we find out which bandwidths to use. The problem of this approach is that we do not have such a test for the sign at hand.

The procedure we propose is a feasible bandwidth selection method that automatically adjusts to each of dichotomous cases. The proposed method for bandwidth selection can be considered as a generalization of the traditional plug-in method. See, e.g., Wand and Jones (1994, Section 3.6). The feasible method we propose is based on the $MMSE$. Let $\hat{f}(\cdot)$, $\hat{f}^{(2)}(\cdot)$ and $\hat{f}^{(4)}(\cdot)$ be some consistent estimators for $f(\cdot)$, $f^{(2)}(\cdot)$ and $f^{(4)}(\cdot)$. Consider its plug-in version denoted by $\widehat{MMSE}$:

$$\widehat{MMSE}_n(h_1, h_2) = \left\{\frac{\mu_2}{2}\left[\hat{f}^{(2)}(x_1)h_1^2 - \hat{f}^{(2)}(x_2)h_2^2\right]\right\}^2 + \left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)h_1^4 - \hat{f}^{(4)}(x_2)h_2^4\right]\right\}^2 + \frac{\nu_0}{n}\left[\frac{\hat{f}(x_1)}{h_1} + \frac{\hat{f}(x_2)}{h_2}\right]. \tag{16}$$

As we noted above, the most important advantage of the MMSE over the AMSE with the second-bias term is that the MMSE has the nonvanishing trade-off between bias and variance. There also

exists two important byproducts for focusing on the MMSE. First it restricts the minimizers of the MMSE to be of order either $n^{-1/5}$ or $n^{-1/9}$. This simplifies the theoretical developments of the feasible bandwidth selection procedure. Second, observe that we do not need prior knowledge on the relationships between the second derivatives to construct the MMSE. Theorem 2.1 shows that the resulting bandwidths adjust to each situation automatically. Let $(\hat{h}_1, \hat{h}_2)$ be a combination of bandwidths that minimizes the MMSE. In the next theorem, we shall show that $(\hat{h}_1, \hat{h}_2)$ is asymptotically as good as the first-order optimal bandwidths in the sense of Hall (1983). From Lemma 2.2, we know that the asymptotically first-optimal bandwidths are of the order either $n^{-1/5}$ or $n^{-1/9}$ and the MMSE is constructed to yield bandwidths of order $n^{-1/5}$ or $n^{-1/9}$. For this reason we confine attention to the bandwidths (h_1, h_2) in the interval H_n where

$$H_n = H_{1n} \cup H_{2n}$$

where

$$H_{1n} = \left[\delta_{11} n^{-1/5}, \delta_{12} n^{-1/5}\right] \times \left[\delta_{11} n^{-1/5}, \delta_{12} n^{-1/5}\right],$$
$$H_{2n} = \left[\delta_{21} n^{-1/9}, \delta_{22} n^{-1/9}\right] \times \left[\delta_{21} n^{-1/9}, \delta_{22} n^{-1/9}\right]$$

for arbitrarily small δ_{11} and δ_{21}, and large δ_{12} and δ_{22}. See Hall (1983) for a similar discussion in the context of cross validation.

THEOREM 2.1 *Suppose that the same conditions as Lemma 2.1 hold for each case. Assume further that, for $j = 1, 2$, the initial bandwidths and the kernel functions used to define $\hat{f}(x_j)$ and $\hat{f}^{(4)}(x_j)$ satisfy $\hat{f}(x_j) \to f(x_j)$ and $\hat{f}^{(4)}(x_j) \to f^{(4)}(x_j)$ in probability, respectively. The initial bandwidth and the initial kernel function to estimate $\hat{f}^{(2)}(x_j)$ for $j = 1, 2$ satisfy $\hat{f}^{(2)}(x_j) = f^{(2)}(x_j) + o_p(n^{-2/9})$. Let $\hat{\mathbf{h}}$ be a combination of bandwidths that minimizes the modified AMSE defined in (16). Then*

(i) *when $f^{(2)}(x_1) f^{(2)}(x_2) < 0$*

$$\frac{\hat{h}_1}{h_1^*} \to 1, \quad \textit{and} \quad \frac{\hat{h}_2}{h_2^*} \to 1$$

in probability and when $f^{(2)}(x_1) f^{(2)}(x_2) > 0$ and $f^{(2)}(x_2)^2 f^{(4)}(x_1) \neq f^{(2)}(x_1)^2 f^{(4)}(x_2)$

$$\frac{\hat{h}_1}{h_1^{**}} \to 1, \quad \textit{and} \quad \frac{\hat{h}_2}{h_2^{**}} \to 1$$

in probability.

(ii) *when $f^{(2)}(x_1) f^{(2)}(x_2) < 0$*

$$\frac{\widehat{MMSE}_n(\hat{h}_1, \hat{h}_2)}{MSE_n(h_1^*, h_2^*)} \to 1$$

in probability and when $f^{(2)}(x_1) f^{(2)}(x_2) > 0$ and $f^{(2)}(x_2)^2 f^{(4)}(x_1) \neq f^{(2)}(x_1)^2 f^{(4)}(x_2)$

$$\frac{\widehat{MMSE}_n(\hat{h}_1, \hat{h}_2)}{MSE_n(h_1^{**}, h_2^{**})} \to 1$$

in probability.

We outline the proof of Theorem 2.1. To do so, denote $\mathbf{h} = (h_1, h_2)$, $\hat{\mathbf{h}} = (\hat{h}_1, \hat{h}_2)$, $\mathbf{h}^* = (h_1^*, h_2^*)$, $\mathbf{h}^{**} = (h_1^{**}, h_2^{**})$, $\hat{\mathbf{h}}/\mathbf{h}^* = (\hat{h}_1/h_1^*, \hat{h}_2/h_2^*)$ and $\hat{\mathbf{h}}/\mathbf{h}^{**} = (\hat{h}_1/h_1^{**}, \hat{h}_2/h_2^{**})$. For $\varepsilon > 0$, denote an ε-neighborhood of 1 by $N_1(\varepsilon, 1)$ and $N_1(\varepsilon) = N_1(\varepsilon, 1) \times N_1(\varepsilon, 1)$. First, we consider the case where $f^{(2)}(x_1) f^{(2)}(x_2) < 0$. We show

$$\Pr\left\{\inf_{\mathbf{h} \in H_{1n}, \mathbf{h}/\mathbf{h}^* \notin N_1(\varepsilon)} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^*) \leq 0\right\} \to 0 \tag{17}$$

and

$$\Pr\left\{\inf_{\mathbf{h}\in H_{2n}} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^*) \leq 0\right\} \to 0 \tag{18}$$

as n goes to infinity. Remember that Lemma 2.2 shows that the order of the asymptotically first-order optimal bandwidths is $n^{-1/5}$ when $f^{(2)}(x_1)f^{(2)}(x_2) < 0$. The claim (17) means that minimum cannot be attained outside of the neighborhood of $\mathbf{h}^*$ asymptotically even if the correct order for the bandwidths is taken. This is similar to what usually arise in the context of extremum estimators and can be shown by demonstrating the uniform convergence of $n^{4/5}\widehat{MMSE}(\mathbf{h})$ to $n^{4/5}AMSE_{1n}(\mathbf{h})$ in probability. The claim (18) implies that minimum cannot be reached asymptotically for any values of $\mathbf{h}$ when the rate is specified incorrectly. This is true since the slower rate of convergence in bias components prevents from achieving the minimum when the incorrect order for the bandwidths is taken.

Next, consider the case where $f^{(2)}(x_1)f^{(2)}(x_2) > 0$. As in the previous case, we show

$$\Pr\left\{\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon)} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^{**}) \leq 0\right\} \to 0 \tag{19}$$

and

$$\Pr\left\{\inf_{\mathbf{h}\in H_{1n}} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^{**}) \leq 0\right\} \to 0 \tag{20}$$

as n goes to infinity. We note that Lemma 2.2 reveals that the order for the first-order optimal bandwidths is $n^{-1/9}$ for this case. Claim (20) is similar to the claim (18) and is interpreted as that the minimum cannot be attained asymptotically for any values of $\mathbf{h}$ when the rate is specified incorrectly. This comes, however, from the fact that the variance components converges at a slower rate when the incorrect order is specified rather than the bias components. Second, we discuss why the claim (19) holds. Apart from similarity between (18) and (20), the mechanism behind (19) is substantially different from that behind (17). The claim (19) is decomposed further into two, i.e.,

$$\Pr\left\{\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon),h_1/h_2=h_1^{**}/h_2^{**}} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^{**}) \leq 0\right\} \to 0. \tag{21}$$

and

$$\Pr\left\{\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon),h_1/h_2\neq h_1^{**}/h_2^{**}} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^{**}) \leq 0\right\} \to 0, \tag{22}$$

Recall that the first-order bias term can be removed asymptotically by choosing the ratio between h_1 and h_2 equal to that between h_1^{**} and h_2^{**} without h_1 and h_2 being equal to h_1^{**} and h_2^{**}. Then the claims (21) means that minimum cannot be attained outside of the neighborhood of $\mathbf{h}^*$ asymptotically even if the correct order for the bandwidths is taken and even if the first-order bias term is removed. This can be shown in the same manner as (18). On the other hand, the claim (22) means that minimum cannot be attained outside of the neighborhood of $\mathbf{h}^*$ asymptotically even if the correct order for the bandwidths is taken when the first-order bias term is not removed. It turns out that $n^{8/9}\widehat{MMSE}(\mathbf{h})$ does not uniformly converge to $n^{8/9}AMSE_{2n}(\mathbf{h})$ in probability although $n^{8/9}\widehat{MMSE}(\mathbf{h}^{**})$ does converge to $n^{8/9}AMSE_{2n}(\mathbf{h}^{**})$ in probability. In fact, we show that $n^{4/9}\widehat{MMSE}(\mathbf{h})$ is $O_p(1)$ when $\mathbf{h} \neq \mathbf{h}^{**}$. This indicates that when the order of bandwidths is specified correctly, the penalty to choose $\mathbf{h}$ other than $\mathbf{h}^{**}$ becomes huge and this works as driving force to make the claim (22) true. We provide a detailed proof in the Appendix.

We can propose another feasible bandwidths that might be more intuitive. The bandwidths are based on the expression of the asymptotically first-order bandwidths given in Lemma 2.2. The proposed bandwidths are more direct version of the plug-in bandwidths defined by

$$\begin{aligned}
\tilde{h}_1 &= \hat{\theta}_1 n^{-1/5} 1_{\{\hat{f}^{(2)}(x_1)\hat{f}^{(2)}(x_2)<0\}} + \hat{\theta}_2 n^{-1/9}\left(1 - 1_{\{\hat{f}^{(2)}(x_1)\hat{f}^{(2)}(x_2)<0\}}\right), \\
\tilde{h}_2 &= \hat{\theta}_1\hat{\lambda}_1 n^{-1/5} 1_{\{\hat{f}^{(2)}(x_1)\hat{f}^{(2)}(x_2)<0\}} + \hat{\theta}_2\hat{\lambda}_2 n^{-1/9}\left(1 - 1_{\{\hat{f}^{(2)}(x_1)\hat{f}^{(2)}(x_2)<0\}}\right)
\end{aligned}$$

where

$$\hat{\theta}_1 = \left\{ \frac{\nu_0 \hat{f}(x_1)}{\mu_2^2 \hat{f}^{(2)}(x_1) \left[\hat{f}^{(2)}(x_1) - \hat{\lambda}^2 \hat{f}^{(2)}(x_2) \right]} \right\}^{1/5}, \quad \hat{\lambda}_1 = \left[-\frac{\hat{f}(x_2)\hat{f}^{(2)}(x_1)}{\hat{f}(x_1)\hat{f}^{(2)}(x_2)} \right]^{1/3},$$

$$\hat{\theta}_2 = \left\{ \frac{72\nu_0[\hat{f}(x_1) + \hat{f}(x_2)/\hat{\lambda}]}{\mu_4^2 \left[\hat{f}^{(4)}(x_1) - \hat{\lambda}^4 \hat{f}^{(4)}(x_2) \right]^2} \right\}^{1/9}, \quad \text{and} \quad \hat{\lambda}_2 = \left[\frac{\hat{f}^{(2)}(x_1)}{\hat{f}^{(2)}(x_2)} \right]^{1/2}.$$

These bandwidths switches from one to another depending on the estimated sign of of the product of the second derivatives. We call these the direct plug-in bandwidths. It can be also shown that the direct plug-in bandwidths are asymptotically as good as the asymptotically first-order bandwidths. The following result can be derived in the similar manner as Theorem 2.1.

COROLLARY 2.1 *Suppose that the same conditions as Theorem 2.1 hold for each case. Then the results that hold for* $(\hat{h}_1, \hat{h}_2)$ *also holds for* $(\tilde{h}_1, \tilde{h}_2)$

The assumptions of Theorem 2.1 call for pilot estimates of $f(x_j)$, $f^{(2)}(x_j)$, and $f^{(4)}(x_j)$ for $j = 1, 2$ and we discuss estimators that satisfy the requirements. For estimators of $f(x_j)$ and $f^{(4)}(x_j)$, consistency suffices. The standard kernel density estimator as in (1) can be used for the pilot estimate of $f(x_j)$. Epanechnikov kernel function $K(x) = 3/4 \cdot (1 - x^2) 1_{\{|x|<1\}}$ with normal scale bandwidths would be a reasonable choice. See Wand and Jones (1994) for the choice of kernel functions and bandwidths to obtain pilot estimates. For $f^{(2)}(x_j)$, and $f^{(4)}(x_j)$, we can use the kernel density derivative estimator

$$\hat{f}^{(r)}(x) = \frac{1}{nh^{r+1}} \sum_{j=1}^{n} K^{(r)} \left(\frac{x - X_j}{h} \right)$$

where $K^{(r)}(\cdot)$ is the r-th derivative of the kernel function K. For the kernel function K, we can us any kernel functions with sufficient differentiability. For $f^{(2)}(x_j)$, the assumption of Theorem 2.1 requires $\hat{f}^{(2)}(x_j) - f^{(2)}(x_j) = o_p(n^{-2/9})$ but the kernel density derivative estimators with second-order kernel functions do not satisfy this requirement since we have $\hat{f}^{(2)}(x_j) - f^{(2)}(x_j) = O_p(n^{-2/9})$ when the optimal bandwidths of the order $n^{-1/9}$ is used. One possible solution is to use the higher-order kernel function with sufficient differentiability. For example, the higher-order kernel function derived from the normal density kernel function is given by

$$K^{(2)}(x) = \frac{1}{2}(-x^4 + 8x^2 - 5)\phi(x)$$

where $\phi(x)$ is the normal density function. The estimators obtained by employing this kernel function with the bandwidths of the order $n^{-1/9}$ satisfy the requirement. For both $\hat{f}^{(2)}(x_j)$, and $\hat{f}^{(4)}(x_j)$, the normal scale bandwidth can be used again. See Wand and Jones (1994) for basic treatment on density derivative estimation.

We also note that as it is clear from the proof of Theorem 2.1, the result provided in Theorem 2.1 when $f^{(2)}(x_1)f^{(2)}(x_2) < 0$ can be proved under Assumption 2.2 rather than Assumption 2.4. That is, twice continuous differentiability of $f(\cdot)$ is sufficient to prove the consistency of the proposed bandwidths when $f^{(2)}(x_1)f^{(2)}(x_2) < 0$. We assume Assumption 2.4 that include four times continuous differentiability since we proceed without the prior knowledge on the relationship between the second order derivatives and since it is a sufficient condition for both cases.

Throughout the discussion in this section, we consider the difference of kernel density estimators for a "single" density at two distinct points. A straightforward generalization shows that we can apply the proposed method to select bandwidth for the difference of kernel density estimators of two distinct densities, f and g, at two points, x and y

$$f(x) - g(y)$$

based on the two random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. It would be worthwhile noting that we can still exploit the relationship between the second order derivatives of the densities even if two random samples are independent.

3 Nonparametric Estimation for Differences of Regression Functions

In this section we extend the approach proposed in the previous section to the nonparametric estimation of the difference of regression functions. The nonparametric regression estimators we consider are local linear estimators since, as shown in Fan (1992), they have desirable properties in terms of asymptotic efficiency, design adaptive property as well as boundary effects. Let Y_i be a scalar random variable and X_i is a scalar variable having common density $f(\cdot)$. Throughout this section, we assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed observations. We use the notation $\sigma^2(x)$ for the conditional variance of Y given $X = x$. Suppose we are interested in estimating

$$m(x) = E(Y_i|X_i = x)$$

at a scalar $x \in \text{supp}(f)$ where $\text{supp}(f)$ stands for the support of f and $\text{supp}(f) \subset R$. The local linear regression estimator for the conditional mean function is the solution for α to the following problem:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{Y_i - \alpha - \beta(X_i - x)\}^2 K\left(\frac{X_i - x}{h}\right), \tag{23}$$

where $K(\cdot)$ is a kernel function and h is a bandwidth. The solution to minimize (23) can be expressed as

$$\begin{bmatrix} \hat{\alpha}_h(x) \\ \hat{\beta}_h(x) \end{bmatrix} = \left(X(x)'W(x)X(x)\right)^{-1} X(x)'W(x)Y \tag{24}$$

where

$$X(x) = \begin{bmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix},$$

$Y = (Y_1, \ldots, Y_n)'$, $X = (X_1, \ldots, X_n)'$, $W(x) = \text{diag}(K_h(X_i - x))$ and $K_h(\cdot) = K(\cdot/h)/h$. The local least squares estimator of $m(x)$ can also be written as

$$\hat{\alpha}_h(x) = e_1' \left(X(x)'W(x)X(x)\right)^{-1} X(x)'W(x)Y$$

where e_1 is the 2×1 vector having 1 in the first entry and the other entry 0. Then the difference of the regression functions evaluated at two distinct points x_1 and x_2, $m(x_1) - m(x_2)$ is estimated by

$$\hat{\alpha}_{h_1}(x_1) - \hat{\alpha}_{h_2}(x_2).$$

We first consider a case where x_1 and x_2 are interior points of the support of f and then a case where they are near the boundary. In the standard local linear regression, basic characteristics of bias and variance for interior points are the same as those for boundary points. However, we show that essentially different behaviors arise since we take higher-order terms into consideration as we have done for density estimation.

3.1 Differences of Local Linear Regression Estimators at Interior Points

In this subsection, we proceed under the following assumptions.

ASSUMPTION 3.1 *The density $f(\cdot)$ is a bounded function at points x_1 and x_2 with $f(x_1) > 0$ and $f(x_2) > 0$. It is also twice continuously differentiable in a neighborhood of x_1 and x_2.*

ASSUMPTION 3.2 *$\sigma^2(\cdot)$ are bounded functions at points x_1 and x_2. Also assume $\sigma^2(x_1) > 0$ and $\sigma^2(x_2) > 0$.*

ASSUMPTION 3.3 *The conditional mean function $m(\cdot)$ is bounded and twice continuously differentiable in a neighborhood of x_1 and x_2.*

Let $m^{(j)}(\cdot)$ stand for the j-th derivative of $m(\cdot)$. A straightforward extension of Theorem 1 in Fan (1992) shows under Assumptions 2.1, 2.3, 3.1, 3.2, and 3.3

$$\begin{aligned} MSE_n(h_1, h_2) &= E\left[\{(\hat{\alpha}_{h_1}(x_1) - \hat{\alpha}_{h_2}(x_2)) - (m(x_1) - m(x_2))\}^2 \middle| X \right] \\ &= \left[\frac{\mu_2}{2} \left\{ m^{(2)}(x_1) h_1^2 - m^{(2)}(x_2) h_2^2 \right\} \right]^2 + \frac{\nu_0}{n} \left\{ \frac{\sigma^2(x_1)}{h_1 f(x_1)} + \frac{\sigma^2(x_2)}{h_2 f(x_2)} \right\} \\ &\quad + o\left(h_1^4 + h_2^4 + \frac{1}{nh_1} + \frac{1}{nh_2} \right) \end{aligned}$$

where $X = (X_1, X_2, \ldots, X_n)'$ and μ_j and ν_j are defined in (3). This implies that we have exactly the same problem as before when we try to minimize the AMSE based on this MSE. Hence, as in the case of density estimation, we need to take higher order terms into consideration. The result on the higher order approximation of the MSE is provided in Fan et al. (1996). However, their result is up to an order that disappears when we use a symmetric kernel function. The next lemma that is an analogous result to Lemma 2.1 shows the further generalization on the higher order approximation of the MSE. We proceed under the following assumption:

ASSUMPTION 3.4 *The conditional mean function $m(\cdot)$ is bounded and four times continuously differentiable in a neighborhood of x_1 and x_2.*

LEMMA 3.1 *Suppose Assumptions 2.1, 2.3, 3.1, and 3.4 hold. Then it follows that*

$$\begin{aligned} MSE_n(h_1, h_2) &= \left[\frac{\mu_2}{2} \left\{ m^{(2)}(x_1) h_1^2 - m^{(2)}(x_2) h_2^2 \right\} + \left\{ b(x_1) h_1^4 - b(x_2) h_2^4 \right\} \right]^2 \\ &\quad + \frac{\nu_0}{n} \left\{ \frac{\sigma^2(x_1)}{h_1 f(x_1)} + \frac{\sigma^2(x_2)}{h_2 f(x_2)} \right\} + o\left(h_1^8 + h_2^8 + \frac{1}{nh_1} + \frac{1}{nh_2} \right) \end{aligned} \tag{25}$$

where

$$b(x) = \frac{1}{4} \left\{ \frac{m^{(2)}(x)}{f(x)^2} (\mu_4 - \mu_2) \left(f^{(2)}(x) f(x) - f^{(1)}(x)^2 \right) + \frac{m^{(4)}(x)}{6} \mu_4 \right\}.$$

Based on the MSE provided in Lemma 3.1, the asymptotically first-order optimal bandwidths to estimate the difference of regression functions at two interior points are given in the same manner as Definition 2.1.

DEFINITION 3.1 *(Asymptotically First-Order Optimal Bandwidths for Difference of Regression Functions at Interior Points)*

(i) *When $m^{(2)}(x_1) m^{(2)}(x_2) < 0$, choose the bandwidths that minimize*

$$AMSE_{1n}(h_1, h_2) \stackrel{def}{=} \left\{ \frac{\mu_2}{2} \left[m^{(2)}(x_1) h_1^2 - m^{(2)}(x_2) h_2^2 \right] \right\}^2 + \frac{\nu_0}{n} \left[\frac{\sigma^2(x_1)}{h_1 f(x_1)} + \frac{\sigma^2(x_2)}{h_2 f(x_2)} \right]. \tag{26}$$

(ii) *When* $m^{(2)}(x_1)m^{(2)}(x_2) > 0$ *and* $m^{(2)}(x_2)^2 b(x_1) \neq m^{(2)}(x_1)^2 b(x_2)$, *choose the bandwidths that minimize*

$$AMSE_{2n}(h_1, h_2) \stackrel{def}{=} \left[b(x_1)h_1^4 - b(x_2)h_2^4\right]^2 + \frac{\nu_0}{n}\left[\frac{\sigma^2(x_1)}{h_1 f(x_1)} + \frac{\sigma^2(x_2)}{h_2 f(x_2)}\right] \tag{27}$$

subject to the restriction $m^{(2)}(x_1)h_1^2 - m^{(2)}(x_2)h_2^2 = 0$

We present next Lemma without proof since it is totally parallel to Lemma 2.2.

LEMMA 3.2 *Suppose that the same conditions as Lemma 3.1 hold for each case. Then it follows*

(i) *when* $m^{(2)}(x_1)m^{(2)}(x_2) < 0$, $AMSE_{1n}(h_1, h_2)$ *has a unique minimum with respect to* h_1 *and* h_2 *and unique minimizers, denoted by* h_1^* *and* h_2^*, *are given by*

$$h_1^* = \theta^* n^{-1/5}, \quad h_2^* = \lambda^* h_1^*$$

where

$$\theta^* = \left\{\frac{\nu_0 \sigma^2(x_1)}{\mu_2^2 f(x_1) m^{(2)}(x_1)\left[m^{(2)}(x_1) - \lambda^{*2} m^{(2)}(x_2)\right]}\right\}^{1/5}, \quad \lambda^* = \left[-\frac{\sigma^2(x_2) f(x_1) m^{(2)}(x_1)}{\sigma^2(x_1) f(x_2) m^{(2)}(x_2)}\right]^{1/3}$$

(ii) *when* $m^{(2)}(x_1)m^{(2)}(x_2) > 0$ *and* $m^{(2)}(x_2)^2 b(x_1) \neq m^{(2)}(x_1)^2 b(x_2)$, $AMSE_{2n}(h_1, h_2)$ *subject to the restriction* $m^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2 = 0$ *has a unique minimum with respect to* h_1 *and* h_2 *and unique minimizers, denoted by* h_1^{**} *and* h_2^{**}, *are given by*

$$h_1^{**} = \theta^{**} n^{-1/9}, \quad h_2^{**} = \lambda^{**} h_1^{**}$$

where

$$\theta^{**} = \left\{\frac{\nu_0}{8\left[m^{(4)}(x_1) - \lambda^{**4} m^{(4)}(x_2)\right]^2}\left[\frac{\sigma^2(x_1)}{f(x_1)} + \frac{\sigma^2(x_2)}{\lambda^{**} f(x_2)}\right]\right\}^{1/9}, \quad \lambda^{**} = \left[\frac{m^{(2)}(x_1)}{m^{(2)}(x_2)}\right]^{1/2}$$

Again, the dichotomous behaviors of the first-order optimal bandwidths are present. In proposing the feasible automatic bandwidths, the MMSE for the present context is defined by

$$MMSE_n(h_1, h_2) = \left\{\frac{\mu_2}{2}\left[m^{(2)}(x_1)h_1^2 - m^{(2)}(x_2)h_2^2\right]\right\}^2 + \left[b(x_1)h_1^4 - b(x_2)h_2^4\right]^2 + \frac{\nu_0}{n}\left[\frac{\sigma^2(x_1)}{h_1 f(x_1)} + \frac{\sigma^2(x_2)}{h_2 f(x_2)}\right]$$

and its plug-in version by

$$\widehat{MMSE}_n(h_1, h_2) = \left\{\frac{\mu_2}{2}\left[\hat{m}^{(2)}(x_1)h_1^2 - \hat{m}^{(2)}(x_2)h_2^2\right]\right\}^2 + \left[\hat{b}(x_1)h_1^4 - \hat{b}(x_2)h_2^4\right]^2 + \frac{\nu_0}{n}\left[\frac{\hat{\sigma}^2(x_1)}{h_1 \hat{f}(x_1)} + \frac{\hat{\sigma}^2(x_2)}{h_2 \hat{f}(x_2)}\right] \tag{28}$$

where $\hat{m}^{(2)}(x_j)$, $\hat{b}(x_j)$, $\hat{\sigma}^2(x_j)$, and $\hat{f}(x_j)$ are consistent estimators for $m^{(2)}(x_j)$, $b(x_j)$, $\sigma^2(x_j)$, and $f(x_j)$ for $j = 1, 2$, respectively. Precise conditions for these estimators are provided in the next theorem. Let $(\hat{h}_1, \hat{h}_2)$ be a combination of bandwidths that minimizes the MMSE given in (28). Then the next theorem which is analogous to Theorem 2.1 holds.

THEOREM 3.1 *Suppose that the same conditions as Lemma 3.1 hold for each case. Assume further that, for $j = 1, 2$, the initial bandwidths and the kernel functions used to define $\hat{b}(x_j)$, $\hat{f}(x_j)$ and $\hat{\sigma}^2(x_j)$ satisfy $\hat{b}(x_j) \to b(x_j)$, $\hat{f}(x_j) \to f(x_j)$ and $\hat{\sigma}^2(x_j) \to \sigma^2(x_j)$ in probability, respectively. The initial bandwidth, the initial kernel function and the order of polynomial regression to estimate $\hat{m}^{(2)}(x_j)$ for $j = 1, 2$ satisfy $\hat{m}^{(2)}(x_j) = m^{(2)}(x_j) + o_p(n^{-2/9})$. Let $\hat{\mathbf{h}}$ be a combination of bandwidths that minimizes the modified AMSE defined in (28). Then*

(i) *when $m^{(2)}(x_1)m^{(2)}(x_2) < 0$*

$$\frac{\hat{h}_1}{h_1^*} \to 1, \quad \textit{and} \quad \frac{\hat{h}_2}{h_2^*} \to 1$$

in probability and when $m^{(2)}(x_1)m^{(2)}(x_2) > 0$ and $m^{(2)}(x_2)^2 b(x_1) \neq m^{(2)}(x_1)^2 b(x_2)$

$$\frac{\hat{h}_1}{h_1^{**}} \to 1, \quad \textit{and} \quad \frac{\hat{h}_2}{h_2^{**}} \to 1$$

in probability.

(ii) *when $m^{(2)}(x_1)m^{(2)}(x_2) < 0$*

$$\frac{\widehat{MMSE}_n(\hat{h}_1, \hat{h}_2)}{MSE_n(h_1^*, h_2^*)} \to 1$$

in probability and when $m^{(2)}(x_1)m^{(2)}(x_2) > 0$ and $m^{(2)}(x_2)^2 b(x_1) \neq m^{(2)}(x_1)^2 b(x_2)$

$$\frac{\widehat{MMSE}_n(\hat{h}_1, \hat{h}_2)}{MSE_n(h_1^{**}, h_2^{**})} \to 1$$

in probability.

As in the case for densities we can propose another feasible bandwidths that might be more intuitive. The bandwidths are based on the expression of the asymptotically first-order bandwidths given in Lemma 3.2. The proposed bandwidths are more direct version of the plug-in bandwidths defined by

$$\begin{aligned}
\tilde{h}_1 &= \hat{\theta}_1 n^{-1/5} 1_{\{\hat{m}^{(2)}(x_1)\hat{m}^{(2)}(x_2)<0\}} + \hat{\theta}_2 n^{-1/9} \left(1 - 1_{\{\hat{m}^{(2)}(x_1)\hat{m}^{(2)}(x_2)<0\}}\right), \\
\tilde{h}_2 &= \hat{\theta}_1 \hat{\lambda}_1 n^{-1/5} 1_{\{\hat{m}^{(2)}(x_1)\hat{m}^{(2)}(x_2)<0\}} + \hat{\theta}_2 \hat{\lambda}_2 n^{-1/9} \left(1 - 1_{\{\hat{m}^{(2)}(x_1)\hat{m}^{(2)}(x_2)<0\}}\right)
\end{aligned}$$

where

$$\hat{\theta}_1 = \left\{ \frac{\nu_0 \hat{\sigma}^2(x_1)}{\mu_2^2 \hat{f}(x_1)\hat{m}^{(2)}(x_1)\left[\hat{m}^{(2)}(x_1) - \hat{\lambda}^2 \hat{m}^{(2)}(x_2)\right]} \right\}^{1/5}, \quad \hat{\lambda}_1 = \left[-\frac{\hat{\sigma}^2(x_2)\hat{f}(x_1)\hat{m}^{(2)}(x_1)}{\hat{\sigma}^2(x_1) f(x_2)\hat{m}^{(2)}(x_2)} \right]^{1/3},$$

$$\hat{\theta}_2 = \left\{ \frac{\nu_0}{8\left[\hat{m}^{(4)}(x_1) - \hat{\lambda}_2^4 \hat{m}^{(4)}(x_2)\right]^2} \left[\frac{\hat{\sigma}^2(x_1)}{\hat{f}(x_1)} + \frac{\hat{\sigma}^2(x_2)}{\hat{\lambda}_2 \hat{f}(x_2)} \right] \right\}^{1/9}, \quad \text{and} \quad \hat{\lambda}_2 = \left[\frac{\hat{m}^{(2)}(x_1)}{\hat{m}^{(2)}(x_2)} \right]^{1/2}.$$

These bandwidths switches from one to another depending on the estimated sign of of the product of the second derivatives. It can be also shown that the direct plug-in bandwidths are asymptotically as good as the asymptotically first-order bandwidths. The following result can be derived in the similar manner as Theorem 3.1.

COROLLARY 3.1 *Suppose that the same conditions as Theorem 3.1 hold for each case. Then the results that hold for $(\hat{h}_1, \hat{h}_2)$ also holds for $(\tilde{h}_1, \tilde{h}_2)$*

As Theorem 2.1, Theorem 3.1 requires pilot estimates for $m^{(2)}(x_j)$, $b(x_j)$, $f(x_j)$ and $\sigma^2(x_j)$. Note that $b(x_j)$ consists of $f(x_j)$, $f^{(1)}(x_j)$, $f^{(2)}(x_j)$, $m^{(2)}(x_j)$ and $m^{(4)}(x_j)$. $f(x_j)$, $f^{(1)}(x_j)$, $f^{(2)}(x_j)$ can be estimated consistently by the method mentioned in the previous section. Remember that Theorem 3.1 only require the consistency of $m^{(4)}(x_j)$. For example, it can be estimated by the fourth-order local polynomial regression with the bandwidths of the order $n^{-1/11}$. For $\sigma^2(x_j)$, we can estimate it individually as in equation (4.8) or $\sigma^2(x_j)/f(x_j)$ as a single object by equation (4.10) of Fan and Gijbels (1996). See Fan and Gijbels (1996, Chapter 4) for more discussions on pilot estimates in implementing the local polynomial regression. Only for $m^{(2)}(x_j)$, Theorem 2.1 demand more than consistency. As in the previous case of density estimation, the requirement can be achieved by employing the higher-order kernel such as $K(x) = \frac{1}{2}(3 - x^2)\phi(x)$ where $\phi(x)$ is the density function of the standard normal random variable. For example, the pilot estimate can be obtained by the third-order local polynomial regression with the bandwidths of the order $n^{-1/9}$ as well as the higher-order kernel function.

3.2 Differences of Local Linear Regression Estimators Near The Boundary

In this subsection, we consider estimating the difference of functions at points near the boundary by the difference of local linear estimators of functions. In the cases of the differences of kernel density estimators and local linear regression estimators at interior points, we have remarked that the results can be generalized to the cases where the estimand is the difference of two distinct densities or regression curves. As it will be clear, it is also true for the difference of regression curves near boundary points. However, for boundary cases, we have more varieties since a boundary point can be either left or right boundary. Here as a leading case, we consider the problem of the sharp regression discontinuity design (RDD). Define $m(z) = E(Y_i|X_i = z)$. Suppose that the limits $\lim_{z\to x+} m(z)$ and $\lim_{z\to x-} m(z)$ exist where $z \to x+$ and $z \to x-$ mean taking limit from the right and the left, respectively. When $m(\cdot)$ is discontinuous at x, the parameter of interest in the analysis of the sharp RDD is given by

$$\lim_{z\to x+} m(z) - \lim_{z\to x-} m(z).$$

See Imbens and Lemieux (2008) for a survey on the RDD. The local linear regression becomes especially attractive to estimate these limits since it is known to have automatic boundary adaptive property as shown by Fan and Gijbels (1992). The local linear estimator for $\lim_{z\to x+} m(z)$ is given by $\hat{\alpha}_{h+}(x)$ where

$$\left(\hat{\alpha}_{h+}(x), \hat{\beta}_{h+}(x)\right) \equiv \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \{Y_i - \alpha - \beta'(X_i - x)\}^2 K\left(\frac{X_i - x}{h}\right) 1_{\{X_i \geq x\}}, \tag{29}$$

where $K(\cdot)$ is a kernel function and h is a bandwidth. The solution to minimize (29) can be expressed as

$$\begin{bmatrix} \hat{\alpha}_{h+}(x) \\ \hat{\beta}_{h+}(x) \end{bmatrix} = (X(x)'W_1(x)X(x))^{-1} X(x)'W_1(x)Y \tag{30}$$

where $W_1(x) = \text{diag}(K_{1h}(X_i - x))$ and $K_{1h}(\cdot) = K(\cdot/h)1_{\{\cdot\geq 0\}}/h$, and $X(x)$ and Y are those defined in the previous subsection. Similarly, the local linear estimator for $\lim_{z\to x-} m(x)$ denoted by $\hat{\alpha}_{h-}(x)$ can be obtained by replacing $W_1(x)$ in equation (30) with $W_2(x)$ where $W_2(x) = \text{diag}(K_{2h}(X_i - x))$ and $K_{2h}(\cdot) = K(\cdot/h)1_{\{\cdot\leq 0\}}/h$. Denote $\hat{\alpha}_{h+}$ and $\hat{\alpha}_{h-}$ by $\hat{m}_1(x)$ and $\hat{m}_2(x)$, respectively. Then $\lim_{z\to x+} m(z) - \lim_{z\to x-} m(z)$ is estimated by $\hat{m}_1(x) - \hat{m}_2(x)$ and its conditional MSE given X is given by

$$MSE_n(h_1, h_2) = E\left[\{(\hat{m}_1(x) - \hat{m}_2(x)) - (m_1(x) - m_2(x))\}^2 |X\right]$$

To derive the AMSE based on this $MSE_n(h_1, h_2)$, define

$$\sigma_1^2(x) = \lim_{z \to x+} \sigma^2(z), \quad \sigma_2^2(x) = \lim_{z \to x-} \sigma^2(z), \quad m_1(x) = \lim_{z \to x+} m(z), \quad m_2(x) = \lim_{z \to x-} m(z),$$
$$m_1^{(2)}(x) = \lim_{z \to x+} m^{(2)}(z), \quad m_2^{(2)}(x) = \lim_{z \to x-} m^{(2)}(z), \quad m_1^{(3)}(x) = \lim_{z \to x+} m^{(3)}(z),$$
$$m_2^{(3)}(x) = \lim_{z \to x-} m^{(3)}(z), \quad \mu_{j,0} = \int_0^\infty u^j K(u) du, \quad \nu_{j,0} = \int_0^\infty u^j K^2(u) du.$$

To be precise, we proceed under the following assumption.

ASSUMPTION 3.5 *The density f is bounded above and away from zero. It is also continuously differentiable in a neighborhood of x.*

ASSUMPTION 3.6 *For $z > x$ the conditional mean function $m(z)$ is bounded and three-times continuously differentiable and $\sigma_1^2(x)$, $m_1(x)$, $m_1^{(2)}(x)$ and $m_1^{(3)}(x)$ exist and are bounded. Similarly, for $z < x$ the conditional mean function $m(z)$ is also bounded and three-times continuously differentiable and $\sigma_2^2(x)$, $m_2(x)$, $m_2^{(2)}(x)$ and $m_2^{(3)}(x)$ exist and are bounded.*

Under Assumptions 2.1, 2.3, 3.5, and 3.6, we can generalize the result by Fan and Gijbels (1992) easily to get

$$MSE_n(h_1, h_2) = \left[\frac{b_1}{2} \left\{ m_1^{(2)}(x) h_1^2 - m_2^{(2)}(x) h_2^2 \right\} \right]^2 + \frac{v}{nf(x)} \left\{ \frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_2^2(x)}{h_2} \right\} + o\left(h_1^4 + h_2^4 + \frac{1}{nh_1} + \frac{1}{nh_2} \right)$$

where

$$b_1 = \frac{\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0}}{\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2}, \quad \text{and} \quad v = \frac{\mu_{2,0}^2 \nu_{0,0} - 2\mu_{1,0}\mu_{2,0}\nu_{1,0} + \mu_{1,0}^2 \nu_{2,0}}{(\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2)^2}.$$

Again it is evident that the trade-off between bias and variance can break down when we try to minimize the AMSE based on this MSE. Hence we need to consider the MSE that includes the second-order bias term.

LEMMA 3.3 *Suppose Assumptions 2.1, 2.2 and 3.4 hold. Then it follows that*

$$MSE_n(h_1, h_2) = \left[\frac{b_1}{2} \left\{ m_1^{(2)}(x) h_1^2 - m_2^{(2)}(x) h_2^2 \right\} \right]^2 + \left[\left\{ b_{2,1}(x) h_1^3 - b_{2,2}(x) h_2^3 \right\} \right]^2 + \frac{v}{nf(x)} \left\{ \frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_2^2(x)}{h_2} \right\} + o\left(h_1^8 + h_2^8 + \frac{1}{nh_1} + \frac{1}{nh_2} \right) \tag{31}$$

where

$$b_{2,j}(x) = (-1)^{j+1} \left\{ c_1 \left(\frac{m_j^{(2)}(x)}{2} \frac{f^{(1)}(x)}{f(x)} + \frac{m_j^{(3)}(x)}{6} \right) - c_2 \frac{m_j^{(2)}(x)}{2} \frac{f^{(1)}(x)}{f(x)} \right\}$$
$$c_1 = \frac{\mu_{2,0}\mu_{3,0} - \mu_{1,0}\mu_{4,0}}{\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2}, \quad \textit{and} \quad c_2 = \frac{(\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0})(\mu_{0,0}\mu_{3,0} - \mu_{1,0}\mu_{2,0})}{(\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2)^2}.$$

The result given above is essentially different from the one at interior points since the second order bias terms now involve h^3 rather than h^4. This is because the terms that disappear due to symmetry of the kernel functions remain for one-sided kernel. Based on the MSE provided in Lemma 3.3, the first-order optimal bandwidths to estimate the difference of regression functions near the boundary can be defined.

DEFINITION 3.2 *(Asymptotically First-Order Optimal Bandwidths for Difference of Regression Functions Near The Boundary)*

(i) *When* $m_1^{(2)}(x)m_2^{(2)}(x) < 0$*, choose the bandwidths that minimize*

$$AMSE_{1n}(h_1, h_2) \stackrel{def}{=} \left\{ \frac{b_1}{2} \left[m_1^{(2)}(x)h_1^2 - m_2^{(2)}(x)h_2^2 \right] \right\}^2 + \frac{v}{nf(x)} \left[\frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_2^2(x)}{h_2} \right]. \quad (32)$$

(ii) *When* $m_1^{(2)}(x)m_2^{(2)}(x) > 0$ *and* $m_2^{(2)}(x)^2 b_{2,1}(x) \neq m_1^{(2)}(x)^2 b_{2,2}(x)$*, choose the bandwidths that minimize*

$$AMSE_{2n}(h_1, h_2) \stackrel{def}{=} \left[b_{2,1}(x)h_1^3 - b_{2,2}(x)h_2^3 \right]^2 + \frac{v}{nf(x)} \left[\frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_2^2(x)}{h_2} \right] \quad (33)$$

subject to the restriction $m_1^{(2)}(x)h_1^2 - m_2^{(2)}(x)h_2^2 = 0$

Next Lemma is presented without proof since its proof is analogous to that of Lemma 2.2.

LEMMA 3.4 *Suppose that the same conditions as Lemma 3.3 hold for each case. Then it follows*

(i) *when* $m_1^{(2)}(x)m_2^{(2)}(x) < 0$*,* $AMSE_{1n}(h_1, h_2)$ *has a unique minimum with respect to* h_1 *and* h_2 *and unique minimizers, denoted by* h_1^* *and* h_2^**, are given by*

$$h_1^* = \theta^* n^{-1/5}, \quad h_2^* = \lambda^* h_1^*$$

where

$$\theta^* = \left\{ \frac{v\sigma_1^2(x)}{b_1^2 f(x) m_1^{(2)}(x) \left[m_1^{(2)}(x) - \lambda^{*2} m_2^{(2)}(x) \right]} \right\}^{1/5}, \quad \lambda^* = \left[-\frac{\sigma_2^2(x)m_1^{(2)}(x)}{\sigma_1^2(x)m_2^{(2)}(x)} \right]^{1/3}$$

(ii) *when* $m^{(2)}(x_1)m^{(2)}(x_2) > 0$ *and* $m^{(2)}(x_2)^2 b(x_1) \neq m^{(2)}(x_1)^2 b(x_2)$*,* $AMSE_{2n}(h_1, h_2)$ *subject to the restriction* $m^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2 = 0$ *has a unique minimum with respect to* h_1 *and* h_2 *and unique minimizers, denoted by* h_1^{**} *and* h_2^{**}*, are given by*

$$h_1^{**} = \theta^{**} n^{-1/7}, \quad h_2^{**} = \lambda^{**} h_1^{**}$$

where

$$\theta^{**} = \left\{ \frac{v \left[\sigma_1^2(x) + \sigma_2^2(x)/\lambda^{**} \right]}{6f(x) \left[b_{2,1}(x) - \lambda^{**3} b_{2,2}(x) \right]^2} \right\}^{1/7}, \quad \lambda^{**} = \left[\frac{m_1^{(2)}(x)}{m_2^{(2)}(x)} \right]^{1/2}$$

We see that the first-order optimal bandwidths still exhibits the dichotomous behaviors. The AMSEs with the first-order optimal bandwidths are given by

$$AMSE_{1n}(h_1^*, h_2^*) = n^{-4/5} \left\{ \left(\frac{b_1 \theta^{*2}}{2} \right)^2 \left[m_1^{(2)}(x) - \lambda^{*2} m_2^{(2)}(x) \right]^2 + \frac{v}{\theta^* f(x)} \left[\sigma_1^2(x) + \frac{\sigma_2^2(x)}{\lambda^*} \right] \right\} \quad (34)$$

and

$$AMSE_{2n}(h_1^{**}, h_2^{**}) = n^{-6/7} \left\{ \theta^{**6} \left[b_{2,1}(x) - \lambda^{**3} b_{2,2}(x) \right]^2 + \frac{v}{\theta^{**} f(x)} \left[\sigma_1^2(x) + \frac{\sigma_2^2(x)}{\lambda^{**}} \right] \right\} \quad (35)$$

We note that, when the sign of the second derivatives are positive, the order of the bandwidths is $n^{-1/7}$ that is closer to the order of the bandwidths for the other case than those for interior points.

Then in discussing the asymptotic properties of the proposed bandwidths, we confine attention to the bandwidths (h_1, h_2) in the interval H_n where

$$H_n = H_{1n} \cup H_{2n}$$

where

$$H_{1n} = \left[\delta_{11} n^{-1/5}, \delta_{12} n^{-1/5}\right] \times \left[\delta_{11} n^{-1/5}, \delta_{12} n^{-1/5}\right],$$
$$H_{2n} = \left[\delta_{21} n^{-1/7}, \delta_{22} n^{-1/7}\right] \times \left[\delta_{21} n^{-1/7}, \delta_{22} n^{-1/7}\right]$$

for arbitrarily small δ_{11} and δ_{21}, and large δ_{12} and δ_{22}.

THEOREM 3.2 *Suppose that the same conditions as Lemma 3.3 hold for each case. Assume further that, for $j = 1, 2$, the initial bandwidths and the kernel functions used to define $\hat{b}_{2,j}(x)$, $\hat{f}(x)$ and $\hat{\sigma}_j^2(x)$ satisfy $\hat{b}_{2,j}(x_j) \to b_{2,j}(x)$, $\hat{f}(x) \to f(x)$ and $\hat{\sigma}_j^2(x) \to \sigma_j^2(x)$ in probability, respectively. The initial bandwidth, the initial kernel function and the order of polynomial regression to estimate $\hat{m}_j^{(2)}(x)$ for $j = 1, 2$ satisfy $\hat{m}_j^{(2)}(x) = m_j^{(2)}(x) + o_p(n^{-1/7})$. Let $\hat{\mathbf{h}}$ be a combination of bandwidths that minimizes the modified AMSE defined in (28). Then*

(i) *when $m_1^{(2)}(x) m_2^{(2)}(x) < 0$*

$$\frac{\hat{h}_1}{h_1^*} \to 1, \quad \text{and} \quad \frac{\hat{h}_2}{h_2^*} \to 1$$

in probability and when $m_1^{(2)}(x) m_2^{(2)}(x) > 0$ and $m_2^{(2)}(x)^2 b_{2,1}(x) \neq m_1^{(2)}(x)^2 b_{2,2}(x)$

$$\frac{\hat{h}_1}{h_1^{**}} \to 1, \quad \text{and} \quad \frac{\hat{h}_2}{h_2^{**}} \to 1$$

in probability.

(ii) *when $m_1^{(2)}(x) m_2^{(2)}(x) < 0$*

$$\frac{\widehat{MMSE}_n(\hat{h}_1, \hat{h}_2)}{MSE_n(h_1^*, h_2^*)} \to 1$$

in probability and when $m_1^{(2)}(x) m_2^{(2)}(x) > 0$ and $m_2^{(2)}(x)^2 b_{2,1}(x) \neq m_1^{(2)}(x)^2 b_{2,2}(x)$

$$\frac{\widehat{MMSE}_n(\hat{h}_1, \hat{h}_2)}{MSE_n(h_1^{**}, h_2^{**})} \to 1$$

in probability.

As in the case for regression functions at interior points we can propose another feasible bandwidths that might be more intuitive. The bandwidths are based on the expression of the asymptotically first-order bandwidths given in Lemma 3.4. The proposed bandwidths are more direct version of the plug-in bandwidths defined by

$$\tilde{h}_1 = \hat{\theta}_1 n^{-1/5} 1_{\{\hat{m}_1^{(2)}(x)\hat{m}_2^{(2)}(x) < 0\}} + \hat{\theta}_2 n^{-1/7} \left(1 - 1_{\{\hat{m}_1^{(2)}(x)\hat{m}_2^{(2)}(x) < 0\}}\right),$$
$$\tilde{h}_2 = \hat{\theta}_1 \hat{\lambda}_1 n^{-1/5} 1_{\{\hat{m}_1^{(2)}(x)\hat{m}_2^{(2)}(x) < 0\}} + \hat{\theta}_2 \hat{\lambda}_2 n^{-1/7} \left(1 - 1_{\{\hat{m}_1^{(2)}(x)\hat{m}_2^{(2)}(x) < 0\}}\right)$$

where

$$\hat{\theta}_1 = \left\{ \frac{v\hat{\sigma}_1^2(x)}{b_1^2 \hat{f}(x)\hat{m}_1^{(2)}(x)\left[\hat{m}_1^{(2)}(x) - \hat{\lambda}^2\hat{m}_2^{(2)}(x)\right]} \right\}^{1/5}, \quad \hat{\lambda}_1 = \left[-\frac{\hat{\sigma}_2^2(x)\hat{m}_1^{(2)}(x)}{\hat{\sigma}_1^2(x)\hat{m}_2^{(2)}(x)} \right]^{1/3},$$

$$\hat{\theta}_2 = \left\{ \frac{v\left[\hat{\sigma}_1^2(x) + \hat{\sigma}_2^2(x)/\hat{\lambda}_2\right]}{6\hat{f}(x)\left[\hat{b}_{2,1}(x) - \hat{\lambda}_2^3\hat{b}_{2,2}(x)\right]^2} \right\}^{1/7}, \quad \text{and} \quad \hat{\lambda}_2 = \left[\frac{\hat{m}_1^{(2)}(x)}{\hat{m}_2^{(2)}(x)} \right]^{1/2}.$$

These bandwidths switches from one to another depending on the estimated sign of of the product of the second derivatives. It can be also shown that the direct plug-in bandwidths are asymptotically as good as the asymptotically first-order bandwidths. The following result can be derived in the similar manner as Theorem 3.2.

COROLLARY 3.2 *Suppose that the same conditions as Theorem 3.2 hold for each case. Then the results that hold for* $(\hat{h}_1, \hat{h}_2)$ *also holds for* $(\tilde{h}_1, \tilde{h}_2)$

As Theorem 2.1 and Theorem 3.1, Theorem 3.2 requires pilot estimates for $m_j^{(2)}(x)$, $b_{2,j}(x)$, $f(x)$ and $\sigma_j^2(x)$. One notable difference between Theorem 3.2 and previous two is that Theorem 3.2 only require the substantially weaker condition for $\hat{m}_j^{(2)}(x)$, i.e., $\hat{m}_j^{(2)}(x) = m_j^{(2)}(x) + o_p(n^{-1/7})$. This is true because the second-order bias term is involved in h^3 rather than h^4. An important consequence of this is that we do not have to employ the higher-order kernel function in estimating $m_j^{(2)}(x)$. Third-order local polynomial regression with the triangular kernel function $K(x) = (1 - |x|)I_{[-1,1]}(x)$ and the bandwidth of the order $n^{-1/9}$ yields $\hat{m}_j^{(2)}(x) = m_j^{(2)}(x) + O_p(n^{-2/9})$, satisfying the restriction. See, e.g., Fan and Gijbels (1996, Chapter 3) for the order of the bias and variance in local polynomial regression. We employ the triangular weight functions since they are optimal in a weak minimax sense as shown in Cheng et al. (1997). We briefly mention how to obtain other pilot estimates. The derivatives of $f(x)$ and $\sigma_j^2(x)$ can be estimated by the method described in Subsection 2.3 and $m_j^{(3)}(x)$ by the third-order local polynomial regression with the triangular kernel function and the bandwidths of the order $n^{-1/9}$.

Not to mention, the case of the sharp RDD is just one special case of applications and our proposed method can be generalized to cover other cases. For example those include cases to estimate

$$m_1(x_1) - m_2(x_2)$$

where x_1 and x_2 can be taken as either the left boundary or the right boundary point. Throughout this section, cases where one is near the boundary and another is an interior point are excluded. This is somewhat similar to the situation where one of the second derivative is equal to zero since the second-order bias term involve two terms with different orders in the bandwidth. The discussion given in Section 5 might provide some insight for this case.

4 Simulation

To investigate finite sample performance of the proposed method, we conduct simulation experiment. We concentrate on the case of the RDD since this would be one of the most empirically relevant cases.

4.1 Data Generating Processes

The objective of the RDD application is to estimate

$$\lim_{z \to x+} m(z) - \lim_{z \to x-} m(z).$$

where $m(z) = E(Y_i|X_i = z)$. We consider four DGPs considered in Imbens and Kalyanaraman (2012) and an original one that is a modified version of the first DGP in Imbens and Kalyanaraman (2012). The DGPs considered are as follows:

1. Lee data

$$m(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } x < 0 \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{if } x \geq 0 \end{cases}$$

2. Quadratic

$$m(x) = \begin{cases} 3x^2 & \text{if } x < 0 \\ 4x^2 & \text{if } x \geq 0 \end{cases}$$

3. Constant Additive Treatment Effect 1

$$m(x) = \ 0.42 + 0.1 \cdot 1_{\{x \geq 0\}} + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5$$

4. Constant Additive Treatment Effect 2

$$m(x) = \ 0.42 + 0.1 \cdot 1_{\{x \geq 0\}} + 0.84x + 7.99x^3 - 9.01x^4 + 3.56x^5$$

5. Modified Lee data

$$m(x) = \begin{cases} 0.48 + 1.27x - 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } x < 0 \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{if } x \geq 0 \end{cases}$$

See Imbens and Kalyanaraman (2012) for DGP 1-4. In DGP 5, we flip the sign of the coefficient of x^2 in the negative range of the forcing variable so that the second derivatives have the same sign.

For each DGP, we consider two specifications of the forcing variable and the additive error term. The first one is exactly the same as that considered by Imbens and Kalyanaraman (2012), the forcing variable is generated by a Beta distribution. More precisely, let Z have a Beta distribution with parameters $\alpha = 2$ and $\beta = 4$. Then the forcing variable X is given by $2Z - 1$. An additive error is the normal error with variance 0.1295^2.

In the second specification, the forcing variable X has a normal distribution with mean -0.1 and variance 1. We use the same error term as the first one for data with $X_i < 0$ and the normal error with $5 \cdot 0.1295^2$ for data with $X_i \geq 0$.

We call the first and the second specifications Case 1 and Case 2, respectively. We use data sets of size 500 and the results are drawn from 1000 replications.

4.2 An Algorithm

In implementing the proposed method, we need pilot estimates for values of the density and the first derivative of density at the discontinuity point, the second and the third derivatives of the regression functions. We obtain these pilot estimates by the following steps.

4.2.1 Step 1

First we calculate the density of the forcing variable at discontinuity point $f(0)$. It is estimated by a kernel density estimator with Epanechnikov kernel. A pilot bandwidth for the kernel density estimation is chosen by the normal scale rule. (See Wand and Jones (1994) for the normal scale rules.) The first derivative of the density is estimated by the method proposed by Jones (1994). Again a pilot bandwidth is obtained by the normal scale rule. The conditional variance at the threshold $\sigma_+^2(0)$ and $\sigma_+^2(0)$ are estimated by the same manner as Imbens and Kalyanaraman (2012).

4.2.2 Step 2

Second we obtain a pilot bandwidth for the 4th order local polynomial regression. To do so we fit a fifth order polynomial to data with $X_i \geq 0$ to get a pilot estimate of the fifth derivative denoted by $\hat{m}_+^{(5)}(0)$. We repeat the same for data with $X_i < 0$ to get $\hat{m}_-^{(5)}(0)$. Then the plug-in bandwidth is calculated by

$$h_{p,s} = C_{3,4}(K) \left(\frac{\sigma_s^2}{\hat{f}(0) \cdot \hat{m}_s^{(5)}(0) \cdot N_s} \right)^{1/11}$$

where $s = +$ or $-$, $C_{3,4}(K)$ is defined in Fan and Gijbels (1996), and N_+ is the number of observations with $X_i \geq 0$ and N_- with $X_i < 0$.

4.2.3 Step 3

Next we estimate the fourth derivatives at the threshold by 4th order local polynomial regression with the pilot bandwidths obtained in Step 2. They are denoted by $\hat{m}_1^{(4)}(0)$, $\hat{m}_2^{(4)}(0)$ following the notations introduced in the previous section.

4.2.4 Step 4

Then the pilot estimate of the fourth derivatives are used to refine the bandwidths following the procedure proposed by Fan and Gijbels (1995). Then the refined bandwidth is used for the third order local polynomial regression to obtain the second and the third derivatives $\hat{m}_1^{(2)}(0)$, $\hat{m}_2^{(2)}(0)$, $\hat{m}_1^{(3)}(0)$, and $\hat{m}_2^{(3)}(0)$.

4.2.5 Step 5

Finally plug all pilot estimates in the modified MSE considered in the previous section to obtain bandwidth that minimizes the estimated MSE.

4.3 Results

The simulation results for Case 1 and Case 2 are presented in Tables 1 and 2. The result using the bandwidth proposed by Imbens and Kalyanaraman (2012) is also presented for comparison purpose.

Table 1 shows somewhat disappointing result for the proposed bandwidth. The result shows superior performance of the proposed method concerning bias but poor performance in terms of RMSE. In fact, this may not be very surprising given the fact that the number of observations with $X_i > 0$ is less than 100 on average and that the proposed method require pilot estimates of third derivatives. Also observe that the second and the third derivatives of DGPs 3 and 4 are equal, implying the theoretically optimal bandwidths on positive and negative sides are equal. Since the bandwidth proposed by Imbens and Kalyanaraman (2012) is derived under the restriction of the same bandwidth on both sides, these DGPs would work favorably for it.

Table 2 shows superior performance of the proposed bandwidth in terms of bias. The proposed bandwidth also works better in terms of RMSE for Designs 1, 3, and 5. Even for non-standard designs such as Design 2 and 4, the performance are comparable to that of Imbens and Kalyanaraman (2012). The main source of the results are that we have plenty of observations on both sides. In fact, there are more than 200 observations on both sides on average.

5 Discussion and Topics for Future Researches

In this section, we provide some discussion and topics for future researches.

Table 1: Bias and RMSE for the Sharp Regression Discontinuity Design, Case 1

Design	IK		Proposed Bandwidth	
	Bias	RMSE	Bias	RMSE
Design 1	0.0439	0.0587	0.0203	0.0678
Design 2	0.0065	0.0394	0.0033	0.0807
Design 3	0.0560	0.0832	0.0352	0.1447
Design 4	0.0574	0.0844	0.0471	0.1378
Design 5	-0.0214	0.0651	-0.0054	0.0814

Table 2: Bias and RMSE for the Sharp Regression Discontinuity Design, Case 2

Design	IK		Proposed Bandwidth	
	Bias	RMSE	Bias	RMSE
Design 1	0.0179	0.0574	0.0120	0.0516
Design 2	-0.0365	0.0508	0.0078	0.0519
Design 3	0.0425	0.0781	0.0235	0.0733
Design 4	0.0458	0.0819	0.0186	0.1016
Design 5	0.0086	0.0929	0.0006	0.0775

5.1 What will happen when only one of the second derivatives is equal to zero?

We consider what will happen when the product of the second derivatives is equal to zero. For simplicity we restrict our attention to the case of density estimation. However, as we generalized the proposed method for density estimation to regression function estimation in the previous section, our discussion can be generalized to more general situations. It would be clear that the standard procedure dealing with the AMSE only with the first-order bias term fails when either $f^{(2)}(x_1)$ or $f^{(2)}(x_2)$ is equal to 0 or when both $f^{(2)}(x_1)$ and $f^{(2)}(x_2)$ are equal to 0. We mentioned how to deal with a case where both $f^{(2)}(x_1)$ and $f^{(2)}(x_2)$ are equal to 0 in Subsection 2.2. Here we consider what will be the property of the AMSE with the higher-order bias term and whether the automatic bandwidth selection method proposed in the section 2.2 is applicable when only one of the second derivatives is equal to 0. In the following we consider a case where $f^{(2)}(x_2) = 0$ since the problem is symmetric.

We will make the following assumption since we will need further smoothness of the density in some situation.

ASSUMPTION 5.1 *The density f is bounded and six times continuously differentiable in the neighborhood of x_1 and x_2.*

We present the following result that is a direct extension of Lemma 2.1 without proof.

LEMMA 5.1 *Suppose Assumptions 2.1 and 5.1 hold. Then*

$$\begin{aligned} MSE_n(h_1, h_2) = &\left[\frac{\mu_2}{2}\left\{f^{(2)}(x_1)h_1^2 - f^{(2)}(x_2)h_2^2\right\} + \frac{\mu_4}{4!}\left\{f^{(4)}(x_1)h_1^4 - f^{(4)}(x_2)h_2^4\right\}\right. \\ &\left. + \frac{\mu_6}{6!}\left\{f^{(6)}(x_1)h_1^6 - f^{(6)}(x_2)h_2^6\right\}\right]^2 \\ &+ \nu_0\left\{\frac{f(x_1)}{nh_1} + \frac{f(x_2)}{nh_2}\right\} + o\left(h_1^{12} + h_2^{12} + \frac{1}{nh_1} + \frac{1}{nh_2}\right). \end{aligned} \tag{36}$$

In the following corollary we present the asymptotic properties of the MSE when the leading bias term is removed if possible.

COROLLARY 5.1 *Suppose Assumptions 2.1 hold.*

(i) *Further suppose Assumption 2.4 hold and* $f^{(2)}(x_1)f^{(4)}(x_2) < 0$ *and* $f^{(2)}(x_2) = 0$*, Then we obtain that*

$$MSE_n(h_1, h_2) = \left[\frac{\mu_2}{2} f^{(2)}(x_1) h_1^2 - \frac{\mu_4}{4!} f^{(4)}(x_2) h_2^4\right]^2 + \frac{\nu_0}{n}\left\{\frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2}\right\} + o\left(h_1^4 + h_1^2 h_2^4 + h_2^8 + \frac{1}{nh_1} + \frac{1}{nh_2}\right). \tag{37}$$

(ii) *Further suppose Assumption 5.1 hold and* $f^{(2)}(x_1)f^{(4)}(x_2) > 0$*,* $f^{(4)}(x_1)f^{(6)}(x_2) < 0$*,* $f^{(2)}(x_2) = 0$ *and* $f^{(6)}(x_2) \neq 0$*. Then we obtain that*

$$MSE_n(h_1, h_2) = \left[\frac{\mu_4}{4!} f^{(4)}(x_1) h_1^4 - \frac{\mu_6}{6!} f^{(6)}(x_2) h_2^6\right]^2 + \frac{\nu_0}{n}\left\{\frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2}\right\} + o\left(h_1^8 + h_1^4 h_2^6 + h_2^{12} + \frac{1}{nh_1} + \frac{1}{nh_2}\right) \tag{38}$$

where h_1 *and* h_2 *satisfy the relationship*

$$\frac{1}{2}\mu_2 f^{(2)}(x_1) h_1^2 - \frac{1}{4!}\mu_4 f^{(4)}(x_2) h_2^4 = 0.$$

The assumption of $f^{(4)}(x_1)f^{(6)}(x_2) < 0$ in Corollary 5.1 (ii) is made for simplicity. This assumption excludes the case where the first- and the second-order term vanish simultaneously. If this is violated, we must proceed for the MSE with the higher-order bias term. In Corollary 5.1, the case where one of the second order derivatives is equal to zero is divided into two cases according to the signs for the second and the fourth derivatives. Basic implications of Corollary 5.1 is that there exits an opportunity to eliminate the first-order bias term when $f^{(2)}(x_1)f^{(4)}(x_2) > 0$ as in the case discussed in Section 2. A notable difference from Section 2 is that two terms in the bias expression are of different orders.

Next we consider the result of minimizing the corresponding AMSE. These cases are more involved than the previously discussed cases. The first lemma of the two discuss the case where $f^{(2)}(x_1)f^{(4)}(x_2) < 0$.

LEMMA 5.2 *Suppose Assumptions 2.1 and 2.4 hold. Further suppose* $f^{(2)}(x_1)f^{(4)}(x_2) < 0$ *and* $f^{(2)}(x_2) = 0$*. Then minimizing the MSE given in (37) ignoring higher order terms leads to the choice of*

$$h_1 = O\left(n^{-1/5}\right), \quad h_2 = O\left(n^{-3/25}\right).$$

Thus the MSE given in (37) ignoring higher order terms can be written as

$$n^{-\frac{4}{5}}\left(\left\{\frac{\mu_2}{2} f^{(2)}(x_1)\theta_1^2\right\}^2 + \frac{\nu_0 f(x_1)}{\theta_1}\right) \tag{39}$$

where θ_1 *and* θ_2 *satisfy the relationship* $\mu_2 f^{(2)}(x_1) f(x_2)\theta_1^3 - 6\mu_4 f(x_1) f^{(4)}(x_2)\theta_2^5 = 0$.

Denote the AMSE (39) by $Q_n^*(h_1, h_2)$. Interesting characteristics of Lemma 5.2 is that we now have two bandwidth with different rates and that the MSE is asymptotically dominated by the AMSE of $\hat{f}(x_1)$.

The next lemma considers the case where $f^{(2)}(x_1)f^{(4)}(x_2) > 0$.

LEMMA 5.3 *Suppose Assumptions 2.1 and 5.1 hold. Also suppose* $f^{(2)}(x_1)f^{(4)}(x_2) > 0$, $f^{(2)}(x_2) = 0$ *and* $f^{(6)}(x_2) \neq 0$. *Then minimizing the MSE given in (37) ignoring higher order terms leads to the choice of*

$$h_1 = O(n^{-1/7}), \quad h_2 = O(n^{-1/14}).$$

Thus the MSE given in (38) ignoring higher order terms can be written as

$$n^{-\frac{6}{7}} \left(\left\{ \frac{\mu_6}{6!} f^{(6)}(x_2)\theta_2^6 \right\}^2 + \frac{\nu_0 f(x_1)}{\theta_1} \right) \tag{40}$$

where θ_1 *and* θ_2 *satisfy the relationship*

$$\frac{\mu_2}{2} f^{(2)}(x_1)\theta_1^2 - \frac{\mu_4}{4!} f^{(4)}(x_2)\theta_2^4 = 0.$$

Denote the AMSE (40) by $Q_n^{**}(h_1, h_2)$. Lemma 5.3 implies that we have two bandwidth with different rates as the results of Lemma 5.2. However, these rates are distinct from those given in Lemma 5.2. What is more interesting is that the MSE is asymptotically dominated by the component due to bias of one and to variance of the other kernel density estimator.

Next lemma confirms that the problems minimizing $Q_n^*(h_1, h_2)$ and $Q_n^{**}(h_1, h_2)$ is well-defined. We present the lemma without proof since it is analogous to Lemma 2.2.

LEMMA 5.4 *Suppose that the same conditions as those in Lemma 5.1 hold for each case. Then both* $Q_n^*(h_1, h_2)$ *and* $Q_n^{**}(h_1, h_2)$ *have a unique minimum with respect to* h_1 *and* h_2.

Above discussion shows that we may use the AMSEs to decide the two bandwidth parameters in these two cases as well. The question is whether we can extend the proposed method in the previous sections to cover these cases. Unfortunately, it turns out to be a very difficult task. For example, consider the case where $f^{(2)}(x_1)f^{(4)}(x_2) < 0$ and $f^{(2)}(x_2) = 0$. Key observations made in Lemma 5.2 is that the MSE is asymptotically dominated by the MSE of $\hat{f}(x_1)$ and h_2 is determined by the bias due to the cross product term of $f^2(x_1)$ and $f^{(4)}(x_2)$. Since the cross product term is abstracted from the MMSE, h_2 cannot be chosen properly. There is possibility for the direct plug-in to work but at this point this is a topic of our future research.

5.2 Generalization and Topics for Future Researches

Next, we discuss the generalization of the proposed bandwidth selection method and its extensions. The followings are the topics of our research agenda.

- Extend the proposed method to deal with LATE, fuzzy RDD

6 Appendix: Proofs

Proof of Lemma 2.1: Remember that we have by Taylor expansion that

$$f(x+hu) = f(x) + f^{(1)}(x)hu + \frac{1}{2}f^{(2)}(x)h^2u^2 + \frac{1}{3!}f^{(3)}(x)h^3u^3 + \frac{1}{4!}f^{(4)}(x)h^4u^4 + o(h^4)$$

where $f^{(r)}(\cdot)$ stands for r-th derivative of $f(\cdot)$. Observe that

$$\begin{aligned} E\left[\hat{f}_h(x)\right] - f(x) &= \frac{1}{h}\int K\left(\frac{x-u}{h}\right) f(u)du - f(x) \\ &= \int K(u)\left\{f(x-hu) - f(x)\right\} du \\ &= \frac{\mu_2}{2} f^{(2)}(x)h^2 + \frac{\mu_4}{4!} f^{(4)}(x)h^4 + o(h^4) \end{aligned}$$

where we used the facts that $K(\cdot)$ is symmetric and $\int K(u)du = 1$. Then it follows that the bias for (2) is given by

$$\begin{aligned}
\text{Bias}(h_1, h_2) &= \left\{ E\left[\hat{f}_{h_1}(x_1) - \hat{f}_{h_2}(x_2) \right] \right\} - \{ f(x_1) - f(x_2) \} \\
&= \left\{ E\left[\hat{f}_{h_1}(x_1) \right] - f(x_1) \right\} - \left\{ E\left[\hat{f}_{h_2}(x_2) \right] - f(x_2) \right\} \\
&= \left\{ \frac{\mu_2}{2} f^{(2)}(x_1) h_1^2 + \frac{\mu_4}{4!} f^{(4)}(x_1) h_1^4 \right\} - \left\{ \frac{\mu_2}{2} f^{(2)}(x_2) h_2^2 + \frac{\mu_4}{4!} f^{(4)}(x_2) h_2^4 \right\} + o(h_1^4 + h_2^4) \\
&= \frac{\mu_2}{2} \left\{ f^{(2)}(x_1) h_1^2 - f^{(2)}(x_2) h_2^2 \right\} + \frac{\mu_4}{4!} \left\{ f^{(4)}(x_1) h_1^4 - f^{(4)}(x_2) h_2^4 \right\} + o\left(h_1^4 + h_2^4 \right). \quad (41)
\end{aligned}$$

This gives a component due to bias. Next we consider a contribution from variance. Observe that

$$\begin{aligned}
&\text{Var}(h_1, h_2) = \text{var}\left(\hat{f}_{h_1}(x_1) - \hat{f}_{h_2}(x_2) \right) \\
&= \frac{1}{n} \text{var} \left\{ \frac{1}{h_1} K\left(\frac{x_1 - X}{h_1} \right) - \frac{1}{h_2} K\left(\frac{x_2 - X}{h_2} \right) \right\} \\
&= \frac{1}{n} E\left[\left\{ \frac{1}{h_1} K\left(\frac{x_1 - X}{h_1} \right) - \frac{1}{h_2} K\left(\frac{x_2 - X}{h_2} \right) \right\}^2 \right] - \frac{1}{n} \left\{ E\left(\frac{1}{h_1} K\left(\frac{x_1 - X}{h_1} \right) - \frac{1}{h_2} K\left(\frac{x_2 - X}{h_2} \right) \right) \right\}^2 \\
&= \frac{1}{nh_1} \int K^2(u) f(x_1 - h_1 u) du + \frac{1}{nh_2} \int K^2(u) f(x_2 - h_2 u) du \\
&\qquad - \frac{2}{nh_2} \int K(u) K\left(\frac{x_1 - x_2 - h_1 u}{h_2} \right) f(x_1 + h_1 u) du - \frac{1}{n} \{ f(x_1) - f(x_2) + \text{Bias}(h_1, h_2) \}^2 \\
&= \frac{\nu_0}{n} \left\{ \frac{f(x_1)}{h_1} + \frac{f(x_2)}{h_2} \right\} + o\left(\frac{1}{nh_1} + \frac{1}{nh_2} \right). \quad (42)
\end{aligned}$$

Combining (41) and (42) gives the required result.

Proof of Lemma 2.2: (i) First order conditions are given by

$$\left. \frac{\partial AMSE_{1n}(h_1, h_2)}{\partial h_1} \right|_{h_1 = h_1^*, h_2 = h_2^*} = \mu_2^2 f^{(2)}(x_1) h_1^* \left[f^{(2)}(x_1) h_1^{*2} - f^{(2)}(x_2) h_2^{*2} \right] - \frac{\nu_0}{n} \frac{f(x_1)}{h_1^{*2}} = 0, \quad (43)$$

$$\left. \frac{\partial AMSE_{2n}(h_1, h_2)}{\partial h_2} \right|_{h_1 = h_1^*, h_2 = h_2^*} = -\mu_2^2 f^{(2)}(x_2) h_2^* \left[f^{(2)}(x_1) h_1^{*2} - f^{(2)}(x_2) h_2^{*2} \right] - \frac{\nu_0}{n} \frac{f(x_2)}{h_2^{*2}} = 0 \quad (44)$$

Dividing (44) by (43) yields

$$\frac{h_2^*}{h_1^*} = \left\{ \frac{f(x_2)}{f(x_1)} \left[-\frac{f^{(2)}(x_1)}{f^{(2)}(x_2)} \right] \right\}^{1/3}$$

or

$$h_2^* = \lambda^* h_1^*. \quad (45)$$

Substituting (45) into (43) and solving it for h_1 gives the result (10).

To see h_1^* and h_2^* are global minimizers, it suffices to show that $AMSE_{1n}(h_1, h_2)$ is strictly convex with respect to h_1 and h_2. For the strict convexity we show that the Hessian matrix is positive definite, i.e. we show that

$$\frac{\partial^2 AMSE_{1n}(h_1, h_2)}{\partial h_1^2} > 0, \quad \frac{\partial^2 AMSE_{1n}(h_1, h_2)}{\partial h_1^2} \cdot \frac{\partial^2 AMSE_{1n}(h_1, h_2)}{\partial h_2^2} - \left[\frac{\partial^2 AMSE_{1n}(h_1, h_2)}{\partial h_1 \partial h_2} \right]^2 > 0.$$

It follows by noting $f^{(2)}(x_1)$ and $f^{(2)}(x_2)$ have different signs that

$$\frac{\partial^2 AMSE_{1n}(h_1, h_2)}{\partial \theta_1^2} = \mu_2^2 f^{(2)}(x_1) \left[f^{(2)}(x_1) h_1^2 - f^{(2)}(x_2) h_2^2 \right] + 2 \left[\mu_2 f^{(2)}(x_1) h_1 \right]^2 + \frac{2\nu_0 f(x_1)}{nh_1^3} > 0$$

since $f(\cdot)$, μ_2, ν_0, n, h_1 and h_2 are all positive. We can also show that

$$\begin{aligned}
&\frac{\partial^2 AMSE_{1n}(h_1,h_2)}{\partial h_1^2}\cdot\frac{\partial^2 AMSE_{1n}(h_1,h_2)}{\partial h_2^2}-\left[\frac{\partial^2 AMSE_{1n}(h_1,h_2)}{\partial h_1\partial h_2}\right]^2\\
&=\left\{\mu_2^2 f^{(2)}(x_1)\left[f^{(2)}(x_1)h_1^2-f^{(2)}(x_2)h_2^2\right]+2\left[\mu_2 f^{(2)}(x_1)h_1\right]^2+\frac{2\nu_0 f(x_1)}{nh_1^3}\right\}\\
&\quad\times\left\{-\mu_2^2 f^{(2)}(x_2)\left[f^{(2)}(x_1)h_1^2-f^{(2)}(x_2)h_2^2\right]+2\left[\mu_2 f^{(2)}(x_2)h_2\right]^2+\frac{2\nu_0 f(x_2)}{nh_2^3}\right]\\
&\quad-\left[2\mu_2^2 f^{(2)}(x_1)f^{(2)}(x_2)h_1h_2\right]^2.
\end{aligned}$$

Note that if we ignore the first and the third terms in the two brackets of the first term on the right-hand side, it coincides with the last term on the right-hand side. But the both the first and the third terms are positive as discussed earlier. Thus the difference of the two terms are positive.

(ii) Next, we consider $AMSE_{2n}(h_1,h_2)$. With the restriction $f^{(2)}(x_1)h_1^2-f^{(2)}(x_2)h_2^2=0$, $AMSE_{2n}(h_1,h_2)$ in (9) can be written as

$$AMSE_{2n}(h_1)=\left\{\frac{\mu_4}{4!}\left[f^{(4)}(x_1)-\lambda^{**4}f^{(4)}(x_2)\right]h_1^4\right\}^2+\frac{\nu_0}{nh_1}\left[f(x_1)+\frac{f(x_2)}{\lambda^{**}}\right].$$

The first order condition becomes

$$\left.\frac{\partial AMSE_{2n}(h_1)}{\partial h_1}\right|_{h_1=h_1^{**}}=\frac{1}{72}\mu_4^2\left[f^{(4)}(x_1)-\lambda^{**4}f^{(4)}(x_2)\right]^2 h_1^{**7}-\frac{\nu_0}{nh_1^{**2}}\left[f(x_1)+\frac{f(x_2)}{\lambda^{**}}\right]=0.$$

Solving this with respect to h_1^{**} yields (11). To see that $AMSE_{2n}(h_1)$ has a unique minimum, observe that

$$\frac{\partial^2 AMSE_{2n}(h_1)}{\partial h_1^2}=\frac{7}{56}\mu_4^2\left[f^{(4)}(x_1)-\lambda^{**4}f^{(4)}(x_2)\right]^2 h_1^6+\frac{2\nu_0}{h_1^3}\left[f(x_1)+\frac{f(x_2)}{\lambda^{**}}\right].$$

Noting both terms on the right-hand side are positive proves the strict convexity.

Proof of Theorem 2.1: Denote $\mathbf{h}=(h_1,h_2)$, $\hat{\mathbf{h}}=(\hat{h}_1,\hat{h}_2)$, $\mathbf{h}^*=(h_1^*,h_2^*)$, $\mathbf{h}^{**}=(h_1^{**},h_2^{**})$, $\hat{\mathbf{h}}/\mathbf{h}^*=(\hat{h}_1/h_1^*,\hat{h}_2/h_2^*)$ and $\hat{\mathbf{h}}/\mathbf{h}^{**}=(\hat{h}_1/h_1^{**},\hat{h}_2/h_2^{**})$. For $\varepsilon>0$, denote an ε-neighborhood of 1 by $N_1(\varepsilon,1)$ and $N_1(\varepsilon)=N_1(\varepsilon,1)\times N_1(\varepsilon,1)$. First, suppose $f^{(2)}(x_1)f^{(2)}(x_2)<0$. Observe that

$$\begin{aligned}
&\Pr\left\{\inf_{\mathbf{h}\in H_n}\widehat{MMSE}_n(\mathbf{h})\le\widehat{MMSE}_n(\mathbf{h}^*)\right\}\\
&\le\Pr\left\{\inf_{\mathbf{h}\in H_n}\widehat{MMSE}_n(\mathbf{h})\le\widehat{MMSE}_n(\mathbf{h}^*),\hat{\mathbf{h}}/\mathbf{h}^*\in N_1(\varepsilon),\hat{\mathbf{h}}\in H_{1n}\right\}\\
&\quad+\Pr\left\{\inf_{\mathbf{h}\in H_n}\widehat{MMSE}_n(\mathbf{h})\le\widehat{MMSE}_n(\mathbf{h}^*),\hat{\mathbf{h}}/\mathbf{h}^*\notin N_1(\varepsilon),\hat{\mathbf{h}}\in H_{1n}\right\}\\
&\quad+\Pr\left\{\inf_{\mathbf{h}\in H_n}\widehat{MMSE}_n(\mathbf{h})\le\widehat{MMSE}_n(\mathbf{h}^*),\hat{\mathbf{h}}\in H_{2n}\right\}
\end{aligned}$$

where $A\backslash B=\{x|x\in A\text{ and }x\notin B\}$ for given sets A and B. Then it follows

$$\begin{aligned}
&\Pr\left\{\inf_{\mathbf{h}\in H_n}\widehat{MMSE}_n(\mathbf{h})\le\widehat{MMSE}_n(\mathbf{h}^*)\right\}\le\Pr\left\{\hat{\mathbf{h}}/\mathbf{h}^*\in N_1(\varepsilon),\hat{\mathbf{h}}\in H_{1n}\right\}\\
&\quad+\Pr\left\{\inf_{\mathbf{h}\in H_{1n},\mathbf{h}/\mathbf{h}^*\notin N_1(\varepsilon)}\widehat{MMSE}_n(\mathbf{h})-\widehat{MMSE}_n(\mathbf{h}^*)\le 0\right\}\\
&\quad+\Pr\left\{\inf_{\mathbf{h}\in H_{2n}}\widehat{MMSE}_n(\mathbf{h})-\widehat{MMSE}_n(\mathbf{h}^*)\le 0\right\}. \qquad (46)
\end{aligned}$$

The LHS of equation (46) is equal to 1 whenever the parameter space includes $\mathbf{h}^*$ and this is true since we take arbitrarily small δ_{11} and large δ_{12}. Then the first term on the right-hand side of equation (46) converges to 1 if the last two terms on the right-hand side of equation (46) converge to 0. Then it suffices to show

$$\Pr\left\{\inf_{\mathbf{h}\in H_{1n},\mathbf{h}/\mathbf{h}^*\notin N_1(\varepsilon)} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^*) \leq 0\right\} \to 0 \tag{47}$$

and

$$\Pr\left\{\inf_{\mathbf{h}\in H_{2n}} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^*) \leq 0\right\} \to 0 \tag{48}$$

as n goes to infinity. We prove (47). Note that

$$\begin{aligned}
&\Pr\left\{\inf_{\mathbf{h}\in H_{1n},\mathbf{h}/\mathbf{h}^*\notin N_1(\varepsilon)} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^*) \leq 0\right\} \\
&= \Pr\left\{\inf_{\mathbf{h}\in H_{1n},\mathbf{h}/\mathbf{h}^*\notin N_1(\varepsilon)} \left[\widehat{MMSE}_n(\mathbf{h}) - AMSE_{1n}(\mathbf{h}) + AMSE_{1n}(\mathbf{h})\right]\right. \\
&\qquad\qquad \left. \leq AMSE_{1n}(\mathbf{h}^*) - AMSE_{1n}(\mathbf{h}^*) + \widehat{MMSE}_n(\mathbf{h}^*)\right\} \\
&\leq \Pr\left\{\inf_{\mathbf{h}\in H_{1n},\mathbf{h}/\mathbf{h}^*\notin N_1(\varepsilon)} \left[\widehat{MMSE}_n(\mathbf{h}) - AMSE_{1n}(\mathbf{h})\right] \leq \widehat{MMSE}_n(\mathbf{h}^*) - AMSE_{1n}(\mathbf{h}^*)\right. \\
&\qquad\qquad \left. - \inf_{\mathbf{h}\in H_{1n},\mathbf{h}/\mathbf{h}^*\notin N_1(\varepsilon)} [AMSE_{1n}(\mathbf{h}) - AMSE_{1n}(\mathbf{h}^*)]\right\} \\
&\leq \Pr\left\{2\sup_{\mathbf{h}\in H_{1n}} \left[\widehat{MMSE}_n(\mathbf{h}) - AMSE_{1n}(\mathbf{h})\right]\right. \\
&\qquad\qquad \left. \geq \inf_{\mathbf{h}\in H_{1n},\mathbf{h}/\mathbf{h}^*\notin N_1(\varepsilon)} [AMSE_{1n}(\mathbf{h}) - AMSE_{1n}(\mathbf{h}^*)]\right\} \\
&\leq \Pr\left\{2\sup_{\mathbf{h}\in H_{1n}} n^{4/5}\left[\widehat{MMSE}_n(\mathbf{h}) - AMSE_{1n}(\mathbf{h})\right]\right. \\
&\qquad\qquad \left. \geq \inf_{\mathbf{h}\in H_{1n},\mathbf{h}/\mathbf{h}^*\notin N_1(\varepsilon)} n^{4/5}[AMSE_{1n}(\mathbf{h}) - AMSE_{1n}(\mathbf{h}^*)]\right\} \\
&\leq \Pr\left\{2\sup_{\mathbf{h}\in H_{1n}} n^{4/5}\left[\widehat{MMSE}_n(\mathbf{h}) - AMSE_{1n}(\mathbf{h})\right] > \delta_\varepsilon\right\}
\end{aligned} \tag{49}$$

for some $\delta_\varepsilon > 0$ where the last inequality follows since for any $\varepsilon > 0$ there exists $\delta_\varepsilon > 0$ such that

$$\inf_{\mathbf{h}\in H_{1n},\mathbf{h}/\mathbf{h}^*\notin N_1(\varepsilon)} n^{4/5} AMSE_{1n}(\mathbf{h}) > n^{4/5} AMSE_{1n}(\mathbf{h}^*) + \delta_\varepsilon$$

by Lemma 2.2 and (12). For $\mathbf{h} \in H_{1n}$ we observe that by consistency of $\hat{f}(\cdot)$, $\hat{f}^{(2)}(\cdot)$ and $\hat{f}^{(4)}(\cdot)$,

$$\begin{aligned}
&n^{4/5}\left[\widehat{MMSE}_n(\mathbf{h}) - AMSE_{1n}(\mathbf{h})\right] \\
&\quad = n^{4/5}\cdot\frac{\mu_2^2}{2}\cdot\left[\bar{f}_{21}h_1^2 - \bar{f}_{22}h_2^2\right]\left\{h_1^2\left[\hat{f}^{(2)}(x_1) - f^{(2)}(x_1)\right] - h_2^2\left[\hat{f}^{(2)}(x_2) - f^{(2)}(x_2)\right]\right\} \\
&\quad + n^{4/5}\left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)h_1^4 - \hat{f}^{(4)}(x_2)h_2^4\right]\right\}^2 + \frac{\nu_0}{n^{1/5}}\left[\frac{\hat{f}(x_1) - f(x_1)}{h_1} + \frac{\hat{f}(x_2) - f(x_2)}{h_2}\right]
\end{aligned} \tag{50}$$

converges uniformly to zero in probability where $\bar{f}_{21}$ and $\bar{f}_{22}$ lies between $f^{(2)}(x_1)$ and $\hat{f}^{(2)}(x_1)$ and between $f^{(2)}(x_2)$ and $\hat{f}^{(2)}(x_2)$, respectively. Then the last expression in (49) converges to 0 by the uniform convergence of (50), proving (47) .

Next, we prove (48). Note that

$$\begin{aligned}
&\Pr\left\{\inf_{\mathbf{h}\in H_{2n}} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^*) \le 0\right\} \\
&\quad = \Pr\left\{\inf_{\mathbf{h}\in H_{2n}} \widehat{MMSE}_n(\mathbf{h}) \le \widehat{MMSE}_n(\mathbf{h}^*) - AMSE_{1n}(\mathbf{h}^*) + AMSE_{1n}(\mathbf{h}^*)\right\} \\
&\quad = \Pr\left\{\inf_{\mathbf{h}\in H_{2n}} n^{4/5}\widehat{MMSE}_n(\mathbf{h}) \le n^{4/5}\left[\widehat{MMSE}_n(\mathbf{h}^*) - AMSE_{1n}(\mathbf{h}^*)\right] + n^{4/5}AMSE_{1n}(\mathbf{h}^*)\right\}
\end{aligned}$$

Without loss of generality we can represent $\mathbf{h} = (h_1, h_2) \in H_{2n}$ by

$$h_1 = \theta_1 n^{-1/9} + o\left(n^{-1/9}\right), \quad h_2 = \theta_2 n^{-1/9} + o\left(n^{-1/9}\right)$$

for some $\theta_1 > 0$ and $\theta_2 > 0$. Using these, we get

$$\begin{aligned}
\widehat{MMSE}_n(\mathbf{h}) &= n^{-\frac{4}{9}}\left\{\frac{\mu_2}{2}\left[\hat{f}^{(2)}(x_1)\theta_1^2 - \hat{f}^{(2)}(x_2)\theta_2^2\right]\right\}^2 \\
&\quad + n^{-\frac{8}{9}}\left\{\left(\frac{\mu_4}{4!}\right)^2\left[\hat{f}^{(4)}(x_1)\theta_1^4 - \hat{f}^{(4)}(x_2)\theta_2^4\right]^2 + \nu_0\left[\frac{\hat{f}(x_1)}{\theta_1} + \frac{\hat{f}(x_2)}{\theta_2}\right]\right\} + o_p\left(n^{-4/9}\right). \qquad (51)
\end{aligned}$$

Because $\hat{f}(\cdot)$, $\hat{f}^{(2)}(\cdot)$ and $\hat{f}^{(4)}(\cdot)$ are consistent and $f^{(2)}(x_1)f^{(2)}(x_2) < 0$, $\inf_{\mathbf{h}\in H_{2n}} n^{4/9}\widehat{MMSE}_n(\mathbf{h})$ converges to a positive constant in probability. Combining this with (12) and the uniform convergence of (50) to zero in probability proves

$$\Pr\left\{n^{16/45}\inf_{\mathbf{h}\in H_{2n}} n^{4/9}\widehat{MMSE}_n(\mathbf{h}) \le n^{4/5}\left[\widehat{MMSE}_n(\mathbf{h}^*) - AMSE_{1n}(\mathbf{h}^*)\right] + n^{4/5}AMSE_{1n}(\mathbf{h}^*)\right\}$$

converges to zero, finishing the proof for (48).

Next, we turn to the case of $f^{(2)}(x_1)f^{(2)}(x_2) > 0$. The same argument as the previous case shows that it suffices to show

$$\Pr\left\{\inf_{\mathbf{h}\in H_{2n}, \mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon)} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^{**}) \le 0\right\} \to 0 \qquad (52)$$

and

$$\Pr\left\{\inf_{\mathbf{h}\in H_{1n}} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^{**}) \le 0\right\} \to 0 \qquad (53)$$

as n goes to infinity. The proof of (52) proceeds in a substantially different manner as that of (47). Notice that

$$\begin{aligned}
&\Pr\left\{\inf_{\mathbf{h}\in H_{2n}, \mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon)} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^{**}) \le 0\right\} \\
&\quad = \Pr\Bigg\{\inf_{\mathbf{h}\in H_{2n}, \mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon), h_1/h_2 \neq h_1^{**}/h_2^{**}} n^{8/9}\widehat{MMSE}_n(\mathbf{h}) \\
&\qquad\qquad \le n^{8/9}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] + n^{8/9}AMSE_{2n}(\mathbf{h}^{**})\Bigg\} \\
&\quad + \Pr\Bigg\{\inf_{\mathbf{h}\in H_{2n}, \mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon), h_1/h_2 = h_1^{**}/h_2^{**}} n^{8/9}\widehat{MMSE}_n(\mathbf{h}) \\
&\qquad\qquad \le n^{8/9}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] + n^{8/9}AMSE_{2n}(\mathbf{h}^{**})\Bigg\}. \qquad (54)
\end{aligned}$$

We show that both terms on the right-hand side of (54) converges to zero. We first show that

$$n^{8/9}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] \to 0 \tag{55}$$

in probability. Using the definition of $\mathbf{h}^{**}$ given in (11), we obtain that

$$\begin{aligned}
&n^{8/9}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] \\
&\quad = n^{4/9}\left(\frac{\mu_2\theta^{**2}}{2}\right)^2\left\{\left[\hat{f}^{(2)}(x_1) - f^{(2)}(x_1)\right] - \lambda^{**2}\left[\hat{f}^{(2)}(x_2) - f^{(2)}(x_2)\right]\right\}^2 \\
&\qquad + 2\left(\frac{\mu_4\theta^{**4}}{4!}\right)^2\left[\bar{f}_{41} - \lambda^{**4}\bar{f}_{42}\right]\left\{\left[\hat{f}^{(4)}(x_1) - f^{(4)}(x_1)\right] - \lambda^{**4}\left[\hat{f}^{(4)}(x_2) - f^{(4)}(x_2)\right]\right\} \\
&\qquad + \frac{\nu_0}{\theta^{**}}\left\{\left[\hat{f}(x_1) - f(x_1)\right] + \frac{1}{\lambda^{**}}\left[\hat{f}(x_2) - f(x_2)\right]\right\}
\end{aligned}$$

where the equality follows from the fact that $f^{(2)}(x_1)h_1^{**} - f^{(2)}(x_2)h_2^{**} = 0$ and the Taylor expansion with $\bar{f}_{41}$ and $\bar{f}_{42}$ lying between $\hat{f}^{(4)}(x_1)$ and $f^{(4)}(x_1)$ and between $\hat{f}^{(4)}(x_2)$ and $f^{(4)}(x_2)$, respectively. Then the result (55) follows since $\hat{f}(\cdot)$ and $\hat{f}^{(4)}(\cdot)$ are consistent estimators of $f(\cdot)$ and $f^{(4)}(\cdot)$ and since $\hat{f}^{(2)}(\cdot) - f^{(2)}(\cdot) = o_p\left(n^{-2/9}\right)$ by the assumptions of Theorem 2.1. Next, we get by (51) that

$$\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon),h_1/h_2\neq h_1^{**}/h_2^{**}} n^{4/9}\widehat{MMSE}_n(\mathbf{h}) = O_p(1) \tag{56}$$

since the condition $h_1/h_2 \neq h_1^{**}/h_2^{**}$ implies that the first term on the right-hand side of (51) does not vanish and dominates the remaining terms. The first term on the right-hand side of (54) converges to zero from (13), (55) and (56). Define

$$\widehat{AMSE}_{2n}(\mathbf{h}) = \left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)h_1^4 - \hat{f}^{(4)}(x_2)h_2^4\right]\right\}^2 + \frac{\nu_0}{n}\left[\frac{\hat{f}(x_1)}{h_1} + \frac{\hat{f}(x_2)}{h_2}\right].$$

Then we have for the second term of the right-hand side of (54) that

$$\begin{aligned}
&\Pr\Bigg\{\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon),h_1/h_2=h_1^{**}/h_2^{**}} n^{8/9}\widehat{MMSE}_n(\mathbf{h}) \\
&\qquad \le n^{8/9}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] + n^{8/9}AMSE_{2n}(\mathbf{h}^{**})\Bigg\} \\
&\le \Pr\Bigg\{\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon),h_1/h_2=h_1^{**}/h_2^{**}} n^{8/9}\widehat{AMSE}_{2n}(\mathbf{h}) \\
&\qquad \le n^{8/9}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] + n^{8/9}AMSE_{2n}(\mathbf{h}^{**})\Bigg\} \\
&\le \Pr\Bigg\{-\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon),h_1/h_2=h_1^{**}/h_2^{**}} n^{8/9}\left[\widehat{AMSE}_{2n}(\mathbf{h}) - AMSE_{2n}(\mathbf{h})\right] \\
&\qquad + n^{8/9}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] > \delta_\varepsilon\Bigg\}
\end{aligned} \tag{57}$$

where the last inequality follows since for any $\varepsilon > 0$ there exists $\delta_\varepsilon > 0$ such that

$$\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon),h_1/h_2=h_1^{**}/h_2^{**}} n^{8/9}AMSE_{2n}(\mathbf{h}) > n^{8/9}AMSE_{2n}(\mathbf{h}^{**}) + \delta_\varepsilon$$

by the definition of $\mathbf{h}^{**}$. Then the last object in (57) converges to zero by (55) and the uniform convergence of $n^{8/9}\widehat{AMSE}_{2n}(\mathbf{h})$ to $n^{8/9}AMSE_{2n}(\mathbf{h})$ in probability, proving the claim (52). This completes the proof for the second part of Theorem 2.1 (i).

It remains to prove (53). Observe that

$$\inf_{\mathbf{h}\in H_{1n}} n^{8/9}\widehat{MMSE}_n(\mathbf{h}) \geq \inf_{\mathbf{h}\in H_{1n}} \frac{\nu_0}{n^{1/9}}\left[\frac{\hat{f}(x_1)}{h_1}+\frac{\hat{f}(x_2)}{h_2}\right]$$
$$= n^{4/45}\left\{\frac{\nu_0}{\delta_{12}}\left[\hat{f}(x_1)+\hat{f}(x_2)\right]+o_p(1)\right\}. \tag{58}$$

Then it follows from (13), (55) and (58) that

$$\Pr\left\{\inf_{\mathbf{h}\in H_{2n}} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^{**}) \leq 0\right\}$$
$$= \Pr\left\{\inf_{\mathbf{h}\in H_{2n}} n^{8/9}\widehat{MMSE}_n(\mathbf{h}) \leq n^{8/9}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] + n^{8/9}AMSE_{2n}(\mathbf{h}^{**})\right\}$$
$$\to 0.$$

This complete the first assertion of Theorem 2.1.

Next we prove (ii). When $f^{(2)}(x_1)f^{(2)}(x_2) < 0$, we show

$$\frac{\widehat{MMSE}(\hat{h}_1,\hat{h}_2) - MSE_n(h_1^*,h_2^*)}{MSE_n(h_1^*,h_2^*)} \to 0$$

in probability. We have by the definition of the MSE given in (5) and the AMSE given in (8) that

$$n^{4/5}MSE_n(h_1^*,h_2^*) = n^{4/5}AMSE_{1n}(h_1^*,h_2^*) + O\left(n^{-4/9}\right). \tag{59}$$

Given the property of the $AMSE_{1n}(h_1^*,h_2^*)$ in (12), it suffices to show

$$n^{4/5}\left[\widehat{MMSE}(\hat{h}_1,\hat{h}_2) - MSE_n(h_1^*,h_2^*)\right] \to 0$$

in probability. Notice that

$$n^{4/5}\left[\widehat{MMSE}(\hat{h}_1,\hat{h}_2) - MSE_n(h_1^*,h_2^*)\right]$$
$$= n^{4/5}\left[\widehat{MMSE}(\hat{h}_1,\hat{h}_2) - \widehat{MMSE}_n(h_1^*,h_2^*)\right] + n^{4/5}\left[\widehat{MMSE}(h_1^*,h_2^*) - MSE_n(h_1^*,h_2^*)\right]$$

and the second term on the right-hand side converges to zero in probability from the uniform convergence of (50) and (59). Hence, it remains to prove that the first term on the right-hand side converges to zero in probability. By the Taylor expansion, we obtain that

$$n^{4/5}\left[\widehat{MMSE}(\hat{h}_1,\hat{h}_2) - \widehat{MMSE}_n(h_1^*,h_2^*)\right]$$
$$= n^{4/5}\mu^2\hat{f}^{(2)}(x_1)\bar{h}_1\left[\hat{f}^{(2)}(x_1)\bar{h}_1^2 - \hat{f}^{(2)}(x_2)\bar{h}_2^2\right]\left(\hat{h}_1 - h_1^*\right)$$
$$- n^{4/5}\mu^2\hat{f}^{(2)}(x_2)\bar{h}_2\left[\hat{f}^{(2)}(x_1)\bar{h}_1^2 - \hat{f}^{(2)}(x_2)\bar{h}_2^2\right]\left(\hat{h}_2 - h_2^*\right)$$
$$+ n^{4/5}\left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)\hat{h}_1^4 - \hat{f}^{(4)}(x_2)\hat{h}_2^4\right]\right\}^2 - n^{4/5}\left\{\frac{\mu_4}{4!}\left[\hat{f}^{(4)}(x_1)h_1^{*4} - \hat{f}^{(4)}(x_2)h_2^{*4}\right]\right\}^2$$
$$+ \frac{\nu_0}{n^{1/5}}\left[\frac{\hat{f}(x_1)}{h_1^*}+\frac{\hat{f}(x_2)}{h_2^*}\right]\left(\frac{h_*}{\hat{h}_1}-1\right)$$

where $\bar{h}_1$ and $\bar{h}_2$ lie between $\hat{h}_1$ and h_1^* and between $\hat{h}_2$ and h_2^*, respectively. All terms on the right-hand side converges to zero in probability from the result in the first part of Theorem 2.1 and

the fact that both h_1^* and h_2^* are the order of $n^{-1/5}$ as shown in Lemma 2.2, proving the result for the case where $f^{(2)}(x_1)f^{(2)}(x_2) < 0$.

Next, we prove (ii) when $f^{(2)}(x_1)f^{(2)}(x_2) > 0$. Given $MSE_n(h_1^{**}, h_2^{**}) = AMSE_{2n}(h_1^{**}, h_2^{**})$ and (13), it suffices to show that

$$n^{8/9}\left[\widehat{MMSE}_n(\hat{h}_1, \hat{h}_2) - AMSE_{2n}(h_1^{**}, h_2^{**})\right] \to 0$$

in probability. Observe that

$$\begin{aligned}
n^{8/9}\left[\widehat{MMSE}_n(\hat{h}_1, \hat{h}_2) - AMSE_{2n}(h_1^{**}, h_2^{**})\right] &= n^{8/9}\left\{\frac{\mu_2}{2}\left[\hat{f}^{(2)}(x_1)\hat{h}_1^2 - \hat{f}^{(2)}(x_2)\hat{h}_2^2\right]\right\}^2 \\
&+ 2n^{8/9}\left(\frac{\mu_4}{4!}\right)^2\left[\bar{f}_{41}\hat{h}_1^4 - \bar{f}_{42}\hat{h}_2^4\right]\left\{\hat{h}_1^4\left[\hat{f}^{(4)}(x_1) - f^{(4)}(x_1)\right] - \hat{h}_2^4\left[\hat{f}^{(4)}(x_2) - f^{(4)}(x_2)\right]\right\} \\
&+ 8n^{8/9}\left(\frac{\mu_4}{4!}\right)^2\left[f^{(4)}(x_1)\bar{h}_1^4 - f^{(4)}(x_2)\bar{h}_2^4\right]\left[f^{(4)}(x_1)\bar{h}_1^3\left(\hat{h}_1 - h_1^{**}\right) - f^{(4)}(x_2)\bar{h}_2^3\left(\hat{h}_2 - h_2^{**}\right)\right] \\
&+ \frac{\nu_0}{n^{1/9}}\left\{\left[\frac{\hat{f}(x_1) - f(x_1)}{\hat{h}_1} + \frac{\hat{f}(x_2) - f(x_2)}{\hat{h}_2}\right] + \left[\frac{f(x_1)}{h_1^{**}}\left(\frac{h_1^{**}}{\hat{h}_1} - 1\right) + \frac{f(x_2)}{h_2^{**}}\left(\frac{h_2^{**}}{\hat{h}_2} - 1\right)\right]\right\} \quad (60)
\end{aligned}$$

where $\bar{h}_1$ and $\bar{h}_2$ lie between $\hat{h}_1$ and h_1^{**} and between $\hat{h}_1$ and h_1^{**}, respectively, and $\bar{f}_{41}$ and $\bar{f}_{42}$ are as before. All terms except the first on the right-hand side of (60) converge to zero in probability by consistency of $\hat{f}(\cdot)$ and $\hat{f}^{(4)}(\cdot)$ and the first part of Theorem 2.1. It remains to show the first term converges to zero in probability. Notice that the first term of the right-hand side of (60) equals

$$\begin{aligned}
n^{8/9}\left(\frac{\mu_2}{2}\right)^2\Big\{&\left[\hat{f}^{(2)}(x_1) - f^{(2)}(x_1)\right]\hat{h}_1^2 - \left[\hat{f}^{(2)}(x_2) - f^{(2)}(x_2)\right]\hat{h}_2^2 \\
&+ f^{(2)}(x_1)\left(\hat{h}_1^2 - h_1^{**2}\right) - f^{(2)}(x_2)\left(\hat{h}_2^2 - h_2^{**2}\right)\Big\}^2.
\end{aligned}$$

Then by the assumption that $\hat{f}^{(2)}(x_j) - f^{(2)}(x_j) = o_p\left(n^{-2/9}\right)$ for $j = 1, 2$ and the first part of Theorem 2.1, it is enough to show

$$n^{8/9}\left[f^{(2)}(x_1)\left(\hat{h}_1^2 - h_1^{**2}\right) - f^{(2)}(x_2)\left(\hat{h}_2^2 - h_2^{**2}\right)\right]^2 = o_p(1),$$

or

$$f^{(2)}(x_1)\left(\hat{h}_1^2 - h_1^{**2}\right) - f^{(2)}(x_2)\left(\hat{h}_2^2 - h_2^{**2}\right) = o_p\left(n^{-4/9}\right). \quad (61)$$

It is easily seen from the first part of Theorem 2.1 that

$$\left.\frac{\partial\widehat{AMSE}_{2n}(\mathbf{h})}{\partial h_1}\right|_{\mathbf{h}=\hat{\mathbf{h}}} = O_p\left(n^{-7/9}\right) \quad (62)$$

since

$$\frac{\partial\widehat{AMSE}_n(\mathbf{h})}{\partial h_1} = 8\left(\frac{\mu_4}{4!}\right)^2\hat{f}^{(4)}(x_1)h_1^3\left[\hat{f}^{(4)}(x_1)h_1^4 - \hat{f}^{(4)}(x_2)h_2^4\right] - \frac{\nu_0}{n}\frac{\hat{f}(x_1)}{h_1^2}.$$

Define

$$\widehat{F}_n(\mathbf{h}) = \left\{\frac{\mu_2}{2}\left[\hat{f}^{(2)}(x_1)h_1^2 - \hat{f}^{(2)}(x_2)h_2^2\right]\right\}^2.$$

Then it follows that

$$\begin{aligned}
\left.\frac{\partial\widehat{F}_n(\mathbf{h})}{\partial h_1}\right|_{\mathbf{h}=\hat{\mathbf{h}}} &= \mu_2^2\hat{f}^{(2)}(x_1)\hat{h}_1\left[\hat{f}^{(2)}(x_1)\hat{h}_1^2 - \hat{f}^{(2)}(x_2)\hat{h}_2^2\right] \\
&= \mu_2^2\hat{f}^{(2)}(x_1)\hat{h}_1\left\{\left[\hat{f}^{(2)}(x_1) - f^{(2)}(x_1)\right]\hat{h}_1^2 - \left[\hat{f}^{(2)}(x_2) - f^{(2)}(x_2)\right]\hat{h}_2^2\right\} \\
&\quad + \mu_2^2\hat{f}^{(2)}(x_1)\hat{h}_1\left[f^{(2)}(x_1)\hat{h}_1^2 - f^{(2)}(x_2)\hat{h}_2^2\right] \quad (63)
\end{aligned}$$

By the assumption that $\hat{f}^{(2)}(x_j) - f^{(2)}(x_j) = o_p\left(n^{-2/9}\right)$ for $j = 1, 2$ and the first part of Theorem 2.1, we get

$$\mu_2^2 \hat{f}^{(2)}(x_1)\hat{h}_1 \left\{ \left[\hat{f}^{(2)}(x_1) - f^{(2)}(x_1)\right] \hat{h}_1^2 - \left[\hat{f}^{(2)}(x_2) - f^{(2)}(x_2)\right] \hat{h}_2^2 \right\} = o_p\left(n^{-5/9}\right). \tag{64}$$

We obtain by the Taylor expansion that

$$f^{(2)}(x_1)\hat{h}_1^2 - f^{(2)}(x_2)\hat{h}_2^2 = \bar{h}_1 f^{(2)}(x_1)\left(\hat{h}_1 - h_1^{**}\right) + \bar{h}_2 f^{(2)}(x_2)\left(\hat{h}_2 - h_2^{**}\right) \tag{65}$$

where the equality follows from the fact that h_1^{**} and h_2^{**} satisfy $f^{(2)}(x_1)h_1^{**} - f^{(2)}(x_2)h_2^{**} = 0$ and $\bar{h}_1$ and $\bar{h}_2$ are as defined above. Observe that

$$\widehat{MMSE}_n(\mathbf{h}) = \widehat{F}_n(\mathbf{h}) + \widehat{AMSE}_{2n}(\mathbf{h}).$$

Then it follows by the definition of $\hat{\mathbf{h}}$ that

$$0 = \left.\frac{\partial \widehat{MMSE}_n(\mathbf{h})}{\partial h_1}\right|_{\mathbf{h}=\hat{\mathbf{h}}} = \left.\frac{\partial \widehat{F}_n(\mathbf{h})}{\partial h_1}\right|_{\mathbf{h}=\hat{\mathbf{h}}} + \left.\frac{\partial \widehat{AMSE}_{2n}(\mathbf{h})}{\partial h_1}\right|_{\mathbf{h}=\hat{\mathbf{h}}}.$$

Combining this relationship with (62), (63), (64) and (65) proves (61). This complete the proof of (ii) when $f^2(x_1)f^{(2)}(x_2) > 0$.

Proof of Lemma 3.1: A contribution to the MSE from a variance component is standard (see Fan and Gijbels 1996). Here we derive a contribution from the bias component. Denote

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\alpha}_h(x) \\ \hat{\beta}_h(x) \end{bmatrix}.$$

The conditional bias is given by

$$\text{Bias}(\hat{\boldsymbol{\beta}}|X) = (X(x)'W(x)X(x))^{-1}X(x)W(x)(m - X(x)\boldsymbol{\beta})$$

where $m = (m(X_1), \ldots, m(X_n))'$ and $\boldsymbol{\beta} = (m(x), m^{(1)}(x))'$. Let

$$s_{n,j} = \sum_{t=1}^{n} K_h(X_i - x)(X_i - x)^j.$$

We will use the following notations.

$$S_{n,j} = \begin{bmatrix} s_{n,j} & s_{n,j+1} \\ s_{n,j+1} & s_{n,j+2} \end{bmatrix}, \quad S_j = \begin{bmatrix} \mu_j & \mu_{j+1} \\ \mu_{j+1} & \mu_{j+2} \end{bmatrix}, \quad c_{n,j} = \begin{bmatrix} s_{n,j} \\ s_{n,j+1} \end{bmatrix}, \quad c_j = \begin{bmatrix} \mu_j \\ \mu_{j+1} \end{bmatrix} \tag{66}$$

Note that $S_{n,0} = X(x)'W(x)X(x)$. The argument in Fan et al. (1996) can be generalized to yield

$$s_{n,j} = nh^j \left\{ f(x)\mu_j + hf^{(1)}(x)\mu_{j+1} + \frac{h^2 f^{(2)}(x)}{2}\mu_{j+2} + o_p\left(h^2\right) \right\}. \tag{67}$$

Then it follows

$$S_{n,0} = nH \left\{ f(x)S_0 + hf^{(1)}(x)S_1 + \frac{h^2 f^{(2)}(x)}{2}S_2 + o_p\left(h^2\right) \right\} H$$

where $H = \text{diag}(1, h)$. Using the fact that

$$(A + hB + h^2C)^{-1} = A^{-1} - hA^{-1}BA^{-1} - h^2A^{-1}CA^{-1} + h^2A^{-1}BA^{-1}BA^{-1} + o\left(h^2\right),$$

we obtain

$$S_{n,0}^{-1} = n^{-1}H^{-1}\left\{\frac{1}{f(x)}A_0 - \frac{hf^{(1)}(x)}{f(x)^2}A_1 - \frac{h^2f^{(2)}(x)}{2f(x)^2}A_2 + \frac{h^2f^{(1)}(x)^2}{f(x)^3}A_3 + o_p\left(h^2\right)\right\}H^{-1} \tag{68}$$

where

$$A_0 = \begin{bmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} \mu_2 & 0 \\ 0 & \mu_4/\mu_2^2 \end{bmatrix}, \quad A_1 = \begin{bmatrix} \mu_2 & 0 \\ 0 & 1 \end{bmatrix}$$

Here the matrix structure is simplified considerably by a symmetric kernel function.

Next, we consider $X(x)W(x)(m - X(x)\boldsymbol{\beta})$. We obtain from Taylor expansion of $m(\cdot)$

$$X(x)W(x)(m - X(x)\boldsymbol{\beta}) = \frac{m^{(2)}(x)}{2}c_{n,2} + \frac{m^{(3)}(x)}{3!}c_{n,3} + \frac{m^{(4)}(x)}{4!}c_{n,4} + o_p\left(nh^4\right). \tag{69}$$

The definition of $c_{n,j}$ in (71) with (72) yields

$$c_{n,j} = nh^jH\left\{f(x)c_j + hf^{(1)}(x)c_{j+1} + \frac{h^2f^{(2)}(x)}{2}c_{j+2} + o_p\left(h^2\right)\right\} \tag{70}$$

Combining (73) with (74) and (75) and extracting the first element gives

$$\text{Bias}(\hat{\alpha}_h(x)|X) = \frac{h^2m^{(2)}(x)}{2}\mu_2 + \frac{h^4}{4}\left\{\frac{m^{(2)}(x)}{f(x)^2}(\mu_4 - \mu_2)\left(f^{(2)}(x)f(x) - f^{(1)}(x)^2\right) + \frac{m^{(4)}(x)}{3!}\mu_4\right\}.$$

This expression together with the argument used to show Lemma 2.1 gives the required result.

Proof of Lemma 3.3: Again, we consider a contribution only from the bias component since the variance component is standard. Denote

$$\hat{\boldsymbol{\beta}}_\ell = \begin{bmatrix} \hat{\alpha}_{h,\ell}(x) \\ \hat{\beta}_{h,\ell}(x) \end{bmatrix}.$$

and let $\boldsymbol{\beta}_\ell$ be the population analogue (check). The conditional bias is given by

$$\text{Bias}(\hat{\boldsymbol{\beta}}_\ell|X) = (X(x)'W(x)X(x))^{-1}X(x)W(x)(m_1 - X(x)\boldsymbol{\beta}_\ell)$$

where $m = (m(X_1), \ldots, m(X_n))'$ and $\boldsymbol{\beta} = (m(x), m^{(1)}(x))'$. Let

$$s_{n,j} = \sum_{t=1}^{n} K_h(X_i - x)(X_i - x)^j.$$

We will use the following notations.

$$S_{n,j} = \begin{bmatrix} s_{n,j} & s_{n,j+1} \\ s_{n,j+1} & s_{n,j+2} \end{bmatrix}, \quad S_j = \begin{bmatrix} \mu_j & \mu_{j+1} \\ \mu_{j+1} & \mu_{j+2} \end{bmatrix}, \quad c_{n,j} = \begin{bmatrix} s_{n,j} \\ s_{n,j+1} \end{bmatrix}, \quad c_j = \begin{bmatrix} \mu_j \\ \mu_{j+1} \end{bmatrix} \tag{71}$$

Note that $S_{n,0} = X(x)'W(x)X(x)$. The argument in Fan et al. (1996) can be generalized to yield

$$s_{n,j} = nh^j\left\{f(x)\mu_j + hf^{(1)}(x)\mu_{j+1} + \frac{h^2f^{(2)}(x)}{2}\mu_{j+2} + o_p\left(h^2\right)\right\}. \tag{72}$$

Then it follows

$$S_{n,0} = nH\left\{f(x)S_0 + hf^{(1)}(x)S_1 + \frac{h^2f^{(2)}(x)}{2}S_2 + o_p\left(h^2\right)\right\}H$$

where $H = \mathrm{diag}(1, h)$. Using the fact that

$$(A + hB + h^2C)^{-1} = A^{-1} - hA^{-1}BA^{-1} - h^2A^{-1}CA^{-1} + h^2A^{-1}BA^{-1}BA^{-1} + o\left(h^2\right),$$

we obtain

$$S_{n,0}^{-1} = n^{-1}H^{-1}\left\{\frac{1}{f(x)}A_0 - \frac{hf^{(1)}(x)}{f(x)^2}A_1 - \frac{h^2f^{(2)}(x)}{2f(x)^2}A_2 + \frac{h^2f^{(1)}(x)^2}{f(x)^3}A_3 + o_p\left(h^2\right)\right\}H^{-1} \tag{73}$$

where

$$A_0 = \begin{bmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} \mu_2 & 0 \\ 0 & \mu_4/\mu_2^2 \end{bmatrix}, \quad A_1 = \begin{bmatrix} \mu_2 & 0 \\ 0 & 1 \end{bmatrix}$$

Here the matrix structure is simplified considerably by a symmetric kernel function.

Next, we consider $X(x)W(x)(m - X(x)\boldsymbol{\beta})$. We obtain from Taylor expansion of $m(\cdot)$

$$X(x)W(x)(m - X(x)\boldsymbol{\beta}) = \frac{m^{(2)}(x)}{2}c_{n,2} + \frac{m^{(3)}(x)}{3!}c_{n,3} + \frac{m^{(4)}(x)}{4!}c_{n,4} + o_p\left(nh^4\right). \tag{74}$$

The definition of $c_{n,j}$ in (71) with (72) yields

$$c_{n,j} = nh^jH\left\{f(x)c_j + hf^{(1)}(x)c_{j+1} + \frac{h^2f^{(2)}(x)}{2}c_{j+2} + o_p\left(h^2\right)\right\} \tag{75}$$

Combining (73) with (74) and (75) and extracting the first element gives

$$\mathrm{Bias}(\hat{\alpha}_h(x)|X) = \frac{h^2m^{(2)}(x)}{2}\mu_2 + \frac{h^4}{4}\left\{\frac{m^{(2)}(x)}{f(x)^2}(\mu_4 - \mu_2)\left(f^{(2)}(x)f(x) - f^{(1)}(x)^2\right) + \frac{m^{(4)}(x)}{3!}\mu_4\right\}.$$

This expression together with the argument used to show Lemma 2.1 gives the required result.

Proof of Theorem 3.2: (i) When $m_1^{(2)}(x)m_2^{(2)}(x) < 0$, the same discussion provided in the proof of Theorem 2.1 shows that it suffices to show

$$\Pr\left\{\inf_{\mathbf{h}\in H_{1n},\mathbf{h}/\mathbf{h}^*\notin N_1(\varepsilon)}\widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^*) \le 0\right\} \to 0 \tag{76}$$

and

$$\Pr\left\{\inf_{\mathbf{h}\in H_{2n}}\widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^*) \le 0\right\} \to 0 \tag{77}$$

as n goes to infinity. The claim (76) can be proved in the same manner as (47). To prove (77), we represent $\mathbf{h} = (h_1, h_2) \in H_{2n}$ by

$$h_1 = \theta_1 n^{-1/7} + o\left(n^{-1/7}\right), \quad h_2 = \theta_2 n^{-1/7} + o\left(n^{-1/7}\right)$$

for some $\theta_1 > 0$ and $\theta_2 > 0$. With these we can write $\widehat{MMSE}_n(\mathbf{h})$ as

$$\begin{aligned}\widehat{MMSE}_n(\mathbf{h}) = {} & n^{-\frac{4}{7}}\left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(x)\theta_1^2 - \hat{m}_2^{(2)}(x)\theta_2^2\right]\right\}^2 \\ & + n^{-\frac{6}{7}}\left\{\left[\hat{b}_{2,1}(x)\theta_1^3 - \hat{b}_{2,2}(x)\theta_2^3\right]^2 + \frac{v}{\hat{f}(x)}\left[\frac{\hat{\sigma}_1^2(x)}{\theta_1} + \frac{\hat{\sigma}_2^2(x)}{\theta_2}\right]\right\} + o_p\left(n^{-4/7}\right).\end{aligned} \tag{78}$$

Because $\hat{m}_1^{(2)}(x)$, $\hat{m}_2^{(2)}(x)$, $\hat{b}_{2,1}(x)$, $\hat{b}_{2,2}(\cdot)$, $\hat{f}(\cdot)$, $\sigma_1^2(x)$ and $\sigma_2^2(x)$ are consistent and $m_1^{(2)}(x)m_2^{(2)}(x) < 0$, $\inf_{\mathbf{h}\in H_{2n}} n^{4/7}\widehat{MMSE}_n(\mathbf{h})$ converges to a positive constant in probability. Then (77) follows since

$$\Pr\left\{n^{8/35}\inf_{\mathbf{h}\in H_{2n}} n^{4/7}\widehat{MMSE}_n(\mathbf{h}) \le n^{4/5}\left[\widehat{MMSE}_n(\mathbf{h}^*) - AMSE_{1n}(\mathbf{h}^*)\right] + n^{4/5}AMSE_{1n}(\mathbf{h}^*)\right\}$$

converges to zero by (12) and the uniform convergence of $n^{4/5}\left[\widehat{MMSE}_n(\mathbf{h}^*) - AMSE_{1n}(\mathbf{h}^*)\right]$ to zero in probability.

Next, consider the case of $m_1^{(2)}(x)m_2^{(2)}(x) > 0$. As in the proof of Theorem 2.1, it suffices to show

$$\Pr\left\{\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon)} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^{**}) \le 0\right\} \to 0 \tag{79}$$

and

$$\Pr\left\{\inf_{\mathbf{h}\in H_{1n}} \widehat{MMSE}_n(\mathbf{h}) - \widehat{MMSE}_n(\mathbf{h}^{**}) \le 0\right\} \to 0 \tag{80}$$

as n goes to infinity. The proof of (79) can be carried out in the same manner as that of (52). We prove (79) by showing

$$\begin{aligned}\Pr\Bigg\{&\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon),h_1/h_2\neq h_1^{**}/h_2^{**}} n^{6/7}\widehat{MMSE}_n(\mathbf{h}) \\ &\le n^{6/7}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] + n^{6/7}AMSE_{2n}(\mathbf{h}^{**})\Bigg\}\to 0\end{aligned} \tag{81}$$

and

$$\begin{aligned}\Pr\Bigg\{&\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon),h_1/h_2=h_1^{**}/h_2^{**}} n^{6/7}\widehat{MMSE}_n(\mathbf{h}) \\ &\le n^{6/7}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] + n^{6/7}AMSE_{2n}(\mathbf{h}^{**})\Bigg\}\to 0.\end{aligned} \tag{82}$$

Observe first that the argument used to obtain (55) yields

$$n^{6/7}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] \to 0 \tag{83}$$

in probability. Notice that for $\mathbf{h} \in H_{2n}$ we obtain by (78) that

$$\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon),h_1/h_2\neq h_1^{**}/h_2^{**}} n^{4/7}\widehat{MMSE}_n(\mathbf{h}) = O_p(1) \tag{84}$$

due to the argument used to get (56). Then (81) follows from (35), (83) and (84). Next we prove (82). The same argument used to obtain (57) shows that it suffices to show

$$\begin{aligned}\Pr\Bigg\{&-\inf_{\mathbf{h}\in H_{2n},\mathbf{h}/\mathbf{h}^{**}\notin N_1(\varepsilon),h_1/h_2=h_1^{**}/h_2^{**}} n^{6/7}\left[\widehat{AMSE}_{2n}(\mathbf{h}) - AMSE_{2n}(\mathbf{h})\right] \\ &+ n^{6/7}\left[\widehat{MMSE}_n(\mathbf{h}^{**}) - AMSE_{2n}(\mathbf{h}^{**})\right] > \delta_\varepsilon\Bigg\}.\end{aligned}$$

This follows from (83) and the uniform convergence of $n^{6/7}\widehat{AMSE}_{2n}(\mathbf{h})$ to $n^{6/7}AMSE_{2n}(\mathbf{h})$ in probability, proving the claim (82).

It remains to prove (80). Notice that

$$\begin{aligned}\inf_{\mathbf{h}\in H_{1n}} n^{6/7}\widehat{MMSE}_n(\mathbf{h}) &\ge \inf_{\mathbf{h}\in H_{1n}} \frac{v}{n^{1/7}\hat{f}(x)}\left[\frac{\hat{\sigma}_1^2(x)}{h_1} + \frac{\hat{\sigma}_2^2(x)}{h_2}\right] \\ &= n^{2/35}\left\{\frac{v}{\delta_{12}\hat{f}(x)}\left[\hat{\sigma}_1^2(x) + \hat{\sigma}_2^2(x)\right] + o_p(1)\right\}.\end{aligned} \tag{85}$$

Then (80) follows from (35), (83) and (85).

Now we prove (ii) of Theorem 3.2. When $m_1^{(2)}(x)m_2^{(2)}(x) < 0$, the proof is the same as that of Theorem 2.1 and it is omitted. To prove (ii) when $m_1^{(2)}(x)m_2^{(2)}(x) > 0$, it is enough to show

$$n^{6/7}\left[\widehat{MMSE}_n(\hat{h}_1,\hat{h}_2) - AMSE_{2n}(h_1^{**},h_2^{**})\right] \to 0$$

in probability. Observe that

$$\begin{aligned}
n^{6/7}&\left[\widehat{MMSE}_n(\hat{h}_1,\hat{h}_2) - AMSE_{2n}(h_1^{**},h_2^{**})\right] = n^{6/7}\left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(x)\hat{h}_1^2 - \hat{m}_2^{(2)}(x)\hat{h}_2^2\right]\right\}^2 \\
&+ 2n^{6/7}\left[\bar{b}_{2,1}\hat{h}_1^3 - \bar{b}_{2,2}\hat{h}_2^3\right]\left\{\hat{h}_1^3\left[\hat{b}_{2,1}(x) - b_{2,1}(x_1)\right] - \hat{h}_2^3\left[\hat{b}_{2,2}(x) - b_{2,2}(x)\right]\right\} \\
&+ 6n^{6/7}\left[b_{2,1}(x)\bar{h}_1^3 - b_{2,2}(x)\bar{h}_2^3\right]\left[b_{2,1}(x)\bar{h}_1^2\left(\hat{h}_1 - h_1^{**}\right) - b_{2,2}(x)\bar{h}_2^2\left(\hat{h}_2 - h_2^{**}\right)\right] \\
&+ \frac{v}{n^{1/7}}\left\{\frac{1}{\hat{f}(x)}\left[\frac{\hat{\sigma}_1^2(x)-\sigma_1^2(x)}{\hat{h}_1} + \frac{\hat{\sigma}_2^2(x)-\sigma_2^2(x)}{\hat{h}_2}\right] + \frac{1}{f(x)}\left[\frac{\sigma_1^2(x)}{h_1^{**}}\left(\frac{h_1^{**}}{\hat{h}_1} - 1\right) + \frac{\sigma_2^2(x)}{h_2^{**}}\left(\frac{h_2^{**}}{\hat{h}_2} - 1\right)\right]\right. \\
&\left. + \frac{1}{\bar{f}}\left[\frac{\sigma_1^2(x)}{\hat{h}_1} + \frac{\sigma_2^2(x)}{\hat{h}_2}\right]\left[\hat{f}(x) - f(x)\right]\right\} \qquad (86)
\end{aligned}$$

where $\bar{h}_1$ and $\bar{h}_2$ lie between $\hat{h}_1$ and h_1^{**} and between $\hat{h}_1$ and h_1^{**}, respectively, $\bar{b}_{2,1}$ and $\bar{b}_{2,2}$ lie between $\hat{b}_{2,1}(x)$ and $b_{2,1}(x)$ and between $\hat{b}_{2,2}(x)$ and $b_{2,2}(x)$, respectively, and $\bar{f}$ lies between $\hat{f}(x)$ and $f(x)$. All terms except the first on the right-hand side of (86) converge to zero in probability by consistency of $\hat{b}_{2,1}(x)$, $\hat{b}_{2,2}(x)$, and $\hat{\sigma}_1^2(x)$, $\hat{\sigma}_2^2(x)$ and $\hat{f}(x)$ and the first part of Theorem 3.2. To show that the first term of the right-hand side of (86) to zero in probability, note that it equals

$$\begin{aligned}
n^{6/7}\left(\frac{b_1}{2}\right)^2\left\{\left[\hat{m}_1^{(2)}(x) - m_1^{(2)}(x)\right]\hat{h}_1^2 - \left[\hat{m}_2^{(2)}(x) - m_2^{(2)}(x)\right]\hat{h}_2^2\right. \\
\left. + m_1^{(2)}(x)\left(\hat{h}_1^2 - h_1^{**2}\right) - m_2^{(2)}(x)\left(\hat{h}_2^2 - h_2^{**2}\right)\right\}^2.
\end{aligned}$$

Then by the assumption that $\hat{m}_j^{(2)}(x) - m_j^{(2)}(x) = o_p\left(n^{-1/7}\right)$ for $j = 1, 2$ and the first part of Theorem 2.1, it is enough to show

$$m_1^{(2)}(x)\left(\hat{h}_1^2 - h_1^{**2}\right) - m_2^{(2)}(x)\left(\hat{h}_2^2 - h_2^{**2}\right) = o_p\left(n^{-3/7}\right). \qquad (87)$$

It is easily seen from the first part of Theorem 2.1 that

$$\left.\frac{\partial\widehat{AMSE}_{2n}(\mathbf{h})}{\partial h_1}\right|_{\mathbf{h}=\hat{\mathbf{h}}} = O_p\left(n^{-5/7}\right) \qquad (88)$$

Define

$$\widehat{F}_n(\mathbf{h}) = \left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(x)h_1^2 - \hat{m}_2^{(2)}(x)h_2^2\right]\right\}^2.$$

Then it follows that

$$\begin{aligned}
\left.\frac{\partial\widehat{F}_n(\mathbf{h})}{\partial h_1}\right|_{\mathbf{h}=\hat{\mathbf{h}}} &= b_1^2\hat{m}_1^{(2)}(x)\hat{h}_1\left[\hat{m}_1^{(2)}(x)\hat{h}_1^2 - \hat{m}_2^{(2)}(x)\hat{h}_2^2\right] \\
&= b_1^2\hat{m}_1^{(2)}(x)\hat{h}_1\left\{\left[\hat{m}_1^{(2)}(x) - m_1^{(2)}(x)\right]\hat{h}_1^2 - \left[\hat{m}_2^{(2)}(x) - m_2^{(2)}(x)\right]\hat{h}_2^2\right\} \\
&\quad + b_1^2\hat{m}_1^{(2)}(x)\hat{h}_1\left[m_1^{(2)}(x)\hat{h}_1^2 - m_2^{(2)}(x)\hat{h}_2^2\right] \qquad (89)
\end{aligned}$$

By the assumption that $\hat{m}_j^{(2)}(x) - m_j^{(2)}(x) = o_p\left(n^{-1/7}\right)$ for $j = 1, 2$ and the first part of Theorem 2.1, we get

$$b_1^2 \hat{m}_1^{(2)}(x)\hat{h}_1 \left\{ \left[\hat{m}_1^{(2)}(x) - m_1^{(2)}(x)\right] \hat{h}_1^2 - \left[\hat{m}_2^{(2)}(x) - m_2^{(2)}(x)\right] \hat{h}_2^2 \right\} = o_p\left(n^{-4/7}\right). \tag{90}$$

We obtain by the Taylor expansion that

$$m_1^{(2)}(x)\hat{h}_1^2 - m_2^{(2)}(x)\hat{h}_2^2 = \bar{h}_1 m_1^{(2)}(x)\left(\hat{h}_1 - h_1^{**}\right) + \bar{h}_2 m_2^{(2)}(x)\left(\hat{h}_2 - h_2^{**}\right) \tag{91}$$

where the equality follows from the fact that h_1^{**} and h_2^{**} satisfy $m_1^{(2)}(x)h_1^{**} - m_2^{(2)}(x)h_2^{**} = 0$ and $\bar{h}_1$ and $\bar{h}_2$ are as defined above. Observe that

$$\widehat{MMSE}_n(\mathbf{h}) = \widehat{F}_n(\mathbf{h}) + \widehat{AMSE}_{2n}(\mathbf{h}).$$

Then it follows by the definition of $\hat{\mathbf{h}}$ that

$$0 = \left.\frac{\partial \widehat{MMSE}_n(\mathbf{h})}{\partial h_1}\right|_{\mathbf{h}=\hat{\mathbf{h}}} = \left.\frac{\partial \widehat{F}_n(\mathbf{h})}{\partial h_1}\right|_{\mathbf{h}=\hat{\mathbf{h}}} + \left.\frac{\partial \widehat{AMSE}_{2n}(\mathbf{h})}{\partial h_1}\right|_{\mathbf{h}=\hat{\mathbf{h}}}.$$

Combining this relationship with (88), (89), (90) and (91) proves (87). This complete the proof of (ii) when $m_1^2(x)m_2^{(2)}(x) > 0$.

Proof of Lemma 5.2: Denote

$$F_n(h_1, h_2) = \left[\frac{\mu_2}{2} f^{(2)}(x_1)h_1^2 - \frac{\mu_4}{4!} f^{(4)}(x_2)h_2^4\right]^2 + \nu_0 \left\{ \frac{f(x_1)}{nh_1} + \frac{f(x_2)}{nh_2} \right\}$$

The first order conditions for minimization is provided by

$$\frac{\partial F_n}{\partial h_1} = 4\left(\frac{\mu_2 f^{(2)}(x_1)}{2}\right)^2 h_1^3 - 4\left(\frac{\mu_2 f^{(2)}(x_1)}{2}\right)\left(\frac{\mu_4 f^{(4)}(x_2)}{4!}\right) h_1 h_2^4 - \frac{\nu_0 f(x_1)}{nh_1^2}$$

and

$$\frac{\partial F_n}{\partial h_2} = -8\left(\frac{\mu_2 f^{(2)}(x_1)}{2}\right)\left(\frac{\mu_4 f^{(4)}(x_2)}{4!}\right) h_1^2 h_2^3 + 8\left(\frac{\mu_4 f^{(4)}(x_2)}{4!}\right)^2 h_2^7 - \frac{\nu_0 f(x_1)}{nh_2^2}.$$

The sufficient condition for minimization is provided in Lemma 5.4. Setting these equal to zeros yields

$$4\left(\frac{\mu_2 f^{(2)}(x_1)}{2}\right) h_1^3 \left\{ \left(\frac{\mu_2 f^{(2)}(x_1)}{2}\right) h_1^2 - \left(\frac{\mu_4 f^{(4)}(x_2)}{4!}\right) h_2^4 \right\} = \frac{\nu_0 f(x_1)}{n}$$

and

$$-8\left(\frac{\mu_4 f^{(4)}(x_2)}{4!}\right) h_2^5 \left\{ \left(\frac{\mu_2 f^{(2)}(x_1)}{2}\right) h_1^2 - \left(\frac{\mu_4 f^{(4)}(x_2)}{4!}\right) h_2^4 \right\} = \frac{\nu_0 f(x_2)}{n}.$$

Combining these two equations leads to the relationship

$$\mu_2 f^{(2)}(x_1) f(x_2) h_1^3 - 6\mu_4 f(x_1) f^{(4)}(x_2) h_2^5 = 0. \tag{92}$$

This implies that $h_1^3 \sim h_2^5$ where $a_n \sim b_n$ stands for a_n/b_n converges to a nonzero constant. With this relationship, the leading terms in the MSE (37) due to bias and variance are given by

$$\left\{\frac{\mu_2}{2} f^{(2)}(x_1)h_1^2\right\}^2, \quad \frac{\nu_0 f(x_1)}{nh_1}.$$

Then minimizing the sum of these terms show $h_1 = O(n^{-1/5})$, implying $h_2 = O(n^{-3/25})$. The relationship between θ_1 and θ_2 is given by equation (92)

$$\mu_2 f^{(2)}(x_1) f(x_2) \theta_1^3 - 6\mu_4 f(x_1) f^{(4)}(x_2) \theta_2^5 = 0$$

giving the required result.

Proof of Lemma 5.3: The same reasoning provided in the proof of Lemma 5.2 shows that

$$h_1 \sim h_2^2.$$

In contrast with the situation of Lemma 5.3, this relationship implies that the leading terms in the MSE (37) due to bias and variance are given by

$$\left\{ \frac{\mu_6}{6!} f^{(6)}(x_2) h_2^6 \right\}^2, \quad \frac{\nu_0 f(x_1)}{n h_1}.$$

Then minimizing the sum of these terms show $h_1 = O(n^{-1/7})$, implying $h_2 = O(n^{-1/14})$.

References

Cheng, M.-Y., Fan, J., and Marron, J. S. (1997). On automatic boundary corrections. *Annals of Statistics*, 25:1691–1708.

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

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics*, 20:2008–2036.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption. *Journal of the Royal Statistical Society, Series B*, 57:371–394.

Fan, J. and Gijbels, I. (1996). *Local polynomial modeling and its applications*. Chapman & Hall.

Fan, J., Gijbels, I., Hu, T.-C., and Huang, L.-S. (1996). A study of variable bandwidth selection fro local polynomial regression. *Statistica Sinica*, 6:113–127.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*, 11:1156–1174.

Imbens, G. W. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, forthcoming.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142:615–635.

Jones, M. C. (1994). On kernel density derivative estimation. *Communications in Statistics, Theory and Methods*, 23:2133–2139.

Prakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation*. Academic Press, Orland, Florida.

Silverman, B. W. (1986). *Density estimation for statistics and data analyasis*. Chapman & Hall, London.

Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing*. Chapman & Hall.