Adaptive Estimation of the Regression Discontinuity Model

Yixiao Sun*

Department of Economics
Univeristy of California, San Diego
La Jolla, CA 92093-0508

Feburary 2005

*Email: yisun@ucsd.edu; Tel: 858-534-4692

Abstract

In order to reduce the finite sample bias and improve the rate of convergence, local polynomial estimators have been introduced into the econometric literature to estimate the regression discontinuity model. In this paper, we show that, when the degree of smoothness is known, the local polynomial estimator achieves the optimal rate of convergence within the Hölder smoothness class. However, when the degree of smoothness is not known, the local polynomial estimator may actually inflate the finite sample bias and reduce the rate of convergence. We propose an adaptive version of the local polynomial estimator which selects both the bandwidth and the polynomial order adaptively and show that the adaptive estimator achieves the optimal rate of convergence up to a logarithm factor without knowing the degree of smoothness. Simulation results show that the finite sample performance of the locally cross-validated adaptive estimator is robust to the parameter combinations and data generating processes, reflecting the adaptive nature of the estimator. The root mean squared error of the adaptive estimator compares favorably to local polynomial estimators in the Monte Carlo experiments.

Keywords: Adaptive estimator, local cross validation, local polynomial, minimax rate, optimal bandwidth, optimal smoothness parameter

JEL Classification Numbers: C13, C14

1 Introduction

In this paper, we consider the regression discontinuity model:

$$y = m(x) + \alpha d + \varepsilon \tag{1}$$

where $m(x)$ is a continuous function of x, $d = 1\{x \geq x^*\}$, and $E(\varepsilon|x, d) = 0$. Such a model has been used in the empirical literature to identify the treatment effect when there is a discontinuity in the treatment assignment. A partial list of examples include Angrist and Lavy (1999), Black (1999), Battistin and Rettore (2002), Van der Klaauw (2002), DiNardo and Lee (2004), and Chay and Greenstone (2005).

Given the iid data $\{x_i, y_i\}_{i=1}^n$, our objective is to develop a good estimator of α, the treatment effect at a known cut-off point x^*. In order to maintain generality of the response pattern, we do not impose a specific functional form on $m(x)$. Instead, we take $m(x)$ to belong to a family that is characterized by regularity conditions near the cut-off point. This is a semiparametric approach to estimating the regression discontinuity model.

Semiparametric estimation of the regression discontinuity model is closely related to the estimation of conditional expectation at a boundary point. In both settings, the widely used Nadaraya-Watson (NW) estimator has a large finite sample bias and slow rate of convergence. To reduce the finite sample bias and improve the rate of convergence, Hahn, Todd and Van der Klaauw (2001) and Porter (2003) propose using a linear function or a polynomial to approximate $m(x)$ in a small neighborhood of the cut-off point. Porter (2003) obtains the optimal rate of convergence using Stone's (1980) criterion and shows that the local polynomial estimator achieves the optimal rate when the degree of smoothness of $m(x)$ is known.

In this paper, we show that the local polynomial estimator with the asymptotic MSE optimal bandwidth may actually inflate the finite sample bias and reduce the rate of convergence when the degree of smoothness of $m(x)$ is not known. In particular, this will happen if the order of the local polynomial is too large relative to the degree of smoothness. Hence, a drawback of the local polynomial estimator is that the optimal rate of convergence can not be achieved because it depends on the unknown quantity. This calls for an estimator that is adaptive to the unknown smoothness. We require the estimator to be adaptive not just at a fixed model, but also at a sequence of models near it. The adaptive rate refers not just to pointwise convergence, but rather to convergence uniformly over models that are very close to some particular model of interest.

The problem of adaptive estimation of a nonparametric function from noisy data has been studied in a number of papers including Lepski (1990,1991,1992), Donoho and John-

stone (1995), Birge and Massart (1997) and the references cited therein. Various approaches have been proposed, among which Lepski's method has been widely used in the statistical literature; see for example, Lepski and Spokoiny (1997), Lepski, Mammen and Spokoiny (1997) and Spokoiny (2000). These papers study adaptive bandwidth choice in local constant or linear regression for estimating the drift function in a Gaussian white noise model or a nonparametric diffusion model. More specifically, Lepski and Spokoiny (1997) work with the Gaussian white noise model and consider pointwise estimation using a kernel method with the Hölder smoothness class, assuming that the order of smoothness is less than 2. Lepski, Mammen and Spokoiny (1997) extend the pointwise estimation to global estimation using a high order kernel method with the Bosev class. In addition, Lepski's method has been used in several papers on semiparametric estimation of long memory in the time series literature including Giritis, Robinson, and Samarov (2000), Hurvich, Moulier and Soulier (2002), Ioudisky, Moulier and Soulier (2002), Andrews and Sun (2004) and Guggenberger and Sun (2004).

In this paper, we use Lepski's method to construct a rate-adaptive estimator of the regression discontinuity model. In doing so, we extend Lepski's method in several important ways.

First, we consider the local polynomial estimators instead of kernel estimators. The estimation of the regression discontinuity model is similar to the estimation of conditional expectation on the boundary. It is well known that local polynomial estimators have some optimality properties for the boundary estimation problem.

Second, a direct application of Lepski's approach to the present framework involves using a polynomial of a pre-specified order and comparing local polynomial estimators with different bandwidths. More specifically, one has to first choose the order of the polynomial to be larger than the upper bound s^* of the smoothness parameter. Such a strategy is not optimal. If the underlying smoothness parameter s is less than s^*, then it is better to use a polynomial of order $\lfloor s \rfloor$, the largest integer strictly smaller than s. Using a polynomial of a higher order will only inflate the asymptotic variance without the benefit of bias reduction. In contrast, our adaptive method chooses both the bandwidth and the order of the polynomial adaptively The chosen polynomial in the adaptive estimator is indeed of order $\lfloor s \rfloor$.

Third, our adaptive rule does not use the lower and upper bounds for s while the adaptive rule in Lepski (1990) uses them explicitly. In consequence, the rate of convergence of our adaptive estimator can be arbitrarily close to the parametric rate in the infinitely smooth case while that of Lepski's estimator is capped by the upper bound s^*. This advan-

tage of our adaptive estimator is partly due to the use of the zero-one loss rather than the squared-error loss. Results for the zero-one loss are sufficient to obtain the optimal rate of convergence, which is the item of greatest interest here.

Finally, one drawback of Lepski's approach is that there are constants in the adaptive procedure that are arbitrary. This is true for other adaptive procedures although some procedures may fix their constants at certain ad hoc values and seemingly remove the need to choose any constant. In this paper, we propose using local cross validation to select the constants and provide a practical strategy to implement the adaptive estimator.

We compare the root mean-squared error (RMSE) performance of the adaptive estimator with the local constant, local linear, local quadratic and local cubic estimators. We consider three groups of models with different response functions $m(x)$. In the first group, $m(x)$ is the sum of a third order polynomial and a term containing $(x - x^*)^{s_0}$ for some non-integer s_0. Response functions in this group are designed to have finite smoothness s_0. By choosing different s_0, we can get response functions that have different degrees of smoothness. The second group is the same as the first group except that $m(x)$ is perturbed by an additive sine function such that the response function has a finer structure. For the third group, we take $m(x)$ to be a constant, linear, quadratic or cubic function. This group is designed to give each of the local polynomial estimators the best advantage.

The Monte Carlo results show that the RMSE performance of the adaptive estimator is very robust to the data generating process, reflecting its adaptive nature. Its RMSE is either the lowest or among the three lowest ones for the parameter combinations and data generating processes considered. In contrast, a local polynomial estimator may perform very well in some scenario but disastrously in other scenarios. The best estimator in an overall sense seems to be the adaptive estimator.

The rest of the paper is organized as follows. Section 2 overviews the local polynomial estimator and examines its asymptotic properties when the order of the polynomial is larger than the underlying smoothness. Section 3 establishes the optimal rate of convergence within the Hölder smoothness class and shows that the local polynomial estimator achieves the optimal rate when the degree of smoothness is known. Section 4 introduces the adaptive local polynomial estimator. It is shown that the adaptive estimator achieves the optimal rate for known smoothness up to a logarithm factor when the smoothness is not known. For a given response function $m(x)$, it is also shown that the adaptive procedure provides a consistent estimator of the smoothness index defined in that section. The subsequent section contains the simulation results that compare the finite sample performance of the adaptive estimator with those of the local polynomial estimators. Proofs and additional

technical results are given in the Appendix.

Throughout the paper, $1\{\cdot\}$ is the indicator function and $||\cdot||$ signifies the Euclidean norm. C is a generic constant that may be different across different lines.

2 Local Polynomial Estimation

Consider the regression discontinuity design model $y = m(x) + \alpha d + \varepsilon$ where $m(x)$ is a unknown function of x, $E(\varepsilon|x,d) = 0$ and $d = 1\{x \geq x^*\}$. Given the iid data (x_i, y_i), $i = 1, 2, ..., m$, our objective is to estimate α without assuming the functional form of $m(\cdot)$. However, it is necessary to assume that $m(x)$ belongs to some smoothness class.

Definition: Let $s = \ell + \tau$ where ℓ is the largest integer strictly less than s and $\tau \in (0, 1]$. If a function defined on the interval $[x^*, x^* + \delta)$ is ℓ times differentiable,

$$\sup_{x \in [x^*, x^* + \delta)} \left| m^{(j)}(x) \right| \leq K \text{ for } j = 0, 1, 2, 3, ..., \ell$$

and

$$\left| m^{(\ell)}(x_1) - m^{(\ell)}(x_2) \right| \leq K \left| x_1 - x_2 \right|^{\tau} \text{ for } x_1, x_2 \in [x^*, x^* + \delta)$$

where $m^{(j)}(x)$ is the j-th order derivative and $m^{(j)}(x^*)$ is the j-th order right hand derivative at x^*, then we say $m(x)$ is smooth of order s on $[x^*, x^* + \delta)$. Denote this class of functions by $\mathcal{M}_+(s, \delta, K)$. Similarly, we can define $\mathcal{M}_-(s, \delta, K)$ as the class of functions that satisfy the above two conditions with $[x^*, x^* + \delta)$ replaced by $(x^* - \delta, x^*]$ and $m^{(j)}(x^*)$ being the left hand derivative at x^*.

Assumption 1: $m(x) \in \mathcal{M}(s, \delta, K)$ where

$$\mathcal{M}(s, \delta, K) := \{m : m \in \mathcal{M}_+(s, \delta, K) \cap \mathcal{M}_-(s, \delta, K) \cap C^0(x^* - \delta, x^* + \delta)\}$$

and $C^0(x^* - \delta, x^* + \delta)$ is the set of continuous functions on $(x^* - \delta, x^* + \delta)$.

Assumption 1 allows us to develop an ℓ term Taylor expansion of $m(x)$ on each side of x^*. Without loss of generality, we focus on $x \geq x^*$, in which case we have

$$m(x) = m(x^*) + \sum_{j=1}^{\ell} b_j^+ (x - x^*)^j + \tilde{e}^+(x), \tag{2}$$

where $b_j^+ = \frac{1}{j!} \frac{d^j}{dx^j} m(x)|_{x = x^*+}$ is the (normalized) j-th order right hand derivative of $m(x)$ at x^* and

$$\tilde{e}^+(x) = \frac{1}{\ell!} \left(m^{(\ell)}(\tilde{x}) - m^{(\ell)}(x^*) \right) (x - x^*)^{\ell} \tag{3}$$

for some $\tilde{x}$ between x and x^*. Under Assumption 1, $\tilde{e}^+(x)$ satisfies

$$\left|\tilde{e}^+(x)\right| \leq K\left(\ell!\right)^{-1}\left|x-x^*\right|^s \text{ for all } x \in [x^*, x^*+\delta). \tag{4}$$

We break up the Taylor expansion into the part that will be captured by the local polynomial regression and the remainder:

$$m(x) = m(x^*) + \sum_{j=1}^{\min(r,\ell)} b_j^+(x-x^*)^j + R^+(x),\ x \geq x^* \tag{5}$$

where

$$\begin{aligned} R^+(x) &= \sum_{j=\min(r,\ell)+1}^{\ell} b_j^+(x-x^*)^j + \tilde{e}^+(x) \\ &:= 1\{\ell \geq r+1\} b_{r+1}^+(x-x^*)^{r+1} + e^+(x), \end{aligned} \tag{6}$$

$$\left|e^+(x)/\left(x-x^*\right)^q\right| = O(1) \text{ uniformly over } x \in [x^*, x^*+\delta), \tag{7}$$

and $q = \min\left\{s, r+2\right\}$.

Let $b^+(r)$ denote the column r-vector whose j-th element is b_j^+ for $j = 1, 2, ..., \min(r, \ell)$ and 0 for $j = \min(r,\ell)+1, ..., r$. Let $z_{ir} = (1, (x_i - x^*), ..., (x_i - x^*)^r)$ be the row $(r+1)$-vector, $\left(\theta_r^+\right)' = (c^+, (b^+(r))')$ and $c^+ = \alpha + m(x^*)$. Then for $x_i \geq x^*$, we have

$$y_i = z_{ir}\theta_r^+ + R^+(x_i) + \varepsilon_i \tag{8}$$

To estimate θ_r^+, we minimize

$$\sum_{i=1}^{n} k_h(x_i - x^*) d_i \left(y_i - z_{ir}\theta_r\right)^2 \tag{9}$$

with respect to θ_r, where $d_i = 1\{x_i \geq x^*\}$, $k_h(x_i - x^*) = 1/hk((x_i - x^*)/h)$ and h is the bandwidth parameter. Let Y^+ and Z_r^+ be the data matrix that collects the values of y_i and z_{ir} respectively with the corresponding value of $x_i \geq x^*$. Then (8) can be written in the vector form:

$$Y^+ = Z_r^+\theta_r^+ + R^+ + \varepsilon^+ \tag{10}$$

and the objective function in (9) becomes

$$\left(Y^+ - Z_r^+\theta_r\right)' W^+ \left(Y^+ - Z_r^+\theta_r\right) \tag{11}$$

where $W^+ = diag(\{hk_h(x_i - x^*)\})_{x_i \geq x^*}$. Minimizing the preceding quantity gives

$$\hat{\theta}_r^+ = \left(\hat{c}_r^+, (\hat{b}^-(r))'\right)' = \left(Z_r^{+\prime} W^+ Z_r^+\right)^{-1}\left(Z_r^{+\prime} W^+ Y^+\right). \tag{12}$$

Defining Y^-, Z_r^-, W^- analogously using the observations satisfying $x_i < x^*$, we have

$$Y^- = Z_r^- \theta_r^- + R^- + \varepsilon^- \tag{13}$$

where $\left(\theta_r^-\right)' = (c^-, (b^-(r))')$, $c^- = m(x^*)$ and $b^-(r)$ is similarly defined but with the right hand derivatives replaced by the left hand derivatives. Minimizing $(Y^- - Z_r^- \theta_r)' W^- (Y^- - Z_r^- \theta_r)$ with respect to θ_r gives an estimate for θ_r^- :

$$\hat{\theta}_r^- = \left(\hat{c}_r^-, (\hat{b}^-(r))'\right)' = \left(Z_r^{-\prime} W^- Z_r^-\right)^{-1} \left(Z_r^{-\prime} W^- Y^-\right). \tag{14}$$

The difference between $\hat{c}_r^+$ and $\hat{c}_r^-$ gives an estimate for α :

$$\hat{\alpha}_r = \hat{c}_r^+ - \hat{c}_r^-. \tag{15}$$

To investigate the asymptotic properties of $\hat{\alpha}_r$, we maintain the following two additional assumptions.

Assumption 2: (a) $E(\varepsilon|x, d) = 0$.

(b) $\sigma^2(x) = E(\varepsilon^2|x)$ is continuous for $x \neq x^*$ and the right and left hand limits exist at x^*.

(c) For some $\zeta > 0$, $E(|\varepsilon|^{2+\zeta}|x)$ is uniformly bounded on $[x^* - \delta, x^* + \delta]$.

(d)The marginal density $f(x)$ of x is continuous on $[x^* - \delta, x^* + \delta]$.

Assumption 3: The kernel $k(\cdot)$ is even, bounded and has a bounded support.

Theorem 1 *Let Assumptions 1-3 hold. If $n \to \infty$ and $h \to 0$ such that $nh \to \infty$, then*

$$\sqrt{nh}(\hat{\alpha}_r - \alpha) - B \Rightarrow N\left(0, \omega^2 \lambda_r^2\right)$$

where

$$\omega^2 = \frac{\sigma^{2+}(x^*) + \sigma^{2-}(x^*)}{f(x^*)}, \ \lambda_r^2 = e_1' \Gamma_r^{-1} V_r \Gamma_r^{-1} e_1,$$

$$B = 1\{s > r+1\} \frac{\left(e_1' \Gamma_r^{-1} \mu_r\right) \left[b_{r+1}^+ - (-1)^{r+1} b_{r+1}^-\right]}{f(x^*)} h^{r+1} \sqrt{nh}(1 + o_p(1)) + O_p\left(h^q \sqrt{nh}\right),$$

$$\Gamma_r = \left(\gamma_{i+j-2}\right)_{(r+1)\times(r+1)} = \begin{pmatrix} \gamma_0 & \cdots & \gamma_r \\ \vdots & & \vdots \\ \gamma_r & \cdots & \gamma_{2r} \end{pmatrix},$$

$$V_r = \left(v_{i+j-2}\right)_{(r+1)\times(r+1)} = \begin{pmatrix} v_0 & \cdots & v_r \\ \vdots & & \vdots \\ v_r & \cdots & v_{2r} \end{pmatrix},$$

$e_1 = (1, 0, ..., 0)'$, $\mu_r = (\gamma_{r+1}, ..., \gamma_{2r+1})'$, $\gamma_j = \int_0^\infty k(u) u^j du$ *and* $v_j = \int_0^\infty k^2(u) u^j du$.

Remarks

1. When $s > r+1$, Theorem 1 is the same as Theorem 3(a) in Porter (2003). The proof is straightforward and uses part of Porter's result.

2. If $s > r+1$, the "asymptotic bias" of $\hat{\alpha}_r$, defined as $B/\sqrt{nh}$, is of order h^{r+1}. In contrast, the asymptotic bias of $\hat{\alpha}_0$ is of order h. The asymptotic bias of $\hat{\alpha}_r$ for $r \geq 1$ is smaller than that of $\hat{\alpha}_0$ by an order of magnitude provided that $m(x)$ is smooth of order $s > r+1$.

3. If $s > r+1$, then the "asymptotic MSE" of $\hat{\alpha}_r$ is

$$AMSE(\hat{\alpha}_r) = C_1 h^{2r+2} + \frac{C_2}{nh}. \tag{16}$$

   Assume that $C_1 > 0$ and $C_2 > 0$, then minimizing $AMSE(\hat{\alpha}_r)$ over h gives the AMSE-optimal choice for h :

$$h^* = \left(\frac{C_2}{(2r+2)C_1} \right)^{1/(2r+3)} n^{-1/(2r+3)}. \tag{17}$$

   For this AMSE-optimal choice of h, $AMSE(\hat{\alpha}_r)$ is proportional to

$$\left(\left(e_1' \Gamma_r^{-1} \mu_r \mu_r' \Gamma_r^{-1} e_1 \right) \left(e_1' \Gamma_r^{-1} V_r \Gamma_r^{-1} e_1 \right)^{2(r+1)} \right)^{1/(2r+3)} n^{-2(r+1)/(2r+3)}. \tag{18}$$

   So $\hat{\alpha}_r$ converges to α at the rate of $n^{-(r+1)/(2r+3)}$. In particular, $\hat{\alpha}_0$ converges to α at the rate of $n^{-1/3}$. As a consequence, by appropriate choice of h, one has asymptotic normality of $\hat{\alpha}_r$ with a faster rate of convergence (as a function of the sample size n) than is possible with $\hat{\alpha}_0$.

4. When $s > r+1$ and $h = h^*$, the asymptotic mean squared error depends on the kernel only through the quantity

$$\Xi(k) = \left(e_1' \Gamma_r^{-1} \mu_r \mu_r' \Gamma_r^{-1} e_1 \right) \left(e_1' \Gamma_r^{-1} V_r \Gamma_r^{-1} e_1 \right)^{2(r+1)}. \tag{19}$$

   This quantity is the same as $T_{p+1,\nu}$ defined in equation (7) in Cheng, Fan and Marron (1997, p. 1695). Using their proof without change, we can show that the kernel that minimizes $\Xi(k)$ over the class of kernels defined by

$$\mathcal{K} = \left\{ k(x) : k(x) \geq 0, \ \int_{-\infty}^{\infty} k(x)dx = 1, \ |k(x) - k(y)| \leq C |x-y| \text{ for some } C > 0 \right\}$$

   is simply the Bartlett kernel $k(x) = (1 - |x|) \, 1\{|x| \leq 1\}$ for all r. This is an unusual result because the optimal kernel does not depend on the order of the local polynomial.

5. Consider the case that $s \leq r+1$ and h is proportional to the AMSE optimal rate $n^{-1/(2r+3)}$. For such a configuration, the asymptotic bias dominates the asymptotic variance. The estimator $\hat{\alpha}_r$ converges to the true α at the rate of $n^{-\frac{s}{2r+3}}$. The larger r is, the slower the rate of convergence is. For example, when $2r+3 \geq 3s$, the rate of convergence is slower than $n^{-1/3}$, the rate that is obtainable using the Nadaraya-Watson estimator. By fitting a high order polynomial, it is possible that we inflate the boundary effect instead of reducing it.

Theorem 1 shows that the local polynomial estimation has the potential to reduce the boundary bias problem and deliver a faster rate of convergence when the response function is smooth enough. In the next section, we establish the optimal rate of convergence when the degree of smoothness is known. It is shown that the local polynomial estimator with appropriately chosen bandwidth achieves this optimal rate.

3 Optimal Rate of Convergence

To obtain the optimal rate of convergence, we cast the regression discontinuity model into the following general framework:

Suppose $\mathcal{P}$ is a family of probability models on some fixed measurable space $(\Omega, \mathcal{A})$. Let α be a functional defined on $\mathcal{P}$, taking values in $\mathbb{R}$. An estimator of α is a measurable map $\hat{\alpha} : \Omega \to \mathbb{R}$. For a given loss function $L(\hat{\alpha}, \alpha)$, the maximum expected loss over $P \in \mathcal{P}$ is defined to be

$$R(\hat{\alpha}, \mathcal{P}) = \sup_{P \in \mathcal{P}} E_P L(\hat{\alpha}, \alpha(P)) \tag{20}$$

where E_P is the expectation operator under the probability measure P. Our goal is to find an achievable lower bound for the minimax risk defined by

$$\inf_{\hat{\alpha}} R(\hat{\alpha}, \mathcal{P}) = \inf_{\hat{\alpha}} \sup_{P \in \mathcal{P}} E_P L(\hat{\alpha}, \alpha(P)). \tag{21}$$

If we add a subscript n to $\hat{\alpha}$, P, and $\mathcal{P}$ where n is the sample size, the achievable lower bound will translate into the best rate of convergence of $R(\hat{\alpha}, \mathcal{P})$ to zero. This best rate is called the minimax rate of convergence as it is derived from the minimax criterion. It is also commonly referred to as the optimal rate of convergence.

Now let us put the regression discontinuity model in the above general framework. Let $f(\cdot)$ be a probability density function of x and $\varphi_x(\cdot)$ be a conditional density of ε for a given x such that $E(\varepsilon|x) = 0$. For both densities the dominating measures are the usual

Lesbegue measures. Define

$$\begin{aligned}\mathcal{P}(s,\delta,K) &= \Big\{P_{m,\alpha} : \frac{dP_{m,\alpha}}{d\mu} = f(x)\varphi_x\left(y - m(x)\right)1\left\{x < x^*\right\} \\ &\quad + f(x)\varphi_x\left(y - m(x) - \alpha\right)1\left\{x \geq x^*\right\},\ m(x) \in \mathcal{M}(s,\delta,K), |\alpha| \leq K\Big\}\end{aligned}$$

where μ is the Lesbegue measure on $\mathbb{R}^2$. For this family of models, the marginal distribution of x and the conditional distribution of ε are the same across all members. The difference among members lies in the conditional mean of y for a given x. In other words, the function $m(\cdot)$ and the constant α characterize the probability model in the family $\mathcal{P}(s,\delta,K)$. To reflect this, we use subscripts m,α to differentiate the probability model in $\mathcal{P}(s,\delta,K)$. For the regression discontinuity model, the functional of interest is $\alpha(P_{m,\alpha}) = \alpha$. For a given loss function $L(\cdot,\cdot)$, we want to design an estimator $\hat{\alpha}$ to minimize

$$\sup_{P_{m,\alpha}\in\mathcal{P}(s,\delta,K)} E_{m,\alpha}L(\hat{\alpha},\alpha) \tag{22}$$

where $E_{m,\alpha}L(\hat{\alpha},\alpha) := E_{P_{m,\alpha}}L(\hat{\alpha},\alpha)$ and $E_{P_{m,\alpha}}$ is the expectation operator under $P_{m,\alpha}$.

One common choice of $L(\cdot,\cdot)$ is the quadratic loss function

$$L(\hat{\alpha},\alpha) := L\left(\hat{\alpha} - \alpha\right) = (\hat{\alpha} - \alpha)^2, \tag{23}$$

in which case $R(\hat{\alpha},\mathcal{P})$ is the maximum expected mean squared error. Another common choice is the 0-1 loss function

$$L(\hat{\alpha},) =: L\left(\hat{\alpha} - \alpha\right) = 1\left\{|\hat{\alpha} - \alpha| > \epsilon/2\right\} \tag{24}$$

for some fixed $\epsilon > 0$, in which case, $R(\hat{\alpha},\mathcal{P})$ is the maximum probability that $\hat{\alpha}$ is not in the $\epsilon/2$-neighborhood of α. Since the expected mean squared error may not exist for the local polynomial estimator, we use the 0-1 loss for convenience in this paper. The use of the 0-1 loss is innocuous if the optimal rate of convergence is the item of greatest interest.

The derivation of a minimax rate of convergence for an estimator involves a series of minimax calculations for different sample sizes. There is no initial advantage in making the dependence on the sample size explicit. Consider then the problem of finding a lower bound for the minimax risk $\inf_{\hat{\alpha}} \sup_{P\in\mathcal{P}} E_P L(\hat{\alpha},\alpha)$. The simplest method for finding such a bound is to identify an estimator with a test between simple hypotheses. The whole argument could be cast in the language of Neyman-Pearson testing. Let P, Q be probability measures defined on the same measurable space $(\Omega,\mathcal{A})$. Then the testing affinity (Le Cam (1986) and Donoho and Liu (1991)) of two probability measures is defined to be

$$\pi(P,Q) = \inf(E_P\phi + E_Q(1-\phi)) \tag{25}$$

where the infimum is taken over the measurable function ϕ such that $0 \leq \phi \leq 1$. In other words, $\pi(P,Q)$ is the smallest sum of type I and type II errors of any test between P and Q. It is a natural measure of the difficulty of distinguishing P and Q. Suppose μ is a measure dominating both P and Q with corresponding densities p and q. It follows from the Neyman-Pearson lemma that the infimum is achieved by setting $\phi = 1\{p \leq q\}$ and

$$\begin{aligned} \pi(P,Q) &= \int 1\{p \leq q\}\, p d\mu + \int 1\{p > q\}\, q d\mu \\ &= 1 - \frac{1}{2}\int |p-q|\, d\mu := 1 - \frac{1}{2}\,||P-Q||_1 \end{aligned} \tag{26}$$

where $||P-Q||_1 = \int |p-q|\, d\mu$ is the L_1 distance between two probability measures.

Now consider a pair of probability models $P, Q \in \mathcal{P}$ such that $\alpha(P) - \alpha(Q) \geq \epsilon$. Then for any estimator $\hat{\alpha}$

$$1\{|\hat{\alpha} - \alpha(P)| > \epsilon/2\} + 1\{|\hat{\alpha} - \alpha(Q)| > \epsilon/2\} \geq 1. \tag{27}$$

Let

$$\phi = \frac{1\{|\hat{\alpha} - \alpha(P)| > \epsilon/2\}}{1\{|\hat{\alpha} - \alpha(P)| > \epsilon/2\} + 1\{|\hat{\alpha} - \alpha(Q)| > \epsilon/2\}}, \tag{28}$$

then $0 \leq \phi \leq 1$ and

$$\begin{aligned} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(|\hat{\alpha} - \alpha(\mathbb{P})| &> \epsilon/2) \geq \frac{1}{2}\{P(|\hat{\alpha} - \alpha(P)| > \epsilon/2) + Q(|\hat{\alpha} - \alpha(Q)| > \epsilon/2)\} \\ &\geq \frac{1}{2}E_P\phi + \frac{1}{2}E_Q(1-\phi) \geq \frac{1}{2}\pi(P,Q). \end{aligned} \tag{29}$$

Therefore

$$\inf_{\hat{\alpha}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\{|\hat{\alpha} - \alpha| > \epsilon/2\} \geq \frac{1}{2}\pi(P,Q) \tag{30}$$

for any P and Q such that $\alpha(P) - \alpha(Q) \geq \epsilon$.

Inequality (30) suggests a simple way to get a good lower bound for the minimax probability error: search for the pair (P,Q) to minimize $\pi(P,Q)$, subject to the constraint $\alpha(P) - \alpha(Q) \geq \epsilon$.

To obtain a lower bound with a sequence of independent observations, we let $(\Omega, \mathcal{A})$ be the product space and $\mathcal{P}$ be a family of probability models on such a space. Then for any pair of finite-product measures $P = \Pi_{i=1}^n P_i$ and $Q = \Pi_{i=1}^n Q_i$, the minimax risk satisfies

$$\inf_{\hat{\alpha}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\{|\hat{\alpha} - \alpha| > \epsilon/2\} \geq \frac{1}{2}\left(1 - \frac{1}{2}\,||\Pi_{i=1}^n P_i - \Pi_{i=1}^n Q_i||_1\right) \tag{31}$$

provided that $\alpha(P) - \alpha(Q) \geq \epsilon$.

We now turn to the regression discontinuity model. Our objective is to search for two probability models P and Q that are difficult to distinguish by the independent observations

(x_i, y_i), $i = 1, 2, ..., n$. Note that it is not restrictive to consider only particular distributions for ε_i and x_i for the purpose of obtaining a lower bound. The minimax risk for a larger class of probability models must not be smaller than that for a smaller class of probability models. Therefore, if the lower bound holds for a particular distributional assumption, then it also holds for a wider class of distributions. To simplify the calculation, we assume that ε_i is iid $N(0, \sigma^2)$ and x_i is iid uniform $[x^* - \delta, x^* + \delta]$ under both P and Q. More details on the construction of P and Q are given in the proof of the following theorem:

Theorem 2 *Let Assumption 2 hold.*

(a) For any finite constants s, δ and K, we have

$$\liminf_{n\to\infty} \inf_{\hat{\alpha}} \sup_{P_{m,\alpha}\in\mathcal{P}(s,\delta,K)} P_{m,\alpha}\left(\left|n^{\frac{s}{2s+1}}(\hat{\alpha}-\alpha)\right| > \frac{\epsilon}{2}\right) \geq C$$

for some positive constant C and a small $\epsilon > 0$.

(b) Suppose Assumption 3 also holds. Let $h = \psi_1 n^{-1/(2s+1)}$ for some constant ψ_1, then

$$\lim_{\epsilon\to\infty} \limsup_{n\to\infty} \sup_{P_{m,\alpha}\in\mathcal{P}(s,\delta,K)} P_{m,\alpha}\left(\left|n^{\frac{s}{2s+1}}(\hat{\alpha}_\ell-\alpha)\right| > \frac{\epsilon}{2}\right) = 0.$$

Remarks

1. Part (a) of the theorem shows that there exists no estimator $\hat{\alpha}$ that converges to α at a rate faster than $n^{-s/(2s+1)}$ uniformly over the class of probability models $\mathcal{P}(s, \delta, K)$. Part (b) of the theorem shows that the rate $n^{-s/(2s+1)}$ is achieved by the local polynomial estimator provided that $r = \ell$ and h is chosen appropriately. Because of Parts (a) and (b), the rate $n^{-s/(2s+1)}$ is called the minimax optimal rate of convergence.

2. This results of the theorem extends Porter (2003) who considers a class of functions that are ℓ times continuously differentiable. Our result is more general as we consider the Hölder smoothness class, which is larger than what Porter (2003) has considered. Our method for calculating the lower bound for the minimax risk is also simpler than that of Stone (1980), which is adopted in Porter (2003).

3. An alternative proof of the minimax rate is to use the asymptotic equivalence of nonparametric regression models and Gaussian noise models (see Brown and Low (1996)). The Gaussian noise model is defined by $dY = S(t)dt + \varepsilon dW(t)$ where $W(t)$ is the standard Brownian motion. Ibragimov and Khasminskii (1981) show that the optimal minimax rate for estimating the drift function $S(t)$ is $\varepsilon^{2s/(2s+1)}$. Since ε in

the Gaussian noise model corresponds to $1/\sqrt{n}$ in a nonparametric regression with n copies of iid data, we infer that the optimal minimax rate in the nonparametric regression is $n^{-s/(2s+1)}$. Our proof is in the spirit of Donoho and Liu (1991) and involves only elementary calculations.

4 A Rate Adaptive Estimator

The previous section establishes the optimal rate of convergence when the degree of smoothness is known. In this section, we propose a local polynomial estimator that achieves the optimal rate of convergence up to a logarithm factor when the degree of smoothness is not known.

Let $[s_*, s^*]$ for some $s_* > 0$ and $s^* \in [s_*, \infty)$ be the range of smoothness. For each $\tau \in [s_*, s^*]$, we define a local polynomial estimator $\hat{\alpha}_\tau = \hat{c}_\tau^+ - c_\tau^-$, by setting

$$\begin{aligned} h_\tau &= \psi_1 n^{-1/(2\tau+1)} \text{ and} \\ r_\tau &= w \text{ for } \tau \in (w, w+1] \text{ for } w = 0, 1, ... \end{aligned} \tag{32}$$

where ψ_1 is a positive constant. Equivalently, r_τ is the largest integer that is strictly less than τ. Note that the subscript on $\hat{\alpha}$, $\hat{c}^+$ and $\hat{c}^-$ indicates the order of the local polynomial in the previous sections while it now indicates the underlying smoothing parameter that generates the bandwidth and the order of the polynomial given in (32).

Let $g := 1/\log n$ and $\mathcal{S}_g$ be the g-net of the interval $[s_*, \infty)$: $\mathcal{S}_g = \{\tau : \tau = s_* + jg, j = 0, 1, 2, ...\}$. For a positive constant ψ_2, define

$$\hat{s} = \sup \left\{ \tau_2 \in \mathcal{S}_g : |\hat{\alpha}_{\tau_1} - \hat{\alpha}_{\tau_2}| \leq \psi_2 \left(n h_{\tau_1}\right)^{-1/2} \lambda_{\tau_1} \zeta(n) \text{ for all } \tau_1 \leq \tau_2, \tau_1 \in \mathcal{S}_g \right\}, \tag{33}$$

where $\zeta(n) = (\log n)(\log\log(n))^{1/2}$. Intuitively, $\hat{s}$ is the largest smoothness parameter such that the associated local polynomial estimator does not differ significantly from the local polynomial estimator with a smaller smoothness parameter. Graphically, one can view the bound in the definition of $\hat{s}$ as a function of τ_1. Then, $\hat{s}$ is the largest value of $\tau_2 \in \mathcal{S}_g$ such that $|\hat{\alpha}_{\tau_1} - \hat{\alpha}_{\tau_2}|$ lies below the bound for all $\tau_1 \leq \tau_2$, $\tau_1 \in \mathcal{S}_g$. Calculation of $\hat{s}$ is carried out by considering successively larger τ_2 values s_*, $s_* + g$, $s_* + 2g$, ..., until for some τ_2 the deviation $|\hat{\alpha}_{\tau_1} - \hat{\alpha}_{\tau_2}|$ exceeds the bound for some $\tau_1 \leq \tau_2$, $\tau_1 \in \mathcal{S}_g$.

Finally, we set the adaptive estimator to be

$$\hat{\alpha}_A = \hat{\alpha}_{\hat{s}}. \tag{34}$$

The proposed adaptive procedure is based on the comparison of local polynomial estimators with different smoothness parameters from the g-net $\mathcal{S}_g$. The total number of smoothness parameters in $\mathcal{S}_g$ is of order $\log(n)$ and the resolution of the g-net $\mathcal{S}_g$ is $1/\log n$. As the sample size increases, the grid of $\mathcal{S}_g$ becomes finer and finer. However, given the structure of $\mathcal{S}_g$, it is not possible to distinguish smoothness parameters whose difference is less than $1/\log n$. This is why the proposed estimator can not achieve the best rate of convergence $n^{-s/(2s+1)}$ for known smoothness.

To further understand the adaptive procedure, consider a function $m(\cdot) \in \mathcal{M}(s,\delta,K)$ but $m(\cdot) \notin \mathcal{M}(s',\delta,K)$ for any $s' > s$. In other words, $m(\cdot)$ is smooth to at most order s. For any $\tau_1 \leq \tau_2 \leq s$, it follows from Theorem 1 that the asymptotic bias of $\sqrt{nh_{\tau_1}}(\hat{\alpha}_{\tau_1} - \alpha)$ is

$$\begin{aligned} & asymbias\left(\sqrt{nh_{\tau_1}}(\hat{\alpha}_{\tau_1} - \alpha)\right) \\ = \; & O\left(\sqrt{nh_{\tau_1}} h_{\tau_1}^{r_{\tau_1}+1}\right) = O\left(n^{[\tau_1 - \min(r_{\tau_1}+1, s)]/(2\tau_1+1)}\right) = O(1). \end{aligned} \tag{35}$$

Similarly, the asymptotic bias of $\sqrt{nh_{\tau_1}}(\hat{\alpha}_{\tau_2} - \alpha)$ is

$$\begin{aligned} & asymbias\left(\sqrt{nh_{\tau_1}}(\hat{\alpha}_{\tau_2} - \alpha)\right) \\ = \; & O\left(n^{\tau_1/(2\tau_1+1)} n^{-\min(r_{\tau_2}+1, s)/(2\tau_2+1)}\right) \\ = \; & O\left(n^{\tau_1/(2\tau_1+1) - \tau_2/(2\tau_2+1)} n^{[\tau_2 - \min(r_{\tau_2}+1, s)]/(2\tau_2+1)}\right) = o(1). \end{aligned} \tag{36}$$

Therefore, the asymptotic bias of $\sqrt{nh_{\tau_1}}\,|\hat{\alpha}_{\tau_1} - \hat{\alpha}_{\tau_2}|$ is bounded. On the other hand, $\sqrt{nh_{\tau_1}}\,|\hat{\alpha}_{\tau_1} - \hat{\alpha}_{\tau_2}|$ is no larger than

$$\sqrt{nh_{\tau_1}}\,|\hat{\alpha}_{\tau_1} - \alpha| + \sqrt{nh_{\tau_1}}\,|\hat{\alpha}_{\tau_2} - \alpha| \tag{37}$$

whose asymptotic variance is of order $O(1)$. As a consequence, when $\tau_1 \leq \tau_2 \leq s$, $\sqrt{nh_{\tau_1}}\,|\hat{\alpha}_{\tau_1} - \hat{\alpha}_{\tau_2}|$ is stochastically bounded in large samples and$\sqrt{nh_{\tau_1}}\,|\hat{\alpha}_{\tau_1} - \hat{\alpha}_{\tau_2}| \leq \psi_2 \lambda_{\tau_1} \zeta(n)$ holds with probability approaching 1. This heuristic argument suggests that the probability that $\hat{s}$ is less than s is small in large samples. Next, consider $\tau_1 = s$ and $\tau_2 > s$, the asymptotic bias of $\sqrt{nh_s}(\hat{\alpha}_{\tau_2} - \alpha)$ is of order $O\left(n^{s/(2s+1)} n^{-s/(2\tau_2+1)}\right) = O\left(n^{\tau_2 - s}\right)$ which will be larger than $\psi_2 \lambda_{\tau_1} \zeta(n)$ in general if $\tau_2 - s$ is sufficiently large. This suggests that $\hat{s}$ can not be too far away from s from above. Rigorous arguments are given in the proofs of the next two Theorems in the Appendix.

Theorem 3 *Let Assumptions 2–3 hold. Assume that* $\min_{r \in [r_{s_*}, r_{s^*}]} \{\mu_{\min}(\Gamma_r)\} > 0$ *where* $\mu_{\min}(\Gamma_r)$ *is the smallest eigenvalue of* Γ_r. *For all* $s^* \in [s_*, \infty)$ *with* $s_* > 0$, *we have*

$$\lim_{C_1 \to \infty} \limsup_{n \to \infty} \sup_{s \in [s_*, s^*]} \sup_{P_{m,\alpha} \in \mathcal{P}(s,\delta,K)} P_{m,\alpha}\left(n^{\frac{s}{2s+1}} \zeta^{-1}(n)\,|\hat{\alpha}_A - \alpha| \geq C_1\right) = 0.$$

Remarks

1. Theorem 2 shows that the optimal rate of convergence for the estimation of α is given by $n^{-s/(2s+1)}$ when s is finite and known. Theorem 3 shows that the adaptive estimator achieves this rate up to a logarithm factor $\zeta(n)$ when s is finite and *not* known.

2. When s is not known, the optimal rate of $n^{-s/(2s+1)}$ for known smoothness can not be achieved in general. For the Gaussian noise model and quadratic loss, Lepski (1990) shows that an extra $(\log n)^{s/(2s+1)}$ factor is needed. This result has been recently challenged by Cai and Low (2003) who show that under the 0-1 loss the achievable lower bound for unknown smoothness is the same as that is possible with known smoothness. However, their results are obtained under the assumption that there are a finite number of different values of the smoothness parameter. This assumption does not hold for the problem at hand. As a result, the extra logarithm factor may not be removed in general for the 0-1 loss. This extra logarithmic factor is an unavoidable price for adaptation and most (if not all) adaptive estimators of linear functionals share this property.

3. If the function $m(x)$ is not smooth to the same order on the two sides of x^*, say $m(x) \in \mathcal{M}_+(s_1, \delta, K) \cap \mathcal{M}_-(s_2, \delta, K)$, then we can estimate c^+ and c^- adaptively on each side of the cutoff point x^*. For a constant $\psi_2^+ > 0$, let

$$\hat{s}_+ = \sup\left\{\tau_2 \in \mathcal{S}_g : \left|\hat{c}_{\tau_1}^+ - \hat{c}_{\tau_2}^+\right| \leq \psi_2^+ \left(nh_{\tau_1}\right)^{-1/2} \lambda_{\tau_1}\zeta(n) \text{ for all } \tau_1 \leq \tau_2, \tau_1 \in \mathcal{S}_g\right\}$$

   where $\hat{c}_\tau^+$ is the local polynomial estimator of c^+ when $h = \psi_1^+ n^{-1/(2\tau+1)}$ and $r = r_\tau$, the largest integer strictly less than τ. The adaptive estimator $\hat{c}_A^+$ of c^+ is given by $\hat{c}_{\hat{s}+}$. The adaptive estimator $\hat{c}_A^-$ of c^- can be analogously defined. Finally, the adaptive estimator of $\hat{\alpha}$ is set to be $\hat{\alpha}_A = \hat{c}_A^+ - \hat{c}_A^-$. In this case, the rate of the convergence of $\hat{\alpha}_A$ is easily seen to be $\zeta(n)\exp\left(-\frac{\min(s_1,s_2)}{2\min(s_1,s_2)+1}\log n\right)$. In other words, the slower rate of convergence of $\hat{c}_A^+$ and $\hat{c}_A^+$ dictates.

4. Through $\hat{s}$, the adaptive estimator depends on several user-chosen constants, namely ψ_1, ψ_2, s_*, and s^*. In Section 5 we use local cross validation to choose ψ_1 and ψ_2. For the bounds s_* and s^* we suggest using $1/\log(n)$ and ∞, respectively.

Theorems 2 and 3 suggest that $\hat{s}$ provides a consistent estimator of s if $m(x) \in \mathcal{M}(s, \delta, K)$. However, s is not well defined. According to our definition of smoothness,

a function that is smooth of order s_1 is also smooth of order s_2 whenever $s_1 > s_2$. The rate-optimal polynomial order and bandwidth are increasing functions of the smoothness and we are therefore interested in defining a class of functions with a unique smoothness index.

Before defining the new function class, recall that any function $m(x) \in \mathcal{M}(s, \delta, K)$ admits Taylor expansions of the form:

$$m(x) = m(x^*) + \sum_{j=1}^{\ell} b_j^+ (x - x^*)^j + \tilde{e}^+(x) \text{ for } x \geq x^* \tag{38}$$

$$m(x) = m(x^*) + \sum_{j=1}^{\ell} b_j^- (x - x^*)^j + \tilde{e}^-(x) \text{ for } x < x^* \tag{39}$$

with the remainder terms satisfying

$$\left|\tilde{e}^+(x)\right| / (x - x^*)^s \leq (\ell!)^{-1} K \text{ for } x \geq x^*, \ \left|\tilde{e}^-(x)\right| / \left|x - x^*\right|^s \leq (\ell!)^{-1} K \text{ for } x < x^*. \tag{40}$$

Let $\tilde{e}^+ = \{\tilde{e}^+(x_i)\}_{x_i \geq x^*}$ and $\tilde{e}^- = \{\tilde{e}^-(x_i)\}_{x_i < x^*}$ be the vectors that contain the remainder terms. The following definition imposes an additional condition on $\tilde{e}^+(x)$ and $\tilde{e}^-(x)$.

Definition 4 *Let $s_0 = \ell_0 + \tau_0$ where ℓ_0 is the largest integer strictly less than s_0 and $\tau_0 \in (0, 1]$. Let $\mathcal{M}_0(s_0, \delta, K)$ be the class of functions satisfying*

(i) $m(x) \in \mathcal{M}(s_0, \delta, K)$ but $m(x) \notin \mathcal{M}(s, \delta, K)$ for any $s > s_0$.

(ii) Let $D_{n\ell_0} = \sqrt{nh}\, diag(1, h, h^2, ..., h^{\ell_0})$. The remainder terms $\tilde{e}^+(x)$ and $\tilde{e}^-(x)$ of the ℓ_0-th order Taylor expansion of $m(x)$ around x^* satisfy

$$(nh)^{-1/2} h^{-s_0} \left\| \left(D_{n\ell_0}^{-1} Z_{\ell_0}^{+\prime} W^+ \tilde{e}^+ \right) - (-1)^{\ell_0 + 1} \left(D_{n\ell_0}^{-1} Z_{\ell_0}^{-\prime} W^- \tilde{e}^- \right) \right\| \geq C$$

for a constant $C > 0$ with probability approaching 1 as $n \to \infty, h \to 0$ such that $nh \to \infty$.

The first requirement in the above definition determines the 'maximum degree of smoothness' of a function. For an infinitely differentiable function, there is no s_0 such that the first requirement is met. In this case, we define s_0 to be ∞. In other words, $\mathcal{M}_0(\infty, \delta, K)$ is the set of infinitely differentiable functions. The second requirement asks for a lower bound for the asymptotic bias of the local polynomial estimator with order ℓ_0. These two requirements make $\mathcal{M}_0(s_0, \delta, K)$ a subset of $\mathcal{M}(s_0, \delta, K)$ which is the most difficult to estimate. Heuristically, if $m(x) \in \mathcal{M}_0(s_0, \delta, K)$, then there exists no estimator $\hat{\alpha}$ with the rate of convergence faster than $n^{-2s_0/(2s_0+1)+\Delta}$ for any $\Delta > 0$. For a function $m(x) \in \mathcal{M}(s_0, \delta, K) \cap \mathcal{M}(s, \delta, K)$ with $s > s_0$, it is easy to see that the estimator $\hat{\alpha}_s$ converges to α at the rate of $n^{-2s/(2s+1)}$

which is faster than the rate $n^{-2s_0/(2s_0+1)}$. To rule out this case, we impose the first requirement. On the other hand, when the first requirement is met but the asymptotic bias of $\hat{\alpha}_{s_0}$ diminishes as $n \to \infty$, possibly due to the cancellation of the asymptotic biases from the two sides, we can choose a large bandwidth without inflating the asymptotic bias and thus obtain a rate of convergence that is faster than $n^{-2s_0/(2s_0+1)}$. To rule out this case, we thus impose the second requirement.

Sufficient conditions for the second requirement are (i) $K_1 |x - x^*|^{s_0} \leq |\tilde{e}^+(x)| \leq K_2 |x - x^*|^{s_0}$ and $K_1 |x - x^*|^{s_0} \leq |\tilde{e}^-(x)| \leq K_2 |x - x^*|^{s_0}$ for some $K_1 > 0, K_2 > 0$ (ii) $\tilde{e}^+(x) \neq \tilde{e}^-(x)$ when ℓ_0 is odd.

The following theorem shows that $\hat{s}$ provides a consistent estimate for the maximal degree of smoothness.

Theorem 5 *Let the assumptions of Theorem 3 hold. If $m(x) \in \mathcal{M}_0(s_0, \delta, K)$ with $s_0 \geq s_* > 0$, then*

$$\hat{s} = \min(s_0, s^*) + O_p\left(\frac{\log\log n}{\log n}\right) \text{ as } n \to \infty.$$

Remarks

1. The theorem shows that $\hat{s}$ consistently estimates the maximal degree of smoothness s_0 when it is finite and s_* and s^* are appropriately chosen.

2. A direct implication of Theorem 5 is that $\hat{s}$ converges to s^* when $s^* \leq s_0$. As a result, when the sample size is not large in practical applications, we can set an upper bound that is relatively small. This will prevent us from using high order polynomials for small sample sizes. For example, when $s^* = 3$, the adaptive procedure effectively provides a method to choose between the local constant, local linear and local quadratic estimators. In the simulation study, we choose $s^* = 4$, which we feel is a reasonable choice for sample size 500.

3. The adaptive estimator $\hat{\alpha}_A$ is not necessarily asymptotically normal. At the cost of a slower rate of convergence, Theorem 5 enables us to define a new adaptive estimator that is asymptotically normal with zero asymptotic bias. More specifically, after obtaining $\hat{s}$ using the above adaptive procedure, we define

$$\hat{\alpha}^*_{\hat{s}} := \alpha_{\hat{s}}(r_{\hat{s}}, h^*_{\hat{s}}), \text{ where } h^*_s = \psi_1 n^{-1/(2r_s+1)}. \tag{41}$$

   If $s_0 < \infty$ and s_0 is not an integer, Theorem 5 implies that $r_{\hat{s}} = r_{s_0}$ with probability approaching one. Thus, both $r_{\hat{s}}$ and $h^*_{\hat{s}}$ are essentially non-random for large n. In

consequence, the adaptive estimator $\hat{\alpha}^*_{\hat{s}}$ is asymptotically normal:

$$\sqrt{nh^*_{\hat{s}}}\,(\hat{\alpha}^*_{\hat{s}} - \alpha) \to_d N(0, \omega^2 \lambda^2_{r_{\hat{s}}}). \tag{42}$$

Of course, one would expect that a given level of accuracy of approximation by the normal distribution would require a larger sample size when r and h are adaptively selected than otherwise.

4. The only unknown quantity in (42) is $\omega^2 = \left(\sigma^{2+}(x^*) + \sigma^{2-}(x^*)\right)/f(x^*)$. The density of x at the cut-off point, $f(x^*)$, can be estimated consistently by kernel methods. Given a consistent estimate $\tilde{\alpha}$, we define the estimated residual by

$$\tilde{\varepsilon}_i = y_i - \tilde{m}(x_i) - \tilde{\alpha} d_i \tag{43}$$

where

$$\tilde{m}(x_i) = \frac{\sum_{i=1}^n k_h(x - x_i)\,[y_i - \tilde{\alpha} d_i]}{\sum_{i=1}^n k_h(x - x_i)} \tag{44}$$

Porter (2003) shows that, under some regularity conditions,

$$\hat{\sigma}^{2+}(x^*) = \frac{2\sum_{i=1}^n k_h(x_i - x^*) d_i \tilde{\varepsilon}_i^2}{\sum_{i=1}^n k_h(x_i - x^*)} \text{ and } \hat{\sigma}^{2-}(x^*) = \frac{2\sum_{i=1}^n k_h(x_i - x^*)(1 - d_i)\tilde{\varepsilon}_i^2}{\sum_{i=1}^n k_h(x_i - x^*)} \tag{45}$$

are consistent for $\sigma^{2+}(x^*)$ and $\sigma^{2-}(x^*)$ respectively. Plugging $\hat{\sigma}^{2+}(x^*)$, $\hat{\sigma}^{2-}(x^*)$ and $\hat{f}(x^*) = 1/n \sum_{i=1}^n k_h(x_i - x^*)$ into the definition of ω^2 produces a consistent estimator for it. The adaptive estimator $\hat{\alpha}_{\hat{s}}$ or $\hat{\alpha}^*_{\hat{s}}$ can be used to compute the estimated residual in (43).

5 Monte Carlo Experiments

In this section, we propose a practical strategy to select the constants ψ_1 and ψ_2 in the adaptive procedure and provide some simulation evidence on the finite sample performance of the adaptive estimator.

The empirical strategy we use is based on the squared-error cross validation, which has had considerable influence on nonparametric estimation. Since our objective is to estimate the discontinuity at a certain point, we use a local version of cross validation proposed by Hall and Schuany (1989) for density estimation.

For each combination of (ψ_1, ψ_2), we first use the adaptive rule to determine $\hat{s}$, $h_{\hat{s}}$, and $r_{\hat{s}}$. We then use the local polynomial estimator with bandwidth $h_{\hat{s}}$ and polynomial order $r_{\hat{s}}$ to estimate the conditional mean of y_i at $x = x_i$ leaving the observation (x_i, y_i) out. Denote

the estimate by $\hat{y}_{-i}(\psi_1, \psi_2)$, where we have made it explicit that $\hat{y}_{-i}$ depends on (ψ_1, ψ_2). Let $\{x_{i_1}^+, ..., x_{i_m}^+\}$ and $\{x_{i_1}^-, ..., x_{i_m}^-\}$ be the closest m observations that are larger and smaller than x^* respectively. We choose ψ_1 and ψ_2 to minimize the local cross validation function:

$$CV(\psi_1, \psi_2) = \sum_{k=1}^{m} (y_{i_k}^+ - \hat{y}_{-i_k}^+(\psi_1, \psi_2))^2 + \sum_{k=1}^{m} (y_{i_k}^- - \hat{y}_{-i_k}^-(\psi_1, \psi_2))^2 \tag{46}$$

Finally we use the cross validation choice $(\hat{\psi}_1, \hat{\psi}_2)$ of (ψ_1, ψ_2) to compute the adaptive estimator, which is denoted by $\hat{\alpha}_A(\hat{\psi}_1, \hat{\psi}_2)$.

In this paper, we do not provide asymptotic results for $\hat{\alpha}_A(\hat{\psi}_1, \hat{\psi}_2)$, but we do give some simple results for an estimator based on a data-dependent method that is close to $(\hat{\psi}_1, \hat{\psi}_2)$. Let $\Psi = \{\Psi_1, ..., \Psi_U\}$ be a finite grid of positive real numbers. Take $(\tilde{\psi}_1, \tilde{\psi}_2)$ to be the closest point in $\Psi \times \Psi$ to $(\hat{\psi}_1, \hat{\psi}_2)$. Let $\hat{\alpha}_A(\tilde{\psi}_1, \tilde{\psi}_2)$ denote the adaptive estimator based on $(\tilde{\psi}_1, \tilde{\psi}_2)$. One can take the grid size of Ψ to be sufficiently small that the minimum of $CV(\psi_1, \psi_2)$ over $(\psi_1, \psi_2) \in \Psi \times \Psi$ is quite close to its minimum over $\mathbb{R}^+ \times \mathbb{R}^+$, at least if one has knowledge of suitable lower and upper bounds for ψ_1 and ψ_2.

The asymptotic behavior of $\hat{\alpha}_A(\tilde{\psi}_1, \tilde{\psi}_2)$ is relatively easy to obtain. First, Theorem 3 holds for $\hat{\alpha}_A(\tilde{\psi}_1, \tilde{\psi}_2)$ under Assumptions 2 and 3. The reasons are that the theorem holds for $\hat{\alpha}_A$ for each combination $(\psi_1, \psi_2) \in \Psi \times \Psi$ and that there are a finite number of such combinations. So, $\hat{\alpha}_A(\tilde{\psi}_1, \tilde{\psi}_2)$ is consistent and has the rate of convergence given by $n^{\frac{s}{2s+1}} \zeta^{-1}(n)$. Second, suppose the value $(\hat{\psi}_1, \hat{\psi}_2)$ is not equidistant to any two points in $\Psi \times \Psi$ (which fails only for a set of points with Lebesgue measure zero) and assume that $(\hat{\psi}_1, \hat{\psi}_2)$ converges to (ψ_1^*, ψ_2^*) in large samples. Let (ψ_1^o, ψ_2^o) be the closest point in $\Psi \times \Psi$ to (ψ_1^*, ψ_2^*) . Let $\hat{\alpha}_A(\psi_1^o, \psi_2^o)$ and $\hat{\alpha}_A(\psi_1^*, \psi_2^*)$ denote the adaptive estimators based on (ψ_1^o, ψ_2^o) and (ψ_1^*, ψ_2^*) respectively. Then, the asymptotic distribution of $\hat{\alpha}_A(\tilde{\psi}_1, \tilde{\psi}_2) - \alpha$ is the same as that of $\hat{\alpha}_A(\psi_1^o, \psi_2^o) - \alpha$. This holds because $(\tilde{\psi}_1, \tilde{\psi}_2) = (\psi_1^o, \psi_2^o)$ with probability that goes to 1 as $n \to \infty$ by the discreteness of Ψ. After a simple modification along the line of (41), we have

$$\sqrt{nh_{\hat{s}}^*} \left(\hat{\alpha}_A^*(\tilde{\psi}_1, \tilde{\psi}_2) - \alpha \right) \to_d N(0, \omega^2 \lambda_{r_{\hat{s}}}^2). \tag{47}$$

where $\hat{\alpha}_A^*(\tilde{\psi}_1, \tilde{\psi}_2)$ is the same as $\hat{\alpha}_A(\tilde{\psi}_1, \tilde{\psi}_2)$, except that the bandwidth $h_{\hat{s}} = \tilde{\psi}_1 n^{-1/(2\hat{s}+1)}$ is replaced by $h_{\hat{s}}^* = \tilde{\psi}_1 n^{-1/(2r_{\hat{s}}+1)}$.

The above theoretical results for $\hat{\alpha}_A^*(\tilde{\psi}_1, \tilde{\psi}_2)$ are not entirely satisfactory because they require the use of the somewhat artificial grid Ψ. Nevertheless, in the absence of asymptotic results for $\hat{\alpha}_A(\hat{\psi}_1, \hat{\psi}_2)$, they should be useful. Since our cross validation algorithm is based on a grid search, we effectively use the estimator $\hat{\alpha}_A(\tilde{\psi}_1, \tilde{\psi}_2)$ in our simulations.

In our Monte Carlo experiment, we let $s^* = 4$, $m = 0.1n$, and $\Psi = \{0.1, 0.5, 1, 5\}$ to compute the adaptive estimator. To evaluate the finite sample performance of the adaptive

estimator $\hat{\alpha}_A(\tilde{\psi}_1, \tilde{\psi}_2)$, we compare it with the local constant, local linear, local quadratic and local cubic estimators, each of them using the locally cross-validated bandwidth. For these local polynomial estimators, we use the AMSE-optimal bandwidth $h = cn^{-1/(2r+3)}$ and choose c over the set $\mathcal{C} = (0.1, 0.2, ..., 1) \cup (2, 3, 4, ..., 10)$ via cross validation. For each estimator the cross validation is based on the same neighborhood observations $\{x_{i_1}^+, ..., x_{i_m}^+\}$ and $\{x_{i_1}^-, ..., x_{i_m}^-\}$ and uses the grid search method. We have considered other choices of m, Ψ and $\mathcal{C}$ but the qualitative results are similar.

We consider three groups of experiments. In the first group, the data generating process is $y_i = m(x_i) + \alpha \times 1\{x_i > x^*\} + \varepsilon_i$ where $\alpha = 1$ and

$$m(x_i) = \begin{cases} \sum_{i=1}^{3}(x_i - x^*)^i + \kappa |x_i - x^*|^{s_0} & \text{for } x_i \geq x^*; \\ \sum_{i=1}^{3}(x_i - x^*)^i - \kappa |x_i - x^*|^{s_0} & \text{for } x_i < x^*. \end{cases} \tag{48}$$

Both x_i and ε_i are iid standard normal. $\{x_i\}_{i=1}^n$ is independent $\{\varepsilon_i\}_{i=1}^n$. We set $x^* = 0$ without loss of generality. We consider several values for s_0, i.e. $s_0 = 1/2, 3/2, 5/3, 7/2$ and two values for κ, i.e. $\kappa = 1$ and 5. s_0 characterizes the smoothness of $m(x)$ while κ determines the importance of the not-so-smooth component in $m(x)$.

For the second group of experiments, the data generating process is the same as the one above except that a sine wave is added to $m(x)$, leading to

$$m(x_i) = \begin{cases} \sum_{i=1}^{3}(x_i - x^*)^i + 5\sin 10(x_i - x^*) + \kappa |x_i - x^*|^{s_0} & \text{for } x_i \geq x^* \\ \sum_{i=0}^{3}(x_i - x^*)^i + 5\sin 10(x_i - x^*) - \kappa |x_i - x^*|^{s_0} & \text{for } x_i < x^* \end{cases} \tag{49}$$

The response function we just defined has a finer structure than that given in (48). Such a response function may not be realistic in empirical applications but it is used to examine the finite sample performances of different estimators in the worst situations.

For the last group of experiments, the data generating process is

$$m(x_i) = \sum_{i=0}^{k} 10(x_i - x^*)^i, \text{ for } k = 0, 1, 2 \text{ or } 3. \tag{50}$$

Since $m(x_i)$ is a constant, linear, quadratic or cubic function, we expect the local constant, local linear, local quadratic and local cubic estimators to have the best finite sample performances in the respective cases of $k = 0, 1, 2$ and 3. The motivation for considering this group is to 'crash' test the adaptive estimator against the local polynomial estimators.

For each group of the Monte Carlo experiments, we compute the bias, standard deviation (SD) and root mean square error (RMSE) of all estimators considered. The number of replication is 1000 and the sample size is 500. More specifically, for an estimator $\hat{\alpha}$, the

bias, SD, and RMSE are computed according to

$$\text{bias} = \overline{\hat{\alpha}} - \alpha,\ \text{SD} = \frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\alpha}_i - \overline{\hat{\alpha}}\right)^2 \text{ and RMSE} = \sqrt{(\text{bias})^2 + (\text{SD})^2} \tag{51}$$

where $\overline{\hat{\alpha}} = 1/1000\sum_{m=1}^{1000}\hat{\alpha}_m$ and $\hat{\alpha}_m$ is the estimate for the m-th replication.

Table I presents the results for the first group of experiments. It is clear that the local constant estimator has the smallest standard deviation and the largest bias. When $s_0 = 3/2, 5/2, 7/2$, the slope of $m(x)$ is relatively flat at $x = x^*$. As a result, the effect of the standard deviation outweighs that of the bias. It is not surprising that the local constant estimator has the smallest RMSE in these cases. However, when $s_0 = 1/2$, the function $m(x)$ becomes very steep at $x = x^*$. As expected, the local constant estimator has a large upward bias and the largest RMSE. Next, for the rest of the local polynomial estimators, the absolute values of the biases are in general comparable while the standard deviation decreases with the order of the polynomial. The latter result seems to be counter-intuitive at first sight. However, as the order of the polynomial increases, the cross-validated bandwidth also increases. Note that the bandwidth and polynomial order have opposite effects on the variance of the local polynomial estimators. In finite samples, it is likely that the variance reduction from using a larger bandwidth dominates the variance inflation from using a higher order polynomial. This is the case for the first group of data generating processes we consider. Finally, the performance of the adaptive estimator is very robust to the parameter configurations. When the underlying process is not so smooth ($s_0 = 1/2, \kappa_0 = 1$), the adaptive estimator has the smallest RMSE. In other cases, the RMSE of the adaptive estimator is only slightly larger than the smallest RMSE. It is important to note that the smallest RMSE is achieved by different estimators for different parameter combinations.

Table II reports the results for the second group of experiments. We report only the case $\kappa = 1$ as it is representative of the case $\kappa = 5$. Due to the rapid slope changes in the response function, all estimators have much larger RMSE's than those given in Table I. While the local constant estimator has a satisfactory RMSE performance in Table I, its RMSE performance is the poorest because of the large bias. The best estimator, according to the RMSE criterion, is the local linear estimator whose absolute bias is the smallest among the local polynomial estimators and standard deviation is only slightly larger than that of the local constant estimator. Compared with the local polynomial estimators, the adaptive estimator has the smallest bias for all parameter combinations while its variance is comparable to that of the local linear estimator. As a consequence, the RMSE performance of the adaptive estimator is quite satisfactory.

Table III gives the result for the last group of experiments. As expected, when the

response function is a polynomial with order r, the local polynomial estimator with the same order has the best finite sample performance in general. An exception is the local linear estimator whose RMSE is larger than that of the local quadratic and cubic estimators. The performance of the adaptive estimator is very encouraging. Its RMSE is either the smallest or slightly larger than that of the estimator which is most suitable for the underlying data generating process.

To sum up, the RMSE of the adaptive estimator is either the smallest or among the smallest ones. The performance of the adaptive estimator is robust to the underlying data generating process. In contrast, a local polynomial estimator may have the best performance in one scenario and disastrous performances in other scenarios. For example, the local constant estimator performs well in the first group of experiments but performs poorly in the second group of experiments. The local linear estimator has a satisfactory performance in the second group of experiments but its performance is the worst in the first group of experiments. The adaptive estimator seems to be the best estimator in an overall sense.

Table I: Finite Sample Performances of Different Estimators
When $m(x_i) = \sum_{i=1}^{3}(x_i - x^*)^i + \kappa\,|x_i - x^*|^{s_0}\, sign(x_i - x^*)$

	Adaptive Estimator	Local Constant	Local Linear	Local Quadratic	Local Cubic
$(s_0, \kappa) = (1/2, 1)$					
Bias	0.2710	0.5085	0.1688	0.3075	0.3097
SD	0.4807	0.4798	0.6722	0.5813	0.4861
RMSE	0.5517^{1}	0.6990	0.6927	0.6574^{3}	0.5762^{2}
$(s_0, \kappa) = (3/2, 1)$					
Bias	-0.0639	0.1646	-0.1086	0.0775	0.0157
SD	0.4894	0.4259	0.6983	0.5507	0.4456
RMSE	0.4933^{3}	0.4564^{2}	0.7063	0.5558	0.4459^{1}
$(s_0, \kappa) = (5/2, 1)$					
Bias	-0.0537	0.1392	-0.0929	0.0933	0.0301
SD	0.4818	0.4129	0.7006	0.5571	0.4473
RMSE	0.4845^{3}	0.4356^{1}	0.7064	0.5646	0.4480^{2}
$(s_0, \kappa) = (7/2, 1)$					
Bias	-0.0663	0.1349	-0.0776	0.0922	-0.0450
SD	0.4776	0.4049	0.6979	0.5654	0.4498
RMSE	0.4819^{3}	0.4275^{1}	0.7019	0.5726	0.4518^{2}
$(s_0, \kappa) = (1/2, 5)$					
Bias	1.1136	1.5467	1.0278	1.0667	1.1755
SD	0.7618	0.6399	0.8771	0.7423	0.6884
RMSE	1.3490^{2}	1.6737	1.3509	1.2994^{1}	1.3621^{3}
$(s_0, \kappa) = (3/2, 5)$					
Bias	-0.0668	0.2178	-0.1801	0.0248	0.0017
SD	0.5002	0.4667	0.7462	0.5323	0.4938
RMSE	0.5044^{2}	0.5148^{3}	0.7672	0.5326	0.4938^{1}
$(s_0, \kappa) = (5/2, 5)$					
Bias	0.0318	0.1373	-0.1122	0.1130	0.0906
SD	0.4643	0.4262	0.7404	0.5749	0.4548
RMSE	0.4651^{3}	0.4476^{1}	0.7485	0.5856	0.4635^{2}
$(s_0, \kappa) = (7/2, 5)$					
Bias	-0.1153	0.1310	-0.0794	0.0879	-0.0838
SD	0.5387	0.4131	0.7095	0.6064	0.4970
RMSE	0.5506^{3}	0.4332^{1}	0.7136	0.6124	0.5038^{2}

The superscripts $1, 2, 3$ indicate the smallest, second smallest, and third smallest RMSE in each row, respectively

Table II: Finite Sample Performances of Different Estimators
When $m(x_i) = \sum_{i=1}^{3}(x_i - x^*)^i + 5\sin 10(x_i - x^*) + \kappa\,|x_i - x^*|^{s_0}\, sign(x_i - x^*)$

	Adaptive Estimator	Local Constant	Local Linear	Local Quadratic	Local Cubic
$(s_0, \kappa) = (1/2, 1)$					
Bias	0.0203	1.8541	0.1991	0.2331	0.4653
SD	1.2396	0.9792	1.0803	1.1507	1.7428
RMSE	1.2398^3	2.0965	1.0979^1	1.1735^2	1.8030
$(s_0, \kappa) = (3/2, 1)$					
Bias	-0.0596	1.6518	0.0738	0.1394	0.3368
SD	1.2646	0.9398	1.0732	1.1760	1.6929
RMSE	1.2654^3	1.9002	1.0752^1	1.1836^2	1.7253
$(s_0, \kappa) = (5/2, 1)$					
Bias	-0.0573	1.6481	0.0756	0.1491	0.3326
SD	1.2651	0.9369	1.0749	1.1811	1.6782
RMSE	1.2657^3	1.8956	1.0770^1	1.1899^2	1.7100
$(s_0, \kappa) = (7/2, 1)$					
Bias	-0.0560	1.6476	0.0769	0.1487	0.3284
SD	1.2680	0.9370	1.0755	1.1810	1.6761
RMSE	1.2686^3	1.8952	1.0777^1	1.1897^2	1.7072

The superscripts $1, 2, 3$ indicate the smallest, second smallest, and third smallest values in each row, respectively

Table III Finite Sample Performances of Different Estimators
for Different Response Functions

	Adaptive Estimator	Local Constant	Local Linear	Local Quadratic	Local Cubic
			$m(x) = 0$		
Bias	-0.0287	-0.0228	-0.0128	-0.0089	-0.0250
SD	0.4287	0.3554	0.6243	0.5015	0.4437
RMSE	0.4294^{2}	0.3559^{1}	0.6242	0.5014	0.4442^{3}
			$m(x) = 10(x - x^*)$		
Bias	0.0204	0.5789	-0.0127	0.0563	0.1073
SD	0.5273	0.5411	0.6244	0.5355	0.5816
RMSE	0.5274^{1}	0.7922	0.6245	0.5382^{2}	0.5911^{3}
			$m(x) = 10(x - x^*) + 10(x - x^*)^2$		
Bias	0.0198	0.5876	-0.0079	0.0522	0.1198
SD	0.5304	0.5491	0.7809	0.5380	0.5881
RMSE	0.5305^{1}	0.8040	0.7809	0.5402^{2}	0.5999^{3}
			$m(x) = 10(x - x^*) + 10(x - x^*)^2 + 10(x - x^*)^3$		
Bias	0.0991	0.5763	-0.0115	0.1630	0.1177
SD	0.6317	0.5436	0.7470	0.6464	0.5949
RMSE	0.6391^{2}	0.7920	0.7471	0.6663^{3}	0.6064^{1}

6 Appendix of Proofs

Proof of Theorem 1. It is easy to show that

$$\hat{\theta}_r^+ - \theta_r^+ = \left(Z_r^{+\prime} W^+ Z_r^+\right)^{-1} Z_r^{+\prime} W^+ \varepsilon^+ + \left(Z_r^{+\prime} W Z_r^+\right)^{-1} Z_r^{+\prime} W^+ R^+ \tag{A.1}$$

Let

$$D_{nr} = \sqrt{nh}\text{diag}(1, h, h^2, ..., h^r). \tag{A.2}$$

Then

$$\begin{aligned} & D_{nr}\left(\hat{\theta}_r^+ - \theta_r^+\right) \qquad (A.3) \\ = \; & \left(D_{nr}^{-1} Z_r^{+\prime} W^+ Z_r^+ D_{nr}^{-1}\right)^{-1} D_{nr}^{-1} Z_r^{+\prime} W^+ \varepsilon^+ + \left(D_{nr}^{-1} Z_r^{+\prime} W^+ Z_r^+ D_{nr}^{-1}\right)^{-1} D_{nr}^{-1} Z_r^{+\prime} W^+ R^+. \end{aligned}$$

It follows from the proof of Lemma A.1(a) below that

$$\text{p}\lim_{n\to\infty} D_{nr}^{-1} Z_r^{+\prime} W^+ Z_r^+ D_{nr}^{-1} = f(x^*)\Gamma_r. \tag{A.4}$$

Porter (2003) shows that, under Assumption 2,

$$D_{nr}^{-1} Z_r^{+\prime} W^+ \varepsilon^+ \Rightarrow N\left(0,\ \frac{\sigma^{2+}(x^*)}{f(x^*)} V_r\right). \tag{A.5}$$

Combining (A.3), (A.4) and (A.5) gives

$$\begin{aligned} & D_{n,r}\left(\hat{\theta}_r^+ - \theta_r^+\right) - \left(D_{nr}^{-1} Z_r^{+\prime} W^+ Z_r^+ D_{nr}^{-1}\right)^{-1} D_{nr}^{-1} Z_r^{+\prime} W^+ R^+ \\ \Rightarrow \; & N\left(0, \frac{\sigma^{2+}(x^*)}{f(x^*)} \Gamma_r^{-1} V_r \Gamma_r^{-1}\right), \end{aligned} \tag{A.6}$$

which implies

$$\sqrt{nh}(\hat{c}_r^+ - c^+) - B^+ \Rightarrow N\left(0,\ \frac{\sigma^{2+}(x^*)}{f(x^*)} e_1' \Gamma_r^{-1} V_r \Gamma_r^{-1} e_1\right), \tag{A.7}$$

where $B_+ = e_1'\left(D_{nr}^{-1} Z_r^{+\prime} W^+ Z_r^+ D_{nr}^{-1}\right)^{-1} D_{nr}^{-1} Z_r^{+\prime} W^+ R^+$.

Similarly, we can show that

$$\sqrt{nh}(\hat{c}_r^- - c^-) - B^- \Rightarrow N\left(0,\ \frac{\sigma^{2-}(x^*)}{f(x^*)} e_1' \Gamma_r^{-1} V_r \Gamma_r^{-1} e_1\right). \tag{A.8}$$

By the independence of $\sqrt{nh}(\hat{c}_r^+ - c^+)$ and $\sqrt{nh}(\hat{c}_r^- - c^-)$, we get

$$\sqrt{nh}(\hat{\alpha}_r - \alpha) - (B^+ - B^-) \Rightarrow N\left(0, \frac{\sigma^{2+}(x^*) + \sigma^{2-}(x^*)}{f(x^*)} e_1' \Gamma_r^{-1} V_r \Gamma_r^{-1} e_1\right). \tag{A.9}$$

When $\ell \geq r+1$,

$$D_{nr}^{-1} Z_r^{+\prime} W^+ R^+ = h^{r+1} \sqrt{nh} b_{r+1}^+ \mu_r (1 + o_p(1)). \tag{A.10}$$

When $\ell \leq r$,

$$D_{nr}^{-1} Z_r^{+\prime} W^+ R^+ = \sqrt{nh} \left(D_{nr}^{-1} Z_r^{+\prime} W^+ e^+ \right) = O_p \left(h^q \sqrt{nh} \right). \tag{A.11}$$

Therefore

$$B^+ = 1\{\ell \geq r+1\} \frac{\left(e_1' \Gamma_r^{-1} \mu_r\right) b_{r+1}^+}{f(x^*)} h^{r+1} \sqrt{nh} (1 + o_p(1)) + O_p \left(h^q \sqrt{nh} \right). \tag{A.12}$$

Similarly

$$B^- = 1\{\ell \geq r+1\} \frac{(-1)^{r+1} \left(e_1' \Gamma_r^{-1} \mu_r\right) b_{r+1}^-}{f(x^*)} h^{r+1} \sqrt{nh} (1 + o_p(1)) + O_p \left(h^q \sqrt{nh} \right). \tag{A.13}$$

Let $B = B^+ - B^-$, then

$$\begin{aligned} B &= 1\{\ell \geq r+1\} \frac{\left(e_1' \Gamma_r^{-1} \mu_r\right) \left[b_{r+1}^+ - (-1)^{r+1} b_{r+1}^-\right]}{f(x^*)} h^{r+1} \sqrt{nh} (1 + o_p(1)) \\ &\quad + O_p \left(h^q \sqrt{nh} \right). \end{aligned} \tag{A.14}$$

Combining (A.14) and (A.9) leads to the desired result. ■

Proof of Theorem 2. Part (a). The proof uses the following result from Pollard (1993): Let $P = \Pi_{i=1}^n P_i$ and $Q = \Pi_{i=1}^n Q_i$ be the finite products of probability measures such that Q_i has density $1 + \Delta_i(\cdot)$ with respect to P_i. If $\nu_i^2 = E_{P_i} \Delta_i^2$ is finite for each i, then

$$||\Pi_{i=1}^n P_i - \Pi_{i=1}^n Q_i||_1 \leq \exp\left(\sum_{i=1}^n \nu_i^2 \right) - 1. \tag{A.15}$$

Using this result and (31), we have

$$\inf_{\hat{\alpha}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(|\hat{\alpha} - \alpha| \geq \epsilon/2) \geq \frac{1}{2} \left(\frac{3}{2} - \exp\left(\sum_{i=1}^n \nu_i^2 \right) \right), \tag{A.16}$$

provided that $\alpha(P) - \alpha(Q) > \epsilon$.

To get a good lower bound for the minimax risk, we consider two probability models P and Q. Under the model P, the data is generated according to

$$Y = m_P(X) + \alpha_P d + \varepsilon \tag{A.17}$$

where $Y = (y_1, y_2, ..., y_n)'$, $m_P(X) = (m_P(x_1), ..., m_P(x_n))$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$, $x_i \sim iid$ uniform$(x^* - \delta, x^* + \delta)$, $\varepsilon_i \sim iid\ N(0,1)$ and ε_i is independent of x_j for all i and j. The

data generating process under Q is defined analogously with $m_P(X) + \alpha_P d$ replaced by $m_Q(X) + \alpha_Q d$. It is obvious that both models P and Q satisfy Assumption 2.

We now specify m and α for each model. For the probability model P, we let $m_P(x) = 0$ and $\alpha_P = 0$. For the probability model Q, we let

$$m_Q(x) = -\xi\eta^s \phi\left(\left(x - x^*\right)/\eta\right) \text{ and } \alpha_Q = \xi\eta^s \tag{A.18}$$

where $\eta = n^{-1/(2s+1)}$ and ϕ is an infinitely differentiable function satisfying (i) $0 \leq \phi(x) \leq 1$, (ii) $\phi(x) = 0$ for $x \leq 0$ and (iii) $\phi(x) = 1$ for $x \geq \delta$.

Obviously $m_P \in \mathcal{M}(s, \delta, K)$. We next verify that $m_Q \in \mathcal{M}(s, \delta, K)$. First, by construction, m_Q is continuous on $[x^* - \delta, x^* + \delta]$. Second, the i-th order derivative of $m_Q^{(i)}$ is $\xi\eta^{s-i}\phi^{(i)}\left(\left(x - x^*\right)/\eta\right)$ which is obviously bounded by K when n is large enough for all $i \leq \ell$. Third, we verify the Hölder condition for the ℓ-th order derivative. It suffices to consider the case when $x_1 \in [x^*, x^* + \delta]$ and $x_2 \in [x^*, x^* + \delta]$ as the Hölder condition holds trivially when $x_1 \in [x^* - \delta, x^*]$ and $x_2 \in [x^* - \delta, x^*]$. We consider three cases: (i) when x_1, $x_2 \in [x^*, x^* + \delta\eta]$, the ℓ-th order derivative satisfies

$$\begin{aligned}
& \left|\xi\eta^{s-\ell}\phi^{(\ell)}\left(\left(x_1 - x^*\right)/\eta\right) - \xi\eta^{s-\ell}\phi^{(\ell)}\left(\left(x_2 - x^*\right)/\eta\right)\right| \\
\leq\ & \xi\eta^{s-\ell}\phi^{(\ell+1)}\left(\eta^{-1}\tilde{x}\right)\eta^{-1}\left|x_1 - x_2\right| \\
=\ & \xi\eta^{s-\ell-1}\phi^{(\ell+1)}\left(\eta^{-1}\tilde{x}\right)\left|x_1 - x_2\right|^{\ell+1-s}\left|x_1 - x_2\right|^{s-\ell} \\
\leq\ & C\xi\eta^{s-\ell-1}\eta^{\ell+1-s}\delta^{\ell+1-s}\left|x_1 - x_2\right|^{\tau} \\
\leq\ & K\left|x_1 - x_2\right|^{\tau}
\end{aligned} \tag{A.19}$$

if ξ is small enough; (ii) when $x_1 \in [x^*, x^* + \eta\delta]$ and $x_2 \geq x^* + \eta\delta$,

$$\begin{aligned}
& \left|\xi\eta^{s-\ell}\phi^{(\ell)}\left(\left(x_1 - x^*\right)/\eta\right) - \xi\eta^{s-\ell}\phi^{(\ell)}\left(\left(x_2 - x^*\right)/\eta\right)\right| \\
=\ & \left|\xi\eta^{s-\ell}\phi^{(\ell)}\left(\left(x_1 - x^*\right)/\eta\right) - \xi\eta^{s-\ell}\phi^{(\ell)}\left(\left(x^* + \eta\delta - x^*\right)/\eta\right)\right| \\
\leq\ & K\left|x_1 - x^* - \eta\delta\right|^{\tau} \leq K\left|x_1 - x_2\right|^{\tau}
\end{aligned} \tag{A.20}$$

when the first inequality follows from (A.19); (iii) when $x_1 \geq x^* + \eta\delta$ and $x_2 \geq x^* + \eta\delta$, we have $\phi^{(\ell)}\left(\left(x_1 - x^*\right)/\eta\right) = \phi^{(\ell)}\left(\left(x_2 - x^*\right)/\eta\right) = 0$. Again the Hölder condition holds trivially.

It remains to compute the L_1 distance between the two measures. Let the density of Q_i with respect to P_i be $1 + \Delta_i(x_i, y_i)$, then

$$\Delta_i(x_i, y_i) = \begin{cases} \varphi\left(y_i - m_Q(x_i) - \alpha_Q\right)/\varphi(y_i) - 1, & \text{if } x_i \in [x^*, x^* + \eta\delta) \\ 0, & \text{otherwise} \end{cases} \tag{A.21}$$

where $\varphi(\cdot)$ is the standard normal pdf. Therefore,

$$\begin{aligned} E_{P_i}\Delta_i^2 &= \frac{1}{2\delta}\int_{x^*}^{x^*+\eta\delta}\int_{-\infty}^{\infty}\left[\varphi\left(y-m_Q(x)-\alpha_Q\right)\varphi^{-1}(y)-1\right]^2\varphi(y)dydx \\ &= \frac{1}{2\delta}\int_{x^*}^{x^*+\eta\delta}\int_{-\infty}^{\infty}\varphi^2\left(y-m_Q(x)-\alpha_Q\right)\varphi^{-1}(y)dydx \\ &\quad -\frac{1}{\delta}\int_{x^*}^{x^*+\eta\delta}\int_{-\infty}^{\infty}\varphi\left(y-m_Q(x)-\alpha_Q\right)dydx+\frac{1}{2}\eta \\ &= \frac{1}{2\delta}\int_{x^*}^{x^*+\eta\delta}\int_{-\infty}^{\infty}\varphi^2\left(y-m_Q(x)-\alpha_Q\right)\varphi^{-1}(y)dydx-\frac{1}{2}\eta. \end{aligned} \tag{A.22}$$

Plugging the standard normal pdf yields:

$$\begin{aligned} E_{P_i}\Delta_i^2 &= \frac{1}{2\delta}\int_{x^*}^{x^*+\eta\delta}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{2\left(y-m_Q(x)-\alpha_Q\right)^2}{2}+\frac{y^2}{2}\right)dydx-\frac{1}{2}\eta \\ &= \frac{1}{2\delta}\int_{x^*}^{x^*+\eta\delta}\exp\left(m_Q(x)+\alpha_Q\right)^2dx-\frac{1}{2}\eta \\ &= \frac{1}{2\delta}\int_{x^*}^{x^*+\eta\delta}\exp\left\{\xi^2\eta^{2s}\left[1-\phi(\eta^{-1}\left(x-x^*\right))\right]^2\right\}dx-\frac{1}{2}\eta \\ &\leq \frac{1}{2}\eta\exp\left(\xi^2\eta^{2s}\right)-\frac{1}{2}\eta=\frac{1}{2}\eta\left(\exp\left(\xi^2\eta^{2s}\right)-1\right) \\ &= \frac{1}{2}\xi^2\eta^{2s+1}(1+o(1))\leq\xi^2/(2n) \end{aligned} \tag{A.23}$$

when n is large enough.

When ξ is small enough, say $\xi^2/2\leq\log(5/4)$, we have

$$\exp\left(\sum_{i=1}^{n}\nu_i^2\right)\leq\exp\left(\xi^2/2\right)<\frac{5}{4}. \tag{A.24}$$

It follows from (A.16) that

$$\inf_{\hat{\alpha}}\sup_{P_{m,\alpha}\in\mathcal{P}(s,\delta,K)}P_{m,\alpha}\left(\left|n^{\frac{s}{2s+1}}\left(\hat{\alpha}-\alpha\right)\right|\geq\epsilon/2\right)\geq\frac{1}{2}(\frac{3}{2}-\frac{5}{4})=\frac{1}{8}\geq C \tag{A.25}$$

on choosing $C\leq 1/8$. Here the second inequality holds because $\alpha(P)-\alpha(Q)=\xi n^{-\frac{s}{2s+1}}\geq\epsilon n^{-\frac{s}{2s+1}}$ for a small ϵ.

Part (b). It follows from Theorem 1 that $\lim_{\epsilon\to\infty}P(n^{\frac{s}{2s+1}}\left[\hat{\alpha}_\ell-\alpha\right]\geq\epsilon/2)=0$ for a single probability model and a single bandwidth. This is because Theorem 1 holds and when $h=\psi_1 n^{-1/(2s+1)}$, the bias term satisfies $B=O_p\left(h^s\sqrt{nh}\right)=O_p(1)$. Hence, it suffices to show that the results of Theorem 1 hold uniformly over $P_{m,\alpha}\in\mathcal{P}(s,\delta,K)$. We focus on the case $x\geq x^*$ as the case for the $x<x^*$ follows in a similar way. Inspection of the proof of Theorem 1 shows that all quantities except $\left(D_{nr}^{-1}Z_r^{+\prime}W^+Z_r^+D_{nr}^{-1}\right)^{-1}D_{nr}^{-1}Z_r^{+\prime}W^+R^+$ are independent of m and α. So we only need to show that $\left(D_{nr}^{-1}Z_r^{+\prime}W^+Z_r^+D_{nr}^{-1}\right)^{-1}D_{nr}^{-1}Z_r^{+\prime}W^+R^+$

is stochastically bounded uniformly over $m \in \mathcal{M}(s,\delta,K)$ and $\alpha \in [-K,K]$. This is obvious as (i) $\left(D_{nr}^{-1}Z_r^{+\prime}W^+Z_r^+D_{nr}^{-1}\right)^{-1}$ does not depend on m and α and $D_{nr}^{-1}Z_r^{+\prime}W^+R^+$ is uniformly bounded because

$$\left|e^+(x)/\left(x-x^*\right)^q\right| = O(1) \text{ uniformly over } x \in [x^*, x^*+\delta].$$

■

To prove Theorems 3 and 5, we need the following two lemmas. For notational convenience, when $r = r_\tau, h_\tau = \psi_1 n^{-1/(2\tau+1)}$, we write $Z_\tau^+ = Z_{r_\tau}^+, D_{n\tau}^+ = D_{nr_\tau}^+$ and $W_\tau^+ = W^+ = h_\tau diag(k_{h_\tau}(x_i - x^*))_{x_i \geq x^*}$. Define Z_τ^-, $D_{n\tau}^-$ and W_τ^- analogously. Let $\sup_{(s,P_{m,\alpha})}$ abbreviate $\sup_{s\in[s_*,s^*]}\sup_{P_{m,\alpha}\in\mathcal{P}(s,\delta,K)}$ throughout the rest of the proof.

Lemma A.1 *Let Assumptions 2(d) and 3 hold. If* $\min_{r\in[r_{s*},r_{s^*}]}\{\mu_{\min}(\Gamma_r)\} > 0$, *then for some constant* C_2 *and any constant* C_3 *we have, as* $n \to \infty$,

(a) $\sup_{(s,P_{m,\alpha})} P_{m,\alpha}\left(\inf_{\tau\in[s_*,s^*]}\left\{\mu_{\min}\left(D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+Z_\tau^+D_{n\tau}^{-1}\right)\right\} \leq C_2\right) = o(1)$,

(b) $\sup_{(s,P_{m,\alpha})} P_{m,\alpha}\left(\inf_{\tau\in[s_*,s^*]}\left\{\mu_{\min}\left(D_{n\tau}^{-1}Z_\tau^{-\prime}W_\tau^-Z_\tau^-D_{n\tau}^{-1}\right)\right\} \leq C_2\right) = o(1)$,

(c) $\sup_{(s,P_{m,\alpha})} P_{m,\alpha}\left(\sup_{\tau\in[s_*,s^*]}\left\|D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+R_\tau^+\right\| > C_3\zeta(n)\right) = o(1)$,

(d) $\sup_{(s,P_{m,\alpha})} P_{m,\alpha}\left(\sup_{\tau\in[s_*,s^*]}\left\|D_{n\tau}^{-1}Z_\tau^{-\prime}W_\tau^-R_\tau^-\right\| > C_3\zeta(n)\right) = o(1)$,

where $\mu_{\min}(A)$ *is the smallest eigenvalue of matrix A.*

Proof of Lemma A.1. **Part (a)** Let $\Gamma_{n\tau} = D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+Z_\tau^+D_{n\tau}^{-1}$, then the (i,j)-th element of $\Gamma_{n\tau}$ is

$$\Gamma_{n\tau}(i,j) = \frac{1}{nh_\tau}\frac{1}{h_\tau^{i+j-2}}\sum_{k=1}^n k(\frac{x_k - x^*}{h_\tau})\left(x_k - x^*\right)^{i+j-2}1\left\{x_k \geq x^*\right\}. \tag{A.26}$$

Note that

$$E\Gamma_{n\tau}(i,j) = \int_0^\infty k(z)z^{i+j+2}f(x^* + zh_\tau)dz \tag{A.27}$$

and

$$\begin{aligned}
& Var(\Gamma_{n\tau}(i,j)) \\
= & \frac{1}{nh_\tau^2}\frac{1}{h_\tau^{2(i+j-2)}}Var\left\{k(\frac{X-x^*}{h_\tau})\left(X-x^*\right)^{i+j-2}1\left\{X \geq x^*\right\}\right\} \\
\leq & \frac{1}{nh_\tau^2}\frac{1}{h_\tau^{2(i+j-2)}}\int_0^\infty k^2(\frac{x-x^*}{h_\tau})\left(x-x^*\right)^{2(i+j-2)}1\left\{x \geq x^*\right\}f(x)dx \\
= & \frac{1}{nh_\tau}\int_0^\infty k^2(z)z^{2(i+j-2)}f(x^* + zh_\tau)dz,
\end{aligned} \tag{A.28}$$

we have, as $n \to \infty$, $nh_\tau \to \infty$,

$$E\Gamma_{n\tau}(i,j) = \int_0^\infty k(z) z^{i+j-2} f(x^*) dz (1 + o(1)) \tag{A.29}$$

and

$$Var(\Gamma_{n\tau}(i,j)) \leq \frac{1}{nh_\tau} \max_{x \in [x^*, x^*+\delta]} f(x) \int_0^\infty k^2(z) z^{2(i+j-2)} dz = O(\frac{1}{nh_\tau}) \tag{A.30}$$

uniformly over $\tau \in [s_*, s], P_{m,\alpha} \in \mathcal{P}(s, \delta, K)$ and $s \in [s_*, s^*]$. The uniformity over $P_{m,\alpha} \in \mathcal{P}(s, \delta, K)$ is trivial because $\Gamma_{n\tau}(i,j)$ does not depend on $m(\cdot)$ or α. The uniformity over τ and s holds because $\max_{x \in [x^*, x^*+\delta]} f(x) \int_0^\infty k^2(z) z^{2(i+j-2)} dz$ does not depend on τ or s.

Invoking the Markov inequality yields, for any $\epsilon > 0$,

$$P_{m,\alpha} \left(|\Gamma_{n\tau}(i,j) - E\Gamma_{n\tau}(i,j)| > \epsilon \right) = O(\frac{1}{nh_\tau}) \tag{A.31}$$

uniformly over $\tau \in [s_*, s], P_{m,\alpha} \in \mathcal{P}(s, \delta, K)$ and $s \in [s_*, s^*]$.

Let $\Gamma(i,j) = \int_0^\infty k(z) z^{i+j+2} f(x^*) dz$. By the dominating convergence theorem, we have

$$\lim_{h_\tau \to 0} E\Gamma_{n\tau}(i,j) = \Gamma(i,j). \tag{A.32}$$

Combining this with (A.31), we get

$$P_{m,\alpha} \left(|\Gamma_{n\tau}(i,j) - \Gamma(i,j)| > \epsilon \right) = O(\frac{1}{nh_\tau}) \tag{A.33}$$

uniformly over $\tau \in [s_*, s], P_{m,\alpha} \in \mathcal{P}(s, \delta, K)$ and $s \in [s_*, s^*]$.

Denote $\Gamma_{r_\tau} = (\Gamma(i,j))$, the $(r_\tau + 1) \times (r_\tau + 1)$ matrix with the (i,j)-th element being $\Gamma(i-1, j-1)$. Then

$$P_{m,\alpha} \left(|\mu_{\min} (\Gamma_{n\tau}) - \mu_{\min}(\Gamma_{r_\tau})| > \epsilon \right) = O(\frac{1}{nh_\tau}) \tag{A.34}$$

and thus

$$P_{m,\alpha} \left(\mu_{\min} (\Gamma_{n\tau}) < C_{r_\tau} \right) = O(n^{-\frac{2\tau}{2\tau+1}}) \tag{A.35}$$

for some positive constant $C_{r_\tau} \leq \mu_{\min}(\Gamma_{r_\tau}) - \epsilon$. Note that for $\tau \in [s_*, s^*]$, there is only a finite number of limiting matrices Γ_{r_τ} and constants C_{r_τ}. Let $C_2 = \min_{\tau \in [s_*, s^*]} C_{r_\tau}$. Then

$$P_{m,\alpha} \left(\inf_{\tau \in [s_*, s^*]} \mu_{\min} (\Gamma_{n\tau}) \leq C_2 \right) = o(1) \tag{A.36}$$

uniformly over $\tau \in [s_*, s], P_{m,\alpha} \in \mathcal{P}(s, \delta, K)$ and $s \in [s_*, s^*]$.

Part (b) The proof is similar to that of part (a). Details are omitted.

Part (c) Let $\mathcal{B}_\tau = D_{n\tau}^{-1} Z_\tau^{+\prime} W_\tau^+ R_\tau^+$, then the i-th element of $\mathcal{B}_\tau$ is

$$\mathcal{B}_\tau(i) = \frac{1}{\sqrt{nh_\tau}} \frac{1}{h_\tau^{i-1}} \sum_{k=1}^{n} k(\frac{x_k - x^*}{h_\tau}) (x_k - x^*)^{i-1} R_\tau^+(x_k). \tag{A.37}$$

But

$$R_\tau^+(x_k) = 1\{\ell \geq r_\tau + 1\} b_{r_\tau+1}^+ (x_k - x^*)^{r_\tau+1} + e^+(x_k), \tag{A.38}$$

where

$$\left| e^+(x_k) / (x_k - x^*)^{\min\{s, r_\tau+2\}} \right| < C \tag{A.39}$$

for a constant C that is independent of x_k and τ. Hence $|R_\tau^+(x_k)| \leq C |x_k - x^*|^{\min(r_\tau+1, s)}$. As a consequence

$$\begin{aligned} & \left(\sqrt{nh_\tau} h^{\min(r_\tau+1,s)}\right)^{-1} |\mathcal{B}_\tau(i)| \qquad (A.40) \\ \leq \quad & \frac{C}{nh_\tau} \frac{1}{h_\tau^{i-1+\min(r_\tau+1,s)}} \sum_{k=1}^{n} k(\frac{x_k - x^*}{h_\tau}) (x_k - x^*)^{i-1} (x_k - x^*)^{\min(r_\tau+1,s)} \end{aligned}$$

Using the same argument in the proof of part (a), we can show that the above upper bound converges to

$$\int_0^\infty k(z) z^{i-1+\min(r_\tau+1,s)} f(x^*) dz \tag{A.41}$$

uniformly over $\tau \in [s_*, s], P_{m,\alpha} \in \mathcal{P}(s, \delta, K)$ and $s \in [s_*, s^*]$. Note that

$$\begin{aligned} \sqrt{nh_\tau} h_\tau^{\min(r_\tau+1,s)} &= \psi_1^{0.5+\min(r_\tau+1,s)} n^{1-\frac{1}{2\tau+1}} n^{\frac{\min(r_\tau+1,s)}{2\tau+1}} \\ &= C n^{\frac{1}{2}(1-\frac{1}{2\tau+1})} n^{\frac{-\min(r_\tau+1,s)}{2\tau+1}} \\ &= C n^{\frac{\tau-\min(r_\tau+1,s)}{2\tau+1}} = O(1) \text{ uniformly.} \end{aligned} \tag{A.42}$$

Therefore $|\mathcal{B}_\tau(i)|$ is bounded above uniformly over $\tau \in [s_*, s], P_{m,\alpha} \in \mathcal{P}(s, \delta, K)$ and $s \in [s_*, s^*]$. Combining this with the divergence of $\zeta(n)$ yields the desired result.

Part (d) The proof is similar to that of part (c). Details are omitted. ■

Let A_n be the union of events

$$\begin{aligned} & \left\{ \inf_{\tau \in [s_*, s^*]} \mu_{\min} \left(D_{n\tau}^{-1} Z_\tau^{+\prime} W_\tau^+ Z_\tau^+ D_{n\tau}^{-1} \right) \leq C_2 \right\} \cup \\ & \left\{ \inf_{\tau \in [s_*, s^*]} \mu_{\min} \left(D_{n\tau}^{-1} Z_\tau^{-\prime} W_\tau^- Z_\tau^- D_{n\tau}^{-1} \right) \leq C_2 \right\} \cup \\ & \left\{ \sup_{\tau \in [s_*, s^*]} \left\| D_{n\tau}^{-1} Z_\tau^{+\prime} W_\tau^+ R_\tau^+ \right\| > C_3 \zeta(n) \right\} \cup \\ & \left\{ \sup_{\tau \in [s_*, s^*]} \left\| D_{n\tau}^{-1} Z_\tau^{-\prime} W_\tau^- R_\tau^- \right\| > C_3 \zeta(n) \right\} \end{aligned}$$

whose probabilities are specified in Lemma A.1. Let A_n^c denote its complement.

Lemma A.2 *Let the assumptions of Theorem 3 hold.*

(a) For a constant C such that $C > 4C_3C_2^{-1}$, we have

$$\sup_{s\in[s_*,s^*]}\sup_{P_{m,\alpha}\in\mathcal{P}(s,\delta,K)}\sup_{\tau\in[s_*,s]} P_{m,\alpha}\left(\sqrt{nh_\tau}|\hat{\alpha}_\tau-\alpha|>C\zeta(n),A_n^c\right)=O\left(\zeta^{-2}(n)\right).$$

(b) Let $\tau_0 := s_0+(\rho s_0)(\log\log n)/\log(n)$ with $\sqrt{2\rho}>2+1/s_0$. If $m(x)\in\mathcal{M}_0(s_0,\delta,K)$ for some $s_0<\infty$, then for any constant $C>0$,

$$P_{m,\alpha}\left(\sqrt{nh_{s_0}}|\hat{\alpha}_{\tau_0}-\alpha|\le C\zeta(n),A_n^c\right)=o(1).$$

Proof of Lemma A.2. Part (a) Note that

$$\begin{aligned}
& P_{m,\alpha}\left(\sqrt{nh_\tau}|\hat{\alpha}_\tau-\alpha|>C\zeta(n),A_n^c\right)\\
= \;& P_{m,\alpha}\left(\sqrt{nh_\tau}|\left[\hat{c}_\tau^+-\alpha-m(x^*)\right]-\left[\hat{c}_\tau^--m(x^*)\right]|>C\zeta(n),A_n^c\right)\\
\le \;& P_{m,\alpha}\left(\sqrt{nh_\tau}\left|\hat{c}_\tau^+-\alpha-m(x^*)\right|>\frac{C}{2}\zeta(n),A_n^c\right)\\
& +P_{m,\alpha}\left(\sqrt{nh_\tau}|\hat{c}_\tau^--m(x^*)|>\frac{C}{2}\zeta(n),A_n^c\right) \qquad (A.43)
\end{aligned}$$

We now consider each of the two terms. It follows from the proof of Theorem 1 that

$$\hat{c}_\tau^+-\alpha-m(x^*)=e_1'\left(Z_\tau^{+\prime}W_\tau^+Z_\tau^+\right)^{-1}Z_\tau^{+\prime}W_\tau^+\varepsilon^+ + e_1'\left(Z_\tau^{+\prime}W_\tau^+Z_\tau^+\right)^{-1}Z_\tau^{+\prime}W_\tau^+R_\tau^+. \qquad (A.44)$$

So

$$\begin{aligned}
& P_{m,\alpha}\left(\sqrt{nh_\tau}\left|\hat{c}_\tau^+-\alpha-m(x^*)\right|>\frac{C}{2}\zeta(n),A_n^c\right)\\
\le \;& P_{m,\alpha}\left(\left|e_1'\left(D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+Z_\tau^+D_{n\tau}^{-1}\right)^{-1}D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+R_\tau^+\right|>\frac{C}{4}\zeta(n),A_n^c\right)\\
& +P_{m,\alpha}\left(\left|e_1'\left(D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+Z_\tau^+D_{n\tau}^{-1}\right)^{-1}D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+\varepsilon^+\right|>\frac{C}{4}\zeta(n),A_n^c\right)\\
\le \;& P_{m,\alpha}\left(\left[\mu_{\min}\left(D_{n\tau}^{-1}Z_\tau^{+\prime}W^+Z_\tau^+D_{n\tau}^{-1}\right)\right]^{-1}\left\|D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+R_\tau^+\right\|>\frac{C}{4}\zeta(n),A_n^c\right)\\
& +P_{m,\alpha}\left(\left[\mu_{\min}\left(D_{n\tau}^{-1}Z_\tau^{+\prime}W^+Z_\tau^+D_{n\tau}^{-1}\right)\right]^{-1}\left\|D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+\varepsilon^+\right\|>\frac{C}{4}\zeta(n),A_n^c\right)\\
\le \;& P_{m,\alpha}\left(\left\|D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+R_\tau^+\right\|>C_3\zeta(n),A^c\right)+P_{m,\alpha}\left(\left\|D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+\varepsilon^+\right\|>C_3\zeta(n)\right)\\
= \;& P_{m,\alpha}\left(\left\|D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+\varepsilon^+\right\|>C_3\zeta(n)\right) \qquad (A.45)
\end{aligned}$$

The last equality holds because on A^c, $\left\|D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+R_\tau^+\right\|\le C_3\zeta(n)$ for all τ. Let $\Sigma^+=$

$diag(\sigma^2(x_i)|_{x_i>x^*})$, then

$$\begin{aligned}
& P_{m,\alpha}\left(\left\|D_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+\varepsilon^+\right\| > C_3\zeta(n)\right) \\
\leq\ & C_3^{-2}\zeta^{-2}(n)trace(ED_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+\varepsilon^+\varepsilon^{+\prime}W_\tau^{+\prime}Z_\tau^{+\prime}D_{n\tau}^{-1}) \\
=\ & C_3^{-2}\zeta^{-2}(n)trace(ED_{n\tau}^{-1}Z_\tau^{+\prime}W_\tau^+\Sigma^+W_\tau^{+\prime}Z_\tau^{+\prime}D_{n\tau}^{-1}) \\
=\ & C_3^{-2}\zeta^{-2}(n)\sum_{i=1}^{r_\tau+1}\frac{1}{nh_\tau}\frac{1}{h_\tau^{2i-2}}\sum_{k=1}^{n}Ek^2(\frac{x_k-x^*}{h_\tau})\sigma^2(x_k)\left(x_k-x^*\right)^{2i-2}1\left\{x_k\geq x^*\right\} \\
=\ & C_3^{-2}\zeta^{-2}(n)\sum_{i=1}^{r_\tau+1}\int_0^\infty k^2(z)z^{2i-2}f(x^*+h_\tau z)\sigma^2(x^*+h_\tau z)dz \\
\leq\ & C_3^{-2}\zeta^{-2}(n)\max_{x\in[x^*,x^*+\delta]}\left\{f(x)\sigma^2(x)\right\}\sum_{i=1}^{r_\tau+1}\int_0^\infty k^2(z)z^{2i-2}dz \\
=\ & O\left(\zeta^{-2}(n)\right) \qquad \text{(A.46)}
\end{aligned}$$

uniformly over $\tau\in[s_*,s]$, $P_{m,\alpha}\in\mathcal{P}(s,\delta,K)$ and $s\in[s_*,s^*]$. Therefore

$$P_{m,\alpha}\left(\sqrt{nh_\tau}\left|\hat{c}_\tau^+-\alpha-m(x^*)\right|>\frac{C}{2}\zeta(n),A_n^c\right)=O\left(\zeta^{-2}(n)\right) \qquad \text{(A.47)}$$

uniformly. Similarly, we can prove that

$$P_{m,\alpha}\left(\sqrt{nh_\tau}|\hat{c}_\tau^--m(x^*)|>\frac{C}{2}\zeta(n),A_n^c\right)=O\left(\zeta^{-2}(n)\right) \qquad \text{(A.48)}$$

uniformly. Combining (A.43), (A.47) and (A.48) leads to the require result.

Part (b) Note that

$$\begin{aligned}
& P_{m,\alpha}\left(\sqrt{nh_{s_0}}|\hat{\alpha}_{\tau_0}-\alpha|\leq C\zeta(n),A_n^c\right) \\
=\ & P_{m,\alpha}\left(\sqrt{nh_{\tau_0}}|\hat{\alpha}_{\tau_0}-\alpha|\leq n^{-\frac{s_0}{2s_0+1}}n^{\frac{\tau_0}{2\tau_0+1}}C\zeta(n),A_n^c\right) \\
=\ & P_{m,\alpha}\left(\sqrt{nh_{\tau_0}}|\hat{\alpha}_{\tau_0}-\alpha|\leq C\left(\log n\right)^{\rho s_0/(2s_0+1)(2\tau_0+1)}\zeta(n),A_n^c\right). \qquad \text{(A.49)}
\end{aligned}$$

The last equality holds because

$$\begin{aligned}
n^{-\frac{s_0}{2s_0+1}+\frac{\tau_0}{2\tau_0+1}} &= \exp\left\{\frac{\rho s_0\log^{-1}n\log\log n}{(2s_0+1)(2\tau_0+1)}\log n\right\} \\
&= (\log n)^{\rho s_0/[(2s_0+1)(2\tau_0+1)]}. \qquad \text{(A.50)}
\end{aligned}$$

Let

$$\begin{aligned}
G_{\tau_0}^+(\varepsilon^+) &= e_1'\left(D_{n\tau_0}^{-1}Z_{\tau_0}^{+\prime}W_{\tau_0}^+Z_{\tau_0}^+D_{n\tau_0}^{-1}\right)^{-1}D_{n\tau_0}^{-1}Z_{\tau_0}^{+\prime}W_{\tau_0}^+\varepsilon^+ \qquad \text{(A.51)} \\
G_{\tau_0}^+(\tilde{e}^+) &= e_1'\left(D_{n\tau_0}^{-1}Z_{\tau_0}^{+\prime}W_{\tau_0}^+Z_{\tau_0}^+D_{n\tau_0}^{-1}\right)^{-1}D_{n\tau_0}^{-1}Z_{\tau_0}^{+\prime}W_{\tau_0}^+\tilde{e}^+ \qquad \text{(A.52)}
\end{aligned}$$

and define $G^-_{\tau_0}(\varepsilon^-)$ and $G^-_{\tau_0}(\tilde{e}^-)$ analogously. Then

$$\sqrt{nh_{\tau_0}}|\hat{\alpha}_{\tau_0}-\alpha| = G^+_{\tau_0}(\varepsilon^+) - G^-_{\tau_0}(\varepsilon^-) + G^+_{\tau_0}(\tilde{e}^+) - G^-_{\tau_0}(\tilde{e}^-) \tag{A.53}$$

and

$$\begin{aligned}
& P_{m,\alpha}\left(\sqrt{nh_{\tau_0}}|\hat{\alpha}_{\tau_0}-\alpha| \le C\left(\log n\right)^{\rho s_0/(2s_0+1)(2\tau_0+1)}\zeta(n), A_n^c\right) \\
\le\ & P_{m,\alpha}\left\{\left|G^+_{\tau_0}(\tilde{e}^+) - G^-_{\tau_0}(\tilde{e}^-)\right| \le \left|G^+_{\tau_0}(\varepsilon^+) - G^-_{\tau_0}(\varepsilon^-)\right| + \right. \\
& \left. C\left(\log n\right)^{\rho s_0/(2s_0+1)(2\tau_0+1)}\zeta(n), A_n^c\right\} \\
\le\ & P_{m,\alpha}\left(\left|G^+_{\tau_0}(\tilde{e}^+) - G^-_{\tau_0}(\tilde{e}^-)\right| \le C\zeta(n) + C\left(\log n\right)^{\rho s_0/[(2s_0+1)(2\tau_0+1)]}\zeta(n), A_n^c\right) \\
& + P_{m,\alpha}\left(\left|G^+_{\tau_0}(\varepsilon^+) - G^-_{\tau_0}(\varepsilon^-)\right| \ge C\zeta(n), A_n^c\right) \\
\le\ & P_{m,\alpha}\left(\left|G^+_{\tau_0}(\tilde{e}^+) - G^-_{\tau_0}(\tilde{e}^-)\right| \le C\zeta(n) + C\left(\log n\right)^{\rho s_0/[(2s_0+1)(2\tau_0+1)]}\zeta(n), A_n^c\right) \\
& + P_{m,\alpha}\left(\left\|G^+_{\tau_0}(\varepsilon^+)\right\| \ge \frac{C}{2}\zeta(n), A_n^c\right) + P_{m,\alpha}\left(\left\|G^-_{\tau_0}(\varepsilon^-)\right\| \ge \frac{C}{2}\zeta(n), A_n^c\right).
\end{aligned} \tag{A.54}$$

But

$$\begin{aligned}
& P_{m,\alpha}\left(\left\|G^+_{\tau_0}(\varepsilon^+)\right\| \ge \frac{C}{2}\zeta(n), A_n^c\right) \\
=\ & P_{m,\alpha}\left(\left|e_1'\left(D^{-1}_{n\tau_0}Z^{+\prime}_{\tau_0}W^+_{\tau_0}Z^+_{\tau_0}D^{-1}_{n\tau_0}\right)^{-1}D^{-1}_{n\tau_0}Z^{+\prime}_{\tau_0}W^+_{\tau_0}\varepsilon^+\right| \ge \frac{C}{2}\zeta(n), A_n^c\right) \\
\le\ & P_{m,\alpha}\left(\left[\mu_{\min}(D^{-1}_{n\tau_0}Z^{+\prime}_{\tau_0}W^+_{\tau_0}Z^+_{\tau_0}D^{-1}_{n\tau_0}\right]^{-1}\left\|D^{-1}_{n\tau_0}Z^{+\prime}_{\tau_0}W^+_{\tau_0}\varepsilon^+\right\| \ge \frac{C}{2}\zeta(n), A_n^c\right) \\
\le\ & P_{m,\alpha}\left(\left[\mu_{\min}(D^{-1}_{n\tau_0}Z^{+\prime}_{\tau_0}W^+_{\tau_0}Z^+_{\tau_0}D^{-1}_{n\tau_0}\right]^{-1}\left\|D^{-1}_{n\tau_0}Z^{+\prime}_{\tau_0}W^+_{\tau_0}\varepsilon^+\right\| \ge \frac{C}{2}\zeta(n), A_n^c\right) \\
=\ & P_{m,\alpha}\left(\left\|D^{-1}_{n\tau_0}Z^{+\prime}_{\tau_0}W^+_{\tau_0}\varepsilon^+\right\| > C\zeta(n)\right) \\
=\ & O(\zeta^{-2}(n))
\end{aligned} \tag{A.55}$$

where the last line follows from (A.46). Similarly

$$P_{m,\alpha}\left(\left\|G^-_{\tau_0}(\varepsilon^-)\right\| \ge \frac{C}{2}\zeta(n), A_n^c\right) = O(\zeta^{-2}(n)). \tag{A.56}$$

As a consequence

$$\begin{aligned}
& P_{m,\alpha}\left(\sqrt{nh_{s_0}}|\hat{\alpha}_{\tau_0}-\alpha| \le C\zeta(n), A_n^c\right) \\
\le\ & P_{m,\alpha}\left(\left|G^+_{\tau_0}(\tilde{e}^+) - G^-_{\tau_0}(\tilde{e}^-)\right| \le C\zeta(n) + C\left(\log n\right)^{\rho s_0/[(2s_0+1)(2\tau_0+1)]}\zeta(n), A_n^c\right) + o(1).
\end{aligned} \tag{A.57}$$

Using the definition of $\mathcal{M}_0$, we have

$$\begin{aligned} & \left|G^+_{\tau_0}(\tilde{e}^+) - G^-_{\tau_0}(\tilde{e}^-)\right| \\ = & \left|e_1'\Gamma_{s_0}\left(D^{-1}_{ns_0}Z^{+\prime}_{s_0}W^+_{s_0}\tilde{e}^+ - (-1)^{\ell_0+1}D^{-1}_{ns_0}Z^{-\prime}_{s_0}W^-_{s_0}\tilde{e}^-\right)\right|(1+o(1)) \\ \geq & C\sqrt{nh_{\tau_0}}h^{s_0}_{\tau_0}(1+o(1)) = Cn^{\frac{\tau_0-s_0}{2\tau_0+1}}(1+o(1)) \qquad \text{(A.58)} \\ = & (\log n)^{(\rho s_0)/(2s_0+1)}(C+o(1)). \end{aligned}$$

However, since

$$\begin{aligned} & (\log n)^{(\rho s_0)/[(2s_0+1)(2\tau_0+1)]}\zeta(n) = C\,(\log n)^{(\rho s_0)/(2s_0+1)^2}\zeta(n)(1+o(1)). \\ = & o((\log n)^{(\rho s_0)/(2s_0+1)}) \qquad \text{(A.59)} \end{aligned}$$

provided that $\sqrt{2\rho} > 2 + 1/s_0$, we have

$$P\left(\left|G^+_{\tau_0}(\tilde{e}^+) - G^-_{\tau_0}(\tilde{e}^-)\right| \leq C\zeta(n) + C\,(\log n)^{8s_0/[(2s_0+1)(2\tau_0+1)]}\zeta(n)\right) = o(1), \qquad \text{(A.60)}$$

when n is large enough. Combining this with (A.57) leads to the stated result. ■

Proof of Theorem 3. Using Lemma A.1, we write

$$\begin{aligned} & P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)|\hat{\alpha}_{\hat{s}} - \alpha| \geq C_1\right) \\ = & P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)|\hat{\alpha}_{\hat{s}} - \alpha| \geq C_1, A^c_n\right) + P_{m,\alpha}(A_n) \\ :\quad & = \Pi^+_n + \Pi^-_n + o(1) \qquad \text{(A.61)} \end{aligned}$$

where

$$\begin{aligned} \Pi^+_n &= P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)|\hat{\alpha}_{\hat{s}} - \alpha| \geq C_1,\ \hat{s} \geq s, A^c_n\right) \text{ and} \\ \Pi^-_n &= P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)|\hat{\alpha}_{\hat{s}} - \alpha| \geq C_1,\ \hat{s} < s, A^c_n\right). \qquad \text{(A.62)} \end{aligned}$$

We want to show that $\lim_{C_1\to\infty}\limsup_{n\to\infty}\sup_{s,P_{m,\alpha}}\Pi^+_n = 0$ and likewise for Π^-_n.

We consider Π^+_n first. By the triangle inequality and the definition of $\hat{s}$, we have

$$\begin{aligned} \Pi^+_n \leq & P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)|\hat{\alpha}_{\hat{s}} - \hat{\alpha}_s| \geq C_1/2,\ \hat{s} \geq s, A^c_n\right) \\ & + P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)|\hat{\alpha}_s - \alpha| \geq C_1/2, A^c_n\right) \\ \leq & P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)(nh_s)^{-1/2}\psi_2\lambda_s\zeta(n) \geq C_1/2, A^c_n\right) \\ & + P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)|\hat{\alpha}_s - \alpha| \geq C_1/2, A^c_n\right) \\ \leq & P_{m,\alpha}(\psi_1^{-1/2}\psi_2\lambda_{s^*} \geq C_1/2, A^c_n) + P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)|\hat{\alpha}_s - \alpha| \geq C_1/2, A^c_n\right) \\ :\quad & = \Pi^+_{n,1} + \Pi^+_{n,2} \qquad \text{(A.63)} \end{aligned}$$

where we have used that λ_s is non-decreasing in s. Obviously, $\lim_{C_1\to\infty}\limsup_{n\to\infty}\sup_{s,P_{m,\alpha}}\Pi^+_{n,1}=0$. It follows from Lemma A.2(a) that

$$\lim_{C_1\to\infty}\limsup_{n\to\infty}\sup_{(s,P_{m,\alpha})}\Pi^+_{n,2}=0. \tag{A.64}$$

In consequence, $\lim_{C_1\to\infty}\limsup_{n\to\infty}\sup_{(s,P_{m,\alpha})}\Pi^+_n=0$.

Next, we consider Π^-_n. We have

$$\begin{aligned}
\Pi^-_n &= \sum_{\tau\in\mathcal{S}_g:\tau+g<s} P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)|\hat\alpha_\tau-\alpha|\geq C_1,\ \hat s=\tau, A^c_n\right)\\
&\quad + P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)|\hat\alpha_{\tau_s}-\alpha|\geq C_1,\ \hat s=\tau_s, A^c_n\right)\\
&\leq \sum_{\tau\in\mathcal{S}_g:\tau+g<s} P_{m,\alpha}(\hat s=\tau, A^c_n) + P_{m,\alpha}\left(n^{\frac{s}{2s+1}}\zeta^{-1}(n)|\hat\alpha_{\tau_s}-\alpha|\geq C_1, A^c_n\right)\\
&:= \Pi^-_{n,1}+\Pi^-_{n,2},
\end{aligned} \tag{A.65}$$

where $\tau_s\in\mathcal{S}_g$ and $s-g\leq\tau_s<s$.

Now, we bound $P_{m,\alpha}(\hat s=\tau, A^c_n)$. By the definition of $\hat s$, if $\hat s=\tau$, there exists $\tilde\tau\leq\tau$, $\tilde\tau\in\mathcal{S}_g$ such that $|\hat\alpha(\tau+g)-\hat\alpha(\tilde\tau)|>\psi_2\left(nh_{\tilde\tau}\right)^{-1/2}\lambda_{\tilde\tau}\zeta(n)$. As a consequence, for all $\tau\in\mathcal{S}_g$ with $\tau+g<s$,

$$\begin{aligned}
&P_{m,\alpha}(\hat s=\tau, A^c_n)\\
&\leq \sum_{\tilde\tau\in\mathcal{S}_g:\tilde\tau\leq\tau} P_{m,\alpha}\left(|\hat\alpha_{\tau+g}-\hat\alpha_{\tilde\tau}|>\psi_2\left(nh_{\tilde\tau}\right)^{-1/2}\lambda_{\tilde\tau}\zeta(n), A^c_n\right)\\
&\leq \sum_{\tilde\tau\in\mathcal{S}_g:\tilde\tau\leq\tau} P_{m,\alpha}\left(\left(nh_{\tau+g}\right)^{1/2}|\hat\alpha_{\tau+g}-\alpha|>\frac{1}{2}\psi_2\lambda_{s_*}\zeta(n), A^c_n\right)\\
&\quad + \sum_{\tilde\tau\in\mathcal{S}_g:\tilde\tau\leq\tau} P_{m,\alpha}\left(\left(nh_{\tilde\tau}\right)^{1/2}|\hat\alpha_{\tilde\tau}-\alpha|>\frac{1}{2}\psi_2\lambda_{s_*}\zeta(n), A^c_n\right)\\
&\leq 2(s^*-s_*)(\log n)\sup_{\tilde\tau<s} P_{m,\alpha}\left(\left(nh_{\tilde\tau}\right)^{1/2}|\hat\alpha_{\tilde\tau}-\alpha|>\frac{1}{2}\psi_2\lambda_{s_*}\zeta(n), A^c_n\right),
\end{aligned} \tag{A.66}$$

where the third inequality holds because there are at most $(s^*-s_*)(\log n)$ elements $\tilde\tau\in\mathcal{S}_g$ for which $\tilde\tau\leq\tau$. Note that the third inequality only applies for τ such that $\tau+g<s$. It is for this reason that we decompose Π^-_n into $\Pi^-_{n,1}+\Pi^-_{n,2}$ in (A.65).

Equations (A.65), (A.66) and Lemma A.2(a) give: for some $C<\infty$,

$$\begin{aligned}
\sup_{s\in[s_*,s^*]}\sup_{P_{m,\alpha}\in\mathcal{P}(s,\delta,K)}\Pi^-_{n,1} &\leq 2(s^*-s_*)^2(\log n)^2C\zeta^{-2}(n)\\
&= O((\log\log n)^{-1}) = o(1) \text{ as } n\to\infty.
\end{aligned} \tag{A.67}$$

Next, we have

$$n^{s/(2s+1)}n^{-\tau_s/(2\tau_s+1)} \leq n^{s/(2s+1)}n^{-(s-g)/(2s-2g+1)} = n^{\kappa_{s,g}g} \leq n^g = n^{\log^{-1} n} = e, \quad \text{(A.68)}$$

where $\kappa_{s,g} = (2s+1)^{-1}(2s-2g+1)^{-1} \leq 1$. This, $\tau_s < s$, and (A.65) give: for some $C < \infty$, $n^{s/(2s+1)} \leq \psi_1^{-1/2}(nh_{\tau_s})^{1/2}e$ and

$$\begin{aligned}\Pi_{n,2}^- &\leq P_{m,\alpha}\left((nh_{\tau_s})^{1/2}|\hat{\alpha}_{\tau_s} - \alpha| \geq C_1\psi_1^{1/2}e^{-1}\zeta(n)\right) \\ &\leq C\zeta^{-2}(n) = o(1) \text{ as } n \to \infty. \end{aligned} \quad \text{(A.69)}$$

This completes the proof of the theorem. ■

Proof of Theorem 5. Set $\overline{s} := \min(s_0, s^*)$. We first bound $P(\hat{s} < \overline{s} - g)$. We have

$$P(\hat{s} < \overline{s} - g) = \sum_{\tau \in \mathcal{S}_g : \tau + g < \overline{s}} P(\hat{s} = \tau, A_n^c) + o(1) = o(1), \quad \text{(A.70)}$$

where the second $o(1)$ term follows from the same proof for $\Pi_{n,1}^- = o(1)$; see equation (A.67). Here we do not need the uniformity result as we focus on a particular function in $\mathcal{M}_0(s_0, \delta, K)$.

If $s_0 \geq s^*$, then (A.70) clearly implies the result. Therefore, from now on we can assume $s_0 < s^*$. We now prove that $P(\hat{s} > \tau_0) = o(1)$ where $\tau_0 := s_0 + (\rho s_0)(\log\log n)/\log(n)$ with $\sqrt{2\rho} > 2 + 1/s_0$ as defined in Lemma A.2(b). Assume without loss of generality that $\tau_0 \in \mathcal{S}_g$. By the definition of $\hat{s}$,

$$\begin{aligned} & P(\hat{s} > \tau_0) = P(\hat{s} > \tau_0, A_n^c) + o(1) \\ \leq\ & P\left(\sqrt{nh_{s_0}}|\hat{\alpha}_{\tau_0} - \hat{\alpha}_{s_0}| \leq \psi_2\lambda_{s_0}\zeta(n), A_n^c\right) + o(1) \\ \leq\ & P\left(\sqrt{nh_{s_0}}|\hat{\alpha}_{\tau_0} - \alpha| \leq \psi_2\lambda_{s_0}\zeta(n) + \sqrt{nh_{s_0}}|\hat{\alpha}_{s_0} - \alpha|, A_n^c\right) + o(1) \\ \leq\ & P\left(\sqrt{nh_{s_0}}|\hat{\alpha}_{\tau_0} - \alpha| \leq \psi_2\lambda_{s_0}\zeta(n)) + C\zeta(n), A_n^c\right) \\ & \quad + P\left(\sqrt{nh_{s_0}}|\hat{\alpha}_{s_0} - \alpha| \geq C\zeta(n), A_n^c\right) \ + o(1) \\ =\ & o(1), \end{aligned} \quad \text{(A.71)}$$

where the last line uses both parts of Lemma A.2. In the above proof, we implicitly assume that $s_0 \in \mathcal{S}_g$. If this is not the case, we can bound $P(\hat{s} > \tau_0)$ by

$$P\left(\sqrt{nh_{s_0^*}^*}|\hat{\alpha}_{\tau_0} - \hat{\alpha}_{s_0^*}| \leq \psi_2\lambda_{s_0}^*\zeta(n), A_n^c\right) + o(1) \quad \text{(A.72)}$$

where $s_0^* := \max\{s : s \in \mathcal{S}_g, s \leq s_0\}$. The rest of the proof goes through with obvious changes.

Combining (A.70) and (A.71), we get $\hat{s} = \min(s_0, s^*) + O_p(\log\log n/\log n)$ as desired, completing the proof of Theorem 5. ■

References

[1] Andrews, D. W. K. and Y. Sun (2004): "Adaptive Local Polynomial Whittle Estimation of Long-range Dependence," *Econometrica* 72(2), 569-614.

[2] Angrist, J. D. and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics,* 114(2), 533-575.

[3] Battistin, E. and E. Rettore (2002): "Testing for Programme Effects in a Regression Discontinuity Design with Imperfect Compliance," *Journal of the Royal Statistical Society A*, 165(1), 39-57.

[4] Black, S. E. (1999): "Do 'Better' Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics* 114(2), 577-599.

[5] Brown, L. D. and M. G. Low (1996): "Asymptotic Equivalence of Nonparametric Regression and White Noise," *The Annals of Statistics*, Vol. 24(6), 2384-2398.

[6] Birge, L. and P. Massart (1997): "From Model Selection to Adaptive Estimation." In Festschrift for Lucien Le Cam (D. Pollard, E. Torgersen, G. L. Yang, eds.) 55-87. Springer-Verlag, New York.

[7] Cai, T. and M. G. Low (2003): "Adaptation Under Probabilistic Error for Estimating Linear Functionals," Technical report, Department of Statistics, University of Pennsylvania.

[8] Chay, K. and M. Greenstone (2005): "Does Air Quality Matter? Evidence from the Housing Market," *Journal of Political Economy*, 113(2).

[9] Cheng, M-Y, J. Fan, and J. S. Marron (1997): "On Automatic Boundary Corrections," *The Annals of Statistics,* Vol. 25, No. 4, 1691-1708.

[10] DiNardo, J. and D. Lee (2004): "Economic Impacts of New Unionization on Private Sector Employers: 1984-2001," *Quarterly Journal of Economics*, 119(4), 1383-1441.

[11] Donoho D. L. and R. C. Liu (1991): "Geometrizing Rates of Convergence II," *The Annals of Statistics*, Vol. 19(2), 633-667.

[12] Donoho, D. L. and I. M. Johnstone (1995): "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of American Statistics Association,* 90, 1200-1224.

[13] Fan, J. and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications,* Chapman and Hall, London.

[14] Giraitis, L., P. M. Robinson and A. Samarov (2000): "Adaptive Semiparametric Estimation of the Memory Parameter," *Journal of Multivariate Analysis,* 72, 183–207.

[15] Guggenberger, P. and Y. Sun (2003): "Bias-Reduced Log-Periodogram and Whittle Estimation of the Long-Memory Parameter Without Variance Inflation," Discussion Paper 2004-14, Department of Economics, University of California, San Diego.

[16] Hahn, J., P. Todd and W. Van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1), 201-209.

[17] Hall, P. and W. R. Schucany (1989): "A Local Cross-validation Algorithm," *Statistics and Probability Letters*, 8, 109-117.

[18] Hurvich, C. M., E. Moulines and P. Soulier (2002): "The FEXP Estimator for Potentially Non-stationary Linear Time Series," *Stochastic Processes and Their Applications,* 97, 307-340

[19] Ibragimov, I. A. and R. Z. Khasminskii (1981): *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.

[20] Iouditsky, A., E. Moulines and P. Soulier (2001): "Adaptive Estimation of the Fractional Differencing Coefficient," *Bernoulli* 7, 699-731.

[21] Pollard, D. (1993): "Asymptotics for a Binary Choice Model," Preprint, Department of Statistics, Yale University. Available at http://www.stat.yale.edu/OLD/Preprints/1993/93oct-1.pdf.

[22] Porter, J. (2003): "Estimation in the Regression Discontinuity Model," manuscript, Department of Economics, University of Wisconsin, Madison.

[23] Le Cam, L. (1986): *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

[24] Lepski, O. V. (1990): "On a Problem of Adaptive Estimation in Gaussian White Noise," *Theory of Probability and Its Applications,* 35, 454-466.

[25] —— (1991): "Asymptotically Minimax Adaptive Estimation I: Upper Bounds. Optimally Adaptive Estimates," *Theory of Probability and Its Applications,* 36, 682-697.

[26] —— (1992): "Asymptotically Minimax Adaptive Estimation II: Schemes Without Optimal Adaptation: Adaptive Estimator," *Theory of Probability and Its Applications,* 37, 433-468.

[27] Lepski, O. V., E. Mammen and V. G. Spokoiny (1997): "Optimal Spatial Adaptation to Inhomogeneous Smoothness: An Approach Based on Kernel Estimates with Variable Bandwidth Selectors," *The Annals of Statistics*, Vol. 25(3), 929-947.

[28] Lepski, O. V. and V. G. Spokoiny (1997): "Optimal Pointwise Adaptive Methods in Nonparametric Estimation," *The Annals of Statistics,* Vol. 25(6), 2512-2546.

[29] Stone, C. J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators, *The Annals of Statistics,* 8, 1348-1360.

[30] Spokoiny, V. G. (2000): "Adaptive Drift Estimation for Nonparametric Diffusion Model," *The Annals of Statistics*, Vol. 28(3), 815-836.

[31] Van der Klaauw, W. (2002): "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-discontinuity Approach," *International Economic Review*, 43(4), 1249-1287.