Estimation in the Regression Discontinuity Model*

Jack Porter
Harvard University
Department of Economics
Littauer Center 121
Cambridge, MA 02138
jporter@harvard.edu

May 7, 2003

Abstract

The regression discontinuity model has recently become a commonly applied framework for empirical work in economics. Hahn, Todd, and Van der Klaauw (2001) provide a formal development of the identification of a treatment effect in this framework and also note the potential bias problems in its estimation. This bias difficulty is the result of a particular feature of the regression discontinuity treatment effect estimation problem that distinguishes it from typical semiparametric estimation problems where smoothness is lacking. Here, the discontinuity is not simply an obstacle to overcome in estimation; instead, the size of discontinuity is itself the object of estimation interest. In this paper, I derive the optimal rate of convergence for estimation of the regression discontinuity treatment effect. The optimal rate suggests that with appropriate choice of estimator the bias difficulties are no worse than would be found in the usual nonparametric conditional mean estimation problem (at an interior point of the covariate support). Two estimators are proposed that attain the optimal rate under varying conditions. One new estimator is based on Robinson's (1988) partially linear estimator. The other estimator uses local polynomial estimation and is optimal under a broader set of conditions.

Keywords: Regression Discontinuity Design, Optimal Convergence Rate, Asymptotic Bias

JEL Classification: C13, C14

*I am grateful to Gary Chamberlain, Michael Murray, Whitney Newey, numerous seminar participants, and especially Guido Imbens for comments and suggestions. I thank the National Science Foundation for research support under grant SES-0112095. This paper has also circulated under the title "Asymptotic Bias and Optimal Convergence Rates for Semiparametric Kernel Estimators in the Regression Discontinuity Model."

1 Introduction

The regression discontinuity [RD] design has recently become a commonly applied framework for identification of treatment effects in economics. The RD design can be useful when there is a "cut-off" point in treatment assignment or in the probability of treatment assignment. Under weak smoothness conditions, the assignment near the cut-off behaves almost as if random. Regression discontinuity models typically use only the information very close to the shift to identify the treatment effect of interest without reliance on functional form assumptions. For example, Van der Klaauw (1996) examined the effect of a university's scholarship offer on the probability of an applicant choosing to attend that university. The scholarship amount offered was based on an underlying index of various individual characteristics observable to the econometrician. Cut-off points in the index were used by the college to group applicants into categories, and scholarship offers were given depending on the applicant category. This admissions procedure led to discontinuous shifts in the scholarship amount with respect to the index at the cut-off points. Van der Klaauw used the data near the discontinuities to estimate the causal effect of the scholarship amount on the probability of acceptance. The empirical literature applying the RD design in other contexts where a discontinuity in the treatment assignment exists includes Angrist and Lavy (1999), Battistin and Rettore (2002), Black (1999), Chay and Greenstone (1998), Lee (2001), and Pence (2002). The formal theoretical work in the econonometrics literature on identification and estimation in the RD model is apparently limited to the work of Hahn, Todd, and Van der Klaauw (2001). The present paper proposes two new estimators, derives the optimal convergence rate for semiparametric estimation of the treatment effect identified in the RD model, and finds that the proposed estimators attain the bound under varying conditions.

The lack of smoothness inherent in the RD model is not simply an obstacle to overcome in estimation; instead, the size of discontinuity itself is the object of interest. Semiparametric estimation of the discontinuity size is akin to estimation of a conditional expectation at two boundary points and differencing the results. Given this connection to boundary estimation, it is not surprising to find that a nonparametric estimator, such as the Nadaraya-Watson estimator, for the RD treatment effect has poor asymptotic bias behavior. This poor bias behavior forces the bandwidth to shrink at a fast rate and hence leads to a slow convergence rate. Hahn, Todd, and Van der Klaauw (2001) suggest dealing with the bias by using a local linear estimator for the RD treatment effect and provide the associated distributional results.

In this paper, I establish that the optimal convergence rate for estimation of the treatment effect in the RD model is the same as Stone's (1980) optimal rate for estimation of a conditional expectation at an interior point of the covariate support. This result suggests that the apparent bias difficulties are completely surmountable. I propose two optimal kernel-based semiparametric estimators: a new estimator based on Robinson's (1988) partially linear estimator, and a local polynomial estimator that is the natural extension of the local linear estimator covered in Hahn, Todd, and Van der Klaauw (2001). While the usual intuition motivating Robinson's estimator in the partially linear model breaks down in the RD model, it still turns out that this estimator achieves the optimal rate of convergence for estimation of the RD treatment effect under a particular smoothness condition (namely, that the derivatives of the conditional expectation have identical right and left-hand limits at the discontinuity point). Moreover, the local polynomial estimator attains the optimal convergence rate even more generally. These estimators not only improve the asymptotic bias to the order of typical conditional expectation estimators at an interior point of the support, but in some cases actually exhibit further unexpected bias reductions. Both of these estimators are familiar from their use in other nonparametric settings, but their limiting distributions in the RD context are different than in previously studied models.

Section 2 discusses RD treatment effect identification results relevant to our semiparametric estimation problem and formally demonstrates that estimation of treatment effects in RD models is equivalent to estimation of the size of a discontinuity in a conditional expectation. In particular, given a dependent variable y and a scalar covariate x, the jump size is given by

$$\alpha = \lim_{x \downarrow \bar{x}} E(y|x) - \lim_{x \uparrow \bar{x}} E(y|x) \tag{1}$$

where $\bar{x}$ is the known discontinuity point of the conditional expectation $E(y|x)$. No finite parametrization of the conditional expectation is assumed, just smoothness away from the discontinuity, so estimating the jump size requires nonparametric techniques.

The connection to the conditional expectation boundary problem in nonparametric estimation is clear from equation (1).[1] The usual bias issues in the boundary problem carry over to the RD case. Also, equation (1) makes apparent that only data near the discontinuity will be useful in estimating α. Data far from the discontinuity could only contribute to estimation in the limit if more is known about the functional form of $E(y|x)$ thus violating the nonparametric assumption. From this intuition, we should expect a slower than parametric $\sqrt{n}$-rate for estimation of α. This intuition is confirmed in sections 3 and 4. In section 3, we find that, under certain conditions, the partially linear and the local polynomial estimators have bias behavior like the difference of a conditional expectation at two *interior* points of support, where the differencing results in still further bias reduction. Due to this bias behavior, the limiting distribution of the local polynomial estimator does not follow from known results on the limiting distribution of such estimators at a boundary point (Fan and Gijbels 1996, Ruppert and Wand 1994). The new asymptotic distribution theory is given given in section 3. The asymptotic behavior of the partially linear estimator in the RD model is wholly different than its behavior in the canonical partially linear model (Robinson 1988). The new limiting distribution theory for this estimator is also covered in section 3.

In section 4, the optimal convergence rates are derived under a strong and weak set of smoothness conditions. The optimal rate is attained by the partially linear and local polynomial estimators. The relation of these results on best rates to the literature on optimal nonparametric conditional mean estimation, especially, optimal estimation at boundaries is also described in section 4. Section 5 discusses extending the results to more general RD models with multivariate covariates or unknown cut-off point. Section 6 concludes.

2 Regression Discontinuity Models

The theoretical connections between the regression discontinuity model and the treatment effect literature have been formally established in Hahn, Todd, and Van der Klaauw (2001). From

[1] see Härdle and Linton (1994) for further discussion of the nonparametric boundary problem.

its inception in Thistlewaite and Campbell (1960), the regression discontinuity model has been termed "quasi-experimental", reflecting its intuitive connection to purely randomized experimental methods. Trochim (1984) distinguishes between two different incarnations of the discontinuity model, depending on whether the treatment assignment is related to the observed variable by a deterministic function (sharp design) or a stochastic function (fuzzy design).

If treatment assignment is given by the indicator variable $D \in \{0, 1\}$, then the sharp design can be characterized by $D = \begin{cases} 1 & \text{if } Z \geq \bar{Z} \\ 0 & \text{if } Z < \bar{Z} \end{cases}$ where Z is observed and the cut-off point $\bar{Z}$ is known. Let Y_1 and Y_0 be the potential outcomes corresponding to the two treatment assignments, and as usual, $Y = DY_1 + (1 - D)Y_0$ is the observed outcome. Then under the smoothness assumption that $E(Y_j|Z)$ is continuous at $\bar{Z}$ for $j = 0, 1$, the expected causal effect of the treatment on the outcome is identified at the discontinuity point. In particular

$$\begin{aligned}
\alpha &= E(Y_1 - Y_0|\bar{Z}) \\
&= E(Y_1|\bar{Z}) - E(Y_0|\bar{Z}) \\
&= \lim_{Z \downarrow \bar{Z}} E(Y_1|Z) - \lim_{Z \uparrow \bar{Z}} E(Y_0|Z) \\
&= \lim_{Z \downarrow \bar{Z}} E(Y|Z) - \lim_{Z \uparrow \bar{Z}} E(Y|Z).
\end{aligned}$$

Hence, the discontinuity in treatment assignment at $\bar{Z}$ provides an opportunity to observe the average difference in potential outcomes from points just on either side of the discontinuity. In the limit, this average difference represents the expected causal effect evaluated at the discontinuity, $Z = \bar{Z}$. Identification is achieved assuming only smoothness in expected potential outcomes at the discontinuity (and a positive density in a neighborhood of the discontinuity) without any parametric functional form restrictions.

In the fuzzy design, $E(D|Z)$ is discontinuous at $\bar{Z}$. Hahn, Todd, and Van der Klaauw (2001) discusses conditions under which the expected causal effect at $\bar{Z}$ is identified. For instance, the assumption that $Y_1 - Y_0$ is independent of D conditional on Z is sufficient for identification.[2]

[2]Hahn, Todd, and Van der Klaauw (2001) and Battistin and Rettore (2002) also discuss the case where $D(Z)$ is a deterministic function that varies by individual, see Imbens and Angrist (1994). If $(Y_{i,1} - Y_{i,0}, D_i(Z))$ is jointly independent of Z_i for all Z, $D_i(Z)$ is discontinuous at $\bar{Z}$ for some i, and treatment assignment is monotone in Z, then the local average treatment effect (LATE) is identified at the discontinuity. Importantly, this set of identify-

In this case, we obtain $E(Y_1 - Y_0|Z = \bar{Z}) = [\lim_{Z\downarrow\bar{Z}} E(Y|Z) - \lim_{Z\uparrow\bar{Z}} E(Y|Z)] / [\lim_{Z\downarrow\bar{Z}} E(D|Z) - \lim_{Z\uparrow\bar{Z}} E(D|Z)]$. The key point for our estimation purposes is that in all cases the causal effect is identified from an expression involving only the sizes of discontinuities in conditional expectations.

A special case of both sharp and fuzzy designs is the common treatment effects model, where $Y_{i,1} - Y_{i,0} = \alpha$ is constant across individuals.[3] In this case identification of the treatment effect is achieved for sharp and fuzzy designs by only assuming smoothness in expected potential outcomes at $Z = \bar{Z}$. This case is important because with common treatment effects smoothness in expected potential outcomes translates directly into smoothness of $m(\cdot)$ in the corresponding estimation model (3) below (e.g. equal right and left hand side derivatives) at the discontinuity.

From the above expressions it is clear that the regression discontinuity approach only identifies a causal effect at the discontinuity. Without some additional functional form assumptions (e.g. common treatment effects) or additional discontinuity points, the result does not generalize to identify the average effect or effects at other values of Z. Nevertheless, the regression discontinuity model provides a useful framework for certain applications, and the above theoretical developments show where the regression discontinuity model fits into the treatment evaluation literature.

3 Semiparametric Kernel Estimators

For the purposes of the current work, the key point to take from the previous section is the finding that for each set of identifying assumptions, estimation of the RD treatment effect is equivalent to the problem of estimating the jump size of a discontinuity in a conditional expectation. We focus on estimation in the simplest RD model where the observed index in which the discontinuity occurs is scalar and no additional covariates are included. Below three different kernel-based approaches to semiparametric estimation are discussed. Since estimation in the RD model is analogous to nonparametric estimation at a boundary, where bias problems have received some attention, special consideration is given to the asymptotic bias behavior of each estimator. In particular, two proposed

ing assumptions also lead to an expression for the treatment effect in terms of just discontinuities in conditional expectations.

[3] Note that a local common treatment effects assumption would also be sufficient for these results. That is, it is sufficient that the common treatment effect assumption only hold for observations in a neighborhood of $\bar{Z}$.

discontinuity jump size esimators are based on partially linear and local polynomial estimators which are often applied in different nonparametric models. The asymptotic bias of each of these two jump size estimators will be different than what would be expected from their limiting behavior in other nonparametric settings. These new asymptotic results are important for inference in RD applications. In the next section, an additional reason for emphasizing the partially linear and local polynomial estimators will be presented when obtaining optimal rates of convergence.

3.1 Assumptions

Before discussing the semiparametric kernel estimators in the following subsections, we introduce the main assumptions necessary to derive the limiting distributions for the estimators.

The first assumption concerns the choice of kernel for estimation:

Assumption 1 *(a) $k(\cdot)$ is a symmetric, bounded, Lipschitz function, zero outside a bounded set; $\int k(u)du = 1$.*
(b) For positive integer s, $\int k(u)u^j du = 0$, $1 \leq j \leq s-1$.

This assumption is fairly standard, allowing for higher-order bias-reducing kernels (an s^{th} order kernel above). For $s \geq 3$, the function $k(\cdot)$ is necessarily somewhere negative to satisfy condition (b). By assuming that the kernel has a bounded support set, the remaining assumptions on the model only need to describe behavior local to the discontinuity, and the need for a trimming function is avoided.[4]

The second assumption concerns the RD model itself. Let f_0 denote the marginal density of x and $m(x)$ denote the conditional expectation of y given x minus the discontinuity. Formally, $m(x) = E(y|x) - \alpha\mathbf{1}\{x \geq \bar{x}\}$.

Assumption 2 *Suppose the data $(y_i, x_i)_{i=1,\ldots,n}$ is i.i.d. and α is defined by equation (1).*
(a) For some compact interval $\mathcal{N}$ of $\bar{x}$ with $\bar{x} \in int(\mathcal{N})$, f_0 is l_f times continuously differentiable and bounded away from zero; $m(x)$ is l_m times continuously differentiable for $x \in \mathcal{N}\backslash\{\bar{x}\}$, and m

[4]Alternatively one could have an unbounded kernel support and trimming that becomes negligible as the sample size increases, as in Robinson (1988).

is continuous at $\bar{x}$ with finite right and left-hand derivatives to order l_m.

(b) Right and left-hand derivatives of m to order l_m are equal at $\bar{x}$.

Assumption 2(a) assures sufficient smoothness on either side of the discontuity, but allows for unequal right and left-hand derivatives of m (equivalently $E(y|x)$) at $\bar{x}$. Assumption 2(b) imposes more smoothness in m at $\bar{x}$. This part of the assumption will not affect the limiting distribution result for the Nadaraya-Watson estimator below, but will play an important role in the asymptotic bias of subsequent estimators. Part (b) would be implied by, for instance, smoothness in the derivatives of $E(Y_0|X)$ to order l_m and a common treatment effect assumption for observations in a neighborhood of $\bar{x}$. Bounding the density away from zero in a neighborhood of $\bar{x}$ assures that there is positive density near the discontinuity and thus sufficient data for estimating the jump size. It also allows us to avoid the technical problems of a near-zero denominator.

The third assumption concerns the behavior of the moments of the outcome variable. Define $\varepsilon = y - E(y|x) = y - m(x) - \alpha \mathbf{1}\{x \geq \bar{x}\}$.

Assumption 3 *(a) $\sigma^2(x) = E(\varepsilon^2|x)$ is continuous for $x \neq \bar{x}$, $x \in \mathcal{N}$, and the right and left-hand limits at $\bar{x}$ exist.*

(b) For some $\zeta > 0$, $E[|\varepsilon|^{2+\zeta}|x]$ is uniformly bounded on $\mathcal{N}$.

This assumption places no restrictions on the unconditional moments. Instead, only the behavior of conditional moments evaluated near the discontinuity is restricted. Part (b) of the assumption bounds the moments so that a central limit theorem can be applied to derive the limit distribution of the different estimators. Part (a) simply allows us to write the asymptotic variance expressions in a tractable form.

3.2 Nadaraya-Watson Estimation

The simplest kernel estimator of the discontinuity jump takes a kernel weighted average of observations on each side of the discontinuity and differences. Such an estimator is referred to as a Nadaraya-Watson estimator and is discussed in Hahn, Todd, and Van der Klaauw (2001). Here we briefly cover this estimator to formally establish the bias difficulty in estimation of the discontinuity

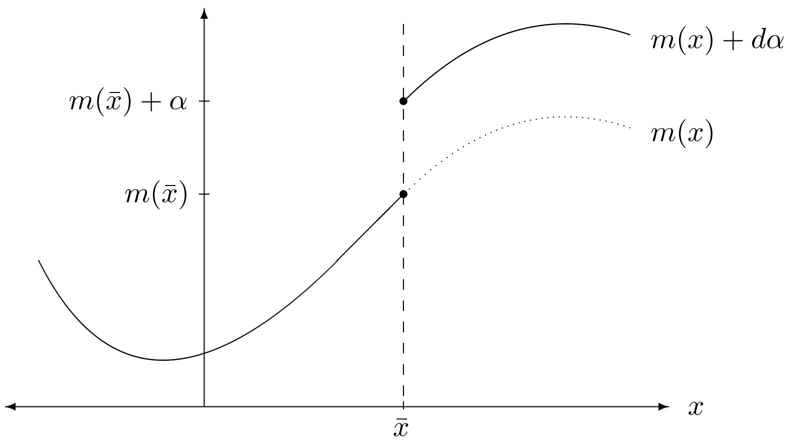

Figure 1

jump size. Our discussion then serves as a point of reference for the subsequent estimators with improved bias properties.

Usual Nadaraya-Watson estimation of α would proceed from a sample version of equation (1). To estimate the first term in the difference, one would average the values of y corresponding to the observations with the x's just to the right of the discontinuity. The second term in the difference would be estimated similarly, and then the estimate of α would be given by the difference of the estimated first and second terms. The idea of using local averages to estimate the desired conditional expectations is captured by kernel regression. A specific kernel version of the estimation method just described follows. Given data $(y_i, x_i)_{i=1,\ldots,n}$, let

$$\tilde{\alpha} = \frac{\frac{1}{n}\sum_i k_h(\bar{x} - x_i) y_i d_i}{\frac{1}{n}\sum_j k_h(\bar{x} - x_j) d_j} - \frac{\frac{1}{n}\sum_i k_h(\bar{x} - x_i) y_i (1 - d_i)}{\frac{1}{n}\sum_j k_h(\bar{x} - x_j)(1 - d_j)} \tag{2}$$

where $d_i = \mathbf{1}\{x_i \geq \bar{x}\}$, $\mathbf{1}\{A\}$ is an indicator for the event A, $k_h(\cdot) = \frac{1}{h} k\left(\frac{\cdot}{h}\right)$, $k(\cdot)$ is a kernel function, and h denotes a bandwidth that controls the size of the local neighborhood to be averaged over. The first term in $\tilde{\alpha}$ is a weighted average of the dependent variable values for data just to the right of the discontinuity, where the weights depend on the distance from the discontinuity $(\bar{x} - x_i)$, the bandwidth h, and the kernel k. Similarly, the second term estimates the left-hand side limit of the conditional expectation at the discontinuity.

Such an estimator is subject to the problems of many nonparametric methods at a boundary. Here, the bias at the boundary is $O(h)$, where h is the bandwidth, whereas typical kernel regression estimates at an interior point of the covariate support have a bias that is $O(h^2)$. For the example in Figure 1, due to the positive slope of m near $\bar{x}$, observations just to the right of the discontinuity are more likely to be greater than the "intercept" $m(\bar{x}) + \alpha$. So an average of such observations will generally provide an upward biased estimate of $m(\bar{x}) + \alpha$. Similarly, an average of observations just to the left of the discontinuity would provide a downward biased estimate of $m(\bar{x})$. A kernel regression estimate at an interior point behaves like the sum of these two averages, so that the biases from each side tend to cancel giving the lower order of bias. On the other hand, estimating the discontinuity jump involves differencing these averages which, if anything, compounds the bias problem. Higher-order bias-reducing kernels do not improve this bias problem. These ideas are formalized in the following theorem, which is a straightforward corollary of standard results on the limiting distribution of Nadaraya-Watson estimators at a boundary point.

THEOREM 1 *Suppose Assumptions 1(a), 2(a), and 3 hold with l_m any positive integer and l_f any nonnegative integer. If $h \longrightarrow 0$, $nh \longrightarrow \infty$ and $h\sqrt{nh} \longrightarrow C$, where $0 \leq C < \infty$, then*

$$\sqrt{nh}(\tilde{\alpha} - \alpha) \xrightarrow{d} N\left(C2K_1(0)(m'^{+}(\bar{x}) + m'^{-}(\bar{x})), 4\delta_0 \frac{\sigma^{2+}(\bar{x}) + \sigma^{2-}(\bar{x})}{f_0(\bar{x})}\right)$$

where

$$\delta_0 = \int_0^\infty k^2(u)du \qquad \text{and} \qquad K_1(w) = \int_w^\infty k(u)udu.$$

PROOF: see Appendix A □

Since the observations used in the first and second terms of the difference defining $\tilde{\alpha}$ are independent, the asymptotic variance consists of the sum of asymptotic variances for each term. Its form is familiar from the usual asymptotic variance formula for Nadaraya-Watson estimators at interior points.

The asymptotic bias is written in a form that underscores its dependence on the rate at which the bandwidth approaches zero. When $C = 0$, we are "under-smoothing" and the asymptotic bias is zero. When $h \sim n^{-1/3}$, $\tilde{\alpha}$ achieves its fastest rate of convergence ($n^{1/3}$) and the asymptotic bias term is non-negligible. From this formula, we see formally that the bias of $\tilde{\alpha}$ is of order $O(h)$. If Assumptions 1(b) and 2(b) are added to the conditions of Theorem 1, the asymptotic distribution does not change. That is, higher-order bias-reducing kernels do not affect the order of the asymptotic bias, so that s does not enter into the limiting distribution.[5] Also, equality or inequality of the right and left-hand derivatives of m at $\bar{x}$ does not affect the order of the asymptotic bias (unless the sum of the right and left-hand derivatives happened to exactly equal zero).

The bias behavior of the Nadaraya-Watson estimator follows from the form of the RD treatment effect in equation (1). It is exactly what one would expect from taking the difference of two boundary estimates. This expected order $O(h)$ asymptotic bias will serve as a basis for comparison with the estimators discussed next.

3.3 Partially Linear Estimation

The discontinuous conditional expectation given in (1) can be rewritten in a slightly different form,

$$y = m(x) + d\alpha + \varepsilon, \qquad \text{where } E(\varepsilon|x,d) = 0 \text{ and } d = \mathbf{1}\{x \geq \bar{x}\}. \tag{3}$$

Assuming $m(\cdot)$ is continuous at $\bar{x}$, α represents the jump size of the discontinuity at $\bar{x}$. Written in this form, the model appears to be of the usual partially linear type, as considered by Engle, Granger, Rice, and Weiss (1986), where the discontinuity jump corresponds to the linear parametric component. Estimation of partially linear models has been considered in Chen (1988), Heckman (1986), Rice (1986), Robinson (1988), and Speckman (1988). Stock (1989) applied these semiparametric techniques. Estimation in the partially linear model is one of the prime successes of semiparametrics. In particular, despite the presence of a nonparametric component, the parametric component can be estimated at a $\sqrt{n}$-rate. Below we focus on Robinson's (1988) estimation method. We will first point out that the model in (3) differs from the traditional partially linear model in important ways, making the use of Robinson's partially linear estimator questionable.

The partially linear estimator is motivated by subtracting the conditional expectation with

[5] The kernel could be modified to achieve higher-order type bias reduction. In particular, if the kernel is symmetric and satisfies $\int k(u)|u|^j du = 0$ for $1 \leq j < s$, then boundary bias reduction occurs to order s. Such kernels are called boundary kernels.

respect to x from both sides of (3),

$$y - E(y|x) = \alpha(d - E(d|x)) + \varepsilon \tag{4}$$

This partialling out idea is the natural nonparametric extension of the same idea in the fully linear model. The intuition for the partially linear estimator is to first use nonparametric regression to estimate the conditional expectations and form the residuals $\hat{\eta}_i = y_i - \hat{E}(y|x_i)$ and $\hat{\nu}_i = d_i - \hat{E}(d|x_i)$. Then apply least squares to the residuals to obtain an estimate of α, $\hat{\alpha} = [\sum_i \hat{\nu}_i \hat{\nu}_i']^{-1} \sum_i \hat{\nu}_i \hat{\eta}_i$. Robinson shows that the fact that the conditional expectations in the partially linear analog of (4) are estimated before applying least squares does not affect the limiting distribution of the estimator. That is, the limiting distribution of $\hat{\alpha}$ is the same as the least squares estimate from (4) with known $E(y|x)$ and $E(d|x)$. Under certain regularity conditions, Robinson provides a necessary and sufficient condition for $\sqrt{n}$-consistency with asymptotic normality.

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, \sigma^2 E[(d - E(d|x))^2]^{-1})$$ [6]

if and only if

$$E[(d - E(d|x))^2] \text{ is positive definite.}$$

The specification (3), and in particular the condition $d = \mathbf{1}\{x \geq \bar{x}\}$, present some problems with this approach that ultimately lead to failure of the necessary and sufficient positive definiteness condition in the RD model. First, $E(d|x)$ is discontinuous and so violates the smoothness assumption needed for consistent nonparametric estimation. As a result, the usual nonparametric regression estimator $\hat{E}(d|x)$ would be inconsistent at the discontinuity point $\bar{x}$, which would appear to be a stumbling block in estimation of α. Second, d is a deterministic function of x ($d = \mathbf{1}(x \geq \bar{x})$), and so $d - E(d|x) = 0$. Thus we lose the intuition from the partialling out equation (4), which can now be written more simply as $y - E(y|x) = \varepsilon$. Partialling out with respect to x completely partials out the linear (parametric) term too. If we proceeded with partialling out, the residuals ($\hat{\nu}$) would be $d - \hat{E}(d|x) = [d - E(d|x)] + [E(d|x) - \hat{E}(d|x)] = E(d|x) - \hat{E}(d|x)$. In the usual partially linear framework, $E(d|x) - \hat{E}(d|x)$ is small enough that the residual, $\hat{\nu}$, is mostly approximating the term

[6] For notational convenience and to match Robinson's result exactly, I've assumed homoskedasticity for the moment.

$d - E(d|x)$. In the discontinuity model, $d - E(d|x) = 0$ so the residual, $\hat{\nu} = E(d|x) - \hat{E}(d|x)$, is just the error in estimating a discontinuous conditional expectation by nonparametric regression. Stated differently, $d - \hat{E}(d|x)$ is simply a noisy estimate of $0 = d - E(d|x)$. Moreover, note that the necessary and sufficient condition (positive definiteness of $E(d - E(d|x))^2$) in Robinson's asymptotic normality result for the partially linear estimator fails in the discontinuity model (3).

The partially linear estimator is known for obtaining root-n consistent estimates of the parameter of the linear component. In the discontinuity model, α is the parameter of the linear part. From the earlier intuition that only data near the discontinuity is useful in estimation of α, it's clear that α is only estimable at a nonparametric rate. In section 4, this intuition is formalized by providing results on optimal rates of convergence for estimation of α.

We now turn to a different intuition for estimating α from (3) which will lead to an alternative motivation for using Robinson's approach in the RD model. Consider rewriting (3) as $(y - d\alpha) = m(x) + \varepsilon$, and treating $\tilde{y} = (y - d\alpha)$ as the dependent variable. If we knew α, then nonparametric estimation of m using $\tilde{y}$ would be a standard exercise. And the resulting estimate $\tilde{m}$ of m could then be used to provide a least squares estimate of α by subtracting that estimate from y and choosing the estimate of α to minimize the squared difference between $y - \tilde{m}(x)$ and $d\alpha$. Clearly this procedure is infeasible since we do not know α to start with. But we can follow basically the same approach for unknown α, essentially using a profiled estimator. More specifically, estimate α by finding the value that minimizes the average squared deviation between the dependent variable $(y - d\alpha)$ and the nonparametric estimate of $m(\cdot)$ (using $(y - d\alpha)$ as the dependent variable). Then our estimate of α comes from the following optimization problem.

$$\min_{\alpha} \sum_{i=1}^{n} [y_i - \alpha d_i - \sum_{j=1}^{n} w_j^i (y_j - \alpha d_j)]^2 \tag{5}$$

$$\text{where} \quad w_j^i = \frac{k_h(x_i - x_j)}{\sum_{l=1}^{n} k_h(x_i - x_l)}$$

A closer look at the individual terms in the objective function's sum reveals the intuition behind this estimator. First, we seek which terms in the sum are most important to the optimization, i.e. have more weight in the sum. Let $\hat{f}_+(x) = \frac{1}{n}\sum_{j=1}^{n} k_h(x - x_j)d_j$ and $\hat{f}_-(x) = \frac{1}{n}\sum_{j=1}^{n} k_h(x - x_j)(1 -$

d_j). Then the kernel density estimate at x is $\hat{f}_0(x) = \hat{f}_+(x) + \hat{f}_-(x)$, and $\hat{f}_+(x)$ is the part of the density estimate that comes from data to the right of the discontinuity and $\hat{f}_-(x)$ is the part that comes from the left. So if x is very close to the discontinuity point $\bar{x}$, then we would expect $\hat{f}_+(x)$ and $\hat{f}_-(x)$ to be close in magnitude. On the other hand, if $x >> \bar{x}$, then we would expect $\hat{f}_+(x)$ $>> \hat{f}_-(x)$. The ratio of $\hat{f}_+(x)$ and $\hat{f}_-(x)$ determines the import of individual terms in the objective function sum. From the objective function (5), consider the portion of the deviation affected by the choice of α: $(d_i - \sum_j w_j^i d_j)\alpha = \left(d_i \frac{1}{1+(\hat{f}_+(x)/\hat{f}_-(x))} - (1-d_i)\frac{1}{1+(\hat{f}_-(x)/\hat{f}_+(x))}\right)\alpha$. So if $d_i = 1$, i.e. $x_i > \bar{x}$, then the weight on observation i is small if $\hat{f}_+(x)/\hat{f}_-(x)$ is large, i.e. $x_i >> \bar{x}$. If $d_i = 1$ and x_i is close to $\bar{x}$, then $\hat{f}_+(x)/\hat{f}_-(x)$ is small and the weight on observation i is relatively larger. Similarly, if $d_i = 0$, the observations with x_i very close to $\bar{x}$ have the greatest weight in the sum. Just as α is defined by the limiting behavior of the conditional expectation on either side of the discontinuity, the estimator defined in (5) is mostly determined by the data closest to the discontinuity.

If the true value of α were known, then we could let $\tilde{y}_i = y_i - d_i\alpha$, and $\tilde{m}(x_i) = \sum_j w_j^i \tilde{y}_j$ would be a consistent Nadaraya-Watson kernel regression estimate of $m(x_i)$. Thus in each term of the sum in (5), $\sum_j w_j^i (y_j - d_j\alpha)$ is close to $m(x_i)$ for appropriately chosen α. Note here that the data from *both* sides of the discontinuity is used in regression estimation at each data point, in contrast to the Nadaraya-Watson estimator of α given in 3.2. In fact, from above we know that the data closest to the discontinuity will have the most weight and such points will make most use of data from the other side of the discontinuity. The whole term $[y_i - d_i\alpha - \sum_j w_j^i (y_j - d_j\alpha)]$ then gives the deviation of $y_i - d_i\alpha$ from an estimate of $m(x_i)$ where α is chosen to make this difference as small as possible.

Finally, of course, the objective function (5) has an analytic solution.

$$\begin{aligned}
\hat{\alpha} &= \left[\sum_i (d_i - \sum_j w_j^i d_j)^2\right]^{-1} \sum_i [(d_i - \sum_j w_j^i d_j)(y_i - \sum_j w_j^i y_j)] \qquad (6)\\
&= \left[\sum_i \left(d_i \frac{\hat{f}_-^2(x_i)}{\hat{f}_0^2(x_i)} + (1-d_i)\frac{\hat{f}_+^2(x_i)}{\hat{f}_0^2(x_i)}\right)\right]^{-1} \sum_i \left[\left(d_i \frac{\hat{f}_-(x_i)}{\hat{f}_0(x_i)} - (1-d_i)\frac{\hat{f}_+(x_i)}{\hat{f}_0(x_i)}\right)\left(y_i - \sum_j w_j^i y_j\right)\right]\\
&= \alpha + \left[\sum_i \left(d_i \frac{\hat{f}_-^2(x_i)}{\hat{f}_0^2(x_i)} + (1-d_i)\frac{\hat{f}_+^2(x_i)}{\hat{f}_0^2(x_i)}\right)\right]^{-1} \sum_i \left[\left(d_i \frac{\hat{f}_-(x_i)}{\hat{f}_0(x_i)} - (1-d_i)\frac{\hat{f}_+(x_i)}{\hat{f}_0(x_i)}\right)\left(\tilde{y}_i - \sum_j w_j^i \tilde{y}_j\right)\right]
\end{aligned}$$

$$\begin{aligned}
\hat{\alpha} - \alpha = &\left[\frac{1}{nh}\sum_i \left(d_i \frac{\hat{f}_-^2(x_i)}{\hat{f}_0^2(x_i)} + (1-d_i)\frac{\hat{f}_+^2(x_i)}{\hat{f}_0^2(x_i)}\right)\right]^{-1} \qquad (7)\\
&\cdot \frac{1}{nh}\sum_i \left[\left(d_i \frac{\hat{f}_-(x_i)}{\hat{f}_0(x_i)} - (1-d_i)\frac{\hat{f}_+(x_i)}{\hat{f}_0(x_i)}\right)[\varepsilon_i - (\tilde{m}(x_i) - m(x_i))]\right]
\end{aligned}$$

From (6), it's clear that $\hat{\alpha}$ is the partially linear estimator for the model (3) treating $d_i\alpha$ as the linear part and partialling out with respect to the nonparametric part, $m(x_i)$. The Robinson estimator allows one to partial out with respect to the additive nonparametric part and still estimate the parametric part at a $n^{1/2}$-rate. In the present setting, the discontinuity jump cannot be estimated at a parametric rate, yet the same partialling out method still allows us to estimate the jump at a nonparametric rate. This result is perhaps unexpected, since we are able to nonparametrically partial out on the right hand side of equation (3) despite the fact that $E(d|x)$ is identically equal to d and thus is discontinuous at $\bar{x}$.

The expression in (7) suggests that $\hat{\alpha}$ has reasonable asymptotic properties. The numerator of the expression contains a weighted average of $\varepsilon_i - (\tilde{m}(x_i) - m(x_i))$, which should approach zero if appropriately normalized. Next we formalize this intuition with an asymptotic normality result for $\hat{\alpha}$.

Theorem 2 *Suppose Assumptions 1 and 3 hold, and assume* $s \geq 2$ *and* $\frac{n^{\frac{\zeta}{2+\zeta}}h}{\ln n} \longrightarrow \infty$.
(a) If Assumption 2(a) holds with $l_f \geq s$, $l_m \geq 2$, *and* $h^2\sqrt{nh} \longrightarrow C_a$, *where* $0 \leq C_a < \infty$, *then*

$$\sqrt{nh}(\hat{\alpha} - \alpha) \xrightarrow{d} N\left(C_a b_a, v_k \frac{\sigma^{2+}(\bar{x}) + \sigma^{2-}(\bar{x})}{4f_0(\bar{x})}\right)$$

(b) Let $I_s = \mathbf{1}\{s \text{ odd}\}$*. If Assumption 2 holds with* $l_f \geq s+I_s$, $l_m \geq s+1+I_s$, *and* $h^{s+1+I_s}\sqrt{nh} \longrightarrow C_b$, *where* $0 \leq C_b < \infty$, *then*

$$\sqrt{nh}(\hat{\alpha}-\alpha) \xrightarrow{d} N\left(C_b b_b, v_k \frac{\sigma^{2+}(\bar{x}) + \sigma^{2-}(\bar{x})}{4f_0(\bar{x})}\right)$$

where

$$\begin{aligned}
v_k &= \left(\int_0^\infty K_0^2(w)dw\right)^{-2} \int_0^\infty (K_0(w) + L(w) - L(-w))^2 dw, \\
b_a &= \left(2f_0(\bar{x})\int_0^\infty K_0^2(w)dw\right)^{-1} \int_0^\infty \big[2K_1^2(v)\mu_1 - K_0(v)(K_2(0) - K_2(v))(\mu_0 + 2\mu_1) \\
&\qquad -2vK_0(v)K_1(v)(\mu_0 + 3\mu_1) + v^2 K_0^2(v)(\mu_0 + 2\mu_1)\big]dv, \\
b_b &= 2K_{s+I_s}(0)\left(f_0(\bar{x})\int_0^\infty K_0^2(w)dw\right)^{-1}\left[\frac{f_0'(\bar{x})}{f_0(\bar{x})} g_{s+I_s}(\bar{x})\int_0^\infty K_1(v)dv \right. \\
&\qquad \left. -g_{s+I_s}'(\bar{x})\int_0^\infty K_0(v)v dv\right], \\
K_j(w) &= \int_w^\infty k(u)u^j du \quad \text{and} \quad L(w) = \int_0^\infty K_0(u)k(u+w)du, \\
\mu_0 &= f_0(\bar{x})(m''^{+}(\bar{x}) - m''^{-}(\bar{x})) \quad \text{and} \quad \mu_1 = f_0'(\bar{x})(m'^{+}(\bar{x}) - m'^{-}(\bar{x})), \text{ and} \\
g_t(x) &= \sum_{j=1}^{t} \frac{m^{(j)}(x) f_0^{(t-j)}(x)}{j!(t-j)!}.
\end{aligned}$$

PROOF: see Appendix A □

This asymptotic normality result for $\hat{\alpha}$ will be complemented by a consistent variance estimator in section 3.5 to provide the basic tools for inference. Formally, the rate conditions above provide no practical guide to bandwidth choice (only a rate condition). While this theorem does not incorporate data-determined bandwidth choices, one could imagine using a leave-one-out cross-validation criterion evaluated at points outside a bandwidth neighborhood of the discontinuity. The resulting bandwidth could be used as a guide for bandwidth choice.

While we've shown that the estimator $\hat{\alpha}$ is the partially linear estimator, it is important to notice that the asymptotic variance in Theorem 2 differs from the corresponding expression for Robinson's estimator for the parametric component in the traditional partially linear model.

Look first at the denominator in (6). In the usual partially linear model, the analogous denominator would converge in probability to $E[(d - E(d|x))^2]$. In the discontinuity case, $d = E(d|x)$, so $E[(d - E(d|x))^2] = 0$ corresponding to the slower than $\sqrt{n}$-rate in this model. Here, $\frac{1}{nh}\sum_i (d_i - \sum_j w_j^i d_j)^2 = 2f_0(\bar{x}) \int_0^\infty (\int_w^\infty k(t)dt)^2 dw + o_p(1)$. In the numerator of the expression (6) for $\sqrt{nh}(\hat{\alpha} - \alpha)$, we have $\frac{1}{\sqrt{nh}}\sum_i (d_i - \sum_j w_j^i d_j)\ [\varepsilon_i - (\tilde{m}(x_i) - m(x_i))]$. In Robinson's partially linear case, the latter part of the expression $\tilde{m}(x_i) - m(x_i)$ makes no contribution to the influence function. In the discontinuity case, *both* terms ε_i and $\tilde{m}(x_i) - m(x_i)$, contribute to the influence function. In the numerator, the difference from the usual partially linear case is the leading term in the product, $d_i - \sum_j w_j^i d_j$. Again in the discontinuity case this term converges slowly to zero, so that as the sample size increases the influence of any fixed data point shrinks to zero (and more weight is given to new data points closer to the discontinuity) which leads to the nonparametric convergence rate.

To prove Theorem 2, I linearize the numerator of (7) and use the uniform convergence results of Appendix B to show the remainder ("quadratic") terms are asymptotically negligible.[7] Andrews (1995) notes that $\sqrt{n}$-consistent semiparametric estimators often require uniform convergence results of the form $n^{1/4}\sup_x |\hat{f}_0(x) - f_0(x)| = o_p(1)$. However, in the regression discontinuity context, the slower rate of convergence would seem to analogously require the much stronger result that $\left(\frac{n}{h}\right)^{1/4}\sup_x |\hat{f}_0(x) - f_0(x)| = o_p(1)$. In fact, the assumptions of Theorem 2 need not be so strong and only the weaker uniform convergence condition, explored by Andrews (1995) for the $\sqrt{n}$-consistent semiparametric case, is needed here. The reason that only the weaker result is required here is that the averages in these remainder terms are only taken over observations i with x_i in a shrinking neighborhood of the cut-off $\bar{x}$. As a result, these averages are $O_p(h)$ (rather than $O_p(1)$ as in the $\sqrt{n}$-rate semiparametric cases), so that only the less stringent uniform convergence rate result is needed in Theorem 2.[8]

[7] The lemmas of Appendix B contain apparently new boundary uniform convergence results and a uniform convergence result for regression functions that only assumes bounds on conditional moments.

[8] The mathematical details of this argument are given following equation (14) in the proof in Appendix A.

Bias

Since the asymptotic variance expression is the same in parts (a) and (b) of Theorem 2, the order of the asymptotic bias determines the fastest convergence rate for the partially linear estimator. In part (b) of the theorem, continuity of the derivatives of m at $\bar{x}$ is assumed (Assumption 2(b)), while in part (a) it is not assumed. The order of the bias depends crucially on this condition. When smoothness of the derivatives of m at $\bar{x}$ is not assumed, as in part (a), the order of the bias is $O(h^2)$. Higher-order bias-reducing kernels do not affect this rate. Comparing this rate to the order of bias for the Nadaraya-Watson estimator from Theorem 1, $O(h)$, the partially linear estimator already shows an improvement. That is, while the Nadaraya-Watson estimator behaves as one would expect from a boundary estimation problem, the partially linear estimator achieves an order of bias commensurate with kernel regression at an interior point of support, without higher-order kernels or derivative continuity in m at $\bar{x}$.

When continuity in the derivatives of m at $\bar{x}$ is assumed in part (b) of the theorem, the bias behavior becomes more interesting. In this case, the higher-order kernels are effective in reducing the bias further. In fact, given enough smoothness, the order of bias is reduced beyond what would be expected from a given higher-order kernel. If an s^{th}-order kernel is used in a typical interior point kernel regression problem, the asymptotic bias is of order $O(h^s)$. In the discontinuity problem, if s is even, the partially linear estimator yields a bias of order $O(h^{s+1})$. If s is odd, the order of the bias is further improved at $O(h^{s+2})$. Since symmetry of the kernels is assumed, for odd s, $\int k(u)u^s du = 0$. So it is not surprising that an order s (odd) kernel achieves the same bias reduction as an $s+1$-order kernel given sufficient smoothness in m. However, in the next section on local polynomial estimation, it will be seen that similar odd/even bias reduction behavior occurs even without higher-order kernels.

Using the Nadaraya-Watson estimator to estimate the discontinuity jump, the bias can be characterized as in a boundary point estimation problem. On the other hand, if the derivatives of m are continuous at $\bar{x}$, the bias of the partially linear estimator behaves quite differently. In the model (3), observations with $x \geq \bar{x}$ provide data on $m(x) + \alpha + \varepsilon$, while observations with $x < \bar{x}$ provide data on $m(x) + \varepsilon$. Hence, observations to the right of the discontinuity can be used to estimate $m(\bar{x})+\alpha$ and observations to the left of $\bar{x}$ can be used to estimate $m(\bar{x})$. Now consider a different model, where for some $\eta \in (0,1)$

$$y = m(x) + d\alpha(x) + \varepsilon, \qquad \text{where } E(\varepsilon|x,d) = 0 \text{ and } \Pr(d=1|x) \in [\eta, 1-\eta] \text{ for all } x. \tag{8}$$

Now suppose the parameter of interest is $\alpha(\bar{x})$. In this model, observations on $m(x)+\alpha(x)+\varepsilon$ are available for x's to the right and left of the discontinuity, and similarly for $m(x)+\varepsilon$. The observations with $d=1$ can then be used to provide an interior point Nadaraya-Watson kernel regression estimate of $m(\bar{x})+\alpha(\bar{x})$, and the $d=0$ observations can be used to estimate $m(\bar{x})$. Each of these estimators would have the asymptotic bias associated with a typical interior support point kernel regression problem. Under certain conditions, when the estimators are differenced to obtain an estimator of $\alpha(\bar{x})$, the leading bias terms from each estimator cancel to yield a bias improvement over the interior point kernel regression estimators. The partially linear estimator in the discontinuity problem mimics the bias behavior of the Nadaraya-Watson kernel regression estimator in the above setup. That is, instead of inheriting the poor bias behavior of a boundary problem, the bias of the partially linear estimator behaves more like there's data available to estimate both $m(\bar{x})+\alpha$ and $m(\bar{x})$ on *both* sides of the discontinuity. The partially linear estimator achieves this performance by actually using data on both sides of the discontinuity to estimate each point of the regression function. This use of the data allows the higher-order kernels to be effective in bias reduction. In the next subsection, it is seen that the local polynomial estimator, like the Nadaraya-Watson estimator, uses the data on each side of the discontinuity separately and then differences the level estimates. Yet, the local polynomial estimator achieves the same order of bias reduction as the partially linear estimator under Assumption 2(b) and improves on the partially linear bias performance when Assumption 2(b) does not hold. Next, a more mathematical description of the bias reduction for the partially linear estimator is provided.

From (7), the bias is approximately proportional to $\frac{1}{h}\ E\big[d\frac{\bar{f}_-(x)}{f_0(x)}\ (\bar{m}(x)-m(x))\big]\ -\frac{1}{h}\ E\big[(1-d_i)\frac{\bar{f}_+(x_i)}{f_0(x_i)}\ (\bar{m}(x)-m(x))\big]$, where $\bar{m}(x) = (f_0(x))^{-1}\int \frac{1}{h}k(\frac{x-u}{h})f_0(u)m(u)du$, $\bar{f}_+(x) = \int d(u)\frac{1}{h}k(\frac{x-u}{h})$ $f_0(u)du$, and $\bar{f}_-(x) = \int(1-d(u))\frac{1}{h}k(\frac{x-u}{h})f_0(u)du$. Importantly, this bias expression involves the difference in the bias in estimates of $m(\cdot)$ at an interior point, in contrast to the Nadaraya-Watson estimator of α which involves the nonparametric bias at a boundary. This advantage for $\hat{\alpha}$ comes from the use of data on both sides of the discontinuity in estimation of $m(\cdot)$. As n gets large the weights $(1-d_i)\frac{\bar{f}_+(x_i)}{f_0(x_i)}$ and $d_i\frac{\bar{f}_-(x_i)}{f_0(x_i)}$, on each term favor points closer to the discontinuity, and so these weights on each side tend to be close to a half. If the kernel regression bias $\bar{m}(x)-m(x)$ is smooth in a neighborhood of the discontinuity, then this difference can be approximated by a Taylor expansion. The leading, linear term of this expansion is zeroed out by the symmetry of the kernel, and the quadratic terms cancel in the differencing of the two bias terms. Thus without the use of higher-order bias-reducing kernels, the order of the bias for $\hat{\alpha}-\alpha$ is $O(h^3)$. Considering that the usual Nadaraya-Watson estimator $\tilde{\alpha}$ has bias of order $O(h)$ due to the boundary, the result that $\hat{\alpha}$ not only improves the bias to the usual order for kernel regression at an interior point ($O(h^2)$) but actually improves upon that order is rather remarkable.

3.4 Local Polynomial Estimation

Local polynomial estimators (Fan 1992) are known for their nice boundary behavior (Fan and Gijbels 1996, Ruppert and Wand 1994). A direct application of local polynomial estimation to the regression discontinuity setting would simply estimate the level of the conditional expectation on each side of the boundary and subtract, as in Nadaraya-Watson estimation. However, by accounting for the polynomial type behavior of the conditional expectation near the boundary, the estimate of the level at the boundary is free of the biases associated with Nadaraya-Watson estimation. The bias reduction here is not accomplished by higher-order bias-reducing kernels, but by the polynomial "correction."

The order p local polynomial estimator can be defined as follows. Suppose $(\bar{\alpha}_{p+}, \bar{\beta}_{p+})$ is the solution to the following minimization problem.

$$\min_{a,b_1,\ldots,b_p} \frac{1}{n}\sum_{i=1}^{n} k_h(x_i-\bar{x})d_i[y_i-a-b_1(x_i-\bar{x})-\cdots-b_p(x_i-\bar{x})^p]^2 \tag{9}$$

Similarly $(\bar{\alpha}_{p-}, \bar{\beta}_{p-})$ minimizes the analogous criterion with $1-d_i$ replacing d_i. Then

$$\bar{\alpha}_p = \bar{\alpha}_{p+} - \bar{\alpha}_{p-}.$$

Notice that if $p=0$, then this estimator (with no local polynomial correction) is identical to the

Nadaraya-Watson estimator. We, then, will focus on the cases with $p \geq 1$. Even when $p \geq 1$, the local polynomial estimator is very similar to the Nadaraya-Watson estimator in that $\bar{\alpha}_{p+}$ (and analogously $\bar{\alpha}_{p-}$) is simply a weighted average of the dependent variable from data just to one side of the discontinuity.

For our purposes, the polynomial term coefficients will be treated as nuisance parameters. Still it is worth pointing out that the optimal value $\bar{\beta}_{p-,j}$ of b_j from (9) estimates the j^{th} order right-hand derivative of m at $\bar{x}$. It is in this sense that local polynomial estimation directly accounts for the shape of the conditional expectation when estimating its level.

Hahn, Todd, and Van der Klaauw (2001) provide asymptotic distribution results for $p = 1$ (local linear estimation) under Assumption 2(a) and a slightly different set of rate and moment conditions. The theorem below describes the asymptotic distribution for the local polynomial estimator, $\bar{\alpha}_p$, with $p \geq 1$. In the next section, it is shown that by considering "higher-order" polynomials, $\bar{\alpha}_p$ can achieve the optimal rate of convergence. Also, the asymptotic bias results for the local polynomial estimator in Theorem 3 provide an interesting comparison to the higher-order kernel results for the partially linear estimator in Theorem 2.

Theorem 3 *Suppose Assumptions 1(a) and 3 hold.*
(a) If Assumption 2(a) holds with l_f nonnegative, $l_m \geq p + 1$, and $nh \longrightarrow \infty$, $h^{p+1}\sqrt{nh} \longrightarrow C_a$, where $0 \leq C_a < \infty$, then

$$\sqrt{nh}(\bar{\alpha}_p - \alpha) \xrightarrow{d} N\left(B_a, \frac{\sigma^{2+}(\bar{x}) + \sigma^{2-}(\bar{x})}{f_0(\bar{x})} e_1{}'\Gamma^{-1}\Delta\Gamma^{-1}e_1\right)$$

(b) If p is odd, Assumption 2 holds with l_f a positive integer, $l_m \geq p + 2$, and $nh^3 \longrightarrow \infty$, $h^{p+2}\sqrt{nh} \longrightarrow C_b$, where $0 \leq C_b < \infty$, then

$$\sqrt{nh}(\bar{\alpha}_p - \alpha) \xrightarrow{d} N\left(B_b, \frac{\sigma^{2+}(\bar{x}) + \sigma^{2-}(\bar{x})}{f_0(\bar{x})} e_1{}'\Gamma^{-1}\Delta\Gamma^{-1}e_1\right)$$

where

$$
\begin{aligned}
B_a &= \frac{C_a}{(p+1)!}\left[m^{(p+1)+}(\bar{x}) - (-1)^{p+1} m^{(p+1)-}(\bar{x})\right] e_1{}'\Gamma^{-1}\begin{pmatrix}\gamma_{p+1}\\ \vdots \\ \gamma_{2p+1}\end{pmatrix},\\
B_b &= 2C_b\left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!}\frac{f_0{}'(\bar{x})}{f_0(\bar{x})} + \frac{m^{(p+2)}(\bar{x})}{(p+2)!}\right) e_1{}'\Gamma^{-1}\begin{pmatrix}\gamma_{p+2}\\ \vdots \\ \gamma_{2p+2}\end{pmatrix}\\
&\quad -2C_b\frac{m^{(p+1)}(\bar{x})}{(p+1)!}\frac{f_0{}'(\bar{x})}{f_0(\bar{x})} e_1{}'\Gamma^{-1}\Gamma_{(+1)}\Gamma^{-1}\begin{pmatrix}\gamma_{p+1}\\ \vdots \\ \gamma_{2p+1}\end{pmatrix},
\end{aligned}
$$

and

$$
\begin{aligned}
\Gamma &= \begin{bmatrix}\gamma_0 & \cdots & \gamma_p\\ \vdots & & \vdots\\ \gamma_p & \cdots & \gamma_{2p}\end{bmatrix}, \Gamma_{(+1)} = \begin{bmatrix}\gamma_1 & \cdots & \gamma_{p+1}\\ \vdots & & \vdots\\ \gamma_{p+1} & \cdots & \gamma_{2p+1}\end{bmatrix}, \quad and \quad \Delta = \begin{bmatrix}\delta_0 & \cdots & \delta_p\\ \vdots & & \vdots\\ \delta_p & \cdots & \delta_{2p}\end{bmatrix},\\
e_1 &= (1,0,\ldots,0)', \quad \gamma_j = \int_0^\infty k(u)u^j du \quad and \quad \delta_j = \int_0^\infty k^2(u)u^j du.
\end{aligned}
$$

PROOF: see Appendix A □

The asymptotic variance in this result is of the same form as in usual local polynomial estimation for conditional means. The part of the limiting distribution that is special to RD treatment effects is the asymptotic bias. From Theorem 3, we see the identical convergence rate behavior as in Theorem 2(b) which characterizes the partially linear estimator. That is, when Assumption 2(b) holds giving continuity in the derivatives of m at $\bar{x}$, the local polynomial estimator also achieves an unexpectedly low order of asymptotic bias. For usual conditional expectation estimation, an order p local polynomial has an asymptotic bias of oder $O(h^p)$. For the RD model, which would appear to have a problematic bias problem due to its boundary-like features, a local polynomial of order p with p even has improved asymptotic bias of order $O(h^{p+1})$. Under Assumption 2(b), when p is odd, the bias is $O(h^{p+2})$, just as for the p^{th}-order kernel partially linear estimator from Theorem 2(b).

The results in Theorem 3(a) present a different picture. Whereas the partially linear estimator has bias of order $O(h^2)$ when Assumption 2(b) fails, the order p local polynomial estimator has bias of order $O(h^{p+1})$. Thus even without continuity in the derivatives of m at $\bar{x}$, the local polynomial estimator improves on the bias behavior familiar from interior point conditional expectation estimation with local polynomials. That is, regardless of whether Assumption 2(b) is presumed to hold, the bias of the local polynomial estimator behaves analogous to kernel regression estimation of $\alpha(\bar{x})$ in the smooth version of model (8), as described in 3.3. In section 4, it will be seen that due to this bias behavior, the local polynomial estimator can be used to attain the optimal rate of convergence regardless of whether Assumption 2(b) holds or not.

Note that under Assumption 2(b), the local polynomial estimator could be modified to constrain the derivative estimates to be the same on either side of the discontinuity. In that case, an alternative estimator would be the solution for a from the following optimization problem.[9]

$$\min_{c,a,b_1,\ldots,b_p} \frac{1}{n}\sum_{i=1}^{n} k_h(x_i-\bar{x})[y_i-c-d_i a-b_1(x_i-\bar{x})-\cdots-b_p(x_i-\bar{x})^p]^2 \tag{10}$$

If Assumption 2(b) holds, this estimator might be preferred over $\bar{\alpha}_p$ since it uses the information contained in the derivative restriction in that assumption. However, like the partially linear estimator, this estimator will not have the dramatic bias reductions achieved by $\bar{\alpha}_p$ when Assumption 2(b) does not hold.

3.5 Variance Estimation

To perform inference using any of the estimators covered above, consistent estimates of the expressions in the limiting distribution are often desired. The asymptotic bias expressions in Theorems 1, 2, and 3 involve complicated functionals of the kernel and the derivatives of f_0 and m at $\bar{x}$. Consistent estimation of the levels and derivatives of f_0 and m (or the right and left-hand limits and derivatives) is a straightforward nonparametric exercise, see Härdle (1990) or Pagan and Ullah (1999). As mentioned in 3.4, the relevant estimates for the derivatives of m are a byproduct of the optimization in (9).

The only parts of the asymptotic variance expressions from Theorems 1, 2, and 3 that have to be estimated are $\sigma^{2+}(\bar{x})$, $\sigma^{2-}(\bar{x})$, and $f_0(\bar{x})$. The remaining part of each asymptotic variance expression is a functional of the kernel and can be calculated exactly. The density at the discontinuity, $f_0(\bar{x})$ can

[9]A formal limiting distribution result for this case (analogous to Theorem 3) is available from the author.

be estimated consistently by kernel density estimation, as in Silverman (1992). The right and left-hand limits of the variance of the residual at the discontinuity can be estimated straightforwardly. Next a plug-in estimator is given for variance estimation that can be used with any of the RD estimators discussed above or others.

Suppose $\breve{\alpha}$ is a consistent estimator for α. Define right and left-hand variance estimators by

$$\breve{\sigma}^{2+}(\bar{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\left(\frac{\bar{x}-x_i}{h}\right)d_i\breve{\varepsilon}_i^2}{\frac{1}{2}\hat{f}_0(\bar{x})} \quad \text{and} \quad \breve{\sigma}^{2-}(\bar{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\left(\frac{\bar{x}-x_i}{h}\right)(1-d_i)\breve{\varepsilon}_i^2}{\frac{1}{2}\hat{f}_0(\bar{x})}$$

where

$$\breve{\varepsilon}_i = y_i - \breve{m}(x_i) - d_i\breve{\alpha} \quad \text{and}$$
$$\breve{m}(x) = \frac{\frac{1}{n}\sum_{j=1}^{n}\frac{1}{h}k\left(\frac{x-x_j}{h}\right)(y_j - d_j\breve{\alpha})}{\hat{f}_0(x)}.$$

THEOREM 4 *Suppose Assumptions 1(a), 2(a), and 3 with l_f and l_m positive integers and $\zeta \geq 2$. If $\frac{\sqrt{nh^2}}{\ln n} \longrightarrow \infty$ and $\breve{\alpha} \xrightarrow{p} \alpha$, then $\breve{\sigma}^{2+}(\bar{x}) \xrightarrow{p} \sigma^{2+}(\bar{x})$ and $\breve{\sigma}^{2-}(\bar{x}) \xrightarrow{p} \sigma^{2-}(\bar{x})$.*

PROOF: see Appendix A □

This result requires a more stringent moment condition (since $\zeta \geq 2$) than required for the limiting distribution results for estimators of α in the previous subsections, but otherwise presumes only the minimal parts of Assumptions 1, 2, and 3. In particular, this variance estimator and the above consistency theorem do not require continuity in the derivatives of m at the discontinuity, as in Assumption 2(b), or the higher-order kernels of Assumption 1(b). Hence, this variance estimator is applicable under any of the conditions given in Theorems 1, 2, and 3. Further, $\breve{\alpha}$ could be set equal to any of the estimators discussed in subsections 3.2 - 3.4, since it is simply required to be some consistent estimator of α.

3.6 Fuzzy Design

From the discussion in section 2, estimation of the RD treatment effect in the sharp design is exactly equivalent to estimation of α in (1). Hence, the estimators discussed above are directly applicable to the sharp design case. In the fuzzy design, the causal effect of interest can be expressed as the ratio of the discontinuity jumps in the conditional expectations of the outcome (y) and treatment (t). The discontinuity jump in the conditional expectation of the treatment is given by β in $t = j(x) + d\beta + \eta$, where $E(\eta|x) = 0$. If α and β denote these two discontinuity jump sizes, then taking the quotient of estimates of α and β gives an estimate of the fuzzy design RD treatment effect α/β. The asymptotic distribution of any such estimator can then be computed straightforwardly by application of the Delta Method. Let $\check{\alpha}$ and $\check{\beta}$ be estimators of α and β.

Proposition 1 *If*

$$\begin{pmatrix} \sqrt{nh}(\check{\alpha} - \alpha) \\ \sqrt{nh}(\check{\beta} - \beta) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} B_\alpha \\ B_\beta \end{pmatrix}, \begin{pmatrix} V_\alpha & C_{\alpha\beta} \\ C_{\alpha\beta} & V_\beta \end{pmatrix} \right),$$

then

$$\sqrt{nh}\left(\frac{\check{\alpha}}{\check{\beta}} - \frac{\alpha}{\beta}\right) \xrightarrow{d} N\left(\frac{1}{\beta}B_\alpha - \frac{\alpha}{\beta^2}B_\beta, \frac{1}{\beta^2}V_\alpha - 2\frac{\alpha}{\beta^3}C_{\alpha\beta} + \frac{\alpha^2}{\beta^4}V_\beta\right).$$

PROOF: By the Delta Method. □

If α and β are estimated using one of the estimators from 3.2 - 3.4, then Theorem 1, 2, or 3 can be used to provide the limiting marginal distributions for $\sqrt{nh}(\check{\alpha} - \alpha)$ and $\sqrt{nh}(\check{\beta} - \beta)$. Hence, analytic expressions for B_α, B_β, V_α, and V_β are given by the appropriate theorem. To complete the description of the limiting distribution for the RD treatment effect in the fuzzy design as given by $\sqrt{nh}\left(\frac{\check{\alpha}}{\check{\beta}} - \frac{\alpha}{\beta}\right)$ in Proposition 1, the joint limiting distribution of $\sqrt{nh}(\check{\alpha} - \alpha)$ and $\sqrt{nh}(\check{\beta} - \beta)$ is required. It follows that only the asymptotic covariance $C_{\alpha\beta}$ remains to be specified for each of the RD estimators.

Corollary 1 *Suppose $\sigma_{\varepsilon\eta}(x) = E(\varepsilon\eta|x)$ is continuous for $x \neq \bar{x}$, $x \in \mathcal{N}$, and the right and left-hand limits at $\bar{x}$ exist.*

(a) If $\check{\alpha}$ and $\check{\beta}$ are Nadaraya-Watson estimators and the conditions of Theorem 1 (along with the analogous conditions for the treatment t) are satisfied, then

$$C_{\alpha\beta} = 4\delta_0 \frac{\sigma^+_{\varepsilon\eta}(\bar{x}) + \sigma^-_{\varepsilon\eta}(\bar{x})}{f_0(\bar{x})}$$

(b) If $\breve{\alpha}$ and $\breve{\beta}$ are partially linear estimators and the conditions of Theorem 2(a) or (b) (along with the analogous conditions for the treatment t) are satisfied, then

$$C_{\alpha\beta} = v_k \frac{\sigma^+_{\varepsilon\eta}(\bar{x}) + \sigma^-_{\varepsilon\eta}(\bar{x})}{4f_0(\bar{x})}$$

(c) If $\breve{\alpha}$ and $\breve{\beta}$ are local polynomial estimators and the conditions of Theorem 3(a) or (b) (along with the analogous conditions for the treatment t) are satisfied, then

$$C_{\alpha\beta} = \frac{\sigma^+_{\varepsilon\eta}(\bar{x}) + \sigma^-_{\varepsilon\eta}(\bar{x})}{f_0(\bar{x})} e_1{}'\Gamma^{-1}\Delta\Gamma^{-1}e_1.$$

PROOF: In each case, the Cramer-Wold device can be used to establish the joint limiting distribution of $\sqrt{nh}(\breve{\alpha} - \alpha)$ and $\sqrt{nh}(\breve{\beta} - \beta)$ following closely the scalar derivation in the proof of the corresponding theorem. The derivations of the covariance expressions follow derivations of the expressions for V_α in the proof of the corresponding theorem. □

This result is written to incorporate whatever convergence rates are called for by the smoothness and moments conditions of the relevant theorem. Note that covariance estimation proceeds just as variance estimation in 3.5. One could also imagine using different estimators for α and β, and their limiting joint distribution could then be derived from the influence functions given in the proofs of the corresponding theorems.

4 Optimal Rates of Convergence

The asymptotic results of the previous section seem to confirm the intuition that α can only be estimated at a slower than parametric $\sqrt{n}$-rate. In this section, we derive the optimal rate of convergence, under Stone's (1980) definition, for estimation of the discontinuity jump size.

Efficient nonparametric estimation of a conditional mean at an *interior* point of support has been studied extensively and corresponding optimal rates for a given smoothness class are well known. Conditional mean *boundary* estimation is considered in Cheng, Fan, and Marron (1997). Their efficiency criteria is minimaxity with respect to mean-squared-error risk over linear smoothers.[10]

[10] Cheng, Fan, and Marron (1997) also develop optimal kernel choices in this context.

This criteria allows them to characterize both best rates and best constants (and show the efficiency of local polynomial estimation at the boundary). Their efficiency results might be extended beyond the class of linear estimators by the modulus of continuity approach developed in Donoho and Liu (1987, 1991). Below we consider best rates as defined by Stone (1980). Stone rates do not depend on a choice of loss function and are established relative to all estimators (not just linear smoothers). They are essentially the rate at which a best test has nondegenerate local power.

Note that the Cheng-Fan-Marron best rate for conditional mean boundary point estimation among linear smoothers yields a lower bound for the corresponding Stone rate. On the other hand, optimal rates for estimation of a conditional mean at an interior point yields an upper bound on the optimal rate at the boundary. That is, for estimation purposes, an interior point could be treated as if it were a boundary point, so that boundary point estimators are included in the class of interior point estimators. Finally, note that these upper and lower bounds are equal, so the optimal rate for estimating a regression function at a boundary point of support is the same as the optimal rate at an interior point.

Below we derive results on best rates for estimation of a discontinuity jump size. From (1), the jump size can be expressed as the difference of conditional mean boundary points. Hence, the optimal rate for conditional mean boundary point estimation provides a lower bound on the optimal rate for jump size estimation and, in turn, for RD treatment effects. Theorem 5 below shows that even under Assumption 2(b), giving continuity in the derivatives of the regression function at the discontinuity point, this lower bound is tight. It follows that the partially linear and local estimators are, in fact, optimal as described below.

Stone (1980) establishes the optimal convergence rate for estimation of a conditional mean at a point. The best rate depends on the smoothness of the conditional mean (as well as the dimension of the conditioning variable). Define $\theta(x) = m(x) + \alpha\mathbf{1}\{x \geq \bar{x}\}$ and $\alpha(\theta) = \lim_{x\downarrow\bar{x}} \theta(x) - \lim_{x\uparrow\bar{x}} \theta(x) = \alpha$, where the last equality assumes continuity of m at $\bar{x}$. Then the optimal rate for estimation of α will depend on the class of allowable functions Θ. Stone defines r as an upper bound to the rate of convergence if for every sequence of estimators $\{\check{\alpha}_n\}$,

$$\liminf_{n\longrightarrow\infty}\sup_{\theta\in\Theta} P_\theta(|\check{\alpha}_n - \alpha(\theta)| > cn^{-r}) > 0 \quad \text{for all } c > 0 \tag{11}$$

and

$$\lim_{c\longrightarrow 0}\liminf_{n\longrightarrow\infty}\sup_{\theta\in\Theta} P_\theta(|\check{\alpha}_n - \alpha(\theta)| > cn^{-r}) = 1. \tag{12}$$

If r is an upper bound to the rate of convergence and there exists an estimator $\check{\alpha}_n$ such that

$$\lim_{c\longrightarrow\infty}\limsup_{n\longrightarrow\infty}\sup_{\theta\in\Theta} P_\theta(|\check{\alpha}_n - \alpha(\theta)| > cn^{-r}) = 0, \tag{13}$$

then r is also an optimal rate of convergence. We will find that under Assumption 2(b), continuity in the derivatives of m at $\bar{x}$, the partially linear estimator and the local polynomial estimator achieve the optimal rate of convergence. Moreover, when Assumption 2(b) fails, the local polynomial estimator still achieves the optimal rate, while the partially linear estimator generally does not. However, the order of the local polynomial estimator used to achieve optimality depends on whether Assumption 2(b) holds or not.

Let $\mathcal{C}_p$ denote the class of p times continuously differentiable functions on $\mathcal{N}$. Suppose m_0 is a fixed continuous function; α_0 is a fixed real number; and $\theta_0(x) = m_0(x) + \alpha_0\mathbf{1}\{x \geq \bar{x}\}$. Define $\Theta = \{\theta_0 + \theta : \theta(x) = g(x) + \gamma\mathbf{1}\{x \geq \bar{x}\}, g \in \mathcal{C}_p, \gamma \in \mathbb{R}\}$. Given that θ is an unknown element of Θ, the problem is to estimate $\alpha(\theta)$.

For the case when Assumption 2(b) fails, we also define $\bar{\mathcal{C}}_p$ to be the class of continuous functions on $\mathcal{N}$ and p times continuously differentiable at $x \in \mathcal{N}\backslash\{\bar{x}\}$ with finite right and left-hand derivatives to order p at $\bar{x}$. Correspondingly, let $\bar{\Theta} = \{\theta_0 + \theta : \theta(x) = g(x) + \gamma\mathbf{1}\{x \geq \bar{x}\}, g \in \bar{\mathcal{C}}_p, \gamma \in \mathbb{R}\}$.

Given these definitions, optimal convergence rates for α are determined by equations (11), (12), and (13) for each of the two classes, Θ and $\bar{\Theta}$, of allowable conditional expectation functions. For each of the limiting distribution results in section 3, conditions are given on the conditional distribution of y given x that generates the observed data. To obtain an optimal rate result, a *class* of potential distributions indexed by θ must be considered. The next assumption specifies the manner in which the conditional distribution of y given x may depend on θ.

Assumption 4 *(a) The conditional distribution of y given x can be written as $f(y|x, \theta(x))$.*[11] *If $\mathcal{N}$ is a neighborhood of $\bar{x}$ as in Assumption 2, let T denote an open interval containing $\{\theta(x) : \theta \in \Theta$ and $x \in \mathcal{N}\}$. Suppose $f(y|x, t)$ is positive and twice continuously differentiable in t on the support for y and $\mathcal{N} \times T$, and $\int sup_{t \in T} |\frac{\partial}{\partial t} f(y|x, t)|dy$ and $\int sup_{t \in T} |\frac{\partial^2}{\partial t^2} f(y|x, t)|dy$ are finite.*
(b) For some positive constants τ_0 and C, there is a function $M(y|x, t)$ such that on T, $|l''(y|x, t+\tau)| \leq M(y|x, t)$ for $|\tau| \leq \tau_0$ and $\int M(y|x, t)f(y|x, t)dy \leq C$.

The first part of this assumption assures that the order of integration and differentiation can be interchanged in the equation $\int f(y|x, t)dy = 1$ to yield $\int \frac{\partial}{\partial t} f(y|x, t)dy = 0$ and $\int \frac{\partial^2}{\partial t^2} f(y|x, t)dy = 0$. These equations in conjunction with part (b) of the assumption will allow us to bound the first order term in a Taylor expansion of the likelihood function via the Cauchy-Schwarz Inequality and the information matrix equality. With Assumption 4 in hand, an optimal convergence rate result can now be stated.

The proof of the following result follows the approach outlined in Stone (1980). Given a true $\theta \in \Theta$, one constructs a sequence $\theta_n \in \Theta$ that approaches θ. If θ and θ_n are difficult to distinguish using a best test, then typically $\check{\alpha}_n - \alpha$ is bounded below by $|\alpha(\theta_n) - \alpha(\theta)|/2$. For appropriate choice of a "least favorable" sequence θ_n, a meaningful upper bound to the rate of convergence is established. It then remains to show that some estimator achieves this upper rate to prove optimality of the rate.

Theorem 5 *Suppose x is absolutely continuous and its density f_0 and $\sigma^2(x) = Var(y|x)$ are bounded away from zero and infinity on $int(\mathcal{N})$.*
(a) If $\theta \in \Theta$ and the distribution of (y, x) satisfies Assumption 4, then the optimal rate of convergence for estimation of $\alpha(\theta)$ is $\frac{p}{2p+1}$;
(b) If $\theta \in \bar{\Theta}$ and the distribution of (y, x) satisfies Assumption 4, then the optimal rate of convergence for estimation of $\alpha(\theta)$ is $\frac{p}{2p+1}$.

[11] More formally, we could let the conditional distribution be $f(y|x, \theta(x))\psi(dy)$ for some measure ψ on $\mathbb{R}$. As usual, for notational simplicity, we'll suppress ψ and understand that subsequent integrals with respect to y are taken over a given measure that does not need to be Lebesgue (and could include discrete distributions).

PROOF: see Appendix A □

These optimal rates are the same as those established by Stone for estimation of a conditional expectation at an interior point of the support. Both partially linear and local polynomial estimators achieve the optimal rate of part (a) of Theorem 5, while the partially linear estimator does not generally achieve the optimal rate under the conditions in (b). What this theorem does not reveal is the interesting way in which these estimators achieve the rate in part (a). Specifically, from Theorem 2, the partially linear estimator with bias-reducing kernels of order only $p-1$ (or only $p-2$ if p is odd) achieve the optimal rate. Similarly, the local polynomial estimator only requires polynomials of order $p-1$ (or $p-2$ if p is odd). In contrast, from Stone (1980), to achieve the above rates in estimation of a conditional mean at an interior point of support, one uses either bias-reducing kernels of order p (using the Nadaraya-Watson estimator) or polynomials of order p. To attain the rate in part (b) of this theorem, the local polynomial estimator still only requires use of a polynomial of order $p-1$.

Note that the optimal convergence rate result of Theorem 5 can be extended naturally to the case of higher-dimensional covariates. In these cases, the optimal rate depends on the form of the discontinuity frontier.

5 Extensions

This paper has concentrated exclusively on the univariate RD model. The results here naturally extend to the multivariate model, where the covariate space is partitioned by a discontinuity *frontier*. The goal would be to estimate the size of the discontinuity jump in a conditional expectation along the frontier. The appropriate estimator would depend on the variation in the jump size along the frontier. When the discontinuity jump is not constant along the frontier, the size of the jump can be estimated at any point along the frontier. The local polynomial estimator extends straightforwardly to this case by simply using a multivariate local polynomial in estimation. On the other hand, the partially linear estimator naturally extends to the constant discontinuity frontier case. The same partially linear estimator as above can be used with a multivariate kernel and d now denoting an indicator for one part of the covariate space bounded by the discontinuity frontier. In both of these cases, the additional bias reductions seen in some of the univariate settings would require symmetry around the point or frontier where estimation occurs.[12] Another natural extension to the partially linear estimator occurs when the additional covariate dimensions enter the conditional expectation in the form of an additively separable linear (or parametric) term. Then, estimation of the linear component at a parametric rate would not affect the limiting distribution of the discontinuity size estimates.

Another extension is the model with *unknown* discontinuity point (or frontier). As in the time series break point problem, estimators of the cut-off point have a faster convergence rate than the estimators of the discontinuity size and so do not affect the limiting distribution results given here. Results in this case would allow empirical applications of the RD design to the common case where the cut-off point is unknown.

6 Conclusion

As a semiparametric problem, estimation in the RD model differs from the typical case found in econometrics. Here the inherent non-smoothness given by the discontinuity is itself the object of estimation interest. Because nonparametric boundary estimation commonly encounters bias problems, RD estimation might be expected face bias difficulties, too. I find RD estimation can suffer bias problems, but we can overcome these problems by an appropriate choice of estimator.

In particular, I propose two estimators that attain the optimal convergence rate under varying conditions. Following the approach of Stone (1980), we establish upper bounds on the rate of convergence for estimation of the RD treatment effect under two sets of conditions. One set of conditions assumes smoothness in the derivatives of the conditional expectation at the discontinuity point, while under the other set of conditions, that smoothness assumption is relaxed. When the smoothness in the derivatives at the discontinuity holds, the upper bound can be attained using a partially linear estimator. The order of the bias-reducing kernel used in the partially linear estimator

[12]The interested reader may obtain from the author multivariate results circulated in an earlier draft of this paper under the title "Semiparametric Estimation of Regression Discontinuity Models."

to achieve a given order of bias reduction is actually lower than the order of the kernel required for the same bias reduction in the typical nonparametric problem of estimating a conditional mean at an interior point of the covariate support. Similar bias-reduction properties are enjoyed by the local polynomial estimator. In fact, the local polynomial estimator attains the upper bound on the rate of convergence under either set of smoothness conditions, making it a potentially more robust estimator.

A Proofs

In the proofs below, let C denote a generic positive constant. Also, given $\mathcal{N}$ as defined in Assumption 2, let $\mathcal{N}_0$ be a compact interval such that $\bar{x} \in int(\mathcal{N}_0)$ and $\mathcal{N}_0 \subset int(\mathcal{N})$. In accordance with Assumption 1(a), suppose the support of the kernel k is $[-M, M]$.

PROOF OF THEOREM 1: As stated in section 3, the conclusion of Theorem 1 follows from standard results in nonparametric kernel estimation, as follows.

$$\begin{aligned}\sqrt{nh}(\tilde{\alpha}-\alpha) &= \frac{\frac{1}{\sqrt{nh}}\sum_i k\left(\frac{\bar{x}-x_i}{h}\right) d_i[m(x_i)-m(\bar{x})+\varepsilon_i]}{\frac{1}{n}\sum_j \frac{1}{h}k\left(\frac{\bar{x}-x_j}{h}\right) d_j} \\ &\quad - \frac{\frac{1}{\sqrt{nh}}\sum_i k\left(\frac{\bar{x}-x_i}{h}\right)(1-d_i)[m(x_i)-m(\bar{x})+\varepsilon_i]}{\frac{1}{n}\sum_j \frac{1}{h}k\left(\frac{\bar{x}-x_j}{h}\right)(1-d_j)}\end{aligned}$$

First consider the denominator of the first term. Show $(nh)^{-1}\sum_j k((\bar{x}-x_j)/h)d_j \xrightarrow{p} f_0(\bar{x})/2$.

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{n}\sum_j \frac{1}{h}k\left(\frac{\bar{x}-x_j}{h}\right)d_j\right) &\leq \frac{1}{nh^2}E\left[k^2\left(\frac{\bar{x}-x}{h}\right)d\right] \\ &= \frac{1}{nh}\int_0^M k^2(u)f_0(\bar{x}+uh)du \\ &= O\left(\frac{1}{nh}\right) \;=\; o(1).\end{aligned}$$

Then, by Chebyshev's Inequality,

$$\begin{aligned}\frac{1}{n}\sum_j \frac{1}{h}k\left(\frac{\bar{x}-x_j}{h}\right)d_j &= E\left[\frac{1}{h}k\left(\frac{\bar{x}-x}{h}\right)d\right] \;+\; o_p(1) \\ &= f_0(\bar{x})\int_0^M k(u)du \;+\; o_p(1) \\ &= \frac{1}{2}f_0(\bar{x}) \;+\; o_p(1)\end{aligned}$$

Similarly, $(nh)^{-1}\sum_j k((\bar{x}-x_j)/h)(1-d_j) \xrightarrow{p} f_0(\bar{x})/2$.

Next verify Liapunov's condition. For large enough n,

$$\begin{aligned}
\sum_i E\left|\frac{1}{\sqrt{nh}}k\left(\frac{\bar{x}-x_i}{h}\right)d_i\varepsilon_i\right|^{2+\zeta} &= \frac{1}{(nh)^{\zeta/2}}E\left[\frac{1}{h}\left|k\left(\frac{\bar{x}-x}{h}\right)\right|^{2+\zeta} dE(|\varepsilon|^{2+\zeta}|x)\right] \\
&\leq \frac{1}{(nh)^{\zeta/2}}\left[\sup_{x\in\mathcal{N}} E(|\varepsilon|^{2+\zeta}|x)\right]\int_0^M |k(u)|^{2+\zeta} f_0(\bar{x}+uh)du \\
&= o(1).
\end{aligned}$$

Thus the CLT can be applied. Derive the asymptotic variance,

$$\begin{aligned}
\sum_j \mathrm{Var}\left(\frac{1}{\sqrt{nh}}k\left(\frac{\bar{x}-x_j}{h}\right)d_j\varepsilon_j\right) &= E\left[\frac{1}{h}k^2\left(\frac{\bar{x}-x}{h}\right)d\sigma^2(x)\right] \\
&= \int_0^M k^2(u)\sigma^2(\bar{x}+uh)f_0(\bar{x}+uh)du \\
&= \sigma^{2+}(\bar{x})f_0(\bar{x})\int_0^M k^2(u)du \;+\; o(1).
\end{aligned}$$

Then by Liapunov's CLT,

$$\frac{\frac{1}{\sqrt{nh}}\sum_i k\left(\frac{\bar{x}-x_i}{h}\right)d_i\varepsilon_i}{\frac{1}{n}\sum_j \frac{1}{h}k\left(\frac{\bar{x}-x_j}{h}\right)d_j} - \frac{\frac{1}{\sqrt{nh}}\sum_i k\left(\frac{\bar{x}-x_i}{h}\right)(1-d_i)\varepsilon_i}{\frac{1}{n}\sum_j \frac{1}{h}k\left(\frac{\bar{x}-x_j}{h}\right)(1-d_j)} \xrightarrow{d} N\left(0, 4\frac{\sigma^{2+}(\bar{x})+\sigma^{2-}(\bar{x})}{f_0(\bar{x})}\int_0^M k^2(u)du\right).$$

Finally, consider the bias of the estimator.

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{\sqrt{nh}}\sum_i k\left(\frac{\bar{x}-x_i}{h}\right)d_i[m(x_i)-m(\bar{x})]\right) &\leq \frac{1}{h}E\left[k^2\left(\frac{\bar{x}-x}{h}\right)d[m(x)-m(\bar{x})]^2\right) \\
&\leq \left[\sup_{v\in[0,M]}|m(\bar{x}+vh)-m(\bar{x})|\right]^2\int_0^M k^2(u)f_0(\bar{x}+uh)du \\
&= o(1).
\end{aligned}$$

So again by Chebyshev's Inequality,

$$\begin{aligned}
&\frac{\frac{1}{\sqrt{nh}}\sum_i k\left(\frac{\bar{x}-x_i}{h}\right)d_i[m(x_i)-m(\bar{x})]}{\frac{1}{n}\sum_j \frac{1}{h}k\left(\frac{\bar{x}-x_j}{h}\right)d_j} - \frac{\frac{1}{\sqrt{nh}}\sum_i k\left(\frac{\bar{x}-x_i}{h}\right)(1-d_i)[m(x_i)-m(\bar{x})]}{\frac{1}{n}\sum_j \frac{1}{h}k\left(\frac{\bar{x}-x_j}{h}\right)(1-d_j)} \\
&= \frac{2}{f_0(\bar{x})}\sqrt{\frac{n}{h}}\left(E\left[k\left(\frac{\bar{x}-x}{h}\right)d[m(x)-m(\bar{x})]\right] - E\left[k\left(\frac{\bar{x}-x}{h}\right)(1-d)[m(x)-m(\bar{x})]\right]\right) \;+\; o_p(1) \\
&= \frac{2\sqrt{nh}}{f_0(\bar{x})}\int_0^M k(u)\left\{[m(\bar{x}+uh)-m(\bar{x})]f_0(\bar{x}+uh) - [m(\bar{x}-uh)-m(\bar{x})]f_0(\bar{x}-uh)\right\}du \;+\; o_p(1) \\
&= \sqrt{nh}\frac{2}{f_0(\bar{x})}\left\{h(m'^{+}(\bar{x})+m'^{-}(\bar{x}))f_0(\bar{x})\int_0^M k(u)udu \;+\; o(h)\right\} \;+\; o_p(1).
\end{aligned}$$

The conclusion of the theorem follows. □

PROOF OF THEOREM 2:

From (7),

$$\sqrt{nh}(\hat{\alpha}-\alpha) = \left[\frac{1}{nh}\sum_i \left(d_i \frac{\hat{f}_-^2(x_i)}{\hat{f}_0^2(x_i)} + (1-d_i)\frac{\hat{f}_+^2(x_i)}{\hat{f}_0^2(x_i)}\right)\right]^{-1}$$
$$\cdot \frac{1}{\sqrt{nh}}\sum_i \left[\left(d_i \frac{\hat{f}_-(x_i)}{\hat{f}_0(x_i)} - (1-d_i)\frac{\hat{f}_+(x_i)}{\hat{f}_0(x_i)}\right)[\varepsilon_i - (\tilde{m}(x_i) - m(x_i))]\right].$$

For the numerator, take an expansion to linearize.

$$\begin{aligned}
&\frac{1}{\sqrt{nh}}\sum_i \left[\left(d_i \frac{\hat{f}_-(x_i)}{\hat{f}_0(x_i)} - (1-d_i)\frac{\hat{f}_+(x_i)}{\hat{f}_0(x_i)}\right)[\varepsilon_i - (\tilde{m}(x_i) - m(x_i))]\right] \\
&= \frac{1}{\sqrt{nh}}\sum_i \left[\frac{d_i\varepsilon_i}{f_0(x_i)}(\hat{f}_-(x_i) - \bar{f}_-(x_i)) - \frac{(1-d_i)\varepsilon_i}{f_0(x_i)}(\hat{f}_+(x_i) - \bar{f}_+(x_i))\right] \\
&\quad + \frac{1}{\sqrt{nh}}\sum_i \left\{\left(d_i\frac{\bar{f}_-(x_i)}{f_0(x_i)} - (1-d_i)\frac{\bar{f}_+(x_i)}{f_0(x_i)}\right)\left(\frac{-\varepsilon_i}{f_0(x_i)}(\hat{f}_0(x_i) - f_0(x_i))\right.\right. \qquad (14) \\
&\qquad \left.\left. + \frac{m(x_i)}{f_0(x_i)}(\hat{f}_0(x_i) - f_0(x_i)) - \frac{1}{f_0(x_i)}(\tilde{r}(x_i) - r(x_i)) + \varepsilon_i\right)\right\} + Q_n.
\end{aligned}$$

Since $d(x)\hat{f}_-(x) = 0$ if $x \notin [\bar{x}, \bar{x}+Mh]$, d_i can be replaced everywhere in the above expression by $\mathbf{1}\{\bar{x} \le x_i \le \bar{x}+Mh\}$. Similarly, $1-d_i$ can be replaced by $\mathbf{1}\{\bar{x}-Mh \le x_i < \bar{x}\}$. Q_n denotes the quadratic terms in the expansion,

$$\begin{aligned}
Q_n &= \frac{1}{\sqrt{nh}}\sum_i \left\{ \left(d_i\bar{f}_-(x_i) - (1-d_i)\bar{f}_+(x_i)\right)\left[\frac{(\varepsilon_i\hat{f}_0(x_i) - m(x_i)(\hat{f}_0(x_i) + f_0(x_i)))}{f_0^2(x_i)\hat{f}_0^2(x_i)}\right.\right. \\
&\qquad \left. \cdot(\hat{f}_0(x_i) - f_0(x_i))^2 + \frac{\hat{f}_0(x_i) + f_0(x_i)}{f_0^2(x_i)\hat{f}_0^2(x_i)}(\tilde{r}(x_i) - r(x_i))(\hat{f}_0(x_i) - f_0(x_i))\right] \\
&\qquad + \frac{1}{f_0(x_i)\hat{f}_0(x_i)}[(\tilde{m}(x_i) - \varepsilon_i)(\hat{f}_0(x_i) - f_0(x_i)) - (\tilde{r}(x_i) - r(x_i))] \\
&\qquad \left. \cdot[d_i(\hat{f}_-(x_i) - \bar{f}_-(x_i)) - (1-d_i)(\hat{f}_+(x_i) - \bar{f}_+(x_i))]\right\},
\end{aligned}$$

for $\tilde{r}(x) = \frac{1}{nh}\sum_j k(\frac{x-x_j}{h})\tilde{y}_j$ and $r(x) = m(x)f_0(x)$. Here $\hat{f}_+(x)$ $[\hat{f}_-(x)]$ is linearized around $\bar{f}_+(x)$ $[\bar{f}_-(x)]$ rather than its limit, since for fixed $x < \bar{x}$ $[x \ge \bar{x}]$, the limit is zero.

Show $Q_n = o_p(1)$. Note that $\sup_{x\in\mathcal{N}_0} |1/\hat{f}_0(x)| = O_p(1)$ since $\hat{f}_0$ is uniformly consistent on $\mathcal{N}_0$ by Lemmas B1 and B4 and f_0 is bounded away from zero on $\mathcal{N}$ by Assumption 2. Also, by boundedness of k and f_0, there exists $C < \infty$ such that $|d(x)\bar{f}_-(x)| < C$ and $|(1-d(x))\bar{f}_+(x)| < C$

for all x. For some $C < \infty$ and n large enough (h small enough),

$$\begin{aligned}
\frac{1}{h}E\left[|\varepsilon|\mathbf{1}\{\bar{x} \leq x \leq \bar{x}+Mh\}\right] &\leq \frac{1}{h}\left[\sup_{x\in\mathcal{N}} E(|\varepsilon||x)\right] E\mathbf{1}\{\bar{x} \leq x \leq \bar{x}+Mh\} \\
&= \left[\sup_{x\in\mathcal{N}} E(|\varepsilon||x)\right] \int_0^M f_0(\bar{x}+uh)du \\
&< C
\end{aligned}$$

The last inequality follows by Assumptions 2(a) and 3(b). So, by Markov's Inequality, $\frac{1}{nh}\sum_i |\varepsilon_i|$ $\mathbf{1}\{\bar{x} \leq x_i \leq \bar{x}+Mh\} = O_p(1)$. Similarly, $\frac{1}{nh}\sum_i \mathbf{1}\{\bar{x} \leq x_i \leq \bar{x}+Mh\} = O_p(1)$.

First consider the quadratic terms (in Q_n) that include the difference $\tilde{r}(x_i) - r(x_i)$. The convergence of such terms differs under the assumptions of parts (a) and (b) of the theorem. Let $\bar{r}(x)$ $= E\tilde{r}(x) = \int \frac{1}{h}k(\frac{x-u}{h})r(u)du$. Under the assumptions of part (a), for n large enough,

$$\begin{aligned}
&\left|\frac{1}{\sqrt{nh}}\sum_i \frac{d_i\bar{f}_-(x_i)(\hat{f}_0(x_i)+f_0(x_i))}{f_0^2(x_i)\hat{f}_0^2(x_i)}(\tilde{r}(x_i)-r(x_i))(\hat{f}_0(x_i)-f_0(x_i))\right| \qquad (15)\\
\leq\;& \sqrt{nh}\left[\sup_x |d\bar{f}_-(x)|\right]\left[\sup_{x\in\mathcal{N}_0}\left|\frac{\hat{f}_0(x)+f_0(x)}{f_0^2(x)\hat{f}_0^2(x)}\right|\right]\left[\sup_{x\in\mathcal{N}_0}|\tilde{r}(x)-\bar{r}(x)| + \sup_{\bar{x}\leq x\leq\bar{x}+Mh}|\bar{r}(x)-r(x)|\right] \\
&\cdot\left[\sup_{x\in\mathcal{N}_0}|\hat{f}_0(x)-f_0(x)|\right]\frac{1}{nh}\sum_i \mathbf{1}\{\bar{x}\leq x_i\leq\bar{x}+Mh\} \\
=\;& \sqrt{nh}\;O_p(1)\;O_p\left(\sqrt{\frac{\ln n}{nh}}+h\right)\;O_p\left(\sqrt{\frac{\ln n}{nh}}+h^s\right) \\
=\;& O_p\left(\frac{\ln n}{\sqrt{nh}} + h\sqrt{\ln n} + \sqrt{nh^{2s+3}}\right)
\end{aligned}$$

The first equality follows by Lemmas B1 - B4, and this quadratic term is $o_p(1)$ under the bandwidth rate conditions in part (a). The other terms in Q_n that include the difference $\tilde{r}(x_i) - r(x_i)$ follow similarly under the assumptions of part (a) of the theorem.

If the conditions of part (b) hold, then by Lemma B4, $\sup_{x\in\mathcal{N}} |\bar{r}(x)-r(x)| = O_p(h^s)$. With this modification, the above argument can be used to show that the quadratic term examined above in (15) is $O_p(\ln n/\sqrt{nh} \;\; + h^s\sqrt{\ln n} \;\; + \sqrt{nh^{4s+1}})$, which is $o_p(1)$ under the rate conditions of part (b). The other terms in Q_n that include the difference $\tilde{r}(x_i) - r(x_i)$ follow similarly.

The argument for convergence of the remaining quadratic terms is the same under the assumptions of part (a) or (b) of the theorem. Take a typical term, for n large enough,

$$\begin{aligned}
&\left|\frac{1}{\sqrt{nh}}\sum_i \frac{d_i\varepsilon_i\bar f_-(x_i)}{f_0^2(x_i)\hat f_0(x_i)}(\hat f_0(x_i)-f_0(x_i))^2\right| \\
\leq\ & \sqrt{nh}\left[\sup_{x\in\mathcal{N}_0}\frac{1}{f_0^2(x)\hat f_0(x)}\right]\left[\sup_x |d_i\bar f_-(x_i)|\right]\left[\sup_{x\in\mathcal{N}_0}|\hat f_0(x)-f_0(x)|\right]^2\frac{1}{nh}\sum_i|\varepsilon_i|\mathbf{1}\{\bar x\leq x_i\leq \bar x+Mh\} \\
=\ & O_p\left(\frac{\ln n}{\sqrt{nh}}+h^s\sqrt{\ln n}+\sqrt{nh^{4s+1}}\right).
\end{aligned}$$

The last equality follows by uniform convergence of the kernel density estimator on $\mathcal{N}_0$ from Lemmas B1 and B4. The rate conditions in both parts (a) and (b) imply that this term is $o_p(1)$. The remaining quadratic terms are shown $o_p(1)$ similarly.

Next a U-statistic projection is applied to the first two terms of the linearization in (14) to show that they are $o_p(1)$. Let $z_i = (x_i, \varepsilon_i)$. Then $\frac{1}{\sqrt{nh}}\sum_i \frac{d_i\varepsilon_i}{f_0(x_i)}\ \hat f_-(x_i) = \sqrt{\frac{n}{h}}\frac{1}{n^2}\sum_{i,j} b_n(z_i,z_j)$, where $b_n(z_i,z_j) = d_i\frac{\varepsilon_i}{f_0(x_i)}\ \frac{1}{h}k(\frac{x_i-x_j}{h})(1-d_j)$. Note that $b_n(z_i,z_i)=0$ so that this term is a U-statistic. By Assumptions 1(a), 2(a), and 3(b), $E(b_n(z_i,z_j)^2) = \int_0^M\int_{-M}^{-w}\frac{\sigma^2(\bar x+wh)}{f_0(\bar x+wh)}k^2(u)f_0(\bar x+wh+uh)dudw < \infty$. Then by standard U-statistic projection results, see Serfling (1980) section 5.3, $\frac{1}{\sqrt{nh}}\sum_i\frac{d_i\varepsilon_i}{f_0(x_i)}\ (\hat f_-(x_i)-\bar f_-(x_i)) = \sqrt{\frac{n}{h}}O_p(\frac{[E(b_n(z_i,z_j)^2)]^{1/2}}{n}) = O_p(1/\sqrt{nh}) = o_p(1)$. The second term follows similarly.

Now consider the next term in (14). Show $\frac{1}{\sqrt{nh}}\sum_i\left(d_i\frac{\bar f_-(x_i)}{f_0(x_i)}-(1-d_i)\frac{\bar f_+(x_i)}{f_0(x_i)}\right)\left(\frac{-\varepsilon_i}{f_0(x_i)}(\hat f_0(x_i)-f_0(x_i))\right)$ $= o_p(1)$. Let $b_n(z_i,z_j) = d_i\frac{\bar f_-(x_i)\varepsilon_i}{f_0^2(x_i)}\frac{1}{h}k(\frac{x_i-x_j}{h})$. Then $\frac{1}{\sqrt{nh}}\sum_i d_i\frac{\bar f_-(x_i)\varepsilon_i}{f_0^2(x_i)}\hat f_0(x_i) = \sqrt{\frac{n}{h}}\frac{1}{n^2}\sum_{i,j}b_n(z_i,z_j)$. As above, $E(b_n(z_i,z_j)^2)<\infty$. Also,

$$\begin{aligned}
E|b_n(z_i,z_i)| &= E\left|d_i\frac{\bar f_-(x_i)\varepsilon_i}{f_0^2(x_i)}\frac{1}{h}k(0)\right| \\
&\leq \left[\sup_{x\in\mathcal{N}}E(|\varepsilon||x)\right]\left[\sup_x|d(x)\bar f_-(x)|\right]\left[\sup_{x\in\mathcal{N}}\frac{1}{f_0^2(x_i)}\right]|k(0)|\frac{1}{h}E\mathbf{1}\{\bar x\leq x\leq\bar x+Mh\} \\
&< \infty
\end{aligned}$$

for all n large enough. By a V-statistic projection theorem, as in Newey and McFadden (1994) Lemma 8.4, $\frac{1}{\sqrt{nh}}\sum_i d_i\frac{\bar f_-(x_i)\varepsilon_i}{f_0^2(x_i)}\ (\hat f_0(x_i)-\bar f_0(x_i)) = \sqrt{\frac{n}{h}}\ O_p(\frac{[E(b_n(z_i,z_j)^2)]^{1/2}}{n}+\frac{E|b_n(z_i,z_j)|}{n}) = O_p(\frac{1}{\sqrt{nh}})$ $= o_p(1)$. Next note that

$$E\left[d_i\frac{\bar f_-(x_i)\varepsilon_i}{f_0^2(x_i)}(\bar f_0(x_i)-f_0(x_i))\right]=0$$

and

$$
\begin{aligned}
& V\left(\frac{1}{\sqrt{nh}}\sum_i d_i \frac{\bar{f}_-(x_i)\varepsilon_i}{f_0^2(x_i)}(\bar{f}_0(x_i)-f_0(x_i))\right) \\
= \;& \frac{1}{h}E\left[d_i \frac{\sigma^2(x_i)\bar{f}_-^2(x_i)}{f_0^4(x_i)}(\bar{f}_0(x_i)-f_0(x_i))^2\right] \\
\leq \;& \left[\sup_{x\in\mathcal{N}_0}|\bar{f}_0(x)-f_0(x)|\right]^2 \int_0^M \frac{\sigma^2(\bar{x}+wh)}{f_0^3(\bar{x}+wh)}\left(\int_{-M}^{-w} k(u)f_0(\bar{x}+wh+uh)du\right)^2 dw \\
= \;& O(h^{2s}) = o(1)
\end{aligned}
$$

Then by Chebyshev's Inequality, $\frac{1}{\sqrt{nh}}\sum_i d_i \frac{\bar{f}_-(x_i)\varepsilon_i}{f_0^2(x_i)}(\bar{f}_0(x_i)-f_0(x_i)) = o_p(1)$. Similarly for the remaining part of the expression.

Another application of V-statistic projections and Chebyshev's Inequality leads to

$$
\begin{aligned}
& \frac{1}{\sqrt{nh}}\sum_i \left[\left(d_i\frac{\hat{f}_-(x_i)}{\hat{f}_0(x_i)}-(1-d_i)\frac{\hat{f}_+(x_i)}{\hat{f}_0(x_i)}\right)[\varepsilon_i-(\tilde{m}(x_i)-m(x_i))]\right] \\
& = \sqrt{\frac{n}{h}}E_x\left[\left(d(x)\frac{\bar{f}_-(x)}{f_0^2(x)}-(1-d(x))\frac{\bar{f}_+(x)}{f_0^2(x)}\right)\left[m(x)(\hat{f}_0(x)-\bar{f}_0(x))-(\tilde{r}(x)-\bar{r}(x))\right]\right] \\
& \quad +\frac{1}{\sqrt{nh}}\sum_i\left(d_i\frac{\bar{f}_-(x_i)}{f_0(x_i)}-(1-d_i)\frac{\bar{f}_+(x_i)}{f_0(x_i)}\right)\varepsilon_i + B_n + o_p(1).
\end{aligned}
$$

where

$$
B_n = \sqrt{\frac{n}{h}}E_x\left[\left(d(x)\frac{\bar{f}_-(x)}{f_0^2(x)}-(1-d(x))\frac{\bar{f}_+(x)}{f_0^2(x)}\right)\left[m(x)(\bar{f}_0(x)-f_0(x))-(\bar{r}(x)-r(x))\right]\right].
$$

Here $\hat{f}_0$ and $\tilde{r}$ are treated as fixed functions inside the expectation. The term B_n composes the bias, which will be considered later. Note that $\sqrt{\frac{n}{h}}E_x\big[\big(d(x)\frac{\bar{f}_-(x)}{f_0^2(x)}-(1-d(x))\frac{\bar{f}_+(x)}{f_0^2(x)}\big)m(x)\hat{f}_0(x)\big]$ $= \frac{1}{\sqrt{nh}}\sum_i \delta_{1h}(x_i)$, where $\delta_{1h}^a(z) = E_x\big[\big(d(x)\frac{\bar{f}_-(x)}{f_0^2(x)}\big)m(x)\frac{1}{h}k(\frac{x-z}{h})\big]$, $\delta_{1h}^b(z) = E_x\big[\big((1-d(x))\frac{\bar{f}_+(x)}{f_0^2(x)}\big)$ $m(x)\frac{1}{h}k(\frac{x-z}{h})\big]$, and $\delta_{1h}(z) = \delta_{1h}^a(z)-\delta_{1h}^b(z)$. Similarly, let $\delta_{2h}^a(z) = E_x\big[\big(d(x)\frac{\bar{f}_-(x)}{f_0^2(x)}\big)\frac{1}{h}k(\frac{x-z}{h})\big]$, $\delta_{2h}^b(z) = E_x\big[\big((1-d(x))\frac{\bar{f}_+(x)}{f_0^2(x)}\big)\frac{1}{h}k(\frac{x-z}{h})\big]$, and $\delta_{2h}(z) = \delta_{2h}^a(z)-\delta_{2h}^b(z)$. So,

$$
\begin{aligned}
& \frac{1}{\sqrt{nh}}\sum_i \big[\big(d_i\frac{\hat{f}_-(x_i)}{\hat{f}_0(x_i)}-(1-d_i)\frac{\hat{f}_+(x_i)}{\hat{f}_0(x_i)}\big)\cdot[\varepsilon_i-(\tilde{m}(x_i)-m(x_i))]\big] \\
& = \frac{1}{\sqrt{nh}}\sum_i\big[\delta_{1h}(x_i)-\delta_{2h}(x_i)\tilde{y}_i+\big(d_i\frac{\bar{f}_-(x_i)}{f_0(x_i)}-(1-d_i)\frac{\bar{f}_+(x_i)}{f_0(x)}\big)\varepsilon_i\big] \qquad (16) \\
& \quad -\frac{1}{\sqrt{nh}}\sum_i[E(\delta_{1h}(x_i))-E(\delta_{2h}(x_i)\tilde{y}_i)]+B_n+o_p(1).
\end{aligned}
$$

A limiting distribution for (16) can be derived using the Liapunov Central Limit Theorem (along with the limit for the bias term B_n to be derived subsequently). To apply the Liapunov CLT, it suffices that for some $\zeta > 0$, $\sum_i E\left|\frac{1}{\sqrt{nh}}\left[\delta_{1h}(x_i) - \delta_{2h}(x_i)\tilde{y}_i + \left(d_i\frac{\bar{f}_-(x_i)}{f_0(x)} - (1-d_i)\frac{\bar{f}_+(x_i)}{f_0(x)}\right)\varepsilon_i\right]\right|^{2+\zeta} \longrightarrow 0$. This expression is bounded above by $C\sum_i \left[E\left|\frac{1}{\sqrt{nh}}\delta^a_{1h}(x_i)\right|^{2+\zeta} + \cdots + E\left|\frac{1}{\sqrt{nh}}(1-d_i)\frac{\bar{f}_+(x_i)}{f_0(x)}\varepsilon_i\right|^{2+\zeta}\right]$. Then,

$$
\begin{aligned}
&\sum_i E\left|\frac{1}{\sqrt{nh}}(1-d_i)\frac{\bar{f}_+(x_i)}{f_0(x)}\varepsilon_i\right|^{2+\zeta} \\
&\leq \frac{1}{(nh)^{\zeta/2}}\left[\sup_{x\in\mathcal{N}} E(|\varepsilon|^{2+\zeta}|x)\right]\left[\sup_x |(1-d(x))\bar{f}_+(x)|^{2+\zeta}\right]\left[\sup_{x\in\mathcal{N}}\frac{1}{f_0^{2+\zeta}(x)}\right]\frac{1}{h}E\mathbf{1}\{\bar{x} - Mh \leq x \leq \bar{x}\} \\
&\leq O\left(\frac{1}{(nh)^{\zeta/2}}\right) = o(1).
\end{aligned}
$$

The other terms are bounded similarly, so the Liapunov condition is satisfied.

The asymptotic variance can be derived directly from (16). For instance, since $\frac{1}{\sqrt{h}}E(\delta^a_{1h}(x_i))$ $= O(\sqrt{h})$,

$$
\begin{aligned}
&\frac{1}{h}\mathrm{Var}(\delta^a_{1h}(x_i)) \\
&\quad= \frac{1}{h}E(\delta^a_{1h}(x_i)^2) + o(1) \\
&\quad= \frac{1}{h}\int\left[\int d(x)\frac{\bar{f}_-(x)}{f_0(x)}m(x)\frac{1}{h}k(\frac{x-z}{h})dx\right]^2 f_0(z)dz + o(1) \\
&\quad= \int\left[\int_0^M \frac{m(\bar{x}+uh)}{f_0(\bar{x}+uh)}k(u-w)\left(\int_{-M}^{-u} k(t)f_0(\bar{x}+uh+th)dt\right)du\right]^2 f_0(\bar{x}+wh)dw + o(1) \\
&\quad= m^2(\bar{x})f_0(\bar{x})\int\left[\int_0^M k(u-w)\left(\int_{-M}^{-u} k(t)dt\right)du\right]^2 dw + o(1),
\end{aligned}
$$

where the last equality follows by continuity of m and f_0 near $\bar{x}$, the dominated convergence theorem, f_0 bounded away from zero on $\mathcal{N}$, and the conditions on k in Assumption 1. The variances and covariances of the other terms can be similarly derived. Hence,

$$
\begin{aligned}
&\mathrm{Var}\left(\frac{1}{\sqrt{nh}}\sum_i\left[\delta_{1h}(x_i)-\delta_{2h}(x_i)\tilde{y}_i+\left(d_i\frac{\bar{f}_-(x_i)}{f_0(x_i)}-(1-d_i)\frac{\bar{f}_+(x_i)}{f_0(x)}\right)\varepsilon_i\right]\right)\\
=\;& 2m^2(\bar{x})f_0(\bar{x})\int\left[\int_0^\infty k(u-w)\left(\int_{-\infty}^{-u}k(t)dt\right)du\right]^2dw\\
&+(2m^2(\bar{x})+\sigma^{2+}(\bar{x})+\sigma^{2-}(\bar{x}))f_0(\bar{x})\int\left[\int_0^\infty k(u-w)\left(\int_{-\infty}^{-u}k(t)dt\right)du\right]^2dw\\
&+(\sigma^{2+}(\bar{x})+\sigma^{2-}(\bar{x}))f_0(\bar{x})\int_0^\infty\left(\int_{-\infty}^{-u}k(t)dt\right)^2du\\
&-2m^2(\bar{x})f_0(\bar{x})\int\left[\int_0^\infty k(u-w)\left(\int_{-\infty}^{-u}k(t)dt\right)du\right]\\
&\qquad\cdot\left[\int_{-\infty}^0 k(v-w)\left(\int_{-v}^{\infty}k(z)dz\right)dv\right]dw\\
&-(2m^2(\bar{x})+\sigma^{2+}(\bar{x})+\sigma^{2-}(\bar{x}))f_0(\bar{x})\int\left[\int_0^\infty k(u-w)\left(\int_{-\infty}^{-u}k(t)dt\right)du\right]\\
&\qquad\cdot\left[\int_{-\infty}^0 k(v-w)\left(\int_{-v}^{\infty}k(z)dz\right)dv\right]dw\\
&+4m^2(\bar{x})f_0(\bar{x})\int\left[\int_0^\infty k(u-w)\left(\int_{-\infty}^{-u}k(t)dt\right)du\right]\\
&\qquad\cdot\left[\int_{-\infty}^0 k(v-w)\left(\int_{-v}^{\infty}k(z)dz\right)dv\right]dw\\
&+2(\sigma^{2+}(\bar{x})+\sigma^{2-}(\bar{x}))f_0(\bar{x})\int_0^\infty\left[\int_{-\infty}^0 k(v-w)\left(\int_{-v}^{\infty}k(z)dz\right)dv\right]\left(\int_{-\infty}^{-w}k(t)dt\right)dw\\
&-4m^2(\bar{x})f_0(\bar{x})\int\left[\int_0^\infty k(u-w)\left(\int_{-\infty}^{-u}k(t)dt\right)du\right]^2dw\\
&-2(\sigma^{2+}(\bar{x})+\sigma^{2-}(\bar{x}))f_0(\bar{x})\int_0^\infty\left[\int_0^\infty k(u-w)\left(\int_{-\infty}^{-u}k(t)dt\right)du\right]\left(\int_{-\infty}^{-w}k(z)dz\right)dw\\
&+\;o(1)\\
=\;&(\sigma^{2+}(\bar{x})+\sigma^{2-}(\bar{x}))f_0(\bar{x})\int_0^\infty(K_0(w)+L(w)-L(-w))^2dw+\;o(1)
\end{aligned}
$$

where

$$
K_j(w)=\int_w^\infty k(u)u^j du\quad\text{and}\quad L(w)=\int_{-\infty}^0 K_0(-u)k(u-w)du.
$$

Next consider the bias term, B_n. Under the assumptions of part (a),

$$
\begin{aligned}
&\sqrt{nh}\frac{1}{h}E\left[\frac{d\bar{f}_-(x)}{f_0^2(x)}\big(m(x)\bar{f}_0(x)-\bar{r}(x)\big)\right]-\sqrt{nh}\frac{1}{h}E\left[\frac{(1-d)\bar{f}_+(x)}{f_0^2(x)}\big(m(x)\bar{f}_0(x)-\bar{r}(x)\big)\right]\\
= \quad & -\sqrt{nh}\int_0^M \frac{1}{f_0(\bar{x}+vh)}\left(\int_{-M}^{-v}k(u)f_0(\bar{x}+uh+vh)du\right)\\
&\qquad\qquad \cdot\int_{-M}^{M}k(w)f_0(\bar{x}+vh+wh)[m(\bar{x}+vh+wh)-m(\bar{x}+vh)]dwdv\\
&+\sqrt{nh}\int_{-M}^{0}\frac{1}{f_0(\bar{x}+vh)}\left(\int_{-v}^{M}k(u)f_0(\bar{x}+uh+vh)du\right)\\
&\qquad\qquad \cdot\int_{-M}^{M}k(w)f_0(\bar{x}+vh+wh)[m(\bar{x}+vh+wh)-m(\bar{x}+vh)]dwdv\\
= \quad & -\sqrt{nh}hf_0(\bar{x})\int_0^M\left(\int_{-M}^{-v}k(u)du\right)\Bigg[\left(\int_{-v}^{M}k(w)wdw\right)m'^{+}(\bar{x})\\
&\qquad\qquad +\int_{-M}^{-v}k(w)[m'^{-}(\bar{x})w+(m'^{-}(\bar{x})-m'^{+}(\bar{x}))v]dw\Bigg]dv\\
&+\sqrt{nh}hf_0(\bar{x})\int_{-M}^{0}\left(\int_{-v}^{M}k(u)du\right)\Bigg[\int_{-v}^{M}k(w)[m'^{+}(\bar{x})w+(m'^{+}(\bar{x})-m'^{-}(\bar{x}))v]dw\\
&\qquad\qquad +\left(\int_{-M}^{-v}k(w)wdw\right)m'^{-}(\bar{x})\Bigg]dv\\
&-\sqrt{nh}h^2\int_0^M\Bigg\{\left(\int_{-M}^{-v}k(u)du\right)\Bigg[\int_{-v}^{M}k(w)[f_0(\bar{x})m''^{+}(\bar{x})\left(vw+\frac{w^2}{2}\right)\\
&\qquad +f_0'(\bar{x})m'^{+}(\bar{x})(vw+w^2)]dw+\int_{-M}^{-v}k(w)[f_0(\bar{x})m''^{-}(\bar{x})\left(vw+\frac{w^2}{2}\right)\\
&\qquad +f_0(\bar{x})(m''^{-}(\bar{x})-m''^{+}(\bar{x}))\frac{v^2}{2}+f_0'(\bar{x})m'^{-}(\bar{x})(vw+w^2)\\
&\qquad +f_0'(\bar{x})(m'^{-}(\bar{x})-m'^{+}(\bar{x}))(v^2+vw)]dw\Bigg]\\
&\quad +\left(\int_{-M}^{-v}k(u)udu\right)\Bigg[\left(\int_{-v}^{M}k(w)wdw\right)f_0'(\bar{x})m'^{+}(\bar{x})\\
&\qquad +\int_{-M}^{-v}k(w)[f_0'(\bar{x})m'^{-}(\bar{x})w+f_0'(\bar{x})(m'^{-}(\bar{x})-m'^{+}(\bar{x}))v]dw\Bigg]\Bigg\}dv\\
&+\sqrt{nh}h^2\int_{-M}^{0}\Bigg\{\left(\int_{-v}^{M}k(u)du\right)\Bigg[\int_{-v}^{M}k(w)[f_0(\bar{x})m''^{+}(\bar{x})\left(vw+\frac{w^2}{2}\right)\\
&\qquad +f_0(\bar{x})(m''^{+}(\bar{x})-m''^{-}(\bar{x}))\frac{v^2}{2}+f_0'(\bar{x})m'^{+}(\bar{x})(vw+w^2)\\
&\qquad +f_0'(\bar{x})(m'^{+}(\bar{x})-m'^{-}(\bar{x}))(v^2+vw)]dw\\
&\qquad +\int_{-M}^{-v}k(w)[f_0(\bar{x})m''^{-}(\bar{x})\left(vw+\frac{w^2}{2}\right)+f_0'(\bar{x})m'^{-}(\bar{x})(vw+w^2)]dw\Bigg]\\
&\quad +\left(\int_{-v}^{M}k(u)udu\right)\Bigg[\int_{-v}^{M}k(w)[f_0'(\bar{x})m'^{+}(\bar{x})w+f_0'(\bar{x})(m'^{+}(\bar{x})-m'^{-}(\bar{x}))v]dw\\
&\qquad +\left(\int_{-M}^{-v}k(w)wdw\right)f_0'(\bar{x})m'^{-}(\bar{x})\Bigg]\Bigg\}dv+o(h^2)
\end{aligned}
$$

$$
\begin{aligned}
&= \sqrt{nh}h^2 \int_0^M \Bigg\{ \left(\int_v^M k(u)du\right)\left(\int_v^M k(w)wdw\right) v\big[2f_0(\bar{x})(m''^{-}(\bar{x}) - m''^{+}(\bar{x})) \\
&\qquad +6f_0'(\bar{x})(m'^{-}(\bar{x}) - m'^{+}(\bar{x}))\big] \\
&\qquad + \left[\left(\int_v^M k(u)du\right)\left(\int_0^v k(w)w^2dw\right) - \left(\int_v^M k(u)du\right)^2 v^2\right] \\
&\qquad \cdot\big[f_0(\bar{x})(m''^{-}(\bar{x}) - m''^{+}(\bar{x})) + 2f_0'(\bar{x})(m'^{-}(\bar{x}) - m'^{+}(\bar{x}))\big] \\
&\qquad + \left(\int_v^M k(u)udu\right)^2 \big[2f_0'(\bar{x})(m'^{+}(\bar{x}) - m'^{-}(\bar{x}))\big]\Bigg\} dv + o(h^2)
\end{aligned}
$$

Let $g_t(x) = \sum_{j=1}^t m^{(j)}(x)f_0^{(t-j)}(x)/(j!(t-j)!)$. Under the assumptions of part (b) for s even,

$$
\begin{aligned}
&\sqrt{nh}\frac{1}{h}E\left[\frac{d\bar{f}_-(x)}{f_0^2(x)}\big(m(x)\bar{f}_0(x) - \bar{r}(x)\big)\right] - \sqrt{nh}\frac{1}{h}E\left[\frac{(1-d)\bar{f}_+(x)}{f_0^2(x)}\big(m(x)\bar{f}_0(x) - \bar{r}(x)\big)\right] \\
= \;& -\sqrt{nh}\int_0^M \frac{1}{f_0(\bar{x}+vh)}\left(\int_{-M}^{-v} k(u)f_0(\bar{x}+uh+vh)du\right) \\
&\quad \cdot\Bigg\{h\left[\left(\int k(w)wdw\right) f_0(\bar{x}+vh)m'(\bar{x}+vh)\right] + h^2\Bigg[\left(\int k(w)w^2dw\right) \\
&\quad \cdot g_2(\bar{x}+vh)\Bigg] + \cdots + h^s\left[\left(\int k(w)w^sdw\right) g_s(\bar{x}+vh)\right] \\
&\quad + h^{s+1}\left[\left(\int k(w)w^{s+1}dw\right) g_{s+1}(\bar{x}+vh)\right] + O(h^{s+2})\Bigg\} dv \\
& +\sqrt{nh}\int_{-M}^0 \frac{1}{f_0(\bar{x}+vh)}\left(\int_{-v}^{M} k(u)f_0(\bar{x}+uh+vh)du\right) \\
&\quad \cdot\Bigg\{h\left[\left(\int k(w)wdw\right) f_0(\bar{x}+vh)m'(\bar{x}+vh)\right] + h^2\Bigg[\left(\int k(w)w^2dw\right) \\
&\quad \cdot g_2(\bar{x}+vh)\Bigg] + \cdots + h^s\left[\left(\int k(w)w^sdw\right) g_s(\bar{x}+vh)\right] \\
&\quad + h^{s+1}\left[\left(\int k(w)w^{s+1}dw\right) g_{s+1}(\bar{x}+vh)\right] + O(h^{s+2})\Bigg\} dv \\
= \;& h^s\left(\int k(w)w^sdw\right)\Bigg\{\int_0^M \left(\int_v^M k(u)du\right)[g_s(\bar{x}-vh) - g_s(\bar{x}+vh)]dv \\
&\quad + h\int_0^M \left(\int_v^M k(u)udu\right)\left[\frac{f_0'(\bar{x}+vh)}{f_0(\bar{x}+vh)}g_s(\bar{x}+vh) + \frac{f_0'(\bar{x}-vh)}{f_0(\bar{x}-vh)}g_s(\bar{x}-vh)\right]dv\Bigg\} \\
& + o(h^{s+1}) \\
= \;& h^{s+1}\left(\int k(w)w^sdw\right)\Bigg[2\frac{f_0'(\bar{x})}{f_0(\bar{x})}g_s(\bar{x})\int_0^M \left(\int_v^M k(u)udu\right)dv \\
&\quad -2g_s'(\bar{x})\int_0^M \left(\int_v^M k(u)du\right)dv\Bigg] + o(h^{s+1})
\end{aligned}
$$

The bias under the assumptions of part (b) for s odd is similarly derived.

Finally, for the denominator in (7), expand to linearize.

$$
\begin{aligned}
&\frac{1}{nh}\sum_i\left(d_i\frac{\hat{f}_-^2(x_i)}{\hat{f}_0^2(x_i)}+(1-d_i)\frac{\hat{f}_+^2(x_i)}{\hat{f}_0^2(x_i)}\right)\\
&= \frac{1}{nh}\sum_i\left(d_i\frac{\bar{f}_-^2(x_i)}{f_0^2(x_i)}+(1-d_i)\frac{\bar{f}_+^2(x_i)}{f_0^2(x_i)}\right)\\
&\quad+\frac{1}{nh}\sum_i\left[d_i\frac{\hat{f}_-(x_i)+\bar{f}_-(x_i)}{\hat{f}_0^2(x_i)}(\hat{f}_-(x_i)-\bar{f}_-(x_i))+(1-d_i)\frac{\hat{f}_+(x_i)+\bar{f}_+(x_i)}{\hat{f}_0^2(x_i)}(\hat{f}_+(x_i)-\bar{f}_+(x_i))\right]\\
&\quad-\frac{1}{nh}\sum_i\frac{(\hat{f}_0(x_i)+f_0(x_i))}{f_0^2(x_i)\hat{f}_0^2(x_i)}\left[d_i\bar{f}_-^2(x_i)(\hat{f}_0(x_i)-f_0(x_i))+(1-d_i)\bar{f}_+^2(x_i)(\hat{f}_0(x_i)-f_0(x_i))\right]
\end{aligned}
$$

Use the uniform convergence results in Lemmas B1 and B4 to show the linear terms are $o_p(1)$. Take one such term,

$$
\begin{aligned}
&\left|\frac{1}{nh}\sum_i d_i\frac{\hat{f}_-(x_i)+\bar{f}_-(x_i)}{\hat{f}_0^2(x_i)}(\hat{f}_-(x_i)-\bar{f}_-(x_i))\right|\\
&\le \left[\sup_{x\in\mathcal{N}_0}\frac{1}{\hat{f}_0^2(x)}\right]\left[\sup_{x\in\mathcal{N}_0} d(x)|\hat{f}_-(x)-\bar{f}_-(x)|+2\sup_x|d(x)\bar{f}_-(x)|\right]\left[\sup_{x\in\mathcal{N}_0} d(x)|\hat{f}_-(x)-\bar{f}_-(x)|\right]\\
&\quad\cdot\frac{1}{nh}\sum_i \mathbf{1}\{\bar{x}\le x_i\le \bar{x}+Mh\}\\
&= O_p\left(\sqrt{\frac{\ln n}{nh}}\right)=o_p(1).
\end{aligned}
$$

The other linear terms are shown $o_p(1)$ similarly.

By Assumptions 1 and 2,

$$
\begin{aligned}
&\mathrm{Var}\left(\frac{1}{nh}\sum_i\left[d_i\frac{\bar{f}_-^2(x_i)}{f_0^2(x_i)}+(1-d_i)\frac{\bar{f}_+^2(x_i)}{f_0^2(x_i)}\right]\right)\\
&\le \frac{1}{nh^2}E\left[d_i\frac{\bar{f}_-^4(x_i)}{f_0^4(x_i)}+(1-d_i)\frac{\bar{f}_+^4(x_i)}{f_0^4(x_i)}\right]\\
&= \frac{1}{nh}\int_0^M\frac{1}{f_0^3(\bar{x}+wh)}\left(\int_{-M}^{-w}k(u)f_0(\bar{x}+wh+uh)du\right)^4dw\\
&\quad+\frac{1}{nh}\int_{-M}^0\frac{1}{f_0^3(\bar{x}+wh)}\left(\int_{-w}^{M}k(u)f_0(\bar{x}+wh+uh)du\right)^4dw\\
&= O\left(\frac{1}{nh}\right)=o(1).
\end{aligned}
$$

Then by Chebyshev's Inequality,

$$
\begin{aligned}
\frac{1}{nh}\sum_i\left(d_i\frac{\hat{f}_-^2(x_i)}{\hat{f}_0^2(x_i)}+(1-d_i)\frac{\hat{f}_+^2(x_i)}{\hat{f}_0^2(x_i)}\right) &= \frac{1}{h}E\left[d_i\frac{\bar{f}_-^2(x_i)}{f_0^2(x_i)}+(1-d_i)\frac{\bar{f}_+^2(x_i)}{f_0^2(x_i)}\right]+o_p(1)\\
&= 2f_0(\bar{x})\int_0^M K_0^2(w)dw+o_p(1)
\end{aligned}
$$

using the dominated convergence theorem. The theorem follows. $\square$

PROOF OF THEOREM 3:

Begin by noting

$$
\begin{aligned}
\sqrt{nh}(\bar{\alpha}_+ - \bar{\alpha}_- - \alpha) &= \sqrt{nh}[(\bar{\alpha}_+ - (m(\bar{x}) + \alpha)) - (\bar{\alpha}_- - m(\bar{x}))] \\
&= e_1{}' D_{n+} \left\{ \frac{1}{\sqrt{nh}} \sum_i \left(k\left(\frac{x_i - \bar{x}}{h}\right) d_i Z_i y_i^+ - E\left[k\left(\frac{x_i - \bar{x}}{h}\right) d_i Z_i y_i^+ | x_i \right] \right) \right\} && (17) \\
&+ e_1{}' D_{n+} \sqrt{nh} \left\{ \frac{1}{n} \sum_i \frac{1}{h} E\left[k\left(\frac{x_i - \bar{x}}{h}\right) d_i Z_i y_i^+ | x_i \right] - E\left[\frac{1}{n} \sum_i \frac{1}{h} k\left(\frac{x_i - \bar{x}}{h}\right) d_i Z_i y_i^+ \right] \right\} && (18) \\
&+ e_1{}' D_{n+} \sqrt{nh} \left\{ E\left[\frac{1}{n} \sum_i \frac{1}{h} k\left(\frac{x_i - \bar{x}}{h}\right) d_i Z_i y_i^+ \right] - h^{p+1} B_{n+} \right\} && (19) \\
&- e_1{}' D_{n-} \left\{ \frac{1}{\sqrt{nh}} \sum_i \left(k\left(\frac{x_i - \bar{x}}{h}\right) (1 - d_i) Z_i y_i^- - E\left[k\left(\frac{x_i - \bar{x}}{h}\right) (1 - d_i) Z_i y_i^- | x_i \right] \right) \right\} && (20) \\
&- e_1{}' D_{n-} \sqrt{nh} \Big\{ \frac{1}{n} \sum_i \frac{1}{h} E\left[k\left(\frac{x_i - \bar{x}}{h}\right) (1 - d_i) Z_i y_i^- | x_i \right] \\
&\qquad - E\left[\frac{1}{n} \sum_i \frac{1}{h} k\left(\frac{x_i - \bar{x}}{h}\right) (1 - d_i) Z_i y_i^- \right] \Big\} && (21) \\
&- e_1{}' D_{n-} \sqrt{nh} \left\{ E\left[\frac{1}{n} \sum_i \frac{1}{h} k\left(\frac{x_i - \bar{x}}{h}\right) (1 - d_i) Z_i y_i^- \right] - h^{p+1} B_{n-} \right\} && (22) \\
&+ e_1{}' D_{n+} \sqrt{nh} h^{p+1} B_{n+} - e_1{}' D_{n-} \sqrt{nh} h^{p+1} B_{n-} && (23)
\end{aligned}
$$

where

$$
\begin{aligned}
e_1 &= (1, 0, \ldots, 0)', \qquad Z_i = \left(1, \frac{x_i - \bar{x}}{h}, \ldots, \left(\frac{x_i - \bar{x}}{h}\right)^p\right)' \\
D_{n+} &= \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{h} k\left(\frac{x_i - \bar{x}}{h}\right) d_i Z_i Z_i{}' \right]^{-1}, \qquad D_{n-} = \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{h} k\left(\frac{x_i - \bar{x}}{h}\right) (1 - d_i) Z_i Z_i{}' \right]^{-1} \\
y_i^+ &= m(x_i) - m(\bar{x}) - m'^+(\bar{x})(x_i - \bar{x}) - \cdots - \frac{1}{p!} m^{(p)+}(\bar{x})(x_i - \bar{x})^p + \varepsilon_i \\
y_i^- &= m(x_i) - m(\bar{x}) - m'^-(\bar{x})(x_i - \bar{x}) - \cdots - \frac{1}{p!} m^{(p)-}(\bar{x})(x_i - \bar{x})^p + \varepsilon_i
\end{aligned}
$$

We will proceed by applying a CLT to terms (17) and (20). Term (23) is the bias term, and its limit depends on whether the conditions of part (a) or (b) are assumed to hold. The remaining terms (18), (19), (21), and (22) are shown to be asymptotically negligible. The proof for each term will again differ under the assumptions of part (a) or part (b). The terms B_{n+} and B_{n-} will be defined below (and will be defined differently based on whether we are assuming the conditions of part (a) or (b) hold).[13] They are sucscripted by n to allow for (non-stochastic) sequences, which is useful when part (b) conditions are assumed.

First, we establish the probability limit for the denominator terms D_{n+} and D_{n-}. Show $D_{n+} \xrightarrow{p} \frac{1}{f_0(\bar{x})}\Gamma_+^{-1}$, where Γ_+ is the $(p+1)\times(p+1)$ matrix with (j,l) element Γ_{j+l-2} for $j,l = 1,\ldots,p+1$. Similarly, $[\Gamma_-]_{j,l} = (-1)^{j+l}\Gamma_{j+l-2}$. Define $\Gamma_j = \int_0^\infty k(u)u^j du$. By the dominated convergence theorem, $E[n^{-1}\sum_i k_h(x_i-\bar{x})d_i\left(\frac{x_i-\bar{x}}{h}\right)^l] = \int_0^M k(u)u^l f_0(\bar{x}+uh)du \longrightarrow f_0(x)\int_0^M k(u)u^l du$. Also, $V(n^{-1}\sum_i k_h(x_i-\bar{x})d_i\left(\frac{x_i-\bar{x}}{h}\right)^l) \leq \frac{1}{nh}\int_0^M k^2(u)u^{2l}f_0(\bar{x}+uh)du \longrightarrow 0$ by the dominated convergence theorem. Continuity of matrix inversion completes the argument. Similarly $D_{n-} \xrightarrow{p} \frac{1}{f_0(\bar{x})}\Gamma_-^{-1}$.

Terms (17) and (20):

Show $\frac{1}{\sqrt{nh}}\sum_i\left(k\left(\frac{x_i-\bar{x}}{h}\right)d_i Z_i y_i^+ - E\left[k\left(\frac{x_i-\bar{x}}{h}\right)d_i Z_i y_i^+|x_i\right]\right) \xrightarrow{d} N(0, f_0(\bar{x})\sigma^{2+}(\bar{x})\Delta_+)$, where $[\Delta_+]_{j,l} = \delta_{j+l-2}$ and $\delta_j = \int_0^\infty k^2(u)u^j du$. The Cramer-Wold device will be applied to derive the asymptotic normality, so let λ be a nonzero, finite vector and define $U_{ni} = k\left(\frac{x_i-\bar{x}}{h}\right)d_i\lambda' Z_i\varepsilon_i/\sqrt{nh}$. $E[U_{ni}] = 0$. $h^{-1}E[k^2\left(\frac{x-\bar{x}}{h}\right)d\varepsilon^2\left(\frac{x-\bar{x}}{h}\right)^j] \longrightarrow f_0(\bar{x})\sigma^{2+}(\bar{x})\delta_j$ by the dominated convergence theorem, so $\sum_j V(U_{nj}) \longrightarrow f_0(\bar{x})\sigma^{2+}(\bar{x})\lambda'\Delta\lambda$. For $\zeta > 0$ and constants $C, C' > 0$,

$$\begin{aligned}
\sum_i E|U_{ni}|^{2+\zeta} &= \sum_i \left(\frac{1}{nh}\right)^{\zeta/2}\frac{1}{nh}E\left[\left|k\left(\frac{x_i-\bar{x}}{h}\right)\right|^{2+\zeta} d_i|\lambda' Z_i|^{2+\zeta}|\varepsilon_i|^{2+\zeta}\right] \\
&\leq C\left(\frac{1}{nh}\right)^{\zeta/2}\frac{1}{h}E\left[\left|k\left(\frac{x-\bar{x}}{h}\right)\right|^{2+\zeta} dE\left(|\varepsilon|^{2+\zeta}|x\right)\sum_{l=0}^p\left|\lambda_l\left(\frac{x-\bar{x}}{h}\right)^l\right|^{2+\zeta}\right] \\
&\leq C'\left(\frac{1}{nh}\right)^{\zeta/2}\left[\sup_{x\in\mathcal{N}}E\left(|\varepsilon|^{2+\zeta}|x\right)\right]\int_0^M |k(u)|^{2+\zeta} f_0(\bar{x}+uh)\sum_{l=0}^p\left|\lambda_l u^l\right|^{2+\zeta}du \\
&= O\left(\frac{1}{(nh)^{\zeta/2}}\right) = o(1)
\end{aligned}$$

Application of Liapunov's CLT and the Cramer-Wold device completes the argument. The limit distribution of $\frac{1}{\sqrt{nh}}\sum_i\left(k\left(\frac{x_i-\bar{x}}{h}\right)(1-d_i)Z_i y_i^- - E\left[k\left(\frac{x_i-\bar{x}}{h}\right)(1-d_i)Z_i y_i^-|x_i\right]\right)$ follows similarly.

Define Δ_- by $[\Delta_-]_{j,l} = (-1)^{j+l}\delta_{j+l-2}$ for $j,l = 1,\ldots,p+1$ and note that $e_1'\Gamma_-^{-1}\Delta_-\Gamma_-^{-1}e_1 = e_1'\Gamma_+^{-1}\Delta_+\Gamma_+^{-1}e_1$. Then,

$$(17) + (20) \xrightarrow{d} N\left(0, \frac{\sigma^{2+}(\bar{x}) + \sigma^{2-}(\bar{x})}{f_0(\bar{x})}e_1'\Gamma_+^{-1}\Delta_+\Gamma_+^{-1}e_1\right)$$

[13] Actually, the definitions of B_{n+} and B_{n-} under part (b) could be used under part (a) as well, but the higher order terms would simply drop out asymptotically.

Terms (18) and (21) (under part (a)):

Show $n^{-1}\sum_i E\left[k_h(x_i-\bar{x})d_iZ_iy_i^+|x_i\right] = E\left[n^{-1}\sum_i k_h(x_i-\bar{x})d_iZ_iy_i^+\right] + o_p(h^{p+1})$. Let $\mu_{j,+}(x) = m(x) - [m(\bar{x}) + m'^+(\bar{x})(x-\bar{x}) + \cdots + m^{(j)+}(\bar{x})(x-\bar{x})^j/(j!)]$. For $l=0,\ldots,p$,

$$\begin{aligned}
& V\left(\frac{1}{nh^{p+1}}\sum_i E\left[\frac{1}{h}k\left(\frac{x_i-\bar{x}}{h}\right)d_i\left(\frac{x_i-\bar{x}}{h}\right)^l y_i^+|x_i\right]\right) \\
\leq\ & C\frac{1}{nh^{2(p+1)}}E\left[\frac{1}{h^2}k^2\left(\frac{x-\bar{x}}{h}\right)d\left(\frac{x-\bar{x}}{h}\right)^{2l}\left(\mu_{p+1,+}^2(x)+\left(\frac{m^{(p+1)+}(\bar{x})}{(p+1)!}\right)^2(x-\bar{x})^{2(p+1)}\right)\right] \\
\leq\ & C\frac{1}{nh^{2(p+1)}}\left[\sup_{x\in\mathcal{N}}\mu_{p+1,+}^2(x)\right]E\left[\frac{1}{h^2}k^2\left(\frac{x-\bar{x}}{h}\right)d\left(\frac{x-\bar{x}}{h}\right)^{2l}\right] \\
& +C\left(\frac{m^{(p+1)+}(\bar{x})}{(p+1)!}\right)^2\frac{1}{n}E\left[\frac{1}{h^2}k^2\left(\frac{x-\bar{x}}{h}\right)d\left(\frac{x-\bar{x}}{h}\right)^{2(l+p+1)}\right] \\
\leq\ & C\frac{1}{nh}\left[\frac{1}{h^{p+1}}\sup_{x\in\mathcal{N}}|\mu_{p+1,+}(x)|\right]^2\int_0^M k^2(u)u^{2l}f_0(\bar{x}+uh)du \\
& +C\left(\frac{m^{(p+1)+}(\bar{x})}{(p+1)!}\right)^2\frac{1}{nh}\int_0^M k^2(u)u^{2(l+p+1)}f_0(\bar{x}+uh)du \\
=\ & o\left(\frac{1}{nh}\right)+O\left(\frac{1}{nh}\right) \qquad \text{by the dominated convergence theorem} \\
=\ & o(1)
\end{aligned}$$

The result follows by Chebyshev's Inequality. The analogous result for (21) holds similarly, so that (18) $+$(21) $= o_p(\sqrt{nh}\ h^{p+1})$.

Terms (19) and (22) (under part (a)):

Show $h^{-(p+1)}E\left[n^{-1}\sum_i k_h\left(x_i-\bar{x}\right)d_iZ_iy_i^+\right] - B_{n+} = o(1)$, where $B_{n+} = \frac{m^{(p+1)+}(\bar{x})}{(p+1)!}f_0(\bar{x})\left(\Gamma_{p+1},\ldots,\Gamma_{2p+1}\right)'$. For $l=0,\ldots,p$,

$$
\begin{aligned}
&\left| \frac{1}{h^{p+1}} E\left[\frac{1}{n} \sum_i \frac{1}{h} k\left(\frac{x_i - \bar{x}}{h} \right) d_i \left(\frac{x_i - \bar{x}}{h} \right)^l y_i^+ \right] - \frac{m^{(p+1)+}(\bar{x})}{(p+1)!} f_0(\bar{x}) \Gamma_{l+p+1} \right| \\
&\leq \left| E\left[\frac{1}{h} k\left(\frac{x - \bar{x}}{h} \right) d \left(\frac{x - \bar{x}}{h} \right)^{l+p+1} \frac{m^{(p+1)+}(\bar{x})}{(p+1)!} \right] - \frac{m^{(p+1)+}(\bar{x})}{(p+1)!} f_0(\bar{x}) \Gamma_{l+p+1} \right| \\
&\quad + \left| \frac{1}{h^{p+1}} E\left[\frac{1}{h} k\left(\frac{x - \bar{x}}{h} \right) d \left(\frac{x - \bar{x}}{h} \right)^l \mu_{p+1,+}(x) \right] \right| \\
&\leq \left| \frac{m^{(p+1)+}(\bar{x})}{(p+1)!} \int_0^M k(u) u^{l+p+1} f_0(\bar{x} + uh) du - \frac{m^{(p+1)+}(\bar{x})}{(p+1)!} f_0(\bar{x}) \Gamma_{l+p+1} \right| \\
&\quad + \frac{1}{h^{p+1}} \left[\sup_{x \in \mathcal{N}} |\mu_{p+1,+}(x)| \right] \int_0^M |k(u)| u^l f_0(\bar{x} + uh) du \\
&= o(1) \quad \text{by the dominated convergence theorem}
\end{aligned}
$$

The analogous result for (22) holds similarly.

Assume the conditions of part (b) also hold. Terms (17) and (20) converge in distribution as above.

Terms (18) and (21) (under part (b)):

Show $n^{-1} \sum_i E\left[k_h(x_i - \bar{x}) d_i Z_i y_i^+ | x_i \right] = E\left[n^{-1} \sum_i k_h(x_i - \bar{x}) d_i Z_i y_i^+ \right] + o_p(h^{p+2})$. Let $\mu_j(x) = m(x) - [m(\bar{x}) + m'(\bar{x})(x - \bar{x}) + \cdots + m^{(j)}(\bar{x})(x - \bar{x})^j/(j!)]$. For $l = 0, \ldots, p$,

$$
\begin{aligned}
&V\left(\frac{1}{nh^{p+2}} \sum_i E\left[\frac{1}{h} k\left(\frac{x_i - \bar{x}}{h} \right) d_i \left(\frac{x_i - \bar{x}}{h} \right)^l y_i^+ | x_i \right] \right) \\
&\leq C \frac{1}{nh^{2(p+2)}} E\Bigg[\frac{1}{h^2} k^2 \left(\frac{x - \bar{x}}{h} \right) d \left(\frac{x - \bar{x}}{h} \right)^{2l} \\
&\qquad \cdot \left(\mu_{p+2}^2(x) + \left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!} \right)^2 (x - \bar{x})^{2p+2} + \left(\frac{m^{(p+2)}(\bar{x})}{(p+2)!} \right)^2 (x - \bar{x})^{2p+4} \right) \Bigg] \\
&\leq C \frac{1}{nh} \left[\frac{1}{h^{p+2}} \sup_{x \in \mathcal{N}} \mu_{p+2}(x) \right]^2 E\left[\frac{1}{h} k^2 \left(\frac{x - \bar{x}}{h} \right) d \left(\frac{x - \bar{x}}{h} \right)^{2l} \right] \\
&\quad + C \frac{1}{nh^3} \left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!} \right)^2 E\left[\frac{1}{h} k^2 \left(\frac{x - \bar{x}}{h} \right) d \left(\frac{x - \bar{x}}{h} \right)^{2(l+p+1)} \right] \\
&\quad + C \frac{1}{nh} \left(\frac{m^{(p+2)}(\bar{x})}{(p+2)!} \right)^2 E\left[\frac{1}{h} k^2 \left(\frac{x - \bar{x}}{h} \right) d \left(\frac{x - \bar{x}}{h} \right)^{2(l+p+2)} \right]
\end{aligned}
$$

$$
\begin{aligned}
&= o\left(\frac{1}{nh}\right) + O\left(\frac{1}{nh^3}\right) + O\left(\frac{1}{nh}\right) \quad \text{by the dominated convergence theorem} \\
&= o(1)
\end{aligned}
$$

Apply Chebyshev's Inequality and proceed as above.

Terms (19) and (22) (under part (b)):

Show $h^{-(p+1)}E\left[n^{-1}\sum_i k_h(x_i-\bar{x})\,d_i Z_i y_i^+\right] - B_{n+} = o(h)$, where $B_{n+} = \frac{m^{(p+1)}(\bar{x})}{(p+1)!} f_0(\bar{x})\,(\Gamma_{p+1},\ldots,\Gamma_{2p+1})'$ $+h(\frac{m^{(p+1)}(\bar{x})}{(p+1)!}f'(\bar{x}) + \frac{m^{(p+2)}(\bar{x})}{(p+2)!}f_0(\bar{x}))\ (\Gamma_{p+2},\ldots,\Gamma_{2p+2})'$. For $l = 0,\ldots,p$,

$$
\begin{aligned}
&\left| \frac{1}{h^{p+1}} E\left[\frac{1}{n}\sum_i \frac{1}{h} k\left(\frac{x_i-\bar{x}}{h}\right) d_i \left(\frac{x_i-\bar{x}}{h}\right)^l y_i^+\right] - \frac{m^{(p+1)}(\bar{x})}{(p+1)!} f_0(\bar{x})\Gamma_{l+p+1} \right.\\
&\qquad \left. -h\left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!}f'(\bar{x}) + \frac{m^{(p+2)}(\bar{x})}{(p+2)!}f_0(\bar{x})\right)\Gamma_{l+p+2}\right| \\
&\leq \left| E\left[\frac{1}{h}k\left(\frac{x-\bar{x}}{h}\right) d \left\{\left(\frac{x-\bar{x}}{h}\right)^{l+p+1}\frac{m^{(p+1)}(\bar{x})}{(p+1)!} + h\left(\frac{x-\bar{x}}{h}\right)^{l+p+2}\frac{m^{(p+2)}(\bar{x})}{(p+2)!}\right\}\right]\right. \\
&\qquad \left. -\frac{m^{(p+1)}(\bar{x})}{(p+1)!}f_0(\bar{x})\Gamma_{l+p+1} - h\left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!}f'(\bar{x}) + \frac{m^{(p+2)}(\bar{x})}{(p+2)!}f_0(\bar{x})\right)\Gamma_{l+p+2}\right| \\
&\qquad + \left|\frac{1}{h^{p+1}}E\left[\frac{1}{h}k\left(\frac{x-\bar{x}}{h}\right) d\left(\frac{x-\bar{x}}{h}\right)^l \mu_{p+2}(x)\right]\right| \\
&\leq \left|\frac{m^{(p+1)+}(\bar{x})}{(p+1)!}\right| \left|\int_0^M k(u)u^{l+p+1}[f_0(\bar{x}) + f'(\tilde{x}_u)uh]du - f_0(\bar{x})\Gamma_{l+p+1} - hf'(\bar{x})\Gamma_{l+p+2}\right| \\
&\qquad + h\left|\frac{m^{(p+2)+}(\bar{x})}{(p+2)!}\right| \left|\int_0^M k(u)u^{l+p+2}f_0(\bar{x}+uh)du - f_0(\bar{x})\Gamma_{l+p+2}\right| \\
&\qquad + h\frac{1}{h^{p+2}}\left[\sup_{x\in\mathcal{N}}|\mu_{p+2}(x)|\right]\int_0^M |k(u)|u^l f_0(\bar{x}+uh)du \\
&= o(h) \quad \text{by the dominated convergence theorem}
\end{aligned}
$$

where $\tilde{x}_u$ is a mean-value between $\bar{x}$ and $\bar{x}+uh$. The analogous result for (22) holds similarly with

$$
\begin{aligned}
B_{n-} &= \frac{m^{(p+1)}(\bar{x})}{(p+1)!}f_0(\bar{x})((-1)^{p+1}\Gamma_{p+1},\ldots,(-1)^{2p+1}\Gamma_{2p+1})' \\
&\quad + h(\frac{m^{(p+1)}(\bar{x})}{(p+1)!}f'(\bar{x}) + \frac{m^{(p+2)}(\bar{x})}{(p+2)!}f_0(\bar{x}))((-1)^{p+2}\Gamma_{p+2},\ldots,(-1)^{2p+2}\Gamma_{2p+2})'.
\end{aligned}
$$

Term (23) (under part (b)):

Show $D_{n+} - (\frac{1}{f_0(\bar{x})}\Gamma_+^{-1} + hG_+) = o_p(h)$, where

$$\begin{aligned}
G_+ &= -\frac{f'(\bar{x})}{f^2(\bar{x})}\Gamma_+^{-1}\begin{bmatrix} \Gamma_1 & \cdots & \Gamma_{p+1} \\ \vdots & & \vdots \\ \Gamma_{p+1} & \cdots & \Gamma_{2p+1} \end{bmatrix}\Gamma_+^{-1} \quad \text{and} \\
G_- &= -\frac{f'(\bar{x})}{f^2(\bar{x})}\Gamma_-^{-1}\begin{bmatrix} -\Gamma_1 & \cdots & (-1)^{p+1}\Gamma_{p+1} \\ \vdots & & \vdots \\ (-1)^{p+1}\Gamma_{p+1} & \cdots & (-1)^{2p+1}\Gamma_{2p+1} \end{bmatrix}\Gamma_-^{-1}.
\end{aligned}$$

For $j = 0, \ldots, 2p$,

$$\begin{aligned}
V\left(\frac{1}{nh}\sum_i \frac{1}{h}k\left(\frac{x_i-\bar{x}}{h}\right)d_i\left(\frac{x_i-\bar{x}}{h}\right)^j\right) &\leq \frac{1}{nh^3}E\left[\frac{1}{h}k^2\left(\frac{x-\bar{x}}{h}\right)d\left(\frac{x-\bar{x}}{h}\right)^{2j}\right] \\
&= o(1)\left[f_0(\bar{x})\int_0^M k^2(u)u^{2j}du + o(1)\right] = o(1)
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{h}\left|E\left[\frac{1}{h}k\left(\frac{x-\bar{x}}{h}\right)d\left(\frac{x-\bar{x}}{h}\right)^j\right] - f_0(\bar{x})\Gamma_j - hf'(\bar{x})\Gamma_{j+1}\right| \\
&= \frac{1}{h}\left|\int_0^M k(u)u^j(f_0(\bar{x}+uh) - f_0(\bar{x}))du - hf'(\bar{x})\Gamma_{j+1}\right| \\
&= \left|\int_0^M k(u)u^{j+1}(f'(\tilde{x}_u) - f'(\bar{x}))du\right| \\
&= o(1)
\end{aligned}$$

Then Lemma B5 gives the result and similarly the corresponding result, $D_{n-} - (\frac{1}{f_0(\bar{x})}\Gamma_-^{-1} + hG_-)$ $= o_p(h)$.

Finally, note

$$\begin{aligned}
&{e_1}'D_{n+}\frac{1}{h}B_{n+} - {e_1}'D_{n-}\frac{1}{h}B_{n-} \\
&= {e_1}'D_{n+}\frac{1}{h}\left[\frac{m^{(p+1)}(\bar{x})}{(p+1)!}f_0(\bar{x})\begin{pmatrix}\Gamma_{p+1} \\ \vdots \\ \Gamma_{2p+1}\end{pmatrix} + h\left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!}f'(\bar{x}) + \frac{m^{(p+2)}(\bar{x})}{(p+2)!}f_0(\bar{x})\right)\begin{pmatrix}\Gamma_{p+2} \\ \vdots \\ \Gamma_{2p+2}\end{pmatrix}\right] \\
&\quad -{e_1}'D_{n-}\frac{1}{h}\left[\frac{m^{(p+1)}(\bar{x})}{(p+1)!}f_0(\bar{x})\begin{pmatrix}(-1)^{p+1}\Gamma_{p+1} \\ \vdots \\ (-1)^{2p+1}\Gamma_{2p+1}\end{pmatrix}\right. \\
&\qquad \left. +h\left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!}f'(\bar{x}) + \frac{m^{(p+2)}(\bar{x})}{(p+2)!}f_0(\bar{x})\right)\begin{pmatrix}(-1)^{p+2}\Gamma_{p+2} \\ \vdots \\ (-1)^{2p+2}\Gamma_{2p+2}\end{pmatrix}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{h}\frac{m^{(p+1)}(\bar{x})}{(p+1)!}\left(e_1{}'\Gamma_+^{-1}\begin{pmatrix}\Gamma_{p+1}\\ \vdots \\ \Gamma_{2p+1}\end{pmatrix} - e_1{}'\Gamma_-^{-1}\begin{pmatrix}(-1)^{p+1}\Gamma_{p+1}\\ \vdots \\ (-1)^{2p+1}\Gamma_{2p+1}\end{pmatrix}\right)\\
&\quad + \left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!}\frac{f'(\bar{x})}{f_0(\bar{x})} + \frac{m^{(p+2)}(\bar{x})}{(p+2)!}\right)\left(e_1{}'\Gamma_+^{-1}\begin{pmatrix}\Gamma_{p+2}\\ \vdots \\ \Gamma_{2p+2}\end{pmatrix} - e_1{}'\Gamma_-^{-1}\begin{pmatrix}(-1)^{p+2}\Gamma_{p+2}\\ \vdots \\ (-1)^{2p+2}\Gamma_{2p+2}\end{pmatrix}\right)\\
&\quad + \frac{m^{(p+1)}(\bar{x})}{(p+1)!}f_0(\bar{x})\left(e_1{}'G_+\begin{pmatrix}\Gamma_{p+1}\\ \vdots \\ \Gamma_{2p+1}\end{pmatrix} - e_1{}'G_-\begin{pmatrix}(-1)^{p+1}\Gamma_{p+1}\\ \vdots \\ (-1)^{2p+1}\Gamma_{2p+1}\end{pmatrix}\right)\\
&\quad + h\left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!}f'(\bar{x}) + \frac{m^{(p+2)}(\bar{x})}{(p+2)!}f_0(\bar{x})\right)\left(e_1{}'G_+\begin{pmatrix}\Gamma_{p+2}\\ \vdots \\ \Gamma_{2p+2}\end{pmatrix} - e_1{}'G_-\begin{pmatrix}(-1)^{p+2}\Gamma_{p+2}\\ \vdots \\ (-1)^{2p+2}\Gamma_{2p+2}\end{pmatrix}\right)\\
&= 2\left(\frac{m^{(p+1)}(\bar{x})}{(p+1)!}\frac{f'(\bar{x})}{f_0(\bar{x})} + \frac{m^{(p+2)}(\bar{x})}{(p+2)!}\right)e_1{}'\Gamma_+^{-1}\begin{pmatrix}\Gamma_{p+2}\\ \vdots \\ \Gamma_{2p+2}\end{pmatrix}\\
&\quad + 2\frac{m^{(p+1)}(\bar{x})}{(p+1)!}f_0(\bar{x})e_1{}'G_+\begin{pmatrix}\Gamma_{p+1}\\ \vdots \\ \Gamma_{2p+1}\end{pmatrix} + o(1)
\end{aligned}
$$

which complees the proof. □

PROOF OF THEOREM 4:

$$
\check{\sigma}^{2+}(\bar{x}) = \frac{\frac{1}{n}\sum_i \frac{1}{h}k\left(\frac{\bar{x}-x_i}{h}\right)d_i\check{\varepsilon}_i^2}{\frac{1}{2}\hat{f}_0(\bar{x})}.
$$

Consider the numerator,

$$
\begin{aligned}
&\frac{1}{n}\sum_i \frac{1}{h}k\left(\frac{\bar{x}-x_i}{h}\right)d_i\check{\varepsilon}_i^2\\
&\quad = \frac{1}{n}\sum_i \frac{1}{h}k\left(\frac{\bar{x}-x_i}{h}\right)d_i\varepsilon_i^2\\
&\qquad + \frac{1}{n}\sum_i \frac{1}{h}k\left(\frac{\bar{x}-x_i}{h}\right)d_i\left[\left(m(x_i) - \frac{\tilde{r}(x_i)}{\hat{f}_0(x_i)}\right)^2 + \frac{\hat{f}_-^2(x_i)}{\hat{f}_0^2(x_i)}(\check{\alpha}-\alpha)^2\right] \qquad (24)\\
&\qquad + 2\frac{1}{n}\sum_i \frac{1}{h}k\left(\frac{\bar{x}-x_i}{h}\right)d_i\varepsilon_i\left[\left(m(x_i) - \frac{\tilde{r}(x_i)}{\hat{f}_0(x_i)}\right) - \frac{\hat{f}_-(x_i)}{\hat{f}_0(x_i)}(\check{\alpha}-\alpha)\right]\\
&\qquad - 2\frac{1}{n}\sum_i \frac{1}{h}k\left(\frac{\bar{x}-x_i}{h}\right)d_i\frac{\hat{f}_-(x_i)}{\hat{f}_0(x_i)}(\check{\alpha}-\alpha)\left(m(x_i) - \frac{\tilde{r}(x_i)}{\hat{f}_0(x_i)}\right)
\end{aligned}
$$

We show the last three terms are $o_p(1)$. By Lemmas B1, B2, B3, and B4 (with $s = 1$),

$$\begin{aligned}
\sup_{x\in\mathcal{N}_0} |d\hat{f}_-(x)| &\leq \sup_{x\in\mathcal{N}_0} d|\hat{f}_-(x) - E\hat{f}_-(x)| + \sup_{x\in\mathcal{N}_0} |d\bar{f}_-(x)| = O_p(1) \\
\sup_{x\in\mathcal{N}_0} |\hat{f}_0(x) - f_0(x)| &\leq \sup_{x\in\mathcal{N}_0} |\hat{f}_0(x) - \bar{f}_0(x)| + \sup_{x\in\mathcal{N}_0} |\bar{f}_0(x) - f_0(x)| = o_p(1) \\
\sup_{\bar{x}\leq x\leq \bar{x}+Mh} |\tilde{r}(x) - r(x)| &\leq \sup_{x\in\mathcal{N}_0} |\tilde{r}(x) - \bar{r}(x)| + \sup_{\bar{x}\leq x\leq \bar{x}+Mh} |\bar{r}(x) - r(x)| = o_p(1) \\
\sup_{\bar{x}\leq x\leq \bar{x}+Mh} \left| m(x) - \frac{\tilde{r}(x)}{\hat{f}_0(x)} \right| &\leq \sup_{x\in\mathcal{N}_0} \left| \frac{m(x)}{\hat{f}_0(x)} \right| \cdot \sup_{x\in\mathcal{N}_0} |\hat{f}_0(x) - f_0(x)| \\
&\quad + \sup_{x\in\mathcal{N}_0} \left| \frac{1}{\hat{f}_0(x)} \right| \cdot \sup_{\bar{x}\leq x\leq \bar{x}+Mh} |\tilde{r}(x) - r(x)| = o_p(1).
\end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{n} \sum_i \left| \frac{1}{h} k\left(\frac{\bar{x} - x_i}{h} \right) d_i \varepsilon_i \right| \right) &\leq \frac{1}{nh^2} E\left[k^2\left(\frac{\bar{x} - x}{h} \right) d\sigma^2(x) \right] \\
&= \frac{1}{nh} \int_0^M k^2(u)\sigma(\bar{x} + uh) f_0(\bar{x} + uh) du \\
&= o(1).
\end{aligned}$$

Similarly, $\mathrm{Var}(n^{-1} \sum_i |k_h(\bar{x} - x_i)d_i|) = o_p(1)$. Thus by Chebyshev's Inequality, the uniform convergence results above, and the assumption that $\check{\alpha} - \alpha = o_p(1)$, the last three terms in (24) are $o_p(1)$.

The conclusion that $\check{\sigma}^{2+}(\bar{x}) \xrightarrow{p} \sigma^{2+}(\bar{x})$ follows by another application of Chebyshev's Inequality to the first term of (24) using $\sup_{\mathcal{N}} E(\varepsilon^4|x) < \infty$. The other conclusion of the theorem follows similarly. $\square$

PROOF OF THEOREM 5: The proof method below follows Stone (1980). First we define a sequence of functions $\theta_n \in \Theta$ that approach θ_0. The sequence is chosen to be difficult to distinguish from θ_0 given a sample of size n in the sense that the Bayes risk can be bounded away from zero. Then the Bayes risk of a decision rule that takes its action based on an estimator $\hat{\alpha}$ of α is also bounded below. By appropriate choice of such a decision rule such a bound can be re-written to show (11) holds. Equation (12) holds by a modification of this argument. Without loss of generality, we'll take $\bar{x} = 0$. Let P_θ denote the distribution of the observation (x, y) under θ with marginal density for x, f_x.

Suppose $\phi(x)$ is an infinitely differentiable function such that $\phi(x) = 0$ for $x \leq \underline{K} < 0$ and $\phi(x) = \zeta > 0$ for $x \geq \bar{K} > 0$. For $\delta \in (0,1]$ and $N > 0$, define

$$g_{n,\delta}(x) = -\delta N^p n^{-\frac{p}{2p+1}} \phi\left(\frac{n^{\frac{1}{2p+1}} x}{N}\right) + \delta N^p n^{-\frac{p}{2p+1}} \zeta \mathbf{1}\{x \geq 0\}.$$

Then a sequence of functions is defined $\theta_{n,\delta} = \theta_0 + g_{n,\delta} \in \Theta$.

Now note for n large enough,

$$\begin{aligned}
nEg_{n,1}^2(x) &= n \int_{Nn^{\frac{-1}{2p+1}}\underline{K}}^{0} N^{2p} n^{\frac{-2p}{2p+1}} \phi^2\left(\frac{n^{\frac{1}{2p+1}} x}{N}\right) f_x(x)dx \\
&\quad + n \int_0^{Nn^{\frac{-1}{2p+1}}\bar{K}} N^{2p} n^{\frac{-2p}{2p+1}} \left(\zeta - \phi\left(\frac{n^{\frac{1}{2p+1}} x}{N}\right)\right)^2 f_x(x)dx \\
&= nN^{2p} n^{\frac{-2p}{2p+1}} N n^{\frac{-1}{2p+1}} \left[\int_{\underline{K}}^{0} \phi^2(u) f_x(Nn^{-\frac{1}{2p+1}}u)du + \int_0^{\bar{K}} (\zeta - \phi(u))^2 f_x(Nn^{-\frac{1}{2p+1}}u)du\right] \\
&\leq N^{2p+1}\left[\sup_{\underline{K} \leq u \leq 0} \phi^2(u) + \sup_{0 \leq u \leq \bar{K}} (r - \phi(u))^2\right]\left[\sup_{x \in \mathcal{N}} f_x(x)\right](\bar{K} - \underline{K}) \\
&< C < \infty
\end{aligned}$$

Also, $g_{n,\delta} = \delta g_{n,1}$, so for the same constant C as above, $nEg_{n,\delta}^2(x) < \delta^2 C$. Then, for any fixed $\delta \in (0,1]$,

$$\limsup_n nEg_{n,\delta}^2(x) < \infty \tag{25}$$

and

$$\lim_{\delta \longrightarrow 0} \limsup_n nEg_{n,\delta}^2(x) = 0 \tag{26}$$

Let μ_n and ν_n denote the joint distribution of $(x_1, y_1), \ldots, (x_n, y_n)$ under P_{θ_0} and P_{θ_n}. Let $L_n = d\nu_n/d\mu_n$ (the Radon-Nikodym derivative) and $l_n = \ln L_n$. Next show that

$$\limsup_n E_{\theta_0}|l_n| < \infty \tag{27}$$

and

$$\lim_{\delta \longrightarrow 0} \limsup_n E_{\theta_0}|l_n| = 0. \tag{28}$$

Since ϕ is (uniformly) continuous on $[\underline{K}, \bar{K}]$ and $\sup_{x \notin [\underline{K}, \bar{K}]} |\phi(x)| = \zeta$, there exists a constant $A > 0$ such that $\max\{|\phi|, |\zeta - \phi|\} \leq A$. Choose n large enough that $\delta N^p n^{\frac{-p}{2p+1}} A < \tau_0$.

If U is an open neighborhood of the origin, one can choose n large enough that $\{Nn^{\frac{-1}{2p+1}}x|x \in [\underline{K}, \bar{K}]\} \subset U$. Then $g_{n,\delta} = 0$ on U^c.

Now a Taylor expansion of l_n gives an expression in terms of $g_{n,\delta}$, so that (25) and (26) can be used to prove (27) and (28). Let l' and f' denote the partial derivatives of $l(y|x,t)$ and $f(y|x,t)$ with respect to t (and similarly for l'' and f'').

$$\begin{aligned} l_n &= \sum_i [l(y_i|x_i, \theta_0(x_i) + g_{n,\delta}(x_i)) - l(y_i|x_i, \theta_0(x_i))] \\ &= \sum_{i:x_i \in U} g_{n,\delta}(x_i) l'(y_i|x_i, \theta_0(x_i)) + \frac{1}{2} \sum_{i:x_i \in U} g^2_{n,\delta}(x_i) l''(y_i|x_i, \theta_0(x_i) + \tau_i) \end{aligned}$$

for some τ_i with $|\tau_i| \leq \tau_0$.

The expectation of the second term in the summand above can be bounded directly,

$$\begin{aligned} E_{\theta_0} \left| \frac{1}{2} \sum_{i:x_i \in U} g^2_{n,\delta}(x_i) l''(y_i|x_i, \theta_0(x_i) + \tau_i) \right| &\leq E_{\theta_0} \left[\frac{1}{2} \sum_{i:x_i \in U} g^2_{n,\delta}(x_i) M(y_i|x_i, \theta_0(x_i)) \right] \\ &\leq \frac{Cn}{2} E[g^2_{n,\delta}(x)] \end{aligned}$$

By assumption,

$$\int l'(y|x,t) f(y|x,t) \psi(dy) = \int f'(y|x,t) dy = 0.$$

So,

$$E_{\theta_0} \left[\sum_{i:x_i \in U} g_{n,\delta}(x_i) l'(y_i|x_i, \theta_0(x_i)) \right] = 0$$

As in the proof of the information matrix equality,

$$\begin{aligned} \int (l'(y|x,t))^2 f(y|x,t) dy &= \int f''(y|x,t) dy - \int l''(y|x,t) f(y|x,t) dy \\ &\leq \int |l''(y|x,t)| f(y|x,t) dy \\ &\leq \int |M(y|x,t)| f(y|x,t) dy \\ &\leq C. \end{aligned}$$

Thus,

$$\begin{aligned}
E_{\theta_0}\left|\sum_{i:x_i\in U} g_{n,\delta}(x_i)l'(y_i|x_i,\theta_0(x_i))\right| &\leq \left[E_{\theta_0}\left(\sum_{i:x_i\in U} g_{n,\delta}(x_i)l'(y_i|x_i,\theta_0(x_i))\right)^2\right]^{1/2} \\
&= \left[nE\left(g_{n,\delta}^2(x)E_{\theta_0}[(l'(y|x,\theta_0(x))^2|x,\theta_0(x)]\right)\right]^{1/2} \\
&\leq \left[CnEg_{n,\delta}^2(x)\right]^{1/2}
\end{aligned}$$

Then,

$$E_{\theta_0}|l_n| \leq \left[CnEg_{n,\delta}^2(x)\right]^{1/2} + \frac{Cn}{2}E[g_{n,\delta}^2(x)].$$

Hence, equations (27) and (28) follow by (25) and (26).

By (27), for some $0 < B < \infty$, $\limsup_n E_{\theta_0}|\ln L_n| < B$. Choose $\epsilon \in (0, \frac{1}{2})$ such that $\frac{\epsilon}{1-\epsilon} < e^{-2B}$. By Markov's Inequality,

$$\begin{aligned}
\limsup_n \mu_n\left(\left\{L_n > \frac{1-\epsilon}{\epsilon}\right\} \cup \left\{L_n < \frac{\epsilon}{1-\epsilon}\right\}\right) &\leq \limsup_n \mu_n(|\ln L_n| > 2B) \\
&\leq \frac{1}{2B}\limsup_n E_{\theta_0}|\ln L_n| \\
&< \frac{1}{2}.
\end{aligned}$$

It follows that $\liminf_n \mu_n\left(\frac{\epsilon}{1-\epsilon} \leq L_n \leq \frac{1-\epsilon}{\epsilon}\right) > \frac{1}{2}$, so choose n large enough that

$$\mu_n\left(\frac{\epsilon}{1-\epsilon} \leq L_n \leq \frac{1-\epsilon}{\epsilon}\right) > \frac{1}{2}.$$

Now consider the decision problem corresponding to a test of a simple hypothesis between θ_n and θ_0. The above inequality will be used to derive a lower bound on the Bayes' risk of Bayes' rule. Any other rule based on estimators of θ will necessarily have greater Bayes' risk, which provides the desired upper bound on the rate of convergence.

Define the following loss function as follows,

$$L(\theta, a) = \begin{cases} 1 & \text{if } \theta \neq a \\ 0 & \text{if } \theta = a \end{cases}.$$

Let (x^n, y^n) denote the first n data points $(x_1, y_1), \ldots, (x_n, y_n)$. Given a prior

$$\pi(\theta) = \begin{cases} \frac{1}{2} & \text{if } \theta = \theta_n \\ \frac{1}{2} & \text{if } \theta = \theta_0 \end{cases}$$

the Bayes risk of a decision rule $a(x^n, y^n)$ is half the sum of the error probabilities

$$B(\pi, a(x^n, y^n)) = \frac{1}{2} P_{\theta_n}(a(x^n, y^n) = \theta_0) + \frac{1}{2} P_{\theta_0}(a(x^n, y^n) = \theta_n).$$

The posterior is

$$P(\theta = \theta_n | x^n, y^n) = \frac{P_{\theta_n}(x^n, y^n)\pi(\theta_n)}{P_{\theta_n}(x^n, y^n)\pi(\theta_n) + P_{\theta_0}(x^n, y^n)\pi(\theta_0)} = \frac{L_n}{L_n + 1}.$$

Then, Bayes' Rule is

$$b(x^n, y^n) = \begin{cases} \theta_n & \text{if } \frac{L_n}{L_n+1} > \frac{1}{2} \\ \theta_0 & \text{otherwise} \end{cases} = \begin{cases} \theta_n & \text{if } L_n > 1 \\ \theta_0 & \text{otherwise} \end{cases}$$

and the corresponding Bayes risk can be bounded below, as follows

$$\begin{aligned} B(\pi, b(x^n, y^n)) &= \frac{1}{2} P_{\theta_n}(L_n \le 1) + \frac{1}{2} P_{\theta_0}(L_n > 1) \\ &\ge \frac{1}{2}\epsilon P_{\theta_n}\left(\frac{\epsilon}{1-\epsilon} \le L_n \le 1\right) + \frac{1}{2} P_{\theta_0}\left(\frac{1-\epsilon}{\epsilon} \ge L_n > 1\right) \\ &= \frac{1}{2}\epsilon P_{\theta_n}\left(\frac{\epsilon}{1-\epsilon} \le L_n \le 1\right) + \frac{1}{2}\int_{1 > \frac{1}{L_n} \ge \frac{\epsilon}{1-\epsilon}} \frac{1}{L_n} dP_{\theta_n} \\ &\ge \frac{1}{2}\epsilon P_{\theta_n}\left(\frac{\epsilon}{1-\epsilon} \le L_n \le 1\right) + \frac{1}{2}\epsilon \int_{1 > \frac{1}{L_n} \ge \frac{\epsilon}{1-\epsilon}} dP_{\theta_n} \\ &= \frac{1}{2}\epsilon \mu_n\left(\frac{\epsilon}{1-\epsilon} \le L_n \le \frac{1-\epsilon}{\epsilon}\right) \\ &> \frac{\epsilon}{4}. \end{aligned}$$

Now consider a decision rule based on some estimator $\check{\alpha}_n$:

$$a(x^n, y^n) = \begin{cases} \theta_n & \text{if } \check{\alpha}_n \ge \frac{\alpha(\theta_0) + \alpha(\theta_n)}{2} \\ \theta_0 & \text{otherwise} \end{cases}$$

Since Bayes' rule minimizes Bayes' risk,

$$\begin{aligned} \frac{\epsilon}{4} &< \frac{1}{2} P_{\theta_n}\left(\check{\alpha}_n < \frac{\alpha(\theta_0) + \alpha(\theta_n)}{2}\right) + \frac{1}{2} P_{\theta_0}\left(\check{\alpha}_n \ge \frac{\alpha(\theta_0) + \alpha(\theta_n)}{2}\right) \\ &\le \frac{1}{2} P_{\theta_n}\left(|\check{\alpha}_n - \alpha(\theta_n)| > \frac{\alpha(\theta_n) - \alpha(\theta_0)}{2}\right) + \frac{1}{2} P_{\theta_0}\left(|\check{\alpha}_n - \alpha(\theta_0)| \ge \frac{\alpha(\theta_n) - \alpha(\theta_0)}{2}\right) \\ &\le \sup_{\theta \in \Theta} P_\theta\left(|\check{\alpha}_n - \alpha(\theta)| \ge \frac{\alpha(\theta_n) - \alpha(\theta_0)}{2}\right) \end{aligned}$$

Since N can be chosen arbitrarily large and

$$\frac{\alpha(\theta_n) - \alpha(\theta_0)}{2} = \frac{\alpha + \delta N^p n^{\frac{-p}{2p+1}}\zeta - \alpha}{2} = c n^{\frac{-p}{2p+1}},$$

(11) holds, as desired.

Next show (12). Below it will be shown that

$$\liminf_n \sup_{\theta\in\Theta} P_\theta\left(|\breve{\alpha}_n - \alpha(\theta)| \geq \frac{\delta N^p \zeta}{2(J-1)} n^{\frac{-p}{2p+1}}\right) \geq 1 - \frac{2}{J} \tag{29}$$

Given $0 < \eta < 1$, choose J large enough that $\frac{2}{J} < \eta$ and $C = \frac{\delta N^p \zeta}{4(J-1)}$. Then for all $c \leq C \leq \frac{\delta N^p \zeta}{2(J-1)}$,

$$\liminf_n \sup_{\theta\in\Theta} P_\theta(|\breve{\alpha}_n - \alpha(\theta)| > cn^{\frac{-p}{2p+1}}) \geq \liminf_n \sup_{\theta\in\Theta} P_\theta\left(|\breve{\alpha}_n - \alpha(\theta)| > \frac{\delta N^p \zeta}{2(J-1)} n^{\frac{-p}{2p+1}})\right) > 1 - \eta.$$

Hence, showing (29) will be sufficient to prove (12).

Given J above, let $\xi = \frac{1}{8J^3}$. From (26), $\delta > 0$ can be chosen small enough that

$$\limsup_n E_{\theta_0}|l_n| \leq \left[CnEg^2_{n,\delta}(x)\right]^{1/2} + \frac{Cn}{2} E[g^2_{n,\delta}(x)] < -\xi \ln(1-\xi).$$

Then for $j = 1, \ldots, J$, set

$$\theta_{n_j} = \theta_0 + \frac{j-1}{J-1} g_{n,\delta}.$$

If ν_{n_j} denotes the joint distribution of (x^n, y^n) under $P_{\theta_{n_j}}$, let L_{n_j} denote the Radon-Nikodym derivative $d\nu_{n_j}/d\nu_{n_1}$ and set $l_{n_j} = \ln L_{n_j}$.

By the same arguments as above,

$$E_{\theta_0}|l_{n_j}| \leq \frac{j-1}{J-1}\left[CnEg^2_{n,\delta}(x)\right]^{1/2} + \frac{Cn}{2} E\left[\left(\frac{j-1}{J-1}\right)^2 g^2_{n,\delta}(x)\right] < -\xi \ln(1-\xi).$$

Hence, for $j = 1, \ldots, J$,

$$\begin{aligned}
\liminf_n \mu_n\left(1-\xi \leq L_{n_j} \leq \frac{1}{1-\xi}\right) &\geq 1 - \limsup_n \mu_n(|\ln L_{n_j}| > -\ln(1-\xi)) \\
&\geq 1 - \limsup_n \frac{1}{-\ln(1-\xi)} E_{\theta_0}|l_{n_j}| \\
&> 1 - \xi.
\end{aligned}$$

Choose n large enough that $P_{\theta_0}\left(1-\xi \leq L_{n_j} \leq \frac{1}{1-\xi}\right) > 1 - \xi$ for all j.

Then for $j \neq k$ and $j > 1$,

$$\begin{aligned}
P_{\theta_{n_k}}\left((1-\xi)^2 \le \frac{L_{n_j}}{L_{n_k}} \le \frac{1}{(1-\xi)^2}\right) &\ge P_{\theta_{n_k}}\left(1-\xi \le L_{n_j} \le \frac{1}{1-\xi} \bigcap 1-\xi \le L_{n_k} \le \frac{1}{1-\xi}\right) \\
&= \int_{1-\xi \le L_{n_j} \le \frac{1}{1-\xi}} \mathbf{1}\{1-\xi \le L_{n_k} \le \frac{1}{1-\xi}\} L_{n_k} dP_{\theta_0} \\
&\ge (1-\xi) P_{\theta_0}\left(1-\xi \le L_{n_j} \le \frac{1}{1-\xi}\right) \\
&> (1-\xi)^2 \ge 1-3\xi.
\end{aligned}$$

Similarly,

$$P_{\theta_{n_k}}\left((1-\xi)^2 \le \frac{1}{L_{n_k}} \le \frac{1}{(1-\xi)^2}\right) > (1-\xi)^3 \ge 1-3\xi.$$

Given a prior $\pi(\theta) = \begin{cases} \frac{1}{J} & \text{if } \theta = \theta_{n_j} \\ 0 & \text{otherwise} \end{cases}$, the posterior is $P(\theta = \theta_{n_j} | x^n, y^n) = \frac{L_{n_j}}{\sum_{k=1}^{J} L_{n_k}}$, where $L_{n_1} = 1$.

Now for $j \neq k$ and $k > 1$,

$$\begin{aligned}
& P_{\theta_{n_k}}\left(\frac{1-4\xi}{J} \le P(\theta = \theta_{n_j} | x^n, y^n) \le \frac{1}{(1-4\xi)J}\right) \\
\ge\ & P_{\theta_{n_k}}\left(\frac{(1-\xi)^2}{1+(J-1)\frac{1}{(1-\xi)^2}} \le \frac{\frac{L_{n_j}}{L_{n_k}}}{1+\sum_{i \neq k} \frac{L_{n_i}}{L_{n_k}}} \le \frac{\frac{1}{(1-\xi)^2}}{1+(J-1)(1-\xi)^2}\right) \\
\ge\ & P_{\theta_{n_k}}\left(\bigcap_{i \neq k} \left\{(1-\xi)^2 \le \frac{L_{n_i}}{L_{n_k}} \le \frac{1}{(1-\xi)^2}\right\}\right) \\
\ge\ & \sum_{i \neq k} P_{\theta_{n_k}}\left((1-\xi)^2 \le \frac{L_{n_i}}{L_{n_k}} \le \frac{1}{(1-\xi)^2}\right) - (J-2) \\
\ge\ & (J-1)(1-3\xi) - (J-2) = 1 - 3\xi(J-1).
\end{aligned}$$

By similar arguments,

$$P_{\theta_{n_k}}\left(\frac{1-4\xi}{J} \le P(\theta = \theta_{n_k} | x^n, y^n) \le \frac{1}{(1-4\xi)J}\right) \ge 1 - 3\xi(J-1)$$

$$P_{\theta_0}\left(\frac{1-4\xi}{J} \le P(\theta = \theta_{n_k} | x^n, y^n) \le \frac{1}{(1-4\xi)J}\right) \ge 1 - 3\xi(J-1), \text{ and}$$

$$P_{\theta_0}\left(\frac{1-4\xi}{J} \le P(\theta = \theta_0 | x^n, y^n) \le \frac{1}{(1-4\xi)J}\right) \ge 1 - 3\xi(J-1).$$

Using the same loss function as before, a lower bound for the Bayes' risk can now be established. Consider a decision rule $a(x^n, y^n) = \theta_{n_j}$ with probability $q_j(x^n, y^n)$ for $j = 1, \ldots, J$. Let $P(x^n, y^n)$

$= \sum_{j=1}^{J} P_{\theta_{n_j}}(x^n, y^n)\pi(\theta_{n_j})$.

$$
\begin{aligned}
&B(\pi, a(x^n, y^n)) \\
&= \sum_{j=1}^{J} \pi(\theta_{n_j}) \int L(\theta_{n_j}, a(x^n, y^n)) P_{\theta_{n_j}}(x^n, y^n) d(x^n, y^n) \\
&= \int \left[\sum_{j=1}^{J} (1 - q_j(x^n, y^n)) P(\theta = \theta_{n_j} | x^n, y^n) \right] P(x^n, y^n) d(x^n, y^n) \\
&\geq \left(\frac{1-4\xi}{J} \right) \int_{\cap_{i=1}^{J} \{ \frac{1-4\xi}{J} \leq P(\theta = \theta_{n_i} | x^n, y^n) \leq \frac{1}{J(1-4\xi)} \}} \left[\sum_{j=1}^{J} (1 - q_j(x^n, y^n)) \right] P(x^n, y^n) d(x^n, y^n) \\
&= \frac{1}{J}(1-4\xi)(J-1) P \left(\cap_{i=1}^{J} \{ \frac{1-4\xi}{J} \leq P(\theta = \theta_{n_i} | x^n, y^n) \leq \frac{1}{J(1-4\xi)} \} \right) \\
&\geq (1-4\xi)\left(1 - \frac{1}{J}\right) \left(\sum_{i=1}^{J} \left[\sum_{j=1}^{J} \frac{1}{J} P_{\theta_{n_j}} \left(\frac{1-4\xi}{J} \leq P(\theta = \theta_{n_i} | x^n, y^n) \leq \frac{1}{J(1-4\xi)} \right) \right] - (J-1) \right) \\
&= (1-4\xi)\left(1 - \frac{1}{J}\right) (J(1 - 3\xi(J-1)) - (J-1)) \\
&\geq 1 - \frac{2}{J}
\end{aligned}
$$

Consider the following decision rule based on an estimator $\check{\alpha}_n$:

$$
a(x^n, y^n) = \begin{cases} \theta_0 & \text{if } \check{\alpha}_n < \frac{\alpha(\theta_0) + \alpha(\theta_{n_2})}{2} \\ \theta_{n_j} & \text{if } \frac{\alpha(\theta_{n_{j-1}}) + \alpha(\theta_{n_j})}{2} \leq \check{\alpha}_n < \frac{\alpha(\theta_{n_j}) + \alpha(\theta_{n_{j+1}})}{2} \text{ and } 2 \leq j \leq J-1 \\ \theta_{n_J} & \text{if } \check{\alpha}_n \geq \frac{\alpha(\theta_{n_{J-1}}) + \alpha(\theta_{n_J})}{2} \end{cases}
$$

Note that for $j = 2, \ldots, J$,

$$
\frac{\alpha(\theta_{n_j}) - \alpha(\theta_{n_{j-1}})}{2} = \frac{\delta N^p \zeta}{2(J-1)} n^{\frac{-p}{2p+1}}.
$$

So,

$$
\begin{aligned}
1 - \frac{2}{J} &\leq \frac{1}{J} P_{\theta_0} \left(\check{\alpha}_n \geq \frac{\alpha(\theta_0) + \alpha(\theta_{n_2})}{2} \right) + \frac{1}{J} P_{\theta_{n_J}} \left(\check{\alpha}_n < \frac{\alpha(\theta_{n_{J-1}}) + \alpha(\theta_{n_J})}{2} \right) \\
&\quad + \sum_{j=2}^{J-1} \frac{1}{J} P_{\theta_{n_j}} \left(\left\{ \check{\alpha}_n < \frac{\alpha(\theta_{n_{j-1}}) + \alpha(\theta_{n_j})}{2} \right\} \bigcup \left\{ \check{\alpha}_n \geq \frac{\alpha(\theta_{n_j}) + \alpha(\theta_{n_{j+1}})}{2} \right\} \right) \\
&\leq \sum_{j=1}^{J} \frac{1}{J} P_{\theta_{n_j}} \left(|\check{\alpha}_n - \alpha(\theta_{n_j})| \geq \frac{\delta N^p \zeta}{2(J-1)} n^{\frac{-p}{2p+1}} \right) \\
&\leq \sup_{\theta \in \Theta} P_\theta \left(|\check{\alpha}_n - \alpha(\theta_{n_j})| \geq \frac{\delta N^p \zeta}{2(J-1)} n^{\frac{-p}{2p+1}} \right)
\end{aligned}
$$

which proves (29) and leads to (28) as described before. Thus $\frac{p}{2p+1}$ is an upper bound to the rate of convergence. From Theorem 2(b) and Theorem 3, the partially linear estimator and local polynomial estimator satisfy (13). Part (a) of Theorem 5 follows.

Since $\Theta \subset \bar{\Theta}$,

$$\sup_{\theta\in\bar{\Theta}} P_\theta\left(|\check{\alpha}_n - \alpha(\theta_{n_j})| \geq \frac{\delta N^p \zeta}{2(J-1)} n^{\frac{-p}{2p+1}}\right) \geq \sup_{\theta\in\Theta} P_\theta\left(|\check{\alpha}_n - \alpha(\theta_{n_j})| \geq \frac{\delta N^p \zeta}{2(J-1)} n^{\frac{-p}{2p+1}}\right)$$

and $\frac{p}{2p+1}$ is an upper bound to the rate of convergence for part (b) too. From Theorem 3, the partially linear estimator satisfies (13). Part (b) of the theorem is thus proven. □

B Lemmas

The first four lemmas contain uniform convergence results. Let $\mathcal{N}_1$ and $\mathcal{N}_0$ be compact intervals such that $\bar{x} \in int(\mathcal{N}_0)$ and $\mathcal{N}_0 \subset int(\mathcal{N}_1)$. Define $\mathcal{N}_{1-} = (-\infty, \bar{x}] \cap \mathcal{N}_1$, $\mathcal{N}_{1+} = [\bar{x}, \infty) \cap \mathcal{N}_1$, $\mathcal{N}_{0-} = (-\infty, \bar{x}] \cap \mathcal{N}_0$, and $\mathcal{N}_{0+} = [\bar{x}, \infty) \cap \mathcal{N}_0$.

Lemma B1 *Suppose the kernel k is bounded, symmetric, zero outside of a bounded set $\mathcal{M}$ and Lipschitz, and f_0 is continuous on $\mathcal{N}_1$. If $nh/\ln n \longrightarrow \infty$, then* $\sup_{x\in\mathcal{N}_{0-}} |\hat{f}_+(x) - E\hat{f}_+(x)| = O_p\left(\sqrt{\frac{\ln n}{nh}}\right)$, $\sup_{x\in\mathcal{N}_{0+}} |\hat{f}_-(x) - E\hat{f}_-(x)| = O_p\left(\sqrt{\frac{\ln n}{nh}}\right)$, *and* $\sup_{x\in\mathcal{N}_0} |\hat{f}(x) - E\hat{f}(x)| = O_p\left(\sqrt{\frac{\ln n}{nh}}\right)$.

PROOF: By k Lipschitz, $\sup_{|x-z|<n^{-3}} |\hat{f}_+(x) - \hat{f}_+(z)| \leq Cn^{-3}h^{-2}$. Compact $\mathcal{N}_{0-}$ can be covered by J open balls with centers $x_{j,n}$ and radii n^{-3}. There exists $B > 0$ such that $J \leq Bn^3$. Let $x_{j,n}(x)$ denote the center of a ball containing x.

$$\begin{aligned}\sup_{x\in\mathcal{N}_{0-}} |\hat{f}_+(x) - E\hat{f}_+(x)| \;\leq\;& \sup_{x\in\mathcal{N}_{0-}} |\hat{f}_+(x) - \hat{f}_+(x_{j,n}(x))| + \sup_{x\in\mathcal{N}_{0-}} E|\hat{f}_+(x) - \hat{f}_+(x_{j,n}(x))| \\ &+ \sup_{x\in\mathcal{N}_{0-}} |\hat{f}_+(x_{j,n}(x)) - E\hat{f}_+(x_{j,n}(x))| \\ \leq\;& C'\frac{1}{n^3h^2} + \sup_j |\hat{f}_+(x_{j,n}) - E\hat{f}_+(x_{j,n})|\end{aligned}$$

Note that $2nV(k_h(x_{j,n} - x_i)d_i) \leq 2(n/h)\int_{\mathcal{M}} k^2(u) f_0(x_{j,n} + uh)du \leq C(n/h)$ for n large enough. Also, $M\sqrt{\ln n/(nh)} - C'n^{-3}h^{-2} \geq (M/2)\sqrt{\ln n/(nh)}$ for n large enough. Hence, using Bernstein's inequality,

$$
\begin{aligned}
\Pr\left(\sup_{\mathcal{N}_{0-}}\left|\hat{f}_+(x)-E\hat{f}_+(x)\right|>M\sqrt{\frac{\ln n}{nh}}\right) &\leq \Pr\left(\sup_{j}\left|\hat{f}_+(x_{j,n})-E\hat{f}_+(x_{j,n})\right|>\frac{M}{2}\sqrt{\frac{\ln n}{nh}}\right)\\
&\leq \sum_{j=1}^{J}\Pr\left(\left|\sum_{i=1}^{n}\left[\frac{1}{h}k\left(\frac{x_{j,n}-x_i}{h}\right)d_i-E\left[\frac{1}{h}k\left(\frac{x_{j,n}-x_i}{h}\right)d_i\right]\right]\right|>n\frac{M}{2}\sqrt{\frac{\ln n}{nh}}\right)\\
&\leq (Bn^3)2\exp\left(\frac{-\frac{M^2}{4}n^2\frac{\ln n}{nh}}{2nV\left(\frac{1}{h}k\left(\frac{x_{j,n}-x_i}{h}\right)d_i\right)+\frac{2}{3}\frac{C}{h}n\frac{M}{2}\sqrt{\frac{\ln n}{nh}}}\right)\\
&\leq Cn^{3-C'M^2}=o(1) \quad \text{for } M \text{ large enough,}
\end{aligned}
$$

where the last inequality follows by noting that for n large enough $\frac{M}{3}\sqrt{\frac{\ln n}{nh}}\leq\frac{1}{2}$. The remaining results hold analogously. □

The next lemma is a modification of Newey (1994) Lemma B.1 that only bounds a conditional moment of y, and the proof follows Newey's.

Lemma B2 *Suppose the kernel k is bounded, symmetric, zero outside of a bounded set $\mathcal{M}$ and Lipschitz, f_0 is continuous on $\mathcal{N}_1$, $\sup_{x\in\mathcal{N}_1}E[|y|^{2+\zeta}|x]<\infty$ for some $\zeta>0$, and $n^{\zeta/(2+\zeta)}h/(\ln n)\longrightarrow\infty$. Then $\sup_{x\in\mathcal{N}_0}|\tilde{r}(x)-E\tilde{r}(x)|=O_p\left(\sqrt{\frac{\ln n}{nh}}\right)$.*

PROOF: Note $\tilde{r}(x)=n^{-1}\sum_i k_h(x-x_i)\tilde{y}_i$ and define $\tilde{\tilde{r}}(x)=n^{-1}\sum_i k_h(x-x_i)\tilde{y}_{i,n}$, where $\tilde{y}_{i,n}=\tilde{y}_i$ if $|\tilde{y}_i|\leq Pn^{1/(2+\zeta)}$; $\tilde{y}_{i,n}=Pn^{1/(2+\zeta)}$ if $\tilde{y}_i>Pn^{1/(2+\zeta)}$; $\tilde{y}_{i,n}=-Pn^{1/(2+\zeta)}$ if $\tilde{y}_i<-Pn^{1/(2+\zeta)}$. For any $M_0>0$ and n large enough,

$$
\begin{aligned}
\Pr\left(\sup_{\mathcal{N}_0}\sqrt{\frac{nh}{\ln n}}|\tilde{\tilde{r}}(x)-\tilde{r}(x)|>M_0\right) &\leq \Pr(\tilde{\tilde{r}}(x)\neq\hat{r}(x) \text{ for some } x\in\mathcal{N}_0)\\
&\leq \Pr(\tilde{y}_{i,n}\neq\tilde{y}_i \text{ and } x_i\in\mathcal{N}_1 \text{ for some } i)\\
&\leq n\Pr(\tilde{y}_{i,n}\neq\tilde{y}_i \text{ and } x_i\in\mathcal{N}_1)\\
&\leq n\Pr(\tilde{y}_{i,n}\neq\tilde{y}_i|\ x_i\in\mathcal{N}_1)\\
&\leq \frac{1}{P^{2+\zeta}}E[|\tilde{y}_i|^{2+\zeta}|x_i\in\mathcal{N}_1]. \qquad (30)
\end{aligned}
$$

The last inequality holding by conditional Markov's inequality

For large enough n, if $\Pr(x_i \in \mathcal{N}_1) = 0$, then $|E\tilde{\tilde{r}}(x) - E\tilde{r}(x)| = 0$ for $x \in \mathcal{N}_0$almost surely; otherwise,

$$\begin{aligned}
& \sup_{x\in\mathcal{N}_0} \sqrt{\frac{nh}{\ln n}} \left|E\tilde{\tilde{r}}(x) - E\tilde{r}(x)\right| && (31) \\
\leq\ & \sup_{x\in\mathcal{N}_0} \sqrt{\frac{nh}{\ln n}} \left| E\left[\frac{1}{nh}\sum_i k\left(\frac{x-x_i}{h}\right)(\tilde{y}_i - \tilde{y}_{i,n})\right]\right| \\
\leq\ & \sup_{x\in\mathcal{N}_0} \sqrt{\frac{nh}{\ln n}} \left| E\left[E(\tilde{y}_i - \tilde{y}_{i,n}|x_i)\frac{1}{h}k\left(\frac{x-x_i}{h}\right)\middle| x_i \in \mathcal{N}_1\right]\right| \Pr(x_i \in \mathcal{N}_1) \\
\leq\ & \sup_{x\in\mathcal{N}_0} \sqrt{\frac{nh}{\ln n}} E\left[E\left[\left(\frac{|\tilde{y}_i|}{Pn^{\frac{1}{2+\zeta}}}\right)^{1+\zeta} |\tilde{y}_i||x_i\right]\left|\frac{1}{h}k\left(\frac{x-x_i}{h}\right)\right|\middle| x_i \in \mathcal{N}_1\right] \Pr(x_i \in \mathcal{N}_1) \\
\leq\ & C\sqrt{\frac{nh}{\ln n}} \frac{1}{n^{(1+\zeta)/(2+\zeta)}P^{p-1}} \sup_{x\in\mathcal{N}_0} E\left[\left|k\left(\frac{x-x_i}{h}\right)\right|\middle| x_i \in \mathcal{N}_1\right] \Pr(x_i \in \mathcal{N}_1) \\
\leq\ & C'\sqrt{\frac{h}{n^{\frac{\zeta}{2+\zeta}}\ln n}} \sup_{x\in\mathcal{N}_0} \int_{\mathcal{M}} |k(u)| f_0(x+uh)du \\
=\ & o(1). && (32)
\end{aligned}$$

Cover compact $\mathcal{N}_0$ by J open balls (intervals) with centers $x_{j,n}$ and radii n^{-3}. There exists $B > 0$ such that $J \leq Bn^3$. Let $x_{j,n}(x)$ denote the the center of the ball containing x. Since k is Lipschitz, $\sup_{|x-z|<n^{-3}} |\tilde{\tilde{r}}(x) - \tilde{\tilde{r}}(z)| \leq Cn^{-3+1/(2+\zeta)}h^{-2}$. Then,

$$\begin{aligned}
\sup_{x\in\mathcal{N}_0} |\tilde{\tilde{r}}(x) - E\tilde{\tilde{r}}(x)| \ \leq\ & \sup_{x\in\mathcal{N}_0} |\tilde{\tilde{r}}(x) - \tilde{\tilde{r}}(x_{j,n}(x))| + \sup_{x\in\mathcal{N}_0} E|\tilde{\tilde{r}}(x) - \tilde{\tilde{r}}(x_{j,n}(x))| \\
& + \sup_{x\in\mathcal{N}_0} |\tilde{\tilde{r}}(x_{j,n}(x)) - E\tilde{\tilde{r}}(x_{j,n}(x))| \\
\leq\ & C'\frac{n^{\frac{1}{2+\zeta}-3}}{h^2} + \sup_j |\tilde{\tilde{r}}(x_{j,n}) - E\tilde{\tilde{r}}(x_{j,n})|
\end{aligned}$$

Note that $2nV(k_h(x_{j,n} - x_i)\tilde{y}_{i,n}) \leq 2nh^{-2}E[k^2((x_{j,n} - x_i)/h)\tilde{y}_i^2] \leq 2(n/h)\sup_{x_i\in\mathcal{N}_1} E(\tilde{y}_i^2|x_i)$ $\int_{\mathcal{M}} k^2(u)f_0(x_{j,n} + uh)du \leq C(n/h)$. Also, $M\sqrt{\ln n/(nh)}\ -Cn^{-3+1/(2+\zeta)}h^{-2} \geq (M/2)\sqrt{\ln n/(nh)}$ for n large enough. Similarly, we can choose n large enough that $Mn^{1/(2+\zeta)}\sqrt{\ln n/(nh)} < .5$ by supposition. Hence, using Bernstein's inequality,

$$
\begin{aligned}
\Pr\left(\sup_{\mathcal{N}_0}\left|\tilde{\tilde{r}}(x)-E\tilde{\tilde{r}}(x)\right|>M\sqrt{\frac{\ln n}{nh}}\right) &\leq \Pr\left(\sup_{j}\left|\tilde{\tilde{r}}(x_{j,n})-E\tilde{\tilde{r}}(x_{j,n})\right|>\frac{M}{2}\sqrt{\frac{\ln n}{nh}}\right) \\
\leq \quad & \sum_{j=1}^{J}\Pr\left(\left|\sum_{i=1}^{n}\left[\frac{1}{h}k\left(\frac{x_{j,n}-x_i}{h}\right)\tilde{y}_{i,n}-E\left[\frac{1}{h}k\left(\frac{x_{j,n}-x_i}{h}\right)\tilde{y}_{i,n}\right]\right]\right|>n\frac{M}{2}\sqrt{\frac{\ln n}{nh}}\right) \\
\leq \quad & (Bn^3)2\exp\left(\frac{-\frac{M^2}{4}n^2\frac{\ln n}{nh}}{2nV\left(\frac{1}{h}k\left(\frac{x_{j,n}-x_i}{h}\right)\tilde{y}_{i,n}\right)+\frac{2}{3}\frac{1}{h}Pn^{1/(2+\zeta)}n\frac{M}{2}\sqrt{\frac{\ln n}{nh}}}\right) \\
\leq \quad & Cn^{3-C'M^2}=o(1) \quad \text{for } M \text{ large enough.} \qquad (33)
\end{aligned}
$$

where the last inequality follows by the rate condition giving $\frac{M}{3}Pn^{1/(2+\zeta)}\sqrt{\frac{\ln n}{nh}} < \frac{C}{2}$ for large enough n.

Given $\epsilon > 0$, choose P such that $P^{-(2+\zeta)}E(|\tilde{y}_i|^{2+\zeta}|x_i \in \mathcal{N}_1) < \epsilon/2$. In turn, P determines $\tilde{\tilde{r}}$. From (32), (33), and the triangle inequality, $\sup_{\mathcal{N}_0}|\tilde{\tilde{r}}(x) - E\tilde{r}(x)| = O_p(\sqrt{\ln n/(nh)})$. So, there exists $M > 0$ such that $\Pr(\sqrt{nh/\ln n}\sup_{\mathcal{N}_0}|\tilde{\tilde{r}}(x) - E\tilde{r}(x)| > M/2) < \epsilon/2$ for large enough n. By (30), $\Pr(\sqrt{nh/\ln n}\sup_{\mathcal{N}_0}|\tilde{\tilde{r}}(x) - \tilde{r}(x)| > M/2) < \epsilon/2$. By the triangle inequality, $\Pr(\sqrt{nh/\ln n}\sup_{\mathcal{N}_0}|\tilde{r}(x) - E\tilde{r}(x)| > M) < \epsilon$, and the conclusion follows. □

LEMMA B3 *Suppose the kernel k is bounded, symmetric, zero outside of a bounded set $[-M, M]$. On $\mathcal{N}_1$, f_0 is continuously differentiable; $m(x)$ is continuously differentiable for $x \neq \bar{x}$, $x \in \mathcal{N}_1$, and m is continuous at $\bar{x}$ with finite right and left-hand derivatives. Then $\sup_{x\in[\bar{x},\bar{x}+Mh]}|E\tilde{r}(x) - r(x)| = O(h)$.*

PROOF: For large enough n,

$$
\begin{aligned}
&\sup_{x\in[\bar{x},\bar{x}+Mh]}|E\tilde{r}(x)-r(x)| \\
&= \sup_{x\in[\bar{x},\bar{x}+Mh]}\left|\int_{-M}^{\frac{\bar{x}-x}{h}}k(u)[r(x+uh)-r(\bar{x})-(r(x)-r(\bar{x}))]du\right. \\
&\qquad\left.+\int_{\frac{\bar{x}-x}{h}}^{M}k(u)(r(x+uh)-r(x))du\right|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sup_{x\in[\bar{x},\bar{x}+Mh]}\left[\sup_{z\in\mathcal{N}_{1-}}|r'^{-}(z)|+\sup_{z\in\mathcal{N}_{1+}}|r'^{+}(z)|\right]|x-\bar{x}|\int_{-M}^{\frac{\bar{x}-x}{h}}|k(u)|du \\
&\qquad +h\sup_{z\in\mathcal{N}_{1-}}|r'^{-}(z)|\int_{-M}^{\frac{\bar{x}-x}{h}}|k(u)u|du+h\sup_{z\in\mathcal{N}_{1+}}|r'^{+}(z)|\int_{\frac{\bar{x}-x}{h}}^{M}|k(u)u|du \\
&= O(h)
\end{aligned}
$$

□

LEMMA B4 *Suppose the kernel k is bounded, symmetric, zero outside of a bounded set $[-M,M]$ and $\int_{-M}^{M}k(u)u^j du=0$ for $1\leq j<s$. If f_0 and m are continuously differentiable to order s on $\mathcal{N}_1$, then $\sup_{x\in\mathcal{N}_0}|E\hat{f}_0(x)-f_0(x)|=O_p(h^s)$ and $\sup_{x\in\mathcal{N}_0}|E\tilde{r}(x)-r(x)|=O_p(h^s)$.*

PROOF: For large enough n, and some mean value x_u between x and $x+uh$, $\sup_{\mathcal{N}_0}|E\hat{f}_0(x)-f_0(x)|=\sup_{\mathcal{N}_0}|\int_{\mathcal{M}}k(u)f_0^{(s)}(x_u)u^s h^s/s!du\leq\sup_{\mathcal{N}_1}|f_0^{(s)}(x)/s!|\int_{\mathcal{M}}k(u)u^s du=O(h^s)$ and similarly for the second conclusion. □

LEMMA B5 *If $\{B_n\}$ is a bounded sequence of matrices with $|det(B_n)|\geq\varepsilon\ \forall n\geq N$ for some N and $\varepsilon>0$ and $r_n(A_n-B_n)=o_p(1)$ for a sequence $r_n\longrightarrow\infty$ as $n\longrightarrow\infty$, then $r_n[(A_n^{-1})-(B_n^{-1})]$ $=o_p(1)$.*

PROOF: For bounded sequences, addition and multiplication preserve rates, so $r_n(det(A_n)-det(B_n))=o_p(1)$ and $r_n(det(A_n^{jl})-det(B_n^{jl}))=o_p(1)$ for cofactors A_n^{jl} and B_n^{jl}. Using lower case letters to denote scalars, suppose $r_n(a_n-\alpha_n)=o_p(1)$ and $r_n(d_n-\delta_n)=o_p(1)$. If additionally $|\delta_n|\geq\varepsilon$ for $n\geq N$ and $\{\alpha_n\}$ and $\{\delta_n\}$ are bounded sequences, then

$$
\left|r_n\left(\frac{a_n}{d_n}-\frac{\alpha_n}{\delta_n}\right)\right|\leq\frac{1}{|d_n\delta_n|}\left[|\delta_n||r_n(a_n-\alpha_n)|+|\alpha_n||r_n(d_n-\delta_n)|\right]=O_p(1)\frac{1}{\varepsilon}Co_p(1)=o_p(1)
$$

So, $r_n\left(\frac{det(A_n^{jl})}{det(A_n)}-\frac{det(B_n^{jl})}{det(B_n)}\right)=o_p(1)$. □

References

Andrews, D. (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–596.

Angrist, J., and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533–576.

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

Black, S. (1999): "Do Better Schools Matter? Parental Valuation of Elementary Eduction," *Quarterly Journal of Economics*, 114, 577–599.

Chay, K., and M. Greenstone (1998): "Does Air Quality Matter? Evidence from the Housing Market," NBER Working Paper 6826.

Chen, H. (1988): "Convergence Rates for Parametric Components in a Partly Linear Model," *Annals of Statistics*, 16, 136–146.

Cheng, M., J. Fan, and J. Marron (1997): "On Automatic Boundary Corrections," *Annals of Statistics*, 25, 1691–1708.

Donoho, D., and R. Liu (1987): "Geometrizing Rates of Convergence, I," Discussion Paper 137, UC Berkeley, Department of Statistics.

Engle, R., C. Granger, J. Rice, and A. Weiss (1986): "Semiparametric Estimates of the Relationship Between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310–320.

Fan, J. (1992): "Design Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

FAN, J., AND I. GIJBELS (1996): *Local Polynomial Modelling and its Applications.* Chapman & Hall, New York.

HAHN, J., P. TODD, AND W. VAN DER KLAAUW (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209.

HÄRDLE, W. (1990): *Applied Nonparametric Regression.* Cambridge University Press, New York.

HÄRDLE, W., AND O. LINTON (1994): "Applied Nonparametric Methods," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden, pp. 2295–2339. New York: North-Holland.

HECKMAN, N. (1986): "Spline Smoothing in a Partly Linear Model," *Journal of the Royal Statistical Society, Series B*, 48, 244–248.

IMBENS, G., AND J. ANGRIST (1994): "Identification of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

LEE, D. (2001): "The Electoral Advantage to Incumbency and Voters' Valuation of Politicans' Experience: A Regression Discontinuity Analysis of Elections to the U.S. House," NBER Working Paper 8441.

NEWEY, W., AND D. MCFADDEN (1994): "Estimation and Inference in Large Samples," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden, pp. 2111–2245. New York: North-Holland.

PAGAN, A., AND A. ULLAH (1999): *Nonparametric Econometrics.* Cambridge University Press, New York.

PENCE, K. (2002): "Foreclosing on Opportunity: State Laws and Mortgage Credit," manuscript, Federal Reserve Board of Governors.

RICE, J. (1986): "Convergence Rates for Partially Splined Models," *Statistics and Probability Letters*, 4, 203–208.

Robinson, P. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

Ruppert, D., and M. Wand (1994): "Multivariate Weighted Least Squares Regression," *Annals of Statistics*, 22, 1346–1370.

Serfling, R. (1980): *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Silverman, B. (1992): *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

Speckman (1988): "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society, Series B*, 50, 413–436.

Stock, J. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84, 567–576.

Stone, C. (1980): "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348–1360.

Thistlewaite, D., and D. Campbell (1960): "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, 51, 309–317.

Trochim, W. (1984): *Research Design for Program Evaluation: The Regression Discontinuity Approach*. Sage Publications, Beverly Hills.

Van der Klaauw, W. (1996): "A Regression-Discontinuity Evaluation of the Effect of Financial Aid Offers on College Enrollment," manuscript, New York University.