

Shen, Chan; Klein, Roger

Working Paper

Recursive differencing: Bias reduction with regular kernels

Working Paper, No. 2017-01

Provided in Cooperation with:

Department of Economics, Rutgers University

Suggested Citation: Shen, Chan; Klein, Roger (2017) : Recursive differencing: Bias reduction with regular kernels, Working Paper, No. 2017-01, Rutgers University, Department of Economics, New Brunswick, NJ

This Version is available at:

<https://hdl.handle.net/10419/162934>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Recursive Differencing: Bias Reduction with Regular Kernels

Chan Shen and Roger Klein

1 Introduction

It is well known that it is important to control the bias in estimating conditional expectations in order to obtain asymptotic normality for quantities of interest (e.g. a finite dimensional parameter vector in semiparametric models or averages of marginal effects in the nonparametric case). For this purposes, higher order kernel methods are often employed in developing the theory. However such methods typically do not perform well at moderate sample sizes. Moreover, and perhaps related to their performance, non-optimal windows are selected with undersmoothing needed to ensure the appropriate bias order.

Here, we propose a differences in differences approach to bias reduction for a nonparametric estimator of a conditional expectation. It performs much better at moderate sample sizes than regular or higher order kernels while retaining a bias of any desired order and a convergence rate the same as that of higher order kernels. We also propose an approach to implement this estimator under optimal windows, which ensures asymptotic normality in semiparametric multiple index models of arbitrary dimension. This mechanism further contributes to its very good finite sample performance.

The estimator has a recursive differencing structure, with the order of the bias depending on the stage of the recursion. At any stage, the bias depends on a kernel weighted difference in two bias terms. One term is the bias at a target point of interest, while the other is the bias conditioned on a data point. With kernel weighting ensuring that observations are close to the target point, the biases cancel up to a higher order due to this differencing structure. In this manner, we are able to reduce the bias to any order depending on the stage of the recursion.

To develop this nonparametric expectations estimator and examine its properties, Section 2 defines the estimator and provides the intuition for its theoretical properties. Section 3 obtains the large sample properties of the

proposed estimator and also derives asymptotic normality results for a finite dimension parameter vector in a large class of semiparametric models based on this estimator. Employing a result on semiparametric derivatives, we show that it is possible to obtain normality using optimal windows. If this result is not employed, then undersmoothing is required and we provide the required conditions. Section 4 provides Monte Carlo results that demonstrate that the estimator has very good finite sample properties in triple index models, exhibiting a substantial improvement over both regular and higher order kernels. Section 5 contains our conclusions. The Appendix contains proofs of all theorems and supporting intermediate lemmas.

2 The Estimator

To motivate the form of the bias reduction, we defer discussions of trimming considerations to the next section and consider the model in localized form for the s^{th} observation on the dependent variable Y_s

$$Y_s = M(w) + [M(W_s) - M(w)] + \varepsilon_s,$$

where (W_s) is a vector of continuous and discrete explanatory variables and ε_s is an error with $E(\varepsilon_s|W_s) = 0$. Define $K_s(w)$ as a kernel function that controls the localization error, $[M(W_s) - M(w)]$, by downweighting observations W_s not close to w .¹ The familiar initial nonparametric estimator, which we term as the stage 1 estimator, is given as:

$$\hat{M}_1(w) \equiv \frac{\frac{1}{N} \sum_s Y_s K_s(w)}{\hat{g}(w)}, \quad \hat{g}(w) \equiv \frac{1}{N} \sum_s K_s(w)$$

Substituting the local model for Y_s :

$$\begin{aligned} \hat{M}_1(w) &= M(w) + \frac{\frac{1}{N} \sum_s [M(W_s) - M(w) + \varepsilon_s] K_s(w)}{\hat{g}(w)} \Rightarrow \\ \Delta_1(w) &\equiv \hat{g}(w) \left[\hat{M}_1(w) - M(w) \right] = \frac{1}{N} \sum_s [M(W_s) - M(w) + \varepsilon_s] K_s(w) \end{aligned}$$

¹As defined below, the kernel function, K , will have an indicator component on support points for discrete variables.

As the localization error, $M(W_s) - M(w)$, is responsible for the bias, remove an estimate of it to obtain:

$$\begin{aligned}\hat{M}_2(w) &\equiv \frac{\frac{1}{N} \sum_s \left[Y_s - (\hat{M}_1(W_s) - \hat{M}_1(w)) \right] K_s(w)}{\hat{g}(w)} \\ &= M(w) + \frac{\frac{1}{N} \sum_s \left[(M(W_s) - M(w)) - (\hat{M}_1(W_s) - \hat{M}_1(w)) + \varepsilon_s \right] K_s(w)}{\hat{g}(w)} \Rightarrow \\ \Delta_2(w) &\equiv \hat{g}(w) \left[\hat{M}_2(w) - M(w) \right] = \frac{1}{N} \sum_s \left[\frac{\Delta_1(w)}{\hat{g}(w)} - \frac{\Delta_1(W_s)}{\hat{g}(W_s)} + \varepsilon_s \right] K_s(w)\end{aligned}$$

We will prove that the second stage estimator has "better" bias and convergence properties than the first stage. Accordingly, it would seem desirable to use the second stage estimator to remove a "better" estimate of the bias. Continuing in this manner, for $k > 1$, define the stage k estimator as:

$$\begin{aligned}\hat{M}_k(w) &\equiv \frac{\frac{1}{N} \sum_s \left[Y_s - (\hat{M}_{k-1}(W_s) - \hat{M}_{k-1}(w)) \right] K_s(w)}{\hat{g}(w)} \Rightarrow \\ \Delta_k(w) &\equiv \hat{g}(w) \left[\hat{M}_k(w) - M(w) \right] = \frac{1}{N} \sum_s \left\{ \frac{\Delta_{k-1}(w)}{\hat{g}(w)} - \frac{\Delta_{k-1}(W_s)}{\hat{g}(W_s)} + \varepsilon_s \right\} K_s(w)\end{aligned}$$

When $k = 1$, it is well known that subject to regularity conditions the bias is $O(h^2)$. To intuitively explain why the bias declines with each stage, with $k > 1$ consider the expectation of a term similar to that above with the true g in place of its estimator:

$$\frac{1}{N} \sum_s E \left\{ \left[E \left(\frac{\Delta_{k-1}(w)}{g(w)} \middle| W_s \right) - E \left(\frac{\Delta_{k-1}(W_s)}{g(W_s)} \middle| W_s \right) \right] K_s(w) \right\}$$

It can be shown that for $k > 1$:

$$\begin{aligned}E \left(\frac{\Delta_{k-1}(w)}{g(w)} \middle| W_s \right) &= h^{2(k-1)} B(w) + O(h^{2k}) + o(N^{-=1/2}) \\ E \left(\frac{\Delta_{k-1}(W_s)}{g(W_s)} \middle| W_s \right) &= h^{2(k-1)} B(W_s) + O(h^{2k}),\end{aligned}$$

where $B(\cdot)$ is a bounded function. The $o(N^{-=1/2})$ factor in the first expectation arises from a component of $\Delta_{k-1}(w)$ that depends on W_s . For $k-1 = 1$,

the above orders are those for regular kernels. With the kernel ensuring that $B(w)$ is close to $B(W_s)$, we are able to show that the bias at stage k is smaller than that at stage $k - 1$. The more general case, which we consider in the Appendix, requires us to analyze expressions containing density estimators rather than true densities. While the analysis is substantially more complicated, it is still the case that the recursive differencing structure of the estimator results in the bias declining at each stage.

The nonparametric expectation estimator discussed above can also be used in index models to estimate index parameters. In this context, in addition to the recursive differencing structure that provides bias reduction, we also propose an extra mechanism that further reduces the bias in estimating these parameters. To describe this additional mechanism, note that in a wide class of extremum estimators, a critical step in the asymptotic normality argument requires that the gradient to the objective function of interest have a low order bias. Typically, the gradient contains a multiplicative component that is the derivative of a semiparametric expectation with respect to a finite dimensional parameter vector. As discussed in the next section, Newey has shown that this derivative behaves as a residual in that its expectation conditioned on the true index is zero. This residual property can be exploited as an additional bias control. As a result, using the recursive differencing estimator, normality can be obtained with optimal windows for multiple indices.²

Theorems 1 and 2 in the next section provide the properties of the expectations estimator in both nonparametric and semiparametric contexts. For a wide class of semiparametric models, Theorem 3 provides conditions on the number of stages and the kernel window to obtain asymptotic normality in two cases: a) when an additional bias reduction mechanism is combined with recursive differencing and b) when only recursive differencing is used as the bias reduction method. To obtain \sqrt{N} -asymptotic normality, in case a), optimal windows suffice. However, in case b), undersmoothing is required for the normality result. Theorem 3 provides sufficient conditions on the stage and window parameter to ensure that these results hold.

²See Klein and Shen(2010) for Newey's proof of this result. This paper exploits this result to obtain asymptotic normality in single index models. With recursive differencing, this result extends to multiple index structures.

3 Large Sample Results

To establish large sample results, we require definitions and notation which we introduce below.

3.1 Definitions and Notation

D1) Conditional Expectations. Let $g(x, z)$ be the joint density for (X, Z) , where X and Z are vectors of continuous and discrete random variables respectively, where Z has a finite number of support points. With $W \equiv (X, Z)$ and $w \equiv (x, z)$ a fixed vector, define:

$$M(w) \equiv E(Y|W = w).$$

D2) Trimming. Let $W \equiv (X, Z)$ be i.i.d. from $g(x, z)$ in D1). With λ as a percentile, let $q_x(\lambda)$ be the corresponding population quantile for X . Let $q_x \equiv [q_x(\lambda_1), q_x(\lambda_2)]$. With

$$0 < \lambda_1 < \lambda_1^I < \lambda_2^I < \lambda_2 < 1,$$

let $q_x^I \equiv [q_x(\lambda_1^I), q_x(\lambda_2^I)]$. Then, define exterior and interior trimming sets:

$$\begin{aligned} \mathcal{C}_x(q_x) &\equiv \{x : q_x(\lambda_1) < x < q_x(\lambda_2)\} \\ \mathcal{C}_x(q_x^I) &\equiv \{x : q_x(\lambda_1^I) < x < q_x(\lambda_2^I)\} \end{aligned}$$

D3) Kernel. Referring to D2), let \hat{q} be a sample quantile vector corresponding to q_x . Define:

$$\tau(X_s, \hat{q}) \equiv \begin{cases} 1 & k = 1 \\ 1 \{X_s \in \mathcal{C}_x(\hat{q})\} & k > 1 \end{cases}$$

Let x_l and X_{ls} be the l^{th} components of the vectors x and X_s respectively. Then, with d as the dimension of X , define the kernel function:

$$\begin{aligned} k(w, W_s) &\equiv 1 \{Z_s = z\} \prod_{l=1}^d \frac{1}{s_l h} \phi\left(\frac{x_l - X_{ls}}{s_l h}\right) \\ K_s(w) &\equiv \tau(X_s, \hat{q}) k(w, W_s) \equiv \frac{1}{h^d} f(w, W_s) \end{aligned}$$

where $h = O(N^{-r})$, s_l is a constant,³ and $\phi(z)$ is a density symmetric about 0.

D4) Density Estimator. For $x \in \mathcal{C}_x(q_x^I)$ in D2) and $g(w)$ as the density for W , define the estimator for $g(w)$ as:

$$\hat{g}(w) \equiv \frac{1}{N} \sum_{s=1}^N K_s(w).$$

D5) Nonparametric Expectation Estimator. Define:

$$\hat{M}_k(w) \equiv \begin{cases} \frac{\frac{1}{N} \sum_s Y_s K_s(w)}{\hat{g}(w)} & k = 1 \\ \frac{\frac{1}{N} \sum_s [Y_s - (\hat{M}_{k-1}(W_s) - \hat{M}_{k-1}(w))] K_s(w)}{\hat{g}(w)} & k > 1 \end{cases}$$

$$\Delta_k(w) \equiv \hat{g}(w) [\hat{M}_k(w) - M(w)]$$

D6) Index Functions. Let θ be a finite dimensional parameter vector and $V(W_s; \theta)$ a $d \times 1$ vector of functions. For fixed w , define $v(\theta) \equiv V(w; \theta)$.

D7) Semiparametric Expectation. Define

$$M(v(w; \theta); \theta) \equiv E(Y | V(W_s; \theta) = v(\theta))$$

D8) Estimated Semiparametric Expectation. With the functions $k(\bullet)$ and $\tau(\bullet)$ given in D3), define:

$$K_s(v(\theta); \theta) \equiv \tau(X_s, \hat{q}) k(v(\theta), V(W_s; \theta)), \quad \hat{g}(v(\theta); \theta) \equiv \frac{1}{N} \sum_s K_s(v(\theta); \theta)$$

For $k = 1$:

$$\hat{g}(v(\theta); \theta) \hat{M}_1(v(\theta); \theta) \equiv \frac{1}{N} \sum_s Y_s K_s(v(\theta); \theta)$$

For $k > 1$:

³Note that for each l , s_l can be replaced by an estimate (e.g. a sample standard deviation) that converges in probability to a constant. With our main focus being on bias reduction, the proofs treat s_l as constant.

$$\hat{g}(v(\theta); \theta) \hat{M}_k(v(\theta); \theta) \equiv \frac{1}{N} \sum_s \left\{ Y_s - \left[\begin{array}{c} \hat{M}_{k-1}(V(W_s; \theta); \theta) \\ \hat{M}_{k-1}(v(\theta); \theta) \end{array} \right] \right\} K_s(v(\theta); \theta),$$

$$\Delta_k(v(\theta); \theta) \equiv \hat{g}(v(\theta); \theta) \left[\hat{M}_k(v(\theta); \theta) - M(v(\theta); \theta) \right]$$

To obtain convergence properties for the proposed nonparametric expectation estimator and asymptotic normality for a class of estimators we make the following assumptions.

3.2 Assumptions

A1) The vector $\{Y_s, X_s, Z_s\}$ is i.i.d. over $i = 1, \dots, N$, and takes on values in $\mathfrak{X}_Y \times \mathfrak{X}_X \times \mathfrak{X}_Z \subset \mathfrak{R}^{1+d+d^*}$. The vector X_s is continuous and the discrete vector Z_s has a finite number of support points.

A2) Define $\nabla_x^\alpha(f)$ as the α^{th} derivative of the function f with respect to x , where $\nabla_x^0(f(x)) \equiv f(x)$. Let $g(x|y, z)$ be the conditional density for X conditioned on $Y = y$ and $Z = z$. Then, there exists a constant $c > 0$, for each $x \in \mathcal{C}_x(q_x^I)$ in D2), and $\alpha = 0, 1, \dots, 2k$ such that:

$$\inf_{x \in \mathcal{C}_x(q_x^I)} [g(x|y, z)] > c > 0, \quad |\nabla_x^\alpha g(x|y, z)| = O(1), \quad |\nabla_x^\alpha M(x)| = O(1)$$

A3) For index models, refer to D6) and assume that $\theta \in \Phi$, a compact set. With $V(W_s; \theta)$ as the index vector and $v(w; \theta)$ a fixed index value, interpret $g(v(w; \theta)|y; \theta)$ as the density for $V(W_s; \theta)$ evaluated at $v(w; \theta)$ and conditioned on $Y = y$. Assume that

$$\left| \nabla_v^\alpha g(v(w; \theta)|y; \theta) \right|, \quad \left| \nabla_\theta^\beta g(v(w; \theta)|y; \theta) \right|, \quad \left| \nabla_\theta^\beta v(w; \theta) \right|$$

are all $O(1)$ for all $x \in \mathcal{C}_x(q_x^I)$, z, y , and $\alpha, \beta = 0, 1, \dots, 2k$. Further, for constant $c > 0$:

$$\inf_{x \in \mathcal{C}_x(q_x^I), \theta \in \Phi} [g(v(w; \theta)|y; \theta)] > c$$

A4) The kernel function ϕ in D3) and D8) has window parameter $r < \frac{1}{4d}$. Further, in the semiparametric case (see D8):

$$\phi'(u)/\phi(u), \phi''(u)/\phi(u) = O(1) \text{ for } u = O(1).$$

A5) There exists $m > 4$ such that $E [|Y_s|^m] = O(1)$.

A6) For semiparametric index models:

$$E(Y|W = w) = E (Y|V (W; \theta_0) = v (w, \theta_0))$$

A7) With $r^* = \frac{1}{4k+d}$ as the optimal window and $\varepsilon > 0$ sufficiently small, set k and r to satisfy a) or b) below:

$$\begin{aligned} a) & : k \geq \frac{d+3}{2}, r = r^*. \\ b) & : k \geq \frac{3d+3}{4}, \frac{1}{4k} < r < \frac{(1+\varepsilon)}{3(d+1)}. \end{aligned}$$

3.3 Main Theorems

We begin in Theorem 1 with the convergence properties for the proposed nonparametric expectation estimator.

Theorem 1. Assume A1)-A5). Then, for $x \in \mathcal{C}_x (q_x^I)$ in D2), there exists $\hat{M}_k^*(w)$ such that:

$$\begin{aligned} a) & : \sup_w \left| \hat{g}(w) \left(\hat{M}_k^*(w) - \hat{M}_k(w) \right) \right| = o_p (N^{-1/2}) \\ b) & : \sup_w \left| E \left[\hat{g}(w) \left(\hat{M}_k^*(w) - M(w) \right) \right] \right| = O(h^{2k}) + o(N^{-1/2}) \\ c) & : \sup_w \left| \hat{M}_k(w) - M(w) \right| = O_p(h^{2k}) + O_p \left(N^{-\left(\frac{1}{2}-rd-\frac{1}{m+2}\right)} \right) \end{aligned}$$

For (a-b), the non-linear structure of \hat{M} makes it infeasible to directly study its expectation. However, in using this theory to establish asymptotic normality in econometric models, the form in (a-b) will suffice (see Theorem 3 below). We have written the order for b) in this form because h^{2k} may or may not be smaller than $N^{-1/2}$.

For the semiparametric case, Theorem 2 below provides results similar to those in the nonparametric case in Theorem 1.

Theorem 2. Assume A5-A6). Then, for $x \in \mathcal{C}_x (q_x^I)$ in D2), there exists $\hat{M}_k (v; \theta)$ such that:

$$\begin{aligned}
a) & : \sup_{v, \theta} \left| \hat{g} (v, \theta) \left(\hat{M}_k^* (v; \theta) - \hat{M}_k (v; \theta) \right) \right| = o_p (N^{-1/2}) \\
b) & : \sup_v \left| E \left[\hat{g} (v; \theta_0) \left(\hat{M}_k^* (v; \theta_0) - M (v; \theta_0) \right) \right] \right| = O(h^{2k}) + o(N^{-1/2}) \\
c) & : \sup_v \left| \hat{M}_k (v; \theta_0) - M (v(\theta_0); \theta_0) \right| = O_p(h^{2k}) + O_p(N^{-(\frac{1}{2}-rd-\frac{1}{m+2})})
\end{aligned}$$

For estimating semiparametric models, we will require conditions on the stage k and the window parameter r . For this purpose, we will characterize a wide class of estimators for which we will provide sufficient conditions on the stage and window parameter to obtain \sqrt{N} -asymptotic normality. This class will employ the recursive differencing estimator for conditional expectations discussed earlier. In addition to recursive differencing, it is also possible to exploit a residual property of semiparametric expectations in index models due to Whitney Newey (see Klein and Shen (2010) for a complete statement and proof). In so doing, we will find that we are able to employ optimal windows in the differencing estimator. Under this result :

$$\begin{aligned}
\nabla_{\theta} [M_k (V (W; \theta); \theta)]_{\theta_0} & = \left\{ \begin{array}{l} \nabla_{\theta} [M_k (V (W; \theta); \theta)]_{\theta_0} - \\ E (\nabla_{\theta} [M_k (V (W; \theta); \theta)]_{\theta_0} | V (W; \theta_0)) \end{array} \right\} \\
& \Rightarrow E \left\{ \nabla_{\theta} [M_k (V (W; \theta); \theta)]_{\theta_0} | V (W; \theta_0) \right\} = 0
\end{aligned}$$

In many extremum problems, the gradient to the objective function multiplicatively depends on an estimator for the derivative above weighted by a trimming function and other factors. If trimming is based (asymptotically) on the true index, it is possible to exploit the above result. Below, we provide a definition for a class of estimators that assumes it is possible (asymptotically) to trim on the index. Following this definition, we will discuss existing estimators in the literature that employ a 2-step approach to ensure that trimming is (asymptotically) based on the index in the second step.

A Class of Estimators. With λ - percentiles given in D2), let $q_v(\lambda)$ be the population quantile for the index $V_{i0} \equiv V(W; \theta_0)$. As in D2) set $0 < \lambda_1 < \lambda_1^I < \lambda_2^I < \lambda_2$. Letting $q_v \equiv [q_v(\lambda_1), q_v(\lambda_2)]$ and $q_v^I \equiv$

$[q_v(\lambda_1^I), q_v(\lambda_2^I)]$. Define trimming sets:

$$\begin{aligned}\mathcal{C}_v(q_v) &\equiv \{v : q_v(\lambda_1) < v < q_v(\lambda_2)\} \\ \mathcal{C}_v(q_v^I) &\equiv \{v : q_v(\lambda_1^I) < v < q_v(\lambda_2^I)\}\end{aligned}$$

With $S_i = X_i$ or $V_{0i} \equiv V(W_i, \theta_0)$, define:

$$\bar{\tau}(S_i) \equiv \begin{cases} 1, & k = 1 \\ 1 \{X_i \in \mathcal{C}_x(q_x)\} & k > 1, S_i = X_i \\ 1 \{V_i \in \mathcal{C}_v(q_v)\} & k > 1, S_i = V_i \end{cases}$$

Define $\bar{\tau}_I(S_i)$ by replacing q_x with q_x^I and q_v with q_v^I in the above definition for $\bar{\tau}(S_i)$. Referring to D8), define $\hat{M}_k(V_{i0}, \bar{\tau}(S_i); \theta_0)$ by replacing the trimming function within $\hat{M}_k(V_{i0}; \theta_0)$ by $\bar{\tau}_i(S_i)$. Let:

$$\hat{u}_i \equiv Y_i - \hat{M}_k(V_{i0}, \bar{\tau}(S_i); \theta_0)$$

Then, with $\alpha_i \equiv \alpha(V(W_i; \theta_0))$ a function of the true index and $V_i \equiv V(W; \theta_0)$ define \mathcal{C} as the linear class of estimators:

$$\begin{aligned}\sqrt{N}(\hat{\theta} - \theta_0) &= \hat{A}(\theta^+) \sqrt{N} \hat{G}(\theta_0), \theta^+ \in [\hat{\theta}, \theta_0] \\ \hat{G}(\theta_0) &\equiv \frac{1}{N} \sum_{i=1}^N \bar{\tau}_I(S_i) \hat{u}_i \nabla_{\theta}^1 \left[\hat{M}_k(V_i, \bar{\tau}(S_i); \theta) \right]_{\theta_0} \alpha_i\end{aligned}$$

Further, with $D \equiv \nabla_{\theta}^2 \left[\hat{M}_k(V_i, \bar{\tau}(S_i); \theta) - M(V(W_i; \theta); \theta) \right]$ and $A_0(\theta)$ a constant matrix, the class satisfies:

$$\begin{aligned}i) &: \sup_{\theta} \left| \hat{A}(\theta) - A_0(\theta) \right| = O_p \left(\bar{\tau}_I(S_i) \sup_{\theta} \nabla_{\theta}^2 [D] \right) \\ ii) &: \hat{\theta} \xrightarrow{p} \theta_0.^4\end{aligned}$$

Before proceeding to provide a theorem that provides conditions on the stage and window parameter for asymptotic normality, it is important to note several features of this class. First, it includes multiple index Semiparametric Least-Squares estimators (see Ichimura(1993) and Ichimura and Lee(1991)) with $\hat{\lambda}_i = \lambda_i = 1$. It also includes Quasi-Maximum-Likelihood estimators for semiparametric binary response (see Klein and Spady(1993)), with

$$\alpha_i \equiv \frac{1}{M(V_i(\theta_0); \theta_0) [1 - M(V_i(\theta_0); \theta_0)]}$$

This class can also be easily extended to semiparametric ordered models by defining categorical gradients and weights. It also extends to a system of joint binary models.

As a second remark, the class in the definition also includes estimators for which (asymptotically) trimming is based on the true index. With the index being unknown, in the single index case Klein and Shen (2010) provide a 2 step SLS estimator with X-trimming at the first step and index trimming at the second.⁵ A similar strategy is employed by Klein, Shen, and Vella (2014) for joint binary models.

Third, in characterizing this class, we have taken trimming ($\bar{\tau}$) and weighting (α) functions as known. For the trimming function within the estimator of M, Lemma 2 can be applied to take this function as known. For the trimming function outside of M, Lemma 1.18 of Pakes and Pollard (1989) can be employed to show that it can be taken as given. The analysis for taking the weighting function as known is similar to that employed in the proof of Theorem 3 for taking the estimated derivative of the M-function as known.

Theorem 3 below provides conditions on the stage and window parameter that ensure asymptotic normality for the above class of estimators. When Newey's result is exploited, these conditions permit an optimal choice for the window parameter.

Theorem 3 Asymptotic Normality. Assume A2)-A5) and let $u_i \equiv Y_i - M_{ik}(\theta_0)$. With g_{0i} as the density for $V(W_i; \theta_0)$ and employing the notation

⁵With k set in accordance with Theorem 3 and r set optimally, in the first step, trim on X and employ the SLS estimator. Employ the parameter estimator from the first step to construct an estimated index. The asymptotic form for the gradient in the definition follows from trimming on the estimated index with an adjusted estimator for the M-function. In the adjustment, which is needed for a uniform convergence argument underlying consistency, the density denominator \hat{g} is replaced by $\hat{g} + \delta$, where δ is an adjustment factor that depends on the index evaluated at θ . Near the support boundaries for the index, δ vanishes very slowly away from θ_0 . At θ_0 , where trimming at the true index provides protection against small density denominators, δ is uniformly $o_p(N^{-1/2})$.

in the definition above:

$$\begin{aligned}
G_1(\theta_0) &\equiv \frac{1}{N} \sum_{i=1}^N G_{1i}(\theta_0), \quad G_{1i}(\theta_0) \equiv \bar{\tau}_I(S_i) u_i \nabla_{\theta}^1 [M_{ik}(\theta)]_{\theta_0} \alpha_i \\
G_2(\theta_0) &\equiv \frac{1}{N} \sum_{i=1}^N G_{2i}(\theta_0), \\
G_{2i}(\theta_0) &\equiv \bar{\tau}_I(S_i) \frac{\hat{g}_{0i} \left(\hat{M}_k^*(V_{0i}, S_i; \theta_0) - M_i(\theta_0) \right)}{g_{0i}} \nabla_{\theta}^1 \left[\hat{M}_k^*(V_{0i}, S_i; \theta) \right]_{\theta_0} \lambda_i
\end{aligned}$$

Then, for the class of estimators defined in \mathfrak{E} :

$$a) : \sqrt{N} \left(\hat{\theta} - \theta_0 \right) = A_0 \sqrt{N} [G_1(\theta_0) - G_2(\theta_0)] + o_p(1),$$

If $S_i \equiv V_{0i}$, then for the optimal window and stage in A7a):

$$b) : \sqrt{N} \left(\hat{\theta} - \theta_0 \right) \xrightarrow{d} Z_1 \sim N(0, A_0 E [G_{1i}(\theta_0) G_{1i}(\theta_0)'] A_0'),$$

Assume the window parameter and stage satisfy A7b) and define

$$\gamma_i \equiv \tau(X_i, q_x^I) \nabla_{\theta}^1 [M(V_i(\theta); \theta)]_{\theta_0} \alpha_i \quad (1)$$

$$G^*(\theta_0) \equiv \frac{1}{N} \sum_{i=1}^N G_i^*(\theta_0), \quad G_i^*(\theta_0) \equiv u_i [\gamma_i - E(\gamma_i | V_i(\theta_0))] \quad (2)$$

Then, if $S_i \equiv X_i$:

$$c) : \sqrt{N} \left(\hat{\theta} - \theta_0 \right) \xrightarrow{d} Z^* \sim N(0, A_0 E (G_i^*(\theta_0) G_i^*(\theta_0)') A_0')$$

4 Monte Carlo Results

We conducted Monte Carlo experiments using quadratic and cubic designs. In both designs, we constructed three indices: $V_1 = X_1 + X_4$, $V_2 = X_2 - X_4$, $V_3 = X_3 + X_4$ where X_1, X_2 and X_3 follow standard normal distribution. We also generated an error term ε that follows a standard normal distribution. For the quadratic design the outcome is given as:

$$Y = \alpha V_1 + \beta V_2^2 + \gamma V_1 V_3 + \varepsilon;$$

while for the cubic design

$$Y = \alpha V_1 + \beta V_2^3 + \gamma V_1 V_3 + \varepsilon.$$

The α, β, γ are standardizing constants selected so that in both models, each of the three explanatory components has approximate standard deviation of one.

We examine several different estimators. For the proposed differencing estimator, we provide estimation results for stage 1 (regular kernel), stage 2, and stage 3 estimators. We also report results for a twicing kernel with bias $O(h^6)$ as required for asymptotic normality (see Newey et. al (2004)).

In Table 1, we compare the differencing estimator with the higher order kernel. Due to outlier problems for the higher order kernel estimator, we report robust measures of performance. In both designs, the medians for both estimators are reasonably close to the truth, with the differencing estimator being marginally better. However, the median absolute deviation (MAD) and the median absolute error (MAE) are on the order of 5 times smaller for the differencing estimator.

In Table 2, we report results for each of three stages of the differencing estimator. In practice, it is common for regular kernels to perform better than bias-reducing kernels. However, with the first stage being the regular kernel case, for both quadratic and cubic designs, the RMSE for the sequential differencing estimator declines over the stages, with the third stage estimator being required for asymptotic normality.

Table 1. Monte Carlo results comparing proposed differencing estimator with second order twicing kernel

	Proposed estimator	Second order twicing kernel
Quadratic Design		
Median	0.99	1.07
	-0.96	-0.94
	0.98	1.02
MAD	0.04	0.20
	0.04	0.26
	0.06	0.42
MAE	0.04	0.19
	0.05	0.25
	0.07	0.44
Cubic Design		
Median	0.93	1.04
	-0.88	-1.00
	0.99	0.97
MAD	0.01	0.20
	0.02	0.27
	0.02	0.24
MAE	0.07	0.18
	0.12	0.28
	0.02	0.24

Table 2. Monte Carlo results comparing different stages of the proposed differencing estimator

	1st Stage	2nd Stage	3rd Stage
Quadratic Design			
Mean	0.96	0.97	0.99
	-0.99	-0.98	-0.96
	1.01	1.00	0.99
SD	0.06	0.06	0.05
	0.08	0.07	0.06
	0.13	0.11	0.09
RMSE	0.07	0.07	0.05
	0.08	0.08	0.07
	0.13	0.11	0.09
Cubic Design			
Mean	0.88	0.89	0.93
	-0.86	-0.85	-0.88
	1.02	1.01	0.99
SD	0.02	0.02	0.02
	0.04	0.04	0.03
	0.04	0.04	0.03
RMSE	0.12	0.11	0.07
	0.15	0.16	0.12
	0.05	0.04	0.03

5 Conclusions

In this paper, we propose a nonparametric recursive differencing estimator for conditional expectations. Depending on the stage of the recursion, it is possible to obtain any order for the bias without any change in the variance. Theorems 1 and 2 provide these results for both nonparametric and semiparametric models.

While higher order kernels share the above properties, they differ from the proposed estimator in two important respects. First, the proposed estimator not only performs much better than higher kernels in monte-carlo studies, but it also dominates regular kernels. Accordingly, the proposed estimator is both theoretically valid and performs quite well at moderate sample sizes. Second, in estimating index models, we show that with recursive differencing it is possible to exploit a "residual" property of semiparametric derivatives. In so doing, we obtain asymptotic normality without undersmoothing, regardless of the dimension of the index vector. This theoretical property which is obtained in Theorem 3, may contribute to the very good finite sample performance of the proposed estimator. As there are estimators for which Newey's residual result does not hold (e.g. Klein and Vella (2010)), Theorem 3 shows that asymptotic normality can be obtained with recursive differencing as the sole bias control.

References

- [1] Bhattacharya, P.K. (1967): "Estimation of a Probability Density Function and its Derivatives," *The Indian Journal of Statistics, Series 4*, v. 29, 373-382.
- [2] Ichimura, H. (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single index models," *Journal of Econometrics*, 58, 71-120.
- [3] Ichimura, H., and L. F. Lee (1991): "Semiparametric least squares (SLS) and weighted SLS estimation of multiple index models: Single equation estimation," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. W.Barnett, J.Powell and G.Tauchen, Cambridge University Press.
- [4] Klein, R. (1993), Specification tests for binary choice models based on index quantiles, *Journal of Econometrics* 59, 343-375.
- [5] Klein, R. and C. Shen (2010): "Bias Corrections in Testing and Estimating Semiparametric, Single Index Models," *Econometric Theory*, 1683-1718.
- [6] Klein R, Shen C, Vella F. Estimation of marginal effects in semiparametric selection models with binary outcomes. *Journal of Econometrics* 185(1):82-94, 3/2015.
- [7] Klein, R. and R. Spady (1993): "An Efficient Semiparametric Estimator for the Binary Response Model," *Econometrica*, 61, 387-421.
- [8] Klein, R. and F. Vella (2010): "Estimating a Class of Triangular Simultaneous Equations Models Without Exclusion Restrictions," *Journal of Econometrics* 154 (2010) 154–164.
- [9] Newey, W., F. Hsieh, and J. Robins (2004): "Twicing Kernels and a Small Bias Property of Semiparametric Estimators," *Econometrica*, 72, 947-962.
- [10] Pakes, A., and D.Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1058

6 Appendix

When the proposed estimator is employed in estimating semiparametric index models, we require two types of trimming. We refer to these two types of trimming as interior and exterior, with interior trimming based on a subset of that for exterior trimming. With exterior trimming, we trim to insure that the recursive bias adjustments have desirable properties. We employ interior trimming to control the target observation at which the recursive differencing estimator is evaluated.

In addition, we require external trimming to control the target observation at which the estimator is evaluated. Lemma 3 takes exterior trimming as known and provides conditions under which interior trimming can be taken as known. Lemma 1.17 in Pakes and Pollard (1997) provides the basis for taking interior trimming as known. Based on these results, all remaining lemmas and theorems are then proved on the basis of known trimming.

6.1 Proofs of Main Theorems

$$\begin{aligned}
 a) & : \sup_w \left| \hat{g}(w) \left(\hat{M}_k^*(w) - \hat{M}_k(w) \right) \right| = o_p \left(N^{-1/2} \right) \\
 b) & : \sup_w \left| E \left[\hat{g}(w) \left(\hat{M}_k^*(w) - M(w) \right) \right] \right| = O(h^{2k}) + o(N^{-1/2}) \\
 c) & : \sup_w \left| \hat{M}_k(w) - M(w) \right| = O_p(h^{2k}) + O_p \left(N^{-\left(\frac{1}{2} - rd - \frac{1}{m+2}\right)} \right)
 \end{aligned}$$

Proof of Theorem 1. With $\Delta^*(w)$ defined as in Lemma 3 and $\hat{M}_k^*(w) \equiv \frac{\Delta_{k-1}^*(w)}{g(w)}$, part a) follows from Lemma 3. Part b) follows from Lemma 7. For part c) from the uniform convergence $\hat{g}(v; \theta)$ in Lemma 10, with g uniformly bounded away from zero for w bounded (Assumption A3), and Part a) of Theorem 1, it suffices to consider:

$$\hat{g}(w) \left[\hat{M}_k^*(w) - M(w) \right] \tag{3}$$

For $k = 1$ from Lemma 10 and the uniform bias result in part b), the uniform convergence rate is that given. Assume that the rate holds for stage $k - 1$ and

define $E_{k-1}^*(w)$ as in Lemma 5. Employing the same decomposition as in Lemma 7, an upper bound for the term in (3) is given as:

$$\sup_w \left| \frac{\langle \delta(w) \rangle_{\mathbb{P}}}{g(w)} \right| \left| \frac{1}{N} \sum_s \left[\frac{\Delta_{k-1}^*(w)}{g(w)} - E_{k-1}^*(w) \right] K_s(w) \right| \quad (4)$$

$$+ \sup_w \left| \frac{\langle \delta(w) \rangle_{\mathbb{P}}}{g(w)} \right| \left| \frac{1}{N} \sum_s \left[\frac{\Delta_{k-1}^*(W_s)}{g(W_s)} - E_{k-1}^*(W_s) \right] K_s(w) \right| \quad (5)$$

$$+ \left| \frac{1}{N} \sum_s \left[E_{k-1}^*(w) \frac{\langle \delta(w) \rangle_{\mathbb{P}}}{g(w)} - E_{k-1}^*(W_s) \frac{\langle \delta(W_s) \rangle_{\mathbb{P}}}{g(W_s)} \right] K_s(w) \right| \quad (6)$$

$$+ \left| \frac{1}{N} \sum_s \varepsilon_s K_s(w) \right| \quad (7)$$

An induction argument then establishes the required convergence rates for (4) and (5). The required convergence rate for the term in (6) readily follows from the uniform bias result in part b). The remaining term in (7) is uniformly $O_p \left(N^{-\left(\frac{1}{2} - rd - \frac{1}{m+2}\right)} \right)$ under an argument similar to that for part a) of Lemma 10.

Proof of Theorem 2. With $\Delta_k^*(v, \theta)$ having the same form as $\Delta_k^*(w)$ in Lemma 3 and $\hat{M}_k^*(v, \theta) \equiv \frac{\Delta_{k-1}^*(v, \theta)}{g(v)}$, the proofs for parts a)-c) are identical to a-c) in Theorem 1 for the nonparametric case.

Proof of Theorem 3. Beginning with the characterization in a), we must establish the uniform convergence of $\hat{A}(\theta)$ to $A_0(\theta)$, which will follow from the uniform convergence of $\nabla_{\theta}^{\alpha} \left[\hat{M}_{ik}(\theta) - M_i(\theta) \right]$, $\alpha = 0, 1, 2$, and the consistency of $\hat{\theta}$. From Lemma 11, with the convergence rate being slowest for $\alpha = 2$, convergence follows for $r > 0$ and

$$\frac{1}{2} - r(d+2) - \frac{1}{m+2} > 0 \Leftrightarrow r < \frac{1}{2(d+2)} \frac{m}{m+2}, \quad (8)$$

Since $m > 4$, from A5) : $4/6 < m/(m+2) < 1$. There then exists $\varepsilon > 0$ such that

$$\frac{4}{6}(1 + \varepsilon) < \frac{m}{m+2}$$

Therefore, (8) is satisfied with

$$r < \frac{1}{2(d+2)} \frac{4}{6} (1 + \varepsilon) = \frac{(1 + \varepsilon)}{3(d+2)} \quad (9)$$

Proceeding to the "Gradient" component, recall that $\sqrt{N}\hat{G}(\theta_0)$ is given as:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N [Y_i - M_i(\theta_0)] \tau_i \nabla_{\theta}^1 [\hat{M}_{ik}(\theta)]_{\theta_0} \alpha_i - \quad (10)$$

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N [\hat{M}_{ik}(\theta_0) - M_i(\theta_0)] \tau_i \nabla_{\theta}^1 [\hat{M}_{ik}(\theta)]_{\theta_0} \alpha_i. \quad (11)$$

For the term in (10), with $\varepsilon_i \equiv Y_i - M_i(\theta_0)$ we must show

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \tau_i \varepsilon_i \alpha_i \left[\nabla_{\theta}^1 [\hat{M}_{ik}(\theta)]_{\theta_0} - \nabla_{\theta}^1 [M_i(\theta)]_{\theta_0} \right] = o_p(1),$$

which follows from an extension of a mean-square convergence argument in Klein and Shen(2010). From Cauchy Schwarz :

$$\begin{aligned} & \frac{1}{\sqrt{N}} \sum_{i=1}^N \tau_i \alpha_i \left| \hat{M}_{ik}(\theta_0) - M_i(\theta_0) \right| \left| \nabla_{\theta}^1 [\hat{M}_{ik}(\theta)]_{\theta_0} - \nabla_{\theta}^1 [M_{ik}(\theta)]_{\theta_0} \right| \\ & \leq \sqrt{N} \left\{ \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N \tau_i \alpha_i^2 [\hat{M}_{ik}(\theta_0) - M_i(\theta_0)]^2} \times \sqrt{\frac{1}{N} \sum_{i=1}^N \tau_i \left(\nabla_{\theta}^1 [\hat{M}_{ik}(\theta)]_{\theta_0} - \nabla_{\theta}^1 [M_{ik}(\theta)]_{\theta_0} \right)^2}}{\sqrt{\frac{1}{N} \sum_{i=1}^N \tau_i \left(\nabla_{\theta}^1 [\hat{M}_{ik}(\theta)]_{\theta_0} - \nabla_{\theta}^1 [M_{ik}(\theta)]_{\theta_0} \right)^2}} \right\} \end{aligned}$$

From Lemmas 8-9, in order for this expression to be $o_p(1)$, we must set r and k such that

$$\left\{ \begin{aligned} & \left[O_p(N^{-2rk}) + O_p\left(N^{-\frac{1}{2} + \frac{rd}{2}}\right) \right] \times \\ & \left[O_p(N^{-2rk}) + O_p\left(N^{-\frac{1}{2} + \frac{r(d+2)}{2}}\right) \right] \end{aligned} \right\} = o_p(N^{-1/2}) \quad (12)$$

The conditions in (12) are satisfied with

$$\frac{1}{8k} < r < \frac{1}{2(d+1)}, \quad K > \frac{d+2}{4} \quad (13)$$

For $\varepsilon > 0$ and sufficiently small, the conditions in (13) and (9) hold with

$$\frac{1}{8k} < r < \frac{(1 + \varepsilon)}{3(d + 2)}, K > \frac{d + 2}{4}, \quad (14)$$

which is satisfied under A7a or A7b). . Finally, to establish a), we must set r and k such that

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \tau_i \left[\hat{M}_{ik}(\theta_0) - M_i(\theta_0) \right] \frac{\hat{g}_{0t} - g_{0t}}{g_{0t}} \nabla_{\theta}^1 [M_{ik}(\theta)]_{\theta_0} \lambda_i \quad (15)$$

is $o_p(1)$. Employing Cauchy-Schwarz as above, Lemma 8, and known results for density estimators, the result follows.

For b), under A7a) for some $\varepsilon > 0$ and arbitrarily small:

$$\frac{1}{8k} < r < \frac{(1 + \varepsilon)}{3(d + 2)}, K \geq \frac{d + 3}{2} \quad (16)$$

To complete the proof, note first that this condition in (16) implies that in (14). Second, (16) is satisfied by the optimal window $r^* = \frac{1}{4k+d}$. Finally, since the semiparametric derivative can be taken as known from Part a) of Theorem 3, Newey's residual result can be used to show that the gradient component in the statement of Theorem 3 given in (1)-(2) is a degenerate U-statistic, from which b) immediately follows. Klein and Shen(2010) provide this argument in the single index case under regular kernels. This argument extends to multiple indices under the recursive estimator.

For c), note first that the condition in (A7b) implies that (8). When trimming is based in X , Newey's residual result does not directly apply and the low order bias must be obtained by under-smoothing and setting $r > \frac{1}{4k}$. From above, we now require:

$$\frac{1}{4k} < r < \frac{(1 + \varepsilon)}{3(d + 2)}$$

This interval will be non-empty with k satisfying the restriction in (A7b). The gradient is then a centered U-statistic, from which (c) follows.

6.2 Intermediate Lemmas and Proofs

As a first intermediate lemma, we require a standard result on bias expansions involving the expectation of the product of a function and a kernel. The following lemma summarizes a well-known result in the literature.

Lemma 1. Let $G(w)$ be a function such that $\nabla_v^{2i}G(w)$ is uniformly bounded for $i = 1, \dots, m + 1$.

$$E [G(W_r)K_r(w)] - G(w) = \sum_{i=1}^m h^{2i} \nabla_v^{2i} G(w) + O(h^{2m+2}),$$

The estimator for $M(w)$ requires trimming to control the bias adjustment that is evaluated at data points. The lemma below will enable us to take such trimming as known. The proof of this lemma is based on an adaptation of arguments in Klein and Shen (2010) and Klein (1993).

Lemma 2: For $0 < \delta < 1/2$, set the window parameter r , such that $r < \frac{1-2\delta}{2d}$. Denote $\hat{M}_k(w; \tau(\hat{q}_x))$ as the estimator in D3)-D5), where trimming is based on the estimated quantile vector \hat{q} . Let $\hat{M}_k(w; \tau(q_x))$ be the estimator for M when trimming is based on the true population quantile, q_x . Define $\hat{g}(w; \tau(q_x))$ and $\hat{g}(w; \tau(\hat{q}_x))$ as the density estimators under known and estimated trimming respectively. Let:

$$D(w) = \begin{Bmatrix} \hat{g}(w; \tau(q_x)) \left[\hat{M}_k(w; \tau(q_x)) - M_k(w) \right] - \\ \hat{g}(w; \tau(\hat{q}_x)) \left[\hat{M}_k(w; \tau(\hat{q}_x)) - M_k(w) \right] \end{Bmatrix}$$

Then,

$$\sup_w |D(w)| = o_p(N^{-1/2})$$

Proof. As $\tau(\hat{q}_x) = \tau(q_x) = 1$ for $k = 1$, we begin an induction argument with $k = 2$. Let $\tau^*(X_s, q_0, \hat{q}) \equiv 1 \{X_s \in [\mathcal{C}(q_0) \cup \mathcal{C}(\hat{q})]\}$ be a trimming function on the union of estimated and known trimming sets. Since $[\tau(X_s, \hat{q}_x) - \tau(X_s, q_x)] \tau^*(\bullet) = \tau(X_s, \hat{q}_x) - \tau(X_s, q_x)$, we may write:

$$D(w) = \frac{1}{N} \sum_s \left\{ \begin{array}{c} [\tau(X_s, \hat{q}_x) - \tau(X_s, q_x)] \tau^*(X_s, q_x, \hat{q}_x) \\ \Delta_{1M}(w) - \Delta_{1M}(W_s) \end{array} \right\} k(w, W_s),$$

$$\Delta_{1M}(t) \equiv \hat{M}_1(t) - M(t),$$

with $k(w, W_s)$ defined in D3). Then, with \mathfrak{N} an $o_p(1)$ neighborhood of q_x

$$\sup_w |D(w)| \leq 2 \left\{ \sup_{w, q \in \mathfrak{N}} \frac{1}{N} \sum_s |\tau(X_s, \hat{q}_x) - \tau(X_s, q_x)| \tau^*(x, q_x, q) k(w, W_s) \right\} \times \sup_{w, q \in \mathfrak{N}} |\tau^*(x, q_x, q) \Delta_{1M}(w)| \quad (17)$$

From well known results in the literature, the first component in (17) is $O_p(h^2) + O_p(N^{-(1/2-rd)}) = O_p(N^{-\delta})$ for r given in the lemma. ⁶ For the second component in (17), we bound the absolute difference in indicators by a smooth function. Let the estimation error be given as $e_j \equiv |\hat{q}_x(\lambda_j) - q_x(\lambda_j)| = O_p(N^{-1/2})$ and $d_{js} \equiv |X_s - q_x(\lambda_j)|$ as the distance to the lower ($j = 1$) and upper ($j = 2$) trimming boundaries respectively. Define the smoothed indicator function:

$$S(z) \equiv \left\{ 1 + \exp \left[-N^{(1/2-\frac{\varepsilon}{2})} (z + N^{-(1/2-\varepsilon)}) \right] \right\}^{-1}$$

$$S^*(e_j - d_{js}) \equiv S(e_j - d_{js}) + [1 - S(0)], \quad j = 1, 2$$

Employing an inequality due to Jim Powell and contained Klein (1993), from Klein(Lemma A, 1993):

$$|\tau(X_s, \hat{q}_x) - \tau(X_s, q_x)| \leq S^*(e_1 - d_{1s}) + S^*(e_2 - d_{2s})$$

Therefore, for the second component of D(w) in (17)

$$\sup_{w, q \in \mathfrak{N}} \frac{1}{N} \sum_s |\tau(X_s, \hat{q}_x) - \tau(X_s, q_x)| \tau^*(x, q_x, q) k(w, W_s) \quad (18)$$

$$\leq \sup_{w, q \in \mathfrak{N}} \frac{1}{N} \sum_s [S^*(e_1 - d_{1s}) + S^*(e_2 - d_{2s})] \tau^*(x, q_x, q) k(w, W_s) \quad (19)$$

As the analysis for each of two terms in (19) is the same, we provide the argument for

$$\sup_{w, q \in \mathfrak{N}} D_{21} \equiv \sup_{w, q \in \mathfrak{N}} \frac{1}{N} \sum_s S^*(e_1 - d_{1s}) \tau^*(X_s, q_x, q) k(w, W_s).$$

Denoting $\nabla^l S^*$ as the l^{th} partial derivative of S^* w.r.t. e_j , Taylor expand

⁶If uniformity holds in all $o(1)$ neighborhoods of q_x , it holds in any $o_p(1)$ neighborhood.

$S^*(e_j - d_j)$ about $e_j = 0$ to obtain:

$$\begin{aligned} \sup_{w, q \in \mathfrak{N}} D_{21}(w) &\leq [1 - S(0)] \frac{1}{N} \sup_{w, q \in \mathfrak{N}} \sum_s \tau^*(x, q_x, q) k(w, W_s) + \quad (20) \\ &\frac{1}{N} \sum_{l=0}^{L-1} e_1^l \sup_{w, q \in \mathfrak{N}} \sum_s |\nabla^l S^*(-d_{1s})| \tau^*(x, q_x, q) k(w, W_s) + \\ &e_1^L \sup_{w, q \in \mathfrak{N}} \frac{1}{N} \sum_s |\nabla^L S^*(e_j^+ - d_{s1})| \tau^*(x, q_x, q) k(w, W_s). \end{aligned}$$

For L sufficiently large and finite, the sup of the absolute value of the last term is $o_p(N^{-1/2})$ due to the factor e_1^L . Similarly because of the factor $[1 - S(0)]$ vanishes exponentially, the first term is of this order.

It can be shown that the absolute value of the m^{th} term, $0 \leq m < L$, is bounded above by:

$$\begin{aligned} &|e_j|^m \sup_{w, q \in \mathfrak{N}} \frac{1}{N} \sum_s |\nabla^m S^*(-d_s)| \tau^*(x, q_x, q) k(w, W_s) \quad (21) \\ &= |e_j|^m N^{m(1/2 - \frac{\varepsilon}{2})} \sup_{w, q \in \mathfrak{N}} \frac{1}{N} \sum_s B(d_{1s}) |S^*(-d_{js})| \tau^*(x, q_x, q) k(w, W_s) \end{aligned}$$

where B is uniformly bounded. The strategy for ordering this term will depend on the set in which d_{1s} lies. Define $b_s \equiv 1 \{d_{1s} \geq 2N^{-1/2+\varepsilon}\}$ and bound the sup of the term in (21) by:

$$o_p(1) \left[\sup_{w, q \in \mathfrak{N}} \frac{1}{N} \sum_s b_s B(d_{1s}) |S^*(-d_{js})| \tau^*(x, q_x, q) k(w, W_s) + \right. \\ \left. \sup_{w, q \in \mathfrak{N}} \frac{1}{N} \sum_s (1 - b_s) B(d_{1s}) |S^*(-d_{js})| \tau^*(x, q_x, q) k(w, W_s), \right]$$

For the b_s - terms, it follows from the definition of S^* that:

$$\begin{aligned} &\sup_{w, q \in \mathfrak{N}} \sum_s b_s B(d_{1s}) |S^*(-d_{1s})| \tau^*(x, q_x, q) k(w, W_s) \\ &\leq O(1) \sup_d |b_s S^*(-d)| \sup_{w, q \in \mathfrak{N}} \frac{1}{N} \sum_s \tau^*(x, q_x, q) k(w, W_s) \\ &\leq O\left(\frac{1}{1 + \exp(N^{\varepsilon/2})}\right) O_p(1) = o_p(N^{-1/2}) \end{aligned}$$

For the $(1 - b_s)$ - terms,

$$\begin{aligned} & \sup_w \frac{1}{N} \sum_{s=1}^N (1 - b_s) |S^*(-d_{1s})| \tau^*(x, q_x, q) k(w, W_s) \\ & \leq \frac{1}{N} \sum_{s=1}^N (1 - b_s) \sup_d |S^*(-d)| \sup_{w, sq \in \mathfrak{N}} [(1 - b_s) \tau^*(x, q_x, q) k(w, W_s)] \end{aligned} \quad (22)$$

The first term in (23) is $O(N^{-1/2+\varepsilon})$ because:

$$E(1 - b_s) = \Pr(d_s < 2N^{-1/2+\varepsilon})$$

The second component in (23) is $O(1)$. For the third component in (23), recall from D3) that

$$k(w, W_s) \equiv 1 \{Z_s = z\} \prod_{l=1}^d \frac{1}{s_l h} \phi\left(\frac{x_l - X_{ls}}{s_l h}\right)$$

By assumption, x_l is in the interior of a subset to which τ^* restricts X_{ls} . It then follows that the middle term in (23) is $o_p(N^{-1/2})$ because there exists $c > 0$ such that $|x_l - X_{ls}| > c$. The kernel then rapidly vanishes on this set. The result now readily follows for $k = 2$.

To complete the proof, assume the result holds for $k-1$ and write $\left| \hat{M}_k(w; \tau(\hat{q}_x)) - \hat{M}_k(w; \tau(q_x)) \right|$ as:

$$\left| \frac{1}{N} \sum_s [\tau(X_s, \hat{q}_x) - \tau(X_s, q_x)] \tau^*(X_s, q_x, \hat{q}_x) \left[\hat{M}_{k=1}(W_s) - \hat{M}_1(W_s) \right] K_s(w) \right|$$

The argument is now essentially the same as that above. ■

From Lemma 2, in all subsequent lemmas we will take trimming that controls the bias adjustment as given. To establish the order of the bias and to obtain convergence rates, it is useful to employ an expansion that eliminates estimated denominators. In so doing, we will be able to study the bias of an estimator that is within $o_p(N^{-1/2})$ of the original estimator $\Delta_k(w)$

Lemma 3. Estimated Denominators. With $h = O(N^{-r})$, assume $r < \frac{1}{2d}$, which is implied by A4). Recalling the definition of $g(w)$ and $\hat{g}(w)$ in

D4), let: $\delta(w) \equiv [g(w) - \hat{g}(w)]/g(w)$ and define:

$$\langle \delta(w) \rangle_{\mathbb{P}} \equiv \frac{1}{g(w)} \sum_{p=0}^{\mathbb{P}} \delta(w)^p,$$

with \mathbb{P} set sufficiently large to ensure that

$$\sup_e |\delta(w)|^{\mathbb{P}} = o_p(N^{-1/2})$$

Define:

$$\Delta_k^*(w) \equiv \begin{cases} \Delta_1(w) = \Delta_1^*(w) & k = 1 \\ \Delta_{k-1}^*(w) + \frac{1}{N} \sum_s \left[\frac{\Delta_{k-1}^*(W_s)}{g(W_s)} \langle \delta(W_s) \rangle_{\mathbb{P}} + \varepsilon_s \right] K_s(w) & k > 1 \end{cases}$$

Then,

$$\sup_w |\Delta_k(w) - \Delta_k^*(w)| = o_p(N^{-1/2})$$

Proof. Write:

$$\begin{aligned} \frac{1}{\hat{g}(w)} &= \frac{1}{g(w)} + \delta(w) \frac{1}{\hat{g}(w)} = \frac{1}{g(w)} + \delta(w) \left[\frac{1}{g(w)} + \delta(w) \frac{1}{\hat{g}(w)} \right] \\ &= \frac{1}{g(w)} \sum_{p=0}^{\mathbb{P}} \delta(w)^p + \frac{1}{\hat{g}(w)} (\delta(w))^{\mathbb{P}+1} \end{aligned} \quad (24)$$

For $\inf_w [g(w)]$ bounded away from 0, $\sup_w |\delta(w)| = O_p\left(h^2 + \frac{1}{h^d \sqrt{N}}\right)$.

Therefore for h given as above and for \mathbb{P} sufficiently large

$$\sup_w \left| \frac{1}{\hat{g}(w)} \delta(w)^{\mathbb{P}+1} \right| = o_p(N^{-1/2}) \quad (25)$$

Therefore,

$$\frac{1}{\hat{g}(w)} = \frac{1}{g(w)} \langle \delta(w) \rangle_{\mathbb{P}} + o_p(N^{-1/2}),$$

where the o_p -term is uniform in w . The lemma is now immediate for $k = 1$.

For $k = 2$ and K_{st} as the kernel function in D3):

$$\begin{aligned}
\Delta_2^*(w) &\equiv \Delta_1^*(w) - \frac{1}{N} \sum_s \left[\frac{\Delta_1^*(W_s)}{g(W_s)} \langle \delta(W_s) \rangle_{\mathbb{P}} - \varepsilon_s \right] K_s(w) \\
&= \Delta_1^*(w) - \frac{1}{N} \sum_s \left[\frac{\Delta_1^*(W_s)}{\hat{g}(W_s)} - \varepsilon_s \right] K_s(w) + o_p(N^{-1/2}) \\
&= \frac{1}{N} \sum_s \left[\frac{\Delta_1^*(w)}{\hat{g}(w)} - \frac{\Delta_1^*(W_s)}{\hat{g}_s} + \varepsilon_s \right] K_s(w) + o_p(N^{-1/2}) \\
&= \frac{1}{N} \sum_s \left[\frac{\Delta_1(w)}{\hat{g}(w)} - \frac{\Delta_1(W_s)}{\hat{g}_s} + \varepsilon_s \right] K_s(w) + o_p(N^{-1/2}) \\
&= \Delta_2(w) + o_p(N^{-1/2}),
\end{aligned}$$

where we have employed the expansion above with the $o_p(N^{-1/2})$ term being uniformly of this order in w . Assuming that the result holds for $k - 1$ and employing the same arguments as above, it readily follows that the result holds for stage k . ■

From the recursion above, Lemma 4 expresses $\Delta_k^*(w)$ in terms of first stage estimates, which is useful for studying the order of the bias.

Lemma 4. Characterization. Recalling the definitions of $\langle \delta(w) \rangle_{\mathbb{P}}$ and $\Delta_k^*(w)$ in Lemma 3, there exists integers C_1, \dots, C_{k-1} such that for $k > 1$, $\rho(w) \equiv g(w) / \langle \delta(w) \rangle_{\mathbb{P}}$:

$$\begin{aligned}
\Delta_k^*(w) - \Delta_1^*(w) &= U + \frac{C_1}{N} \sum_{i_1} \frac{\Delta_1^*(W_{i_1})}{\rho(W_{i_1})} K_{i_1}(w) + \\
&\quad \frac{C_2}{O(N^2)} \sum_{i_1} \sum_{i_2 \neq i_1} \frac{\Delta_1^*(W_{i_2})}{\rho(W_{i_1}) \rho(W_{i_2})} K_{i_1}(w) K_{i_1}(W_{i_2}) + \dots + \\
&\quad \frac{C_{k-1}}{O(N^{k-1})} \sum_{i_1} \dots \sum_{i_{k-1} \neq i_{k-2}} \frac{\Delta_1^*(W_{i_{k-1}})}{\prod_{l=1}^{k-1} \rho(W_{i_l})} K_{i_1}(w) \prod_{l=2}^{k-1} K_{i_l}(W_{i_{k-1}}),
\end{aligned}$$

where $E(U|W_1, \dots, W_N) = 0$.

Proof. Since

$$\Delta_k^*(w) - \Delta_1^*(w) = [\Delta_k^*(w) - \Delta_{k-1}^*(w)] + [\Delta_{k-1}^*(w) - \Delta_{k-2}^*(w)] + \dots + [\Delta_2^*(w) - \Delta_1^*(w)],$$

it suffices to show that $\Delta_k^*(w) - \Delta_{k-1}^*(w)$ has the form required by the lemma. The lemma is immediate for $k = 2$. Assuming the result holds for stage $k - 1$, recall from Lemma 3 that

$$\Delta_k^*(w) - \Delta_{k-1}^*(w) = \frac{1}{N} \sum_s \left[\frac{\Delta_{k-1}^*(W_s)}{\rho(W_s)} + \varepsilon_s \right] K_s(w)$$

Replacing $\Delta_{k-1}^*(w)$ and $\Delta_{k-1}^*(W_s)$ with the lemma's characterization completes the proof, with U being sums of kernel-weighted ε -terms, all of which have zero conditional expectation. ■

Lemma 5 below shows that up to $o(N^{-1/2})$, the expectation of the product of kernel terms is equal to the product of their expectations. This result is important as it provides the basis for showing how the bias decreases over stages.

Lemma 5. Expectations of Kernel Products. Assume A4) and let $g_{i_l} \equiv g(W_{i_l})$ and $\delta_{i_l} \equiv \delta(W_{i_l}) \equiv [g(W_{i_l}) - \hat{g}(W_{i_l})] / g(W_{i_l})$. With $\langle \delta_{i_l} \rangle_{\mathbb{P}}$ defined in Lemma 3, for arbitrary L , define:

$$D \equiv \frac{\Delta_1^*(W_{i_l})}{g_{i_l}} \prod_{l=1}^L \langle \delta_{i_l} \rangle_{\mathbb{P}},$$

$$E_{i_l}^* \equiv E \left[\frac{\Delta_1^*(W_{i_l})}{g_{i_l}} | W_{i_l} \right], \quad E_{i_l} \equiv E [\delta_{i_l} | W_{i_l}].$$

Then, with trimming taken as known from Lemma 2:

$$E(D | W_{i_l}, l = 1, \dots, L) = E_{i_l}^* \prod_{l=1}^L \langle E_{i_l} \rangle_{\mathbb{P}} + o(N^{-1/2}),$$

$$\langle E_{i_l} \rangle_{\mathbb{P}} \equiv \sum_{p=1}^{\mathbb{P}} [E_{i_l}]^p.$$

Proof. Write:

$$D \equiv \left\{ \left[\frac{\Delta_1^*(W_{i_l})}{g_{i_l}} - E_{i_l}^* \right] + E_{i_l}^* \right\} \prod_{l=1}^L \langle \{ [\delta_{i_l} - E_{i_l}] + E_{i_l} \} \rangle_{\mathbb{P}}$$

The conditional expectation of D will contain a finite number of terms, one of which is

$$E_{i_l}^* \prod_{l=1}^L \sum_{p=0}^{\mathbb{P}} [E_{i_l}]^p$$

Therefore, the lemma will follow if the expectation of each of the remaining finite number of terms in D is $o(N^{-1/2})$. These remaining terms will involve a product of expectations and centered kernel functions. For any stage k and an arbitrary $T < L$, typical terms are given as:

$$\begin{aligned} & \left[\frac{\Delta_1^*(W_{i_l})}{g_{i_l}} - E_{i_l}^* \right] \prod_{l=1}^T [\hat{g}_{i_l} - E_{i_l}] / g_{i_l} \prod_{l=T+1}^L [E_{i_l}]^p \\ & E_{i_l}^* \prod_{l=1}^T [\hat{g}_{i_l} - E_{i_l}] / g_{i_l} \prod_{l=T+1}^L [E_{i_l}]^p \end{aligned}$$

As the analysis for each of these two typical terms is the same, here we consider the expectation of the second. With all expectations and index densities being bounded, it then suffices to show that :

$$E \left[\prod_{l=1}^T [\hat{g}_{i_l} - E_{i_l}] | W_{i_l}, l = 1, \dots, T \right] = o(N^{-1/2})$$

To prove this result, recall that

$$\hat{g}_{i_l} \equiv \frac{1}{N-1} \sum_{j_l \neq i_l} \frac{1}{h^d} f(W_{i_l}, W_{j_l}).$$

where $f(\bullet)$ is uniformly bounded. Then, with \mathcal{J} as a set of positive from 1 to N excluding $\{j : j = i_1, i_2, \dots, i_T\}$:

$$\mathcal{J} \equiv \{j : j = 1, 2, 3, \dots, N\} \setminus \{j : j = i_1, i_2, \dots, i_L\},$$

define:

$$\hat{g}_{i_l}^* \equiv \frac{1}{N-T} \sum_{j_l \in \mathcal{J}} \frac{1}{h^d} f(W_{i_l}, W_{j_l}).$$

Employing this notation, it suffices to prove:

$$\begin{aligned}
a) & : E \left\{ \prod_{l=1}^T [\hat{g}_{i_l}^* - E_{i_l}] \mid W_{i_l}, l = 1, \dots, T \right\} = o(N^{-1/2}) \\
b) & : E \{ \Delta \mid W_{i_l}, l = 1, \dots, T \} = o(N^{-1/2}), \\
\Delta & \equiv \prod_{l=1}^T [\hat{g}_{i_l} - E_{i_l}] - \prod_{l=1}^T [\hat{g}_{i_l}^* - E_{i_l}]
\end{aligned}$$

For a), write the product of sums as a sum of products, where each product has the following block independent structure. Let $\mathcal{I}_1, \dots, \mathcal{I}_S$ be disjoint sets of integers with:

$$\cup_{s=1}^S \mathcal{I}_s = \{i_l, l = 1, \dots, T\}$$

Then, define Block s as:

$$B_s(\mathcal{I}_s, W_{i_s}) = \prod_{i_l \in \mathcal{I}_s} \frac{1}{h^d} \{f[W_{i_l}, W_{j_s}] - E_{i_l}\}, j_s \in \mathcal{J}.$$

For $j_1 \in \mathcal{J}$, let $\mathcal{J}_2 \equiv \mathcal{J} \setminus j_1$; for $j_2 \in \mathcal{J}_2$, let $\mathcal{J}_3 \equiv \mathcal{J}_2 \setminus j_2$. Continuing to eliminate elements from sets in this manner, for $j_{S-1} \in \mathcal{J}_{S-1}$ let $\mathcal{J}_S \equiv \mathcal{J}_{S-1} \setminus j_{S-1}$. Noting that the blocks are conditionally independent, a typical term in the expectation in a) then has the form:

$$\frac{1}{N^S} \sum_{j_1 \in \mathcal{J}} \sum_{j_2 \in \mathcal{J}_2} \dots \sum_{j_S \in \mathcal{J}_S} \frac{1}{N^{T-S}} \prod_{s=1}^S E[B_s(\mathcal{I}_s, W_{j_s}) \mid W_{i_l}, l = 1, \dots, T].$$

When a block only has one member, we will term such a block as a singleton. For each S , the above expectation will be 0 when there are singleton blocks. As there will always be singleton blocks when $T - S < S$, it suffices to consider the case where $T - S \geq S$. For this case, we can always construct at least one configuration with no singletons. The expectation then has a convergence rate to 0 of

$$\frac{1}{N^{T-S} h^{d(T-S)}}$$

This result follows because fW_{i_l}, W_{j_s}) is a bounded random variable and

$$\frac{1}{N^{T-S}} \prod_{s=1}^S B_s(\mathcal{I}_s, W_{j_s}) = O\left(\frac{1}{N^{T-S} h^{dT}}\right)$$

and in taking the expectation, we lose a factor of h^d for each of the \mathbb{S} blocks. Part a) of the lemma will then follow if:

$$\begin{aligned} \frac{1}{N^{T-\mathbb{S}}h^{d(T-\mathbb{S})}} &< N^{-1/2} \Leftrightarrow N^{T-\mathbb{S}}h^{d(T-\mathbb{S})} > N^{1/2} \\ &\Leftrightarrow T - \mathbb{S} - rd(T - S) > 1/2 \\ &\Leftrightarrow r < \frac{1}{d} - \frac{1}{2(T - S)d} \end{aligned}$$

With $T - \mathbb{S} \geq \mathbb{S} \geq 1$,

$$\frac{1}{2d} \leq \frac{1}{d} - \frac{1}{2(T - S)d}$$

Therefore with $r < \frac{1}{2d}$ (Assumption A), the result follows.

Turning to b), write:

$$\prod_{l=1}^T [\hat{g}_{i_l} - E_{i_l}] = \prod_{l=1}^T [(\hat{g}_{i_l}^* - E_{i_l}) + (\hat{g}_{i_l} - \hat{g}_{i_l}^*)].$$

Referring to the statement of b), It then follows that a typical term Δ is given as:

$$\prod_{i_t \in \mathcal{I}_1} (\hat{g}_{i_t}^* - E_{i_t}) \prod_{i_r \in \mathcal{I}_2} (\hat{g}_{i_r} - \hat{g}_{i_r}^*),$$

where \mathcal{I}_1 and \mathcal{I}_2 are non-empty and non-intersecting sets of integers such that $\mathcal{I}_1 \cup \mathcal{I}_2 = \mathcal{I}$. With there being T_k elements in \mathcal{I}_k and $T_1 + T_2 = T$, Let $\mathcal{I}_2[j]$ refer to the j^{th} integer element of \mathcal{I}_2 . Define $W(\mathcal{I}_1)$ as the vector with j^{th} element $\mathcal{I}_1[j]$, and similarly define $W(\mathcal{I}_2)$. The absolute value of the expectation of the above typical term is then given as:

$$\begin{aligned} &\left| E \left[\left(\prod_{i_t \in \mathcal{I}_1} (\hat{g}_{i_t}^* - E_{i_t}) \prod_{i_r \in \mathcal{I}_2} (\hat{g}_{i_r} - \hat{g}_{i_r}^*) \right) | W(\mathcal{I}_1), W(\mathcal{I}_2) \right] \right| \\ &\leq E \left[\prod_{i_t \in \mathcal{I}_1} (\hat{g}_{i_t}^* - E_{i_t}) | W(\mathcal{I}_1) \right] O \left(\frac{1}{Nh^d} \right)^{T_2} \end{aligned}$$

The result readily follows from a), the definition of \hat{g}_{i_r} and the window condition in A4), ■

Employing Lemma 5, Lemma 6 provides the basis for how expectations are updated over stages.

Lemma 6 .Employing notation given in the statement of Lemma 3:

$$E [\Delta_{k-1}^* (W_s) \langle \delta (W_s) \rangle_{\mathbb{P}} | V] = E [\Delta_{k-1}^* (W_s) | W_s] [\langle E [\delta (W_s)] \rangle_{\mathbb{P}} | W_s] + o(N^{-1/2})$$

Proof. The result is immediate for $k = 1$. For $k > 1$, from Lemma 4, a typical term in $\Delta_{k-1}^* (W_s) \langle \delta (W_s) \rangle_{\mathbb{P}}$ is given as:

$$\frac{C_j}{(N-1)^j} \sum_{i_1} \cdots \sum_{i_j \neq i_{j-1}} \frac{\Delta_1^*(W_{i_1}) \prod_{l=1}^j \langle \delta(W_{i_l}) \rangle_{\mathbb{P}}}{g(W_{i_1}) g(W_{i_2}) \cdots g(W_{i_j})} K_{i_1 s} \prod_{l=2}^j K_{i_l i_{l-1}}$$

where the integer $j < k$. The result now follows from Lemma 5 ■

Employing the above lemmas, Lemma 7 obtains a uniform bias rate and a convergence rate for the proposed estimator.

Lemma 7 With $\Delta_k^*(w)$ defined as in Lemma 3, $B_k(V_t)$ a uniformly bounded function, and trimming taken as known under Lemma 2:

$$\begin{aligned} a) & : \sup_w |E[\Delta_k^*(w)]| = h^{2k} B_k(V_t) + o(N^{-1/2}). \\ b) & : |\Delta_k^*(w)| = O_p\left(h^{2k} + \frac{1}{\sqrt{N}h^d}\right) \end{aligned}$$

Proof. For a), the result is immediate for $k = 1$. Assume the result is true for stage $k - 1$ so that

$$E[\Delta_{k-1}^*(w)] = h^{2k-2} B_{k-1}(w) + O(h^{2k}) + o(N^{-1/2}).$$

Then, from Lemmas 2 and 6:

$$\begin{aligned} E[\Delta_k^*(w)] & = \frac{1}{N} \sum_s E \left\{ \begin{array}{l} E \left[\frac{\Delta_{k-1}^*(w)}{g(w)} \langle \delta(w) \rangle_{\mathbb{P}} \right] \\ - E \left[\frac{\Delta_{k-1}^*(W_s)}{g(W_s)} \langle \delta(W_s) \rangle_{\mathbb{P}} | V_s \right] \end{array} \right\} K_s(w) + o(N^{-1/2}) \\ & = \frac{1}{N} \sum_s E \left\{ \begin{array}{l} h^{2k-2} C_{k-1}(w) \\ - h^{2k-2} C_{k-1}(W_s) + O(h^{2k}) \end{array} \right\} K_s(w) + o(N^{-1/2}), \end{aligned}$$

where C_{k-1} is uniformly bounded. Part a) of the Lemma now follows from Lemma 1.

To prove b), letting $E_k^*(w) \equiv E \left[\frac{\Delta_{k-1}^*(w)}{g(w)} \right]$, from Lemma 3, $\Delta_{k-1}^*(w)$ has the following form:

$$\left[\frac{\Delta_{k-1}^*(w)}{g(w)} - E_{k-1}^*(w) \right] - \quad (26)$$

$$\frac{1}{N} \sum_s \left\{ \left[\frac{\Delta_{k-1}^*(W_s)}{g_s} - E_{k-1}^*(W_s) \right] \frac{\langle \delta(W_s) \rangle_{\mathbb{P}}}{g(W_s)} + \varepsilon_s \right\} K_s(w) + \quad (27)$$

$$E_{k-1}^*(w) - \frac{1}{N} \sum_s \left[E_{k-1}^*(W_s) \frac{\langle \delta(W_s) \rangle_{\mathbb{P}}}{g(W_s)} \right] K_s(w) \quad (28)$$

Employing an induction argument, the term in (26) has the required convergence rate. For the term in (27), its absolute value is bounded above by:

$$\sup_w \left| \frac{\langle \delta(w) \rangle_{\mathbb{P}}}{g(w)} \right| \frac{1}{N} \sum_s \left| \frac{\Delta_{k-1}^*(w)}{g(w)} - E_{k-1}^*(w) \right| K_s(w). \quad (29)$$

From standard convergence results, the first component of (29) is $O_p(1)$ and the second is $O_p(1/\sqrt{Nh^d})$ from an induction argument. and standard convergence results. For the term in (28), from the proof of Lemma 3, we may write it as:

$$\frac{1}{N} \sum_s \left[E_{k-1}^*(w) \frac{\langle \delta(w) \rangle_{\mathbb{P}}}{g(w)} - E_{k-1}^*(W_s) \frac{\langle \delta(W_s) \rangle_{\mathbb{P}}}{g(W_s)} \right] K_s(w) + o_p(N^{-1/2}).$$

The result follows from Part a) and Lemma 3,. ■

To obtain asymptotic normality in semiparametric models, we require additional pointwise and uniform convergence rates. With θ a finite dimensional parameter vector, recall from D6) that observation s on the index is given as $V(W_s; \theta)$ and that a value for the index is denoted by $v(\theta) \equiv V(w; \theta)$. Employing this notation, lemmas 8-11 provide the required additional results for analyzing a class of estimators for index models.

Lemma 8: Mean-Square Convergence rate for expectations. Referring to D8), define $\Delta_k^*(V(w; \theta_0))$ as in Lemma 3 with $V(W_s; \theta_0)$ replacing V_s .

$$\sup_v E \left[\Delta_k^*(v)^2 \right] = O(N^{-4rk}) + O(N^{-(1-rd)}),$$

where the sup is taken over v , where $v \in \mathcal{C}_x(q_x^I) \equiv \{v : q_v(\lambda_1^I) < x < q_v(\lambda_2^I)\}$ from D2).

Proof. As the result is immediate for $k = 1$, consider case where $k > 1$. Letting $V_t \equiv V(W_t; \theta_0)$ and

$$\begin{aligned} u_{k-1}(v, V_s) &\equiv \left[\frac{\Delta_{k-1}^*(v)}{g_0(v)} \langle \delta(v) \rangle_{\mathbb{P}} - \frac{\Delta_{k-1}^*(V_s)}{g_0(V_s)} \langle \delta(V_s) \rangle_{\mathbb{P}} + \varepsilon_s \right] \\ D(v) &\equiv \Delta_{k-1}^*(v) \left[1 - \frac{\langle \delta(v) \rangle_{\mathbb{P}} \hat{g}(v)}{g_0(v)} \right], \end{aligned}$$

recall that

$$\begin{aligned} \Delta_k^*(v) &\equiv \Delta_{k-1}^*(v) - \frac{1}{N} \sum_s \left[\frac{\Delta_{k-1}^*(V_s)}{g_0(V_s)} \langle \delta(V_s) \rangle_{\mathbb{P}} - \varepsilon \right] K_s(v)_s \\ &= D(v) + \frac{1}{N} \sum_s u_{k-1}(v, V_s) K_s(v). \end{aligned}$$

Then,

$$\Delta_k^*(v)^2 \leq 2D^2(v) + \tag{30}$$

$$\frac{2}{N^2} \sum_s [u_{k-1}(v, V_s)]^2 K_s^2(v) + \tag{31}$$

$$\frac{2}{N^2} \sum_s \sum_{r \neq s} u_{k-1}(s, t) u_{k-1}(r, t) K_r(v) K_s(v). \tag{32}$$

Beginning with the term in (40), in an induction argument, it can readily be shown that it uniformly converges at the required rate. Being a single sum, it can also be shown that the squared terms in (31) converge in expectation to 0 faster than the double sum of cross-product terms in (32). Accordingly, it suffices to analyze the cross-product terms.

For the terms in (32), the lemma readily holds for $k = 1$. Assume that it holds for $k - 1, k > 1$. It can be shown that all components that depend upon ε_s and/or ε_t either have 0 expectation or have an expectation that vanishes faster than the rate in the lemma. Letting

$$\begin{aligned} u_s^*(v) &\equiv u_{k-1}(v, V_s) - \varepsilon_s \\ E_s(v) &\equiv E[u_s^*(v) | V_r, V_s] = E[u_{k-1}(v, V_s) | V_r, V_s], \end{aligned}$$

write the remaining terms as:

$$T_1 \equiv \frac{1}{O(N^2)} \sum_s \sum_{r \neq s} E_r(v) [u_s^*(v) - E_s(v)] K_r(v) K_s(v) \quad (33)$$

$$T_2 \equiv \frac{1}{O(N^2)} \sum_s \sum_{r \neq s} E_s(v) [u_r^*(v) - E_r(v)] K_{rt} K_{st} \quad (34)$$

$$T_3 \equiv \frac{1}{O(N^2)} \sum_s \sum_{r \neq s} u_s^*(v) u_r^*(v) K_r(v) K_s(v) \quad (35)$$

$$T_4 \equiv \frac{1}{O(N^2)} \sum_s \sum_{r \neq s} E_s(v) E_r(v) K_r(v) K_s(v) \quad (36)$$

From iterated expectations, $E(T_1) = E(T_2) = 0$. For T_3 , let " \bullet " refer to conditioning on V_r, V_s . Therefore,

$$\begin{aligned} |E(T_3|\bullet)| &\leq \frac{1}{O(N^2)} \sum_s \sum_{r \neq s} |E(u_s^*(v) u_r^*(v) | \bullet)| K_r(v) K_s(v) \\ &\leq \frac{1}{O(N^2)} \sum_s \sum_{r \neq s} E[(u_s^*(v))^2 + (u_r^*(v))^2 | \bullet] K_r(v) K_s(v) \end{aligned}$$

With the lemma holding at stage $k-1$, it follows that $|E(T_3|\bullet)|$ is uniformly $O_p(N^{-(1-rd)})$.

Turning to T_4 , note that:

$$\begin{aligned} E_{st} &= E \left[\frac{\frac{\Delta_{ok-1}^*(V_{0t})}{g_0(V_{0t})} \langle \delta(V_{0t}) \rangle_{\mathbb{P}} - \frac{\Delta_{k-1}^*(V_{0s})}{g_0(V_{0s})} \langle \delta(V_{0s}) \rangle_{\mathbb{P}}}{g_0(V_{0s})} \middle| V_r, V_s, V_t \right] \\ &= \left\{ \begin{array}{l} E \left[\frac{\Delta_{ok-1}^*(V_{0t})}{g_0(V_{0t})} \langle \delta(V_{0t}) \rangle_{\mathbb{P}} \middle| V_t \right] - \\ E \left[\frac{\Delta_{k-1}^*(V_{0s})}{g_0(V_{0s})} \langle \delta(V_{0s}) \rangle_{\mathbb{P}} \middle| V_s \right] \end{array} \right\} + o(N^{-1/2}) \\ &= h^{2(k-1)} \tau(X_t, q_o^c) [B(V_t) - B(V_s)] + o(N^{-1/2}), \end{aligned}$$

where, with B a uniformly bounded function, the last result follows from Lemma 7 and its proof. Therefore, up to $o(N^{-1/2})$:

$$\begin{aligned} E(T_4) &= h^{4(k-1)} E \left[\frac{[B(v) - B(v_s)] \times [B(v) - B(v_r)]}{[B(v) - B(v_s)]} K_s(v) K_r(v) \right] \\ &= h^{4(k-1)} \int \int \left[\frac{[B(v) - B(v_s)] [B(v) - B(v_r)] \times}{K_s(v) K_r(v) g(v_r) g(v_s)} \right] dv_r dv_s \end{aligned}$$

Change the variables of integration to $z_1 \equiv (v_r - v) / h$ and $z_2 \equiv (v_2 - v) / h$. The result then follows as the double integral is $O(h^4)$ ■

In providing results for a class of semiparametric models, we require a uniform result on derivatives, which is provided by Lemma 9.

Lemma 9: Mean Square Convergence for First Derivatives. Recalling that $w \equiv (x, z)$ with $v(\theta) \equiv V(w, \theta)$, for $x \in \mathcal{C}_x(q_x^I) \equiv \{x : q_x(\lambda_1^I) < x < q_x(\lambda_2^I)\}$ from D2).

$$\sup_w E \left[(\nabla_\theta^1 [\Delta_k(v(w, \theta))]_{\theta_0})^2 \right] = O(N^{-4rk}) + O(N^{-(1-r(d+2))})$$

Proof. Employing the notation of Lemma 8, $\nabla_\theta^1 [\Delta_k(v(w, \theta))]$ can be written as:

$$\nabla_\theta^1 D(v) + \frac{1}{N} \nabla_\theta^1 \sum_s u_{k-1}(v, V_s) [K_s(v)]$$

Similar to lemma 8:

$$[\nabla_\theta^1 \Delta_k^*(v)]^2 \leq 2 [\nabla_\theta^1 D(v)]^2 + \tag{37}$$

$$\frac{2}{N^2} \nabla_\theta^1 \sum_s [u_{k-1}(v, V_s)]^2 K_s^2(v) + \tag{38}$$

$$\frac{2}{N^2} \nabla_\theta^1 \sum_s \sum_{r \neq s} u_{k-1}(s, t) u_{k-1}(r, t) K_r(v) K_s(v). \tag{39}$$

The proof for each of these terms follows the same argument as in Lemma 8. ■

Lemma 10. Let X_{ks} be the s^{th} observation on the k^{th} exogenous variable. For a) below, assume A5): $E[|Y_s|^m] = O(1)$ for $m > 4$. Let r be the window parameter and set $\alpha = 0, 1, 2$. then, for v restricted to a compact set:

$$a) : \sup_\theta \left| \frac{\frac{1}{N-1} \sum_s Y_s \nabla_\theta^\alpha [K_s(v(\theta); \theta)]}{E \{Y_s \nabla_\theta^\alpha [K_s(v(\theta); \theta)]\}} - \right| = O_p \left(N^{-\left(\frac{m}{2(m+2)} - r(\alpha+d)\right)} \right)$$

$$b) : \sup_\theta \left| \frac{\frac{1}{N-1} \sum_s \nabla_\theta^\alpha [K_s(v(\theta); \theta)]}{E \{ \nabla_\theta^\alpha [K_s(v(\theta); \theta)] \}} - \right| = O_p \left(N^{-\left(\frac{1}{2} - r(\alpha+d)\right)} \right) \setminus$$

Proof. As the argument for b) is similar to and involves fewer assumptions than that for a), here we provide the proof for a). Let:

$$b_i = \begin{cases} 1 & : |Y_s| > N^{\frac{1}{m+2}} \\ 0 & : \textit{otherwise} \end{cases} .$$

Then, write the term in a) as:

$$\begin{aligned} T_{1-b} &\equiv \frac{1}{N} \sum_s (1 - b_i) Y_s \nabla_\theta^\alpha [K_s(v(\theta); \theta)] - E \{ Y_s \nabla_\theta^\alpha [K_s(v(\theta); \theta)] \} + \\ T_b &\equiv \frac{1}{N} \sum_s b_i Y_s \nabla_\theta^\alpha [K_s(v(\theta); \theta)] - E \{ Y_s \nabla_\theta^\alpha [K_s(v(\theta); \theta)] \} . \end{aligned}$$

For T_{1-b} , from standard results in the literature (e.g. see Bhattacharya (1967), Klein (1993)):

$$\sup_\theta |T_{1-b} - E(T_{1-b})| = O_p \left(N^{-\left(\frac{1}{2} - r(\alpha+d) - \frac{1}{m+2}\right)} \right) . \setminus = O_p \left(N^{-\left(\frac{1}{2(m+2)} - r(\alpha+d)\right)} \right)$$

For T_b , noting that the function $K_s(V(\theta))$ incorporates trimming that can be taken as known (Lemma 2) and that bounds X_s . Then for x restricted to a compact set, it follows that

$$\sup_{\theta, s} |\nabla_\theta^\alpha [K_s(v(\theta))]| = O \left(\frac{1}{h^{\alpha+d}} \right)$$

It then suffices to study

$$\begin{aligned} S_b &\equiv \frac{1}{h^{\alpha+d}} E [b_i | Y_s] \\ &\leq \frac{1}{h^{\alpha+d}} E [b_i]^{1/2} E [|Y_s^2|]^{1/2} \end{aligned}$$

which follows from Cauchy-Schwarz. For the first of the two components in the upper bound:

$$\begin{aligned} E [b_i] &\leq \Pr \left(|Y_s|^m > N^{\frac{m}{m+2}} \right) \\ &\leq N^{-\left(\frac{m}{m+2}\right)} E [|Y_s|^m] = O \left(N^{-\frac{m}{m+2}} \right) \end{aligned}$$

Therefore,

$$S_b = O \left(N^{-\left[\frac{m}{2(m+2)} - r(\alpha+d)\right]} \right)$$

The result now follows.⁷ ■

Finally, for establishing asymptotic normality, we need a uniform convergence lemma for the estimated mean functions and their first two derivatives. Lemma 11 provides the required results.

Lemma 11. Assume A5 and that with r as the window parameter:

$$1/2 - r(d + \alpha) - \frac{1}{m + 2} > 0.$$

Denote $\nabla_{\theta}^{\alpha} f(\theta)$ as the α^{th} partial derivative of the function f w.r.t. θ , with $\nabla_{\theta}^0 f(\theta) \equiv f(\theta)$. Then recalling that $v(\theta) \equiv V(w, \theta)$ from D6), for $\alpha = 0, 1, 2$:

$$\sup_{w, \theta} |d(w, \theta)| \equiv o_p(1), \quad d(w, \theta) \equiv \nabla_{\theta}^{\alpha} \left[\hat{M}_k(v(\theta); \theta) - M(v(\theta); \theta) \right]$$

Proof. From standard results in the literature (or Lemma 10), the lemma is immediate for $k - 1$. For $k \geq 2$, write:

$$\hat{M}_k(v(\theta); \theta) - M(v(\theta); \theta) = \frac{1}{N} \sum \left\{ Y_s - \left[\begin{array}{c} \hat{M}_{k-1}(V_s(\theta); \theta) - \\ \hat{M}_{k-1}(v(\theta); \theta) \end{array} \right] \right\} K_s(v(\theta); \theta)$$

Differentiating the above expression with

$$d(w, \theta) = d_1(w, \theta) - d_{21}(w, \theta) + d_{22}(w, \theta) - d_{23}(w, \theta) + d_3(w, \theta),$$

⁷The b_i indicator was defined to insure that the b-terms and the (1=b)-terms converge to zero at the same rate.

yields:

$$d_1(w, \theta) \equiv \frac{1}{N} \sum Y_s \nabla_\theta^\alpha K_s(v(\theta); \theta) \quad (40)$$

$$d_{21}(w, \theta) \equiv \frac{1}{N} \sum \left\{ \frac{\hat{g}_s [\hat{M}_{k-1}(V_s(\theta); \theta) - M(V_s(\theta); \theta)]}{\times \nabla_\theta^\alpha K_s(v(\theta); \theta)} \right\} \quad (41)$$

$$d_{22}(w, \theta) \equiv \frac{1}{N} \sum \left\{ \frac{\hat{g}_t [\hat{M}_{k-1}(v; \theta) - M(v; \theta)]}{\times \nabla_\theta^\alpha K_s(v(\theta); \theta)} \right\} \quad (42)$$

$$d_{23}(w, \theta) \equiv \frac{1}{N} \sum \left\{ \frac{[M_{k-1}(V_s; \theta) - M_{k-1}(v; \theta)]}{\times \nabla_\theta^\alpha K_s(v(\theta); \theta)} \right\} \quad (43)$$

$$d_3(w, \theta) \equiv \frac{1}{N} \sum \left\{ \nabla_\theta^\alpha \left[\begin{array}{l} \hat{M}_{k-1}(V_s(\theta); \theta) - \\ \hat{M}_{k-1}(v(\theta); \theta) \end{array} \right] \times K_s(v(\theta); \theta) \right\} \quad (44)$$

From Lemma 10, the term in (40) uniformly converges to $\nabla_\theta^\alpha M(v; \theta)$.

For the terms in (41)-(42), the result holds for $k = 1$ from Lemma 10. An induction argument completes the proof. For the term in (43), it can be shown that its expectation uniformly converges to 0. The result that follows from the same argument as in Lemma 10.

Turning to the term in (44), decompose it into three terms as in (41)-(43). Employing Lemma 10, the proof is then essentially the same as that for the previous three terms. ■