Uniform Convergence of Weighted Sums of Non- and Semi-parametric Residuals for Estimation and Testing*

Juan Carlos Escanciano[†] Indiana University

David T. Jacho-Chávez[‡] Emory University

Arthur Lewbel[§] Boston College

First draft: May 2010. This draft: January 2012

Abstract

A new uniform expansion is introduced for sums of weighted kernel-based regression residuals from nonparametric or semiparametric models. This result is useful for deriving asymptotic properties of semiparametric estimators and test statistics with data-dependent bandwidth, random trimming, and estimated weights. An extension allows for generated regressors, without requiring the calculation of functional derivatives. Example applications are provided for a binary choice model with selection, including a new semiparametric maximum likelihood estimator, and a new directional test for correct specification of the average structural function. An extended Appendix contains general results on uniform rates for kernel estimators, additional applications, and primitive sufficient conditions for high level assumptions.

Keywords: Semiparametric regression; Semiparametric residuals; Nonparametric residuals; Uniform-in-bandwidth; Sample selection models; Empirical process theory; Limited dependent variables.

JEL classification: C13; C14; C21; D24

[*] This paper replaces material from an earlier paper, "Identification and Estimation of Semiparametric Two Step Models." We would like to thank Hidehiko Ichimura, Simon Lee, Enno Mammen, Jim Powell, Jean Marc Robin, Christoph Rothe, Ingrid van Keilegom, Adonis Yatchew and participants of many conferences and seminars for helpful suggestions. All errors are our own.

[†] Department of Economics, Indiana University, 105 Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405–7104, USA. E-mail: jescanci@indiana.edu. Web Page: http://mypage.iu.edu/~jescanci/. Research funded by the Spanish Plan Nacional de I+D+I, reference number SEJ2007-62908.

[‡] Department of Economics, Emory University, Rich Building 306, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA. E-mail: djachocha@emory.edu. Web Page: http://userwww.service.emory.edu/~djachoc/.

[§] Corresponding Author: Department of Economics, Boston College, 140 Commonwealth Avenue, Chesnut Hill, MA 02467, USA. E-mail: lewbel@bc.edu. Web Page: http://www2.bc.edu/~lewbel/.

1 Introduction

This paper provides a new uniform expansion for a sum of weighted kernel-based regression residuals from nonparametric or semiparametric models, which has a variety of applications in semiparametric estimation and testing. Consider an independent and identically distributed (iid) data set $\{Y_i, X_i^\top\}_{i=1}^n$ drawn from the joint distribution of the vector-valued random variable $(Y, X^\top)$. Henceforth, $A^\top$ denotes the transpose of A. Let $W := W(X)$ denote a vector of measurable functions of X with $W_i := W(X_i)$, let $\widehat{m}(\cdot|W)$ be a Nadaraya-Watson (NW) kernel estimator of $E[Y|W = \cdot]$, let $\widehat{t}_{ni}(W)$ be a data dependent trimming function that may depend on a kernel density estimator $\widehat{f}(\cdot|W)$ of the unknown density of W, $f(\cdot|W)$, and let $\phi(X)$ be any generic measurable and integrable function of X. Both $\widehat{m}(\cdot|W)$ and $\widehat{f}(\cdot|W)$ use a common possibly data dependent bandwidth $\widehat{h}_n$.

Our primary contribution is a general representation of the empirical process

$$\hat{\Delta}_n(W, \phi) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i - \widehat{m}(W_i|W)\} \widehat{t}_{ni}(W) \phi(X_i) \tag{1}$$

that is uniform in the bandwidth $\widehat{h}_n$, and uniform in both W and ϕ. In addition, when ϕ and W are unknown but can be consistently estimated by $\widehat{\phi}$ and $\widehat{W}$, and if $E[Y|X]$ fulfills the index condition $E[Y|X] = E[Y|W(X)]$ almost surely (a.s.), we also provide a uniform representation of the process $\hat{\Delta}_n(\widehat{W}, \widehat{\phi})$ that accounts for the estimation effects of $\widehat{W}$ and $\widehat{\phi}$ without requiring calculation of pathwise functional derivatives.

For the example applications we list in the following paragraphs (and many others like them), the results here can be used to extend otherwise known asymptotic properties of estimators and tests to allow for data dependent bandwidths and random trimming. We also show that $\hat{\Delta}_n$ can be used to develop inference for some new semiparametric models and objects of interest. Our results include simple theoretically justified data dependent bandwidth choice procedures.

Equation (1) has the form of typical terms that show up in expansions of semiparametric estimators and test statistics. For example, by defining ϕ accordingly, if $W(X) = (X^\top\theta_1, \ldots, X^\top\theta_J)$ for a collection of J-finite dimensional unknown parameters $\theta_1, \ldots, \theta_J$, $\hat{\Delta}_n$ could be the first order conditions for a semiparametric weighted least squares estimator of index parameters as in Ichimura and Lee (1991) or when $J = 1$, $\hat{\Delta}_n$ could be the first order conditions for semiparametric weighted least squares or maximum likelihood estimators as those in Ichimura (1993) and Klein and Spady (1993), respectively. Similarly, if $X := (X_1^\top, X_2^\top, Z_1^\top, Z_2^\top)^\top$ and $W(X) = (Z_1^\top\theta_1 + X_2^\top\theta_2, X_2 - g(Z_1, Z_2))$, then $\hat{\Delta}_n$ could be the first order conditions for semiparametric weighted least squares or maximum likelihood estimators that uses 'control function' approaches as in Escanciano, Jacho-Chávez and Lewbel (2011) and Rothe (2009) respectively. Alternatively, if $W(X) = X_1 \subset X$, $\hat{\Delta}_n$ also has the form of test statistics designed to test nonparametrically the significance of a subset of covariates as in Delgado and González Manteiga (2001).

When $\widehat{W}$ replaces W in (1), we have a generated regressors model, as (parametrically) described by Pagan (1984). Semiparametric models with generated regressors include Ichimura and Lee (1991), Ichimura (1993), Ahn and Powell (1993), Ahn and Manski (1993), Olley and Pakes (1996), Ahn (1997), Heckman, Ichimura and Todd (1998), Newey, Powell and Vella (1999), Pinkse (2001), Li and Wooldridge (2002), Das, Newey, and Vella (2003), Blundell and Powell (2004), Heckman and Vytlacil (2005), Lewbel and Linton (2007), Imbens and Newey (2009), Rothe (2009), and Mammen, Rothe and Schienle (2011b), among others. The asymptotic variance of general estimators within this class of models is studied by Hahn and Ridder (2010). Analyses of the properties of generic nonparametric two step estimators with nonparametric generated regressors, include Song (2008), Sperlich (2009), and Mammen, Rothe and Schienle (2011a).

We contribute to these literatures in several ways. First, we provide results allowing for stochastic bandwidths, which can be difficult to obtain using more standard methods of analysis such as U-statistic or U-processes theory. Second, we show how simple stochastic equicontinuity arguments can be used to derive the impact of generated regressors on inference. Third, we propose a unified method for inference in semiparametric models with generated regressors, including estimation, testing, and bandwidth choice. Fourth, we contribute to the literature on nonparametric two step estimation with nonparametric generated regressors by providing results for kernel estimators that are uniform in the bandwidth. In particular, the Appendix to this paper shows how our new results can be used to prove that, under primitive conditions, the infinite-dimensional nuisance parameter belongs to a certain class of smooth functions. This then provides primitive conditions for a high level assumption that is commonly employed in the semiparametric estimation literature (see e.g, Chen, Linton, and van Keilegom, 2003 and Ichimura and Lee, 2010).

Works devoted to estimation of general semiparametric models include Bickel, Klaassen, Ritov and Wellner (1993), Andrews (1994), Newey (1994), Newey and McFadden (1994), Ai and Chen (2003), Chen, Linton, and van Keilegom (2003), Ichimura and Lee (2010) and references therein. Applications of these general results are frequently difficult because they require an investigation of pathwise functional derivatives (often up to a second order) and their limits. Our uniform representation of the process $\hat{\Delta}_n(\widehat{W}, \widehat{\phi})$ accounts for the estimation effects of $\widehat{W}$ and $\widehat{\phi}$ without requiring the calculation of pathwise functional derivatives. This is possible here by means of an approach based on stochastic equicontinuity arguments. Andrews (1994) also used stochastic equicontinuity for estimating semiparametric models, but he relied on an asymptotic orthogonality condition that does not always hold in our setting. One purpose of our results is to show how stochastic equicontinuity can still be used in situations where the orthogonality condition fails.

Related to our derivation is work on nonparametric and semiparametric estimation with possibly parametric or nonparametric generated covariates. In particular, Mammen, Rothe and Schienle (2011a,b) study these problems using kernel estimators, and characterize the asymptotic contribution of generated regressors to the pointwise distribution of their local linear estimator, as well as to the distribution of optimization estimators. Unlike Mammen, Rothe and Schienle (2011a,b), our results permit data dependent bandwidths and random trimming for both semiparametric estimation and testing. In the Appendix material we also provide sufficient conditions for the uniform (in evaluation point, conditioning variable, and bandwidth) consistency of the NW estimator and related quantities.

Also related is a recent paper by Li and Li (2010) which provides sufficient conditions for the first-order asymptotic properties of a larger class of kernel-based semiparametric estimators and test statistics to hold with data dependent bandwidths. Their method of proof requires one to use an estimated bandwidth first with a 'rule-of-thumb' asymptotic representation, i.e. a constant term times a known power of the sample size, and then establish the stochastic equicontinuity of these generic estimators and test statistics with respect to this constant term. Our development does not require this last step. Instead, our results are shown to hold uniformly over sets of admissible bandwidths which include estimated bandwidths with 'rule-of-thumb' asymptotic representations as a special case. A useful by-product of our proposed method is that bandwidth choice procedures can be readily justified without further calculations under our assumptions.

To illustrate the general applicability of our results for both estimation and testing, we first apply them to a semiparametric binary threshold crossing model with sample selection. This model has the form $Y = \mathbb{I}\left(X^{\top}\theta_0 - e \geq 0\right) D$ with $D = \mathbb{I}\left[g_0\left(X\right) - u \geq 0\right]$, where $\mathbb{I}\left(\cdot\right)$ represents the indicator function that equals one if its argument is true and zero otherwise. Here D is a binary variable that indicates if an individual is selected. An individual who is not selected has both $D = 0$ and $Y = 0$, while selected individuals have $D = 1$ and choose outcome Y to be either zero or one based on a threshold crossing model. The function $g_0\left(X\right)$ and the distribution of e given u are nonparametric, with the motivation that economic theory drives model specification for the outcome Y, but relatively less is known about the selection mechanism. We propose a semiparametric maximum likelihood estimator for θ_0 as in Klein and Spady (1993), but with a nonparametric generated regressor estimated in a first stage. The estimator includes both observation weighting and data dependent trimming to increase efficiency. It also includes a data dependent bandwidth choice that is justified by our asymptotic theory.

For a second application, we construct a directional test for the correct specification of a policy parameter in this model. More precisely, we consider a researcher who is concerned about misspecification of the semiparametric model only to the extent that the misspecification may lead to inconsistent estimates of an average structural function (ASF) parameter. We show how a directional test can be developed for this situation using our uniform expansions. Our uniform expansion permits the use of a data-driven bandwidth choice procedure for this example that leads to a test with better power properties than alternatives that use bandwidths chosen for estimation.

The paper is organized as follows: Section 2 provides our main uniform expansion results, including the extension allowing for generated regressors. Section 3 illustrates the utility of these results by applying them to the new estimator and new test statistic for the binary threshold crossing model with sample selection described above. Section 4 concludes, and the main proofs of Sections 2 and 3 are gathered into an Appendix A.

Appendix B for this paper contains new results on uniform rates of convergence of kernel estimators based on our theorems, and Appendix C contains examples of primitive conditions that suffice to satisfy some high level assumptions. Similarly, Appendix D provides more example applications of our results, including a description of how they could be generically applied to derive the asymptotic properties of semiparametric estimators such as Ichimura (1993), Klein and Spady (1993) and Rothe (2009), while allowing for data-driven bandwidths, data-driven asymptotic trimming, and estimated weights. Similarly, to provide another application of the proposed results for testing, Appendix D also contains a new test for the null hypothesis of zero conditional average treatment effect under selection on observables. The test is justified under minimal regularity conditions.

2 A Uniform Expansion

Let $\{Y_i, X_i^\top\}_{i=1}^n$ represent a random sample of size n from the joint distribution of $(Y, X^\top)$ taking values in $\mathcal{X}_Y \times \mathcal{X}_X \in \mathbb{R}^{1+p}$. Let $(\Omega, \mathcal{F}, P)$ be the probability space in which all the variables of this paper are defined. Henceforth, $\mathcal{X}_\xi$ denotes the support of the generic random vector ξ. Let $\mathcal{W}$ be a class of measurable functions of X with values in $\mathbb{R}^d$, and let $f\left(w|\, W\right)$ denote the Lebesgue density of $W(X)$ evaluated at $w \in \mathcal{X}_W$. Define $\mathcal{X}_\mathcal{W} := \{W(x) \in \mathbb{R}^d : W \in \mathcal{W} \text{ and } x \in \mathcal{X}_X\}$. To simplify notation define $W_i := W(X_i)$ and $W := W\left(X\right)$. We assume that $E\left|Y\right| < \infty$, so that the regression function

$$m\left(w|\, W\right) := E[Y|\, W = w], \qquad w \in \mathcal{X}_W \subset \mathbb{R}^d,$$

is well defined a.s., for each $W \in \mathcal{W}$. Under standard regularity conditions, the function $m\left(w|\, W\right)$ can be consistently estimated by the nonparametric NW kernel estimator

$$\widehat{m}\left(w|\, W\right) := \widehat{T}\left(w|\, W\right)/\widehat{f}\left(w|\, W\right),$$
$$\widehat{T}\left(w|\, W\right) := \frac{1}{n}\sum_{i=1}^n Y_i K_{\widehat{h}_n}\left(w - W_i\right),$$
$$\widehat{f}\left(w|\, W\right) := \frac{1}{n}\sum_{i=1}^n K_{\widehat{h}_n}\left(w - W_i\right),$$

where $K_h\left(w\right) = \prod_{l=1}^d k_h(w_l)$, $k_h\left(w_l\right) = h^{-1}k\left(w_l/h\right)$, $k\left(\cdot\right)$ is a kernel function, $w = (w_1, \ldots, w_d)^\top$ and $\widehat{h}_n$ denotes a possibly data dependent bandwidth parameter satisfying regularity conditions described in Assumption 5 below. Appendix B provides sufficient conditions for the uniform (in w, W and $\widehat{h}_n$) consistency of $\widehat{m}$ and related quantities. These new uniform in bandwidth convergence results should be of some independent interest.

Let $f_X\left(x|\, w, W\right)$ be the density, with respect to a σ-finite measure $\mu_W\left(\cdot\right)$, of X conditional on $W = w$, and evaluated at $x \in \mathcal{X}_X$. Note that X does not need to be absolutely continuous as we do not require $\mu_W\left(\cdot\right)$ to be the Lebesgue measure. To measure the complexity of the class $\mathcal{W}$, we employ covering numbers. For a measurable class of functions $\mathcal{G}$ from $\mathbb{R}^p$ to $\mathbb{R}$, let $\left\|\cdot\right\|$ be a generic pseudo-norm on $\mathcal{G}$, defined as a norm except for the property that $\left\|f\right\| = 0$ does not necessarily imply that $f \equiv 0$. Let $N(\varepsilon, \mathcal{G}, \left\|\cdot\right\|)$ denote the *covering number with respect to* $\left\|\cdot\right\|$, i.e., the minimal number of ε-balls with respect to $\left\|\cdot\right\|$ needed to cover $\mathcal{G}$. Given two functions $l, u \in \mathcal{G}$ a bracket $[l, u]$ is the set of functions $f \in \mathcal{G}$ such that $l \leq f \leq u$. An ε-bracket with respect to $\left\|\cdot\right\|$ is a bracket $[l, u]$ with $\left\|l - u\right\| \leq \varepsilon$, $\left\|l\right\| < \infty$ and $\left\|u\right\| < \infty$. The *covering number with bracketing* $N_{[\cdot]}(\varepsilon, \mathcal{G}, \left\|\cdot\right\|)$ is the minimal number of ε-brackets with respect to $\left\|\cdot\right\|$ needed to cover $\mathcal{G}$. These definitions are extended to classes taking values in $\mathbb{R}^d$, with $d > 1$, by taking the maximum of the covering or bracketing numbers of the coordinate classes. Let $\left\|\cdot\right\|_{2,P}$ be the $L_2(P)$ norm, i.e. $\left\|f\right\|_{2,P}^2 = \int f^2 dP$. When P is clear from the context, we simply write $\left\|\cdot\right\|_2 \equiv \left\|\cdot\right\|_{2,P}$. Let λ denote the Lebesgue measure. Let $\left|\cdot\right|$ denote the Euclidean norm, i.e. $\left|A\right|^2 = A^\top A$. Let $\left\|\cdot\right\|_\infty$ and $\left\|\cdot\right\|_{\mathcal{W},\infty}$ denote the *sup*-norms

$\|f\|_{\infty} := \sup_{x\in\mathcal{X}_X} |f(x)|$ and $\|q\|_{\mathcal{W},\infty} := \sup_{W\in\mathcal{W},w\in\mathcal{X}_W} |q(w|W)|$, respectively. Henceforth, P^* and E^* denote the outer probability and expectation, respectively; see van der Vaart and Wellner (1996). Finally, throughout C denotes a positive constant that may change from expression to expression. We consider the following regularity conditions.

Assumption 1 *The sample observations $\{Y_i, X_i^\top\}_{i=1}^n$ are a sequence of iid variables, distributed as $(Y, X^\top)$, satisfying $E[|Y|^s|X = x] < C$ a.s., for some $s > 2$.*

Assumption 2 *The class $\mathcal{W}$ is such that $\log N(\varepsilon, \mathcal{W}, \|\cdot\|_\infty) \le C\varepsilon^{-v_w}$ for some $v_w < 1$.*

Assumption 3 *For all $W \in \mathcal{W}$ and $x \in \mathcal{X}_X$: $f(w|W)$, $m(w|W)$ and $f_X(x|w, W)$ are r-times continuously differentiable in w, with uniformly (in w, W and x) bounded derivatives (including zero derivatives), where r is as in Assumption 4.*

Assumption 4 *The kernel function $k(t) : \mathbb{R} \to \mathbb{R}$ is bounded, r-times continuously differentiable and satisfies the following conditions: $\int k(t)\,dt = 1$, $\int t^l k(t)\,dt = 0$ for $0 < l < r$, and $\int |t^r k(t)|\,dt < \infty$, for some $r \ge 2$; $|\partial k(t)/\partial t| \le C$ and for some $v > 1$, $|\partial k(t)/\partial t| \le C|t|^{-v}$ for $|t| > L$, $0 < L < \infty$.*

Assumption 5 *The possibly data dependent bandwidth $\widehat{h}_n$ satisfies $P(a_n \le \widehat{h}_n \le b_n) \to 1$ as $n \to \infty$, for deterministic sequences of positive numbers a_n and b_n such that: (i) $b_n \to 0$ and $a_n^d n/\log n \to \infty$; (ii) $nb_n^{2r} \to 0$.*

The conditional bounded moment of Assumption 1 can be relaxed to $E[|Y|^s] < C$ by working with bracketing entropies of weighted L_2-norms instead. Assumption 2 restricts the "size" of the class $\mathcal{W}$ with respect to $\|\cdot\|_\infty$. van der Vaart and Wellner (1996) contains numerous examples of classes $\mathcal{W}$ satisfying Assumption 2. To give an example, define for any vector a of p integers the differential operator $\partial_x^a := \partial^{|a|_1}/\partial x_1^{a_1} \ldots \partial x_p^{a_p}$, where $|a|_1 := \sum_{i=1}^p a_i$. Assume that $\mathcal{X}$ is the finite union of convex, bounded subsets of $\mathbb{R}^p$, with non-empty interior. For any smooth function $h : \mathcal{X} \subset \mathbb{R}^p \to \mathbb{R}$ and some $\eta > 0$, let $\underline{\eta}$ be the largest integer smaller than η, and

$$\|h\|_{\infty,\eta} := \max_{|a|_1\le\underline{\eta}} \sup_{x\in\mathcal{X}} |\partial_x^a h(x)| + \max_{|a|_1=\underline{\eta}} \sup_{x\ne x'} \frac{|\partial_x^a h(x) - \partial_x^a h(x')|}{|x - x'|^{\eta-\underline{\eta}}}.$$

Further, let $C_M^\eta(\mathcal{X})$ be the set of all continuous functions $h : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ with $\|h\|_{\infty,\eta} \le M$. Then, it is known that $\log N(\varepsilon, C_M^\eta(\mathcal{X}), \|\cdot\|_\infty) \le C\varepsilon^{-v_w}$, $v_w = d/\eta$, so if $\mathcal{W} \subset C_M^\eta(\mathcal{X}_X)$, then $d < \eta$ suffices for our Assumption 2 to hold in this example. For extensions to unbounded $\mathcal{X}$ see Nickl and Pötscher (2007).

Assumption 3 is used for controlling the bias of $\widehat{m}$ and related quantities. Assumption 4 is standard in the nonparametric kernel estimation literature, while Assumption 5 permits data dependent bandwidths, as in e.g. Andrews (1995). In particular, our theory allows for plug-in bandwidths of the form $\widehat{h}_n = \widehat{c}h_n$ with $\widehat{c}$ stochastic and h_n a suitable deterministic sequence converging to zero as $n \to \infty$. Andrews (1995) points out that this condition holds in many common data dependent bandwidth selection procedures, such as cross-validation and generalized cross-validation. Similarly, our results also apply to deterministic sequences. In particular if $\widehat{h}_n$ is of the form $\widehat{h}_n = cn^{-\delta}$, for some constant $c > 0$, then Assumption 5 requires that $1/2r < \delta < 1/d$, so r needs to be greater than $d/2$. That is, a simple second-order Gaussian kernel can be used when $d < 4$, in view of Assumption 5.

We now introduce a class of functions that will serve as a parameter space for m. We assume that for each $W \in \mathcal{W}$, $\mathcal{X}_W$ is a finite union of convex, bounded subsets of $\mathbb{R}^d$, with non-empty interior. Let $\mathcal{T}_M^{\eta}$ be a class of measurable functions on $\mathcal{X}_X$, $q\left(W(x)\right| W)$ say, such that $W \in \mathcal{W}$ and q satisfies for a universal constant C_L and each $W_j \in \mathcal{W}$, $j = 1, 2$,

$$\left\| q\left(W_1(\cdot)\right| W_1) - q\left(W_2(\cdot)\right| W_2) \right\|_{\infty} \leq C_L \left\| W_1 - W_2 \right\|_{\infty}. \tag{2}$$

Moreover, assume that for each $W \in \mathcal{W}$, $q\left(\cdot\right| W) \in C_M^{\eta}(\mathcal{X}_W)$, for some $\eta \geq 1$, and $\|q\|_{\mathcal{W},\infty} < \infty$.

Assumption 6 *(i) $m \in \mathcal{T}_M^{\eta_m}$; and (ii) $P(\widehat{m} \in \mathcal{T}_M^{\eta_m}) \to 1$, for some $\eta_m > d/2$ and $M > 0$.*

Assumption 6 is a high level condition, some version of which is commonly required in the literature of semiparametric estimation. See e.g. Assumption 2.4 in Chen, Linton, and van Keilegom (2003) and Assumption 3.4(b) in Ichimura and Lee (2010). Even for simple cases such as standard kernel estimators, the verification of assumptions like 6(ii) is rather involved, see e.g. Akritas and van Keilegom (2001) and Neumeyer and van Keilegom (2010). Appendix C provides primitive conditions for 6(ii) and similar assumptions, showing how they can be used in complex settings (including ours) that can include possibly data dependent bandwidths and generated regressors.

We next introduce some technical conditions to handle the random trimming factor

$$\widehat{t}_{ni}(W) := \mathbb{I}(\widehat{f}\left(W_i\right| W) \geq \tau_n),$$

where τ_n satisfies Assumption 7(ii) below. Define the rates $p_n := E\left[\sup_{W \in \mathcal{W}} \mathbb{I}(f\left(w\right| W) \leq 2\tau_n)\right]$, $\lambda_n := \sup_{W \in \mathcal{W}} \lambda(w : f\left(w\right| W) < 2\tau_n)$ and

$$d_n := \sqrt{\frac{\log a_n^{-d} \vee \log\log n}{na_n^d}} + b_n^r.$$

Note that d_n is the rate of convergence of quantities like $\|\,\widehat{f} - f\,\|_{\mathcal{W},\infty}$, where a_n and b_n are as in Assumption 5.

Assumption 7 *(i) For all $W \in \mathcal{W}$, and all $u > 0$ and $\delta > 0$ sufficiently small, $P(u - 2\delta \leq f(W|W) \leq u + 2\delta) \leq C\delta$; and (ii) τ_n is a sequence of positive numbers satisfying $\tau_n \to 0$, $\lambda_n^2 a_n^{-d} \log n \to 0$ and $n(\tau_n^{-4} d_n^4 + p_n^2) \to 0$.*

Assumption 7 is a strong assumption, but it is hard to relax given that uniform results in $W \in \mathcal{W}$ are pursued. A sufficient condition for Assumption 7(i) is that $f/\dot{f}$ is bounded, where $\dot{f}$ is the derivative of f. In simpler settings, like for instance when the class $\mathcal{W}$ is parametric, these assumptions can be relaxed by using a similar approach to that in Robinson (1988). Note that Assumption 7 is only required if the estimator includes asymptotic trimming, otherwise it can be omitted as in e.g. when using fixed trimming. Assumption 7(ii) will be weakened in Theorem 2.2 below.

We assume that the weight function ϕ lies in a class Φ of real-valued measurable functions of X satisfying the following regularity condition:

Assumption 8 *The class Φ is a class of uniformly bounded functions such that $\log N_{[\cdot]}(\varepsilon, \Phi, \|\cdot\|_2) \leq C\varepsilon^{-v_\phi}$ for some $v_\phi < 2$.*

Assumption 8 restricts the size of the class Φ. The boundedness restriction in Assumption 8 can be relaxed by requiring instead suitable high order bounded moments for errors and weights.

For any generic measurable and integrable function $\phi(\cdot)$ define

$$\phi_W^{\perp}(X_i) := \phi(X_i) - E[\phi(X_i)|W(X)].$$

Define the parameter space $\mathcal{A} := \mathcal{W} \times \Phi$ and a generic element $\alpha := (W, \phi) \in \mathcal{A}$. We are interested in the asymptotic representation of the process (1), i.e.

$$\hat{\Delta}_n(\alpha) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(W_i|W)\} \widehat{t}_{ni}(W) \phi(X_i),$$

that is uniform over $\alpha \in \mathcal{A}$.

Define the error-weighted empirical process $\Delta_n(\alpha)$ as

$$\Delta_n(\alpha) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - m(W_i|W)\} \phi_W^{\perp}(X_i).$$

We prove below that $\hat{\Delta}_n$ and Δ_n are asymptotically uniformly equivalent. This provides a general uniform representation for $\hat{\Delta}_n(\alpha)$ in terms of iid variables. This uniform expansion quantifies the asymptotic effect from estimating true errors by nonparametric kernel regression residuals. To give some informal intuition for the asymptotic equivalence between $\hat{\Delta}_n$ and Δ_n, ignore trimming effects for now and write, for each $\alpha \in \mathcal{A}$,

$$\begin{aligned}\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(W_i|W)\} \phi(X_i) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(W_i|W)\} \phi_W^{\perp}(X_i) \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(W_i|W)\} E[\phi(X_i)|W(X)] =: I_{1n} + I_{2n}.\end{aligned}$$

Roughly, the term $I_{2n} = o_P(1)$ by a nonparametric version of the least squares normal equations, and the term I_{1n} is asymptotically equivalent to $\Delta_n(\alpha)$, because the latter when viewed as a mapping of m is (stochastically) continuous in m, since $\phi_W^{\perp}(X_i)$ is orthogonal to functions of W and m is a function of W. The following theorem formalizes this intuition and extends it to a uniform result in $\alpha \in \mathcal{A}$ and in the (possibly data dependent) bandwidth, while also allowing for random trimming.

Theorem 2.1 *Let Assumptions 1 – 8 hold. Then,*

$$\sup_{a_n \leq \widehat{h}_n \leq b_n} \sup_{\alpha \in \mathcal{A}} |\hat{\Delta}_n(\alpha) - \Delta_n(\alpha)| = o_{P^*}(1).$$

Remark 2.1 *If $\phi(X)$ is such that $E(\phi(X)|W) = 0$ a.s., so $\phi_W^{\perp} \equiv \phi$, then estimation of m has no asymptotic effect in the limit distribution of $\hat{\Delta}_n(\alpha)$.*

Theorem 2.1 has many applications for semiparametric inference as discussed in the Introduction, the next section and Appendix D.

2.1 Generated Regressors and Estimated Weights

In this section, we apply our uniform expansion to a setting where residuals are from a semiparametric index regression model with nonparametric generated regressors, and possibly nonparametrically estimated weights. Specifically, in this section we assume that

$$E[Y|X] = E[Y|v(g_0(X_1), X)] \text{ a.s.},$$

for some d-dimensional known function v of X and a conditional mean function $g_0(x_1) := E[D|X_1 = x_1]$, where D is a random variable and X_1 is a subvector of X, $X_1 \subseteq X$. Thus, the conditioning variable $W_0 := v(g_0(X_1), X)$ is known up to the unknown regression g_0. For notational simplicity we only consider univariate D but the extension to multivariate D is straightforward. Later in Section 3 we further extend the current setting to $W_0 = v(\theta_0, g_0(X_1), X)$ for an unknown finite-dimensional parameter $\theta_0 \in \Theta \subset \mathbb{R}^q$. To simplify the notation, denote $W_{0i} := v(g_0(X_{1i}), X_i)$, $m_{0i} := m(W_{0i}|W_0)$, $g_i := g(X_{1i})$ and $g_{0i} := g_0(X_{1i})$, for $i = 1, ..., n$.

We observe a random sample $\{Y_i, X_i^{\top}, D_i\}_{i=1}^n$ from the joint distribution of $(Y, X^{\top}, D)$ and estimate g_0 by some nonparametric estimator $\widehat{g}$, possibly but not necessarily a kernel estimator. Let ϕ_0 denote a weight function, and let $\widehat{\phi}$ denote a $\|\cdot\|_2$-consistent estimator for ϕ_0. The estimator $\widehat{\phi}$ can be nonparametric, e.g. a kernel or series estimator. We investigate the impact of estimating W_0 by $\widehat{W} = v(\widehat{g}(X_1), X)$ and then estimating ϕ_0 by $\widehat{\phi}$ in the empirical process $\hat{\Delta}_n(\widehat{W}, \widehat{\phi})$. The goal is to provide an expansion in iid terms for the standardized sample mean of weighted and trimmed residuals

$$\hat{\Delta}_n(\widehat{\alpha}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(\widehat{W}_i|\widehat{W})\} \widehat{t}_{ni}(\widehat{W}) \widehat{\phi}(X_i), \tag{3}$$

where $\widehat{\alpha} := (\widehat{W}, \widehat{\phi})$ and $\widehat{W}_i := v(\widehat{g}(X_{1i}), X_i)$. Define $\phi_0^{\perp}(X_i) := \phi_0(X_i) - E[\phi_0(X_i)|W_{0i}]$, $\varepsilon_i := Y_i - m_{0i}$ and $u_i := D_i - g_{0i}$. We show that under regularity conditions

$$\hat{\Delta}_n(\widehat{\alpha}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \varepsilon_i \phi_0^{\perp}(X_i) - u_i E[\partial_{\bar{g}} m(W_{0i}) \phi_0^{\perp}(X_i)|X_{1i}] \right\} + o_P(1), \tag{4}$$

where $\partial_{\bar{g}} m(W_{0i}) := \partial m(v(\bar{g}, X_i)|W_0)/\partial \bar{g}|_{\bar{g}=g_{0i}}$.

A convenient feature of the expansion in (4) is that it does not require an analysis of pathwise functional derivatives, as in, e.g., Newey (1994). This is particularly useful here because the map $g \to E[Y|v(g(X_1), X)]$ has no closed form and can be highly non-linear. Note that $\partial_{\bar{g}} m(W_{0i})$ is a standard (finite-dimensional) derivative of the regression involving the 'true' index W_0, and the derivative is with respect to the evaluation point.

To better understand the expansion in (4) and how we use Theorem 2.1 to obtain it, note that denoting $\widehat{\phi}^{\perp}_{\widehat{W}}(X) := \widehat{\phi}(X) - E[\widehat{\phi}(X_i)|\widehat{W}]$, we can write

$$\begin{aligned}
\hat{\Delta}_n(\widehat{\alpha}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Y_i - m(\widehat{W}_i|\widehat{W})\}\widehat{\phi}^{\perp}_{\widehat{W}}(X_i) + o_P(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\widehat{\phi}^{\perp}_{\widehat{W}}(X_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g_{0i} - \widehat{g}_i)\partial_{\bar{g}}m(W_{0i})\widehat{\phi}^{\perp}_{\widehat{W}}(X_i) \qquad (5) \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(W_{0i}|W_0) - m(\widehat{W}_i|\widehat{W}) - (g_{0i} - \widehat{g}_i)\partial_{\bar{g}}m(W_{0i})\}\widehat{\phi}^{\perp}_{\widehat{W}}(X_i) + o_P(1),
\end{aligned}$$

where the first equality follows from Theorem 2.1 and $P(\widehat{\alpha} \in \mathcal{A}) \to 1$. We then show that the third term in the last expansion is asymptotically negligible. To see this, notice that by adding and subtracting $m(W|W_0)$,

$$\begin{aligned}
&E\left[\{m(W_0|W_0) - m(W|W) - (g_0 - g)\partial_{\bar{g}}m(W_0)\}\phi^{\perp}_{W}(X)\right] \\
&= E\left[\{m(W_0|W_0) - m(W|W_0) - (g_0 - g)\partial_{\bar{g}}m(W_0)\}\phi^{\perp}_{W}(X)\right] \\
&+ E\left[\{m(W|W_0) - m(W|W)\}\phi^{\perp}_{W}(X)\right] \\
&= E\left[\{m(W_0|W_0) - m(W|W_0) - (g_0 - g)\partial_{\bar{g}}m(W_0)\}\phi^{\perp}_{W}(X)\right],
\end{aligned}$$

where the second equality follows from the orthogonality of $\phi^{\perp}_{W}(X)$ and functions of $W(X)$. The absolute value of the last expectation is shown to be of the order $\|g_0 - g\|^2_{\infty}$ (and it can be made of smaller order by accounting for higher order derivatives, see the proof of Theorem 2.2 below). These simple equalities show, without the need to introduce functional derivatives, that there is zero contribution from estimating W_0 in $m(W|W_0)$, and that there is a trade-off between smoothness in $m(v(\cdot, X)|W_0)$ and rates of convergence of $\|g_0 - g\|_{\infty}$. The rest of the proof of the expansion (4) then follows from stochastic equicontinuity of the first and second terms of (5) at α_0, as shown below.

To handle the second summand in (5) for a generic first step estimator $\widehat{g}$, and accounting for higher derivatives in $m(v(\cdot, X)|W_0)$, we define the empirical processes, for $j = 1, ..., \kappa$, $\kappa \geq 1$, and $\alpha = (W, \phi)$,

$$\begin{aligned}
R_{jn}(\alpha) :=& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g_{0i} - g_i)^j\partial^{(j)}_{\bar{g}}m(W_{0i})\phi^{\perp}_{W}(X_i), \text{ and} \\
G_n(\alpha) :=& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_iE[\partial_{\bar{g}}m(W_{0i})\phi^{\perp}_{W}(X_i)|X_{1i}], \qquad (6)
\end{aligned}$$

where $\partial^{(j)}_{\bar{g}}m(W_{0i}) := \partial^j m(v(\bar{g}, X_i)|W_0)/\partial\bar{g}^j\big|_{\bar{g}=g_{0i}}$. Define also the rates $p_{0n} := P(f\,(W_0|\,W_0) \leq 2\tau_n)$, $w_n := \|\widehat{g} - g_0\|_{\infty}$ and $q_n := \tau_n^{-1}d_n + w_n$.

Assumption 9 *The function $m(v(\bar{g}, x)|W_0)$ is κ-times continuously differentiable in $\bar{g}$ with bounded derivatives, for all x.*

Assumption 10 *The estimator $\widehat{\alpha}$ is such that: (i) $|R_{1n}(\widehat{\alpha}) + G_n(\widehat{\alpha})| = o_P(1)$ and $|R_{jn}(\widehat{\alpha})| = o_P(1)$ for all $j = 2, ..., \kappa$; (ii) $E[D^2|X] < C$ a.s., $w_n = o_P(n^{-1/2\kappa})$, $W_0 \in \mathcal{W}$ and $P(\widehat{W} \in \mathcal{W}) \to 1$; and (iii) $\|\,\widehat{\phi} - \phi_0\,\|_2 = o_P(1)$, $\phi_0 \in \Phi$, and $P(\widehat{\phi} \in \Phi) \to 1$.*

Assumption 11 *τ_n is a sequence of positive numbers satisfying $\tau_n \to 0$, $\lambda_n^2 a_n^{-d} \log n \to 0$ and $n(\tau_n^{-4} q_n^6 + q_n^4 + p_{0n}^2) \to 0$.*

Assumption 9 is a standard smoothness condition. Assumption 10(i) is a high level assumption regarding expansions related to the first step estimator. Appendix C provides primitive conditions for Assumption 10 when $\widehat{g}$ is a NW kernel estimator of g_0. Alternative primitive conditions for series estimators can be found in Escanciano and Song (2010).

Assumption 10(ii) describes the trade-off between number of derivatives in $m(v(\cdot, x)|W_0)$ and the required rate for $\|\widehat{g} - g_0\|_\infty$. Similar smoothness trade-offs are noted by Mammen, Rothe and Schienle (2011a). See also Cattaneo, Crump, and Jansson (2011) for a related finding in a different context. The high level assumption 10(iii) can be replaced by rates of convergence on $\| \widehat{\phi} - \phi_0 \|_2$ as shown in the examples of Section 3. As mentioned earlier, Assumption 11 replaces and relaxes Assumption 7(ii). The condition on p_{0n} can be relaxed to $nq_n^2 p_{0n}^2 \to 0$ if, for instance, $\{\varepsilon_i\}_{i=1}^n$ are conditionally uncorrelated given $\{\widehat{\alpha}_i\}_{i=1}^n$.

Theorem 2.2 *Let Assumptions 1 – 7(i) and 8 – 11 hold. Assume that $E[Y|X] = E[Y|W_0]$ a.s. Then the expansion in (4) holds uniformly in $a_n \leq \widehat{h}_n \leq b_n$.*

Theorem 2.2 quantifies the estimation effect of $\widehat{\alpha}$, that is, $\widehat{g}$ and $\widehat{\phi}$, in the empirical process $\hat{\Delta}_n(\widehat{\alpha})$. Although W depends on g_0, the Theorem shows that the estimation error $\widehat{g} - g_0$ only has an asymptotic impact through the first argument in $m(W_i|W)$. For the process $\hat{\Delta}_n(\widehat{\alpha})$, the contribution from the second argument is asymptotically negligible due to the orthogonality of $\phi_W^{\perp}$ with functions of W.

To illustrate the usefulness of Theorem 2.1 and Theorem 2.2 for estimation and testing, and to show how they would be applied in practice, in the next section we use them to derive asymptotic theory for a new estimator of a binary choice model with selection, and then we apply them to the construction of a new directional specification test.

3 Example: A Binary Choice Model with Selection

Suppose a latent binary variable Y^* satisfies the ordinary threshold crossing binary response model $Y^* = \mathbb{I}\left(X^\top \theta_0 - e \geq 0\right)$ with e independent of X, in short $e \perp X$, and the distribution function of e, F_e, may be unknown. Suppose further that we only observe Y^* for some subset of the population, indexed by a binary variable D, i.e. we only observe $Y = Y^* D$. This is a sample selection model with a binary outcome. The econometrician is assumed to know relatively little about selection D other than that it is binary, so let D be given by the nonparametric threshold crossing model $D = \mathbb{I}[g_0(X) - u \geq 0]$ where $u \perp X$ and the function $g_0(X)$ is unknown. Based on Matzkin (1992), we may without loss of generality assume $g_0(X) = E[D|X]$ and u has a uniform distribution, since then $P(D = 1|X) = P[u \leq g_0(X)] = g_0(X)$.

We then have the model

$$D = \mathbb{I}[g_0(X) - u \geq 0] \tag{7}$$

$$Y = \mathbb{I}(X^\top \theta_0 - e \geq 0) D \tag{8}$$

The latent error terms e and u are not independent of each other, so the model does not have selection on observables. When $g_0(\cdot)$ is assumed to be linear in X, e.g. $g_0(X) = X^\top \delta_0$ for some unknown coefficient vector δ_0, and the joint distribution of the errors $(e,u)^\top$ is assumed to be normal, model (7) – (8) is known as the 'Censored Probit Model' or a 'Probit Model with Sample Selection,' see e.g. van de Ven and van Praag (1981) and Meng and Schmidt (1985). Newey (2007) discusses its semiparametric identification within a larger class of models.[1]

Let $(e,u)^\top$ be drawn from an unknown joint distribution function $F(e,u)$ with $e, u \perp X$. Then $g_0(X) = E[D|X]$ and

$$E[Y|X] = m[X^\top \theta_0, g_0(X)]$$

so an index restriction with $W_0 := v(\theta_0, g_0, X) = (X^\top \theta_0, g_0(X))$ holds, and $g_0(X)$ is identified from the selection equation as a conditional expectation. Identification of m and θ_0 in this model, which is possible even without an exclusion restriction, has been studied in Escanciano, Jacho-Chávez and Lewbel (2011), who also propose a semiparametric least squares estimator for this model.[2]

The class of functions

$$\mathcal{W} = \{x \to (x^\top \theta, g(x)) : \theta \in \Theta_0 \subset \mathbb{R}^q,\ g \in \mathcal{G} \subset C_M^{\eta_g}(\mathcal{X}_X),\ \|g - g_0\|_\infty < \delta\}, \tag{9}$$

for an arbitrarily small $\delta > 0$ and $\eta_g > p$ is used for the remaining part of this section.

3.1 Semiparametric Maximum Likelihood Estimation

Following Klein and Spady (1993), we propose a semiparametric maximum likelihood estimator (SMLE) of θ_0 in model (7) – (8), and apply our earlier results to obtain limiting distribution theory for this estimator. Firstly, we have $Y = D = 0$ if $u > g_0(X)$, $Y = D = 1$ if both $e \leq X^\top \theta_0$ and $u \leq g_0(X)$, and otherwise $Y = 0$ and $D = 1$. Therefore, $P(Y = D = 0|X) = P(D = 0|X) = 1 - E[D|X] = 1 - g_0(X)$, $P(Y = D = 1|X) = P(Y = 1|X) = E[Y|X] = m[X^\top \theta_0, g_0(X)]$ and $P(Y = 0, D = 1|X) = 1 - P(Y = D = 0|X) - P(Y = D = 1|X) = E[D|X] - E[Y|X] = g_0(X) - m[X^\top \theta_0, g_0(X)]$. Based on these probabilities, define the following semiparametric log-likelihood objective function

$$\mathcal{L}_n(\theta, \widehat{g}) := \frac{1}{n} \sum_{i=1}^n \{Y_i \log[\widehat{m}_{i\theta}] + (D_i - Y_i) \log[\widehat{g}_i - \widehat{m}_{i\theta}]\} \widetilde{t}_{in} \psi_i, \tag{10}$$

where $\widehat{g}_i := \widehat{g}(X_i)$ is the NW estimator of g_0 with possibly data-driven bandwidth $\widehat{h}_{gn}$, $\widehat{m}_{i\theta} := \widehat{m}(W_i(\theta, \widehat{g})|W(\theta, \widehat{g}))$, $W(\theta, g) := (X^\top \theta, g(X))$, $W_i(\theta, g) := (X_i^\top \theta, g(X_i))$ with possible data-driven bandwidth $\widehat{h}_n$, and $\widetilde{t}_{in}$ is a trimming sequence that also accounts for the possibility that $\widehat{m}$ is close to

[1] Our methods could also be applied to other related models, e.g., if we replaced (8) with $Y = \left(X^\top \theta_0 - e\right) D$, then this would be a semiparametric generalization of the standard Heckman selection model, and if we replaced (8) with $Y = \max\left(X^\top \theta_0 - e, 0\right) D$ then this would be a semiparametric generalization of Cragg's (1971) double hurdle model.

[2] Closely related identification and estimation results include Blundell and Powell (2004) and Ichimura and Lee (1991). If instead of the assumption $e, u \perp X$ we had the more general assumption $u \perp X$ and $e|u, X \sim e|u, W_0$, then the above model would still hold with $F_{eu}(e, u|W_0)$ denoting the conditional distribution of $e, u|W_0$ and the function $m(r, g)$ now defined as $m(r, g) = F_{eu}(r, g|W_0)$.

zero or to $\widehat{g}_i$. More specifically, the trimming has the form

$$\widetilde{t}_{in} := \mathbb{I}(\tau_n \leq \widetilde{m}_i \leq \widehat{g}_i - \tau_n) \times \mathbb{I}(\tau_n \leq \widetilde{f}_i),$$

where τ_n is a sequence of positive numbers with $\tau_n \to 0$ as $n \to \infty$, $\widetilde{m}_i := \widehat{m}(\widetilde{W}_i|\widetilde{W})$, $\widetilde{f}_i := \widehat{f}(\widetilde{W}_i|\widetilde{W})$, $\widetilde{W}_i := W_i(\widetilde{\theta}, \widehat{g})$, $\widetilde{W} := W(\widetilde{\theta}, \widehat{g})$, and $\widetilde{\theta}$ is a preliminary consistent estimator for θ_0. For instance, $\widetilde{\theta}$ can be a SMLE with $\psi_i = 1$ and fixed trimming $\widehat{t}_{ni} = \mathbb{I}(X_i \in A)$ for a compact set $A \subset \mathcal{X}_X$ (or non-data-dependent asymptotic trimming). Note that weighting $\psi_i = 1$ and either fixed or non-data-dependent asymptotic trimming will in general make $\widetilde{\theta}$ inefficient, but that does not violate the required Assumption 14 below for an initial consistent $\widetilde{\theta}$. The estimator for θ_0 we propose is

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}_n \left(\theta, \widehat{g}\right). \tag{11}$$

The estimator $\widehat{\theta}$ extends the related estimator in Klein and Spady (1993) for the single-index binary choice model in two ways. First, the objective function (11) has a nonparametric generated regressor $\widehat{g}$ associated with selection, which complicates the relevant asymptotic theory. Second, and intimately related to the first, is that unlike in Klein and Spady (1993), adaptive weighting is necessary here to improve efficiency due to the presence of the generated regressor.

Sufficient conditions for identification of θ_0 in this model (which do not require exclusion restrictions) are provided by Escanciano, Jacho-Chávez and Lewbel (2011), and given identification it is straightforward to demonstrate consistency of $\widehat{\theta}$. Given consistency, we now apply the results of the previous section to derive limiting distribution theory for $\widehat{\theta}$, allowing for data dependent choice of bandwidth, data dependent asymptotic trimming, and data dependent adaptive weighting for efficiency.

Recall $\varepsilon_i = Y_i - m_{0i}$ and $u_i = D_i - g_{0i}$. Further, define $v_i := \varepsilon_i - u_i \partial_{\bar{g}} m(W_{0i})$, $\sigma^2_{0i} := E[v_i^2|X_i]$, $\psi_i := \psi(W_{0i})$ and $\partial_\theta m(W_{0i}) := \partial m(W_i(\theta, g_0)|W(\theta, g_0))/\partial\theta|_{\theta=\theta_0}$. Also note that

$$\sigma^2_{0i} = m_{0i}(1 - m_{0i}) + \left(\partial_{\bar{g}} m(W_{0i})\right)^2 g_{0i}(1 - g_{0i}) - 2\partial_{\bar{g}} m(W_{0i}) m_{0i}(1 - g_{0i}). \tag{12}$$

We shall assume that the following matrix is non-singular and finite (this is little more than a linear index model identification condition),

$$\Gamma_0 := E\left[\frac{g_{0i}\partial_\theta m(W_{0i})\partial_\theta^\top m(W_{0i})}{m_{0i}(g_{0i} - m_{0i})}\psi_i\right]. \tag{13}$$

Define the rates $q_n' := d_n' + w_n$ and $\tau_{ng} := \inf_{\{x: f(W_0(x)|W_0) < 2\tau_n\}} f_X(x)$, where

$$d_n' := \sqrt{\frac{\log a_n^{-2} \vee \log\log n}{n a_n^4}} + b_n^{r-1}.$$

Finally, to simplify the notation define $m_\theta := m(W(\theta, g_0)|W(\theta, g_0))$.

Assumption 12 *(i) The kernel function satisfying Assumption 4 also satisfies* $\left|\partial^{(j)} k(t)/\partial t^j\right| \leq C\,|t|^{-v}$ *for* $|t| > L_j$, $0 < L_j < \infty$, *for* $j = 1, 2$*; (ii) the sequence* τ_n *is such that* $\tau_n \to 0$, $n\tau_n^2 \to \infty$, $n\tau_n^{-2} q_n'^4 \to 0$ *and* $n q_n'^2 q_n^2 \to 0$ *(iii) The functions* $\sigma^2(\cdot)$, $\inf_{\theta \in \Theta_0} m_\theta$ *and* $\inf_{\theta \in \Theta_0}(g_0 - m_\theta)$ *are bounded away from zero.*

Assumption 13 *(i) The regression function $g_0(X) = E[D|X]$ is estimated by a NW kernel estimator $\widehat{g}$ with a kernel function satisfying Assumption 4 with $r = \rho$ and a possibly stochastic bandwidth $\widehat{h}_{gn}$ satisfying $P(l_n \leq \widehat{h}_{gn} \leq u_n) \to 1$ as $n \to \infty$, for deterministic sequences of positive numbers l_n and u_n such that $u_n \to 0$, $n\tau_{ng}^2(l_n^p/\log n)^{\kappa/(\kappa-1)} \to \infty$ and $n\tau_{ng}^{-2}u_n^{2\rho} \to 0$; (ii) the function g_0 and the density $f_X(\cdot)$ of X are ρ-times continuously differentiable in x, with bounded derivatives. Furthermore, g_0 is bounded away from zero, $g_0 \in \mathcal{G} \subset C_M^{\eta_g}(\mathcal{X}_X)$ and $P(\widehat{g} \in \mathcal{G}) \to 1$ for some $\eta_g > p$.*

Assumption 14 *The parameter space Θ_0 is a compact subset of $\mathbb{R}^p$ and θ_0 is an element of its interior. The estimator $\widetilde{\theta}$ is $\sqrt{n}-$consistent for θ_0 and $\widehat{\theta}$ is consistent. The matrix Γ_0 is non-singular and $E\left[\psi_1^2\right] < \infty$.*

Theorem 3.1 *Let Assumption 1 hold for model (7) – (8), and let Assumptions 3 – 7(i), 9, 11, 12 – 14 hold. Then $\widehat{\theta}$ is asymptotically normal, i.e.*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \longrightarrow_d N(0, \Gamma_0^{-1}\Sigma_0\Gamma_0^{-1}),$$

where Γ_0 is given in (13) and

$$\Sigma_0 := E\left[\frac{\sigma^2(W_{0i})g_{0i}^2\partial_\theta m(W_{0i})\partial_\theta^\top m(W_{0i})}{m_{0i}^2(g_{0i} - m_{0i})^2}\psi_i^2\right].$$

Remark 3.1 *A sufficient primitive condition for the high-level Assumption 6(ii) is that $\tau_n^2 a_n^6 n/\log(n) \to \infty$; see Appendix C.*

Remark 3.2 *Consider a bandwidth of the form $\widehat{h}_n = ch_n$, with c a constant to be chosen and h_n a suitable deterministic sequence satisfying the assumptions in Theorem 3.1 above. Then, a natural data-driven choice for the constant c is one that maximizes an estimated semiparametric likelihood criterion, i.e.*

$$\widehat{c}_n = \underset{c \in [\epsilon, \epsilon^{-1}] : \widehat{h}_n = ch_n}{\arg\max} \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i \log[\widetilde{m}_{i;\widehat{h}_n}] + (D_i - Y_i)\log[\widehat{g}_i - \widetilde{m}_{i;\widehat{h}_n}]\right\}\mathbb{I}(X_i \in A),$$

where ϵ is an arbitrarily small positive number, and we have made explicit the dependence of the leave-one-out version of estimator $\widetilde{m}_i$ on the bandwidth $\widehat{h}_n$. Note that the resulting bandwidth $\widehat{c}_n h_n$ will automatically satisfy our required assumptions by construction. Furthermore, when using this choice in (10) no changes to Theorem 3.1 are needed by virtue of our uniformity-in-bandwidth results.

It can be easily shown that using weights $\psi_i^* := m_{0i}[g_{0i} - m_{0i}]/[\sigma_{0i}^2 g_{0i}]$ leads to a more efficient estimator[3] with asymptotic variance Γ_*^{-1}, where the positive definite matrix Γ_* is given by

$$\Gamma_* := E\left[\frac{\partial_\theta m(W_0)\partial_\theta^\top m(W_0)}{\sigma^2(W_0)}\right].$$

Now let $\widehat{\psi}_i^* = \widetilde{m}_i(\widehat{g}_i - \widetilde{m}_i)/(\widehat{\sigma}_i^2\widehat{g}_i)$, where $\widehat{\sigma}_i^2 = \widetilde{m}_i(1 - \widetilde{m}_i) + (\partial_{\bar{g}}\widetilde{m}_i)^2\widehat{g}_i(1 - \widehat{g}_i) - 2\partial_{\bar{g}}\widetilde{m}_i\widetilde{m}_i(1 - \widehat{g}_i)$ and $\partial_{\bar{g}}\widetilde{m}_i := \partial\widehat{m}(W_i(\widetilde{\theta}, \bar{g})|W(\widetilde{\theta}, \widehat{g}))/\partial\bar{g}|_{\bar{g} = \widehat{g}_i}$. Similarly, let $\widehat{\theta}^*$ be the resulting estimator when the optimal weight $\psi_i = \widehat{\psi}_i^*$ is used in (10). Furthermore, let $\widehat{\Gamma}_* = n^{-1}\sum_{i=1}^n \partial_\theta\widehat{m}_{i\widehat{\theta}^*}\partial_\theta^\top\widehat{m}_{i\widehat{\theta}^*}/\widehat{\sigma}_i^2$, where $\partial_\theta\widehat{m}_{i\widehat{\theta}^*} := \partial\widehat{m}(W_i(\theta, \widehat{g})|W(\theta, \widehat{g}))/\partial\theta|_{\theta = \widehat{\theta}^*}$.

[3] Whether Γ_*^{-1} coincides with the semiparametric efficiency bound of θ_0 in model (7) – (8) is an open question.

Corollary 3.1 *Let the Assumptions of Theorem 3.1 hold. Then* $\sqrt{n}(\widehat{\theta}^{*}-\theta_0)\longrightarrow_d N(0,\Gamma_*^{-1})$*, and* $\widehat{\Gamma}_*\to_P\Gamma_*$.

The proof of Corollary 3.1 is provided in the Appendix A. The proof of Theorem 3.1 itself is almost the same (except simpler, since it does not involve estimated weights), and so is omitted. It is shown that the asymptotic distribution of $\widehat{\theta}^{*}$ is the same as it would be if the optimal weights ψ_i^* and the regression function m were known instead of estimated.

3.2 A Tailor-Made Specification Test

We next illustrate the usefulness of our uniform convergence results for constructing test statistics. A policy parameter of considerable interest in applications of binary choice models is the average structural function (ASF). See, e.g., Stock (1989), Blundell and Powell (2004) and Newey (2007). In our binary choice model with selection, the ASF is given by

$$\gamma^*:=\int\int m(x^\top\theta_0,g)f_g(g)dF_X^*(x)dg,$$

where F_X^* is a particular marginal distribution for X and f_g is the density of g_0. In this context, suppose we are concerned with possible misspecification of the semiparametric binary choice model only to the extent that it leads to inconsistent estimation of the ASF γ^*. Our goal is construction both of a test and an associated bandwidth choice procedure that concentrates power in this direction.

Consider a directional specification test with these alternatives in mind, testing the correct specification of the model

$$H_0:E[Y|X]=m[X^\top\theta_0,g_0\left(X\right)]\text{ a.s,}$$

against alternatives for which

$$E[\{Y-m(X^\top\theta_0,g_0\left(X\right))\}\phi_*(X,g_0(X))]\neq 0, \tag{14}$$

where $\phi_*(X,g_0(X)):=f_g(g_0(X))dF_X^*(X)/f\left(W_0|\,W_0\right)$.

We propose constructing such a test based on

$$T_{n,h_n}:=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Y_i-\widehat{m}(\widehat{W}_i|\widehat{W}_i)\}\widehat{\phi}_*(X_i,\widehat{g}_i)\widehat{t}_{ni},$$

where $\widehat{W}_i:=(X_i^\top\widehat{\theta},\widehat{g}_i)$, $\widehat{\theta}$ is a $\sqrt{n}$-consistent estimator for θ_0, such as (11) from the previous section, $\widehat{\phi}_*(X_i,\widehat{g}_i)=\widehat{f}_{ig}(\widehat{g}_i)dF_X^*(X_i)/\widehat{f}_i$, $\widehat{f}_{ig}$ is a kernel estimator for the density of g_0 resulting from integrating $\widehat{f}_i\equiv\widehat{f}(\widehat{W}_i|\widehat{W}_i)$, $\widehat{t}_{in}=\mathbb{I}(\widehat{f}_i\geq\tau_n)$ and h_n in T_{n,h_n} denotes the bandwidth used in estimating $\widehat{m}$. Set $\sigma^2:=E[\varepsilon_i^2\phi_*^{\perp 2}(X_i,g_{0i})]$, and consider the variance estimator

$$\widehat{\sigma}^2:=\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\widehat{m}(\widehat{W}_i|\widehat{W})\}^2\widehat{\phi}_*^{\perp 2}(X_i,\widehat{g}_i)\widehat{t}_{ni},$$

where $\widehat{\phi}_*^{\perp}$ is based on a uniformly consistent estimator of $E[dF_X^*(X_i)|W_i]$.

Theorem 3.2 *Let Assumption 1 hold for model (7) – (8), and let Assumptions 3 – 7(i), 9, 11 and 13 – 14 hold. Assume also that $\phi_*(\cdot, g_0)$ is bounded. Then, under H_0,*

$$\widehat{\sigma}^{-1} \max_{a_n \leq h_n \leq b_n} T_{n,h_n} \longrightarrow_d N(0,1),$$

whereas under the alternative (14), $\widehat{\sigma}^{-1} \left|\max_{a_n \leq h_n \leq b_n} T_n\right| \longrightarrow_P \infty$.

Remark 3.3 *Let $\widehat{h}_n$ denote the solution to the optimization problem $\max_{a_n \leq h_n \leq b_n} T_{n,h_n}$. Our uniform in bandwidth theorems allow us to choose the bandwidth by this optimization, which leads to a test with better power properties than a test that uses a bandwidth $\widehat{h}_n$ optimized for estimation like that described in remark 3.2. See e.g. Horowitz and Spokoiny (2001) for a related approach in a different context.*

4 Concluding Remarks

We have obtained a new uniform expansion for standardized sample means of weighted regression residuals from nonparametric or semiparametric models, with possibly nonparametric generated regressors. The expansion is uniform in the generated regressor, random bandwidth, and the weights. We have shown by examples how these results are useful for deriving limiting distribution theory for estimators and tests. Additional example applications of our uniform expansions are provided in Appendix D to this paper.

For estimation, we showed that a simple data driven bandwidth choice procedure could be used where the rate is chosen by the practitioner based on theory and the constant is chosen by minimizing the same objective function that is for estimation. A topic for future research is consideration of more general selection rules where the rate might also be chosen by minimizing some estimation criterion.

Less is known about optimal bandwidth rates for testing. We choose the bandwidth to maximize the test statistic in the region of admissible bandwidths, and show that this choice is permitted by our asymptotic results.

The appealing properties of our estimators and tests regarding data driven bandwidths, possibly nonparametric generated regressors, random trimming and estimated weights are made possible by the use of uniform convergence in these aspects. We have shown how this uniform convergence, when combined with standard stochastic equicontinuity arguments, allows us to establish the desired expansions without the need to introduce functional derivatives, which can be difficult to deal with in these contexts.

Our results should have applications beyond the types considered here. For example, expansions of the kind provided by Theorem 2.1 and Theorem 2.2 are the key ingredient in proving the consistency of bootstrap procedures for estimation and testing in semiparametric models.

Appendix A Main Proofs

Before we prove our main results we need some preliminary results from empirical processes theory. Define the generic class of measurable functions $\mathcal{G} := \{z \to m(z,\theta,h) : \theta \in \Theta, h \in \mathcal{H}\}$, where Θ and $\mathcal{H}$ are endowed with the pseudo-norms $|\cdot|_\Theta$ and $|\cdot|_\mathcal{H}$, respectively.

Lemma A.1 *Assume that for all $(\theta_0, h_0) \in \Theta \times \mathcal{H}$, $m(z, \theta, h)$ is locally uniformly $L_2(P)$ continuous, in the sense that*

$$E\left[\sup_{\theta:|\theta_0-\theta|_\Theta<\delta, h:|h_0-h|_{\mathcal{H}}<\delta} |m(Z,\theta,h)-m(Z,\theta_0,h_0)|^2\right] \leq C\delta^s,$$

for all sufficiently small $\delta > 0$, and some constant $s \in (0, 2]$. Then,

$$N_{[\cdot]}(\varepsilon, \mathcal{G}, \|\cdot\|_2) \leq N\left(\left(\frac{\varepsilon}{2C}\right)^{2/s}, \Theta, |\cdot|_\Theta\right) \times N\left(\left(\frac{\varepsilon}{2C}\right)^{2/s}, \mathcal{H}, |\cdot|_{\mathcal{H}}\right).$$

Proof of Lemma A.1: The proof can be found as part of the proof of Theorem 3 in Chen, Linton, and van Keilegom (2003, p. 1597), and therefore is omitted. *Q.E.D.*

Lemma A.2 *Let Assumptions 3 and 8 hold. Then, for each W_1 and W_2 in $\mathcal{W}$, and all $\delta > 0$,*

$$\sup_{\phi\in\Phi} \|E[\phi(X)|W_1(X) = W_1(\cdot)] - E[\phi(X)|W_2(X) = W_2(\cdot)]\|_\infty \leq C\|W_1 - W_2\|_\infty.$$

Proof of Lemma A.2: The proof follows from Lemma A2(ii) in Song (2008), noting that Assumption 3 implies his condition (A.35) with $s = 1$. *Q.E.D.*

Let $\mathcal{S}$ be a class of measurable functions of X. Let $\{\xi_i, X_i^\top\}_{i=1}^n$ denote a random sample from the joint distribution of $(\xi, X^\top)$ taking values in $\mathcal{X}_\xi \times \mathcal{X}_X \in \mathbb{R}^{1+p}$, and define the weighted empirical process, indexed by $s \in \mathcal{S}$,

$$\Psi_n(s) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i s(X_i) - E[\xi_i s(X_i)].$$

We say that Ψ_n is asymptotically uniformly ρ-equicontinuous at $s_0 \in \mathcal{S}$, for a pseudo-metric ρ on $\mathcal{S}$, if for all $\varepsilon > 0$ and $\eta > 0$, there exists $\delta > 0$ such that

$$\limsup_{n\to\infty} P^*\left[\sup_{s_1\in\mathcal{S}:\rho(s_1,s_0)<\delta} |\Psi_n(s_1) - \Psi_n(s_0)| > \varepsilon\right] \leq \eta.$$

The following result gives sufficient conditions for uniform $\|\cdot\|_2$-equicontinuity of Ψ_n. One important implication of the uniform equicontinuity is that $\Psi_n(\widehat{s}) = \Psi_n(s_0) + o_P(1)$, provided $\|\widehat{s} - s_0\|_2 = o_P(1)$.

Lemma A.3 *Assume $E[\xi_i^2|X_i] < L$ a.s., and let $\mathcal{S}$ be a class of uniformly bounded functions such that $\log N_{[\cdot]}(\varepsilon, \mathcal{S}, \|\cdot\|_2) \leq C\varepsilon^{-v_s}$ for some $v_s < 2$. Then, Ψ_n is asymptotically uniformly $\|\cdot\|_2$-equicontinuous at $s_0 \in \mathcal{S}$, for all s_0.*

Proof of Lemma A.3: Define the class of functions $\mathcal{G} := \{(\xi, x) \to \xi s(x) : s \in \mathcal{S}\}$. Let $a^+ := \max\{a, 0\}$ and $a^- := \max\{-a, 0\}$ denote the positive and negative parts of a, respectively. Let $\{[s_{lj}, s_{uj}] : j = 1, ..., N_\varepsilon \equiv N_{[\cdot]}(\varepsilon, \mathcal{S}, \|\cdot\|_2)\}$ be a family of ε-brackets (with respect to $\|\cdot\|_2$) covering $\mathcal{S}$. Then, it holds that $\{[\xi^+ s_{lj} - \xi^- s_{uj}, \xi^+ s_{uj} - \xi^- s_{lj}] : j = 1, ..., N_\varepsilon\}$ is also a family of $L^{1/2}\varepsilon$-brackets covering $\mathcal{G}$. Then, by our assumptions, $\mathcal{G}$ has finite bracketing entropy, and hence, Ψ_n is

$\|\cdot\|_2$-equicontinuous at all points in $\mathcal{S}$. *Q.E.D.*

Throughout the Appendix we use the notation, for $i = 1, ..., n$, and $W \in \mathcal{W}$,

$$t_{ni}(W) := \mathbb{I}(f\,(W_i|\,W) \geq \tau_n/2) \text{ and } \Delta t_{ni}(W) := \widehat{t}_{ni}(W) - t_{ni}(W). \tag{A-15}$$

Similarly, define $\widehat{t}_{ni} \equiv \widehat{t}_{ni}(\widehat{W})$ and $\Delta t_{ni} := \widehat{t}_{ni} - t_{ni}(\widehat{W})$.

Lemma A.4 *Under the Assumptions of Theorem 2.1,* $\sup_{\alpha \in \mathcal{A}} |R_n(\alpha)| = o_{P^*}(1)$*, where*

$$R_n(\alpha) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(W_i|W)\} \Delta t_{ni}(W) \phi(X_i).$$

Proof of Lemma A.4: Write

$$\begin{aligned} R_n(\alpha) &:= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - m(W_i|W)\} \Delta t_{ni}(W) \phi(X_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{m(W_i|W) - \widehat{m}(W_i|W)\} \Delta t_{ni}(W) \phi(X_i) \\ &:= R_{1n}(\alpha) + R_{2n}(\alpha). \end{aligned}$$

We shall prove that $\sup_{\alpha \in \mathcal{A}} |R_{1n}(\alpha)| = o_{P^*}(1)$. By Cauchy inequality

$$\frac{1}{\sqrt{n}} |R_{1n}(\alpha)| \leq \left(\frac{1}{n} \sum_{i=1}^{n} |\Delta t_{ni}(W)| \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(W_i|W)\}^2 |\Delta t_{ni}(W)| \, \phi^2(X_i) \right)^{1/2}.$$

Thus, using the simple inequalities, with $f_i\,(W) \equiv f\,(W_i|\,W)$ and $\widehat{f}_i\,(W) \equiv \widehat{f}\,(W_i|\,W)$,

$$|\Delta t_{ni}(W)| \leq \mathbb{I}(f_i\,(W) \geq \tau_n/2) \mathbb{I}(\widehat{f}_i\,(W) < \tau_n) + \mathbb{I}(\widehat{f}_i\,(W) \geq \tau_n) \mathbb{I}(|\widehat{f}_i\,(W) - f_i\,(W)\,| > \tau_n/2), \tag{A-16}$$

$$\mathbb{I}(\widehat{f}_i\,(W) < \tau_n) \leq \mathbb{I}(f_i\,(W) \leq 2\tau_n) + \mathbb{I}(|\widehat{f}_i\,(W) - f_i\,(W)\,| > \tau_n), \tag{A-17}$$

and the uniform rates for $\|\, \widehat{f} - f \,\|^2_{\mathcal{W},\infty}$, we obtain

$$\begin{aligned} \sup_{\alpha \in \mathcal{A}} |R_{1n}(\alpha)| &= O_{P^*}\left(\sqrt{n} p_n + \sqrt{n} \frac{\|\, \widehat{f} - f \,\|^2_{\mathcal{W},\infty}}{\tau_n^2} \right) \\ &= o_{P^*}(1). \end{aligned}$$

The proof that $\sup_{\alpha \in \mathcal{A}} |R_{2n}(\alpha)| = o_{P^*}(1)$ follows the same steps as for R_{1n}, hence, it is omitted. *Q.E.D.*

Proof of Theorem 2.1: We write, with $\Delta t_{ni}(W)$ defined in (A-15),

$$\begin{aligned} \hat{\Delta}_n(\alpha) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(W_i|W)\} t_{ni}(W) \phi(X_i) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}(W_i|W)\} \Delta t_{ni}(W) \phi(X_i) \\ &=: S_n(\alpha) + R_n(\alpha), \end{aligned}$$

By Lemma A.4, $R_n(\alpha) = o_{P^*}(1)$, uniformly in $\alpha \in \mathcal{A}$.

To handle S_n we shall apply Theorem 2.11.9 in van der Vaart and Wellner (1996) to the array $Z_{ni}(\lambda) = n^{-1/2}(Y_i - m(X_i)) \mathbb{I}(f\,(W_i|\,W) \geq 0.5\tau_n) \phi(X_i)$, where $\lambda = (m, W, \phi) \in \Lambda$, and $\Lambda := \mathcal{T}_M^{\eta_m} \times$

$\mathcal{W} \times \Phi$. By Triangle inequality, Assumption 7, definition of $\mathcal{T}_M^{\eta_m}$ and the monotonicity of the indicator function, it follows that

$$\begin{aligned}
&\sum_{i=1}^{n} E[\sup |Z_{ni}(\lambda_2) - Z_{ni}(\lambda_1)|^2] \\
&\leq C\delta^2 + CE[\mathbb{I}(0.5\tau_n - C\delta^2 \leq f\left(W_{1i}|\, W_1\right) \leq 0.5\tau_n + C\delta^2)] \\
&\leq C\delta^2,
\end{aligned}$$

where the sup is taken over $\lambda_2 = (m_2, W_2, \phi_2) \in \Lambda$ such that $\|m_2 - m_1\|_\infty < \delta$, $\|W_2 - W_1\|_\infty < \delta^2$ and $\|\phi_2 - \phi_1\|_2 < \delta$, for a fixed $\lambda_1 = (m_1, W_1, \phi_1) \in \Lambda$. Then, using the notation of Theorem 2.11.9 in van der Vaart and Wellner (1996), for any $\varepsilon > 0$,

$$N_{[\cdot]}(\varepsilon, \Lambda, L_2^n) \leq N\left(\frac{\varepsilon}{2C}, \mathcal{T}_M^{\eta_m}, \|\cdot\|_\infty\right) \times N\left(\left[\frac{\varepsilon}{2C}\right]^2, \mathcal{W}, \|\cdot\|_\infty\right) \times N\left(\frac{\varepsilon}{2C}, \Phi, \|\cdot\|_2\right).$$

Hence, by Lemma B.2 in Ichimura and Lee (2010) and Assumption 8, Λ satisfies $\int_0^1 \sqrt{N_{[\cdot]}(\varepsilon, \Lambda, L_2^n)} < \infty$. On the other hand, for any $\delta > 0$, by Chebyshev's inequality

$$\begin{aligned}
\sum_{i=1}^{n} E[\|Z_{ni}\|_\Lambda \mathbb{I}(\|Z_{ni}\|_\Lambda > \delta)] &\leq Cn^{1/2} E[|(Y - m(X))|\, \mathbb{I}(|(Y - m(X))| > Cn^{1/2}\delta)] \\
&\leq \frac{CE[|(Y - m(X))|^2]}{n^{1/2}\delta^2} \to 0.
\end{aligned}$$

Hence, the conditions of Theorem 2.11.9 in van der Vaart and Wellner (1996, p. 211) are satisfied and $\sum_{i=1}^n Z_{ni}(\lambda) - E[Z_{ni}(\lambda)]$ is asymptotic stochastic equicontinuous with respect to the pseudo-metric $\rho(\lambda_1, \lambda_2) := \max\{\|m_2 - m_1\|_\infty, \|W_2 - W_1\|_\infty, \|\phi_2 - \phi_1\|_2\}$. The stochastic equicontinuity, Assumption 6 and our results in Appendix B imply that, uniformly in $\alpha \in \mathcal{A}$,

$$\begin{aligned}
S_n(\alpha) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - m(W_i|W)\} t_{ni}(W)\phi(X_i) - \sqrt{n} E[\{\widehat{m}(W_i|W) - m(W_i|W)\} t_{ni}(W)\phi(X_i)] + o_{P^*}(1), \\
&=: \Delta_{0n}(\alpha) - \Delta_{1n}(\alpha) + o_{P^*}(1).
\end{aligned}$$

We shall prove that

$$\sup_{\alpha \in \mathcal{A}} \left| \Delta_{0n}(\alpha) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - m(W_i|W)\}\phi(X_i) \right| = o_{P^*}(1) \tag{A-18}$$

and

$$\sup_{\alpha \in \mathcal{A}} \left| \Delta_{1n}(\alpha) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - m(W_i|W)\} E[\phi(X_i)|W_i] \right| = o_{P^*}(1). \tag{A-19}$$

The equality in (A-18) follows from the same arguments in Lemma A.4. We prove now (A-19). To simplify notation denote $\iota_\alpha(w) := E[\phi(X_i)|W_i = w]$, $\alpha \in \mathcal{A}$, and note that $\Delta_{1n}(\alpha) = E[\iota_\alpha(W_i)\, t_{ni}(W)(\widehat{m}(W_i|W) - m(W_i|W))]$. We write

$$\widehat{m}(w|W) - m(w|W) = a_n\left(w|\, W\right) + r_n(w|\, W),$$

where

$$a_n(w|W) := f^{-1}(w|W)\left(\widehat{T}(w|W) - T(w|W) - m(w|W)\left(\widehat{f}(w|W) - f(w|W)\right)\right),$$

$T(w|W) := m(w|W)f(w|W)$ and

$$r_n(w|W) := -\frac{\widehat{f}(w|W) - f(w|W)}{\widehat{f}(w|W)f(w|W)}a_n(w|W).$$

From the results in Appendix B we obtain that $\sup|r_n(w|W)| = o_{P^*}(n^{-1/2})$ under our assumptions on the bandwidth. It then follows that $\Delta_{1n}(\alpha)$ is uniformly bounded by

$$\int \iota_\alpha(w)\,t_{ni}(w)[\widehat{T}(w|W) - T(w|W)]dw \tag{A-20}$$

$$-\int \iota_\alpha(w)\,t_{ni}(w)m(w|W)[\widehat{f}(w|W) - f(w|W)]dw \tag{A-21}$$

$$+\,o_{P^*}(n^{-1/2}).$$

We now look at terms (A-20)-(A-21). Firstly, it follows from our results in Appendix B that the difference between $T(w|W)$ and $E(\widehat{T}(w|W))$ is $o_{P^*}(n^{-1/2})$. Secondly, under Assumption 7(ii) we can replace $t_{ni}(w)$ by one. Hence, uniformly in $\alpha \in \mathcal{A}$,

$$\begin{aligned}\int \iota_\alpha(w)[\widehat{T}(w|W) - T(w|W)]dw &= \int \iota_\alpha(w)[\widehat{T}(w|W) - E(\widehat{T}(w|W))]dw + o_{P^*}(n^{-1/2})\\ &= \frac{1}{n}\sum_{j=1}^{n} Y_j \int \iota_\alpha(w)K_h(W_j - w)dw - \int \iota_\alpha(w)E(Y_jK_h(W_j - w))dw + o_{P^*}(n^{-1/2}),\\ &= \frac{1}{n}\sum_{j=1}^{n} \iota_\alpha(W_j)Y_j - E[\iota_\alpha(W_j)m(W_j|W)] + o_{P^*}(n^{-1/2}),\end{aligned}$$

where the last equality follows from the change of variables $u = h^{-1}(W_j - w)$, Assumptions 3, 5 and the fact that, uniformly in $\alpha \in \mathcal{A}$, $\int \iota_\alpha(w)K_h(W_j - w)dw = \int \phi(x)\left(\int f_X(x|w,W)K_h(W_j - w)dw\right)dx = \int \phi(x)f_X(x|W_j,W)dx + O(b_n^r)$. Likewise, the term (A-21) becomes $\int \iota_\alpha(w)m(w|W)[\widehat{f}(w|W) - f(w|W)]dw = n^{-1/2}\sum_{j=1}^{n}\iota_\alpha(W_j)m(W_j|W) - E[\iota_\alpha(W_j)m(W_j|W)] + o_{P^*}(n^{-1/2})$. In conclusion, we have uniformly in $\alpha \in \mathcal{A}$, that $\Delta_{1n}(\alpha) = n^{-1/2}\sum_{j=1}^{n}\iota_\alpha(W_j)[Y_j - m(W_j|W)] + o_{P^*}(n^{-1/2})$. This proves (A-19) and hence the result of the Theorem. *Q.E.D.*

Proof of Theorem 2.2: We write, using $\widehat{m}_i := \widehat{m}(\widehat{W}_i|\widehat{W})$,

$$\begin{aligned}\hat{\Delta}_n(\widehat{\alpha}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Y_i - \widehat{m}_i\}t_{ni}\widehat{\phi}(X_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Y_i - \widehat{m}_i\}\Delta t_{ni}\widehat{\phi}(X_i)\\ &=: \widehat{S}_n(\widehat{\alpha}) + \widehat{R}_n(\widehat{\alpha}).\end{aligned}$$

We write $\widehat{R}_n(\widehat{\alpha}) = n^{-1/2}\sum_{i=1}^{n}\varepsilon_i\Delta t_{ni}\widehat{\phi}(X_i) - n^{-1/2}\sum_{i=1}^{n}\{\widehat{m}_i - m_{0i}\}\Delta t_{ni}\widehat{\phi}(X_i) =: \widehat{R}_{1n} - \widehat{R}_{2n}$. Note that, by the arguments of the proof of Lemma A.4, the first term in the last equation satisfies $\widehat{R}_{1n}(\widehat{\alpha}) = o_P(1)$, while the second term can be further decomposed as

$$\begin{aligned}\widehat{R}_{2n} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\widehat{m}_i - m(\widehat{W}_i|\widehat{W})\}\Delta t_{ni}\widehat{\phi}(X_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(\widehat{W}_i|\widehat{W}) - m_{0i}\}\Delta t_{ni}\widehat{\phi}(X_i) \\ &:= \widehat{R}_{2n;a} + \widehat{R}_{2n;b}.\end{aligned}$$

We can further write, using (A-16),

$$\begin{aligned}E|\widehat{R}_{2n;a}| \leq &\parallel (\widehat{m} - m)\mathbb{I}(f \geq \tau_n/2) \parallel_{\mathcal{W},\infty} \sqrt{n}P(\widehat{f}(\widehat{W}_i|\widehat{W}) < \tau_n) \\ &+ \parallel (\widehat{m} - m)\mathbb{I}(\widehat{f} \geq \tau_n) \parallel_{\mathcal{W},\infty} \frac{4\sqrt{n} \parallel \widehat{f} - f \parallel_{\infty}^{2}}{\tau_n^2}.\end{aligned} \tag{A-22}$$

The results in Appendix B and (A-17) yield

$$\sqrt{n}P(\widehat{f}(\widehat{W}_i|\widehat{W}) < \tau_n) = O_P(n^{1/2}(p_n + \tau_n^{-1}d_n)),$$

$\parallel (\widehat{m} - m)\mathbb{I}(f \geq \tau_n/2) \parallel_{\mathcal{W},\infty} = O_P(\tau_n^{-1}d_n)$ and $\parallel (\widehat{m} - m)\mathbb{I}(\widehat{f} \geq \tau_n) \parallel_{\mathcal{W},\infty} = O_P(\tau_n^{-1}d_n)$. Thus, from (A-22) and the previous rates

$$\begin{aligned}\widehat{R}_{2n;a} &= O_P(\tau_n^{-1}d_n)O_P(n^{1/2}\left(p_n + \tau_n^{-1}d_n + \tau_n^{-2}d_n^2\right)) \\ &= o_P(1).\end{aligned}$$

Similarly,

$$\begin{aligned}\widehat{R}_{2n;b} &= O_P(w_n)O_P(n^{1/2}\left(p_n + \tau_n^{-1}d_n + \tau_n^{-2}d_n^2\right)) \\ &= o_P(1).\end{aligned}$$

Recall $\widehat{\phi}_W^{\perp}(X) := \widehat{\phi}(X) - E[\widehat{\phi}(X)|\widehat{W}]$. Then, using $\widehat{R}_n(\widehat{\alpha}) = o_P(1)$ and Theorem 2.1, we can write

$$\begin{aligned}\Delta_n(\widehat{\alpha}) = &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\widehat{\phi}_W^{\perp}(X_i) + \sum_{j=1}^{\kappa}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g_{0i} - \widehat{g}_i)^j\frac{1}{j!}\partial_{\bar{g}}^{(j)}m(W_{0i})\widehat{\phi}_W^{\perp}(X_i) \\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{\kappa}\{m(W_{0i}|W_0) - m(\widehat{W}_i|W_0) - (g_{0i} - \widehat{g}_i)^j\frac{1}{j!}\partial_{\bar{g}}^{(j)}m(W_{0i})\}\widehat{\phi}_W^{\perp}(X_i) \\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(\widehat{W}_i|W_0) - m(\widehat{W}_i|\widehat{W})\}\widehat{\phi}_W^{\perp}(X_i) + o_P(1) \\ =: &\widetilde{\Delta}_n(\widehat{\alpha}) + \sum_{j=1}^{\kappa}R_{jn}(\widehat{\alpha}) + \widetilde{R}_{1n}(\widehat{\alpha}) + \widetilde{R}_{2n}(\widehat{\alpha}) + o_P(1).\end{aligned} \tag{A-23}$$

To handle $\widetilde{\Delta}_n(\widehat{\alpha})$, we apply Lemma A.3 with $\mathcal{S} = \{s(x) = \phi(x) - E[\phi(X)|W = W(x)] : \alpha \in \mathcal{A}\}$. Let $\{W_k : k = 1, ..., N_{1\varepsilon}\}$ be an ε^2-net covering of $\mathcal{W}$ with respect to the sup-norm $\|\cdot\|_{\infty}$. Let $\{[\phi_{lj}, \phi_{uj}] : j = 1, ..., N_{2\varepsilon}\}$ be an ε-bracket covering of Φ. Then, it holds that $\{[\phi_{lj} - E[\phi_{uj}(X)|W_k] - C\varepsilon^2, \phi_{uj} - E[\phi_{lj}(X)|W_k] + C\varepsilon^2]\}$ is an ε-bracket covering of $\mathcal{S}$ with respect to $\|\cdot\|_2$. In fact, for each $E[\phi(X)|W]$ we can find j and k such that $\phi_{lj} \leq \phi \leq \phi_{uj}$ and $\|W - W_k\|_{\infty} < \varepsilon^2$, and by Lemma A.2,

$$E[\phi_{lj}(X)|W_k] - C\varepsilon^2 \leq E[\phi(X)|W] \leq E[\phi_{uj}(X)|W_k] + C\varepsilon^2.$$

Hence, $\log N_{[\cdot]}(\varepsilon, \mathcal{S}, \|\cdot\|_2) \leq C\varepsilon^{-v_s}$ for some $v_s < 2$.

In $\mathcal{A}$ we define the pseudo-metric

$$\rho_{\mathcal{A}}(\alpha_1, \alpha_2) := \max\left\{\|W_1 - W_2\|_\infty, \|\phi_1 - \phi_2\|_2\right\}, \; \alpha_j := (W_j, \phi_j), \; j = 1, 2.$$

A consequence of the stochastic equicontinuity of $\widetilde{\Delta}_n(\alpha)$ and that $\rho_{\mathcal{A}}(\widehat{\alpha}, \alpha_0) = o_P(1)$ is that

$$\widetilde{\Delta}_n(\widehat{\alpha}) = \widetilde{\Delta}_n(\alpha_0) + o_P(1). \tag{A-24}$$

On the other hand, by our Assumption 10(i) and an application of Lemma A.3 with $\mathcal{S} = \{s(x) = E[\partial_{\overline{g}} m(W_{0i})\phi_W^\perp(X)|X_1 = x_1] : \alpha \in \mathcal{A}\}$, we obtain

$$\begin{aligned}\sum_{j=1}^{\kappa} R_{jn}(\widehat{\alpha}) &= -G_n(\widehat{\alpha}) + o_P(1) \\ &= -G_n(\alpha_0) + o_P(1).\end{aligned} \tag{A-25}$$

The proof concludes by showing that both $\widetilde{R}_{1n}(\widehat{\alpha})$ and $\widetilde{R}_{2n}(\widehat{\alpha})$ are $o_P(1)$. It follows from a Taylor expansion that $|\widetilde{R}_{1n}(\widehat{\alpha})| = O_P(\sqrt{n}\, \|\widehat{g}_i - g_0\|_\infty^\kappa) = o_P(1)$. To show that $\widetilde{R}_{2n}(\widehat{\alpha}) = o_P(1)$, we first show that the empirical process $Q_n(\alpha) := n^{-1/2}\sum_{i=1}^n \phi_W^\perp(X_i)m(W_i|W)$ is stochastically $\rho_{\mathcal{A}}$-equicontinuous. To that end, consider the class of functions $\mathcal{H} = \{z \to s(z, \alpha) = \{\phi(x) - E[\phi(x)|W]\}\, m(W|W) : \phi \in \Phi, W \in \mathcal{W}\}$. By the Triangle inequality and Lemma A.2, it follows that $E[\sup|s(Z, \alpha_2) - s(Z, \alpha_1)|^2] \leq C\delta^2$, where the sup is taken over $\alpha_2 \in \mathcal{A}$ such that $\rho_{\mathcal{A}}(\alpha_2, \alpha_1) < \delta$, for a fixed $\alpha_1 \in \mathcal{A}$. Then, the stochastic equicontinuity follows from Lemma A.1. Then, by continuity and $\rho_{\mathcal{A}}(\widehat{\alpha}, \alpha_0) = o_{P^*}(1)$, it follows that $\widetilde{R}_{2n}(\widehat{\alpha}) = n^{-1/2}\sum_{i=1}^n \{m(W_{0i}|W_0) - m(W_{0i}|W_0)\}\phi_0^\perp(X_i) + o_{P^*}(1) = o_{P^*}(1)$. These results along with the equality in (A-23) and the expansions (A-24) – (A-25) yield the desired result. *Q.E.D.*

Proof of Corollary 3.1: Our estimator satisfies the first order condition

$$0 = \frac{1}{n}\sum_{i=1}^n [Y_i\widehat{g}_i - D_i\widehat{m}_{i\widehat{\theta}^*}] \frac{\partial_\theta \widehat{m}_{i\widehat{\theta}^*}}{\widehat{m}_{i\widehat{\theta}^*}(\widehat{g}_i - \widehat{m}_{i\widehat{\theta}^*})}\widehat{\psi}_i^* \widetilde{t}_{in}.$$

Simple algebra and a standard Taylor series expansion around θ_0 yield

$$Y_i\widehat{g}_i - D_i\widehat{m}_{i\widehat{\theta}^*} = [Y_i - \widehat{m}_{i\theta_0}]\widehat{g}_i - [D_i - \widehat{g}_i]\, \widehat{m}_{i\theta_0} - D_i\partial_\theta^\top \widehat{m}_{i\bar{\theta}}(\widehat{\theta}^* - \theta_0),$$

where $\bar{\theta}$ is such that $|\bar{\theta} - \theta_0| \leq |\widehat{\theta} - \theta_0|$ a.s. It then follows that, for a sufficiently large n,

$$\begin{aligned}\sqrt{n}(\widehat{\theta}^* - \theta_0) &= \Gamma_n^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n [Y_i - \widehat{m}_{i\theta_0}] \frac{\widehat{g}_i\partial_\theta \widehat{m}_{i\widehat{\theta}^*}}{\widehat{m}_{i\widehat{\theta}^*}(\widehat{g}_i - \widehat{m}_{i\widehat{\theta}^*})}\widehat{\psi}_i^*\widetilde{t}_{in} - \frac{1}{\sqrt{n}}\sum_{i=1}^n [D_i - \widehat{g}_i]\, \frac{\partial_\theta \widehat{m}_{i\widehat{\theta}^*}}{(\widehat{g}_i - \widehat{m}_{i\widehat{\theta}^*})}\widehat{\psi}_i^*\widetilde{t}_{in}\right) \\ &\equiv \Gamma_n^{-1}(A_{1n} - A_{2n})\end{aligned} \tag{A-26}$$

where

$$\Gamma_n := \frac{1}{n}\sum_{i=1}^n \frac{D_i\partial_\theta \widehat{m}_{i\widehat{\theta}^*}\partial_\theta^\top \widehat{m}_{i\bar{\theta}}}{\widehat{m}_{i\widehat{\theta}^*}(\widehat{g}_i - \widehat{m}_{i\widehat{\theta}^*})}\widehat{\psi}_i^*\widetilde{t}_{in}.$$

We obtain now some rates of convergence for kernel estimates that will be useful in subsequent arguments. Note that

$$\widehat{f}\left(w|\,\theta,g\right)\partial_\theta\widehat{m}(w|\theta,g)=\partial_\theta\widehat{T}\left(w|\,\theta,g\right)-\widehat{m}\left(w|\,\theta,g\right)\partial_\theta\widehat{f}\left(w|\,\theta,g\right).$$

To show the convergence of the right hand side (r.h.s), using a simplified notation, we write

$$\begin{aligned}\sup|\partial_\theta\widehat{f}(w|\theta,g)-\partial_\theta f(w|\theta,g)| &\le \sup|\partial_\theta\widehat{f}(w|\theta,g)-E\partial_\theta\widehat{f}(w|\theta,g)|\\ &\quad+\sup|E\partial_\theta\widehat{f}(w|\theta,g)-\partial_\theta f(w|\theta,g)|\\ &\equiv I_{1n}+I_{2n},\end{aligned}$$

where the sup is over the set $a_n\le\widehat{h}_n\le b_n$, $\theta\in\Theta_0$, $w\in\mathcal{X}_{\mathcal{W}}$ and $g\in\mathcal{G}$. From Lemma B.8, it follows that

$$I_{1n}=O_{P^*}\left(\sqrt{\frac{\log a_n^{-2}\vee\log\log n}{na_n^4}}\right).$$

By the classical change of variables and integration by parts, for any $a_n\le h\le b_n$,

$$\begin{aligned}E\left[\partial_\theta\widehat{f}(w|\theta,g)-\partial_\theta f(w|\theta,g)\right] &= \frac{1}{h^3}E\left[X\partial_{w_1}K\left(\frac{w-W\left(\theta,g\right)}{h}\right)-\partial_\theta f(w|\theta,g)\right]\\ &= \int\partial_{w_1}m(w-uh|\theta,g)K\left(u\right)du-\partial_\theta f(w|\theta,g),\end{aligned}$$

where $m\left(w|\,\theta,g\right)=r\left(w|\,\theta,g\right)f\left(w|\,\theta,g\right)$ and $r\left(w|\,\theta,g\right):=E[X|W\left(\theta,g\right)=w]$. By a Taylor series expansion,

$$I_{2n}=O\left(b_n^{r-1}\frac{1}{r!}\left\|\partial_w^{r-1}\partial_{w_1}m\right\|_{\Theta\times\mathcal{G},\infty}\right)=O\left(b_n^{r-1}\right).$$

The proof for $\widehat{T}$ follows the same arguments as for $\widehat{f}$, and hence is omitted. Therefore by simple but somewhat tedious algebra one can show that

$$\begin{aligned}\left\|\widehat{f}\partial_\theta\widehat{m}_{\widehat{\theta}^*}(\cdot|\,\widehat{\theta}^*,\widehat{g})-f\partial_\theta m_{\theta_0}\left(\cdot|\,W_0\right)\right\|_\infty &= O_{P^*}\left(d_n'\right)+\left\|f\partial_\theta m_{\widehat{\theta}^*}(\cdot|\,\widehat{\theta}^*,\widehat{g})-f\partial_\theta m_{\theta_0}\left(\cdot|\,W_0\right)\right\|_\infty\\ &= o_{P^*}(1),\end{aligned}$$

where the second equality uses that $f\partial_\theta m_{\widehat{\theta}^*}(W)$ is Liptschitz in W, and where

$$d_n'=\sqrt{\frac{\log a_n^{-2}\vee\log\log n}{na_n^4}}+b_n^{r-1}.$$

To simplify the notation, we define

$$\widehat{\varphi}_i:=\frac{\partial_\theta\widehat{m}_{i\widehat{\theta}^*}\partial_\theta^\top\widehat{m}_{i\bar{\theta}}}{\widehat{\sigma}_i^2}\times\frac{\widetilde{m}_i(\widehat{g}_i-\widetilde{m}_i)}{\widehat{m}_{i\widehat{\theta}^*}(\widehat{g}_i-\widehat{m}_{i\widehat{\theta}^*})}=:\widehat{\varphi}_{1i}\times\widehat{\varphi}_{2i},$$

and note that $\Gamma_n=n^{-1}\sum_{i=1}^n D_i\widehat{g}_i^{-1}\widehat{\varphi}_i\widetilde{t}_{in}$ can be written as

$$\begin{aligned}\Gamma_n &= n^{-1}\sum_{i=1}^n\widehat{\varphi}_i\widetilde{t}_{in}+n^{-1}\sum_{i=1}^n(D_i-\widehat{g}_i)\widehat{g}_i^{-1}\widehat{\varphi}_i\widetilde{t}_{in}\\ &= n^{-1}\sum_{i=1}^n\widehat{\varphi}_i\widetilde{t}_{in}+o_P(1).\end{aligned}$$

By the continuous mapping theorem it can be shown that, uniformly in $1 \leq i \leq n$ and $a_n \leq \widehat{h}_n \leq b_n$,

$$\begin{aligned}\widehat{\varphi}_{1i} \equiv \frac{\partial_\theta \widehat{m}_{i\widehat{\theta}^*} \partial_\theta^\top \widehat{m}_{i\bar{\theta}}}{\widehat{\sigma}_i^2} &= \frac{\partial_\theta m(W_{0i}) \partial_\theta^\top m(W_{0i})}{\sigma_{0i}^2} + o_P(1) \\ &=: \varphi_{1i} + o_P(1)\end{aligned}$$

and

$$\begin{aligned}\widehat{\varphi}_{2i} &= \varphi_{2i}(\widehat{\theta}^*) + O_P\left(\tau_n^{-1}\left\{d_n + n^{-1/2} + w_n\right\}\right) \qquad \text{(A-27)} \\ &= 1 + o_P(1),\end{aligned}$$

where $\varphi_{2i}(\theta) := m_{0i}(g_{0i} - m_{0i})/m_{i\theta}(g_{0i} - m_{i\theta})$. Hence, since $n^{-1}\sum_{i=1}^n(\widetilde{t}_{in} - 1) = o_P(1)$, we obtain by the uniform consistency of $\widehat{\varphi}_i$ and the law of large numbers,

$$\begin{aligned}\Gamma_n &= n^{-1}\sum_{i=1}^n \varphi_i + n^{-1}\sum_{i=1}^n(\widehat{\varphi}_i - \varphi_i)(\widetilde{t}_{in} - 1) + n^{-1}\sum_{i=1}^n \varphi_i(\widetilde{t}_{in} - 1) + n^{-1}\sum_{i=1}^n(\widehat{\varphi}_i - \varphi_i) + o_P(1) \\ &= \Gamma + o_P(1).\end{aligned}$$

We now show that A_{1n} in (A-26) has the expansion

$$\begin{aligned}A_{1n} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n [Y_i - \widehat{m}_{i\theta_0}]\phi_{1i}(\widehat{\theta}^*)\varphi_{2i}(\widehat{\theta}^*)\widehat{t}_{in} + o_P(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n v_i\phi_{1i} + o_P(1), \qquad \text{(A-28)}\end{aligned}$$

where $\widehat{\phi}_{1i} \equiv \widehat{\phi}_{1i}(\widehat{\theta}^*) := \partial_\theta \widehat{m}_{i\widehat{\theta}^*}/\widehat{\sigma}_i^2$ is a uniformly consistent estimate of $\phi_{1i} \equiv \phi_{1i}(\theta_0)$, with $\phi_{1i}(\theta) := \partial_\theta m_{i\theta}/\sigma_i^2$, $\theta \in \Theta_0$. To prove the first equality of the last display it suffices to prove that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n [Y_i - \widehat{m}_{i\theta_0}]\widehat{\phi}_{1i}(\widehat{\varphi}_{2i}\widetilde{t}_{in} - \varphi_{2i}(\widehat{\theta}^*)\widehat{t}_{in}) = o_P(1) \qquad \text{(A-29)}$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n [Y_i - \widehat{m}_{i\theta_0}](\widehat{\phi}_{1i} - \phi_{1i}(\widehat{\theta}^*))\varphi_{2i}(\widehat{\theta}^*)\widehat{t}_{in} = o_P(1). \qquad \text{(A-30)}$$

To that end, write the l.h.s of (A-29) as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i\widehat{\phi}_{1i}(\widehat{\varphi}_{2i}\widetilde{t}_{in} - \varphi_{2i}(\widehat{\theta}^*)\widehat{t}_{in}) + \frac{1}{\sqrt{n}}\sum_{i=1}^n [m_{i\theta_0} - \widehat{m}_{i\theta_0}]\widehat{\phi}_{1i}(\widehat{\varphi}_{2i}\widetilde{t}_{in} - \varphi_{2i}(\widehat{\theta}^*)\widehat{t}_{in}).$$

We shall focus on the second term in the last display, as the first term is of smaller order. This second term can be written as

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^n [m_{i\theta_0} - \widehat{m}_{i\theta_0}]\widehat{\phi}_{1i}(\widehat{\varphi}_{2i} - \varphi_{2i}(\widehat{\theta}^*))\widetilde{t}_{in} + \frac{1}{\sqrt{n}}\sum_{i=1}^n [m_{i\theta_0} - \widehat{m}_{i\theta_0}]\widehat{\phi}_{1i}\varphi_{2i}(\widehat{\theta}^*)(\widetilde{t}_{in} - \widehat{t}_{in}) \\ &\equiv C_{1n} + C_{2n}.\end{aligned}$$

The uniform rates in (A-27) and our results on kernel estimates imply that

$$
\begin{aligned}
C_{1n} &= O_P(n^{1/2} q_n^2 \tau_n^{-1}) \\
&= o_P(1).
\end{aligned}
$$

Similarly, since for sufficiently large n,

$$
\left|\widetilde{t}_{in} - \widehat{t}_{in}\right| \leq \mathbb{I}(|\widetilde{m}_i - m_i| > \tau_n) + \mathbb{I}(|(\widehat{g}_i - \widetilde{m}_i) - (g_{0i} - m_i)| > \tau_n),
$$

it can be shown that

$$
\begin{aligned}
C_{2n} &= O_P(n^{1/2} q_n^2 \tau_n^{-1}) \\
&= o_P(1).
\end{aligned}
$$

On the other hand, we write $\widehat{\phi}_{1i} - \phi_{1i}$ as

$$
\widehat{\phi}_{1i} - \phi_{1i}(\widehat{\theta}^*) = \frac{1}{b_{i\widehat{\theta}^*} + \widehat{b}_i - b_{i\widehat{\theta}^*}} \left\{ \widehat{a}_i - a_{i\widehat{\theta}^*} - \phi_{1i}(\widehat{b}_i - b_{i\widehat{\theta}^*}) \right\},
$$

where, with a simplified notation,

$$
\begin{aligned}
\widehat{a}_i &:= \widehat{f}_i^2 \partial_\theta \widehat{m}_{i\widehat{\theta}^*} = \widehat{f}_i \partial_\theta \widehat{T}_i - \widehat{T}_i \partial_\theta \widehat{f}_i, \ a_{i\widehat{\theta}^*} := f_{i\widehat{\theta}^*}^2 \partial_\theta m_{i\widehat{\theta}^*}, \\
\widehat{b}_i &:= \widehat{f}_i^2 \widehat{\sigma}_i^2 = \widehat{T}_i(\widehat{f}_i - \widehat{T}_i) + (\widehat{c}_i)^2 \, \widehat{g}_i(1 - \widehat{g}_i) - 2\widehat{c}_i \widehat{T}_i (1 - \widehat{g}_i),
\end{aligned}
$$

$\widehat{c}_i := \partial_{\bar{g}} \widehat{T}_i - \widehat{m}_i \partial_{\bar{g}} \widehat{f}_i$ and $b_{i\widehat{\theta}^*} := f_{i\widehat{\theta}^*}^2 \sigma_i^2$. Then, from these expressions and the previous rates for derivatives, we obtain that uniformly in $1 \leq i \leq n$ and $a_n \leq \widehat{h}_n \leq b_n$,

$$
\widehat{\phi}_{1i} - \phi_{1i}(\widehat{\theta}^*) = O_P(q_n').
$$

The last display shows (A-30) in a routine fashion.

Finally, to prove the second equality in (A-28) we apply Theorem 2.2 with the class

$$
\Phi = \left\{ x \to \varphi(x, \theta) := \frac{\partial_\theta m_\theta}{m_\theta (g_0 - m_\theta)} \frac{m(g_0 - m)}{\sigma^2(x)} : \theta \in \Theta_0 \right\},
$$

where $\partial_\theta m_\theta \equiv \partial m(W_i(\theta, g_0)|W(\theta, g_0))/\partial\theta|_{\theta=\theta}$ and $m_\theta \equiv m(W(\theta, g_0)|W(\theta, g_0))$. By our assumptions $\varphi(X, \theta)$ is bounded and satisfies

$$
|\varphi(x, \theta_1) - \varphi(x, \theta_2)| \leq C_L(x) \left|\theta_1 - \theta_2\right|,
$$

for all $\theta_1, \theta_2 \in \Theta_0$ and $C_L(\cdot)$ such that $E[C_L^2(X)] < \infty$. Hence, Assumption 8 is satisfied. Then, using $E[\partial_\theta m(W_{i0})|W_0] = 0$ a.s., which can be shown as in Ichimura (1993, Lemma 5.6, p. 95), we conclude (A-28).

Similar arguments to those used above for A_{1n} show that

$$
\begin{aligned}
A_{2n} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n [D_i - \widehat{g}_i] \frac{\widetilde{m}_i(\widehat{g}_i - \widetilde{m}_i) \partial_\theta \widehat{m}_{i\widehat{\theta}^*}}{\widehat{\sigma}_i^2 \widehat{g}_i (\widehat{g}_i - \widehat{m}_{i\widehat{\theta}^*})} \widetilde{t}_{in} = \frac{1}{\sqrt{n}} \sum_{i=1}^n [D_i - \widehat{g}_i] \, \widehat{\phi}_{2i} \widehat{\phi}_{1i} \widetilde{t}_{in} + o_P(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n [D_i - \widehat{g}_i] \, \phi_{2i}(\widehat{\theta}^*) \phi_{1i}(\widehat{\theta}^*) \widehat{t}_{in} + o_P(1) \\
&= o_P(1),
\end{aligned}
$$

where $\widehat{\phi}_{2i} := \widehat{m}_{i\widehat{\theta}^*}/\widehat{g}_i$ estimates consistently $\phi_{2i} \equiv \phi_{2i}(\theta_0)$, with $\phi_{2i}(\theta) := m_{i\theta}/g_{0i}$. The last equality follows from an application of Theorem 2.2 with $\mathcal{W} = \{x \to W(x) = x\}$ and $\Phi = \{X_i \to \phi_{2i}(\theta)\phi_{2i}(\theta) : \theta \in \Theta_0\}$. Thus, we conclude that

$$\sqrt{n}(\widehat{\theta}^* - \theta_0) = \Gamma_n^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i \phi_{1i} + o_P(1),$$

and the result of the corollary follows from the Lindeberg-Lévy Central Limit Theorem and Slutsky's Lemma. The proof that $\widehat{\Gamma}_* = \Gamma_* + o_P(1)$ follows the same arguments as that of $\Gamma_n^{-1} = \Gamma^{-1} + o_P(1)$ and is therefore omitted. *Q.E.D.*

Proof of Theorem 3.2: Using our uniform rates for kernel estimators one can show that $\widehat{\phi}_*(X_i, \widehat{g}_i)$ converges uniformly to $\phi_*(X, g_0(X))$ with a rate $O_P(\tau_n^{-1}(q_n + n^{-1/2}))$, so that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Y_i - \widehat{m}(\widehat{W}_i|\widehat{W}_i)] \left(\widehat{\phi}_*(X_i, \widehat{g}_i) - \phi_*(X_i, g_{0i})\right) \widehat{t}_{ni} = O_P(n^{1/2}\tau_n^{-1} q_n^2).$$

Hence, by Theorem 2.2 in the main text with $\Phi = \{x \to \phi_*(x, g_{0i})\}$, uniformly in $a_n \leq \widehat{h}_n \leq b_n$,

$$T_{n,h} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Y_i - \widehat{m}(\widehat{W}_i|\widehat{W}_i)] \phi_*(X_i, g_{0i}) \widehat{t}_{ni} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v_i \phi_*^{\perp}(X_i, g_{0i}) + o_P(1).$$

On the other hand, it is straightforward to prove that $\widehat{\sigma}^2 = \sigma^2 + o_P(1)$. The limiting null distribution then follows from the Lindeberg-Lévy Central Limit Theorem and Slutsky's Lemma. Under the alternative,

$$\frac{1}{\sqrt{n}} T_{n,h} := \frac{1}{n} \sum_{i=1}^{n} [Y_i - \widehat{m}(\widehat{W}_i|\widehat{W}_i)] \widehat{\phi}_*(X_i, \widehat{g}_i) \widehat{t}_{ni},$$

converges to $E\left(\varepsilon_i \phi_*(X_i, g_{0i})\right) \neq 0$, and hence the consistency follows. *Q.E.D.*

Appendix B Uniform Consistency Results for Kernel Estimators

This section establishes rates for uniform consistency of kernel estimators used in the paper. These auxiliary results complement related ones in Andrews (1995) and Sperlich (2009), among others, but we impose different conditions on the kernel functions and provide alternative methods of proof. Unlike Mammen, Rothe and Schienle (2011a), we consider uniform in bandwidth consistency and rates, though we do not provide uniform limiting distributions. Einmahl and Mason (2005) also study uniform in bandwidth consistency of kernel estimators, but they did not consider the extension to kernel estimators of (possibly nonparametrically) generated observations, as we do. The results of this section, which should be potentially useful in other settings, are more general than required for the proofs of the main the results in the text.

We first state some well-known results from the empirical process literature. Define the generic class of measurable functions $\mathcal{G} := \{z \to m(z, \theta, h) : \theta \in \Theta, h \in \mathcal{H}\}$, where Θ and $\mathcal{H}$ are endowed with the pseudo-norms $|\cdot|_\Theta$ and $|\cdot|_\mathcal{H}$, respectively. The following result is Theorem 2.14.2 in van der Vaart and Wellner (1996, p. 240).

Lemma B.1 *Let $\mathcal{G}$ be a class of measurable functions with a measurable envelope G. Then, there exists a constant C such that*

$$E^*[\|\mathbb{G}_n\|_{\mathcal{G}}] \leq C \|G\|_2 \int_0^1 \sqrt{1 + \log N_{[\cdot]}(\varepsilon, \mathcal{G}, \|\cdot\|_2)} d\varepsilon.$$

The following result is the celebrated Talagrand's inequality (see Talagrand, 1994). Rademacher variables are iid variables $\{\varepsilon_i\}_{i=1}^n$ such that $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$.

Lemma B.2 *Let $\mathcal{G}$ be a class of measurable functions satisfying $\|g\|_\infty \leq M < \infty$ for all $g \in \mathcal{G}$. Then it holds for all $t > 0$ and some universal positive constants A_1 and A_2 that*

$$P^*\left(\max_{1\leq m\leq n} \|\mathbb{G}_m\|_{\mathcal{G}} \geq A_1\left(E\left\|\sum_{i=1}^n \varepsilon_i g(Z_i)\right\|_{\mathcal{G}} + t\right)\right) \leq 2\left\{\exp\left(-\frac{A_2 t^2}{n\sigma_{\mathcal{G}}^2}\right) + \exp\left(-\frac{A_2 t}{M}\right)\right\},$$

where $\{\varepsilon_i\}_{i=1}^n$ is a sequence of iid Rademacher variables, independent of the sample $\{Z_i\}_{i=1}^n$ and $\sigma_{\mathcal{G}}^2 := \sup_{g\in\mathcal{G}} var(g(Z))$.

We now proceed with the main results of this section. Let Υ be a class of measurable real-valued functions of Z and let $\mathcal{W}$ be a class of measurable functions of X with values in $\mathbb{R}^d$. Define $\mathcal{X}_{\mathcal{W}} := \{W(x) \in \mathbb{R}^d : W \in \mathcal{W} \text{ and } x \in \mathcal{X}_X\}$. We denote by $\psi := (\varphi, w, W)$ a generic element of the set $\Psi := \Upsilon \times \mathcal{X}_{\mathcal{W}} \times \mathcal{W}$. Let $\Psi_I := \Upsilon \times I \times \mathcal{W}$, for a compact set $I \subset \mathcal{X}_{\mathcal{W}}$. Let $f(w|W)$ denote the density of $W(X)$ evaluated at w. Define the regression function $c(\psi) := E[\varphi(Z)|W(X) = w]$. Henceforth, we use the convention that a function evaluated outside its support is zero. Then, an estimator for $m(\psi) := c(\psi)f(w|W)$ is given by

$$\widehat{m}_h(\psi) = \frac{1}{nh^d}\sum_{i=1}^n \varphi(Z_i) K\left(\frac{w - W(X_i)}{h}\right),$$

where $K(w) = \prod_{l=1}^d k(w_l)$, $k(\cdot)$ is a kernel function, $h := h_n > 0$ is a bandwidth and $w = (w_1, \ldots, w_d)^\top$. We consider the following regularity conditions on the data generating process, kernel, bandwidth and classes of functions.

Assumption B.1 *The sample observations $\{Z_i := (Y_i^\top, X_i^\top)^\top\}_{i=1}^n$ are a sequence of independent and identically distributed (iid) variables, distributed as $Z \equiv (Y^\top, X^\top)^\top$.*

Assumption B.2 *The class $\mathcal{W}$ is such that $\log N(\varepsilon, \mathcal{W}, \|\cdot\|_\infty) \leq C\varepsilon^{-v_w}$ for some $v_w < 1$.*

Assumption B.3 *The density $f(w|W)$ is uniformly bounded, i.e. $\|f\|_{\mathcal{W},\infty} < C$.*

Assumption B.4 *The kernel function $k(t) : \mathbb{R} \to \mathbb{R}$ is bounded, r-times continuously differentiable and satisfies the following conditions: $\int k(t) dt = 1$, $\int t^l k(t) dt = 0$ for $0 < l < r$, and $\int |t^r k(t)| dt < \infty$, for some $r \geq 2$; $|\partial k(t)/\partial t| \leq C$ and for some $v > 1$, $|\partial k(t)/\partial t| \leq C|t|^{-v}$ for $|t| > L$, $0 < L < \infty$.*

Assumption B.5 *The possibly data-dependent bandwidth h satisfies $P(a_n \leq h \leq b_n) \to 1$ as $n \to \infty$, for deterministic sequences of positive numbers a_n and b_n such that $b_n \to 0$ and $a_n^d n/\log n \to \infty$.*

Given the class $\mathcal{W}$ and the compact set $I \subset \mathcal{X}_{\mathcal{W}}$, we define the class of functions

$$\mathcal{K}_0 := \left\{ x \to K\left(\frac{w - W(x)}{h}\right) : w \in I, W \in \mathcal{W}, h \in (0,1] \right\}.$$

Our first result establishes the complexity of the class $\mathcal{K}_0$, which is crucial for the subsequent analysis.

Lemma B.3 *Under Assumption B.4, for a positive constant C_1,*

$$N_{[\cdot]}\left(C_1\varepsilon, \mathcal{K}_0, \|\cdot\|_2\right) \le C\varepsilon^{-v} N(\varepsilon^2, \mathcal{W}, \|\cdot\|_\infty), \text{ for some } v \ge 1. \tag{B-31}$$

Proof of Lemma B.3: Let $W_1, ..., W_{N_{1\varepsilon}}$ be the centers of an ε^2-cover of $\mathcal{W}$ with respect to $\|\cdot\|_\infty$, where $N_{1\varepsilon} = N(\varepsilon^2, \mathcal{W}, \|\cdot\|_\infty)$. Fix j, $j = 1, ..., N_{1\varepsilon}$, and consider the marginal class

$$\mathcal{K}_{0,j} := \left\{ x \to K\left(\frac{w - W_j(x)}{h}\right) : w \in I, h \in (0,1] \right\}.$$

We will show that under our assumptions, $\mathcal{K}_{0,j}$ is a VC class for each j, hence $N(\varepsilon, \mathcal{K}_{0,j}) \le C\varepsilon^{-v}$ for some $v \ge 1$. Notice that $\mathcal{K}_{0,j} = \prod_{l=1}^{d} \mathcal{K}_{0,j,l}$ where

$$\mathcal{K}_{0,j,l} := \left\{ x \to k\left(\frac{w_l - W_{jl}(x)}{h}\right) : w_l \in I_l, h \in (0,1] \right\},$$

where $I_l := \{w_l : w \in I\}$ and $W_j(x) = (W_{j1}(x), ..., W_{jd}(x))^\top$. Hence, by Lemma 2.6.18 in van der Vaart and Wellner (1996, p. 147) it suffices to prove that $\mathcal{K}_{0,j,l}$ is a VC subgraph class. Moreover, by the same lemma, without loss of generality (as k is of bounded variation), we can assume that k is non-decreasing on $\mathbb{R}$. Recall that $\mathcal{K}_{0,j,l}$ is a VC subgraph class if and only if its class of subgraphs is a VC class of sets, which holds if the class

$$\mathcal{S}_{\mathcal{K}} = \left\{ \left\{ (x,t) : k\left(\frac{w_l - W_{jl}(x)}{h}\right) < t \right\} : w_l \in I_l, h \in (0,1] \right\},$$

is a VC subgraph class. But this follows from an application of Lemma 2.6.18 in van der Vaart and Wellner (1996, p. 147), after noticing that

$$\mathcal{S}_{\mathcal{K}} = \left\{ \left\{ (x,t) : hk^{-1}(t) - w_l + W_{jl}(x) > 0 \right\} : w_l \in I_l, h \in (0,1] \right\},$$

where $k^{-1}(t) = \inf\{u : k(u) \ge t\}$. Then, Lemma 2.6.15 and Lemma 2.6.18(iii) in van der Vaart and Wellner (1996, p. 146-147) imply that the class $\mathcal{K}_{0,j}$ is a VC class for each $j = 1, ..., N_{1\varepsilon}$. Set, for each $\varepsilon > 0$, $N_{2\varepsilon j} := N(\varepsilon, \mathcal{K}_{0,j})$.

On the other hand, our assumptions on the kernel imply that

$$|K(x) - K(y)| \le |x - y| \, K^*(y), \tag{B-32}$$

where $K^*(y)$ is bounded and integrable, see Hansen (2008, p. 741). Hence, for any $K(w - W(\cdot)/h) \in \mathcal{K}_0$, there exist $W_j \in \mathcal{W}$, $w_{jk} \in I$ and $h_{jk} \in (0,1]$, $j = 1, ..., N_{1\varepsilon}$ and $k = 1, ..., N_{2\varepsilon j}$, such that

$$\begin{aligned} E\left[\left|K\left(\frac{w - W(X)}{h}\right) - K\left(\frac{w_{jk} - W_j(X)}{h_{jk}}\right)\right|^2\right] &\leq 2E\left[\left|K\left(\frac{w - W(X)}{h}\right) - K\left(\frac{w - W_j(X)}{h}\right)\right|^2\right] \\ &\quad + 2E\left[\left|K\left(\frac{w - W_j(X)}{h}\right) - K\left(\frac{w_{jk} - W_j(X)}{h_{jk}}\right)\right|^2\right] \\ &\leq C\varepsilon^2 h^{-1} E\left[K^*\left(\frac{w - W_j(X)}{h}\right)\right] + 2\varepsilon^2 \\ &\leq C_1^2 \varepsilon^2, \end{aligned}$$

where recall W_j is such that $\|W - W_j\|_\infty \leq \varepsilon^2$, and the second inequality uses that K is bounded to conclude $|K((w - W(X))/h) - K((w - W_j(X))/h)| \leq C$. Hence, (B-31) follows. *Q.E.D.*

The following lemma extends some results in Einmahl and Mason (2005) to kernel estimators with nonparametric generated regressors.

Lemma B.4 *Let $J = I^\varepsilon = \{w \in \mathcal{X}_\mathcal{W} : |w - v| \leq \varepsilon,\ v \in I\}$, for I a compact set of $\mathcal{X}_\mathcal{W} \subset \mathbb{R}^d$ for some $0 < \varepsilon < 1$. Also assume that Assumptions B.1 – B.5 hold. Further, assume that Υ is a VC class, with envelope function G satisfying*

$$\exists M > 0 : \ G(Z)\mathbb{I}\{W(X) \in J\} \leq M, \ a.s. \tag{B-33}$$

or for some $s > 2$

$$\sup_{(W,w)\in\mathcal{W}\times J} E[G^s(Z)\,|W(X) = w] < \infty. \tag{B-34}$$

Then we have for any $c > 0$ and $b_n \downarrow 0$, with probability 1,

$$\limsup_{n\to\infty} \sup_{c_n^\gamma \leq h \leq b_n} \frac{\sup_{\psi\in\Psi_I} \sqrt{nh^d}\,|\widehat{m}_h(\psi) - E\widehat{m}_h(\psi)|}{\sqrt{((\log(1/h^d)) \vee \log\log n)}} =: Q(c) < \infty,$$

where $c_n := c(\log n/n)$, $\gamma := 1/d$ in the bounded case (B-33) and $\gamma := (1/d - 2/ds)$ under assumption (B-34).

Proof of Lemma B.4: We only prove this lemma for the unbounded case, the proof for the bounded case follows similar steps and therefore is omitted. For any $k = 1, 2, ...,$ and $\varphi \in \Upsilon$, set $n_k := 2^k$, and

$$\varphi_k(Z_i) := \varphi(Z_i)\mathbb{I}\left\{G(Z_i) < c_{n_k}^{-1/s}\right\},$$

where s is as in (B-34).
For fixed h_0, $0 < h_0 < 1$, and for $n_{k-1} \leq n \leq n_k$, $w \in I$, $c_{n_k}^\gamma \leq h \leq h_0$ and $\varphi \in \Upsilon$, let

$$\widehat{m}_h^{(k)}(\psi) = \frac{1}{nh^d}\sum_{i=1}^n \varphi_k(Z_i)\,K\left(\frac{w - W(X_i)}{h}\right).$$

First, we shall prove that under our assumptions there exists a constant $Q_1(c) < \infty$, such that with probability 1,

$$\lim\sup_{k\to\infty} \Delta_k = Q_1(c), \tag{B-35}$$

where

$$\Delta_k := \max_{n_{k-1}\leq n\leq n_k} \sup_{c_{n_k}^{\gamma}\leq h\leq h_0} \frac{\sup_{\psi\in\Psi_I} \sqrt{nh^d}\left|\widehat{m}_h^{(k)}(\psi) - E\widehat{m}_h^{(k)}(\psi)\right|}{\sqrt{((\log(1/h^d)) \vee \log\log n)}}.$$

To that end, for $\psi \in \Psi_I$ and $c_{n_k}^{\gamma} \leq h \leq h_0$, let

$$v_h(Z_i,\psi) := \varphi(Z_i) K\left(\frac{w - W(X_i)}{h}\right) \text{ and } v_h^{(k)}(Z_i,\psi) := \varphi_k(Z_i) K\left(\frac{w - W(X_i)}{h}\right).$$

Define the class $\mathcal{V}_k(h) := \{v_h^{(k)}(\cdot,\psi) : \psi \in \Psi_I\}$ and note that for each $v_h^{(k)} \in \mathcal{V}_k(h)$,

$$\sup_{z\in\mathcal{Z}} \| \, v_h^{(k)}(z,\cdot) \, \|_{\Psi_I} := \sup_{z\in\mathcal{Z}} \sup_{\psi\in\Psi_I} |v_h^{(k)}(z,\psi)| \leq \|K\|_\infty c_{n_k}^{-1/s}.$$

Also, observe that

$$E[|v_h^{(k)}(Z,\psi)|^2] \leq E[|v_h(Z,\psi)|^2] \leq E\left[\left|\varphi(Z_i) K\left(\frac{w - W(X_i)}{h}\right)\right|^2\right].$$

Using a conditioning argument, we infer that the last term is

$$\begin{aligned}
&\leq \int E[G^2(Z)\,|W(X) = w']K^2\left(\frac{w - w'}{h}\right) f\left(w'\middle| W\right) dw' \\
&\leq C\int h^d K^2(u) f\left(w - uh\middle| W\right) du \\
&\leq C\,\|K\|_{2,\lambda}^2 \|f\|_{\mathcal{W},\infty} h^d =: C_1 h^d.
\end{aligned}$$

Thus,

$$\sup_{v\in\mathcal{V}_k(h)} E[|v(Z)|^2] \leq C_1 h^d. \tag{B-36}$$

Set for $j,k \geq 0$, $h_{j,k} := 2^j c_{n_k}^{\gamma}$ and define $\mathcal{V}_{j,k} := \{v_h^{(k)}(\cdot,\psi) : \psi \in \Psi_I \text{ and } h_{j,k} \leq h \leq h_{j+1,k}\}$. Clearly by (B-36),

$$\sup_{v\in\mathcal{V}_{j,k}} E[|v(Z)|^2] \leq Ch_{j,k}^d =: \sigma_{j,k}^2. \tag{B-37}$$

Define the product class of functions $\mathcal{G}_0 := \mathcal{K}_0 \cdot \mathcal{C} \cdot \Upsilon$, where

$$\mathcal{K}_0 = \left\{x \to K\left(\frac{w - W(x)}{h}\right) : w \in I, W \in \mathcal{W}, h \in (0,1]\right\}$$

and $\mathcal{C} = \{z \to f(z) = \mathbb{I}\{G(z) < c\} : c > 0\}$. It is straightforward to prove that, for some positive constant C,

$$N_{[\cdot]}(\varepsilon, \mathcal{G}_0, \|\cdot\|_2) \leq N(C\varepsilon, \mathcal{K}_0, \|\cdot\|_2) \times N(C\varepsilon, \mathcal{C}, \|\cdot\|_2) \times N(C\varepsilon, \Upsilon, \|\cdot\|_2). \tag{B-38}$$

Hence, by Lemma B.3 and our assumptions on the class Υ, we obtain that $\log N_{[\cdot]}(\varepsilon, \mathcal{G}_0, \|\cdot\|_2) \leq C\varepsilon^{-v_0}$, for some $v_0 < 2$. Note that $\mathcal{V}_{j,k} \subset \mathcal{G}_0$, so $\log N_{[\cdot]}(\varepsilon, \mathcal{V}_{j,k}, \|\cdot\|_2) \leq C\varepsilon^{-v_0}$ also holds.

Define $l_k := \max\{j : h_{j,k} \leq 2h_0\}$ if this set is non-empty, which is obviously the case for large enough k. Also, define

$$a_{j,k} := \sqrt{n_k h_{j,k}^d \left(\left|\log(1/h_{j,k}^d)\right| \vee \log\log n_k\right)}.$$

Then, by Lemma B.1 and (B-37), for some positive constant C_3, for all k sufficiently large and all $0 \leq j \leq l_k - 1$,

$$\begin{aligned} E^*\left[\sup_{v\in\mathcal{V}_{j,k}}\left|\sum_{i=1}^{n_k}\varepsilon_i v(Z_i)\right|\right] &\leq C_3\sqrt{n_k h_{j,k}^d} \\ &\leq C_3 a_{j,k} \end{aligned} \tag{B-39}$$

where $\{\varepsilon_i\}_{i=1}^n$ is a sequence of iid Rademacher variables, independent of the sample $\{Z_i\}_{i=1}^n$.

By definition, $2h_{l_k,k} = h_{l_k+1,k} \geq 2h_0$, which implies that for $n_{k-1} \leq n \leq n_k$, $[c_n^\gamma, h_0] \subset [c_{n_k}^\gamma, h_{l_k,k}]$. Thus, for large enough k and for any $\rho > 1$,

$$A_k(\rho) := \{\Delta_k \geq 2A_1(C_3+\rho)\} \subset \bigcup_{j=0}^{l_k-1}\left\{\max_{n_{k-1}\leq n\leq n_k}\left\|\sqrt{n}\mathbb{G}_n\right\|_{\mathcal{V}_{j,k}} \geq A_1(C_3+\rho)a_{j,k}\right\},$$

where C_3 is the constant in (B-39) and A_1 is the universal constant in Lemma B.2.

Set for any $\rho > 1$, $j \geq 0$ and $k \geq 1$,

$$p_{j,k}(\rho) := P^*\left(\max_{n_{k-1}\leq n\leq n_k}\left\|\sqrt{n}\mathbb{G}_n\right\|_{\mathcal{V}_{j,k}} \geq A_1(C_3+\rho)a_{j,k}\right).$$

Note that $\sqrt{n_k h_{j,k}^d}c_{n_k}^{1/s} = 2^{jd/2}\sqrt{n_k c_{n_k}}$, $n_k c_{n_k}\log\log n_k \geq c\left(\log\log n_k\right)^2$ and that $a_{j,k}^2/n_k h_{j,k}^d \geq \log\log n_k$, for all k sufficiently large. Hence, applying Talagrand's inequality, see Lemma B.2, with $\sigma_{\mathcal{G}}^2 = \sigma_{j,k}^2$, $M = \|K\|_\infty c_{n_k}^{-1/s}$ and $t = \rho a_{j,k}$, we obtain

$$\begin{aligned} p_{j,k}(\rho) &\leq 2\left[\exp\left(-\frac{A_2\rho^2 a_{j,k}^2}{n_k C h_{j,k}^d}\right) + \exp\left(-\frac{A_2\rho a_{j,k}c_{n_k}^{1/s}}{\|K\|_\infty}\right)\right] \\ &\leq 2\left[\exp\left(-\frac{A_2\rho^2}{C}\log\log n_k\right) + \exp\left(-\frac{2^{jd/2}A_2\rho}{\|K\|_\infty}\sqrt{n_k c_{n_k}\log\log n_k}\right)\right] \\ &\leq 2\left(\log n_k\right)^{-\frac{A_2\rho^2}{C}} + 2\left(\log n_k\right)^{-\frac{A_2\rho 2^{jd/2}c^{1/2}}{\|K\|_\infty}} \\ &\leq 4\left(\log n_k\right)^{-\rho A_3}, \end{aligned}$$

where $A_3 := A_2\left(1/C \wedge c^{1/2}/\left\|K\right\|_\infty\right)$. Since $l_k \leq 2\log n_k$ for large enough k,

$$P^*(A_k(\rho)) \leq P_k(\rho) := \sum_{j=0}^{l_k-1}p_{j,k}(\rho) \leq 8(\log n_k)^{1-\rho A_3}.$$

Then, (B-35) follows from Borel-Cantelli by taking ρ sufficiently large, e.g. $\rho \geq 3/A_3$.

Next, for $n_{k-1} \leq n \leq n_k$, $w \in I$, $c_{n_k}^{\gamma} \leq h \leq h_0$ and $\varphi \in \Upsilon$, let

$$\overline{m}_h^{(k)}(\psi) = \frac{1}{nh^d} \sum_{i=1}^{n} \overline{\varphi}_k (Z_i) K \left(\frac{w - W(X_i)}{h} \right),$$

where $\overline{\varphi}_k(Z_i) = \varphi(Z_i) \mathbb{I}\left\{ G(Z) \geq c_{n_k}^{-1/s} \right\}$. Then, following the same steps as in Lemma 4 in Einmahl and Mason (2005, p. 1400), we obtain, with probability 1,

$$\lim_{k\to\infty} \max_{n_{k-1}\leq n \leq n_k} \sup_{c_n^{\gamma} \leq h \leq h_0} \frac{\sup_{\psi \in \Psi_I} \sqrt{nh^d} \left| \overline{m}_h^{(k)}(\psi) - E\overline{m}_h^{(k)}(\psi) \right|}{\sqrt{((\log(1/h^d)) \vee \log\log n)}} = 0. \tag{B-40}$$

Finally, (B-35) and (B-40) together prove the result. *Q.E.D.*

Our next results involve uniform convergence rates for kernel estimators. For a_n and b_n as in Assumption B.5 and r as in Assumption B.4 , define

$$d_n := \sqrt{\frac{\log a_n^{-d} \vee \log\log n}{na_n^d}} + b_n^r.$$

The following are classical smoothness conditions that are needed to control bias.

Assumption B.6 *For all $W \in \mathcal{W}$ and $x \in \mathcal{X}_X$: (i) $f(w|W)$ and (ii) $m(w|W)$ and $f_X(x|w, W)$ are r-times continuously differentiable in w, with uniformly (in w, W and x) bounded derivatives (including zero derivatives), where r is as in Assumption B.4.*

Define as in the main text $m(w|W) := E[Y|W = w]$, $w \in \mathcal{X}_W \subset \mathbb{R}^d$, and its nonparametric NW estimator is $\widehat{m}(w|W) := \widehat{T}(w|W)/\widehat{f}(w|W)$, where $\widehat{T}(w|W) := n^{-1}h^{-d}\sum_{i=1}^{n} Y_i K((w - W_i)/h)$ and $\widehat{f}(w|W) := n^{-1}h^{-d}\sum_{i=1}^{n} K((w - W_i)/h)$.

Lemma B.5 *Let Assumptions B.1 – B.6(i) hold. Then, we have,*

$$\sup_{a_n \leq h \leq b_n} \sup_{w \in \mathcal{X}_{\mathcal{W}}; W \in \mathcal{W}} |\widehat{f}(w|W) - f(w|W)| = O_{P^*}(d_n).$$

Proof of Lemma B.5: Write

$$\begin{aligned} \sup |\widehat{f}(w|W) - f(w|W)| &\leq \sup |\widehat{f}(w|W) - E\widehat{f}(w|W)| + \sup |E\widehat{f}(w|W) - f(w|W)| \\ &\equiv I_{1n} + I_{2n}, \end{aligned}$$

where henceforth the sup is over the set $a_n \leq h \leq b_n$, $w \in \mathcal{X}_{\mathcal{W}}$ and $W \in \mathcal{W}$. An inspection of the proof of Lemma B.4 with $\varphi(\cdot) \equiv 1$ shows that we can take $I = \mathcal{X}_{\mathcal{W}}$, so we obtain

$$I_{1n} = O_{P^*}\left(\sqrt{\frac{\log a_n^{-d} \vee \log\log n}{na_n^d}} \right).$$

By the classical change of variables, Taylor expansion and Assumptions B.4 and B.6, $I_{2n} = O_{P^*}(b_n^r)$. *Q.E.D.*

The following results establish rates of convergence for kernel estimates of $m(w|W)$ and $T(w|W)$.

Lemma B.6 *Let Assumptions B.1 – B.6 hold. Then, we have*

$$\sup_{a_n \le h \le b_n} \sup_{w \in \mathcal{X}_{\mathcal{W}}; W \in \mathcal{W}} |\widehat{T}(w|W) - T(w|W)| = O_{P^*}(d_n).$$

Proof of Lemma B.6: The proof for $\widehat{T}$ follows the same arguments as for $\widehat{f}$, and hence, it is omitted. *Q.E.D.*

Define

$$t_n(w|\,W) := \mathbb{I}(f\,(w|\,W) \ge \tau_n) \qquad \text{and} \qquad \widehat{t}_n(w|\,W) := \mathbb{I}(\widehat{f}\,(w|\,W) \ge \tau_n).$$

Lemma B.7 *Let Assumptions B.1 – B.6 hold. Then, we have*

$$\sup_{a_n \le h \le b_n} \sup_{w \in \mathcal{X}_{\mathcal{W}}; W \in \mathcal{W}} |\widehat{m}(w|W) - m(w|W)| t_n(w|\,W) = O_{P^*}(\tau_n^{-1} d_n) + O_{P^*}(\tau_n^{-2} d_n^2)$$

and

$$\sup_{a_n \le h \le b_n} \sup_{w \in \mathcal{X}_{\mathcal{W}}; W \in \mathcal{W}} |\widehat{m}(w|W) - m(w|W)| \widehat{t}_n(w|\,W) = O_{P^*}(\tau_n^{-1} d_n).$$

Proof of Lemma B.7: We write

$$\widehat{m}(w|W) - m(w|W) = a_n\,(w|\,W) + r_n(w|\,W),$$

where

$$a_n\,(w|\,W) := f^{-1}\,(w|\,W) \left(\widehat{T}\,(w|\,W) - T\,(w|\,W) - m(w|W) \left(\widehat{f}\,(w|\,W) - f\,(w|\,W) \right) \right),$$

$T\,(w|\,W) := m\,(w|\,W)\,f\,(w|\,W)$ and

$$r_n(w|\,W) := -\frac{\widehat{f}\,(w|\,W) - f\,(w|\,W)}{\widehat{f}\,(w|\,W)\,f\,(w|\,W)} a_n\,(w|\,W)\,.$$

Since $f^{-1}\,(w|\,W)\,t_n(w|\,W)$ is bounded by τ_n^{-1}, we obtain from previous results $\sup |r_n(w|\,W)| = O_{P^*}(\tau_n^{-2} d_n^2)$. For the second equality, note that

$$\widehat{m}(w|W) - m(w|W) = \left\{ \frac{\widehat{T}\,(w|\,W) - T\,(w|\,W)}{\widehat{f}\,(w|\,W)} - m(w|W) \frac{\widehat{f}\,(w|\,W) - f\,(w|\,W)}{\widehat{f}\,(w|\,W)} \right\}.$$

Hence, since m is uniformly bounded, we obtain the uniform bound

$$\begin{aligned} |\widehat{m}(w|W) - m(w|W)| \widehat{t}_n(w|\,W) &\le \tau_n^{-1} \left| \widehat{T}\,(w|\,W) - T\,(w|\,W) \right| \\ &\quad + \tau_n^{-1} \left| \widehat{f}\,(w|\,W) - f\,(w|\,W) \right| |m(w|W)| \\ &= O_{P^*}(\tau_n^{-1} d_n). \end{aligned}$$

Q.E.D.

We now consider stronger versions of Assumptions B.5 and B.4 that are applicable to derivatives of kernel estimates such as

$$\dot{m}_h(\psi) = \frac{1}{nh^{d+1}} \sum_{i=1}^{n} \varphi\left(Z_i\right) \dot{K}\left(\frac{w - W\left(X_i\right)}{h}\right),$$

where $\dot{K}\left(w/h\right) = \dot{k}(w_1/h)\prod_{l=2}^{d} k(w_l/h)$, where $\dot{k}\left(u\right) = \partial k(u)/\partial u$ is the derivative of the kernel function k.

Assumption B.7 *The possibly data-dependent bandwidth h satisfies $P(a_n \leq h \leq b_n) \to 1$ as $n \to \infty$, for deterministic sequences of positive numbers a_n and b_n such that $b_n \to 0$ and $a_n^{d+2} n/\log n \to \infty$.*

Assumption B.8 *The kernel function $k\left(t\right) : \mathbb{R} \to \mathbb{R}$ is bounded, r-times continuously differentiable and satisfies the following conditions: $\int k\left(t\right) dt = 1$, $\int t^l k\left(t\right) dt = 0$ for $0 < l < r$, and $\int \left|t^r k\left(t\right)\right| dt < \infty$, for some $r \geq 2$, $\left|\partial^{(j)} k(t)/\partial t^j\right| \leq C$ and for some $v > 1$, $\left|\partial^{(j)} k(t)/\partial t^j\right| \leq C\left|t\right|^{-v}$ for $\left|t\right| > L_j$, $0 < L_j < \infty$, for $j = 1, 2$.*

Lemma B.8 *Under the conditions of Lemma B.4 but with B.7 and B.8 replacing B.5 and B.4, respectively, we have for any $c > 0$ and $b_n \downarrow 0$, with probability 1,*

$$\limsup_{n\to\infty} \sup_{c_n^{\gamma} \leq h \leq b_n} \frac{\sup_{\psi \in \Psi_I} \sqrt{nh^{d+2}} \left|\dot{m}_h(\psi) - E\dot{m}_h(\psi)\right|}{\sqrt{\left((\log(1/h^d)) \vee \log\log n\right)}} =: Q\left(c\right) < \infty.$$

Proof of Lemma B.8: The proof follows the same steps as that of Lemma B.4, and hence it is omitted. *Q.E.D.*

Appendix C Some Primitive Conditions

This section provides primitive conditions for some of the high level assumptions in the main text of the paper. These high level conditions can be classified into into three classes: 1. Assumptions on the smoothness and boundedness conditions regarding densities and regression functions; 2. Asymptotic inclusion assumptions for nonparametric estimators; and 3. Other high-level assumptions regarding properties of these estimates. Assumptions 3, 6(i) and 9 in the main text belong to class 1. Assumptions 6(ii) and 10(ii-iii) in the paper are of the type 2, while Assumptions 10(i) in the main text is an example of type 3. Primitive conditions for assumptions in the class 1 are generally model specific, see e.g. Klein and Spady (1993) for parametric generated regressors. Here we focus on primitive conditions for classes 2 and 3.

Assumption 6(ii) in the main text requires that $P(\widehat{m} \in \mathcal{T}_M^{\eta_m}) \to 1$, for some $\eta_m > d/2$ and $M > 0$. That is, it requires one to prove that

$$\sup_{x \in \mathcal{X}_X, W_1, W_2 \in \mathcal{W}} \frac{\left|\widehat{m}\left(W_1(x)\right| W_1\right) - \widehat{m}\left(W_2(x)\right| W_2\right)|}{\left\|W_1 - W_2\right\|_\infty} = O_{P^*}(1) \tag{C-41}$$

and

$$P(\widehat{m}\left(\cdot|\, W\right) \in C_M^{\eta_m}(\mathcal{X}_W)) \to 1 \text{ for all } W \in \mathcal{W}. \tag{C-42}$$

We now provide primitive conditions for (C-41) and (C-42) when densities are bounded away from zero. Similar arguments can be used to find primitive conditions with vanishing densities when random trimming is used, simply multiplying the rates d_{1n} and d_{2n} below by τ_n^{-1}.

To verify (C-41), we write

$$\begin{aligned}\widehat{m}\left(W_1(x)|\, W_1\right) - \widehat{m}\left(W_2(x)|\, W_2\right) &= \frac{\widehat{T}\left(W_1(x)|\, W_1\right) - \widehat{T}\left(W_2(x)|\, W_2\right)}{\widehat{f}\left(W_1(x)|\, W_1\right)} \\ &+ \frac{\left[\widehat{f}\left(W_2(x)|\, W_2\right) - \widehat{f}\left(W_1(x)|\, W_1\right)\right]\widehat{m}\left(W_2(x)|\, W_2\right)}{\widehat{f}\left(W_1(x)|\, W_1\right)}.\end{aligned} \tag{C-43}$$

Define $\psi := (x, W_1, W_2) \in \Psi := \mathcal{X}_X \times \mathcal{W} \times \mathcal{W}$,

$$v_{h,1}(Z_i, \psi) := Y_i \frac{h}{\|W_1 - W_2\|_\infty}\left\{K\left(\frac{W_1(x) - W_1(X_i)}{h}\right) - K\left(\frac{W_2(x) - W_2(X_i)}{h}\right)\right\}$$

and

$$\widehat{m}_{h,1}(x, \psi) := \frac{1}{nh^{d+1}}\sum_{i=1}^n v_{h,1}(Z_i, \psi).$$

It is straightforward to prove that, for each $\psi \in \Psi$,

$$E[|v_{h,1}(Z_i, \psi)|^2] \leq Ch^d.$$

Then, arguing as in Lemma B.4 one can show that

$$\sup_{a_n \leq h \leq b_n} \sup_{\psi \in \Psi} |\widehat{m}_{h,1}(\psi) - E\widehat{m}_{h,1}(\psi)| = O_{P^*}(d_{1n}),$$

where

$$d_{1n} = \sqrt{\frac{\log a_n^{-d} \vee \log\log n}{na_n^{d+2}}}.$$

On the other hand, the typical arguments used to handle the bias of kernel derivatives yield

$$\sup_{a_n \leq h \leq b_n} \sup_{\psi \in \Psi} |E\widehat{m}_{h,1}(\psi)| = O_{P^*}(1).$$

A similar conclusion can be obtained for the other terms in (C-43). Then, a primitive condition for (C-41) is that $d_{1n} = O(1)$.

To give a primitive condition for (C-42) we consider the case $d = 2$, which arises in models such as the binary choice model with selection discussed in the main text. In this case, we can take $\eta_m = 1 + \eta_q$, with $0 < \eta_q < 1$. Then, (C-42) reduces to showing that, for $j = 1, 2$, for each $W \in \mathcal{W}$,

$$\sup_{w \neq w'} \frac{|\partial\widehat{m}\left(w|\, W\right)/\partial w_j - \partial\widehat{m}\left(w'|\, W\right)/\partial w_j|}{|w - w'|^{\eta_q}} = O_{P^*}(1). \tag{C-44}$$

To that end, define for $\psi := (w, w') \in \Psi := \mathcal{X}_W \times \mathcal{X}_W$,

$$\begin{aligned}\widehat{m}_{h,2}(\psi) &:= \frac{1}{nh^{d+2}} \sum_{i=1}^{n} Y_i \frac{h}{|w-w'|^{\eta_q}} \left\{ \dot{K}_j \left(\frac{w - W(X_i)}{h} \right) - \dot{K}_j \left(\frac{w' - W(X_i)}{h} \right) \right\}, \\ &=: \frac{1}{nh^{d+2}} \sum_{i=1}^{n} v_{h,2}(Z_i, \psi),\end{aligned}$$

where $\dot{K}_1(w) = \partial k(w_1)/\partial w_1 k(w_2)$ and $\dot{K}_2(w) = \partial k(w_2)/\partial w_2 k(w_1)$, $w = (w_1, w_2)$. Then, using the arguments of Lemma B.4 one can show that

$$\sup_{a_n \leq h \leq b_n} \sup_{\psi \in \Psi} \left| \widehat{m}_{h,2}(\psi) - E\widehat{m}_{h,2}(\psi) \right| = O_{P^*}(d_{2n}),$$

and

$$\sup_{a_n \leq h \leq b_n} \sup_{\psi \in \Psi} \left| E\widehat{m}_{h,2}(\psi) \right| = O_{P^*}(1),$$

where

$$d_{2n} = \sqrt{\frac{\log a_n^{-d} \vee \log \log n}{n a_n^{d+4}}}.$$

Hence, for $d = 2$ a primitive condition for Assumption 6(ii) is that $na_n^6/\log(n) \to \infty$ and Assumption B.8 above hold. For a general d a similar approach can be used, provided that the corresponding kernel derivatives are of bounded variation. Similar conditions such as Assumptions 10(iii) and 13(ii) can be verified analogously.

Next, we consider primitive conditions regarding the first step estimator in Assumption 10 when $\widehat{g}_i$ is the NW estimator. Consider the following assumption:

Assumption C.1 *(i) The regression g_0 is estimated by a NW kernel estimator with a kernel function satisfying Assumption B.4 with $r = \rho$ and a possibly stochastic bandwidth $\widehat{h}_{gn}$ satisfying $P(l_n \leq \widehat{h}_{gn} \leq u_n) \to 1$ as $n \to \infty$, for deterministic sequences of positive numbers l_n and u_n such that $u_n \to 0$, $n(l_n^p/\log n)^{\kappa/(\kappa-1)} \to \infty$ and $nu_n^{2\rho\kappa} \to 0$; (ii) the function g_0 and the density $f_X(\cdot)$ of X are ρ-times continuously differentiable in x, with bounded derivatives. The density $f_X(\cdot)$ is bounded away from zero. Furthermore $g_0 \in \mathcal{G} \subset C_M^{\eta_g}(\mathcal{X}_X)$, $P(\widehat{g} \in \mathcal{G}) \to 1$, for some $\eta_g > d$.*

Examples of random bandwidths that satisfy our assumptions are plug-in bandwidths of the form $\widehat{h}_{gn} = \widehat{c}h_{gn}$ with $\widehat{c}$ is bounded in probability and h_n a suitable deterministic sequence. If $h_{gn} = cn^{-\delta}$, for some constant $c > 0$, then Assumption C.1(i) requires that $1/2\rho < \delta < (\kappa - 1)/p\kappa$, so ρ needs to be greater than $p\kappa/2(\kappa - 1)$.

We now prove that under the primitive condition C.1 above, the high level Assumption 10(i) in the main text and $\|\widehat{g} - g_0\|_\infty = o_P(n^{-1/2\kappa})$ hold. To that end, using our Theorem 2.1, without trimming and with the class $W(x) = \{x_1\}$, we obtain that

$$\begin{aligned}
R_{1n}(\widehat{\alpha}) &:= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g_{0i} - \widehat{g}_i)\partial_{\bar{g}} m(W_{0i})\widehat{\phi}_W^{\perp}(X_i) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - \widehat{g}_i)\partial_{\bar{g}} m(W_{0i})\widehat{\phi}_W^{\perp}(X_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - g_{0i})\partial_{\bar{g}} m(W_{0i})\widehat{\phi}_W^{\perp}(X_i) \\
&= \frac{-1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - g_{0i})E(\partial_{\bar{g}} m(W_{0i})\widehat{\phi}_W^{\perp}(X_i)|X_{1i}) + o_P(1).
\end{aligned}$$

Hence, $|R_{1n}(\widehat{\alpha}) + G_n(\widehat{\alpha})| = o_P(1)$. Similarly, for $2 \leq j \leq \kappa$,

$$\begin{aligned}
R_{jn}(\widehat{\alpha}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g_{0i} - \widehat{g}_i)(g_{0i} - \widehat{g}_i)^{j-1}\partial_{\bar{g}}^{(j)} m(W_{0i})\widehat{\phi}_W^{\perp}(X_i) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - \widehat{g}_i)(g_{0i} - \widehat{g}_i)^{j-1}\partial_{\bar{g}}^{(j)} m(W_{0i})\widehat{\phi}_W^{\perp}(X_i) \\
&- \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - g_{0i})(g_{0i} - \widehat{g}_i)^{j-1}\partial_{\bar{g}}^{(j)} m(W_{0i})\widehat{\phi}_W^{\perp}(X_i) \\
&= \frac{-1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - g_i)E((g_{0i} - \widehat{g}_i)^{j-1}\partial_{\bar{g}}^{(j)} m(W_{0i})\widehat{\phi}_W^{\perp}(X_i)|X_{1i}) + o_P(1) \\
&= o_P(1),
\end{aligned}$$

where the last equality follows from Lemma A.4 in Appendix A. On the other hand, an application of Lemma B.7 implies

$$\|\widehat{g} - g_0\|_\infty = \sqrt{\frac{\log n}{n l_n^p}} + u_n^\rho = o_P(n^{-1/2\kappa}).$$

Finally, the condition $P(\widehat{g} \in C_M^{\eta_g}(\mathcal{X}_X)) \to 1$ can be verified as in (C-42) above.

Appendix D Some Generic Applications

In this section, we illustrate the general applicability of Theorems 2.1 and 2.2 in the main text to a variety of settings in semiparametric estimation and testing. In particular, we summarize how the asymptotic distribution of semiparametric estimators such as Ichimura (1993), Klein and Spady (1993) and Rothe (2009) may be derived using our results, which allow for data-driven bandwidths, random trimming and estimated weights. We also propose a new test for the null hypothesis of zero conditional average treatment effect, and discuss its properties based on the main results of the paper. Throughout this section technicalities are omitted for the sake of clarity, since our goal here is only to sketch how our results could be used for classes of applications beyond the specific ones provided in the main text.

D1 Ichimura's (1993) Estimator

Consider the class of functions $\mathcal{W} = \{x \to W(\theta) := v(\theta, x) : \theta \in \Theta \subset \mathbb{R}^{d_X}\}$, where $v(\cdot, \cdot)$ is a known function, i.e. $v(\theta, x) = x^\top \theta$. The class $\mathcal{W}$ trivially satisfies Assumption 2 in the main text. Denote

$W_0 := W(\theta_0)$, $W_{i0} := v(\theta_0, X_i)$ and $W_i(\theta) := v(\theta, X_i)$ for $i = 1, \ldots, n$, where $\left\{Y_i, X_i^\top\right\}_{i=1}^n$ is a random sample from the joint distribution of $\left(Y, X^\top\right)^\top$ that fulfills the index restriction $E[Y|X] = E[Y|W_0] =: m(W_0)$ for $\theta_0 \in \Theta$. Consider the following semiparametric least squares function

$$\mathcal{S}_n(\theta) := \frac{1}{n}\sum_{i=1}^n \widehat{t}_{ni}\{Y_i - \widehat{m}_{i\theta}\}^2, \tag{D-45}$$

where $\widehat{m}_{i\theta} := \widehat{m}(W_i(\theta)|W(\theta))$, $\widehat{t}_{ni} := \mathbb{I}(\widehat{f}(W_i(\widetilde{\theta})|W(\widetilde{\theta})) \geq \tau_n)$, $\tau_n \to 0$ as $n \to \infty$ at a rate that satisfies Assumption 11 in the main text, and $\widetilde{\theta}$ is a preliminary consistent estimator for θ_0, i.e. $\widetilde{\theta}$ could be an estimator that minimizes (D-45) but with $\widehat{t}_{ni} = \mathbb{I}(X_i \in A)$ for a compact set $A \subset \mathcal{X}_X$, and both $\widehat{m}$ and $\widehat{f}$ defined as in Section 2 in the main text. The proposed estimator $\widehat{\theta}$ of θ_0 is the minimizer of this objective function:

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \mathcal{S}_n\left(\theta\right). \tag{D-46}$$

The asymptotic distribution of the estimator will be established here by a combination of standard methods and our Theorem 2.1. Consider the first order conditions

$$0 = \sqrt{n}\partial_\theta \mathcal{S}_n(\widehat{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{Y_i - \widehat{m}_{i\widehat{\theta}}\}\partial_\theta \widehat{m}_{i\widehat{\theta}}\widehat{t}_{ni}, \tag{D-47}$$

where $\partial_\theta \widehat{m}_{i\widehat{\theta}} := \partial\widehat{m}(W_i(\theta)|W(\theta))/\partial\theta|_{\theta=\widehat{\theta}}$. By a Taylor series expansion,

$$\sqrt{n}(\widehat{\theta} - \theta_0) = G_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \{Y_i - \widehat{m}_{i\theta_0}\}\partial_\theta \widehat{m}_{i\theta_0}\widehat{t}_{ni} + o_P\left(1\right),$$

where $G_n = n^{-1}\sum_{i=1}^n \widehat{t}_{ni}\partial_\theta \widehat{m}_{i\widehat{\theta}}\partial_\theta^\top \widehat{m}_{i\bar{\theta}}$ and $\bar{\theta}$ is such that $|\bar{\theta} - \theta_0| \leq |\widehat{\theta} - \theta_0|$ a.s. By the uniform consistency results in Appendix B, and the continuous mapping theorem, it follows that

$$G_n \to_P \Lambda_0 =: E[\partial_\theta m(W_{i0})\partial_\theta^\top m(W_{i0})], \tag{D-48}$$

where $\partial_\theta m(W_{i0}) := \partial m(W_i(\theta)|W(\theta))/\partial\theta|_{\theta=\theta_0}$. By another application of the results in the main text with the uniform consistency of $\partial_\theta \widehat{m}_{i\theta_0}$ shown in Appendix B, we have

$$\Lambda_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \{Y_i - \widehat{m}_{i\theta_0}\}\partial_\theta \widehat{m}_{i\theta_0}\widehat{t}_{ni} = \Lambda_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \{Y_i - m(W_{0i})\}\partial_\theta m(W_{i0}) + o_P\left(1\right),$$

where the equality above follows from the fact that $E[\partial_\theta m(W_{i0})|W_0] = 0$, see Ichimura (1993, Lemma 5.6, p. 95). An application of Linderberg-Lévy CLT then yields

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d N(0, \Lambda_0^{-1}\Xi\Lambda_0^{-1}),$$

where $\Xi = E[\mathrm{var}(Y|X = X_i)\partial_\theta m(W_{i0})\partial_\theta^\top m(W_{i0})]$.

Remark D1.1 Delecroix, Hristache, and Patilea (2006) have also derived the asymptotic properties of Ichimura's (1993) estimator with random (non-vanishing) trimming and uniformly in the bandwidth, while Härdle, Hall, and Ichimura (1993) have shown the first order asymptotic properties of the semiparametric least squares estimator are not affected when plugging in a data-dependent bandwidth chosen jointly with $\widehat{\theta}$ in (D-46) using fixed trimming. Our result above essentially combines these features, while extending them to vanishing trimming.

D2 Klein and Spady's (1993) Estimator

Klein and Spady (1993) consider the binary choice model of the form

$$Y = \mathbb{I}(X^\top \theta_0 - u > 0), \tag{D-49}$$

where u and $X \in \mathcal{X}_X \subseteq \mathbb{R}^{d_X}$ are independent. Consider the class of functions $\mathcal{W} = \{x \to W(\theta) := v(\theta, x) : \theta \in \Theta \subset \mathbb{R}^{d_X}\}$, where $v(\cdot,\cdot)$ is a known function, i.e. $v(\theta, x) = x^\top \theta$. As before, denote $W_0 := W(\theta_0)$, $W_{i0} := v(\theta_0, X_i)$ and $W_i(\theta) := v(\theta, X_i)$ for $i = 1, \ldots, n$, where $\left\{Y_i, X_i^\top\right\}_{i=1}^n$ is a random sample from the joint distribution of $\left(Y, X^\top\right)^\top$ generated from model (D-49). In this case, the regression of Y given X satisfies the index restriction $E[Y|X] = E[Y|W_0] =: m(W_0)$ for $\theta_0 \in \Theta$. Our results can be used here to establish the asymptotic properties of a variant of Klein and Spady's (1993) estimator. First, define the semiparametric likelihood function as

$$\mathcal{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \{Y_i \log[\widehat{m}_{i\theta}] + (1 - Y_i) \log[1 - \widehat{m}_{i\theta}]\} \widetilde{t}_{in}, \tag{D-50}$$

where $\widehat{m}_{i\theta}$, and $\widetilde{t}_{in}$ are like those in Section D1 above but with $\widehat{g}$ replaced by one, and $\widetilde{\theta}$ is a preliminary consistent estimator for θ_0 (see Klein and Spady, 1993, footnote 4, p. 399 for examples). Similarly, both $\widehat{m}$ and $\widehat{f}$ are defined as in Section 2. The proposed estimator $\widehat{\theta}$ of θ_0 is the maximizer of this objective function:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta). \tag{D-51}$$

The asymptotic distribution of the estimator can be established here by a combination of standard methods and Theorem 2.1 in the main text. Now consider the first order conditions

$$0 = \sqrt{n} \partial_\theta \mathcal{L}_n(\widehat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{i\widehat{\theta}}\} \widehat{\psi}_i \widetilde{t}_{in} \tag{D-52}$$

where $\widehat{\psi}_{i\widehat{\theta}} := \partial_\theta \widehat{m}_{i\widehat{\theta}} [\widehat{m}_{i\widehat{\theta}}(1 - \widehat{m}_{i\widehat{\theta}})]^{-1}$ and $\partial_\theta \widehat{m}_{i\widehat{\theta}}$ is as in Section D1 above. Now by a Taylor series expansion,

$$\sqrt{n}(\widehat{\theta} - \theta_0) = H_n^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{i\theta_0}\} \widehat{\psi}_{i\widehat{\theta}} \widetilde{t}_{in} + o_P(1),$$

where

$$H_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial_\theta \widehat{m}_{i\widehat{\theta}} \partial_\theta^\top \widehat{m}_{i\bar{\theta}}}{\widehat{m}_{i\widehat{\theta}}(1 - \widehat{m}_{i\widehat{\theta}})} \widetilde{t}_{in}, \text{ and } |\bar{\theta} - \theta_0| \leq |\widehat{\theta} - \theta_0| \text{ a.s.}$$

From results in the main text, the uniform consistency results in Appendix B, and the continuous mapping theorem, it follows that

$$H_n \to_P \Delta_0 \equiv E[\partial_\theta m(W_{i0}) \partial_\theta^\top m(W_{i0}) [m(W_{i0})(1 - m(W_{i0}))]^{-1}], \tag{D-53}$$

where $\partial_\theta m(W_{i0})$ is as in Section D1. By another application of the results in the paper along with the uniform consistency of $\widehat{\psi}_{i\theta}$ we have, using that $E[\partial_\theta m(W_{i0})|W_0] = 0$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - \widehat{m}_{i\theta_0}\} \widehat{\psi}_{i\theta_0} \widetilde{t}_{in} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Y_i - m(W_{0i})\} \psi_{i\theta_0} + o_P(1),$$

where $\psi_{i\theta_0} = \partial_\theta m(W_{i0})/[m(W_{i0})(1-m(W_{i0}))]^{-1}$. It then follows from (D-53) that

$$\sqrt{n}(\widehat{\theta}-\theta_0) = \Delta_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{Y_i - m(W_{i0})}{m(W_{i0})(1-m(W_{i0}))}\right]\partial_\theta m(W_{i0}) + o_P(1),$$

and an application of Linderberg-Lévy CLT yields

$$\sqrt{n}(\widehat{\theta}-\theta_0) \to_d N(0,\Delta_0^{-1}).$$

Remark D2.2 The trimming function $\widetilde{t}_{in}$ here is slightly different from Klein and Spady's (1993) in that the latter behaved like a smoothed indicator. As in Klein and Spady (1993), the asymptotic variance Δ_0^{-1} equals the semiparametric efficiency bound for the binary choice model as originally derived by Chamberlain (1986) and Cosslett (1987). The main novelty here is that this asymptotic distribution is shown to hold for potentially data-driven bandwidths.

D3 Rothe's (2009) Estimator

Rothe (2009) considers the estimation of θ_0 in the 'endogenous' binary choice model of the form

$$Y = \mathbb{I}(\widetilde{X}^\top\theta_0 - u > 0), \tag{D-54}$$

where u is independent of $\widetilde{X} := (\widetilde{X}^e, \widetilde{X}^{-e})^\top \in \mathcal{X}_{\widetilde{X}^e} \times \mathcal{X}_{\widetilde{X}^{-e}} \subseteq \mathbb{R}^{d_{\widetilde{X}}}$ only conditionally on V where $\widetilde{X}^e = g_0(\widetilde{X}^{-e}, Z) + V$, $E[V|\widetilde{X}^{-e}, Z] = 0$, g_0 is a vector of conditional mean functions of each of the $d_{\widetilde{X}^e}$-'endogenous' components of $\widetilde{X}$, i.e. $\widetilde{X}^e$, given the $d_{\widetilde{X}^{-e}}$-'exogenous' components of $\widetilde{X}$, i.e. $\widetilde{X}^{-e}$, and some d_Z–vector of exogenous instruments Z. Notice that $d_{\widetilde{X}} = d_{\widetilde{X}^e} + d_{\widetilde{X}^{-e}}$. Let $X := (\widetilde{X}^\top, Z^\top)^\top$, and consider the class of functions $\mathcal{W} = \{x \to W(\theta, g) := v(\theta, g, x) : \theta \in \Theta \subset \mathbb{R}^{d_{\widetilde{X}}},\ g \in \mathcal{G}\}$, where $v(\cdot,\cdot,\cdot)$ is a $(1 + d_{\widetilde{X}^e})$-dimensional known function, i.e. $v(\theta, g, x) = (\widetilde{x}^\top\theta, \widetilde{x}^e - g(\widetilde{x}^{-e}, z))^\top$. As before, denote $W_0 := W(\theta_0, g_0)$, $W_{i0} := v(\theta_0, g_0, X_i)$ and $W_i(\theta, g) := v(\theta, g, X_i)$ for $i = 1, \ldots, n$, for an iid sample $\{Y_i, X_i^\top\}_{i=1}^n$ from the joint distribution of $(Y, X^\top)$. In this case, the regression of Y given X satisfies the index restriction $E[Y|X] = E[Y|W_0] =: m(W_{i0})$ for $\theta_0 \in \Theta$, and $g_0 \in \mathcal{G}$. Our results can be used here to establish the asymptotic properties of an efficient version of Rothe's (2009) estimator.

Rothe (2009) proposes estimating θ_0 as in (D-51) where $\widehat{m}_{i\theta} := \widehat{m}(W_i(\theta, \widehat{g})|W(\theta, \widehat{g}))$, $\widehat{t}_{ni} = \mathbb{I}(\widetilde{X}_i \in A)$ for a compact set $A \subset \mathcal{X}_{\widetilde{X}}$, and both $\widehat{m}$ and $\widehat{f}$ defined as in Section 2 in the main text. This type of fixed trimming affects the asymptotic distribution of his estimator. Consider instead $\widehat{t}_{ni} := \mathbb{I}(\widehat{f}(W_i(\widetilde{\theta}, \widehat{g})|W(\widetilde{\theta}, \widehat{g})) \geq \tau_n)$, $\tau_n \to 0$ as $n \to \infty$ at a rate that satisfies Assumption 7 in the main text, and $\widetilde{\theta}$ is a preliminary consistent estimator for θ_0, i.e. Rothe's (2009) original estimator, which uses fixed trimming $\widehat{t}_{ni} = \mathbb{I}(\widetilde{X}_i \in A)$ for a compact set $A \subset \mathcal{X}_{\widetilde{X}}$. The first order conditions are like (D-52). Similar arguments as in Section D2 and repeated application of Theorem 2.1 and Theorem 2.2 in the main text yields

$$\begin{aligned}\sqrt{n}(\widehat{\theta}-\theta_0) &= H_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Y_i - \widehat{m}_{i\theta_0}\}\widehat{\psi}_{i\theta_0}\widehat{t}_{ni} + o_P(1), \\ &= \Delta_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Y_i - m(W_{0i})\}\psi_{i\theta_0} - \Delta_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E[\psi_{i\theta_0}\partial_{\overline{g}} m(W_{0i})|\widetilde{X}^{-e}, Z]V_i + o_P(1),\end{aligned}$$

where $\partial_{\overline{g}} m(W_{0i}) := \partial m(W_i(\theta_0, \overline{g})|W_0)/\partial \overline{g}|_{\overline{g}=g_0}$, and the last equality holds uniformly in the bandwidth from Theorem 2.2 in the main text with $\psi_{i\theta_0} = \partial_\theta m(W_{0i})/[m(W_{0i})(1 - m(W_{0i}))]^{-1}$. Finally, an application of Linderberg-Lévy CLT yields

$$\sqrt{n}(\widehat{\theta} - \theta_0) \rightarrow_d N(0, \Delta_0^{-1} + \Delta_0^{-1}\Psi_0\Delta_0^{-1}),$$

where $\Psi_0 = E[\xi(W_{0i})V_iV_i^\top \xi^\top(W_{0i})]$ and

$$\xi(W_{0i}) = E\left[\left. \frac{\partial_\theta m(W_{0i})\partial_{\overline{g}} m(W_{0i})}{m(W_{0i})(1 - m(W_{0i}))} \right| \widetilde{X}^{-e}, Z \right].$$

Remark D3.3 The asymptotic variance of the estimator here is different from Rothe's (2009) because Rothe (2009) uses a fixed trimming function, i.e. $\widehat{t}_{ni} = \mathbb{I}(\widetilde{X}_i \in A)$ for a compact set $A \subset \mathcal{X}_{\widetilde{X}}$, that appears everywhere in the limiting distribution. However, if we neglect his trimming effect, taking $\widehat{t}_{ni} = 1$ for all $i = 1, \ldots, n$, then both expressions for the asymptotic variance will coincide, (see Rothe, 2009, Theorem 3, p. 55). Unlike Rothe's (2009) original calculations that use results in Chen, Linton, and van Keilegom (2003), the results in the main text can be used to allow for plug-in data driven bandwidths and random trimming, while avoiding the need to calculate pathwise derivatives.

D4 A New Nonparametric Test for Treatment Effect Heterogeneity

Consider the potential outcome framework in program evaluation where one observes a random sample $\{Y_i, D_i, X_i^\top\}_{i=1}^n$ from the joint distribution of $(Y, D, X^\top)$ that satisfies

$$Y = Y(1)D + Y(0)(1 - D),$$

where $Y(D)$ denotes the outcome under treatment ($D = 1$) or without it ($D = 0$), and X represents a vector of observed covariates. Suppose we are interested in the null hypothesis of zero average effects conditional on the covariates (CATE) as in Crump, Hotz, Imbens and Mitnik (2008) under unconfoundness (see Imbens and Wooldridge, 2009, for an up-to-date survey). The null hypothesis is

$$H_0 : E[Y(1) - Y(0)|X] = 0, \text{ a.s.}$$

where $Y(D)$ is independent of D given X, and the standard overlapping assumption holds, i.e. $0 < p(x) < 1$, where $p(x) := \Pr(D = 1|X = x)$ is the propensity score.

Crump, Hotz, Imbens and Mitnik (2008) provide a test of H_0. The motivation is that it is useful to know if a treatment has benefits for individuals with specific characteristics X, regardless of whether it affects outcomes or not on average in the population. Here we provide a different test, with a bandwidth that can be tuned in a data dependent way to improve power.

The following proposition demonstrate that H_0 can be tested based on the sample mean

$$\widehat{R}_{n,h}(x) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \{D_i - \widehat{p}(X_i)\} Y_i \mathbb{I}(X_i \leq x) \widehat{f}^2(X_i)$$

where $\widehat{p}(x)$ and $\widehat{f}(x)$ are NW estimators of $p(x)$ and $f(x)$, respectively, estimated with bandwidth h that satisfies certain assumptions in the main text. The justification of the testing procedure is given in the next result, which proof is standard and hence is omitted.

Proposition D.1 *$E[Y(1) - Y(0)|X] = 0$ a.s. if and only if $E[\{D - p(X)\} f^2(X) Y|X] = 0$ a.s.*

From Theorem 2.1 in the main text, the empirical process $\widehat{R}_{n,h}$ is asymptotically equivalent under the null of zero CATE to

$$R_n(x) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{D_i - p(X_i)\}\{Y_i - E[Y_i|X_i]\} f^2(X_i) \mathbb{I}(X_i \leq x),$$

which can be used to approximate the distribution of the Cramér-von Mises functional

$$C_n = \max_{h \in [a_n, b_n]} \int |\widehat{R}_{n,h}(x)|^2 F_n(dx),$$

where F_n represents the empirical distribution function of $\{X_i\}_{i=1}^n$. Let $\widehat{h}$ denote the solution to the optimization problem for C_n. As pointed out in the testing section of the main text, our results permit choosing the bandwidth in this way, which should lead to tests with better power properties than when the bandwidth is chosen with other objectives in mind (such as estimation). See e.g. Horowitz and Spokoiny (2001).

Set $m(\cdot) := E[Y_i|X = \cdot]$ and let $\widehat{m}(x)$ denote the NW estimator of $m(x)$. The asymptotic null distribution of C_n should be well approximated by the bootstrap process

$$R_n^*(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{D_i - \widehat{p}(X_i)\}\{Y_i - \widehat{m}(X_i)\} \widehat{f}^2(X_i) \mathbb{I}(X_i \leq x) \zeta_i,$$

where the variables $\{\zeta_i\}_{i=1}^n$ are independently generated from the original sample from a random variable ζ with bounded support, mean zero and variance one. The theoretical justification of this multiplier-type bootstrap can be demonstrated using standard methods, given our results regarding the process $\widehat{R}_{n,h}$. The conditions for the justification of the proposed test are rather mild. We only require Assumptions 1, 4, 5, and the conditions: (3') $m(x)$ and $T(x) := p(x)f(x)$ are r-times continuously differentiable in x, with bounded derivatives (including zero derivatives); and (6') $f, T \in C_M^\eta(\mathcal{X}_X)$ and $P(\widehat{f}, \widehat{T} \in C_M^\eta(\mathcal{X}_X)) \to 1$ for some $\eta > d/2$, where $\widehat{T}(\cdot) := \widehat{p}(\cdot)\widehat{f}(\cdot)$.

References

Ahn, H. (1997): "Semiparametric Estimation of a Single-Index Model with Nonparametrically Generated Regressors," *Econometric Theory*, 13(1), 3–31.

Ahn, H. and C. F. Manski (1993): "Distribution Theory for the Analysis of Binary Choice under Uncertainty with Nonparametric Estimation of Expectations," *Journal of Econometrics*, 56(3), 291–321.

Ahn, H. and J. L. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Problem," *Journal of Econometrics*, 58(1-2), 3–29.

Ai, C. and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Akritas, M. G.and I. van Keilegom (2001): "Non-parametric Estimation of the Residual Distribution," *Scandinavian Journal of Statistics*, 28(3), 549–567.

Andrews, D. W. K. (1994): "Asymptotics for Semiparametric Models via Stochastic Equicontinuity," *Econometrica*, 62(1), 43–72.

Andrews, D. W. K. (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11(3), 560–596.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Springer-Verlag, New York, 1 edn.

Bierens, H. J. (1982): "Consistent Model Specification Tests," *Journal of Econometrics*, 20(4), 105–134.

Blundell, R. W., and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models,"*Review of Economic Studies*, 71(7), 655–679.

Cattaneo, M. D., Crump, R. K. and Jansson, M. (2011): "Generalized Jacknife Estimators of Weighted Average Derivatives," *unpublished manuscript*.

Chamberlain, G. (1986). "Asymptotic Efficiency in Semiparametrid Models with Censoring," *Journal of Econometrics*, 32(2), 189–218.

Chen, X., O. B. Linton, and I. van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

Cosslett, S. R. (1987). "Efficiency Bounds for Distribution-free Estimators of the Binary Choice and Censored Regression Models," *Econometrica*, 55(3), 559–586.

Cragg, J. G. (1971): "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods,"*Econometrica*, 39(5), 829–44.

Crump, R., V. J. Hotz, G. W. Imbens and O. A. Mitnik (2008). "Nonparametric Tests for Treatment Effect Heterogeneity," *The Review of Economics and Statistics*, 90(3), 389–405.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *The Review of Economic Studies*, 70(1), 33–58.

Delecroix, M., M. Hristache, and V. Patilea (2006): "On Semiparametric M-Estimation in Single-Index Regression," *Journal of Statistical Planning and Inference*, 136(3), 730–769.

Delgado, M. A. and W. González Manteiga (2001): "Significance Testing in Nonparametric Regression Based on the Bootstrap," *Annals of Statistics*, 29(5), 1469–1507.

Einmahl, J. H. J., and D. M. Mason (2005): "Uniform in Bandwidth Consistency of Kernel-Type Function Estimators," *Annals of Statistics*, 33(3), 1380–1403.

Escanciano, J. C., D. T. Jacho-Chávez and A. Lewbel (2011): "Identification and Estimation of Semiparametric Two Step Models," Unpublished manuscript.

Escanciano, J. C., and K. Song (2010): "Testing Single-Index Restrictions with a Focus on Average Derivatives," *Journal of Econometrics*, 156(2), 377–391.

Hahn, J., and G. Ridder (2010): "The Asymptotic Variance of Semiparametric Estimators with Generated Regressors," forthcoming in *Econometrica*.

Hansen, B. (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory,* 24(3), 726–748.

Härdle, W., P. Hall, and H. Ichimura (1993): "Optimal Smoothing in Single-Index Models," *The Annals of Statistics*, 21(1), 157–178.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65(2), 261–294.

Heckman, J. J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.

Horowitz, J. L. and V. G. Spokoiny (2001): "An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model against a Nonparametric Alternative," *Econometrica*, 69(3), 599–631.

Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58(1-2), 71–120.

Ichimura, H., and L. Lee (1991): "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation"," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. A. Barnett, J. Powell, and G. Tauchen, pp. 3–49. Cambridge University Press.

Ichimura, H., and S. Lee (2010): "Characterization of the Asymptotic Distribution of Semiparametric M-Estimators," *Journal of Econometrics*, 159(2), 252–266.

Imbens, G., and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77(5), 1481–1512.

Imbens, G., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

Klein, R. and R. Spady (1993) "An efficient Semiparametric Estimator for Discrete Choice Models", *Econometrica*, 61(2), 387-421.

Lewbel, A., and O. B. Linton (2007): "Nonparametric Matching and Efficient Estimators of Homothetically Separable Functions," *Econometrica*, 75(4), 1209–1227.

Li, D. and Q. Li (2010): "Nonparametric/semiparametric Estimation and Testing of Econometric Models with Data Dependent Smoothing Parameters," *Journal of Econometrics*, 157(1), 179–190.

Li, Q. and J. M. Wooldridge (2002): "Semiparametric Estimation Of Partially Linear Models For Dependent Data With Generated Regressors," *Econometric Theory*, 18(3), 625–645.

Mammen, E., C. Rothe, and M. Schienle (2011a): "Nonparametric Regression with Nonparametrically Generated Covariates," Unpublished manuscript.

Mammen, E., C. Rothe, and M. Schienle (2011b): "Semiparametric Estimation with Generated Covariates," Unpublished manuscript.

Matzkin, R. L. (1992): "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models," *Econometrica*, 60(2), 239–270.

Meng, C.-L., and P. Schmidt (1985): "On the Cost of Partial Observability in the Bivariate Probit Model," *International Economic Review*, 26(1), 71–85.

Neumeyer, N., and Van Keilegom, I. (2010): "Estimating the Error Distribution in Nonparametric Multiple Regression with Applications to Model Testing," *Journal of Multivariate Analysis*, 101(5), 1067–1078.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–1382.

Newey, W. K. (2007): "Nonparametric Continuous/Discrete Choice Models," *International Economic Review*, 48(4), 1429–1439.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by D. McFadden, and R. F. Engle, vol. IV, pp. 2111–2245. Elsevier, North-Holland, Amsterdam.

Newey, W., J. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67(3), 565–603.

Nickl, R. and B. M. Pötscher (2007). "Bracketing Metric Entropy Rates and Empirical Central Limit Theorems for Function Classes of Besov- and Sobolev-Type," *Journal of Theoretical Probability*, 20(2), 177–199.

Olley, S. and A. Pakes (1996). "The Dynamics Of Productivity In The Telecommunications Equipment Industry". *Econometrica*, 64(6), 1263–1297.

Pagan, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25(1), 221–247.

Pinkse, J. (2001): "Nonparametric Regression Estimation using Weak Separability," Unpublished manuscript.

Robinson, P.M. (1988). "Root-n Consistent Semiparametric Regression". *Econometrica,* 56(4), 931–954.

Rothe, C. (2009): "Semiparametric Estimation of Binary Response Models with Endogenous Regressors," *Journal of Econometrics*, 153(1), 51–64.

Sperlich, S. (2009): "A Note on Non-parametric Estimation with Predicted Variables," *The Econometrics Journal*, 12(2), 382–395.

Song, K. (2008): "Uniform Convergence of Series Estimators over Function Spaces," *Econometric Theory*, 24(6), 1463–1499.

Sperlich, S. (2009): "A Note on Non-parametric Estimation with Predicted Variables," *The Econometrics Journal*, 12(2), 382–395.

Stock, J. H. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84(406), 567–575.

Talagrand, M. (1994): "Sharper bounds for Gaussian and empirical processes," *Annals of Probability*, 22(1), 28–76.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer Series in Statistics. Springer-Verlag, New York, 1 edn.

van de Ven, W. and B. van Praag (1981). "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection," *Journal of Econometrics*, 17(2), 229–252.