Instrumental Variables Methods for Recovering Continuous Linear Functionals

Andres Santos*

Department of Economics
University of California, San Diego

First Draft: March, 2008
This Draft: July, 2010

Abstract

This paper develops methods for estimating continuous linear functionals in nonparametric instrumental variables problems. Examples of such functionals include consumer surplus and weighted average derivatives. The estimation procedure is robust to a setting where the underlying model is not identified but the linear functional is. In order to attain such robustness, it is necessary to employ a partially identified nuisance parameter. We address this problem by consistently estimating a unique element of the identified set for nuisance parameters which we then use to construct a $\sqrt{n}$ asymptotically normal estimator for the desired linear functional.

**Keywords:** Instrumental variables, partial identification.

**JEL Codes:** C13, C14, C21.

*Address: 9500 Gilman Drive MS0508, La Jolla, CA 92093-0508. Tel: (858) 534-2407. Email: a2santos@ucsd.edu.*
1 Introduction

Numerous estimation problems in microeconometrics have encountered the challenge of endogenous regressors. The underlying structural relations often imply the estimated models do not fit the classical regression framework but are instead of the form:

\[ Y = m_0(X) + \epsilon \] (1)

where \( E[\epsilon|X] \neq 0 \) but \( E[\epsilon|Z] = 0 \) for some instrument \( Z \). The parametric analysis of \( m_0 \) through instrumental variables (IV) is well understood. Unfortunately, the extension of such procedures to a more robust nonparametric framework has encountered a number of difficulties. As discussed in Newey and Powell (2003), the nonparametric identification of \( m_0 \) requires the availability of an instrument satisfying far more stringent conditions than the usual covariance restrictions of the linear model. A lack of identification of \( m_0 \), however, does not preclude interesting characteristics of the model from being identified. Severini and Tripathi (2006, 2010), for example, argue that certain linear functionals of \( m_0 \) will be identified even when \( m_0 \) is not. In this paper, we develop methods for the \( \sqrt{n} \) estimation of such functionals without requiring \( m_0 \) to be identified.

For \( \mathcal{X} \) the support of \( X \) and \( \nu : \mathcal{X} \rightarrow \mathbb{R} \) a known function, we study functionals of the form:

\[ \langle \nu, m \rangle \equiv \int_{\mathcal{X}} \nu(x)m(x)dx, \] (2)

which include as special cases consumer surplus and weighted average derivatives. Severini and Tripathi (2010) show, under mild assumptions, that a necessary condition for \( \langle \nu, m_0 \rangle \) to be identified and estimable at a \( \sqrt{n} \) rate is the existence of a function \( \theta_0 \) of the instrument \( Z \) such that

\[ E[\theta_0(Z)|x] = \frac{\nu(x)}{f_X(x)} \] (3)

where \( f_X \) is the density of \( X \). Otherwise, Severini and Tripathi (2010) establish, the semiparametric efficiency bound for \( \langle \nu, m_0 \rangle \) is infinite. It is important to emphasize that condition (3) must hold for \( \langle \nu, m_0 \rangle \) to be \( \sqrt{n} \) estimable, regardless of whether \( m_0 \) is identified or not. A \( \sqrt{n} \)-asymptotically
normal estimator of \( \langle \nu, m_0 \rangle \) must therefore (i) assume either explicitly or implicitly that (3) holds or (ii) further restrict the model so the semiparametric efficiency bound is finite. For this reason, we base our estimator off the necessary condition for \( \sqrt{n} \) estimability in (3) and avoid unnecessary assumptions on the identification of \( m_0 \). As a result, our estimator will be robust to a possible lack of identification of the underlying model \( m_0 \).

The identification of \( \langle \nu, m_0 \rangle \) is guaranteed when (3) holds, since for any function \( m : X \to \mathbb{R} \) that agrees with the exogeneity assumption on the instrument we must have:

\[
\langle \nu, m \rangle = E[m(X)\theta_0(Z)] + E[(Y - m(X))\theta_0(Z)] = E[Y\theta_0(Z)].
\] (4)

Moreover, equation (4) suggests that a natural estimator for \( \langle \nu, m_0 \rangle \) is given by the sample analogue

\[
\hat{\langle \nu, m_0 \rangle} \equiv \frac{1}{n} \sum_{i=1}^{n} Y_i \hat{\theta}_0(Z_i),
\] (5)

where \( \hat{\theta}_0 \) is a consistent estimator of \( \theta_0 \). Unfortunately, in most applications the nuisance parameter \( \theta_0 \) will not be identified. Nonparametric identification requires the existence of a unique solution to the integral equation in (3). Equivalently, identification necessitates that there be no nonzero function \( \psi \) such that \( E[\psi(Z)|x] = 0 \). Intuitively, for \( \theta_0 \) to be identified the regressor \( X \) must be able to detect all forms of variation in the instrument \( Z \). This requirement is problematic, as in most instances instruments posses variation that is unrelated to the endogenous regressor. Hence, a general estimation procedure should be robust to the lack of identification of both \( m_0 \) and \( \theta_0 \).

The lack of identification of \( \theta_0 \) presents important technical challenges, but it does not hinder the identification or \( \sqrt{n} \) estimability of \( \langle \nu, m_0 \rangle \). Any solution to (3) provides a valid nuisance parameter for recovering \( \langle \nu, m_0 \rangle \) through (4). We therefore do not assume that \( \theta_0 \) is identified and instead obtain an estimator for a unique element of the set of solutions to (3). This set of solutions constitutes an identified set which we denote by:

\[
\Theta_0 \equiv \left\{ \theta \in \Theta : E[\theta(Z)|x] = \frac{\nu(x)}{f_X(x)} \right\},
\] (6)
where $\Theta$ is a nonparametric set of functions. We obtain a consistent estimator for a unique element of $\Theta_0$ in two steps. First, results in Chernozhukov et al. (2007) are generalized to arbitrary metric spaces to obtain a consistent estimator $\hat{\Theta}_0$ for $\Theta_0$. In addition, extending results in Ai and Chen (2003), it is possible to establish that $\hat{\Theta}_0$ converges to $\Theta_0$ at a $o_p(n^{-\frac{1}{4}})$ rate with respect to an appropriate Hausdorff pseudo-norm. As a second step, we recover a unique element $\theta_0 \in \Theta_0$ by carefully choosing a unique element $\hat{\theta}_0 \in \hat{\Theta}_0$. This procedure is analogous to a classical $M$-estimation problem where the domain $\Theta_0$ is unknown but instead estimated by $\hat{\Theta}_0$. These results provide a general useful technique for recovering nonparametric nuisance parameters when they are not identified and may be of independent interest.

This paper is highly complementary to previous work in Severini and Tripathi (2006, 2010). The authors are the first to explore conditions for the identification of $\langle \nu, m_0 \rangle$ when $m_0$ is not identified and to derive efficiency bounds for its estimation. They, however, provide no estimation procedures. Ai and Chen (2003, 2007) and Darolles et al. (2003) derive asymptotically normal estimators for $\langle \nu, m_0 \rangle$ that assume $m_0$ is identified. Within the larger nonparametric IV literature, Newey and Powell (2003) and Hall and Horowitz (2005) propose consistent estimators for $m_0$, while Horowitz (2007) and Gagliardini and Scaillet (2008) derive the asymptotic distribution for estimators of $m_0$. Santos (2007) proposes test statistics for inference when the model is partially identified. Ai and Chen (2003) and Blundell et al. (2007) examine the properties of a semiparametric specification. In related work, Newey et al. (1999), Chesher (2003, 2005, 2007), Imbens and Newey (2009) and Schennach et al. (2007) explore estimation and identification in triangular systems, while Chalak and White (2006) and White and Chalak (2006) study the identification of causal effects. This paper is also related to the vast partial identification literature that explores the limits of inference without identification. See Manski (1990, 2003) and references within.

The remainder of the paper is organized as follows. Section 2 provides examples of interesting choices for $\nu$ while Section 3 develops a consistent estimator for a unique element of the identified set.
In Section 4 we employ such estimator of the nuisance parameter to obtain an asymptotically normal estimator of $\langle \nu, m_0 \rangle$. The small sample performance of the estimator is analyzed in a Monte Carlo study in Section 5. Section 6 briefly concludes. All proofs are contained in the appendix.

2 Examples

The canonical example of a function $\nu$ for which the functional $\langle \nu, m_0 \rangle$ is of interest, is that of consumer surplus, as the following example illustrates:

**Example 2.1.** Let $q$ denote quantity and $m_0 : \mathbb{R}_+ \rightarrow \mathbb{R}$ be an inverse demand function. Suppose we are interested in estimating approximate consumer surplus at market clearing price and quantity $(p_c, q_c)$. Clearly, if we let $\nu(q) = 1\{0 \leq q \leq q_c\}$, we can immediately obtain

$$\int_0^{q_c} m_0(q) dq - p_c q_c = \langle \nu, m_0 \rangle - p_c q_c \quad (7)$$

and hence if $(p_c, q_c)$ are observable, then $\langle \nu, m_0 \rangle$ is the object of interest. This simple example can readily be extended to allow for an inverse demand function that depends on both $q$ and covariates $X$. If $X$ is discrete valued, then we may estimate $\int_0^{q_c} m_0(q, x_0) dq$ at some fixed point $x_0$. Otherwise, if $X$ is continuously distributed and $w : X \rightarrow \mathbb{R}$ is some specified weight function, then we can recover $\int_X \int_0^{q_c} m_0(q, x) w(x) dq dx$. ■

Proper selection of the function $\nu$ also allows us to estimate weighted average derivatives of $m_0$ by examining functionals of the form $\langle \nu, m_0 \rangle$.

**Example 2.2.** Let $m_0 : [0, 1] \rightarrow \mathbb{R}$ and $m_0^{(s)}(x)$ denote the $s^{th}$ derivative of $m_0$ evaluated at $x$. Suppose the object of interest is the weighted average derivative given by

$$\int_0^1 m_0^{(s)}(x) w(x) dx \quad (8)$$
for some specified weight function \( w \). If \( w \) is \( s \) times differentiable and in addition satisfies \( w^{(k)}(0) = w^{(k)}(1) = 0 \) for all \( 0 \leq k \leq (s - 1) \), then integration by parts yields:

\[
\int_0^1 m_0^{(s)}(x)w(x)dx = (-1)^{(s)} \int_0^1 m_0(x)w^{(s)}(x)dx .
\]  

(9)

Hence, the weighted average derivative in (8) may be expressed in the form \( \langle \nu, m_0 \rangle \) by letting \( \nu(x) = (-1)^{(s)}w^{(s)}(x) \).

As a final example, we show functionals of the form \( \langle \nu, m_0 \rangle \) may also be employed to recover the parametric component of a semiparametric specification.

**Example 2.3.** Let \( X = (X_1, X_2) \) with \( X_1 \in \mathbb{R}, X_2 \in \mathbb{R} \) and \( \mathcal{X} \equiv [0, 1]^2 \). Suppose we are interested in the semiparametric specification \( m_0(x_1, x_2) = x_1\beta_0 + r_0(x_2) \) for some constant \( \beta_0 \) and integrable function \( r_0(x_2) \). By selecting \( \nu(x_1, x_2) = 12(x_1 - \frac{1}{2}) \), we then obtain by direct calculation:

\[
\langle \nu, m_0 \rangle = \int_0^1 \int_0^1 12(x_1 - \frac{1}{2})(x_1\beta_0 + r_0(x_2))dx_1dx_2 = \beta_0 .
\]  

(10)

It is important to note that, unlike in Examples 2.1 and 2.3, the results in Severini and Tripathi (2010) do not apply as the parameter space has been restricted to be semiparametric. In particular, condition (3) is not a necessary condition for the \( \sqrt{n} \) estimability of \( \beta_0 \).

## 3 Nuisance Parameter

The principal challenge in constructing \( \langle \nu, m_0 \rangle \) consists in producing a consistent estimator \( \hat{\theta}_0 \) for a unique element \( \theta_0 \) of the identified set \( \Theta_0 \). Intuitively, if we could observe \( \Theta_0 \), then we would simply select a unique element \( \theta_0 \) from it and employ it to estimate \( \langle \nu, m_0 \rangle \). The set \( \Theta_0 \), however, is itself identified and can therefore be estimated. Hence, a natural estimation strategy is:

(S.1) Construct a consistent estimator \( \hat{\Theta}_0 \) for the set of functions \( \Theta_0 \).
(S.2) Select $\hat{\theta}_0 \in \hat{\Theta}_0$ in such a way as to ensure it converges to a unique element $\theta_0 \in \Theta_0$.

This section develops the theory necessary for showing such an approach is valid. In order to achieve (S.1), we generalize results in Chernozhukov et al. (2007) to arbitrary metric spaces and construct a consistent estimator for $\Theta_0$ as well as obtain its rate of convergence. To carry out (S.2), we adapt the theory of extremum estimators to problems where the parameter space is unknown but consistently estimated. In this way we are able to show that if $M: \Theta \to \mathbb{R}$ is a population criterion function attaining a unique minimum on $\Theta_0$, then the minimizer of a sample analogue $M_n: \Theta \to \mathbb{R}$ over the estimated set $\hat{\Theta}_0$ satisfies (S.2).

### 3.1 Criterion Functions

The identified set can be characterized as the set of minimizers to a criterion function, as in Chernozhukov et al. (2007) and Romano and Shaikh (2010). In particular, for

$$Q(\theta) \equiv E[(E[\nu(X) - \theta(Z)f_X(X)|X])^2],$$

the set $\Theta_0$, as defined in (6), agrees with the set of zeros of the function $Q: \Theta \to \mathbb{R}$:

$$\Theta_0 = \{\theta \in \Theta : Q(\theta) = 0\}.$$  \hspace{1cm} (12)

We require $\Theta$ to be a smooth set of functions, which both ensures consistency and the uniform behavior of the empirical process on the parameter space. In particular, we assume $\Theta$ is bounded in the Sobolev norm $\|\cdot\|_{\infty,\alpha}$. To define $\|\cdot\|_{\infty,\alpha}$, let $Z \in \mathbb{R}^{d_z}$ and $\lambda$ be a $d_z$ dimensional vector of nonnegative integers. Denote $|\lambda| = \sum_i^{d_z} \lambda_i$ and let $D^\lambda \theta(z) = \partial^{|\lambda|} \theta(z)/\partial z_1^{\lambda_1} \cdots \partial z_{d_z}^{\lambda_{d_z}}$. For $\alpha \in \mathbb{R}$, $\alpha$ the greatest integer smaller than $\alpha$ and $Z$ the support of $Z$, the norm $\|\cdot\|_{\infty,\alpha}$ is then given by:

$$\|\theta\|_{\infty,\alpha} \equiv \max_{|\lambda| \leq \alpha} \sup_{z \in Z} |D^\lambda \theta(z)| + \max_{|\lambda| = \alpha} \sup_{z \neq z'} \frac{|D^\lambda \theta(z) - D^\lambda \theta(z')|}{\|z - z'\|^{|\alpha - \alpha|}}.$$  \hspace{1cm} (13)
A function $\theta$ with $\|\theta\|_{\infty, \alpha} < \infty$ has partial derivatives up to order $\alpha$ uniformly bounded, and partial derivatives of order $\alpha$ Lipschitz of order $\alpha - \alpha$. If $\Theta$ is bounded in $\|\cdot\|_{\infty, \alpha}$, then these properties hold uniformly in $\theta \in \Theta$.

Estimation of $\Theta_0$ is equivalent to estimation of the zeros of the criterion function $Q$. In parametric models, Chernozhukov et al. (2007) show the latter can be accomplished by employing the approximate minimizers of a sample analogue $Q_n$. Adapting their results to a nonparametric framework, we let $\Theta_n$ be a sieve for $\Theta$ and define the estimator:

$$\hat{\Theta}_0 \equiv \{\theta \in \Theta_n : Q_n(\theta) \leq b_n/a_n\}$$

(14)

for $b_n/a_n \downarrow 0$ at an appropriate rate. The requirement $b_n/a_n \downarrow 0$ implies that in the limit $\hat{\Theta}_0$ only includes those $\theta_n \in \Theta_n$ that approximate elements of $\theta_0 \in \Theta_0$ well. On the other hand, if $b_n/a_n \downarrow 0$ slowly enough, then we can ensure $\hat{\Theta}_0$ includes all such $\theta_n$. The requirements on the rate at which $b_n/a_n \downarrow 0$ are of course different than in the parametric case.

The construction of an appropriate sample analogue $Q_n$ requires the use of nonparametric estimators for both conditional expectations and the density of $X$. We estimate conditional expectations using a standard series approach. Assume $X \in \mathbb{R}^{d_x}$ and let $\{p_j(\cdot)\}_{j=1}^{\infty}$ be a sequence of known approximating functions. We denote the vector of the first $k_n$ terms in the basis by $p^{k_n}(x) = (p_1(x), \ldots, p_{k_n}(x))$ and let $P = (p^{k_n}(X_1), \ldots, p^{k_n}(X_n))^\prime$. For an i.i.d. sample $\{W_i\}_{i=1}^n$ from a random variable $W \in \mathbb{R}$, the nonparametric estimator $\hat{E}[W|x]$ for $E[W|x]$ is then given by the linear regression of the vector $(W_1, \ldots, W_n)$ on the matrix $P$; that is:

$$\hat{E}[W|x] \equiv p^{k_n}(x)(P'P)^{-1}\sum_{i=1}^{n} p^{k_n}(X_i)W_i .$$

(15)

This series estimator is studied in Newey (1997), Huang (1998, 2003) and Ai and Chen (2003).

For the nonparametric estimator of $f_X$, we follow Hall and Horowitz (2005) and employ product generalized kernel estimators to allow for compact support of $X$. In particular, if $X \in \mathbb{R}^{d_x}$ and
\(X^{(k)}\) denotes the \(k^{th}\) coordinate of \(X\), then the estimator of \(f_X(X_i)\) is given by:

\[
\hat{f}_X(X_i) \equiv \frac{1}{(n-1)h^{d_x}} \sum_{j \neq i} \prod_{k=1}^{d_x} K_h \left( \frac{X^{(k)}_i - X^{(k)}_j}{h}, X^{(k)}_i \right).
\] (16)

The dependence of \(K_h(u,t)\) on \(t\) accommodates the use of boundary kernels for points \(t\) near the boundary of the support. For certain parameter values, it will also be necessary to resort to higher order kernels in order to attain the appropriate rates of convergence. The kernel \(K_h\) is of order \(\alpha\) if for all \(t\) on its domain and all \(h > 0\) the following holds:

\[
\int_{\mathbb{R}} K_h(u,t) du = 1 \quad \int_{\mathbb{R}} u^j K_h(u,t) = 0 \quad \text{if } 1 \leq j \leq \alpha .
\] (17)

The specific assumptions on the kernel \(K_h\) and the bandwidth \(h\) are stated in Section 3.3.

The criterion \(Q_n\) is the plug-in version of \(Q\) using the above estimators. Specifically, we let

\[
Q_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \hat{m}^2(X_i, \theta) \quad \hat{m}(X_i, \theta) \equiv \hat{E}[\nu(X) - \theta(Z) \hat{f}_X(X)|X_i] .
\] (18)

We show in the next section that for an appropriate choice of \(b_n/a_n\), the criterion function \(Q_n\) produces a consistent estimator \(\hat{\Theta}_0\) for the identified set \(\Theta_0\).

### 3.2 Set Consistency

Under regularity conditions to be introduced, it is possible to establish the consistency of \(\hat{\Theta}_0\) for \(\Theta_0\) under a variety of norms. We focus on the family of Hausdorff norms, which is defined by:

\[
d_H(\Theta_1, \Theta_2, || \cdot ||) \equiv \max \{ h(\Theta_1, \Theta_2), h(\Theta_2, \Theta_1) \} \quad h(\Theta_1, \Theta_2) \equiv \sup_{\theta_1 \in \Theta_1} \inf_{\theta_2 \in \Theta_2} ||\theta_1 - \theta_2|| .
\] (19)

Hence, \(\hat{\Theta}_0\) provides a consistent estimator for \(\Theta_0\) under the Hausdorff norm if both the maximal approximation error of \(\hat{\Theta}_0\) by \(\Theta_0\) and of \(\Theta_0\) by \(\hat{\Theta}_0\) converges to zero in probability. Unlike the parametric case, however, using different norms for the projections in (19) implies significantly different Hausdorff norms and rates of convergence.
As in most two stage estimation problems, the nuisance parameter must be estimated at a rate \( o_p(n^{-\frac{1}{4}}) \) under a suitable pseudo-norm in order for the second stage to be \( \sqrt{n} \) consistent. Ai and Chen (2003) show we can focus on the pseudo-norm \( \| \cdot \|_w \), which in the present context is given by:

\[
\| \theta \|_w \equiv \sqrt{E[(E[\theta(Z)|X])^2f_X^2(X)]}.
\]  

(20)

Interestingly, the identified set \( \Theta_0 \) is an equivalence class under the pseudo-norm \( \| \cdot \|_w \), since for any \( \theta_1, \theta_2 \in \Theta_0 \) we have \( \| \theta_1 - \theta_2 \|_w^2 = E[(\nu(X) - \nu(X))^2] = 0 \). Consequently, the result:

\[
d_H(\hat{\Theta}_0, \Theta_0, \| \cdot \|_w) = o_p(n^{-\frac{1}{4}})
\]  

(21)

suffices for showing that \( \| \hat{\theta}_0 - \theta_0 \|_w = o_p(n^{-\frac{1}{4}}) \) for any \( \theta_0 \in \Theta_0 \) and \( \hat{\theta}_0 \in \hat{\Theta}_0 \). This follows, as

\[
\| \hat{\theta}_0 - \theta_0 \|_w = \inf_{\theta \in \Theta_0} \| \hat{\theta}_0 - \theta \| \leq d_H(\hat{\Theta}_0, \Theta_0, \| \cdot \|_w).
\]  

(22)

Exploiting (22), we obtain a rate of convergence for \( \| \hat{\theta}_0 - \theta_0 \|_w \) by deriving one for \( d_H(\hat{\Theta}_0, \Theta_0, \| \cdot \|_w) \).

Convergence under the metric \( d_H(\cdot, \cdot, \| \cdot \|_w) \), however, is not sufficient for in turn consistently estimating a unique element \( \theta_0 \in \Theta_0 \). For this purpose, we require consistency of \( \hat{\Theta}_0 \) under a norm that is able to differentiate between elements in \( \Theta_0 \) (unlike the pseudo-metric \( \| \cdot \|_w \)). For \( \| \cdot \|_\infty \) the usual supremum norm \( \| \theta \|_\infty \equiv \sup_{z \in Z} |\theta(z)| \), we therefore additionally show that:

\[
d_H(\hat{\Theta}_0, \Theta_0, \| \cdot \|_\infty) = o_p(1).
\]  

(23)

We now state assumptions that are sufficient for showing \( \hat{\Theta}_0 \) is consistent under the norm \( d_H(\cdot, \cdot, \| \cdot \|_\infty) \) and obtaining a rate of convergence under the weaker pseudo-metric \( d_H(\cdot, \cdot, \| \cdot \|_w) \).

**Assumption 3.1.**  
(i) \{\( Y_i, X_i, Z_i \)\} is an i.i.d. sample from (1), with \( E[Y^2] < \infty \) and \( E[\epsilon|Z] = 0 \);  
(ii) \( X \in \mathbb{R}^{d_x} \) has support \([0,1]^{d_x}\) and density \( f_X \) bounded away from zero with \( \| f \|_{\infty,r} < \infty \) for \( r > \frac{3d_x}{2} \);  
(iii) \( Z \in \mathbb{R}^{d_z} \) has compact support \( Z \).

\(^1\)Here, \( \| f \|_{\infty,r} \) is applied to a function of \( x \), and is meant to be defined as in (13) with \( X \) in place of \( Z \).
Assumption 3.2. (i) The generalized kernel $K_h$ is of order at least $r > \frac{3d_z}{2}$; (ii) $K_h(u, t)$ is uniformly bounded on $h > 0, (u, t) \in [0, 1]^2$; (iii) For some compact interval $K$, $K_h(\cdot, t)$ is compactly supported on $[(t-1)/h, t/h] \cap K$ for all $t \in [0, 1]$; (iv) $nh^{3d_z} \not\to \infty$ and $h' = o(n^{-\frac{1}{2}})$; (v) For some $K_0(u, t)$ with $\int K_0(u, t)du = 1$, $\lim_{h \to 0} K_h(u, t + hu) = K_0(u, t)$ pointwise in all $(u, t)$ with $t \in (0, 1)$.

Assumption 3.3. (i) The eigenvalues of $E[p^{k_n}(X)p^{k_n}(X)]$ are bounded above and away from zero; (ii) For every $\theta \in \Theta$ there is a $\pi_n(\theta)$ with $\sup_{\theta, x} |f_X(x)E[\theta(Z)|x] - \pi_n(\theta)p^{k_n}(x)| = O(k_n^{-\frac{3}{2}})$ and $\tilde{\pi}_n(\theta)$ with $\sup_{\theta, x} |E[\theta(Z)|x] - \tilde{\pi}_n(\theta)p^{k_n}(x)| = O(k_n^{-\frac{3}{2}})$; (iii) $\xi_{2n}^2 k_n = o(n)$, for $\xi_{jn} = \sup_{|x|=j, |x|} \|D^\lambda p^{k_n}(x)\|$. 

Assumption 3.4. (i) For some $m > d_z/2$ we have $\sup_{\theta} \|\theta\|_{\infty, m} < \infty$, $\Theta_0 \neq \emptyset$, $\Theta_n \subseteq \Theta$ and both $\{\Theta_n\}$ and $\Theta$ closed; (ii) For every $\theta \in \Theta$ there is $\Pi_n \theta \in \Theta_n$ with $\sup_{\Theta} \|\theta - \Pi_n \theta\|_{\infty} = O(\delta_n)$.

Assumption 3.1 states the requirements on the distribution of $(Y, X, Z)$. It is worth pointing out that the only requirement imposed on $m_0$ is that $E[m_0^2(X)] < \infty$. The regressor and instrument $(X, Z)$ are assumed to have compact support. For notational convenience, we set the support of $X$ to be $[0, 1]^{d_z}$, which may require transforming original variables through a bounded strictly monotonic function. Assumptions 3.1(ii) and 3.2(i)-(v) imply $\hat{f}_X$ is pointwise consistent at an appropriate rate; see Remark 3.1 below for examples of valid kernels $K_h$. Notice that the requirement on the bandwidth $h$ in Assumption 3.2(v) is feasible as a result of $r > \frac{3d_z}{2}$. Assumptions 3.3(i)-(iii) are standard in the use of series estimators for conditional mean functions; see Huang (1998, 2003) and Remarks 3.2 and 3.3 below for primitive conditions that ensure these rate requirements hold. In Assumption 3.4(i), we impose that the parameter space $\Theta$ be bounded in the norm $\|\cdot\|_{\infty, m}$ for some $m > d_z/2$. This assumption has strong implications on the estimation of $\Theta_0$; see Remark 3.4 below. Finally, Assumption 3.4(ii) characterizes the bias introduced by employing a sieve $\Theta_n$ for $\Theta$.

Remark 3.1. The kernel $K_h$ may be set to be a regular order $r$ kernel for points $t$ away from $\{0, 1\}$ and a boundary kernel otherwise. For example, let the bounded kernels $K : \mathbb{R} \to \mathbb{R}$ and $L : \mathbb{R} \to \mathbb{R}$
be supported on \([-1,1]\) and \([0,1]\) respectively and satisfy \(\int_{-1}^{1} K(u)du = \int_{0}^{1} L(u)du = 1\) and
\[
\int_{-1}^{1} u^j K(u)du = 0 \quad \int_{0}^{1} u^j L(u)du = 0
\] (24)
for all \(1 \leq j \leq r\). The generalized kernel \(K_h(u,t)\) given by \(K(u)\) for all \(h \leq t \leq 1 - h\), by \(L(u)\) for all \(h > 1 - t\) and by \(L(-u)\) for all \(t < h\) then satisfies Assumption 3.2(ii)-(iii) and 3.2(v) with \(K_0(u,t) = K(u)\) for all \(t \in (0,1)\).

**Remark 3.2.** Assumptions 3.3(ii) and 3.4(ii) impose restrictions on approximation errors. These are easily controlled if the functions being approximated are sufficiently smooth. For example, for:

\[
\Lambda_\gamma^\gamma([0,1]^d_x) = \{ \psi : \mathcal{X} \to \mathbb{R}, \| \psi \|_{\infty,\gamma} \leq B \}
\] (25)

and \(\{p_j(\cdot)\}_{j=1}^{n}\) polynomials or tensor product univariate splines, the approximation error of \(\psi\) by functions of the form \(p^{k^*}\pi\) under \(\| \cdot \|_\infty\) is of the order \(O(k_n^{-\frac{d^*}{4}})\) uniformly on \(\psi \in \Lambda_\gamma^\gamma([0,1]^d_x)\).

Hence, Assumption 3.4(ii) is satisfied if both \(\sup_{\Theta} \|E[\theta(Z)|\cdot]\|_{\infty,\gamma}\) and \(\sup_{\Theta} \|f_X(\cdot)E[\theta(Z)|\cdot]\|_{\infty,\gamma}\) are finite; see Chen (2007) for further discussion.

**Remark 3.3.** Verifying Assumption 3.4(iii) requires us to understand the relationship between \(\xi_0n\) and \(k_n\), which may be sieve specific. When \(\{p_j(\cdot)\}_{j=1}^{\infty}\) are polynomials, for example, we have \(\xi_0n \lesssim k_n^{d^*}\), while if \(\{p_j(\cdot)\}_{j=1}^{\infty}\) are tensor product univariate splines, then \(\xi_0n \lesssim k_n^{d^*_2}\); see Newey (1997) and Huang (1998) for additional details.

**Remark 3.4.** Let the operator \(T(\theta) \equiv E[\theta(Z)|\cdot]\) have domain \(\Theta\), and notice that the equation
\[
T(\theta)(x) = \frac{\nu(x)}{f_X(x)}
\] (26)
implicitly defines \(\Theta_0\). When \(T\) is bijective, \(T^{-1}\) exists but is not continuous unless \(\Theta\) is restricted, for example, to a compact set. This is the role of Assumptions 3.1(iii) and 3.4(i) which together imply \(\Theta\) is compact under \(\| \cdot \|_\infty\). Alternatively, it is possible to circumvent the discontinuity of \(T^{-1}\) employing a regularization approach as in Darolles et al. (2003) or Chen and Pouzo (2008). The compactness
restriction is a special case of Chen and Pouzo (2008) with a possibly suboptimal choice of lagrange multiplier. While we have assumed compactness of $\Theta$, it would also be interesting to extend the aforementioned regularization approaches to the present setting where $T$ is not bijective. Estimation of $\Theta_0$ through such a method would face important challenges as $\Theta_0$ is likely to be unbounded if $\Theta$ is not restricted and hence $d_H(\hat{\Theta}_0, \Theta_0, \| \cdot \|_\infty)$ is potentially infinite. ■

Assumptions 3.1-3.4 allow us to show the consistency of $\hat{\Theta}_0$ and establish its rate of convergence.

**Theorem 3.1.** Let $a_n = O((k_n/n + k_n^{-\frac{2}{3}}) + (nh^{dx})^{-1} + h^{2r} + \delta_n^2)^{-1}$ and $b_n \to \infty$ with $b_n = o(a_n)$. If in addition Assumptions 3.1(i)-(ii), 3.2(i)-(iv), 3.3(i)-(iii) and 3.4(i)-(ii) hold, then it follows that $d_H(\hat{\Theta}_0, \Theta_0, \| \cdot \|_\infty) = o_p(1)$ and furthermore that $d_H(\hat{\Theta}_0, \Theta_0, \| \cdot \|_w) = O_p(\sqrt{b_n/a_n})$.

Theorem 3.1 establishes that the rate of convergence of $\hat{\Theta}_0$ to $\Theta_0$ under the weak norm is balanced by three natural terms: (i) The estimation error from the conditional expectation $(k_n/n + k_n^{-\frac{2}{3}})$, (ii) The estimation error from the density $(nh^{dx} + h^{-2r})$ and (iii) The approximation error from using a sieve for the parameter space $(\delta_n)$. Therefore, it is possible to establish (21) by imposing conditions on these bandwidths. Theorem 3.1 further establishes the consistency of $\hat{\Theta}_0$ to $\Theta_0$ as in (23), which enables us to consistently estimate a unique element from it.

We illustrate how to verify Assumptions 3.1-3.4 by applying Theorem 3.1 to Example 2.2.

**Proposition 3.1.** In Example 2.2, suppose $(X, Z) \in [0, 1]^2$ and further assume that:

(i) $f_X$ is bounded away from zero, $f_X'$ is Lipschitz and Assumption 3.1(i) holds.

(ii) $K_h$ is as in Remark 3.1 with order $r \geq 2$ and $h \asymp n^{-r_h}$ for $1/4 < r_h < 1/3$.

(iii) $|\frac{d}{dx} f_{Z|X}(z|x_1) - \frac{d}{dx} f_{Z|X}(z|x_2)| \leq G(z)|x_1 - x_2|$ with $\int_0^1 G(z)dz < \infty \forall x_1, x_2 \in [0, 1]$; $\{p_j(\cdot)\}_{j=1}^\infty$ are splines of order two with $k_n \asymp n^{r_x}$ for $1/6 < r_x < 1/3$ and Assumption 3.3(i) holds.

(iv) $\Theta = \{\theta : [0, 1] \to \mathbb{R}, \|\theta\|_\infty \leq B \text{ and } |\theta(z_1) - \theta(z_2)| \leq B|z_1 - z_2| \forall z_1, z_2 \in [0, 1]\}$ for some $B > 0$ and $\Theta_0 \neq \emptyset$. Further let $\{r_q(\cdot)\}_{q=1}^\infty$ be splines of order two and define the sieve $\Theta_n = \{\theta \in \Theta : \theta(z) = \sum_{q=1}^{q_n} \beta_q r_q(z)\}$ for $q_n \asymp n^{r_z}$ with $r_z > 1/3$. 

13
If $a_n \sim n^\alpha$ with $2/3 < \alpha < \max\{1-r_x, 4r_x, 1-r_h, 2r_z\}$ and $b_n \sim \log n$, it then follows that:

$$d_H(\hat{\Theta}_0, \Theta_0, \|\cdot\|) = o_p(1) \quad d_H(\hat{\Theta}_0, \Theta_0, \|\cdot\|_w) = o_p(n^{-\frac{1}{4}}).$$

(27)

### 3.3 Element Consistency

Given the estimator $\hat{\Theta}_0$ for $\Theta_0$, the second challenge consists in selecting an element $\hat{\theta}_0 \in \hat{\Theta}_0$ in such a way as to ensure it converges to a unique element $\theta_0 \in \Theta_0$. For this purpose, we adapt the theory of extremum estimators to problems where the parameter space is unknown but consistently estimated. Suppose $M : \Theta \rightarrow \mathbb{R}$ is a population criterion function attaining a unique minimum on $\Theta_0$ and $M_n : \Theta \rightarrow \mathbb{R}$ is its finite sample analogue. Intuitively, if $\theta_0$ is the unique minimizer of $M$ on $\Theta_0$, then the minimizer of $M_n$ over the estimated parameter space $\hat{\Theta}_0$ should provide a consistent estimator for $\theta_0$. In this manner, we can ensure $\hat{\theta}_0 \in \hat{\Theta}_0$ converges to a unique element of the identified set by selecting it to be the solution to an appropriate minimization problem.

Theorem 3.2 formalizes this argument.

**Theorem 3.2.** Let (i) $\Theta_0 \subseteq \Theta$ be closed with $\Theta$ compact and $M : \Theta \rightarrow \mathbb{R}$ have a unique minimum on $\Theta_0$ at $\theta_0$, (ii) $\hat{\Theta}_0 \subseteq \Theta$ satisfy $d_H(\hat{\Theta}_0, \Theta_0, \|\cdot\|) = o_p(1)$, (iii) $M_n : \Theta \rightarrow \mathbb{R}$ and $M : \Theta \rightarrow \mathbb{R}$ be continuous, (iv) $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| = o_p(1)$. If $\hat{\theta} \in \arg\min_{\theta \in \hat{\Theta}_0} M_n(\theta)$, then $\|\hat{\theta} - \theta_0\| = o_p(1).$

**Remark 3.5.** Even though $\theta_0$ is a minimum of $M : \Theta \rightarrow \mathbb{R}$ on $\Theta_0$, it is often not the minimum on the parameter space $\Theta$. In particular, since $\Theta_0$ may have no interior relative to $\Theta$, $\theta_0$ may lie in the boundary of $\Theta_0$ and not even be a local minimum of $M$ on $\Theta$. The requirement that $\hat{\Theta}_0$ converge to $\Theta_0$ under the Hausdorff norm, however, is sufficiently strong to overcome this difficulty. ■

The principal purpose of the criterion function $M$ is to help us construct a consistent estimator for a unique element $\theta_0 \in \Theta_0$. Any function $M$ that attains a unique minimum on $\Theta_0$ is a suitable

---

2Continuity, closeness and compactness in (i)-(iii) are with respect to the metric under which $d_H(\hat{\Theta}_0, \Theta_0, \|\cdot\|) \overset{p}{\rightarrow} 0$. 

14
choice for this goal. Fortunately, when the parameter space $\Theta$ is both compact and convex, the identified set $\Theta_0$ is itself compact and convex. As a result, it is straightforward to guarantee that $M$ has a unique minimizer by choosing it to be strictly convex on $\Theta$ or $\Theta_0$. In this manner, Theorems 3.1 and 3.2 can be used to construct an estimator $\hat{\theta}_0 \in \hat{\Theta}_0$ such that $\|\hat{\theta}_0 - \theta_0\|_\infty = o_p(1)$ for some unique $\theta_0 \in \Theta_0$. Furthermore, since $\Theta_0$ is an equivalence class under the pseudo-metric $\|\cdot\|_w$, Theorem 3.1 also yields a rate of convergence for $\|\hat{\theta}_0 - \theta_0\|_w$ by arguing as in (22).

**Corollary 3.1.** Let Assumptions 3.1(i)-(ii), 3.2(i)-(iv), 3.3(i)-(iii) and 3.4(i)-(ii) hold, $\Theta$ be convex, $M : \Theta \to \mathbb{R}$ strictly convex, $M_n : \Theta \to \mathbb{R}$ continuous and $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| = o_p(1)$. Then, for $\theta_0$ the minimizer of $M : \Theta \to \mathbb{R}$ on $\Theta_0$ and $\hat{\theta}_0$ a minimizer of $M_n : \Theta \to \mathbb{R}$ on $\hat{\Theta}_0$:

$$\|\hat{\theta}_0 - \theta_0\|_\infty = o_p(1) \quad \|\hat{\theta}_0 - \theta_0\|_w = O_p(\sqrt{b_n/a_n}).$$ (28)

Corollary 3.1 implies it is possible to construct an estimator $\hat{\theta}_0$ such that for some unique element $\theta_0 \in \Theta_0$, $\|\hat{\theta}_0 - \theta_0\|_\infty = o_p(1)$ and in addition $\|\hat{\theta}_0 - \theta_0\|_w = o_p(n^{-\frac{1}{4}})$. Both results are instrumental for showing the second stage estimator is both $\sqrt{n}$ consistent and asymptotically normal. We illustrate the implications of Corollary 3.1 by applying it to Example 2.2.

**Proposition 3.2.** Let all assumptions of Proposition 3.1 hold, and further define:

$$M_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \theta^2(Z_i) \quad M(\theta) \equiv E[\theta^2(Z)].$$ (29)

Then for $\theta_0$ the minimizer of $M$ on $\Theta_0$ and $\hat{\theta}_0$ a minimizer of $M_n$ on $\hat{\Theta}_0$, it follows that:

$$\|\hat{\theta}_0 - \theta_0\|_\infty = o_p(1) \quad \|\hat{\theta}_0 - \theta_0\|_w = o_p(n^{-\frac{1}{4}}).$$ (30)

**Remark 3.6.** Without further restrictions on the model, a lack of a solution to (3) implies $\langle \nu, m_0 \rangle$ is not $\sqrt{n}$-estimable, and possibly not identified either. In such instances we may still define:

$$\nu_p(x) \equiv E[\theta_p(Z)|x]f_X(x) \quad \theta_p \in \arg\min_{\theta \in \Theta} E[(\nu(X) - E[\theta(Z)|X])f_X(X)]^2,$$ (31)

\[\text{Notice this rules out setting } M = Q \text{ because } Q(\theta) = 0 \text{ for all } \theta \in \Theta_0.\]
and hence the function $\nu_P$ is the best approximation in mean squared error that can be obtained from functions of the form $E[\theta(Z)|.]f_X(\cdot)$. Unlike the parameter $\langle \nu, m_0 \rangle$, however, the approximation $\langle \nu_P, m_0 \rangle$ is guaranteed to be both identified and $\sqrt{n}$-estimable. Letting $\hat{\theta}$ denote the exact minimizer of $Q_n$ and $\hat{\nu}_P(x) = \hat{E}[\hat{\theta}(Z)|x]\hat{f}_X(x)$, we may define alternative criterion functions to be given by:

$$
\tilde{Q}_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[\hat{\nu}_P(X) - \theta(Z)\hat{f}_X(X)|X_i])^2
$$

$$
\tilde{Q}(\theta) \equiv E[(E[\nu_P(X) - \theta(Z)f_X(X)|X])^2].
$$

(32)

The results in this paper may be extended to show that a procedure employing $\tilde{Q}_n$ instead of $Q_n$ can deliver an estimator $\hat{\theta}_P$ satisfying $\|\hat{\theta}_P - \theta_P\|_w = o_p(n^{-\frac{1}{4}})$ and $\|\hat{\theta}_P - \theta_P\|_\infty = o_p(1)$ which can in turn be employed to construct a $\sqrt{n}$ asymptotically normal estimator for $\langle \nu_P, m_0 \rangle$.4

3.4 Computational Aspects

While in the development of the theory we have presented $\hat{\theta}_0$ as a two stage estimator, in practice its computation can be done in one simple step. In particular, $\hat{\theta}_0$ minimizes the objective function $M_n$ over the estimate set $\hat{\Theta}_0$, if and only if

$$
\hat{\theta}_0 \in \arg \min_{\theta \in \Theta_n} M_n(\theta) \quad \text{s.t.} \quad a_nQ_n(\theta) \leq b_n.
$$

(33)

The constraint function $Q_n$ is quadratic in $\theta$, while the objective function $M_n$ can be chosen to be convex leading to a tractable optimization problem. The only potential difficulty therefore lies in imposing the constraint that $\theta \in \Theta_n$.

A computationally simple choice for $\Theta_n$ are linear sieves. Let $\{r_q(\cdot)\}_{q=1}^{\infty}$ be a set of basis functions and $q_n$ be the number of terms included when the sample size is $n$. Further denoting the vector $r^{q_n}(z) = (r_1(z), \ldots, r_{q_n}(z))$, and letting $\beta \in \mathbb{R}^{q_n}$, a linear sieve is of the form:

$$
\Theta_n = \{\theta \in \Theta : \theta(z) = r^{\beta}(z)\}.
$$

(34)

4The exact minimizer $\hat{\theta}$ of $Q_n$ may not be employed to construct $\langle \nu_P, m_0 \rangle$ as it will satisfy $\|\hat{\theta} - \theta_P\|_w = o_p(n^{-\frac{1}{2}})$ but not necessarily $\|\hat{\theta} - \theta_P\|_\infty = o_P(1)$.
For this choice of $\Theta_n$, defining $\Theta = \{ \theta : \| \theta \|_{\infty,m} \leq B \}$, as in Proposition 3.1, may not be advisable as the constraint $\| r^{q_n} \beta \|_{\infty,m} \leq B$ can be highly nonlinear in $\beta$. In a similar problem, Newey and Powell (2003) instead suggest employing the norm:

$$
\| \theta \|_{2,m_0}^2 \equiv \sum_{|\lambda| \leq m_0} \int_Z [D^\lambda \theta(z)]^2 dz
$$

and letting $\Theta$ denote the closure of the set $\{ \theta : \| \theta \|_{2,m_0} \leq B \}$ under $\| \cdot \|_{\infty,m}$. If $Z$ is sufficiently regular and $m_0 > m + d_z/2$ then such a choice of $\Theta$ will be bounded in $\| \cdot \|_{\infty,m}$ as required by Assumption 3.4(i). Moreover, the constraint $r^{q_n} \beta \in \Theta_n$ is then quadratic in $\beta$, since for

$$
\Lambda_{q_n} \equiv \sum_{|\lambda| \leq m_0} \int_Z [D^\lambda r^{q_n}(z)D^\lambda r^{q_n}(z)] dz
$$

we have $r^{q_n} \beta \in \Theta_n$ if and only if $\beta' \Lambda_{q_n} \beta \leq B^2$. For these parameter specifications, (33) becomes:

$$
\hat{\beta} \in \arg \min_{\beta} M_n(r^{q_n} \beta) \quad \text{s.t.} \quad (a) \ a_n Q_n(r^{q_n} \beta) \leq b_n, \ (b) \ \beta' \Lambda_{q_n} \beta \leq B^2
$$

and $\hat{\theta} = r^{q_n} \hat{\beta}$. The optimization problem in (37) consists of minimizing a convex function of choice subject to two quadratic constraints, which can be computationally easily solved.

### 4 Asymptotic Normality

The results of Section 3 can be used to obtain an estimator $\hat{\theta}_0$ for a unique element $\theta_0$ of the identified set such that $\| \hat{\theta}_0 - \theta_0 \|_\infty = o_p(1)$ and $\| \hat{\theta}_0 - \theta_0 \|_w = o_p(n^{-\frac{1}{4}})$. In this section we show how such an estimator can be employed to construct a $\sqrt{n}$ asymptotically normal estimator for $\langle \nu, m_0 \rangle$.

We first need to introduce additional notation. Let $\overline{V}$ be the closure of the linear span of $\Theta$ under $\| \cdot \|_w$. The vector space $\overline{V}$ is a Hilbert Space with inner product given by:

$$
\langle \theta_1, \theta_2 \rangle_w \equiv E \left[ E[\theta_1(Z) | X] E[\theta_2(Z) | X] f_X^2(X) \right].
$$

The linear functional $\theta \mapsto E[Y \theta(Z)]$ is continuous under $\| \cdot \|_w$ and hence, by the Riesz Representation theorem, there exists a $\hat{v} \in \overline{V}$ such that for all $\theta \in \overline{V}$ we have:

$$
\langle \hat{v}, \theta \rangle_w = E[Y \theta(Z)].
$$
It is important to note that the Riesz representer is unique only up to equivalence classes in $\| \cdot \|_w$.

To be precise, we therefore denote this equivalence class of functions by:

$$\mathcal{V} \equiv \{ v \in \hat{V} : \langle v, \theta \rangle_w = E[Y \theta(Z)] \text{ for all } \theta \in \hat{V} \}.$$  \hspace{1cm} (40)

Since $E[\hat{v}_1(Z)|x] = E[\hat{v}_2(Z)|x]$ for all $\hat{v}_1, \hat{v}_2 \in \mathcal{V}$, we will generically write $E[\hat{v}(Z)|x]$ without referring to a specific $\hat{v} \in \mathcal{V}$.

We introduce the following assumption which will suffice for establishing asymptotic normality.

**Assumption 4.1.** (i) $\mathcal{V} \cap \Theta \neq \emptyset$ and $\| E[\hat{v}(Z)|\cdot] \nu(\cdot)f_X(\cdot) \|_{\infty,r} < \infty$; (ii) These rate requirements hold: $\xi_{0n} \times k_n = o(n^{-\frac{1}{2}})$, $k_n^{\frac{1}{2} - \frac{1}{2r}} = o(n^{-1})$, $k_n^3 = o(n)$, $\xi_{0n}^2 \times k_n^{\frac{1}{2} - \frac{1}{2r}} = o(1)$ and $\delta_n = o(n^{-\frac{1}{4}})$.

Assumption 4.1(i) imposes that there be at least one element of $\mathcal{V}$ also in the parameter space $\Theta$. Since $\mathcal{V} \subset \bar{V}$, the assumption requires that there be a $\hat{v}$ not only in the closure of the linear span of $\Theta$, but in $\Theta$ itself. The norm $\| \cdot \|_{\infty,r}$ in Assumption 4.1(i) corresponds to the one in Assumptions 3.1(ii), 3.2(i) and 3.2(iv). Finally, Assumption 4.1(ii) strengthens the rate requirements on the bandwidths and approximation errors. For illustrative purposes, we note that all the rate requirements of Assumption 4.1(ii) are satisfied in Example 2.2 by the conditions imposed in Proposition 3.1.

As a first step towards establishing normality, we obtain the following asymptotic expansion:

**Theorem 4.1.** Let $\theta_0 \in \Theta_0$ and assume $\hat{\theta}_0 \in \Theta_n$ is such that $\| \hat{\theta}_0 - \theta_0 \|_\infty = o_p(1)$ and in addition $\| \hat{\theta}_0 - \theta_0 \|_w = o_p(n^{-\frac{1}{2}})$. Further let $\hat{v} \in \mathcal{V} \cap \Theta$ and define the remainder term:

$$r_n(\hat{\theta}_0) \equiv \frac{1}{n} \sum_{i=1}^n \hat{E}[\Pi_n \hat{v}(Z)f_X(X)|X_i]\hat{m}(X_i, \hat{\theta}_0).$$  \hspace{1cm} (41)

If Assumptions 3.1(i)-(iii), 3.2(i)-(v), 3.3(i)-(iii), 3.4(i)-(ii) and 4.1(i)-(ii) hold, then:

$$\sqrt{n}\{\langle \nu, m_0 \rangle - \langle \nu, m_0 \rangle\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{Y_i \theta_0(Z_i) - E[\hat{v}(Z)|X_i]f_X^2(X_i)\theta_0(Z_i)\} - \sqrt{n}r_n(\hat{\theta}_0) + o_p(1).$$  \hspace{1cm} (42)

Direct calculation reveals that the remainder term $r_n(\hat{\theta}_0)$ in (41) is related to $Q_n(\hat{\theta}_0)$ through:

$$2r_n(\theta) = \frac{d}{d\tau} Q_n(\hat{\theta}_0 + \tau \Pi_n \hat{v}) \bigg|_{\tau=0}. \hspace{1cm} (43)$$
Hence, \( 2r_n(\hat{\theta}_0) \) is the pathwise derivative of \( Q_n \) from \( \hat{\theta}_0 \) in the direction of \( \Pi_n\tilde{v} \). If \( \hat{\theta}_0 \) is the exact minimizer of \( Q_n \), then \( r_n(\hat{\theta}_0) = 0 \). However, without identification the exact minimizer fails to converge to a unique element of \( \Theta_0 \) and the conditions of Theorem 4.1 are not satisfied. We instead employ the results of Section 3 to construct \( \hat{\theta}_0 \), in which case \( \hat{\theta}_0 \in \hat{\Theta}_0 \) implies \( \hat{\theta}_0 \) is a “near minimizer” of \( Q_n \) but not necessarily the exact minimizer. As a result, the term \( \sqrt{n}r_n(\hat{\theta}_0) \) may not be asymptotically negligible, a problem we address in the next subsection.

4.1 Remainder Correction

The only unobservable element of the remainder \( r_n(\hat{\theta}_0) \) is \( \Pi_n\tilde{v} \), the projection of \( \tilde{v} \) onto the sieve \( \Theta_n \). We construct an extremum estimator for \( \Pi_n\tilde{v} \) by defining the criterion function:

\[
C(\theta) \equiv E[(E[\theta(Z)|X])^2f_X^2(X)] - 2E[Y\theta(Z)] .
\] (44)

Since \( E[Y\theta(Z)] = \langle \tilde{v}, \theta \rangle_w \), it follows that \( C(\theta) = \|\theta - \tilde{v}\|_w^2 - \|\tilde{v}\|_w^2 \), and hence \( \tilde{v} \in \Theta \) is the unique minimizer of \( \theta \mapsto C(\theta) \) (up to equivalence class in \( \| \cdot \|_w \)). We denote the implied estimator by:

\[
\hat{v} \in \arg\min_{\theta \in \Theta_n} C_n(\theta) \quad C_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^n (\hat{E}[\theta(Z)\hat{f}_X(X)|X_i])^2 - \frac{2}{n} \sum_{i=1}^n Y_i\theta(Z_i) .
\] (45)

It is important to note that in order to accurately approximate the remainder term \( r_n(\hat{\theta}_0) \), we do not require consistency for \( \Pi_n\tilde{v} \) under \( \| \cdot \|_\infty \), but only under the weak pseudo-metric \( \| \cdot \|_w \). This is because \( \Pi_n\tilde{v} \) only enters \( r_n(\hat{\theta}_0) \) through the term \( \hat{E}[\Pi_n\tilde{v}(Z)\hat{f}_X(X)|X_i] \), which is asymptotically equivalent to \( E[\tilde{v}(Z)f_X(X)|X_i] \). In turn, \( E[\tilde{v}(Z)f_X(X)|X_i] \) takes the same values for all \( \tilde{v} \in \mathcal{V} \), which is an equivalence class under \( \| \cdot \|_w \).

Given the estimator \( \hat{v} \in \Theta_n \), we define our approximation to the remainder \( r_n(\hat{\theta}_0) \) to be:

\[
\hat{r}_n(\hat{\theta}_0) = \frac{1}{n} \sum_{i=1}^n \hat{E}[^{\hat{v}}(Z)\hat{f}_X(X)|X_i]^{\hat{m}}(X_i, \hat{\theta}_0) .
\] (46)

The following Lemma characterizes the rate of convergence of \( \hat{r}_n(\hat{\theta}_0) \) to \( r_n(\hat{\theta}_0) \) uniformly on \( \hat{\Theta}_0 \),
Lemma 4.1. If Assumptions 3.1(i)-(iii), 3.2(i)-(v), 3.3(i)-(iii), 3.4(i)-(ii) and 4.1(i)-(ii) hold, it then follows that the approximate remainder \( \hat{r}_n(\hat{\theta}_0) \) satisfies:

\[
\sup_{\hat{\theta}_0 \in \hat{\Theta}_0} |\hat{r}_n(\hat{\theta}_0) - r_n(\hat{\theta}_0)| = O_p\left\{ \left( \frac{\ell_n}{n} \right)^{3/4} + \frac{\sqrt{b_n}}{\sqrt{a_n}} \right\}.
\]

(47)

Lemma 4.1 implies that if \( \sqrt{b_n}/\sqrt{a_n} \searrow 0 \) sufficiently fast, then \( \sqrt{n}|\hat{r}_n(\hat{\theta}_0) - r_n(\hat{\theta}_0)| = o_p(1) \).

The rate conditions imposed in Assumption 4.1(ii) ensure that such a choice of \( \sqrt{b_n}/\sqrt{a_n} \) is indeed feasible. Furthermore, since Lemma 4.1 holds uniformly in \( \hat{\Theta}_0 \), the parameter choices are valid for all possible criterion functions \( M_n \). To conclude, we note that the conditions of Proposition 3.1 are sufficient for establishing the remainder term \( \hat{r}_n(\hat{\theta}_0) \) may be estimated sufficiently fast in the context of Example 2.2.

Proposition 4.1. Let all the Assumptions of Proposition 3.1 and Assumption 4.1(i) hold. Then,

\[
\sup_{\hat{\theta}_0 \in \hat{\Theta}_0} \sqrt{n}|\hat{r}_n(\hat{\theta}_0) - r_n(\hat{\theta}_0)| = o_p(1) .
\]

(48)

4.2 Main Result

Theorem 4.1 and Lemma 4.1 establish that it is possible to construct a \( \sqrt{n} \) asymptotically normal estimator for \( \langle \nu, m_0 \rangle \) employing a nuisance parameter \( \hat{\theta}_0 \) constructed as outlined in Section 3. We collect the implications of our analysis and present the main result of the paper:

Theorem 4.2. Let \( \hat{\Theta}_0 \) be as in (14) with \( a_n = O((\ell_n/n)^3 + k_n^{2\gamma} + (nh^{d_1})^{-1} + h^{2r} + \delta_n^2)^{-1}) \) and \( b_n \to \infty \) with \( n^3 \times b_n = o(a_n) \). Suppose \( \Theta \) is convex, \( M : \Theta \to \mathbb{R} \) is strictly convex on \( \hat{\Theta}_0 \) and \( M_n : \Theta \to \mathbb{R} \) is continuous and satisfies \( \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| = o_p(1) \). Further define the parameters:

\[
\hat{\theta}_0 \in \arg \min_{\theta \in \hat{\Theta}_0} M_n(\theta) \quad \theta_0 = \arg \min_{\theta \in \Theta} M(\theta) .
\]

(49)

If Assumptions 3.1(i)-(iii), 3.2(i)-(v), 3.3(i)-(iii), 3.4(i)-(ii), 4.1(i)-(ii) hold, then it follows that:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ Y_i \hat{\theta}_0(Z_i) + \hat{r}_n(\hat{\theta}_0) - \langle \nu, m_0 \rangle \} \xrightarrow{\mathcal{L}} N(0, \sigma^2)
\]

(50)

where \( \hat{r}_n(\hat{\theta}_0) \) is as in (46) and \( \sigma^2 = E[\hat{\theta}_0^2(Z)(Y - E[\hat{\nu}(Z)|X]f_X^2(X))] \).
In order to conduct inference on \( \langle \nu, m_0 \rangle \), we still require a consistent estimator for the asymptotic variance \( \sigma^2 \). Such an estimator is easily computed given our estimators \( \hat{\theta}_0 \), \( \hat{f}_X \) and \( \hat{\nu} \) by:
\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_0^2(Z_i)(Y_i - \hat{E}[\hat{\nu}(Z)\hat{f}_X^2(X)|X_i])^2.
\]
(51)

Lemma 4.2 establishes \( \hat{\sigma}^2 \) is indeed consistent for the asymptotic variance of Theorem 4.2.

**Lemma 4.2.** If Assumptions 3.1(i)-(iii), 3.2(i)-(v), 3.3(i)-(iii), 3.4(i)-(ii) and 4.1(i)-(ii) hold, it then follows that the estimator \( \hat{\sigma}^2 \) is consistent for \( \sigma^2 \).

Theorem 4.2 and Lemma 4.2 enable us to construct pivotal statistics and in this way conduct hypotheses tests on the parameter of interest \( \langle \nu, m_0 \rangle \). We conclude by showing this can readily be accomplished in Example 2.2.

**Proposition 4.2.** Let all the Assumptions of Proposition 4.1 hold and in addition define \( M_n \) as in Proposition 3.2 and let \( \hat{\theta}_0 \) be the minimizer of \( M_n \) on \( \hat{\Theta}_0 \). It then follows that:
\[
\frac{1}{\sqrt{n} \hat{\sigma}} \sum_{i=1}^{n} \{Y_i\hat{\theta}_0(Z_i) + \hat{r}_n(\hat{\theta}_0) - \langle \nu, m_0 \rangle\} \overset{\mathcal{L}}{\to} N(0,1).
\]
(52)

**Remark 4.1.** Under exogeneity, \( \theta_0(z) = \nu(x)/f_X(x) \) and similarly \( E[\hat{\nu}(Z)|x]f_X^2(x) = m_0(x) \). It is then interesting to note that the asymptotic variance from Theorem 4.2 reduces to
\[
E \left[ \frac{\nu^2(X)}{f_X^2(X)} (Y - m_0(X)) \right]^2,
\]
(53)
which coincides with the asymptotic variance of Newey and McFadden (1994). Since the latter is semiparametrically efficient, so is our estimator in the case of exogeneity.\(^5\)

5 Monte Carlo

In order to illustrate the implementation of the outlined procedure and examine its finite sample performance, we conduct a small-scale Monte Carlo study. To facilitate exposition, the Monte

\(^5\)We thank an anonymous referee for pointing this out.
Carlo was designed so that the integral equation in (3) has a closed form solution. We assume 
\( X, Z \in [0, 1]^2 \), and that they are distributed according to the density:
\[
f_{XZ}(x, z) = 3|x - z| \text{ for } (x, z) \in [0, 1]^2.
\]
(54)
We complete the model by assuming that \( Y \) and \( \epsilon \) are generated according to the relationship:
\[
Y = 2 - X^2 + \epsilon \quad \quad \epsilon = -\frac{U}{12} \left( \frac{1}{f_{X|Z}(X|Z)} - 1 \right)
\]
(55)
where \( U \) is uniformly distributed on \([0, 1]\) independent of \((X, Z)\) and \( f_{X|Z}(x|z) \) is the conditional density of \( X \) given \( Z \).\(^6\) By construction, \( E[\epsilon|Z] = 0 \) and \( E[\epsilon|x] = (1 - f_X(x))/24f_X(x) \).

For the linear functional, we study an approximation to the integral of \( m_0 \) between two points \( x_L < x_U \) satisfying \( 0 < x_L < x_U < 1 \). Let \( \Phi \) denote the c.d.f. of the standard normal and define
\[
F_t(x) \equiv \Phi\left(\frac{x - x_L}{t}\right) - \Phi\left(\frac{x - x_U}{t}\right),
\]
(56)
as well as the constants \( A_t \equiv -\frac{1}{2}(F'_t(1) + F'_t(0)) \) and \( B_t \equiv \frac{1}{2}(F'_t(1) - F_t(1) - F_t(0)) \) for some \( t > 0 \).
The functional of interest is then \( \langle \nu_t, m_0 \rangle \), where \( \nu_t \) is pointwise given by:
\[
\nu_t(x) \equiv F_t(x) + A_t x + B_t.
\]
(57)
It can be verified through direct calculation and the dominated convergence theorem that:
\[
\lim_{t \downarrow 0} \langle \nu_t, m_0 \rangle = \int_{x_L}^{x_U} m_0(x) dx
\]
(58)
whenever \( \int_0^1 m_0^2(x) dx < \infty \). Unlike the choice \( \nu(x) = 1\{x_L \leq x \leq x_U\} \) for which no solution to (3) exists, however, Polyanin and Manzhirov (1998) establish that for any \( t > 0 \) the function
\[
\theta_0(z) \equiv \frac{1}{t^2}(\phi'\left(\frac{z - x_L}{t}\right) - \phi'\left(\frac{z - x_U}{t}\right))
\]
(59)
satisfies \( E[\theta_0(Z)|x] = \nu_t(x)/f_X(x) \) (where \( \phi(u) = \frac{d}{du}\Phi(u) \)). For the simulations we set \( x_L = 0.25, x_U = 0.75 \) and \( t = 0.01 \).
\(^6\)\( U\left(\frac{1}{f_{X|Z}(x|z)} - 1\right) \) has significantly fat tails. The scaling by \( 1/12 \) ensures that \( \epsilon \) retains variability while preventing most its large realizations from driving the results.
Following the discussion in Section 3.4, we define the parameter space $\Theta$ to be:

$$\Theta \equiv \text{cl}\{\theta : \|\theta\|_{2,2} \leq B\}$$

(60)

where the closure is under $\|\cdot\|_{\infty, 1}$ and $\|\cdot\|_{2,2}$ is as defined in (35). We employ a linear sieve $\Theta_n$ as in (34) with $\{r_q(\cdot)\}_{q=1}^n$ splines of order 4 with equally spaced knots in the interior and full multiplicity at the endpoints. Similarly, for estimating the conditional expectation we let $\{p_j(\cdot)\}_{j=1}^\infty$ be splines of order 2 with the same knot arrangement. The kernel $K_h$ for estimating the density of $X$ is constructed as in Remark 3.1 setting $K(u) = 0.75(1-u^2)$ for all $|u| \leq 1$ and $L(u) = 6(u-u^2)(6-10u)$ for all $u \in [0,1]$ which Müller (1991) shows corresponds to the generalized Epanechnikov kernel. Finally, we also define the auxiliary criterion functions $M_n$ and $M$ to be given by:

$$M_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^n \theta^2(Z_i) \quad \text{and} \quad M(\theta) \equiv E[\theta^2(Z)].$$

(61)

Arguing as in Proposition 3.1, it is possible to show all rate requirements are satisfied by letting $a_n/b_n$, $h$, $k_n$ and $q_n$ satisfy the assumptions of said proposition. The theory, however, offers little guidance as to how to select the level of the different bandwidths. Since the minimizer of $M$ on $\Theta_0$ is not the global minimizer of $M$, it seems prudent to select $a_n/b_n$ so that the constraint (a) in (37) binds. For $\hat{\theta}_u$ the minimizer of $Q_n$ on $\Theta_n$, we implement the following ad-hoc selection of $a_n/b_n$:

$$\frac{a_n}{b_n} \equiv \gamma \times Q_n(\hat{\theta}_u) + (1 - \gamma) \times Q_n(0)$$

(62)

for $\gamma \in (0,1)$. Since $\theta = 0$ is the minimizer of $M_n$ on $\Theta_n$, setting $\gamma > 0$ ensures that constraint (a) in (37) is binding. In turn, selecting $\gamma > 0$ implies $\hat{\Theta}_0$ contains elements in addition to $\hat{\theta}_u$. Tables 1 and 2 report the bias and standard deviation respectively of the estimator $\langle \hat{\nu}_t, m_0 \rangle + \hat{r}_n(\hat{\theta}_0)$ for different choices of $\gamma$ (as in (62)), $B$ (as in (60)) and $h$. The bandwidths $k_n$ and $q_n$ were both set equal to six. Other choices for these parameters yielded qualitative similar results. The numbers reported in Tables 1 and 2 are based on one thousand replications.
Higher choices of bandwidth $h$ decrease the standard deviation of the estimator and, somewhat surprisingly, do not translate into a higher bias for the ranges of $h$ selected. Conversely, selecting large values of $B$ significantly decreases the bias of the estimator without an important increment in the standard deviation. For the range specified, the estimator does not appear to be very sensitive to the choice of $\gamma$, with both its bias and standard deviation remaining largely invariant to the level of $\gamma$ across specifications. Overall, we find the performance of the estimator to be encouraging.

6 Conclusion

The results in this paper allow for the estimation of continuous linear functionals in an additive separable model with endogenous regressors. Our proposed estimator relies on an assumption that under weak conditions is necessary for $\langle \nu, m_0 \rangle$ to be $\sqrt{n}$ estimable. This procedure does not require $m_0$ nor the nuisance parameter to be identified. The techniques developed in this paper may be of use in other estimation problems with partially identified nonparametric nuisance parameters.

7 Acknowledgements

I would like to thank Peter Robinson and three anonymous referees for comments and suggestions that greatly improved the paper. I also benefited from helpful comments by Graham Elliott, Aprrajit Mahajan, Azeem Shaikh, Hal White and Frank Wolak as well as seminar participants at UC Berkeley, UC Davis, UC Riverside, Ohio State University, the Cowles Conference on Operator Methods and Inverse Problems and the 2008 Econometric Society Winter Meetings,
APPENDIX A - Notation and Definitions

The following is a table of the notation and definitions that will be used throughout the appendix, including many that go beyond the ones already introduced in the main text:

\[
\begin{align*}
p &
\leq b \\
\| \cdot \|_{m} &
\text{The norm } \|g\|_{m} = \max_{|\lambda| \leq m} \sup_{w} |D^{\lambda}g(w)| + \max_{|\lambda|=m} \sup_{w,w'} |D^{\lambda}g(w) - D^{\lambda}g(w')| / w-w' \\
N(\epsilon,F,\| \cdot \|) &
\text{The covering numbers of size } \epsilon \text{ for } F \text{ under the norm } \| \cdot \|
\end{align*}
\]

APPENDIX B - Proofs of Results

**Theorem 1.** Assume (i) \(Q(\theta) \geq 0\) and \(\Theta_0 = \{ \theta \in \Theta : Q(\theta) = 0 \}\) with \(\Theta\) compact in \(\| \cdot \|\), (ii) \(\Theta_n \subseteq \Theta\) are closed and \(\sup_{\Theta_n} \inf_{\Theta_0} \| \theta - \theta_n \| = O(c_{1n})\), (iii) Uniformly on \(\Theta_n\), \(Q_n(\theta) \leq C_1 Q(\theta) + O_p(c_{2n})\) and \(Q(\theta) \leq C_1 Q_n(\theta) + O_p(c_{2n})\) with probability approaching one, (iv) \(Q(\theta) \leq C_2 \inf_{\Theta_0} \| \theta - \theta_0 \|^{\kappa_1}\) for some \(\kappa_1 > 0\). Then for \(a_n = O(\max[c_{1n}, c_{2n}]^{-1})\) and \(b_n \rightarrow \infty\) with \(b_n = o(a_n)\), the set \(\hat{\Theta}_0 = \{ \theta_n \in \Theta_n : Q_n(\theta) \leq b_n/a_n \} \) satisfies \(d_H(\hat{\Theta}_0, \Theta_0, \| \cdot \|) = O_p(1)\). If in addition (v) \(Q(\theta) \geq \inf_{\Theta_0} C_2 \| \theta - \theta_0 \|^{\kappa_2}\) for some \(\kappa_2 > 0\), then \(d_H(\hat{\Theta}_0, \Theta_0, \| \cdot \|) = O_p(\max\{ b_n/a_n \}^{1/\max[\kappa_1, \kappa_2]} c_{1n})\).

**Proof:** Define \(\Theta_{p_n}\) to be the pointwise projection of \(\Theta_0\) onto \(\Theta_n\) under \(\| \cdot \|\). We then obtain:

\[
\sup_{\Theta_{p_n}} Q_n(\theta) \leq \sup_{\Theta_{p_n}} C_1 Q(\theta) + O_p(c_{2n}) \leq \sup_{\Theta_{p_n}} C_1 C_2 \| \theta - \theta_0 \|^{\kappa_1} + O_p(c_{2n}) = O_p(a_n^{-1})
\]

with probability approaching one by (ii), (iii), (iv) and \(a_n = O(\max[c_{1n}, c_{2n}]^{-1})\). Next observe that we also have:

\[
h(\Theta_0, \hat{\Theta}_0) = \sup_{\Theta_0} \inf_{\Theta_0} \| \theta_0 - \hat{\theta}_0 \| \leq \sup_{\Theta_0} \inf_{\Theta_0, \Theta_{p_n}} \{ \| \theta_0 - \theta_{p_n} \| + \| \theta_{p_n} - \hat{\theta}_0 \| \}
\]

\[
\leq \sup_{\Theta_0} \inf_{\Theta_0, \Theta_{p_n}} \| \theta_0 - \theta_{p_n} \| + \sup_{\Theta_0, \Theta_{p_n}} \inf_{\Theta_0} \| \theta_{p_n} - \hat{\theta}_0 \| = \sup_{\Theta_0, \Theta_{p_n}} \inf_{\Theta_0} \| \theta_{p_n} - \hat{\theta}_0 \| + O(c_{1n}) .
\]

Therefore, for \(M > 0\) sufficiently large so that the \(O(c_{1n})\) term in (64) is lost \(Mc_{1n}/2\) we obtain that:

\[
P(h(\Theta_0, \hat{\Theta}_0) < Mc_{1n}) \geq P(\sup_{\Theta_{p_n}} \inf_{\Theta_0, \Theta_{p_n}} \| \theta_{p_n} - \hat{\theta}_0 \| < Mc_{1n}/2) \geq P(\Theta_{p_n} \subseteq \hat{\Theta}_0) .
\]

By definition of \(\hat{\Theta}_0\), \(\Theta_{p_n} \subseteq \hat{\Theta}_0\) if and only if \(Q_n(\theta_{p_n}) \leq b_n/a_n\) for all \(\theta_{p_n} \in \Theta_{p_n}\). Hence, for \(n\) large enough,

\[
P(h(\Theta_0, \hat{\Theta}_0) < Mc_{1n}) \geq P(a_n \sup_{\Theta_{p_n}} Q_n(\theta) \leq b_n) \rightarrow 1
\]

since by (63) \(a_n \sup_{\Theta_{p_n}} Q_n(\theta) = O_p(1)\) and \(b_n \rightarrow \infty\). We therefore obtain that \(h(\Theta_0, \hat{\Theta}_0) = O_p(c_{1n})\).
Next, we examine $h(\hat{\Theta}_0, \Theta_0)$. For $\epsilon_n > 0$, the continuity of $Q(\theta)$ and compactness of $\Theta$ imply that:

$$\delta_n \equiv \inf_{\{\theta \in \Theta : \inf_{a_n} ||\theta-a_n|| > \epsilon_n\}} Q(\theta) > 0.$$  \hspace{1cm} (67)

If $\Theta_0 = 0$, then $h(\hat{\Theta}_0, \Theta_0) = 0$. Therefore, it follows from (67), the definition of $\hat{\Theta}_0$ and (iii) that:

$$P(h(\hat{\Theta}_0, \Theta_0) > \epsilon_n) \leq P(\exists \theta \in \Theta_n : Q_n(\theta) \leq b_n/a_n \text{ and } Q(\theta) > \delta_n) \leq P(\delta_n < c_1b_n/a_n + O_p(c_2n)) + o(1).$$  \hspace{1cm} (68)

If $\epsilon_n$ is constant, then conditions (i)-(iv) and (68) imply $h(\hat{\Theta}_0, \Theta_0) = o_p(1)$ which together with $h(\hat{\Theta}_0, \Theta_0) = O_p(c_1n)$ establishes consistency. If in addition (v) holds, then set $\epsilon_n = (2C_1b_n/a_nC_2)^{-\frac{3}{n} - \frac{1}{2}}$ which implies $\delta_n \geq C_2\epsilon_n^2 \geq 2C_1b_n/a_n$ for $n$ large, and (68) implies $h(\hat{\Theta}_0, \Theta_0) = O_p((b_n/a_n)^{-\frac{3}{n} - \frac{1}{2}})$). The claimed rate of convergence then follows from $h(\hat{\Theta}_0, \Theta_0) = O_p(c_1n)$ and the definition of $d_H(\hat{\Theta}_0, \Theta_0, \|\cdot\|)$.  \hspace{1cm} ■

**Lemma 1.** If Assumptions 3.1(i)-(ii), 3.2(i)-(iii) hold, then almost surely (i) $E[\hat{f}_X(X)] = f_X(X)$ and $\text{Var}(\hat{f}_X(X)) = O((nh^{d_x})^{-1})$; (ii) $E[(\hat{f}_X(X) - f_X(X))^2] = O(h^{2r} + (nh^{d_x})^{-1})$.

**Proof:** The proof proceeds by standard calculations. Let $u = (X_i - x)/h$ and $u^{(k)}$ its $k^{th}$ component. Defining the set $\Gamma(X_i) = \{u \in \mathbb{R}^{d_x} : (X_i^{(k)} - 1)/h \leq u^{(k)} \leq (X_i^{(k)}/h\}$, we then obtain through a change of variables that:

$$E[\hat{f}_X(X_i) | X_i] = \frac{1}{h^{d_x}} \int_{[0,1]^{d_x}} \prod_{k=1}^{d_x} K_h\left(\frac{X_i^{(k)} - x^{(k)}}{h}\right) f_X(x) dx = \int_{\Gamma(X_i)} \prod_{k=1}^{d_x} K_h(u^{(k)}, X_i^{(k)}) f_X(X_i - hu) du$$

$$= f_X(X_i) + h \int_{\Gamma(X_i)} \sum_{|\lambda| = 1} \left\{ D^{\lambda} f_X(X_i(u)) - D^{\lambda} f_X(X_i) \right\} \prod_{k=1}^{d_x} (u^{(k)})^{\lambda_k} K_h(u^{(k)}, X_i^{(k)}) du = f_X(X_i) + O(h^r),$$  \hspace{1cm} (69)

where the third equality holds for some $X_i(u)$ a convex combination of $X_i$ and $X_i - hu$ by a pointwise Taylor expansion, $K_h$ being of order $r$ and the support of $\prod_{k} K_h(\cdot, X_i^{(k)}) \subset \Gamma(X_i)$ by Assumption 3.2(iii). That the final result in (69) holds almost surely in $X_i$ is the result of $\|X_i(u) - X_i\| \leq h\|u\|$ for all $u$, $\|u\|_{r,\infty} < \infty$, the support of $K_h(\cdot, X_i)$ being contained in a compact interval $\mathcal{K}$ by Assumption 3.2(iii) and $K_h(u, t)$ being uniformly bounded.

By similar arguments and exploiting that $K_h$ and $f_X$ are uniformly bounded and $K_h(\cdot, X_i)$ is compactly supported

$$\text{Var}(\hat{f}_X(X_i)|X_i) \leq \frac{1}{nh^{2d_x}} \int_{[0,1]^{d_x}} \prod_{k=1}^{d_x} K_h^2\left(\frac{X_i^{(k)} - x^{(k)}}{h}\right) f_X(x) dx$$

$$= \frac{1}{nh^{d_x}} \int_{\Gamma(X_i)} \prod_{k=1}^{d_x} K_h^2(u^{(k)}, X_i^{(k)}) f_X(X_i - hu) du = O((nh^{d_x})^{-1})$$  \hspace{1cm} (70)

and hence the first claim of the Lemma follows from (69) and (70). The second claim of the Lemma is in turn a direct consequence of the first claim of the Lemma holding almost surely in $X_i$.  \hspace{1cm} ■

**Lemma 2.** For $w : [0,1]^{d_x} \times Z \to \mathbb{R}$, let $E[w(X, Z) | x] = p_k^\pi(x)(P^\pi P)^{-1} \sum_i p_k^\pi(X_i) E[w(X, Z) | X_i]$. If Assumptions 3.1(i)-(iii), 3.2(i)-(iii), 3.3(i)-(iii) and 3.4(i) hold, then (i) $\sup_{\Theta_n} n^{-1} \sum_i (\hat{E}[\hat{f}_X(X) \theta(Z)] - E[f_X(X) \theta(Z)] | X_i)^2 = O_p\left(k_n^{\frac{3}{2}d_x} + (nh^{d_x})^{-1} + h^{2r}\right)$; and (ii) $\sup_{\Theta_n} E[|E[w(X) - f_X(X) \theta(Z)] | X] - E[w(X) - f_X(X) \theta(Z)] | X_i|^2 = O_p(k_n^{\frac{3}{2}d_x})$.  

26
Proof: As noted in Newey (1997), under Assumption 3.3(i) we may assume without loss of generality that $E[p^{k_n}(X)p^{k_n}(X)] = I$. Let $G_n$ denote the linear span of $p^{k_n}$, and observe that for any $g(x) = \sum_{j=1}^{k_n} a_j p_j(x)$:

$$\|g\|_\infty = \sup_x \left| \frac{\sum_{j=1}^{k_n} a_j p_j(x)}{\sum_{j=1}^{k_n} a_j^2} \right| \leq \sup_x \|p^{k_n}(x)\|.$$  \hfill (71)

Hence, for $A_n = \sup_{G_n} \|g\|_\infty / \|g\|_{L^2}$ Assumption 3.3(iii) implies $A_n^2 k_n/n \to 0$ and by Lemma 2.3(i) in Huang (2003):

$$\frac{1}{2} E[g^2(X)] \leq \frac{1}{n} \sum_{i=1}^{n} g^2(X_i) \leq 2E[g^2(X)] \quad \forall g \in G_n$$  \hfill (72)

with probability tending to one. In order to establish the first claim of the Lemma, we use the inequality:

$$\sup_{\Theta_n} \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[f_X(X)\theta(Z)|X_i] - E[f_X(X)\theta(Z)|X_i])^2$$

$$\leq \sup_{\Theta_n} \frac{2}{n} \sum_{i=1}^{n} (\hat{E}[f_X(X)\theta(Z)|X_i] - E[f_X(X)\theta(Z)|X_i])^2 + \sup_{\Theta_n} \frac{2}{n} \sum_{i=1}^{n} (\hat{E}[\theta(Z)(f_X(X) - \hat{f}_X(X))|X_i])^2.$$  \hfill (73)

Let $\epsilon_i(\theta) = f_X(X_i)\theta(Z_i) - E[f_X(X)\theta(Z)|X_i]$ and $\epsilon(\theta) = (\epsilon_1(\theta), \ldots, \epsilon_n(\theta))'$. It then follows that:

$$\sup_{\Theta_n} \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[f_X(X)\theta(Z)|X_i] - \hat{E}[f_X(X)\theta(Z)|X_i])^2 \leq 2\sup_{\Theta_n} E[p^{k_n}(X)(P'P)^{-1}P'\epsilon(\theta)\epsilon'(\theta)P(P'P)^{-1}p^{k_n}(X)]$$

$$\leq 2\sup_{\Theta_n} \|f_X(\theta)\|_\infty^2 \times \text{trace}\{(P'P)^{-1}P'P(P'P)^{-1}\}$$  \hfill (74)

where the first inequality holds by (72) with probability tending to one, and the second by $E[p^{k_n}(X)p^{k_n}(X)] = I$, the i.i.d. assumption, $E[\epsilon_i^2(\theta)] \leq \|f_X(\theta)\|_\infty^2$ and the law of conditional expectations. Let $\gamma_n$ be the largest eigenvalue of $n(P'P)^{-1}$ and note $\text{trace}\{(P'P)^{-1}\} \leq k_n\gamma_n$. It is shown in Theorem 1 in Newey (1997) that $\gamma_n = O_p(1)$ and hence, $\sup_{\Theta_n} \|f_X(\theta)\|_\infty^2 < \infty$ by Assumptions 3.1(ii) and 3.4(i), and result (74) then imply:

$$\sup_{\Theta_n} \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[f_X(X)\theta(Z)|X_i] - E[f_X(X)\theta(Z)|X_i])^2 = O_p(k_n \gamma_n).$$  \hfill (75)

In addition, it follows from Assumptions 3.1(i)-(ii), 3.3(i)-(ii) and Lemma A.1(B) in Ai and Chen (2003) that:

$$\sup_{\Theta_n} \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[f_X(X)\theta(Z)|X_i] - E[f_X(X)\theta(Z)|X_i])^2 = O_p(k_n^{-2\gamma_n}).$$  \hfill (76)

Finally, using the property of least squares projections and Assumption 3.4(i) implying $\sup_{\Theta_n} \|\theta\|_\infty^2 < \infty$ yields,

$$\sup_{\Theta_n} \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[\theta(Z)(f_X(X) - f_X(X))|X_i])^2 \leq \sup_{\Theta_n} \|\theta\|^2_\infty \times \frac{1}{n} \sum_{i=1}^{n} (f_X(X_i) - f_X(X_i))^2 = O_p((nh^d)^{-1} + h^2r)$$  \hfill (77)

by Markov’s inequality and Lemma 1(ii). The first claim then follows from (73) and (75)-(77).

For the second claim of the Lemma, let $\hat{\epsilon}_i(\theta) = E[\nu(X) - f_X(X)\theta(Z)|X_i] - (\pi_n(\theta_0) - \pi_n(\theta)) p^{k_n}(X_i)$ and define $\hat{\epsilon}(\theta) = (\hat{\epsilon}_1(\theta), \ldots, \hat{\epsilon}_n(\theta))'$. Exploiting Assumption 3.3(ii) and $E[p^{k_n}(X)p^{k_n}(X)] = I$ we then obtain that:

$$\sup_{\Theta_n} E[(E[\nu(X) - f_X(X)\theta(Z)|X] - \hat{E}[\nu(X) - f_X(X)\theta(Z)|X])^2] \leq \sup_{\Theta_n} 2E[(p^{k_n}(X)(P'P)^{-1}P'\hat{\epsilon}(\theta))^2] + O(k_n^{-2\gamma_n})$$

$$\leq \sup_{\Theta_n} 2\|P'P)^{-1}P'\hat{\epsilon}(\theta)\|^2 + O(k_n^{-2\gamma_n}) \leq n^{-1}\gamma_n \times \sup_{\Theta_n} \|\hat{\epsilon}(\theta)P(P'P)^{-1}P'\hat{\epsilon}(\theta) + O(k_n^{-2\gamma_n}) = O_p(k_n^{-2\gamma_n})$$  \hfill (78)
where the final inequality is a result of $\gamma_n$, the largest eigenvalue of $n(P'P)^{-1}$, converging in probability to one. The equality in (78) in turn follows from Assumption 3.3(iii) and $P(P'P)^{-1}P'$ being idempotent. ■

**Corollary 1.** Let Assumptions 3.1(i)-(iii), 3.2(i)-(iii), 3.3(i)-(iii) and 3.4(i) hold, $c_n = \frac{b_n}{\sqrt{n}} + k_n^{-\frac{2}{2r}} + (nh_{dx})^{-1} + h^{2r}$. Then, uniformly in $\Theta_n$ with probability tending to one (i) $Q_n(\theta) \leq 16Q(\theta) + Op(c_n)$ and (ii) $Q(\theta) \leq 16Q_n(\theta) + Op(c_n)$.

**Proof:** First apply (75) and (77) together with $E[\nu(X)|X_i] = E[\nu(X)|X_i]$ and (72) to obtain

$$
\frac{1}{n} \sum_{i=1}^{n} (E[\nu(X) - f_X(X)\theta(Z)|X_i])^2 \leq \sup_{\Theta_n} \frac{4}{n} \sum_{i=1}^{n} (E[\nu(X) - f_X(X)\theta(Z)|X_i])^2 + Op(k_n^{-\frac{2}{2r}})
$$

$$
\leq \sup_{\Theta_n} 8E[(E[\nu(X) - f_X(X)\theta(Z)|X])^2] + Op(k_n^{-\frac{2}{2r}}) + (nh_{dx})^{-1} + h^{2r}
$$

(79)

with probability approaching one. The first claim follows by (79) and Lemma 2(ii).

Similarly, for the second claim of the Corollary, exploit Lemma 2(ii) and (72) to obtain that

$$
\sup_{\Theta_n} E[(E[\nu(X) - f_X(X)\theta(Z)|X])^2] \leq \sup_{\Theta_n} 2E[(E[\nu(X) - f_X(X)\theta(Z)|X])^2] + Op(k_n^{-\frac{2}{2r}})
$$

$$
\leq \sup_{\Theta_n} \frac{4}{n} \sum_{i=1}^{n} (E[\nu(X) - f_X(X)\theta(Z)|X_i])^2 + Op(k_n^{-\frac{2}{2r}}) + (nh_{dx})^{-1} + h^{2r}
$$

(80)

with probability tending to one. The result then follows by combining (80) with (75) and (77). ■

**Proof of Theorem 3.1:** We first show consistency under $d_H(\cdot, \cdot, \|\cdot\|_\infty)$ by verifying the conditions of Theorem .1. Conditions (i) and (ii) hold for $\|\cdot\| = \|\cdot\|_{\infty}$ and $c_1n = \delta_n$ by Assumptions 3.4(i)-(ii). In turn, Condition (iii) holds with $c_2n = \frac{b_n}{\sqrt{n}} + k_n^{-\frac{2}{2r}} + (nh_{dx})^{-1} + h^{2r}$ and $C_1 = 16$ by Corollary .1, while the law of conditional expectations yields

$$
Q(\theta) = E[(E[\nu(X) - f_X(X)\theta(Z)|X])^2] = E[f_X^2(X)(E[\theta_0(Z) - \theta(Z)|X])^2] \leq E[f_X^2(X)] \times \inf_{\theta_0} \|\theta_0 - \theta\|^2_{\infty} \quad (81)
$$

implying Condition (iv) holds with $C_2 = E[f_X^2(X)]$ and $\kappa_1 = 2$. The first claim then follows by Theorem .1.

For the second claim of the Theorem, note that $f_X$ bounded implies $\|\cdot\|_w \lesssim \|\cdot\|_{\infty}$. Hence, Conditions (i)-(iii) are verified by the above discussion for the case $\|\cdot\| = \|\cdot\|_{\infty}$. Since $Q(\theta) = \|\theta - \theta_0\|^2_w$ by (81), Conditions (iv)-(v) are immediately verified with $\kappa_1 = \kappa_2 = 2$ and $C_1 = C_2 = 1$. Hence, Theorem .1 yields

$$
d_H(\Theta_0, \Theta_0, \|\cdot\|_w) = Op(\max\{\sqrt{b_n/a_n}, \delta_n\}) \quad (82)
$$

Since $\sqrt{b_n/a_n}/\delta_n \rightarrow \infty$ the second claim of the Theorem then follows. ■

**Proof of Proposition 3.1:** We verify Assumptions 3.1-3.4 and apply Theorem 3.1 to establish the result. Since $f_X$ is continuous, it is also bounded on $[0, 1]$ while $f_X'$ being Lipschitz implies $\|f\|_{\infty, 2} < \infty$. Hence Condition (i) and
Proof of Theorem 3.2:

Therefore, since \(h\) hence a well defined random variable. If the model is identified, so that \(\Theta \mid M\) since (i) and (ii) imply

\[ \sup M \] constitutes a construction, \(\sup M\) being the unique minimum on \(\Theta\) establish the theorem.

\[ \text{Proof of Theorem 3.2: By Example 3.1 in Stinchcombe and White (1992), the estimator } \hat{\theta} \text{ is measurable and hence a well defined random variable. If the model is identified, so that } \Theta_0 = \{\theta_0\}, \text{ then by (ii) we have} \]

\[ \|\hat{\theta} - \theta_0\| \leq d_H(\hat{\Theta}_0, \Theta_0, \|\cdot\|) = o_p(1). \] (84)

Therefore, without loss of generality we assume \(\Theta_0\) is not a singleton. Let \(N_\epsilon(\theta_0)\) be an \(\epsilon\) open neighborhood of \(\theta_0\) so that \(\Theta_0 \cap N_\epsilon(\theta_0) \neq \emptyset\). Since \(\Theta_0 \cap N_\epsilon(\theta_0) \subseteq \Theta\) is closed, it is also compact by virtue of \(\Theta\) being compact. Therefore, the continuity of \(M : \Theta \to R\) and \(\theta_0\) being the unique minimum on \(\Theta_0\) imply that

\[ \delta \equiv \min_{\hat{\Theta}_0 \cap N_\epsilon(\theta_0)} M(\hat{\theta}) - M(\theta_0) > 0. \] (85)

Since (i) and (ii) imply \(M : \Theta \to R\) is uniformly continuous in \(\Theta\), there exists a \(\zeta\) such that if \(\|\theta_1 - \theta_2\| < \zeta\) then \(|M(\theta_1) - M(\theta_2)| < \delta/4\). Letting \(\Theta^c_\zeta\) denote the closed \(\zeta\) blowup of \(\Theta_0\) we then obtain:

\[ \min_{\overline{\Theta}^c_\zeta \cap N_\epsilon(\theta_0)} M(\hat{\theta}) - M(\theta_0) > \frac{3}{4} \delta. \] (86)

Therefore, since \(d_H(\hat{\Theta}_0, \Theta_0, \|\cdot\|) < \zeta\) implies \(\Theta_0 \subset \Theta^c_\zeta\), it follows from (86) that:

\[ P(\|\hat{\theta} - \theta_0\| < \epsilon) \geq P(M(\hat{\theta}) - M(\theta_0) \leq 3\delta/4; \ d_H(\hat{\Theta}_0, \Theta_0, \|\cdot\|) < \zeta). \] (87)

Let \(\theta_p \in \arg\inf_{\theta_0} \|\theta_0 - \theta_p\|\). If \(d_H(\hat{\Theta}_0, \Theta_0, \|\cdot\|) < \zeta\), then \(\|\theta_0 - \theta_p\| < \zeta\) and hence \(M(\theta_p) \leq M(\theta_0) + \delta/4\). By definition of \(\hat{\theta}\), \(M_n(\hat{\theta}_0) \leq M_n(\theta_p)\) which implies \(M(\hat{\theta}) - M(\theta_p) \leq |M(\hat{\theta}) - M_n(\hat{\theta})| + |M(\theta_p) - M_n(\theta_p)|\). Thus, \(M(\theta_p) \leq M(\theta_0) + \delta/4\) and \(|M(\hat{\theta}) - M_n(\hat{\theta})| + |M(\theta_p) - M_n(\theta_p)| \leq \delta/4\) imply that \(M(\hat{\theta}) - M(\theta_0) \leq \delta/2\). Therefore, using (87), we get that:

\[ P(\|\hat{\theta} - \theta_0\| < \epsilon) \geq P(\|M(\hat{\theta}) - M_n(\hat{\theta})\| + \|M(\theta_p) - M_n(\theta_p)\| \leq \delta/4; \ d_H(\hat{\Theta}_0, \Theta_0, \|\cdot\|) < \zeta). \] (88)

Therefore, (88), (iii), (iv) and \(\hat{\theta}, \theta_p \in \Theta_0 \subseteq \Theta\) establish the theorem. ■
Proof of Corollary 3.1: Since $\Theta$ is compact under $\| \cdot \|_\infty$ and $\Theta_0$ is closed it is also compact. Hence, the convexity of $\Theta_0$ and strict convexity of $M : \Theta \to \mathbb{R}$ implies a unique minimum is attained. The first claim of the corollary then follows by Theorems 3.1 and 3.2. The second claim is implied by Theorem 3.1.

Proof of Proposition 3.2: Note $\Theta = \{ \theta : \|\theta\|_{\infty,1} \leq B \}$ is convex, while $M : \Theta \to \mathbb{R}$ is strictly convex on $\Theta$. Next, notice that since $\theta \in \Theta$ are uniformly bounded we obtain for some $K > 0$: 

$$N[1](\epsilon, \Theta^2, \| \cdot \|_{\infty}) \leq N[1](\epsilon/K, \Theta, \| \cdot \|_{\infty}) < \infty$$

where the final inequality holds by Theorem 2.7.1 in van der Vaart and Wellner (1996). Therefore, $\Theta^2$ is a Glivenko-Cantelli class, and by Theorem 2.4.1 in van der Vaart and Wellner (1996) we can conclude:

$$\sup_{\theta} \frac{1}{n} \sum_{i=1}^{n} \theta^2(Z_i) - E[\theta^2(Z)] = o_p(1).$$

(90)

Hence, we have verified the conditions of Corollary 3.1 and the result follows by Proposition 3.1.

Lemma 3. Let $\mathcal{X}_n$ be the $\sigma$-field generated by $\{X_i\}$ and $\{W_{in}\}$ be a triangular array of random variables that is measurable with respect to $\mathcal{X}_n$. If $n^{-1} \sum_i W_{in}^2 = O_p(1)$, $\|\hat{\theta} - \theta_0\|_{\infty} = o_p(1)$ and in addition Assumptions 3.1(i), 3.1(iii), and 3.4(i) hold, then it follows that 

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{in} \{\hat{\theta}(Z_i) - \theta_0(Z_i) - E[\hat{\theta}(Z) - \theta_0(Z)|X_i]\} = o_p(1).$$

(91)

Proof: Define the class $\mathcal{F}_n \equiv \{ \theta - \theta_0 - E[\theta(Z) - \theta_0(Z)] : \theta \in \Theta \text{ and } \|\theta - \theta_0\|_{\infty} \leq \varsigma_n \}$, where $\varsigma_n \searrow 0$ is such that $\|\hat{\theta} - \theta_0\|_{\infty} = o_p(\varsigma_n)$. Notice that since $\{W_{in}\}$ is measurable with respect to $\mathcal{X}_n$, $E[f(X_i, Z_i)W_{in}] = 0$ for any $f \in \mathcal{F}_n$ and that $Z_i, Z_j$ are conditionally independent given $\mathcal{X}_n$ for all $i \neq j$. Applying Markov’s inequality for conditional expectations, $\|\hat{\theta} - \theta_0\|_{\infty} = o_p(\varsigma_n)$ and Lemma 2.3.6 in van der Vaart and Wellner (1996) then yields:

$$P(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{in} \{\hat{\theta}(Z_i) - \theta_0(Z_i) - E[\hat{\theta}(Z) - \theta_0(Z)|X_i]\} > \eta) \leq \frac{1}{\eta} E[\sup_{\mathcal{F}_n} \frac{2}{\sqrt{n}} \sum_{i=1}^{n} W_{in} f(X_i, Z_i)\epsilon_i|\mathcal{X}_n] + o(1)$$

(92)

where $\{\epsilon_i\}$ are i.i.d. Rademacher random variables independent of $\{X_i, Z_i\}_{i=1}^{n}$. Next, define the semimetric on $\mathcal{F}_n$:

$$\|f\|_n = \sqrt{D_n} \times \|f\|_{\infty}$$

(93)

where $D_n = n^{-1} \sum_i W_{in}$. Let $E[\cdot]$ denote expectation over $\{\epsilon_i\}$, use Corollary 2.2.8 in van der Vaart and Wellner (1996), notice that the diameter of $\mathcal{F}_n$ under $\| \cdot \|_n$ is less than or equal to $2\varsigma_n \sqrt{D_n}$ and exploit (93) to conclude:

$$E[\sup_{\mathcal{F}_n} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{in} f(X_i, Z_i)\epsilon_i] \lesssim \int_0^{\infty} \sqrt{\log N(\epsilon, \mathcal{F}_n, \| \cdot \|_n)}d\epsilon \leq \int_0^{2\varsigma_n \sqrt{D_n}} \frac{1}{\sqrt{\log N(\epsilon/\sqrt{D_n}, \mathcal{F}_n, \| \cdot \|_\infty)}}d\epsilon$$

(94)
Recall that for any class $F$ and norm $\| \cdot \|$, $N(\epsilon, F, \| \cdot \|) \leq N_{\|}^{1}(2\epsilon, F, \| \cdot \|)$. Since in addition $N_{\|}^{1}(\epsilon, F \cap \Theta, \| \cdot \|_{\infty}) \leq N_{\|}^{1}(\epsilon/2, \Theta, \| \cdot \|_{\infty})$, (94) and Theorem 2.7.1 in van der Vaart and Wellner (1996) in turn implies that:

$$E_{\epsilon}[\sup_{F_{\epsilon}^{n}} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(X_{i}, Z_{i}) \epsilon_{i} \right\|^{2} \lesssim \int_{0}^{2 \epsilon_{n}} \left( \frac{\sqrt{D_{n}}}{\epsilon} \right)^{d_{n}} d\epsilon = \sqrt{D_{n}} \times (2\epsilon_{n})^{1-d_{n}}.$$  

(95)

Thus, since $\epsilon_{n} \downarrow 0$, $m > d_{n}/2$ and $D_{n} = O_{p}(1)$ by hypothesis, the desired claim follows from (92) and (95).

**Lemma 4.** Let $X_{n}$ be the $\sigma$-field generated by $\{X_{i}\}$ and $\{W_{in}\}$, $\{\tilde{W}_{in}\}$ be triangular arrays of random variables that are measurable with respect to $X_{n}$ such that $n^{-\frac{1}{2}} \sum_{i}(W_{in} - \tilde{W}_{in})^{2} = O_{p}(1)$. If $\hat{\theta} \in \Theta_{n}$ satisfies $||\hat{\theta} - \theta_{0}||_{\infty} = o_{p}(1)$, $||\hat{\theta} - \theta_{0}||_{w} = o_{p}(n^{-\frac{1}{2}})$ and Assumptions 3.1(i)-(iii), 3.2(i)-(iv), and 3.4(i) hold, then:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{in} \nu(X_{i}) - \hat{\theta}(Z_{i}) \hat{f}_{X}(X_{i}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{W}_{in} \nu(X_{i}) - \hat{\theta}(Z_{i}) \hat{f}_{X}(X_{i}) + o_{p}(1).$$  

(96)

**Proof:** First observe that since $\theta \in \Theta_{n}$ are uniformly bounded by Assumption 3.4(i), Cauchy-Schwarz yields:

$$\frac{1}{n} \sum_{i=1}^{n} (W_{in} - \tilde{W}_{in})(\hat{f}_{X}(X_{i}) - f_{X}(X_{i})) \hat{\theta}(Z_{i}) | \lesssim \left[ \frac{1}{n} \sum_{i=1}^{n} (W_{in} - \tilde{W}_{in})^{2} \right]^{\frac{1}{2}} \times \left[ \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_{X}(X_{i}) - f_{X}(X_{i}))^{2} \right]^{\frac{1}{2}} = o_{p}(n^{-\frac{1}{2}})$$  

(97)

where the final result is implied by (77), Assumption 3.2(iv) and $n^{-\frac{1}{2}} \sum_{i}(W_{in} - \tilde{W}_{in})^{2} = O_{p}(1)$ by hypothesis. Since $f_{X}$ is bounded, $n^{-1} \sum_{i}(W_{in} - \tilde{W}_{in})^{2} f_{X}^{2}(X_{i}) = O_{p}(n^{-\frac{1}{2}})$ and hence by Lemma 3.3 we can conclude that:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (W_{in} - \tilde{W}_{in}) f_{X}(X_{i}) \hat{\theta}(Z_{i}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (W_{in} - \tilde{W}_{in}) f_{X}(X_{i}) \{E[\hat{\theta}(Z) - \theta_{0}(Z)|X_{i}] + \theta_{0}(Z_{i})\} + o_{p}(1).$$  

(98)

Further observe, however, that since $f_{X}$ is bounded, we obtain by the Cauchy Schwarz inequality that:

$$\frac{1}{n} \sum_{i=1}^{n} (W_{in} - \tilde{W}_{in}) f_{X}(X_{i}) E[\hat{\theta}(Z) - \theta_{0}(Z)|X_{i}] \lesssim \frac{1}{n} \sum_{i=1}^{n} (W_{in} - \tilde{W}_{in})^{2} \frac{1}{n} \sum_{i=1}^{n} (E[\hat{\theta}(Z) - \theta_{0}(Z)|X_{i}]^{2})^{\frac{1}{2}} = o_{p}(n^{-\frac{1}{2}})$$  

(99)

where the final result follows by Markov's inequality and $||\hat{\theta} - \theta_{0}||_{w} = o_{p}(n^{-\frac{1}{2}})$. Hence, from (97)-(99) we conclude:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (W_{in} - \tilde{W}_{in}) (\nu(X_{i}) - \hat{\theta}(Z_{i}) \hat{f}_{X}(X_{i})) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (W_{in} - \tilde{W}_{in}) (\nu(X_{i}) - \theta_{0}(Z_{i}) f_{X}(X_{i})) + o_{p}(1).$$  

(100)

Further, since $E[\theta_{0}(Z) f_{X}(X)|X_{i}] = \nu(X_{i})$, and $\{W_{in}\}$, $\{\tilde{W}_{in}\}$ are both measurable with respect to $X_{n}$:

$$E[(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (W_{in} - \tilde{W}_{in}) (\nu(X_{i}) - \theta_{0}(Z_{i}) f_{X}(X_{i})))^{2} | X_{i}] = \frac{1}{n} \sum_{i=1}^{n} (W_{in} - \tilde{W}_{in})^{2} E[[\nu(X) - \theta_{0}(Z) f_{X}(X)]^{2} | X_{i}] = o_{p}(1)$$  

(101)

since $\nu$, $f_{X}$ and $\theta_{0}$ are all bounded by hypothesis. The conclusion of the Lemma then follows from (100), (101) and Markov's inequality for conditional expectations.

**Lemma 5.** Let $F$ satisfy $\| f \|_{\infty} \leq M$, $E[f(Z, X)|X] = 0$ for all $f \in F$, and $\int_{0}^{\infty} \sqrt{\log N(\epsilon, F, \| \cdot \|_{\infty})} d\epsilon < \infty$. If $E[\theta(Z)|x] = p^{K_{n}}(x)(P^{P})^{-1} \sum_{i} p^{K_{n}}(X_{i}) E[\theta(Z)|X_{i}]$ and Assumptions 3.1(i)-(iii), 3.3(i)-(iii), 3.3(i), 4.1(ii) hold:

$$\sup_{F, \hat{\theta} \in \Theta_{n}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{E[\theta(Z)|X_{i}] - E[\theta(Z)|X_{i}]\} f(Z_{i}, X_{i}) \right| = o_{p}(1).$$  

(102)
PROOF: Define \( \epsilon_i(\theta) = \theta(Z_i) - E[\theta(Z)|X_i] \), \( \epsilon(\theta) = (\epsilon_1(\theta), \ldots, \epsilon_n(\theta)) \), \( \beta(\theta) = (P'P)^{-1}P'\epsilon(\theta) \) and notice that:

\[
\sup_{F, \theta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \hat{E}[\theta(Z)|X_i] - E[\theta(Z)|X_i] \right) f(Z_i, X_i) \right| = \sup_{F, \theta_n} \left| \frac{1}{n} \sum_{i=1}^{n} p^{k_\alpha}(X_i) \beta(\theta)f(Z_i, X_i) \right|. \tag{103}
\]

Since the smallest eigenvalue of \( E[p^{k_\alpha}(X)p^{k_\alpha}(X)] \) is bounded away from zero, (72) and (75) imply:

\[
\sup_{\theta_n} \| \beta(\theta) \|^2 = \sup_{\theta_n} \epsilon'(\theta)P(P'P)^{-2}P'\epsilon(\theta) \leq \sup_{\theta_n} E[\epsilon'(\theta)P(P'P)^{-1}p^{k_\alpha}(X)p^{k_\alpha}(X)(P'P)^{-1}P'\epsilon(\theta)]
\]

\[
\leq \sup_{\theta_n} \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[\theta(Z)|X_i] - E[\theta(Z)|X_i])^2 = O_p\left( \frac{k_n}{n} \right) \tag{104}
\]

with probability tending to one. Let \( \varsigma_n \prec 0 \) such that \( \varsigma_n \sqrt{n}/\sqrt{\xi_0} \to \infty \) and \( G_{n_0}^{c_0} = \{ p^{k_\alpha} \beta : \| \beta \| \leq \varsigma_n \} \). By results (103) and (104), together with Markov’s inequality we are then able to conclude that:

\[
P(\sup_{F, \theta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \hat{E}[\theta(Z)|X_i] - E[\theta(Z)|X_i] \right) f(Z_i, X_i) \right| > \eta) \leq \frac{1}{\eta} E\left[ \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i, f(Z_i, X_i)) \right| \right] + o(1) \lesssim M_{\varsigma_0} \xi_0 \int_{0}^{1} \sqrt{1 + \log N(\varsigma_0, \xi_0, n, G_{n_0}^{c_0} \times F, \| \cdot \|_{\infty})} \, de \tag{105}
\]

where the second inequality follows by Theorem 2.14.1 in van der Vaart and Wellner (1996) and noting that \( \| gf \|_{\infty} \leq M_{\varsigma_0} \xi_0 \) uniformly in \( g \in G_{n_0}^{c_0} \) and \( f \in F \). Note that every \( g_j \in G_{n_0}^{c_0} \) is of the form \( g_j = p^{k_\alpha} \beta_j \) and apply the triangle and Cauchy-Schwarz inequality to obtain the bound:

\[
\| g_1 f_1 - g_2 f_2 \| \leq \| g_1 - g_2 \| M + \| g_1 \|_\infty \| f_1 - f_2 \| \leq \zeta_0 \| \beta_1 - \beta_2 \| M + \xi_0 \varsigma_0 \| f_1 - f_2 \|. \tag{106}
\]

Let \( B_n^{c_0} \) be a sphere of radius \( \varsigma_n \) in \( R^{k_\alpha} \). We are then able to conclude from (106) that:

\[
N(\epsilon, G_{n_0}^{c_0} \times F, \| \cdot \|_{\infty}) \leq N\left( \frac{\epsilon}{M_{\varsigma_0}}, B_n^{c_0}, \| \cdot \| \right) \times N\left( \frac{\epsilon}{\xi_0 \varsigma_0}, F, \| \cdot \|_{\infty} \right). \tag{107}
\]

Since \( N(\epsilon, B_n^{c_0}, \| \cdot \|) \leq (2\varsigma_0/\epsilon)^{k_\alpha} \) we can combine (107) with (105) and use finiteness of the resulting integral to obtain:

\[
E[\sup_{F, \theta_n} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i, f(Z_i, X_i))] \lesssim \varsigma_0 \xi_0 \int_{0}^{1} \sqrt{1 + k_n \log \left( \frac{2}{\epsilon} \right) + \log N(\epsilon M, F, \| \cdot \|_{\infty})} \, de \lesssim \sqrt{k_n} \varsigma_0 \xi_0. \tag{108}
\]

To conclude, notice that \( \xi_0 k_n/\sqrt{n} \to 0 \) implies we may choose \( \varsigma_0 \) so that \( \sqrt{k_n} \varsigma_0 \xi_0 \to 0 \), and the claim of the Lemma follows from (108) and Markov’s inequality.

**Lemma 6.** Let \( \tilde{v} \in V \cap \Theta, u = \pm \tilde{v} \) and \( u_n = \pm \Pi_n u \). If Assumptions 3.1(i)-(iii), 3.2(i)-(iv), 3.3(i)-(iii), 3.4(i)-(ii), 4.1(i)-(ii) hold, and \( \bar{\theta} \in \Theta_n \) satisfies \( \| \bar{\theta} - \theta_0 \|_w = o_p(n^{-\frac{1}{2}}) \) and \( \| \bar{\theta} - \theta_0 \|_\infty = o_p(1) \), then:

(i) \( n^{-1} \sum_i E[u(Z)f_X(X)|X_i] \hat{m}(X_i, \bar{\theta}) = n^{-1} \sum_i E[u(Z)f_X(X)|X_i](\nu(X_i) - \hat{\theta}(Z_i)\hat{f}_X(X_i)) + o_p(n^{-\frac{1}{2}}) \)

(ii) \( n^{-1} \sum_i E[u(Z)f_X(X)|X_i](\nu(X_i) - \hat{\theta}(Z_i)\hat{f}_X(X_i)) = n^{-1} \sum_i E[u(Z)f_X(X)|X_i] \hat{f}_X(X_i)(\nu(X_i) - \hat{\theta}(Z_i)\hat{f}_X(X_i)) + o_p(n^{-\frac{1}{2}}) \)
(iii) \( \langle u, \theta_0 - \hat{\theta} \rangle_w = n^{-1} \sum_i E[u(Z)f_X(X_i)|X_i][\hat{m}(X_i, \hat{\theta}) - \hat{m}(X_i, \theta_0)] + o_p(n^{-\frac{1}{2}}) \)

**Proof:** For any function \( w : [0,1]^d \times \mathbb{R} \rightarrow \mathbb{R} \), recall \( E[w(X,Z)|x] = p^{\nu}(x)(P'P)^{-1} \sum_i p^{\nu}(X_i)E[w(X,Z)|X_i] \). In order to establish the first claim of the Lemma, then notice that by rearranging terms it follows that:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[u(Z)f_X(X_i)|X_i][\hat{m}(X_i, \hat{\theta}) - (\nu(X_i) - \hat{\theta}(Z_i)f_X(X_i))]
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (E[u(Z)f_X(X_i)|X_i] - E[u(Z)f_X(X_i)|X_i])\nu(X_i) - \hat{\theta}(Z_i)f_X(X_i)) .
\]

(109)

Hence, (76), Assumption 4.1(ii) and Lemma .4 establish the desired result.

In order to establish the second claim of the Lemma, we first aim to establish that:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i] - \bar{E}[u_n(Z)|X_i])f_X(X_i)(\nu(X_i) - \hat{\theta}(Z_i)f_X(X_i)) = o_p(1) .
\]

(110)

For this purpose, first notice that Cauchy-Schwarz, \( \nu \) bounded and results (75) and (77) establish that:

\[
\left| \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i] - \bar{E}[u_n(Z)|X_i])f_X(X_i)\nu(X_i) \right|
\leq \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i] - \bar{E}[u_n(Z)|X_i])^2 \leq \frac{1}{n} \sum_{i=1}^{n} (f_X(X_i) - \nu(X_i))^2 = o_p(n^{-\frac{1}{2}}) .
\]

(111)

where for the final result we have exploited Assumptions 3.2(iv) and 4.1(ii). Similarly, we obtain that:

\[
\left| \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i] - \bar{E}[u_n(Z)|X_i])(f_X(X_i) - \hat{\theta}(Z_i)f_X(X_i)) \right|
\leq \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i] - \bar{E}[u_n(Z)|X_i])^2 \leq \frac{1}{n} \sum_{i=1}^{n} ((f_X(X_i) - f_X(X_i))^2 + (f_X(X_i) - f_X(X_i))^4) = o_p(n^{-\frac{1}{2}}) .
\]

(112)

Through calculations as in Lemma .1, it is further possible to establish that almost surely in \( X_i \):

\[
E[(f_X(X_i) - f_X(X_i))^4]X_i = O((nh^{2d_z})^{-2} + b^{4w}) .
\]

(113)

Hence, by (111), (112) and (113) together with (75) and Assumptions 3.2(iv) and 4.1(ii), we obtain:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i] - \bar{E}[u_n(Z)|X_i])f_X(X_i)(\nu(X_i) - \hat{\theta}(Z_i)f_X(X_i))
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i] - \bar{E}[u_n(Z)|X_i])f_X(X_i)(\nu(X_i) - \hat{\theta}(Z_i)f_X(X_i)) + o_p(1) .
\]

(114)

In turn, since \( \int_0^\infty \sqrt{N(\epsilon, \Theta, \| \cdot \|_\infty)} d\epsilon < \infty \) by Theorem 2.7.1 in van der Vaart and Wellner (1996) and the \( \theta \in \Theta \) are uniformly bounded, we are able to conclude by Lemma .5 that:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i] - \bar{E}[u_n(Z)|X_i])f_X(X_i)\{\hat{\theta}(Z_i) - \theta_0(Z_i) - E[\hat{\theta}(Z) - \theta_0(Z)|X_i]\} = o_p(1) .
\]

(115)
Similarly, also notice that a second application of Lemma .5 additionally implies that:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i] - \hat{E}[u_n(Z)|X_i]) f_X(X_i) (\nu(X_i) - \theta(Z_i) f_X(X_i)) = o_p(1). \tag{116}
\]

Next observe that the Cauchy-Schwarz inequality, (75) and Assumption 4.1(ii) together with \(\|\hat{\theta} - \theta_0\|_\infty = o_p(n^{-\frac{1}{2}})\) and Markov’s inequality allows us to obtain:

\[
\left| \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i] - \hat{E}[u_n(Z)|X_i]) f_X(X_i) E[\hat{\theta}(Z) - \theta_0(Z)|X_i] \right| \\
\leq \left[ \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i] - \hat{E}[u_n(Z)|X_i])^2 \right]^{\frac{1}{2}} \times \left[ \frac{1}{n} \sum_{i=1}^{n} f_X^2(X_i) (E[\hat{\theta}(Z) - \theta_0(Z)|X_i])^2 \right]^{\frac{1}{2}} = o_p(n^{-\frac{1}{2}}). \tag{117}
\]

Claim (110) then follows from (114), (115), (116) and (117). Let \(\varepsilon_i(\theta) = E[\hat{\theta}(Z)|X_i] - \bar{\pi}_n(\theta) p^{k_n}(X_i)\) and \(\varepsilon(\theta) = (\varepsilon_1(\theta), \ldots, \varepsilon_2(\theta))\). Exploiting Assumption 3.3(ii), that the largest eigenvalue of \(n(P^P)^{-1}\) is bounded in probability and that the matrix \(P(P^P)^{-1}P^\prime\) is indempotent, we are then able to establish:

\[
\sup_{\Theta \in [0,1]^{d_x}} (E[\hat{\theta}(Z)|x] - E[\theta(Z)|x])^2 \leq \sup_{\Theta \in [0,1]^{d_x}} 2(p^{k_n}(x)(P^P)^{-1}P^\prime \varepsilon(\theta))^2 + O(k_n^{-\frac{2\gamma}{\tau}}) = O_p(\xi_0 n^{-\frac{2\gamma}{\tau}}). \tag{118}
\]

Since \(\sup_{\Theta \in [0,1]^{d_x}} \|E[\hat{\theta}(Z)|x]\|_\infty \leq \sup_{\Theta \in [0,1]^{d_x}} \|\theta\|_\infty < \infty\), (118) and Assumption 4.1(ii) implies \(\sup_{\Theta \in [0,1]^{d_x}} \|E[\hat{\theta}(Z)|x]\|_\infty = O_p(1)\). Therefore, result (77) and Assumption 3.2(iv) implies the following result:

\[
\frac{1}{n} \sum_{i=1}^{n} (\hat{E}[u_n(Z)|X_i])^2 (\hat{f}_X(X_i) - f_X(X_i))^2 \leq \sup_{\Theta \in [0,1]^{d_x}} \|E[\hat{\theta}(Z)|x]\|_\infty^2 \times \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_X(X_i) - f_X(X_i))^2 = o_p(n^{-\frac{1}{2}}). \tag{119}
\]

Furthermore, also notice that Assumption 4.1(i)-(ii), 3.4(ii) and \(f_X\) bounded imply that:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (E[u(Z) - u_n(Z)|X_i] f_X(X_i))^2 \lesssim \sqrt{n} \|u - u_n\|_{\infty}^2 = o(1). \tag{120}
\]

Exploiting (119), (76), (120) together with Assumption 4.1(ii) and three successive application of Lemma .4 yields:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[u_n(Z)|X_i] \hat{f}_X(X_i) (\nu(X_i) - \hat{\theta}(Z_i)) \hat{f}_X(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[u_n(Z)|X_i] f_X(X_i) (\nu(X_i) - \hat{\theta}(Z_i)) \hat{f}_X(X_i) + o_p(1) \\
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[u_n(Z)|X_i] f_X(X_i) (\nu(X_i) - \hat{\theta}(Z_i)) \hat{f}_X(X_i) + o_p(1) \\
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[u(Z)|X_i] f_X(X_i) (\nu(X_i) - \hat{\theta}(Z_i)) \hat{f}_X(X_i) + o_p(1) \tag{121}
\]

which together with (110) implies the second claim of the Lemma.

In order to establish the third claim of the Lemma, we rearrange as in (109), and exploit Lemma .3 together with \(E[f^2_X(X_i)] < \infty\) and \(\sup_{[0,1]^{d_x}} |E[u(Z)f_X(X)|x]| = O_p(1)\) so as to derive that:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[u(Z)f_X(X)|X_i](\hat{m}(X_i, \hat{\theta}) - \bar{m}(X_i, \theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[u(Z)f_X(X)|X_i] \hat{f}_X(X_i)(\theta_0(Z_i) - \hat{\theta}(Z_i)) \\
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[u(Z)f_X(X)|X_i] \hat{f}_X(X_i) E[\theta_0(Z) - \hat{\theta}(Z)|X_i] + o_p(1). \tag{121}
\]

34
Since $f_X$ is bounded from above and below, $\|\hat{\theta} - \theta_0\|_w^2 \simeq E[(E[\hat{\theta}(Z) - \theta_0(Z)|X])^2]$. Hence, $\|\hat{\theta} - \theta_0\|_w = o_p(n^{-\frac{1}{4}})$, sup\([0,1]^{d_x} |\tilde{E}[u(Z)f_X(X)|x]| = O_p(1)$ the Cauchy-Schwarz inequality, (77) and $(nh^{d_x})^{-1} + h^{2r} = o(n^{-\frac{1}{2}})$ imply:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} |\tilde{E}[u(Z)f_X(X)|X_i] = O_p(1)$$

Similarly, exploiting (76), Assumption 4.1(ii),

For notational simplicity, let $\theta = 0$ and define the class $F = \{E[u(Z)f_X(X)|x]E[\theta(Z) - \hat{\theta}(Z)|X]| \theta \in \Theta\}$. Note that by Assumptions 3.1(ii), 3.4(i) and 4.1(i) $E[u(Z)f_X^2(X)|x]$ is bounded. Therefore, it follows that:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[u(Z)f_X(X)|X_i](\hat{\theta}(Z) - \hat{\theta}(Z)) = o_p(1)$$

Next define the class $\mathcal{F} = \{E[u(Z)f_X^2(X)|x]E[\theta(Z) - \hat{\theta}(Z)|x] : \theta \in \Theta\}$. Theorem 2.5.6 in van der Vaart and Wellner (1996) together with $m > d_x/2$ then imply $\mathcal{F}$ is a Donsker class. Thus, since $\|\hat{\theta} - \theta_0\|_\infty = o_p(1)$ implies sup\([0,1]^{d_x} |\tilde{E}[u(Z)f_X^2(X)|x]| = O_p(1)$, we obtain that:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{E[u(Z)f_X(X)|X_i]E[(\theta(Z) - \hat{\theta}(Z))f_X(X)|X_i] - (u, \theta(Z) - \hat{\theta}(Z))\} = o_p(1)$$

The third claim of the Lemma therefore follows from (125) and (127).

**Lemma 7.** Suppose Assumptions 3.1(i)-(iii), 3.2(i)-(v) and 4.1(i) hold. It then follows that:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E[u(Z)f_X(X)|X_i]E[\hat{\theta}(Z) - \theta_0(Z)] = o_p(1)$$

**Proof:** For notational simplicity, let $g(x, z) = E[\tilde{v}(Z)f_X(X)|x]\theta_0(z)$, $w_i = (x_i, z_i)$ and define the kernel:

$$H_n(W_i, W_j) = g(X_i, Z_i)(\tilde{K}_h(X_i, X_j) - E[|X_i|]h^{d_x}) + g(X_j, Z_j)(\tilde{K}_h(X_j, X_i) - E[|X_j|]h^{d_x})$$
where \( K_h(\frac{X_i - X_j}{n}, X_i) = \prod_{k=1}^{d_x} K_h(\frac{X^{(k)}_i - X^{(k)}_j}{n}, X^{(k)}_i) \). Since \( g(x, z) \) is bounded, Lemma 1(i) then implies that:

\[
|g(X_i, Z_i)(E[\hat{f}_X(X_i)|X_i] - f_X(X_i))| \lesssim |E[\hat{f}_X(X_i)|X_i] - f_X(X_i)| = O(h^r).
\] (130)

Therefore, since \( h^r n^{-\frac{1}{2}} = o(1) \) by Assumption 3.2(iv), we can use (129) to conclude:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i, Z_i)(\hat{f}_X(X_i) - f_X(X_i)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i, Z_i)(\hat{f}_X(X_i) - E[\hat{f}_X(X_i)|X_i]) + o(1)
\]

\[
= \frac{1}{n^{\frac{1}{2}}h^{d_x}} \sum_{i=1}^{n} \sum_{j > i} H_n(W_i, W_j) + o(1).
\] (131)

Next, expand the square, notice that all cross terms have expectation zero and employ Jensen’s inequality to obtain:

\[
E[(\frac{1}{n^{\frac{1}{2}}h^{d_x}} \sum_{i=1}^{n} \sum_{j > i} H_n(W_i, W_j) - E[H_n(W_i, W_j)|W_i] - E[H_n(W_i, W_j)|W_j] + E[H_n(W_i, W_j)])^2]
\]

\[
\leq \frac{4}{n^{\frac{1}{2}}h^{d_x}} \sum_{i=1}^{n} \sum_{j > i} E[H_n^2(W_i, W_j)] = \frac{2(n-1)}{n^{\frac{1}{2}}h^{d_x}}E[H_n^2(W_i, W_j)] = O(nh^{d_x})
\] (132)

where for the final result we have used \( E[H_n^2(W_i, W_j)] = O(h^{d_x}), \) which follows from (70). Hence, since \( nh^{d_x} \to \infty, \)

Chebychev’s inequality and (132) together with \( E[H_n(W_i, W_j)] = 0 \) from (129), yields

\[
\frac{1}{n^{\frac{1}{2}}h^{d_x}} \sum_{i=1}^{n} \sum_{j > i} H_n(W_i, W_j) = \frac{1}{\sqrt{n}h^{d_x}} \sum_{i=1}^{n} E[H_n(W_i, W_j)|W_i] + o_p(1)
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\frac{1}{h^{d_x}} E[g(X_i, Z_i)K_h(\frac{X_j - X_i}{h}, X_j)|X_i] - E[g(X_i, Z_i)\hat{f}_X(X_j)]) + o_p(1),
\] (133)

where the final equality follows by (129) and direct calculation. Moreover, observe that by (130) we also have

\[
E[g(X_j, Z_j)\hat{f}_X(X_j)] = E[g(X_j, Z_j)f_X(X_j)] + O(h^r).
\] (134)

In addition, letting \( \Gamma(X_i) = \{ u : -t/X_i \leq u \leq (1 - t)/X_i \} \) and doing the change of variables \( u = (X_j - X_i)/h: \)

\[
\lim_{h \to 0} E[(E[E[g(X_j, Z_j)|X_j]K_h(\frac{X_j - X_i}{h}, X_j)|X_i]h^{-d_x} - E[g(X_i, Z_i)|X_i]f_X(X_i))^2]
\]

\[
\lim_{h \to 0} E[(\int_{\Gamma(X_i)} E[g(X_j, Z_j)|X_i + hu]K_h(u, X_i + hu)f_X(X_i + hu)du - E[g(X_i, Z_i)|X_i]f_X(X_i))^2] = 0
\] (135)

where the final equality is established by noting the integrand is uniformly bounded by Assumptions 3.1(ii), 3.2(ii) and 4.1(i), appealing to the dominated convergence theorem and using Assumption 3.2(v). Therefore:

\[
E[(\frac{1}{\sqrt{n}h^{d_x}} \sum_{i=1}^{n} E[H_n(W_i, W_j)|W_i] - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (E[g(X_i, Z_i)|X_i]f_X(X_i) - E[g(X, Z)f_X(X)])^2] = o(1)
\] (136)

by results (133), (134) and (135). The claim of the Lemma then follows from (131), (133), (136), Markov’s inequality and \( g(x, z) = E[\tilde{v}(Z)f_X(X)|x]g_0(x) \) which implies \( E[g(X, Z)|x]f_X(x) = E[\tilde{v}(Z)|x]f_X(x)\nu(x). \) □
Proof of Theorem 4.1: Define the class \( \mathcal{F} = \{ g(y, z) = y \theta(z) : \theta \in \Theta \} \). By Theorems 2.7.11 and 2.7.1 in van der Vaart and Wellner (1996), it then follows that:

\[
\int_0^\infty \sqrt{\log N_\varepsilon (\epsilon, \mathcal{F}, \| \cdot \|_{L^2})} \, d\epsilon \leq \int_0^\infty \sqrt{\log N_\varepsilon (\epsilon/\sqrt{E[Y^2]}, \Theta, \| \cdot \|_\infty)} \, d\epsilon < \infty .
\]  

(137)

Therefore, Theorem 2.5.6 in van der Vaart and Wellner (1996) establishes that \( \mathcal{F} \) is a Donsker class. Since by assumption \( \| \hat{\theta}_0 - \theta_0 \|_\infty = o_p(1) \) and in addition \( E[Y \theta_0(Z)] = (\nu, m_0) \) we can further conclude:

\[
\sqrt{n} \{ (\nu, m_0) - (\nu, m_0) \} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{ Y_i \theta_0(Z_i) - E[Y \theta_0(Z)] \} + \sqrt{n} E[Y(\hat{\theta}_0(Z) - \theta_0(Z))] + o_p(1) .
\]  

(138)

In addition, Lemma .6(i) and Lemma .7 together with \( E[\hat{v}(Z)f_X(X)\nu(X)] = E[Y \theta_0(Z)] \) imply the equality:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^n E[\hat{v}(Z)f_X(X_i)\nu(X_i)\theta_0] = \frac{1}{\sqrt{n}} \sum_{i=1}^n E[\hat{v}(Z)f_X(X_i)\nu(X_i) - \theta_0 f_X(X_i)] + o_p(1)
\]  

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^n E[Y \theta_0(Z)] - E[\hat{v}(Z)f_X(X_i)\nu(X_i)\theta_0 f_X(X_i)] + o_p(1) .
\]  

(139)

Therefore, since \( E[Y(\hat{\theta}_0(Z) - \theta_0(Z))] = (\hat{v}, \hat{\theta}_0 - \theta_0) \), combining (138), (139) and Lemma .6(iii) implies:

\[
\sqrt{n} \{ (\nu, m_0) - (\nu, m_0) \}
\]  

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^n \{ Y_i \theta_0(Z_i) - E[\hat{v}(Z)f_X(X_i)\nu(X_i)\theta_0(Z_i)] \} - \frac{1}{\sqrt{n}} \sum_{i=1}^n E[\hat{v}(Z)f_X(X_i)\nu(X_i)\theta_0(Z_i)] + o_p(1) .
\]  

(140)

Furthermore, by the Cauchy-Schwarz inequality, the definition of \( Q_n(\theta) \) and \( f_X \) bounded, it follows that:

\[
\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (E[\hat{v}(Z)f_X(X_i)\nu(X_i)\theta_0(Z_i)] - \hat{E}[\Pi_n \hat{v}(Z)f_X(X_i)\nu(X_i)\theta_0(Z_i)]) \right| \leq n^{\frac{1}{2}} \| v - \Pi_n \hat{v} \|_\infty \times n^{\frac{1}{2}} \| Q_n(\theta_0) \|_2 = o_p(1)
\]  

(141)

where the final result follows by Assumptions 4.1(i)-(ii), 3.2(ii), Corollary .1 and \( \| \hat{\theta}_0 - \theta_0 \|_w = o_p(n^{-\frac{1}{2}}) \). Similarly,

\[
\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (E[\Pi_n \hat{v}(Z)f_X(X_i)\nu(X_i)\theta_0(Z_i)] - \hat{E}[\Pi_n \hat{v}(Z)f_X(X_i)\nu(X_i)]) \right| \leq n^{\frac{1}{2}} \| Q_n(\theta_0) \|_2 \times \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n (E[\Pi_n \hat{v}(Z)f_X(X_i)\nu(X_i)] - \hat{E}[\Pi_n \hat{v}(Z)f_X(X_i)])^2 \right)^{\frac{1}{2}} = o_p(1)
\]  

(142)

where the final result is implied by Lemma .2(i). The theorem then follows by (140), (141) and (142). ■

Proof of Lemma 4.1: As a first step, we obtain a rate of convergence for \( \| \hat{v} - \Pi_n \hat{v} \|_w \). Towards this end, notice that since \( C(\theta) - C(\hat{v}) = \| \theta - \hat{v} \|^2_w \) for all \( \theta \in \Theta \), Assumption 3.4(ii) implies that:

\[
\| \hat{v} - \Pi_n \hat{v} \|^2_w \leq 2 \| \hat{v} - \hat{v} \|^2 + O(\delta_n^2) = 2 \{ C(\hat{v}) - C(\Pi_n \hat{v}) \} + O(\delta_n^2) .
\]  

(143)

Therefore, since by definition \( C_n(\hat{v}) \leq C_n(\Pi_n \hat{v}) \), we can in turn employ (143) to conclude that:

\[
\| \hat{v} - \Pi_n \hat{v} \|^2_w \leq 2 \{ (C(\hat{v}) - C_n(\hat{v})) - (C(\Pi_n \hat{v}) - C_n(\Pi_n \hat{v})) \} + O(\delta_n^2) \leq 2 \sup_{\theta \in \Theta_n} | C(\theta) - C_n(\theta) | + O(\delta_n^2) .
\]  

(144)
Define $\mathcal{F} = \{ f(y,z) = y\theta(z) : \theta \in \Theta \}$ and recall that by (137), the class $\mathcal{F}$ is Donsker. Therefore, we obtain
\[
\sup_{\theta \in \Theta_n} |C(\theta) - C_n(\theta)| \leq \sup_{\theta \in \Theta_n} |E[ (E[\theta(Z)|X])^2 f_X^2(X) ] - \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[\theta(Z)]f_X(X_i)]^2 | + \sup_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} Y_i\theta(Z_i) - E[Y\theta(Z)]| = \sup_{\theta \in \Theta_n} |E[ (E[\theta(Z)|X])^2 f_X^2(X) ] - \frac{1}{n} \sum_{i=1}^{n} (\hat{E}[\theta(Z)]f_X(X_i)]^2 | + O_p(n^{-\frac{1}{2}}). \tag{145}
\]
Next, define the class $\mathcal{G} = \{ g(x) = (E[\theta(Z)|x])^2 f_X^2(x) : \theta \in \Theta \}$ and notice $f_X$, $\theta \in \Theta$ uniformly bounded implies:
\[
(E[\theta_1(Z)|x])^2 f_X^2(x) - (E[\theta_2(Z)|x])^2 f_X^2(x) \leq f_X^2(x)(E[\theta_1(Z) - \theta_2(Z)|x]) (E[\theta_1(Z) + \theta_2(Z)|x]) \lesssim |\theta_1 - \theta_2|_{\infty} \tag{146}
\]
for any $\theta_1, \theta_2 \in \Theta$. Hence, $\mathcal{G}$ is Lipschitz in $\Theta$, and by Theorems 2.7.11, 2.7.1 and 2.5.6 in van der Vaart and Wellner (1996), the class $\mathcal{G}$ is Donsker as well. We therefore conclude that:
\[
\sup_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} (|E[\theta(Z)|X_i)]^2 f_X^2(X_i) - E[|E[\theta(Z)|X]p f_X^2(X)]| = O_p(n^{-\frac{1}{2}}). \tag{147}
\]
Moreover, by Cauchy-Schwarz, $E[(E[\theta(Z)|X_i)]^2 f_X^2(X)]$ uniformly bounded in $(x, \theta) \in [0, 1]^d \times \Theta$ and Lemma .2(i):
\[
\sup_{\Theta_n} \frac{1}{n} \sum_{i=1}^{n} (|E[\theta(Z)|X_i)]^2 f_X^2(X_i) - (E[\theta(Z)]f_X^2(X_i)])^2 = O_p(\frac{\sqrt{n}}{n} + k_n^{-\frac{\gamma}{2\gamma}} + (nh_{d_x})^{-\frac{1}{2}} + h^n). \tag{148}
\]
Therefore, combining (144), (145), (147), (148) and absorbing higher order terms we obtain:
\[
\|\hat{v} - \Pi_n \hat{v}\|_w^2 = O_p(\frac{\sqrt{n}}{n} + k_n^{-\frac{\gamma}{2\gamma}} + (nh_{d_x})^{-\frac{1}{2}}) \tag{149}
\]
Finally, the Cauchy-Schwarz inequality, Lemma .2(i), result (149), Markov’s inequality and the definition of $\hat{\theta}_0$ imply
\[
\sup_{\hat{\theta}_0} |r_n(\hat{\theta}_0) - \hat{\gamma}_n(\hat{\theta}_0)| \leq \frac{1}{n} \sum_{i=1}^{n} (E[\theta(Z)|X_i)]^2 f_X^2(X_i) - (E[\theta(Z)]f_X^2(X_i)])^2 \sup_{\hat{\theta}_0} [Q_n(\hat{\theta}_0)] \tag{150}
\]
which establishes the claim of the Lemma.

**Proof of Proposition 4.1:** It follows by direct calculation and noting that $n^\frac{1}{2} a_n / b_n \rightarrow 0$. 

**Proof of Theorem 4.2:** By Corollary 3.1, it follows that $|\hat{\theta}_0 - \theta_0|_{\infty} = o_p(1)$ and $|\hat{\theta}_0 - \theta_0|_{w} = o_p(n^{-\frac{1}{2}})$. In addition, Lemma 4.1 and Assumption 4.1(ii) imply $\sqrt{n} |\hat{r}_n(\hat{\theta}_0) - r_n(\hat{\theta}_0)| = o_p(1)$. Therefore, by Theorem 4.1,
\[
\sqrt{n} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \theta_0(Z_i) + r_n(\hat{\theta}_0) - \langle \nu, m_0 \rangle \right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ Y_i \theta_0(Z_i) - E[\hat{v}(Z)|X_i)] f_X^2(X_i) \theta_0(Z_i) \} + o_p(1). \tag{151}
\]
The claim of the Theorem then follows by the central limit theorem.

**Proof of Lemma 4.2:** First notice, that the properties of least square projections, $\hat{v}$ and $f_X$ bounded imply
\[
\frac{1}{n} \sum_{i=1}^{n} (E[\hat{v}(Z)]f_X^2(X)|X_i]) - E[\hat{v}(Z)]f_X^2(X)|X_i]))^2 \leq \frac{2}{n} \sum_{i=1}^{n} (f_X^2(X_i) - f_X^2(X_i))^2 + \frac{2}{n} \sum_{i=1}^{n} (E[\hat{v}(Z)]f_X^2(X)|X_i]) - E[\hat{v}(Z)]f_X^2(X)|X_i]))^2 = o_p(1) \tag{152}
\]
where the final result follows by $\hat{v} \in \Theta_n$ bounded, $f_X$ bounded as well, (75), (76), (112) and (113). Therefore, since $\hat{\theta}$ is bounded, $\|\hat{\theta}_0 - \theta_0\|_\infty = o_p(1), \|\hat{v} - \tilde{v}\|_w = o_p(1)$ and Markov’s inequality yields:

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{E}[\hat{v}(Z)f^2_X(X)|X_i] \hat{\theta}_0(Z_i) - E[\hat{v}(Z)f^2_X(X)|X_i] \theta_0(Z_i))^2 \leq \|\hat{\theta}_0 - \theta_0\|_\infty^2 \times \frac{4}{n} \sum_{i=1}^{n} (E[\hat{v}(Z)f^2_X(X)|X_i])^2 + \frac{4}{n} \sum_{i=1}^{n} \theta^2_0(Z_i)(E[\hat{v}(Z) - \tilde{v}(Z)|X_i])^2 f_X^2(X_i) + o_p(1) = o_p(1) \tag{153}$$

In turn, Cauchy-Schwarz, $\|\hat{\theta}_0 - \theta_0\|_\infty = o_p(1), E[Y^2] < \infty$ and Markov’s inequality together with (153) implies:

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\theta}^2_0(Z_i)(Y_i - \hat{E}[\hat{v}(Z)f^2_X(X)|X_i])^2 = \frac{1}{n} \sum_{i=1}^{n} \theta^2_0(Z_i)(Y_i - \hat{E}[\hat{v}(Z)f^2_X(X)|X_i])^2 + o_p(1) \tag{154}$$

The claim of the Lemma then follows by the law of large numbers. ■

**Proof of Proposition 4.2**: Follows directly from Theorem 4.2 and Lemma 4.2. ■
References


Darolles, S., J. Florens, and E. Renault, 2003, Nonparametric instrumental regression, GREMAQ.


Santos, A., 2007, Inference in nonparametric instrumental variables with partial identification, University of California at San Diego.


White, H., and K. Chalak, 2006, Identifying effects of endogenous causes in nonseparable systems using covariates, University of California at San Diego.