Instrumental Variables Methods for Recovering Continuous Linear Functionals

Andres Santos*
Department of Economics
University of California, San Diego
e-mail: a2santos@ucsd.edu

March 13, 2008

Abstract

This paper develops methods for estimating continuous linear functionals in a nonparametric instrumental variables (IV) setting. Examples of such functionals include consumer surplus and applications to tests for shape restrictions like monotonicity, concavity and additive separability. The estimation procedure is robust to a setting where the underlying model is not identified but the linear functional of interest is. In order to attain such robustness, it is necessary to use a nuisance parameter that is not identified. A procedure is proposed that circumvents this challenge and delivers a $\sqrt{n}$ asymptotically normal estimator for the linear functional of interest. A Monte Carlo study examines the finite sample performance of the procedure.

KEYWORDS: Instrumental variables, linear functionals, partial identification.

*I would like to thank Graham Elliot, Ivana Komunjer, Aprajit Mahajan, Peter Reiss, Azeem Shaikh, Hal White and Frank Wolak for comments that helped greatly improve this paper.

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:

$$Y = m_0(X) + \epsilon \tag{1}$$

where $E[\epsilon|X] \neq 0$. The parametric analysis of $m_0(x)$ 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 originally pointed out in Newey & Powell (2003), the nonparametric identification of $m_0(x)$ is hard to attain. Without a parametric assumption, the identification of $m_0(x)$ requires the availability of an instrument satisfying far more stringent conditions than the usual covariance restrictions in the linear case. A lack of identification of $m_0(x)$, however, does not preclude interesting characteristics of the model from being identified. Severini & Tripathi (2006, 2007), for example, argue that certain linear functionals of $m_0(x)$ will be identified even when $m_0(x)$ is not. In this paper I develop methods for the $\sqrt{n}$ estimation of such functionals without requiring $m_0(x)$ to be identified.

The linear functionals this paper focuses on are of the form $\int_{\mathcal{X}} v^*(x)m_0(x)dx$, where $v^*(x)$ is known and $\mathcal{X}$ is the support of X. For notational convenience we will denote:

$$\int_{\mathcal{X}} v^*(x)m_0(x)dx \equiv \langle v^*, m_0 \rangle \tag{2}$$

The canonical example of such a functional is consumer surplus, which is examined in the case where X is exogenous in Newey & McFadden (1994). In addition, by judiciously choosing $v^*(x)$, it is also possible to utilize estimators for $\langle v^*, m_0 \rangle$ to test for shape restrictions on $m_0(x)$ such as monotonicity, concavity and additive separability. For example, monotonicity of $m_0(x)$ will imply that $\langle v^*, m_0 \rangle$ must be positive for particular choices of $v^*(x)$. The framework developed in this paper allows us to test whether $\langle v^*, m_0 \rangle$ is indeed positive and hence examine if $m_0(x)$ is monotone. Examples of how to choose $v^*(x)$ to test for monotonicity, concavity and additive separability are discussed in Section 2.

Severini & Tripathi (2007) establish that a necessary condition for $\langle v^*, m_0 \rangle$ to be identified and estimable at a $\sqrt{n}$ rate is the existence of a function $\theta_0(z)$ of the instrument Z such that:

$$E[\theta_0(Z)|x] = \frac{v^*(x)}{f_X(x)} \tag{3}$$

where $f_X(x)$ is the density of X. It is important to note, as Severini & Tripathi (2006) argue, that (3) may be satisfied even when the model $m_0(x)$ is not identified. Since identification of $m_0(x)$ is difficult to both attain and test for, the estimator for $\langle v^*, m_0 \rangle$ I propose does not assume $m_0(x)$ is identified. Instead, I base estimation off the necessary condition for $\sqrt{n}$ estimability in (3). Under the exogeneity assumption on the instrument $E[\epsilon|Z] = 0$, we obtain through (1) and (3) that:

$$E[Y\theta_0(Z)] = E[m_0(X)\theta_0(Z)] = \langle v^*, m_0 \rangle \tag{4}$$

The estimator developed in this paper is therefore based on a sample analogue to (4) given by $n^{-1}\sum_i y_i\hat{\theta}_0(z_i)$, where $\hat{\theta}_0(z)$ is a nonparametric estimator for $\theta_0(z)$.

Unfortunately, the identification of $\theta_0(z)$ is even more problematic than the identification of $m_0(x)$. The nonparametric identification of $\theta_0(z)$ requires the existence of a unique function $\theta(z)$ satisfying (3). Such requirement is similar to what is necessary for identification of $m_0(x)$, the existence of a unique function $m(x)$ agreeing with the exogeneity assumption on Z: $E[Y - m(X)|Z] = 0$. The conditioning in (3), however, is on X instead of Z, and hence identification of $\theta_0(z)$ necessitates that there be no nonzero function $\theta_{\mathcal{N}}(z)$ such that $E[\theta_{\mathcal{N}}(Z)|x] = 0$. Intuitively, for $\theta_0(z)$ to be identified the regressor X must be able to explain all variation of the instrument Z. This requirement is problematic, as in most instances instruments posses variation that is unrelated to the endogenous regressor X.

The lack of identification of $\theta_0(z)$ presents considerable technical challenges, but it does not hinder the identification and $\sqrt{n}$ estimability of $\langle v^*, m_0 \rangle$. The function $\theta_0(z)$ is simply a nuisance parameter, and hence its identification is irrelevant to the final goal of estimating $\langle v^*, m_0 \rangle$. Any solution to (3) will provide a valid nuisance parameter for recovering $\langle v^*, m_0 \rangle$ through (4). Hence, a general estimation procedure should be robust to the lack of identification of both $m_0(x)$ and $\theta_0(z)$. I will therefore not assume that $\theta_0(z)$ is identified, but instead define the identified set (see Manski (2003)) as:

$$\Theta_0 = \left\{ \theta(z) \in \Theta : E[\theta(Z)|x] = \frac{v^*(x)}{f_X(x)} \right\} \tag{5}$$

where Θ is a nonparametric set of functions. Any function $\theta(z) \in \Theta_0$ provides a suitable nuisance parameter for recovering $\langle v^*, m_0 \rangle$.

I propose two different estimators for $\langle v^*, m_0 \rangle$ that are robust to both $m_0(x)$ and $\theta_0(z)$ not being identified. The first approach I consider is a minimum distance estimator $\hat{\theta}_0(z)$ similar to Newey & Powell (2003) and Ai & Chen (2003). If $\theta_0(z)$ is not identified, then it is still possible to show $\hat{\theta}_0(z)$ is contained in a shrinking neighborhood of Θ_0 with probability tending to one. The estimator $\hat{\theta}_0(z)$, however, will not converge to any particular element in Θ_0. This result is sufficient

for establishing that $n^{-1}\sum_i y_i\hat{\theta}_0(z_i)$ is a $\sqrt{n}$ consistent estimator for $\langle v^*, m_0\rangle$, but not for obtaining the asymptotic normality of $n^{-\frac{1}{2}}\sum_i(y_i\hat{\theta}_0(z_i) - \langle v^*, m_0\rangle)$. Under the additional assumption that $n^{-\frac{1}{2}}\sum_i(y_i\hat{\theta}_0(z_i) - \langle v^*, m_0\rangle)$ has an aymptotic distribution, however, we can resort to subsampling for determining appropriate critical values.

A second, more complex, estimation procedure is able to recover $\sqrt{n}$ asymptotic normality. In order to do so, it is necessary to construct an estimator $\tilde{\theta}_0(z)$ in such a way as to ensure that it converges to some unique $\theta_0(z) \in \Theta_0$. As a first step, results in Chernozhukov, Hong & Tamer (2007) are generalized to arbitrary metric spaces to obtain a consistent estimator $\hat{\Theta}_0$ for Θ_0. Furthermore, building off arguments in Ai & Chen (2003) it is possible to establish that $\hat{\Theta}_0$ converges to Θ_0 at a $o_p(n^{-\frac{1}{4}})$ rate with respect to a weak norm. I then show we can recover a unique element $\theta_0(z) \in \Theta_0$ by carefully choosing a unique element $\tilde{\theta}_0(z) \in \hat{\Theta}_0$. This procedure is analogous to a classical M-estimation problem where the domain Θ_0 is unknown but instead estimated by $\hat{\Theta}_0$. The approach developed in this paper provides a general useful technique for recovering nonparametric nuisance parameters when they are not identified. With this particular construction for $\tilde{\theta}_0(z)$, it is possible to show that $n^{-\frac{1}{2}}\sum_i(y_i\tilde{\theta}_0(z_i) - \langle v^*, m_0\rangle)$ is asymptotically normally distributed.

This paper is highly complementary to previous work in Severini & Tripathi (2006, 2007). The authors are the first to explore conditions for the identification of $\langle v^*, m_0\rangle$ when $m_0(x)$ is not identified and to derive efficiency bounds for its estimation. They, however, provide no estimation procedures. Darolles, Florens & Renault (2003) derive asymptotically normal estimators for $\langle v^*, m_0\rangle$ that assume $m_0(x)$ is identified and are possibly asymptotically biased. Within the larger nonparametric IV literature, Newey & Powell (2003) and Hall & Horowitz (2005) propose consistent estimators for $m_0(x)$, while Horowtiz (2007) derives the asymptotic distribution for estimators of $m_0(x)$. Santos (2007) proposes test statistics for inference when the model is partially identified. Ai & Chen (2003) and Blundell, Chen & Kristensen (2004) examine the properties of a semiparametric specification. In related work, Newey, Powell & Vella (1999), Chesher (2003, 2005, 2007), Imbens & Newey (2006) and Schennach, Chalak & White (2007) explore estimation and identification in triangular systems, while Chalak & White (2006) and White & 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 choices for $v^*(x)$ that allow for interesting inference on $m_0(x)$ by examining $\langle v^*, m_0\rangle$. Section 3 develops the two proposed estimators, while Section 4 analyzes their performance in a Monte Carlo Study. Section 5 briefly concludes. All proofs are contained in a mathematical appendix.

2 Motivating Examples

In this section I provide examples of choices of $v^*(x)$ for which the functional $\langle v^*, m_0\rangle$ is of interest. The first example is that of consumer surplus. Let $m_0(x)$ be an inverse demand function, set $v^*(x) = 1\{0 \leq x \leq q^*\}$ for some quantity q^* and denote the corresponding market clearing price by p^*. Then, we can write consumer surplus for $m_0(x)$ as:

$$\int_0^{q^*} m_0(x)dx - p^*q^* = \int_{\mathcal{X}} v^*(x)m_0(x)dx - p^*q^* = \langle v^*, m_0\rangle - p^*q^* \tag{6}$$

If the market clearing price p^* corresponding to q^* is observable, then the functional $\langle v^*, m_0\rangle$ determines consumer surplus in (6). Hence, the estimators for $\langle v^*, m_0\rangle$ proposed in this paper will allow us to build confidence intervals for consumer surplus when it is identified. By the same arguments, choosing $v^*(x) = 1\{0 \leq x \leq q^*\}$ can also be used to perform inference on firm profit when $m_0(x)$ is a cost function.

A judicious choice of $v^*(x)$ also enables us to use functionals of the form $\langle v^*, m_0\rangle$ to test for shape restrictions on $m_0(x)$. In Lemmas 2.1, 2.2 and 2.3 I provide examples of choices of $v^*(x)$ that allow for tests of monotonicity, concavity and additive separability. These Lemmas assume compact support on $[-\pi, \pi]$ because they utilize Fourier expansions. Knowledge of $\mathcal{X}$ for their implementation is unnecessary, as it is always possible to map a known connected subset of $\mathcal{X}$ into $[-\pi, \pi]$ and test for shape restrictions in such subset. Similar results for other set of restrictions and choices of basis can be derived. Such tests may have better power properties. Lemmas 2.1, 2.2 and 2.3 are meant for illustrative purposes, and are unlikely to be the optimal choices of $v^*(x)$ for testing their respective hypotheses.

Lemma 2.1. *Assume X has compact support on $[-\pi, \pi]$, and let $v_i^*(x) = \sin(ix)$. If $m_0(x)$ is weakly increasing in x, then $sign\{\langle v_i^*, m_0\rangle\} = (-1)^{i+1}$ for all integer i.*

Lemma 2.2. *Assume X has compact support on $[-\pi, \pi]$, and let $v_i^*(x) = \cos(ix)$. If $m_0(x)$ is concave and differentiable, then $sign\{\langle v_i^*, m_0\rangle\} = (-1)^{i+1}$ for all integer i.*

Lemma 2.3. *Assume $X = (X_1, X_2)$ has compact support on $[-\pi, \pi]^2$ and let $v_{ij}^* = \sin(ix_1)\sin(jx_2)$. If $m_0(x) = m_{01}(x_1) + m_{02}(x_2)$, then $\langle v_{ij}^*, m_0\rangle = 0$ for all integer i, j.*

A disadvantage of using functionals of the form $\langle v^*, m_0\rangle$ to test for shape restrictions is the likely lack of consistency of the test. For example, using Lemma 2.1 we may test whether $m_0(x)$ is weakly increasing by examining whether $\int_{-\pi}^{\pi} \sin(x)m_0(x)dx \geq 0$. Numerous non-increasing functions $m_0(x)$

satisfy $\int_{-\pi}^{\pi} \sin(x) m_0(x) dx < 0$, in which case a test based off $\int_{-\pi}^{\pi} \sin(x) m_0(x) dx$ will be consistent. On the other hand, there are also non-increasing functions for which $\int_{-\pi}^{\pi} \sin(x) m_0(x) dx \geq 0$, and hence such functions will asymptotically not be rejected. The lack of consistency simply reflects that the space of alternatives is very large. In a nonparametric setting it is not possible to fully characterize a shape restriction such as monotonicity by a single, or even finite, number of moment restrictions. By increasing the number of functionals $\langle v^*, m_0 \rangle$ we examine, however, it is possible to reduce the set of functions that do not satisfy the shape restriction we test for, but do agree with the implied moment restrictions on $\langle v^*, m_0 \rangle$.

3 Estimation

In this section I develop two $\sqrt{n}$ consistent estimators for $\langle v^*, m_0 \rangle$ that differ on how the nuisance parameter $\theta_0(z)$ is estimated. Section 3.1 introduces the general framework, while Sections 3.2 and 3.3 explore the properties of an estimator that recovers $\theta_0(z)$ using a series procedure similar to Newey & Powell (2003) and Ai & Chen (2003). This estimator provides insight into the nature of the estimation problem by illustrating the challenges in utilizing a regular minimum distance framework when the parameter of interest is not identified. I show that even if $\theta_0(z)$ is not identified, this approach will yield a $\sqrt{n}$ consistent estimator for $\langle v^*, m_0 \rangle$, though one that will not be asymptotically normally distributed. In contrast, in Sections 3.4, 3.5 and 3.6 I examine a more complex estimator for the nuisance parameter that converges to a unique element of Θ_0. This second procedure yields a $\sqrt{n}$ asymptotically normal estimator for $\langle v^*, m_0 \rangle$.

3.1 General Framework

The principal challenge in recovering $\langle v^*, m_0 \rangle$ consists in obtaining a first stage estimator for $\theta_0(z)$ solving (3). Once such an estimator is available, we can estimate $\langle v^*, m_0 \rangle = E[Y\theta_0(Z)]$ by the sample analogue $n^{-1} \sum_i y_i \hat{\theta}_0(z_i)$. Assuming a unique solution to (3) is too strict a requirement and we therefore let the nuisance parameter be partially identified. The identified set can be characterized as the set of minimizers to a criterion function, as in Chernozhukov, Hong & Tamer (2007) and Romano & Shaikh (2006). Hence, we define:

$$\Theta_0 = \{\theta(z) \in \Theta : Q(\theta) = 0\} \qquad Q(\theta) = E\left[\left(E[v^*(X) - \theta(Z) f_X(X) | X]\right)^2\right] \tag{7}$$

where Θ is a nonparametric set of functions. In order to attain consistency and uniform behavior of the empirical process on the parameter space, I will require Θ to be a smooth set of functions.

Let $Z \in \Re^{d_z}$ and define λ to be a d_z dimensional vector of nonnegative integers, also known as a multi-index. In addition, define $|\lambda| = \sum_i^{d_z} \lambda_i$ and let $D^\lambda \theta(z) = \partial^{|\lambda|}\theta(z)/\partial z_1^{\lambda_1} \dots \partial z_{d_z}^{\lambda_{d_z}}$. For $\underline{\omega}$ the greatest integer smaller than ω, and $\mathcal{Z}$ the support of Z, define the norm:

$$||\theta||_{\Lambda^\omega} = \max_{|\lambda| \leq \underline{\omega}} \sup_{z \in \mathcal{Z}} |D^\lambda \theta(z)| + \max_{|\lambda| = \underline{\omega}} \sup_{z \neq z'} \frac{|D^\lambda \theta(z) - D^\lambda \theta(z')|}{||z - z'||^{\omega - \underline{\omega}}} \tag{8}$$

We denote the set of functions that is bounded in this norm by $\Lambda_C^\omega(\mathcal{Z}) = \{\theta(z) : ||\theta||_{\Lambda^\omega} \leq C\}$. The functions $\theta(z) \in \Lambda_C^\omega(\mathcal{Z})$ have partial derivatives up to order $\underline{\omega}$ uniformly bounded, and partial derivatives of order $\underline{\omega}$ Lipschitz of order $\omega - \underline{\omega}$. Throughout the paper I will assume $\Theta = \Lambda_C^\omega(\mathcal{Z})$ for a particular ω and the existence of a $\theta(z) \in \Lambda_C^\omega(\mathcal{Z})$ satisfying $E[\theta(Z)|x] = v^*(x)/f_X(x)$. While these smoothness conditions are often used in nonparametric estimation, in other contexts the requirement often is that the "true" model is sufficiently smooth. These assumptions are imposed on $m_0(x)$, for example, in Newey & Powell (2003) and Santos (2007). In our problem, however, there is no "true" model. All we require is that there exist some solution to (3) that is sufficiently smooth, because any solution can be used to estimate $\langle v^*, m_0 \rangle$.

Both estimation strategies I develop use the characterization of Θ_0 in (7). We will employ a sample analogue $Q_n(\theta)$ to the criterion function $Q(\theta)$ given by:

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^n \hat{m}^2(x_i, \theta) \qquad \hat{m}(x_i, \theta) = \hat{E}[v^*(X) - \theta(Z)\hat{f}_X(X)|x_i] \tag{9}$$

where $\hat{E}[v^*(X) - \theta(Z)\hat{f}_X(X)|x_i]$ and $\hat{f}_X(x_i)$ are nonparametric estimators of $E[v^*(X) - \theta(Z)\hat{f}_X(X)|x_i]$ and $f_X(x_i)$ respectively. For $\hat{E}[v^*(X) - \theta(Z)\hat{f}_X(X)|x_i]$ I will use a traditional series estimator. Assume $X \in \Re^{d_x}$ and let $\{p_j(x)\}_{j=1}^\infty$ be a sequence of known basis functions. Denote the vector of the first k_n terms in the basis by $p^{k_n}(x) = (p_1(x), \dots, p_{k_n}(x))$ and the matrix of $p^{k_n}(x)$ evaluated at the sample by $P = (p^{k_n}(x_1), \dots, p^{k_n}(x_n))'$. The nonparametric estimator $\hat{E}[v^*(X) - \theta(Z)\hat{f}_X(X)|x_i]$ is then given by the linear regression of $(v^*(x_1) - \theta(z_1)\hat{f}_X(x_1), \dots, v^*(x_n) - \theta(z_n)\hat{f}_X(x_n))$ on P. Hence, we define:

$$\hat{E}[v^*(X) - \theta(Z)\hat{f}_X(X)|x] = p^{k_n'}(x)(P'P)^{-1} \sum_{i=1}^n p^{k_n}(x_i)(v^*(x_i) - \theta(z_i)\hat{f}_X(x_i)) \tag{10}$$

If $k_n \to \infty$ at the appropriate rate and in addition the basis $\{p_j(x)\}_{j=1}^\infty$ can approximate the true conditional expectations $E[v^*(X) - \theta(Z)\hat{f}_X(X)|x]$ arbitrarily well, then $\hat{E}[v^*(X) - \theta(Z)\hat{f}_X(X)|x]$ provides a consistent estimator for $E[v^*(X) - \theta(Z)\hat{f}_X(X)|x]$ under a variety of norms. This series estimator is studied in detail in Newey (1997) and Ai & Chen (2003).

For the nonparametric estimator $\hat{f}_X(x_i)$, I will use a Nadaraya-Watson kernel estimator. Suppose

$X \in \Re^{d_x}$, then for $K(u)$ the kernel and h the bandwidth, the kernel estimator of $f_X(x_i)$ is given by:

$$\hat{f}_X(x_i) = \frac{1}{nh^{d_x}} \sum_{j \neq i} K\left(\frac{x_i - x_j}{h}\right) \tag{11}$$

In certain cases it will necessary to resort to higher order kernels in order to attain the appropriate rates of convergence. The kernel $K(u) : \Re^{d_x} \to \Re$ is of order k if $\int_{\Re^{d_x}} K(u)du = 1$, it is bounded and for all multi-indices $|\lambda| \leq \underline{k}$:

$$\max_{|\lambda| \leq \underline{k}} \int_{\Re^{d_x}} ||u||^{k-\underline{k}} |u_1|^{\lambda_1} \dots |u_{d_x}|^{\lambda_{d_x}} |K(u)|du < \infty \qquad \int_{\Re^{d_x}} u_1^{\lambda_1} \dots u_{d_x}^{\lambda_{d_x}} K(u)du = 0 \;\; \forall |\lambda| \leq \underline{k} \tag{12}$$

The assumption on the kernel $K(u)$ and the bandwidth h are stated in Section 3.3.

3.2 Estimation Strategy for $\sqrt{n}$ Consistency

If $\theta_0(z)$ were identified, then a natural estimator for the nuisance parameter would be the minimizer of $Q_n(\theta)$ over the parameter space Θ. Minimizing over the whole parameter space, however, is not only computationally challenging but can also lead to slow rates of convergence. For this reason it is advantageous to minimize over a sieve $\Theta_n \subseteq \Theta$ that grows to be dense in Θ. The resulting estimator for $\theta_0(z)$ is then defined by:

$$\hat{\theta}(z) \in \arg\min_{\Theta_n} Q_n(\theta) \tag{13}$$

If $\theta_0(z)$ is identified, then $\hat{\theta}(z)$ can be consistent under a variety of norms. In this case, it is possible to establish the asymptotic normality of the statistic $n^{-\frac{1}{2}} \sum_i (y_i \hat{\theta}(z_i) - \langle v^*, m_0 \rangle)$.

As argued, however, the identification of $\theta_0(z)$ is not a tenable assumption. Hence, I now study the behavior of the estimator $n^{-1} \sum_i y_i \hat{\theta}(z_i)$ for $\langle v^*, m_0 \rangle$ when identification of $\theta_0(z)$ breaks down. As in most two stage estimation problems where the first stage is nonparametric, the nuisance parameter estimator must satisfy $||\hat{\theta} - \theta_0|| = o_p(n^{-\frac{1}{4}})$ in order for the second stage to be $\sqrt{n}$ consistent. Because the first stage is nonparametric, it is important to specify the norm under which $||\hat{\theta} - \theta_0|| = o_p(n^{-\frac{1}{4}})$. Unlike the parametric case, different choices of norm often imply drastically different rates of convergence. Ai & Chen (2003) show we can focus on the fairly weak norm $||\cdot||_w$, which in the present context is given by:

$$||\theta||_w = \left[E\left[(E[\theta(Z)|X])^2 f_X^2(X)\right]\right]^{\frac{1}{2}} \tag{14}$$

Interestingly, the norm $||\cdot||_w$ makes the whole identified set Θ_0 an equivalence class, since for any $\theta_0(z)$, $\theta_0'(z) \in \Theta_0$ we have $||\theta_0 - \theta_0'||_w^2 = E[(v^*(X) - v^*(X))^2] = 0$. Therefore, when establishing

$||\hat{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$, identification of $\theta_0(z)$ is not relevant because for any $\theta_0(z)$, $\theta_0'(z) \in \Theta_0$, we have $||\hat{\theta} - \theta_0||_w = ||\hat{\theta} - \theta_0'||_w$. The distance between $\hat{\theta}(z)$ and any point $\theta_0(z) \in \Theta_0$ is the same.

The requirement $||\hat{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$ is often a necessary but not sufficient condition for obtaining the asymptotic normality of a two stage estimator. Unfortunately, while the identification of $\theta_0(z)$ is inconsequential for establishing $||\hat{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$, the same is not true regarding the asymptotic normality of $n^{-\frac{1}{2}} \sum_i (y_i \hat{\theta}(z_i) - \langle v^*, m_0 \rangle)$. The asymptotic behavior of two stage estimators is analyzed through an equality that in the present context takes the form:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(y_i \hat{\theta}(z_i) - \langle v^*, m_0 \rangle \Big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(y_i \hat{\theta}(z_i) - E[Y \hat{\theta}(Z)] \Big) + \sqrt{n} E[Y(\hat{\theta}(Z) - \theta_0(Z))] \tag{15}$$

The term $\sqrt{n} E[Y(\hat{\theta}(Z) - \theta_0(Z))]$ in (15) captures the contribution to the asymptotic distribution of not knowing $\theta_0(z)$ and having to estimate it instead. Under identification and the stochastic equicontinuity of the empirical process, it is possible to analyze the first term in (15) by deriving:

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} (y_i \hat{\theta}(z_i) - E[Y \hat{\theta}(Z)]) = n^{-\frac{1}{2}} \sum_{i=1}^{n} (y_i \theta_0(z_i) - E[Y \theta_0(Z)]) + o_p(1) \tag{16}$$

Hence, the term $n^{-\frac{1}{2}} \sum_i (y_i \hat{\theta}(z_i) - E[Y \hat{\theta}(Z)])$ captures the uncertainty in estimating $\langle v^*, m_0 \rangle$, as if $\theta_0(z)$ were known. In order for (16) to hold, however, $\hat{\theta}(z)$ must be a consistent estimator for $\theta_0(z)$ under a norm stronger than $||\cdot||_w$. This dependency is reflected in that the right hand side of (16) will have a different asymptotic distribution when evaluated at alternative $\theta_0(z)$, $\theta_0'(z) \in \Theta_0$. Thus, for (16) to hold, $\hat{\theta}(z)$ must converge to $\theta_0(z)$ under a norm that is able to differentiate the elements in Θ_0. Without such convergence, the statistic $n^{-\frac{1}{2}} \sum_i (y_i \hat{\theta}(z_i) - \langle v^*, m_0 \rangle)$ will not be asymptotically normally distributed.

In a general setting, the results in Santos (2007) imply $\hat{\theta}(z)$ will not converge to a unique element $\theta_0(z) \in \Theta_0$. Santos (2007) studies a statistic $\tilde{Q}_n(\theta)$ that is analogous to $Q_n(\theta)$ but utilizes kernels instead of series estimators to estimate conditional expectations. This statistic satisfies $\tilde{Q}_n(\hat{\theta}) \xrightarrow{\mathcal{L}} \min_{\Theta_0} G(\theta)$, where $G(\theta)$ is a Gaussian process on $l^\infty(\Theta_0)$, the space of bounded functionals on Θ_0. If $\hat{\theta}(z)$ converged to a specific $\theta_0(z) \in \Theta_0$, however, then $\tilde{Q}_n(\hat{\theta})$ would be asymptotically normally distributed. Thus, the asymptotic distribution of $\tilde{Q}_n(\hat{\theta})$ indicates $\hat{\theta}(z)$ fails to converge to a specific $\theta_0(z) \in \Theta_0$. See Theorem 3.2 and Corollary 3.1 in Santos (2007) for a detailed exposition.

3.3 Establishing $\sqrt{n}$ Consistency

I now formalize the discussion in the previous section. First, I establish that $||\hat{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$ even when $\theta_0(z)$ is not identified. As argued, this result is not sufficient for deriving the asymptotic

normality of the statistic $n^{-\frac{1}{2}}\sum_i(y_i\hat{\theta}(z_i) - \langle v^*, m_0\rangle)$. It is still possible, however, to show that $n^{-1}\sum_i(y_i\hat{\theta}(z_i) - \langle v^*, m_0\rangle) = O_p(n^{-\frac{1}{2}})$. If in addition $n^{-\frac{1}{2}}\sum_i(y_i\hat{\theta}(z_i) - \langle v^*, m_0\rangle)$ actually converges in distribution, then subsampling provides a simple procedure for constructing confidence intervals for $\langle v^*, m_0\rangle$ by using the estimator $n^{-1}\sum_i y_i\hat{\theta}(z_i)$.

Assumptions 1-5 are sufficient for showing $||\hat{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$. Assumptions 2(ii), 3(iii), 4(ii), 5(i) and 5(iv) are meant to hold for some δ_0 satisfying $0 < \delta_0 < (2\gamma - d_x)/16\gamma$.

Assumption 1: (i) $\{y_i, x_i, z_i\}_{i=1}^n$ are i.i.d. generated according to (1) with $E[Y^2] < \infty$, $E[\epsilon|X] \neq 0$ and $E[\epsilon|Z] = 0$; (ii) $\mathcal{X}$ is convex and compact with nonempty interior; (iii) $f_X(x)$ is bounded and bounded away from 0.

Assumption 2: (i) The smallest and largest eigenvalues of $E[p^{k_n}(X)p^{k_n'}(X)]$ are bounded and bounded away from zero uniformly in k_n; (ii) For any $\nu(x) \in \Lambda_C^\gamma(\mathcal{X})$ with $\gamma > d_x/2$ there exists $p^{k_n'}(x)\pi \in \Lambda_C^\gamma(\mathcal{X})$ such that $||\nu - p^{k_n'}\pi||_\infty = O\left(k_n^{-\frac{\gamma}{d_x}}\right)$ uniformly in $\Lambda_C^\gamma(\mathcal{X})$ and $k_n^{-\frac{\gamma}{d_x}} = o(n^{-\frac{1}{4}-\frac{\delta_0}{2}})$.

Assumption 3: (i) $\Theta \equiv \Lambda_C^\omega(\mathcal{Z})$ for $\omega > d_z/2$ and $\Theta_0 \neq \emptyset$; (ii) $\Theta_n \subseteq \Theta$ are closed under $||\cdot||_\infty$ and $\dim(\Theta_n) = k_{1n}$; (iii) $\sup_\Theta \inf_{\Theta_n} ||\theta - \theta_n||_\infty = o(n^{-\frac{1}{4}-\frac{\delta_0}{2}})$; (iv) $v^*(x) - E[\theta(Z)|x]f_X(x) \in \Lambda_C^\gamma(\mathcal{X})$ with $\gamma > d_x/2$ for all $\theta(z) \in \Theta$.

Assumption 1 states the distributional assumptions on (Y, X, Z). Assumptions 2(i) and 2(ii) are standard in the use of series estimators for approximating conditional mean functions, while the requirement $k_n^{-\frac{\gamma}{d_x}} = o(n^{-\frac{1}{4}-\frac{\delta_0}{2}})$ is necessary to ensure $\hat{E}[v^*(X) - \theta(Z)f_X(X)|x]$ converges to $E[v^*(X) - \theta(Z)f_X(X)|x]$ at the appropriate rate. Assumption 3(i) restricts Θ to a smooth set of functions and requires that there is at least one $\theta_0(z) \in \Theta$ that is also a solution to (3). Assumptions 3(ii) and 3(iv) are common in the sieves literature, while Assumption 3(iii) is necessary to guarantee the bias present in $\hat{\theta}(z)$ from optimizing over the approximating space Θ_n instead of Θ decreases at the right rate. The approximation rates of numerous sieves such as Fourier and Hermite series for the space $\Lambda_C^\omega(\mathcal{Z})$ are well known. See Chen (2006) for an excellent reference.

In order to show $||\hat{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$ I also need Assumptions 4 and 5. In Assumption 4, $\xi_{jn} = \sup_{x,|\lambda|=j} ||D^\lambda p^{k^n}(x)||$ and $N(\epsilon, \Theta_n, ||\cdot||)$ is the minimal number of balls of size ϵ under the norm $||\cdot||$ that are necessary to cover Θ_n.

Assumption 4: (i) $k_n \geq 1 + k_{1n}$, $k_{1n} \to \infty$ and $k_n/n \to 0$; (ii) $k_{1n} \times \log n \times \xi_{jn}^2 \times n^{-\frac{1}{2}+\delta_0} = o(1)$ for $j \in \{0, 1\}$; (iii) $\log(N(\epsilon, \Theta_n, ||\cdot||_\infty)) \leq C \times k_{1n} \times \log(k_{1n}/\epsilon)$.

Assumption 5: (i) The kernel $K(u)$ is of order $k > d_x$; (ii) The bandwidth h satisfies $h \asymp n^{-\nu}$ with $(2k)^{-1} < \nu < (2d_x)^{-1}$; (iii) The density $f_{XZ}(x, z) \in \Lambda_C^k(\mathcal{X})$ for all z; (iv) $k_n \times n^{-(1-\nu d_x)+\frac{1}{2}+\delta_0} \to 0$.

Assumption 4 helps us control the uniform rate of convergence in Θ_n of $\hat{E}[v^*(X)-\theta(Z)f_X(X)|x]$ to $E[v^*(X) - \theta(Z)f_X(X)|x]$. Assumption 4(ii) prevents the sieve Θ_n from growing too fast. This is at tension with Assumption 3(ii), which requires the approximation error to vanish sufficiently fast. Such tension can be resolved by assuming additional smoothness in Θ. The requirement $\log(N(\epsilon, \Theta_n, ||\cdot||_\infty)) \leq C \times k_{1n} \times \log(k_{1n}/\epsilon)$ in Assumption 4(iii) is satisfied by all standard sieves. Finally, Assumption 5 states the requirements for a higher order kernel to satisfy $\sup_{\Theta_n} n^{-1}\sum_i(\hat{E}[\theta(Z)(\hat{f}_X(X) - f_X(X))])^2 = o_p(n^{-\frac{1}{2}-\frac{\delta_0}{2}})$. Any density estimator meeting this rate requirement can also be used in the analysis.

Assumptions 1-5 allow us to derive the required rate of convergence.

Theorem 3.1. *Under Assumptions 1, 2(i)-(ii), 3(i)-(iv), 4(i)-(iii), 5(i)-(iv),* $||\hat{\theta}-\theta_0||_w = o_p(n^{-\frac{1}{4}})$.

Theorem 3.1 is not sufficient for showing $n^{-1}\sum_i(y_i\hat{\theta}(z_i) - \langle v^*, m_0\rangle) = O_p(n^{-\frac{1}{2}})$. Further assumptions are needed that require the introduction of additional notation. Let $\overline{V}$ be the closure of the linear span of Θ under $||\cdot||_w$. The space $\overline{V}$ is a Hilbert Space with inner product given by:

$$\langle\theta_1, \theta_2\rangle_w = E\left[E[\theta_1(Z)|X]E[\theta_2(Z)|X]f_X^2(X)\right] \tag{17}$$

The linear functional $E[Y\theta(Z)] : \overline{V} \to \Re$ is continuous under $||\cdot||_w$ since by Cauchy-Schwarz and Jensen's inequality, the exogeneity of the instrument and $f_X(x)$ bounded below we have $|E[Y\theta(Z)]| \leq M\left[E[m_0^2(X)]\right]^{\frac{1}{2}}||\theta||_w$ for some $M > 0$. Therefore, the Riesz Representation theorem implies the existence of a $\tilde{v}(z) \in \overline{V}$ such that:

$$E[Y\theta(Z)] = E\left[E[\tilde{v}(Z)|X]E[\theta(Z)|X]f_X^2(X)\right] = \langle\tilde{v}, \theta\rangle_w \tag{18}$$

By equation (18), $\langle\tilde{v}, \theta_0\rangle_w$ provides an alternative representation for the parameter of interest $\langle v^*, m_0\rangle$. In analyzing the asymptotic behavior of our estimator, the representation $\langle\tilde{v}, \theta_0\rangle_w$ can sometimes be the most convenient as it is solely in terms of functions of Z.

We can now introduce the final regularity conditions necessary to establish the $\sqrt{n}$ consistency of $n^{-1}\sum_i y_i\hat{\theta}(z_i)$. Assumption 3(v) requires $\tilde{v}(z)$ to not only be in the closure of the linear span of Θ, but in Θ itself. The smoothness assumption 3(vi) is necessary to ensure $\hat{E}[E[\theta(Z)|X]|x]$ and $\hat{E}[E[\theta(Z)|X]f_X(X)|x]$ converge to $E[\theta(Z)|x]$ and $E[\theta(Z)|X]f_X(X)$ respectively at the appropriate rate under $||\cdot||_{\mathcal{L}^2}$. Assumption 4(iv) guarantees this convergence also holds under $||\cdot||_\infty$. Assumption 5(v) is satisfied by a wide variety of kernels, and it guarantees that $\sup_x|\hat{f}_X(x) - f_X(x)| = o_p(1)$.

Assumption 3: (v) $\tilde{v}(z) \in \Theta$; (vi) $E[\theta(Z)|x]f_X(x)$, $E[\theta(Z)|x] \in \Lambda_C^\gamma(\mathcal{X})$ for all $\theta \in \Theta$.

ASSUMPTION 4: (iv) $k_n \xi_{0n}^2 n^{-1} \to 0$.

ASSUMPTION 5: (v) The kernel $K(u)$ has an absolutely integrable Fourier Transform.

The previous assumptions are standard when employing nonparametric estimators. In contrast, Assumption 6(i) is particular to the partially identified setting. When deriving the asymptotic distribution of nonparametric estimators, it is often necessary to evaluate the derivative of the sample criterion function at the point $\hat{\theta}(z)$. Implicitly, this dictates $\hat{\theta}(z)$ lie in the interior of Θ_n. This requirement is easily met when $\theta_0(z)$ is identified. If $\theta_0(z)$ is an interior point of Θ, then the consistency of $\hat{\theta}(z)$ implies it will lie in the interior of Θ_n with probability tending to one. In a partially identified setting, however, a complication arises as Θ_0 will not lie in the interior of Θ unless it is a singleton.[1] Assumption 6(i) allows us to nonetheless calculate the derivative of the sample criterion function at the point $\hat{\theta}(z)$. Intuitively, Assumption 6(i) demands that $\hat{\theta}(z)$ converge to $\Theta_0 \cap \Theta^\circ$ for Θ° the interior of Θ.

ASSUMPTION 6: (i) Let $u(z) = \pm\tilde{v}(z)$ and $u_n(z) = \arg\min_{\Theta_n} ||u - \theta_n||_\infty$, then there is some $\epsilon_n = o(n^{-\frac{1}{2}})$ such that $P\left(\hat{\theta}(z) + \epsilon_n u_n(z) \in \Theta_n\right) \to 1$.

With the stated assumptions we can now establish the main theorem in this section:

Theorem 3.2. *If Assumptions 1, 2(i)-(ii), 3(i)-(vi), 4(i)-(v), 5(i)-(v) and 6(i) hold, then it follows that* $n^{-1}\sum_i \left(y_i\hat{\theta}(z_i) - \langle v^*, m_0\rangle\right) = O_p(n^{-\frac{1}{2}})$.

Theorem 3.2 does not imply $n^{-\frac{1}{2}}\sum_i(y_i\hat{\theta}(z_i) - \langle v^*, m_0\rangle)$ actually has a limiting distribution. As is shown in the next section, if in addition $\hat{\theta}(z)$ converged to a specific $\theta_0(z) \in \Theta_0$, then $n^{-\frac{1}{2}}\sum_i(y_i\hat{\theta}(z_i) - \langle v^*, m_0\rangle)$ would be asymptotically normally distributed. A reasonable conjecture therefore is that the statistic $n^{-\frac{1}{2}}\sum_i(y_i\hat{\theta}(z_i) - \langle v^*, m_0\rangle)$ converges in distribution to a mixture of normals, where the mixture is taken over Θ_0. The derivation of this asymptotic distribution is a considerable technical challenge beyond the scope of the present paper. Instead, in the next section we focus in deriving a first stage estimator that delivers asymptotic normality.

3.4 Estimation Strategy to Recover Asymptotic Normality

The principal drawback of utilizing $\hat{\theta}(z)$ as the first stage estimator is its failure to converge to a particular element $\theta_0(z) \in \Theta_0$. As a result, the second stage estimator $n^{-\frac{1}{2}}\sum_i(y_i\hat{\theta}(z_i) - \langle v^*, m_0\rangle)$

[1] If $\theta_0(z), \theta_0'(z) \in \Theta_0$, then $\theta_{\mathcal{N}}(z) = \theta_0(z) - \theta_0'(z)$ satisfies $E[\theta_{\mathcal{N}}(Z)|x] = 0$ and $\lambda\theta_{\mathcal{N}} \in \Theta$ for λ small enough. We can choose λ such that $||\theta_0 + \lambda\theta_{\mathcal{N}}||_{\Lambda^w} = C$ for $\Theta = \{\theta : ||\theta||_{\Lambda^w} \leq C\}$. Hence $\theta_0(z) + \lambda\theta_{\mathcal{N}}$ is in both Θ_0 and the boundary of Θ.

is not asymptotically normally distributed. In this section we remedy this problem by constructing a different first stage estimator that converges to a unique element $\theta_0(z) \in \Theta_0$.

3.4.1 General Estimation Procedure

Intuitively, if we could observe Θ_0, then we would simply choose a unique element $\theta_0(z)$ from it and employ it to estimate $\langle v^*, m_0 \rangle$. The set Θ_0, however, is itself identified and can therefore be estimated. Hence, a natural estimation procedure for a unique element $\theta_0(z) \in \Theta_0$ is given by:

1. Construct a consistent estimator $\hat{\Theta}_0$ for Θ_0.
2. Choose $\tilde{\theta}(z) \in \hat{\Theta}_0$ in such a way as to ensure it converge to a unique element $\theta_0(z) \in \Theta_0$.

In order to implement this strategy it is necessary to develop a set estimation procedure for arbitrary metric spaces. Chernozhukov, Hong & Tamer (2007) provide a method for estimating sets in $\Re^d$, but since Θ_0 is an infinite dimensional set of functions their results do not apply. Their estimation framework is generalized in the appendix to the present context. The estimator $\hat{\Theta}_0$ for Θ_0 is defined by:

$$\hat{\Theta}_0 = \{\theta(z) \in \Theta_n : Q_n(\theta) \leq \epsilon_n\} \tag{19}$$

for $\epsilon_n \searrow 0$ at an appropriate rate. Similarly to the parametric case, we examine the set of near-minimizers on Θ_n instead of the exact minimizers. The requirement $\epsilon_n \searrow 0$ ensures that in the limit $\hat{\Theta}_0$ only includes those $\theta_n(z) \in \Theta_n$ that approximate elements $\theta_0(z) \in \Theta_0$ well. On the other hand, if $\epsilon_n \searrow 0$ slowly enough, then we can ensure $\hat{\Theta}_0$ includes all such $\theta_n(z)$. Sieves are employed in (20) not only for computational purposes, but also to allow us to attain the necessary rates of convergence. The requirements on the rate at which $\epsilon_n \searrow 0$ are different than in the parametric case.

Under certain regularity conditions it is possible to obtain the consistency of $\hat{\Theta}_0$ for Θ_0 under a variety of norms. I will focus on the family of Haussdorf norms, which is defined by:

$$d_H(\Theta_1, \Theta_2, ||\cdot||) = \max\{h(\Theta_1, \Theta_2), h(\Theta_2, \Theta_1)\} \qquad h(\Theta_1, \Theta_2) = \sup_{\theta_1(z) \in \Theta_1} \inf_{\theta_2(z) \in \Theta_2} ||\theta_1 - \theta_2|| \tag{20}$$

Hence, $\hat{\Theta}_0$ provides a consistent estimator for Θ_0 under the Haussdorf norm if both the maximal approximation error of $\hat{\Theta}_0$ by Θ_0 and of Θ_0 by $\hat{\Theta}_0$ converges to 0 in probability. Unlike the parametric case, however, using different norms for the projections in (20) implies significantly different Haussdorf norms. For example, since $||\cdot||_w$ makes Θ_0 an equivalence class, Theorem 3.1 actually implies

that $d_H(\hat{\theta}(z), \Theta_0, ||\cdot||_w) \xrightarrow{p} 0$. On the other hand, $\hat{\theta}(z)$ will not be a consistent estimator for Θ_0 under $d_H(\cdot,\cdot,||\cdot||_\infty)$ unless Θ_0 is a singleton. In Section 3.5 I establish $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_w) = o_p(n^{-\frac{1}{4}})$, so that $\hat{\Theta}_0$ converges to Θ_0 at the required rate. Equally important, however, $\hat{\Theta}_0$ also satisfies $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_\infty) \xrightarrow{p} 0$. The consistency of $\hat{\Theta}_0$ to Θ_0 under a norm that can differentiate the elements in Θ_0 is necessary for recovering a unique element $\theta_0(z) \in \Theta_0$.

Given the estimator $\hat{\Theta}_0$ for Θ_0, the second challenge consists in selecting an element $\tilde{\theta}(z) \in \hat{\Theta}_0$ in such a way as to ensure it converges to a unique element $\theta_0(z) \in \Theta_0$. For this purpose I derive a generalization of extremum estimators to problems where the parameter space is unknown but consistently estimated. Suppose $M(\theta)$ is a population criterion function attaining a unique minimum on Θ_0 and $M_n(\theta)$ is the finite sample analogue. Intuitively, if $\theta_0(z)$ is the unique minimizer of $M(\theta)$ on Θ_0, then the minimizer of $M_n(\theta)$ over the estimated parameter space $\hat{\Theta}_0$ should provide a consistent estimator for $\theta_0(z)$. Theorem 3.3 formalizes this argument.

Theorem 3.3. *If (i) Θ_0 is a closed subset of a compact set Θ such that $M(\theta)$ has a unique minimum on Θ_0 at $\theta_0(z)$, (ii) $\hat{\Theta}_0 \subseteq \Theta$ satisfies $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||) \xrightarrow{p} 0$, (iii) $M_n(\theta)$ and $M(\theta)$ are continuous in Θ, and (iv) $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$. Then $\hat{\theta}(z) = \arg\min_{\theta\in\hat{\Theta}_0} M_n(\theta)$ satisfies $||\hat{\theta} - \theta_0|| \xrightarrow{p} 0$.*[2]

There are two technical complications present in the proof of Theorem 3.3. First, since $\hat{\theta}(z)$ is a minimizer over a random set, it is not immediately clear whether $\hat{\theta}(z)$ is measurable. Results from Stinchcombre & White (1992), however, imply measurability is in fact not a problem. Second, even though $\theta_0(z)$ is a minimum of $M(\theta)$ on Θ_0, it is often not the minimum on the larger parameter space Θ. In fact, since Θ_0 often has no interior relative to Θ, $\theta_0(z)$ will lie in the boundary of Θ_0 and hence not even be a local minimum of $M(\theta)$ on Θ. The requirement that $\hat{\Theta}_0$ converge to Θ_0 under the Haussdorf norm, however, is enough to overcome this difficulty and attain consistency for $\theta_0(z)$.

The principal purpose of the criterion function $M(\theta)$ is to help us attain a consistent estimator for a unique element $\theta_0(z) \in \Theta_0$. Any population criterion function $M(\theta)$ with a unique minimizer on Θ_0 is a suitable choice. Under the result $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_\infty) \xrightarrow{p} 0$, Theorem 3.3 can be used to construct an estimator $\tilde{\theta}(z) \in \hat{\Theta}_0$ that converges to a unique element $\theta_0(z) \in \Theta_0$ under the norm $||\cdot||_\infty$. Because under $||\cdot||_w$ the set Θ_0 is an equivalence class, the result $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_w) = o_p(n^{-\frac{1}{4}})$ will immediately imply $||\tilde{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$. Thus, by employing any $M(\theta)$ with a unique minimizer in Θ_0 it will be possible to produce an estimator $\tilde{\theta}(z)$ satisfying $||\tilde{\theta} - \theta_0||_\infty = o_p(1)$ and $||\tilde{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$.

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

3.4.2 Selecting $M(\theta)$ to Achieve Asymptotic Normality

While $||\tilde{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$ and $||\tilde{\theta} - \theta_0||_\infty = o_p(1)$ are sufficient for establishing the asymptotic normality of $n^{-\frac{1}{2}} \sum_i (y_i \tilde{\theta}(z_i) - \langle v^*, m_0 \rangle)$ in a regular two stage procedure, an additional complication arises in the present framework. In establishing the asymptotic distribution of the statistic $n^{-\frac{1}{2}} \sum_i (y_i \tilde{\theta}(z_i) - \langle v^*, m_0 \rangle)$, we employ a standard linearization that is analogous to a Taylor expansion in the parametric setting:

$$\begin{aligned}\langle \tilde{v}, \tilde{\theta} - \theta_0 \rangle_w &= E[E[\tilde{v}(Z)|X]E[\tilde{\theta}(Z) - \theta_0(Z)|X] f_X^2(X)] \\ &= \frac{1}{n} \sum_{i=1}^{n} \hat{E}[\tilde{v}_n(Z) \hat{f}_X(X)|x_i](\hat{m}(x_i, \theta_0) - \hat{m}(x_i, \tilde{\theta})) + o_p(n^{-\frac{1}{2}}) \quad (21)\end{aligned}$$

where $\tilde{v}_n(z) = \arg\min_{\Theta_n} ||\tilde{v} - \theta_n||_\infty$. The similarity to a regular Taylor expansion can be seen by noting that $-2n^{-1} \sum_i \hat{E}[\tilde{v}_n(Z) \hat{f}_X(X)|x_i] \hat{m}(x_i, \tilde{\theta})$ is the pathwise derivative of $Q_n(\theta)$ with respect to $\tilde{\theta}(z)$. More precisely, $-2n^{-1} \sum_i \hat{E}[\tilde{v}_n(Z) \hat{f}_X(X)|x_i] \hat{m}(x_i, \tilde{\theta}) = \frac{\partial Q_n(\tilde{\theta}(z) + \tau \tilde{v}_n(z))}{\partial \tau}\Big|_{\tau=0}$. In a standard two stage estimation problem, we would employ $\hat{\theta}(z)$, the exact minimizer of $Q_n(\theta)$. Therefore, the pathwise derivatives would satisfy $\frac{\partial Q_n(\hat{\theta}(z) + \tau \tilde{v}_n(z))}{\partial \tau}\Big|_{\tau=0} = o_p(n^{-\frac{1}{2}})$ and (22) would then simplify to:

$$\langle \tilde{v}, \hat{\theta} - \theta_0 \rangle_w = \frac{1}{n} \sum_{i=1}^{n} \hat{E}[\tilde{v}_n(Z) \hat{f}_X(X)|x_i] \hat{m}(x_i, \theta_0) + o_p(n^{-\frac{1}{2}}) \quad (22)$$

Since the right hand side of (22) no longer depends on $\hat{\theta}(z)$, it is then possible to establish its asymptotic normality.

The additional complication that arises when employing $\tilde{\theta}(z)$ is that it is the minimizer of $M_n(\theta)$ over $\hat{\Theta}_0$, not of $Q_n(\theta)$ over Θ_n. Therefore, the derivative of $Q_n(\theta)$ evaluated at $\tilde{\theta}(z)$ does not necessarily vanish at the required rate. By virtue of $\tilde{\theta}(z) \in \hat{\Theta}_0$, $\tilde{\theta}(z)$ is one of the near-minimizers of $Q_n(\theta)$. Consequently, $\frac{\partial Q_n(\tilde{\theta}(z) + \tau \tilde{v}_n(z))}{\partial \tau}\Big|_{\tau=0} = o_p(1)$, but without the appropriate rate, such result is not sufficient for obtaining the desired linearization in (22). Ensuring that the derivative of $Q_n(\theta)$ vanishes sufficiently fast for all $\theta(z) \in \hat{\Theta}_0$ necessitates $\epsilon_n \searrow 0$ at prohibitively fast rates.

The approach I implement to solve this problem is to carefully choose $M_n(\theta)$ and perturb $\hat{\Theta}_0$ so that the derivative of $M_n(\theta)$ over the perturbation of $\hat{\Theta}_0$ is similar to that of $Q_n(\theta)$ over Θ_n. If the derivatives are alike, then their respective minimizers will behave in similar ways. In particular, we will be able to prove $\frac{\partial Q_n(\tilde{\theta}(z) + \tau \tilde{v}_n(z)}{\partial \tau}\Big|_{\tau=0} = o_p(n^{-\frac{1}{2}})$ for $\tilde{\theta}(z)$ the minimizer of $M_n(\theta)$ over the perturbation of $\hat{\Theta}_0$. In this way we obtain (22) by using a near-minimizer that is consistent for a unique $\theta_0(z) \in \Theta_0$ instead of the exact minimizer $\hat{\theta}(z)$ that fails to converge to a particular

$\theta_0(z) \in \Theta_0$. With this goal in mind, we define:

$$M(\theta) = E\left[(v^*(X) - \theta(Z)f_X(X))^2\right] \qquad M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(v^*(x_i) - \theta(z_i)\hat{f}_X(x_i))^2 \tag{23}$$

Since $M(\theta)$ is a strictly convex continuous functional on Θ_0 under $||\cdot||_\infty$ and Θ_0 is convex and compact, the criterion function $M(\theta)$ attains a unique minimum on Θ_0. Therefore, Theorem 3.3 implies that under regularity conditions $\tilde{\theta}(z) = \arg\min_{\hat{\Theta}_0} M_n(\theta)$ is a consistent estimator for a unique element $\theta_0(z) \in \Theta_0$. Equally important, however, by using $M_n(\theta)$ we can construct a first stage estimator for which the representation in (22) holds. Simple calculations show that:

$$\begin{aligned}\frac{\partial}{\partial\tau}Q\left(\theta(z) + \tau\tilde{v}(z)\right)\Big|_{\tau=0} &= -2E\left[(v^*(X) - \theta(Z)f_X(X))\,E[\tilde{v}(Z)|X]f_X(X)\right] \\ &= \frac{\partial}{\partial\tau}M\left(\theta(z) + \tau E[\tilde{v}(Z)|x]\right)\Big|_{\tau=0}\end{aligned} \tag{24}$$

Hence, for all $\theta(z) \in \Theta$, the pathwise derivative of $Q(\theta)$ in the direction of $\tilde{v}(z)$ is the same as the pathwise derivative of $M(\theta)$ in the direction of $E[\tilde{v}(Z)|x]$. Lemma .2 in the Appendix establishes that an analogue to (24) exists for $Q_n(\theta)$ and $M_n(\theta)$ when evaluated at $\tilde{\theta}(z)$ satisfying $||\tilde{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$:

$$\begin{aligned}\frac{\partial}{\partial\tau}Q_n(\tilde{\theta}(z) + \tau\tilde{v}_n(z))\Big|_{\tau=0} &= -\frac{1}{n}\sum_{i=1}^{n}\hat{f}_X(x_i)\hat{E}[\tilde{v}_n(Z)|x_i](v^*(x_i) - \tilde{\theta}(z_i)\hat{f}_X(x_i)) + o_p(n^{-\frac{1}{2}}) \\ &= \frac{\partial}{\partial\tau}M_n(\tilde{\theta}(z) + \tau\hat{E}[\tilde{v}_n(Z)|x])\Big|_{\tau=0} + o_p(n^{-\frac{1}{2}})\end{aligned} \tag{25}$$

As (25) shows, the derivatives of $Q_n(\theta)$ and $M_n(\theta)$ are closely related. The minimizer $\tilde{\theta}(z)$ of $M_n(\theta)$ over the proper domain satisfies $\frac{\partial M_n(\tilde{\theta}(z)+\tau\hat{E}[\tilde{v}_n(Z)|x])}{\partial\tau}\Big|_{\tau=0} = o_p(n^{-\frac{1}{2}})$. Therefore, by (25) we can conclude $\frac{\partial Q_n(\tilde{\theta}(z)+\tau\tilde{v}_n(z))}{\partial\tau}\Big|_{\tau=0} = o_p(n^{-\frac{1}{2}})$ and hence for such a $\tilde{\theta}(z)$, $n^{-\frac{1}{2}}\sum_i(y_i\tilde{\theta}(z_i) - \langle v^*, m_0\rangle)$ is asymptotically normally distributed. Minimizing $M_n(\theta)$ over $\hat{\Theta}_0$, however, does not imply this result because $\hat{E}[\tilde{v}_n(Z)|x] \notin \hat{\Theta}_0$. The minimizer $\tilde{\theta}(z)$ does not necessarily zero the pathwise derivative of $M_n(\theta)$ in the direction of $\hat{E}[\tilde{v}_n(Z)|x]$ unless it is allowed to take such values in the optimization problem. For this reason, $\hat{\Theta}_0$ is perturbed and we define the first stage estimator to be:

$$\tilde{\theta}(z, x) = \arg\min_{\hat{\Theta}_0 + \lambda_n\mathcal{E}_n} M_n(\theta) \tag{26}$$

where $\mathcal{E}_n = \{\hat{E}[\theta_n(Z)|x] : \theta_n(z) \in \Theta_n\}$ and $\lambda_n \searrow 0$. In the limit, this perturbation is innocuous. As can be seen in (24), evaluated at any $\theta(z) \in \Theta_0$ the pathwise derivative of $M(\theta)$ in the direction of $E[\theta'(Z)|x]$ is zero for all $\theta'(z) \in \Theta$. It follows that if $\theta_0(z) = \arg\min_{\Theta_0} M(\theta)$, then $\theta_0(z)$ is also the minimum of $M(\theta)$ over $\Theta_0 + \mathcal{E}$ for $\mathcal{E} = \{E[\theta'(Z)|x] : \theta'(z) \in \Theta\}$. Therefore, while in finite samples the perturbation of $\hat{\Theta}_0$ ensures that $\tilde{\theta}(z, x)$ zeroes the derivatives of $M_n(\theta)$ and $Q_n(\theta)$, in the population this perturbation is harmless as expanding Θ_0 to $\Theta_0 + \mathcal{E}$ does not change the minimizer of $M(\theta)$.

3.5 Establishing Asymptotic Normality

The first step in showing the asymptotic normality of the statistic $n^{-\frac{1}{2}} \sum_i (y_i \tilde{\theta}(z_i, x_i) - \langle v^*, m_0 \rangle)$, for $\tilde{\theta}(z, x)$ as defined in (26), is to establish that $\hat{\Theta}_0$ is indeed a consistent estimator for Θ_0. For this purpose I will need the following additional assumption:

ASSUMPTION 3: (vii) $\sup_{\Theta} \inf_{\Theta_n} ||\theta - \theta_n||_\infty = o(n^{-\frac{1}{2}-\frac{\delta_0}{2}})$.

Assumption 3(vii) requires the approximation error from using sieves to decrease at a considerably faster rate than what is needed in the point identified case. This rate can be obtained by assuming sufficient smoothness in the parameter space Θ. In order for $\hat{\Theta}_0$ to be consistent under $d_H(\cdot, \cdot, ||\cdot||_\infty)$ we need to ensure that all $\theta_n(z) \in \Theta_n$ that approximate $\theta_0(z) \in \Theta_0$ well are included in $\hat{\Theta}_0$. Such requirement is met if $\sup_{\Theta_{0p_n}} Q_n(\theta) = o_p(\epsilon_n)$, for Θ_{0p_n} the projection of Θ_0 onto Θ_n under $||\cdot||_\infty$. Unfortunately, in order for $\hat{\Theta}_0$ to also be consistent at the required rate $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_w) = o_p(n^{-\frac{1}{4}})$, the bandwidth ϵ_n must decrease to zero sufficiently fast. Assumption 3(vii) allows us to use such ϵ_n while still satisfying $\sup_{\Theta_{0p_n}} Q_n(\theta) = o_p(\epsilon_n)$ and hence remaining consistent under $d_H(\cdot, \cdot, ||\cdot||_\infty)$.

Assumptions 1-5 are sufficient for establishing that $\hat{\Theta}_0$ is consistent at the proper rates:

Theorem 3.4. *Under Assumptions 1, 2(i)-(ii), 3(i)-(ii), 3(iv)-(vii), 4(i)-(iii) and 5(i)-(iv), if* $\hat{\Theta}_0 = \{\theta_n \in \Theta_n : Q_n(\theta) \leq b_n/a_n\}$ *for* $a_n = C_a n^{\frac{1}{2}+\frac{\delta_0}{2}}$ *and* $b_n = C_b \log n$, *then* $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_\infty) \xrightarrow{p} 0$ *and* $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_w) = o_p(n^{-\frac{1}{4}})$.

The second step in our estimation procedure consists in employing Theorem 3.3 to construct an estimator for a unique element $\theta_0(z) \in \Theta_0$. While Theorem 3.4 establishes $\hat{\Theta}_0$ is a consistent estimator for Θ_0, the proposed first stage estimator solves a minimization problem over $\hat{\Theta}_0 + \lambda_n \mathcal{E}_n$, not $\hat{\Theta}_0$. Assumptions 7(i)-(ii) require $\lambda_n \searrow 0$ at the appropriate rate so that the perturbation $\hat{\Theta}_0 + \lambda_n \mathcal{E}_n$ also satisfies $d_H(\hat{\Theta}_0 + \lambda_n \mathcal{E}_n, \Theta_0, ||\cdot||_w) = o_p(n^{-\frac{1}{4}})$ and $d_H(\hat{\Theta}_0 + \lambda_n \mathcal{E}_n, \Theta_0, ||\cdot||_\infty) = o_p(1)$.

ASSUMPTION 7: (i) $\lambda_n = o(n^{-\frac{1}{4}})$, (ii) $\lambda_n \xi_{jn} \to 0$ for $j \in \{0, 1\}$

Because Θ_0 is an equivalence class under $||\cdot||_w$, $d_H(\hat{\Theta}_0 + \lambda_n \mathcal{E}_n, \Theta_0, ||\cdot||_w) = o_p(n^{-\frac{1}{4}})$ immediately implies $\tilde{\theta}(z, x) \in \hat{\Theta}_0 + \lambda_n \mathcal{E}_n$ also meets the rate requirement $||\tilde{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$. Theorem 3.5 shows that Assumptions 1-5 and 7 imply $\tilde{\theta}(z, x)$ is consistent for a unique $\theta_0(z) \in \Theta_0$ under the stronger norm $||\cdot||_\infty$ as well.

Theorem 3.5. *Under Assumptions 1, 2(i)-(ii), 3(i)-(ii), 3(iv)-(vii), 4(i)-(iii), 5(i)-(iv) and 7(i)-(ii), it follows that* $||\tilde{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$ *and* $||\tilde{\theta} - \theta_0||_\infty = o_p(1)$.

The first stage estimator $\tilde{\theta}(z,x)$ therefore meets the necessary rate requirements. Furthermore, as discussed in Section 3.4, $\tilde{\theta}(z,x)$ also satisfies the needed linearization in (22). Under two additional assumptions, $\tilde{\theta}(z,x)$ can be used to construct an asymptotically normal estimator for $\langle v^*, m_0 \rangle$. Assumption 6(ii) is an interior condition analogous to Assumption 6(i) that allows us to differentiate by ruling out the possibility that $\tilde{\theta}(z,x)$ lies on the boundary of $\hat{\Theta}_0 + \lambda_n \mathcal{E}_n$. Assumption 7(iii) guarantees the endogenous component $\lambda_n \hat{E}[\theta_n(Z)|x_i]$ of $\tilde{\theta}(z,x)$ is asymptotically negligible in the second stage.

ASSUMPTION 6: (ii) Let $\tilde{\theta}(z) = \bar{\theta}(z) + \hat{E}[\theta_n(Z)|x]$, then there is some $\epsilon_n = o(n^{-\frac{1}{2}})$ such that $P(\theta_n(z) + u_n(z)\epsilon_n/\lambda_n \in \Theta_n) \to 1$.

ASSUMPTION 7: (iii) $\lambda_n \times n^{\frac{1}{2}} \times \xi_{0n} \to 0$

Assumptions 1-7 are sufficient for establishing our main result. Theorem 3.6 shows that the two stage estimaor $n^{-\frac{1}{2}} \sum_i (y_i \tilde{\theta}(z_i, x_i) - \langle v^*, m_0 \rangle)$ is asymptotically normally distributed.

Theorem 3.6. *Under Assumptions 1, 2(i)-(ii), 3(i)-(ii), 3(iv)-(vii), 4(i)-(iii), 5(i)-(iv), 6(ii) and 7(i)-(iii),* $n^{-\frac{1}{2}} \left(\sum_i y_i \tilde{\theta}(z_i, x_i) - \langle v^*, m_0 \rangle \right) \xrightarrow{\mathcal{L}} N(0, \sigma^2)$ *where the asymptotic variance is given by* $\sigma^2 = Var\left((Y - E[\tilde{v}(Z)|X] f_X^2(X)) \theta_0(Z)\right)$.

The asymptotic variance obtained in Theorem 3.6 depends on what element $\theta_0(z) \in \Theta_0$ the first stage estimator $\tilde{\theta}(z,x)$ is consistent for. Modifying the criterion functions to improve efficiency is a challenging exercise. The derivatives of $Q(\theta)$ and $M(\theta)$ need to remain linked as in (24) in order to preserve asymptotic normality. Hence, a weight function $w^2(x)$ can only be introduced into $Q(\theta)$ and $M(\theta)$ through the joint modification:

$$Q_w(\theta) = E[(v^*(X) - E[\theta(Z)|X] f_X(X))^2 w^2(X)] \qquad M_w(\theta) = E[(v^*(X) - \theta(Z) f_X(X))^2 w^2(X)] \tag{27}$$

The asymptotic variance of the second stage estimator corresponding to the criterion functions in (27) is given by:[3]

$$\sigma_w^2 = Var((Y - E[\tilde{v}_w(Z)|X] f_X^2(X) w^2(X)) \theta_w(Z)) \qquad \theta_w(z) = \arg\min_{\theta \in \Theta_0} M_w(\theta) \tag{28}$$

Therefore, the weight function $w^2(x)$ enters directly into the formula for σ_w^2, but also indirectly by affecting the element $\theta_w(z) \in \Theta_0$ that the first stage estimator is consistent for.

[3]The proper weak norm for these criterion functions is implied by the dot product $\langle \theta_1, \theta_2 \rangle = E[E[\theta_1(Z)|X] E[\theta_2(Z)|X] f_X^2 w^2(X)]$. The function $\tilde{v}_w(Z)$ in (28) corresponds to the Riesz representor of $E[Y\theta(Z)]$ under this new dot product.

In a special case, the optimal weight function $w^2(x)$ can be found and the resulting estimator is efficient in that it attains the semiparametric efficiency bound derived in Severini & Tripathi (2007). Let $m_0^*(x)$ denote the projection of $m_0(x)$ onto $\mathcal{N}(E[\cdot|Z])^\perp$, where $\mathcal{N}(E[\cdot|Z])$ is the null space of the operator $E[\cdot|Z]$, and define

$$\tilde{\epsilon} = Y - m_0^*(X) \tag{29}$$

If there exists a solution to the system of integral equations (in $\tau(x) \in \mathcal{L}^2(\mathcal{X})$ and $\eta(z) \in \bar{V}$):

$$E[\tilde{\epsilon}^2|z] = E[f_X^2(X)\tau^2(X)|z] \tag{30}$$

$$\frac{m_0^*(x)}{f_X^2(x)w^2(x)} = E[\eta(Z)|x] \tag{31}$$

then the optimal weight function is given by $w(x) = \tau^*(x)$ for $\tau^*(x)$ solving (30), while the Riesz Representor $\tilde{v}_w(z)$ is given by $\tilde{v}_w(z) = \eta^*(z)$ for $\eta^*(z)$ solving (31). With this choice of $w^2(x)$, the asymptotic variance simplifies to:

$$\sigma_w^2 = E[\tilde{\epsilon}^2\theta_w^2(Z)] \qquad \theta_w^2(z) = \arg\min_{\theta\in\Theta_0} E[\theta^2(Z)f_X^2(X)w^2(X)] \tag{32}$$

Under the conditions of Theorem 2.4 in Severini & Tripathi (2007), the semiparametric efficiency bound for estimating $\langle v^*, m_0\rangle$ is given by

$$E[\tilde{\epsilon}^2\vartheta^2(Z)] \tag{33}$$

where $\vartheta(z)$ is a function that solves (3). Hence, if we further assume $\vartheta(z)$ is sufficiently smooth, then $\vartheta(z) \in \Theta_0$. Because $w^2(x)$ solves (30), however, it follows from (32) and $\vartheta(z) \in \Theta_0$ that

$$E[\tilde{\epsilon}^2\theta_w^2(Z)] \leq E[\tilde{\epsilon}^2\vartheta^2(Z)] \tag{34}$$

Therefore, the resulting estimator attains the semiparametric efficiency bound obtained in Severini & Tripathi (2007).

The preceding discussion, however, hinges on the existence to solutions to (30) and (31). Finding the optimal weight function $w^2(x)$ in a general setting is a complicated open problem. We believe that unless solutions to (30) and (31) exist, the optimal weight function will not be able to yield an estimator that attains the semiparametric efficiency bound.

3.6 Estimating the Asymptotic Variance

In order to construct confidence intervals for $\langle v^*, m_0\rangle$ we still require a consistent estimator for the asymptotic variance of the statistic $n^{-\frac{1}{2}}\sum_i(y_i\tilde{\theta}(z_i, x_i) - \langle v^*, m_0\rangle)$. The only remaining complication is obtaining a consistent estimator for $\tilde{v}(z)$. Because $\tilde{v}(z)$ satisfies $E[Y\theta(Z)] = \langle\tilde{v}, \theta\rangle_w$

for all $\theta(z) \in \overline{V}$, it follows that $\tilde{v}(z)$ zeroes all pathwise derivatives of the convex functional $\frac{1}{2}E\left[(E[v(Z)|X])^2 f_X^2(X)\right] - E[Yv(Z)]$ in any direction within $\overline{V}$. Hence, we can characterize $\tilde{v}(z)$ as:

$$\tilde{v}(z) \in \arg\min_{v(z)\in\overline{V}} \frac{1}{2}E\left[(E[v(Z)|X])^2 f_X^2(X)\right] - E[Yv(Z)] \tag{35}$$

In the general case where Θ_0 is not a singleton, $\tilde{v}(z)$ will not be identified either. If Θ_0 is not a singleton, then there exists at least one $\theta_{\mathcal{N}}(z) \in \Theta$ such that $E[\theta_{\mathcal{N}}(Z)|x] = 0$. Therefore, the set of minimizers to (35) includes $\tilde{v}(z) + \lambda\theta_{\mathcal{N}}(z)$ for any $\lambda \in \Re$. The set of minimizers to (35), however, forms an equivalence class under the weak norm $||\cdot||_w$.[4] Fortunately, since $\tilde{v}(z)$ only affects the asymptotic variance σ^2 through the term $E[\tilde{v}(Z)|x]$, consistency to $\tilde{v}(z)$ under $||\cdot||_w$ is sufficient for recovering an estimator for σ^2. Therefore, define the sample analogue to (35) by:

$$\hat{v}(z) \in \arg\min_{\theta_n(z)\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(\hat{E}[\theta_n(Z)|x_i])^2 \hat{f}_X^2(x_i) - y_i\theta_n(z_i) \tag{36}$$

where for $\hat{E}[\theta(Z)|x]$ and $\hat{f}_X(x)$ we continue using a series and a Kernel estimator respectively.

Utilizing the estimators $\tilde{\theta}(z,x)$, $\hat{E}[\hat{v}(Z)|x]$ and $\hat{f}_X(x)$ for the nuisance parameters determining the asymptotic variance $\sigma^2 = Var\left((Y - E[\tilde{v}(Z)|X]f_X^2(X))\theta_0(Z)\right))$, we define the estimator $\hat{\sigma}^2$ as:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left((y_i - \hat{E}[\hat{v}(Z)|x_i]\hat{f}_X^2(x_i))\tilde{\theta}(z_i,x_i)\right)^2 - \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{E}[\hat{v}(Z)|x_i]\hat{f}_X^2(x_i))\tilde{\theta}(z_i,x_i)\right]^2 \tag{37}$$

Lemma 3.1 establishes the consistency of $\hat{\sigma}^2$ for σ^2.

Lemma 3.1. *Under Assumptions 1, 2(i)-(ii), 3(i)-(ii), 3(iv)-(vii), 4(i)-(iii), 5(i)-(iv), 6(ii) and 7(i)-(iii), the estimator $\hat{\sigma}^2$ satisfies $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$.*

4 Monte Carlo Performance

In order to illustrate the implementation of the outlined procedure and examine the finite sample performance we conduct a small-scale Monte Carlo study. The functional of interest is the change in consumer surplus associated with a decrement in price from p_1 to p_2. For $m_0(x)$ an inverse demand

[4]Let $\overline{W} = \{w(x) = E[\theta(Z)|x] : \theta(z) \in \overline{V}\}$. The uniqueness of the Riesz Representation theorem implies all $v(z)$ satisfying $E[v(Z)\theta(Z)] = E[Y\theta(Z)]$ form an equivalence class. Hence, $E[\tilde{v}(Z)|x]$ is the sole function zeroing all pathwise derivatives of $E[w^2(x)f_X^2(x)] - E[m_0(x)w(x)]$ within $\overline{W}$. Because $E[\epsilon|Z] = 0$, this problem is equivalent to (35) and hence all minimizers $v(z)$ satisfy $E[v(Z)|x] = E[\tilde{v}(Z)|x]$.

function and q_1 and q_2 the market clearing quantities corresponding to prices p_1 and p_2, the change in consumer surplus is given by:

$$\int_0^{q_2} m_0(x)dx - p_2q_2 - \int_0^{q_1} m_0(x)dx + p_1q_1 = \langle v^*, m_0\rangle - (p_2q_2 - p_1q_1) \tag{38}$$

where $v^*(x) = 1\{q_1 \leq x \leq q_2\}$. If (p_1, q_1) and (p_2, q_2) are observable, then only $\langle v^*, m_0\rangle$ needs to be estimated. To facilitate exposition, the Monte Carlo was designed so that the integral equation in (3) has a closed form solution. Assume $X, Z \in [0,1]^2$ are distributed according to the density:

$$f_{XZ}(x,z) = 3|x-z| \quad for \quad (x,z) \in [0,1]^2 \tag{39}$$

For this choice of density, there is no solution to the integral equation $E[\theta_0(Z)|x] = v^*(x)/f_X(x)$. Instead, we estimate $\langle v_t, m_0\rangle$ for $v_t(x)$ an approximation to $v^*(x)$. Let $F_t(x) = \Phi\left(\frac{x-q_2}{t}\right) - \Phi\left(\frac{x-q_1}{t}\right)$ where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution and let $A_t = -\frac{1}{2}(F_t^{'}(1) + F_t^{'}(0))$ and $B_t = \frac{1}{2}(F_t^{'}(1) - F_t(1) - F_t(0))$. We approximate $v^*(x)$ with the function $v_t(x)$ defined as:

$$v_t(x) = F_t(x) + A_tx + B_t \tag{40}$$

Notice that pointwise $\lim_{t\to 0} v_t(x) = v^*(x)$ almost everywhere and that $\lim_{t\to 0}\langle v_t, m_0\rangle = \langle v^*, m_0\rangle$. Furthermore, for any choice of $t > 0$ and $\phi(u)$ the density of a standard normal random variable, we have by Polyanin & Manzhirov (1998):

$$E\left[\theta_0(Z)\Big|X = x\right] = \frac{v_t(x)}{f_X(x)} \quad for \quad \theta_0(z) = \frac{1}{t^2}\left(\phi^{'}\left(\frac{z-q_2}{t}\right) - \phi^{'}\left(\frac{z-q_1}{t}\right)\right) \tag{41}$$

Given the specifications in (38), (39) and (40), $\theta_0(z)$ as defined in (41) is actually the unique solution to $E[\theta_0(Z)|x] = v_t(x)/f_X(x)$. In this setting the statistic $n^{-\frac{1}{2}}\sum_i(y_i\hat{\theta}(z_i) - \langle v_t, m_0\rangle)$ proposed in Sections 3.2-3.3 is asymptotically normally distributed. We will therefore first examine the performance of such estimator and then proceed to evaluate the statistic $n^{-\frac{1}{2}}\sum_i(y_i\tilde{\theta}(z_i, x_i) - \langle v_t, m_0\rangle)$ discussed in Sections 3.4-3.5.

By (41), $\theta_0(z)$ is infinitely differentiable, though it has large derivatives in a neighborhood of q_1 and q_2. For the parameter space we set $\Theta = \{\theta(z) : \sum_{|\lambda|\leq\omega}\int_0^1[D^\lambda\theta(z)]^2dz \leq C\}$ for $\omega = 5$ and C large enough to ensure $\theta_0(z) \in \Theta$.[5] The elements in Θ can be approximated arbitrarily well by a sieve Θ_n of B-Splines of order 6. Similarly, for the basis $\{p_j(x)\}_{j=1}^\infty$ we utilize B-Splines of order 3. By results in Chen (2006) and Newey (1997), these choices are compatible with our assumptions if

[5]By the Sobolev Imbedding Theorem $\Theta \subset \Lambda_{C_1}^{\omega-1}([0,1])$ for some constant C_1. Hence, this definition of the parameter space is compatible with Assumption 3. The constraint $\theta(z) \in \Theta$, however, is more tractable as it is quadratic in the coefficients of a linear sieve. See Newey & Powell (2003) for a detailed discussion.

$k_n \asymp n^{\frac{1}{8}}$ and $k_{1n} \asymp n^{\frac{1}{8}}$. Due to the low dimensionality of the problem, the kernel used to estimate $f_X(x)$ only needs to be of order $k > 4/3$. Hence, we use $\phi(u)$, the density of a standard normal random variable, as the choice of kernel $K(u)$ for constructing $\hat{f}_X(x)$.

The computation of the estimator $\hat{\theta}(z)$ is straightforward, as it is defined by the solution to a quadratic programming problem. Because we are using a linear sieve Θ_n, every element $\theta_n(z) \in \Theta_n$ is of the form $\theta_n(z) = q^{k_{1n}'}(z)\beta$ for $\{q_i(z)\}_{i=1}^{\infty}$ an appropriate basis and $\beta \in \Re^{k_{1n}}$. Furthermore, for $\Lambda_{k_{1n}} = \sum_{|\lambda| \leq \omega} \int_0^1 D^\lambda q^{k_{1n}}(z) D^\lambda q^{k_{1n}'}(z) dz$, the constraint $\sum_{|\lambda| \leq \omega} \int_0^1 [D^\lambda \theta_n(z)]^2 dz \leq C$ is equivalent to $\beta' \Lambda_{k_{1n}} \beta \leq C$. Hence, the first stage estimator $\hat{\theta}(z)$, defined in (13) is given by:

$$\hat{\theta}(z) = q^{k_{1n}'}(z)\hat{\beta} \quad \hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(\hat{E}[v_t(X)|x_i] - \hat{E}[\hat{f}_X(X) q^{k_{1n}'}(Z)|x_i]\beta \right)^2 \quad s.t. \;\; \beta' \Lambda_{k_{1n}} \beta \leq C \tag{42}$$

The optimal coefficients $\hat{\beta}$ have a ridge regression form, see Newey & Powell (2003) for a discussion.

The estimation of the nuisance parameter $\theta_0(z)$ does not require observations of Y, which reflects that the proposed procedure imposes strong assumptions on the joint distribution of X and Z but almost none on the economic model $m_0(x)$. For the simulations we created 1000 replications of sample size $n = 1000$ from the model:[6]

$$Y = 2 - X^2 + \epsilon \qquad \epsilon = -\frac{U}{12} \left(\frac{1}{f_{X|Z}(X|Z)} - 1 \right) \tag{43}$$

where $U \sim U[0,1]$ independent of X and Z and $f_{X|Z}(x|z)$ is the conditional density of X given Z. By construction, $E[\epsilon|Z] = 0$, and straightforward calculations show $E[\epsilon|x] = (1 - f_X(x))/24 f_X(x)$.

For the simulations, we set $t = 0.01$, $q_1 = 0.25$ and $q_2 = 0.75$ when constructing $v_t(x)$ as in (40). The resulting approximation to $\int_{q_1}^{q_2} m_0(x)dx$ has a value of $\langle v_t, m_0 \rangle = 0.865$, while the true parameter is $\langle v^*, m_0 \rangle = 0.866$. All simulation results, such as bias and size, are reported with respect to $\langle v_t, m_0 \rangle$. They are almost identical to the results with respect to $\langle v^*, m_0 \rangle$. The assumptions in Section 3.3 impose rate requirements on the different bandwidths, but offer little guidance as to how to choose their level. For the B-Splines of order 6 used to construct Θ_n we used 3 knots placed at $\{0, 1/2, 2\}$, implying $k_{1n} = 7$. Similarly, for the B-Splines of order 3 used to compute $\hat{E}[\cdot|x]$ we placed 7 equally spaced knots including the end-points $\{0, 1\}$, implying $k_n = 10$. Simulations with larger sieves yielded qualitatively similar results. Table 1 reports the simulation results for different choices of the bandwidth h used in constructing $\hat{f}_X(x)$. The last column in Table 1 contains the

[6]The random variable $-U\left(\frac{1}{f_{X|Z}(X|Z)} - 1\right)$ has significantly fat tails: the maximum draw across samples and replications was -214. The scaling by 1/12 ensures that ϵ retains variability while preventing most of its large realization from driving the results.

actual confidence level of a confidence interval of nominal size 0.05 constructed using a normal approximation and $\hat{\sigma}^2$ as in Lemma 3.1.

Table 1: SIMULATIONS FOR $\frac{1}{n}\sum y_i\hat{\theta}(z_i)$

Bandwidth h	Mean$\left(\frac{1}{n}\sum_i y_i\hat{\theta}(z_i)\right)$	Bias	Std$\left(\frac{1}{n}\sum_i y_i\hat{\theta}(z_i)\right)$	Mean$(\hat{\sigma}/\sqrt{n})$	CI size for $\alpha = 0.05$
$h = 0.1$	0.821	-0.044	0.030	0.037	0.773
$h = 0.075$	0.829	-0.036	0.033	0.038	0.872
$h = 0.05$	0.835	-0.029	0.035	0.042	0.918
$h = 0.025$	0.839	-0.025	0.041	0.053	0.963
$h = 0.01$	0.839	-0.025	0.056	0.079	0.979

Table 1 indicates that the usual bias/variance tradeoff present in choosing h when estimating $\hat{f}_X(x)$ translates into a similar tradeoff for the final estimator of $\langle v_t, m_0\rangle$. On the other hand, the estimated standard deviation $\hat{\sigma}/\sqrt{n}$ overstates the actual finite sample variability of $n^{-1}\sum_i y_i\hat{\theta}(z_i)$. These two results affect the actual size of the confidence interval in opposite directions. While the bias in our estimator causes the confidence interval to be improperly centered, and hence increases its actual size, the overstatement of the finite sample variability causes the confidence intervals to be conservative. The consequence of these two effects is reflected in Table 1 as h decreases, with actual size of the confidence interval being larger than $\alpha = 0.05$ for $h > 0.025$, and slightly conservative for $h \leq 0.025$.

We now proceed to evaluate the finite sample performance of the estimator $n^{-1}\sum_i y_i\tilde{\theta}(z_i, x_i)$, which is robust to $\theta_0(z)$ not being identified. The principal practical challenge in implementing this procedure is in the selection of the bandwidths ϵ_n and λ_n. Their respective rate requirements offer no restrictions as to how their levels should be chosen. The simulation analysis therefore focuses on the selection of ϵ_n and λ_n. All other bandwidths are left at the same level as in the simulations for $n^{-1}\sum_i y_i\hat{\theta}(z_i)$ and we set $h = 0.01$.

The first stage estimator $\tilde{\theta}(z, x)$ is easy to compute. Because the sieve Θ_n is linear, the first stage estimator is of the form $\tilde{\theta}(z, x) = q^{k_{1n}'}(z)\hat{\beta}_1 + \lambda_n\hat{E}[q^{k_{1n}'}(Z)|x]\hat{\beta}_2$. The coefficients $(\hat{\beta}_1, \hat{\beta}_2)$ are the solution to the following quadratic programming problem:

$$(\hat{\beta}_1, \hat{\beta}_2) = \arg\min_{\beta_1,\beta_2} \frac{1}{n}\sum_{i=1}^{n}\left(v_t(x_i) - \hat{f}_X(x_i)q^{k_{1n}'}(z_i)\beta_1 - \lambda_n\hat{f}_X(x_i)\hat{E}[q^{k_{1n}'}(Z)|x_i]\beta_2\right)^2$$

$$s.t.\ \ 1)\,\frac{1}{n}\sum_{i=1}^{n}\left(\hat{E}[v_t(X)|x_i] - \hat{E}[\hat{f}_X(X)q^{k_{1n}'}(Z)|x_i]\beta_1\right)^2 \leq \epsilon_n \quad 2)\,\beta_1'\Lambda_{k_{1n}}\beta_1 \leq C \quad 3)\,\beta_2'\Lambda_{k_{1n}}\beta_2 \leq C \quad (44)$$

Constraints 1) and 2) imply $q^{k_{1n}'}(z)\hat{\beta}_1 \in \hat{\Theta}_0$, while constraint 3) ensures $\hat{E}[q^{k_{1n}'}(Z)|x]\hat{\beta}_2 \in \mathcal{E}_n$. Since $\tilde{\theta}(z, x)$ is consistent for $\theta_0(z) = \arg\min_{\Theta_0} M(\theta)$ and $\theta_0(z)$ is generically not the global minimizer of

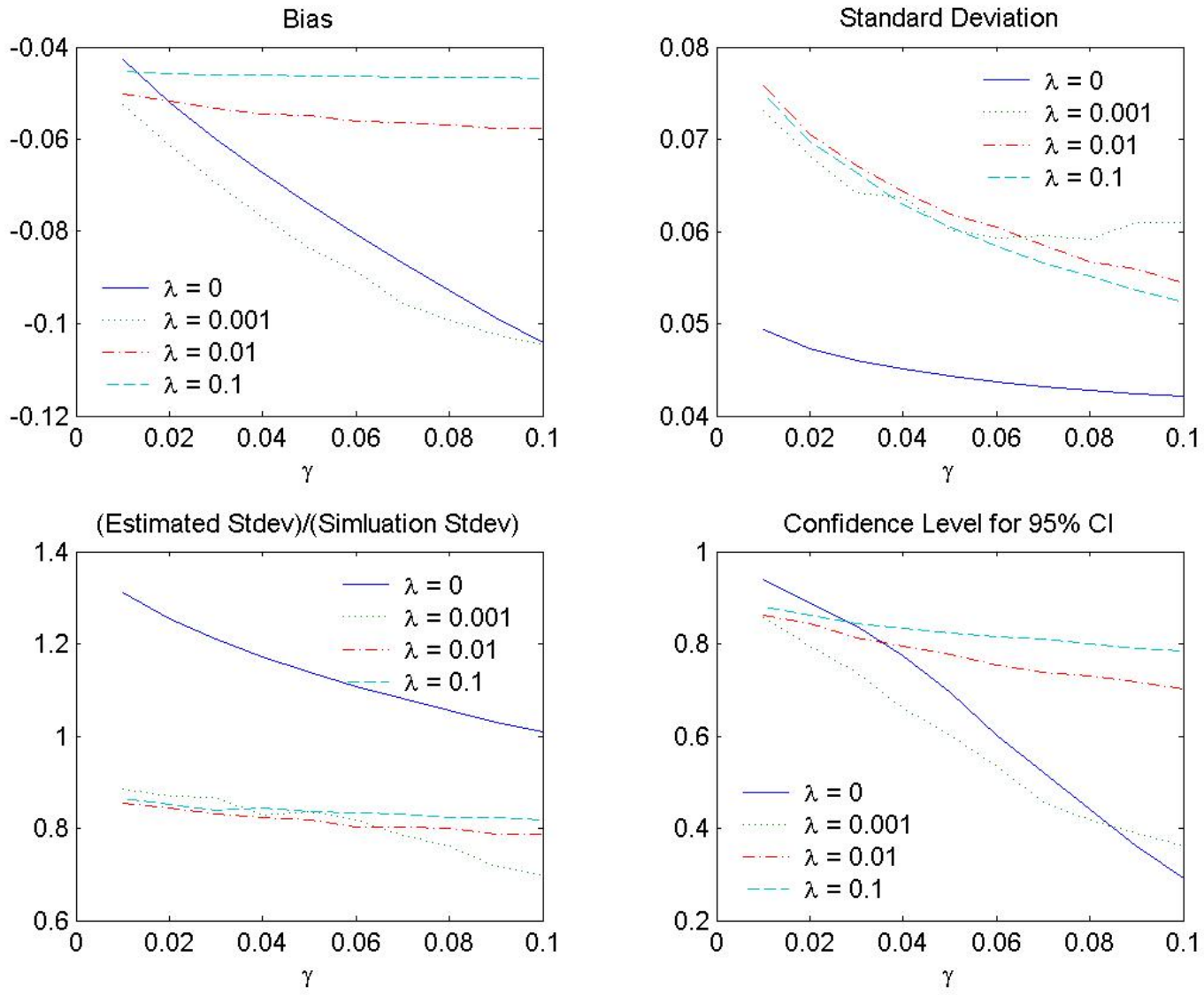

Figure 1: Effects of γ and λ

$M(\theta)$,[7] it follows that constraint 1) asymptotically impacts the solution to (44). It therefore seems prudent to select ϵ_n small enough so that constraint 1) affects the solution to (44) for every n as well. For $\bar{\theta}(z)$ the unconstrained minimizer to (44), we implement the following ad-hoc rule for selecting ϵ_n:

$$\epsilon_n = \gamma Q_n(\bar{\theta}) + (1-\gamma) Q_n(\hat{\theta}) \tag{45}$$

For $\gamma \in (0,1)$, setting ϵ_n according to (45) ensures both that constraint 1) affects the optimal value

[7]Ignoring all constraints, the minimizer to $M(\theta)$ is given by $\tilde{E}[v_h(X)|z]$ where $\tilde{E}[\cdot|z]$ is the conditional expectation with respect to the measure $f_{XZ}(x,z)f_X^2(x)/[\int f_{XZ}(x,z)f_X^2(x)dxdz]$. Hence, the global minimizer is in Θ_0 iff $E[\tilde{E}[v_h(X)|Z]|x] = v_h(x)$, i.e. $v_h(x)$ is the unit eigenfunction for the operator $E[\tilde{E}[\cdot|Z]|X]$. Notice that if this is the case, then $\theta_0(z) = \tilde{E}[v_h(X)|z]$ is a unique identified element in Θ_0, implying we can estimate it directly.

in (44) and that $\hat{\Theta}_0$ contains other elements in addition to $\hat{\theta}(z)$.

Figure 1 reports the results from simulations for different combinations of values for γ and λ_n.[8] Enlarging $\hat{\Theta}_0$ by increasing γ worsens the bias of the estimator $n^{-1}\sum_i y_i\tilde{\theta}(z_i,x_i)$ for $\langle v_t, m_0\rangle$ while at the same time decreasing its variance. In accordance to the theory, however, the perturbation by $\lambda_n\mathcal{E}_n$ greatly helps remedy the situation. For $\lambda_n = 0.01$ and $\lambda_n = 0.001$, the bias of $n^{-1}\sum y_i\tilde{\theta}(z_i,x_i)$ remains stable across values of γ. These values of λ_n represent perturbations small in magnitude. Table 5 in Appendix E shows the endogenous component $\lambda_n\hat{f}_X(x_i)\hat{E}[q^{k'_{1n}}(Z)|x_i]\hat{\beta}_2$ contributes less than 5.5% to the total variability of $\tilde{\theta}(z_i,x_i)$ for all values of γ. The smaller bias causes the confidence intervals constructed using a normal approximation to provide a better control on the actual size for $\lambda_n = 0.01$ and $\lambda_n = 0.1$ than for $\lambda_n = 0$ and $\lambda_n = 0.001$. The simulation results suggest that the introduction of the bandwidth parameter λ_n is indeed necessary to ensure the derivative of $Q_n(\theta)$ evaluated at $\tilde{\theta}(z,x)$ vanishes at the required rate. Without this derivative vanishing, the second stage estimator $n^{-1}\sum_i\tilde{\theta}(z_i,x_i)y_i$ will be asymptotically biased at the $\sqrt{n}$ rate.

5 Conclusion

The results in this paper allow for the estimation of continuous linear functionals in an additive separable model with endogenous regressors. Two estimators are examined that rely on the minimal assumption that $\langle v^*, m_0\rangle$ is $\sqrt{n}$ estimable and do not require the underlying model $m_0(x)$ be identified. These estimators make use of a nuisance parameter that is itself not identified. An adaptation of procedures derived in Newey & Powell (2003) and Ai & Chen (2003) yield a $\sqrt{n}$ consistent estimator that fails to be asymptotically normal. A second more complex procedure is able to reestablish asymptotic normality. The construction of this second estimator required the generalization of set estimation results in Chernozhukov, Hong & Tamer (2007) to arbitrary metric spaces. These techniques should also be applicable to other estimation problems with partially identified nonparametric nuisance parameters.

[8] See Appendix E for an analysis of what the different levels of γ and λ mean.

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:

$a \lesssim b$	$a \leq Mb$ for some constant M which is universal in the context of the proof
$\|\|\theta\|\|_\infty$	The sup-norm $\sup_{z\in\mathcal{Z}} \|\theta(z)\|$
$\|\|\theta\|\|_w$	The norm $\left(E\left[(E[\theta(Z)\|X])^2\right]\right)^{\frac{1}{2}}$
$\langle\theta_1,\theta_2\rangle_w$	The dot product $E\left[E[\theta_1(Z)\|X]E[\theta_2(Z)\|X]\right]$ corresponding to the norm $\|\|\theta\|\|_w$
$\|\|\theta\|\|_{\Lambda^\omega}$	The norm $\max_{\|\lambda\|\leq\underline{\omega}} \sup_{z\in\mathcal{Z}} \|D^\lambda\theta(z)\| + \max_{\|\lambda\|=\underline{\omega}} \sup_{z\neq z'} \frac{\|D^\lambda\theta(z)-D^\lambda\theta(z')\|}{\|\|z-z'\|\|^{\omega-\underline{\omega}}}$
$\rho_n(\theta 1,\theta_2)$	The random semimetric $\left[n^{-1}\sum_i(\theta_1(z_i)-\theta_2(z_i))^2\right]^{\frac{1}{2}}$
$h(\Theta_1,\Theta_2)$	The maximal approximation error $h(\Theta_1,\Theta_2) = \sup_{\theta_1\in\Theta_1} \inf_{\theta_2\in\Theta_2} \|\|\theta_1-\theta_2\|\|$
$d_H(\Theta_1,\Theta_2,\|\|\cdot\|\|)$	The Haussdorf norm under $\|\|\cdot\|\|$ given by $d_H(\Theta_1,\Theta_2,\|\|\cdot\|\|) = \max\{h(\Theta_1,\Theta_2), h(\Theta_2,\Theta_1)\}$
$m(x,\theta)$	The mapping on θ given by $m(x,\theta) = E[v^*(X) - \theta(Z)f_X(X)\|x]$
$N(\epsilon,\mathcal{F},\|\|\cdot\|\|)$	The covering numbers of size ϵ for $\mathcal{F}$ under the norm $\|\|\cdot\|\|$
$N_{[]}(\epsilon,\mathcal{F},\|\|\cdot\|\|)$	The bracketing numbers of size ϵ for $\mathcal{F}$ under the norm $\|\|\cdot\|\|$
$l^\infty(\Theta)$	The space of bounded functionals on Θ
$\Theta_0^{\bar{\epsilon},\|\|\cdot\|\|}$	The closed neighborhood of Θ_0 given by $\Theta_0^{\bar{\epsilon},\|\|\cdot\|\|} = \{\theta\in\Theta : \inf_{\Theta_0}\|\|\theta-\theta_0\|\| \leq \epsilon\}$
$\Theta_{0n}^{\bar{\epsilon},\|\|\cdot\|\|}$	The closed neighborhood of Θ_0 given by $\Theta_{0n}^{\bar{\epsilon},\|\|\cdot\|\|} = \{\theta\in\Theta_n : \inf_{\Theta_0}\|\|\theta-\theta_0\|\| \leq \epsilon\}$
$\mathcal{E}_n$	The set of functions $\mathcal{E}_n = \{\hat{E}[\theta_n(Z)\|x] : \theta_n(z)\in\Theta_n\}$
$\tilde{v}(z)$	The Riesz Representor for the functional $E[Y\theta(Z)]$ under the inner product $\langle\theta_1,\theta_2\rangle_w$
$u(z)$	Define to equal $u(z) = \pm\tilde{v}(z)$
$u_n(z)$	The projection of $u(z)$ onto Θ_n under $\|\|\cdot\|\|_\infty$, given by $u_n(z) = \arg\min_{\Theta_n}\|\|\theta_n-u\|\|_\infty$
$\overline{V}$	The closure of the linear span of Θ under $\|\|\cdot\|\|_w$
ξ_{jn}	Defined by $\xi_{jn} = \sup_{x,\|\lambda\|=j}\|\|D^\lambda p^{k_n}(x)\|\|$

APPENDIX B - Proof of Theorem 3.1, 3.2, Auxiliary Lemmas .1, .2 and Corollary .1

Lemma .1. *Let $m(x,\theta) = E[v^*(X) - \theta(Z)f_X(X)|x]$ and $\Theta_{0n}^{\bar{\epsilon},\|\|\cdot\|\|_w} = \{\theta\in\Theta_n : \inf_{\Theta_0}\|\|\theta-\theta_0\|\|_w \leq \epsilon\}$. Under Assumptions 1, 2(i)-(ii), 3(i)-(iv), 4(i)-(iii) and 5(i)-(iv), a) $\sup_{\Theta_n} n^{-1}\sum_i(\hat{m}(x_i,\theta)-m(x_i,\theta))^2 = o_p(n^{-\frac{1}{2}-\delta_0})$. If $\epsilon_n = o(\eta_n)$ for $\eta_n = n^{-\tau}$ with $1/8 \leq \tau \leq 1/4$ and $\epsilon_n n^{\frac{1}{4}+\frac{\delta_0}{2}} \to \infty$ then b) $\sup_{\Theta_{0n}^{\bar{\epsilon}_n,\|\|\cdot\|\|_w}} n^{-1}\left|\sum_i m^2(x_i,\theta) - E[m^2(x_i,\theta)]\right| = o_p(n^{-\frac{1}{2}-\delta_0})$; c) $\sup_{\Theta_{0n}^{\bar{\epsilon}_n,\|\|\cdot\|\|_w}} n^{-1}\sum_i m^2(x_i,\theta) = o_p(\eta_n^2)$ and $\sup_{\Theta_{0n}^{\bar{\epsilon}_n,\|\|\cdot\|\|_w}} n^{-1}\sum_i \hat{m}^2(x_i,\theta) = o_p(\eta_n^2)$.*

Proof of Lemma .1: To establish part a) I study $\sup_{\Theta_n} n^{-1}\sum_i(\hat{m}(x_i,\theta)-m(x_i,\theta))^2$ through the inequality:

$$\sup_{\Theta_n}\frac{1}{n}\sum_{i=1}^n(\hat{m}(x_i,\theta)-m(x_i,\theta))^2 \leq \sup_{\Theta_n}\frac{2}{n}\sum_{i=1}^n(E[v^*(X)-\theta(Z)f_X(X)|x_i] - \hat{E}[v^*(X)-\theta(Z)f_X(X)|x_i])^2 + \sup_{\Theta_n}\frac{2}{n}\sum_{i=1}^n(\hat{E}[\theta(Z)(\hat{f}_X(X)-f_X(X))|x_i])^2 \quad (46)$$

I will first show $\sup_{\Theta_n} n^{-1}\sum_i(E[v^*(X)-\theta(Z)f_X(X)|x_i] - \hat{E}[v^*(X)-\theta(Z)f_X(X)|x_i])^2 = o_p(n^{-\frac{1}{2}-\delta_0})$ by verifying the conditions in Lemma A.1 in Ai & Chen (2003). Assumptions 3.1 and 3.2(i) are implied by our Assumptions 1 and 2(i). Since $v^*(x) = E[\theta_0(Z)|x]f_X(x)$ for some bounded $\theta_0(z)$ and $f_X(x)$ is bounded, it follows that $v^*(x)$ is bounded. Hence, condition (i) is satisfied for $c_1(Z)$ a constant and $c_{1n} = 1$, while condition (ii) is satisfied with $c_2(Z)$

constant by $f_X(x)$ being bounded, $\kappa = 1$ and $||\cdot||_s = ||\cdot||_\infty$. I verify condition (iii) for $\delta_{1n} = n^{-\frac{1}{4}-\frac{\delta_0}{2}}$. Assumption 4(ii) implies $\xi_{0n}\delta_{1n} \to 0$ and therefore to show condition (iii) we need to establish:

$$\left[d_x \log\left(\xi_{1n} n^{\frac{1}{4}+\frac{\delta_0}{2}}\right) + \log\left(N(\xi_{0n}^{-1} n^{-\frac{1}{4}-\frac{\delta_0}{2}}, \Theta_n, ||\cdot||_\infty)\right)\right] \xi_{0n}^2 n^{-\frac{1}{2}+\delta_0} \to 0 \tag{47}$$

Assumption 4(ii) implies $\xi_{1n} n^{-\frac{1}{4}-\frac{\delta_0}{2}} = o(1)$, and therefore for large n, $\log\left(\xi_{1n} n^{\frac{1}{4}+\frac{\delta_0}{2}}\right) \leq (1/2+\delta_0)\log n$. Assumption 4(iii) implies $\log\left(N(\xi_{0n}^{-1} n^{-\frac{1}{4}-\frac{\delta_0}{2}}, \Theta_n, ||\cdot||_\infty)\right) \lesssim k_{1n} \log\left(k_{1n}\xi_{0n} n^{\frac{1}{4}+\frac{\delta_0}{2}}\right) \lesssim k_{1n} \log n$ because $k_{1n}/n \to 0$ and $\xi_{0n} n^{-\frac{1}{4}-\frac{\delta_0}{2}} = o(1)$. Hence, (47) holds if $k_{1n}\xi_{0n}^2 n^{-\frac{1}{2}+\delta_0} \log n = o(1)$, which is implied by Assumption 4(ii). It therefore follows from Lemma A.1 (A) in Ai & Chen (2003) that:

$$\sup_{x,\Theta_n} p^{k_n'}(x)(P'P)^{-1} \sum_{i=1}^n p^{k_n}(x_i)\left(\theta(z_i) f_X(x_i) - E[\theta(Z)|x_i] f_X(x_i)\right) = o_p(n^{-\frac{1}{4}-\frac{\delta_0}{2}}) \tag{48}$$

In addition, Assumption 2(ii) and 3(iv) imply condition (iv) of Lemma A.1 (B) in Ai & Chen (2003) is satisfied for $\delta_{2n} = n^{-\frac{1}{4}-\frac{\delta_0}{2}}$, and therefore we conclude:

$$\sup_{\Theta_n} \frac{1}{n} \sum_{i=1}^n \left[p^{k_n}(x_i)'(P'P)^{-1} \sum_{j=1}^n p^{k_n}(x_j)(v^*(x_j) - E[\theta(Z)|x_j] f_X(x_j)) - (v^*(x_j) - E[\theta(Z)|x_i] f_X(x_j)) \right]^2 = o_p(n^{-\frac{1}{2}-\delta_0}) \tag{49}$$

Combining (48) and (49) we find that $\sup_{\Theta_n} \frac{1}{n}\sum_i (E[v^*(X) - \theta(Z) f_X(X)|x_i] - \hat{E}[v^*(X) - \theta(Z) f_X(X)|x_i])^2 = o_p(n^{-\frac{1}{2}-\frac{\delta_0}{2}})$, hence controlling the first term in (46). I now examine the term $\sup_{\Theta_n} n^{-1}\sum_i (\hat{E}[\theta(Z)(\hat{f}_X(X) - f_X(X))|x_i])^2$ in (50). The second inequality in (50) follows by the Cauchy-Schwarz inequality and the $\theta \in \Theta$ being uniformly bounded. As shown in the proof of Theorem 1 in Newey (1997), $E[n^{-1}\sum_i p^{k_n'}(x_i)(P'P)^{-1} p^{k_n}(x_i)] = O(k_n/n)$, and hence by Markov's inequality $n^{-1}\sum_i p^{k_n'}(x_i)(P'P)^{-1} p^{k_n}(x_i) = O_p(k_n/n)$. Furthermore, Assumption 5 and Theorem 4.2 2) in Bosq (1998) implies $\sup_x E[(\hat{f}_X(x) - f_X(x))^2] \leq \sup_x 2E[(E[\hat{f}_X(x)] - f_X(x))^2] + \sup_x 2E[(E[\hat{f}_X(x)] - \hat{f}_X(x))^2] = O(n^{-2\nu k}) + O(n^{1-\nu d_x})$, and hence Markov's inequality and $k > d_x$ imply the third equality in (50). Assumption 5(iv) gives us the final result in (50).

$$\begin{aligned}\sup_{\Theta_n} \frac{1}{n}\sum_{i=1}^n (\hat{E}[\theta(Z)(\hat{f}_X(X) - f_X(X))|x_i])^2 &= \sup_{\Theta_n} \frac{1}{n}\sum_{i=1}^n \left(p^{k_n'}(x_i)(P'P)^{-1} \sum_{j=1}^n p^{k_n'}(x_j)\theta(z_j)(\hat{f}_X(x_j) - f_X(x_j)) \right)^2 \\ &\lesssim \left[\frac{1}{n}\sum_{i=1}^n p^{k_n'}(x_i)(P'P)^{-1} p^{k_n}(x_i)\right] \left[\sum_{j=1}^n (\hat{f}_X(x_j) - f_X(x_j))^2\right] = O_p\left(\frac{k_n}{n}\right) \times O_p\left(n^{1-\nu d_x}\right) = o_p(n^{-\frac{1}{2}-\delta_0}) \end{aligned} \tag{50}$$

Together, (46), (48), (49) and (50) establish part a) of the Lemma.

For establishing parts b) and c), first note that Assumption 3(iii), $\epsilon_n n^{\frac{1}{4}+\frac{\delta_0}{2}} \to +\infty$ and $||\theta||_w \lesssim ||\theta||_\infty$ imply $\Theta_{0n}^{\bar{\epsilon}_n, ||\cdot||_w} \neq \emptyset$ for n large enough. Let $\mathcal{A}_n = \left\{\alpha(x) = m^2(x,\theta) - E[m^2(X,\theta)] : \theta \in \Theta_{0n}^{\bar{\epsilon}_n, ||\cdot||_w}\right\}$ and note that in order to establish part b) it is sufficient to show $\sup_{\mathcal{A}_n} n^{-1}\left|\sum_i \alpha(x_i)\right| = o_p(n^{-\frac{1}{2}-\delta_0})$. Let D_n be the diameter of $\mathcal{A}_n$ under the metric ρ_n. Then, the first result in (51) follows by a standard maximal inequality for empirical processes, see for example Theorem .2 in Santos (2007). Let $\mathcal{M} = \{m(x,\theta) : \theta \in \Theta\}$ and $\mathcal{A} = \{\alpha(x) = m^2(x,\theta) - E[m^2(X,\theta)] : m(x,\theta) \in \mathcal{M}\}$. Since $\mathcal{A}_n \subseteq \mathcal{A}$ and $\rho_n(\alpha_1,\alpha_2) \leq ||\alpha_1 - \alpha_2||_\infty$, it follows that $N(\epsilon, \mathcal{A}_n, \rho_n) \leq N_{[]}(\epsilon, \mathcal{A}, ||\cdot||_\infty)$. Furthermore, for $\alpha_i(x) \in \mathcal{A}$ with $\alpha_i(x) = m^2(x,\theta_i) - E[m^2(X,\theta_i)]$, the fact that $m(x,\theta) \in \mathcal{M}$ are uniformly bounded implies $||\alpha_1 - \alpha_2||_\infty \lesssim ||m_1 - m_2||_\infty$ and hence Theorem 2.7.11 in van der Vaart & Wellner implies $N_{[]}(\epsilon, \mathcal{A}, ||\cdot||_\infty) \lesssim N_{[]}(\epsilon, \mathcal{M}, ||\cdot||_\infty)$. Together with $\mathcal{M} \subseteq \Lambda_C^\gamma(\mathcal{X})$ by Assumption 3(iv), this gives us the second inequality in (51). Theorem 2.7.1 in van der Vaart & Wellner, $\gamma > d_x/2$, the definition of D_n and Jensen's inequality

in turn implies the third inequality. For the final inequality we exploit that $\mathcal{A}_n \subseteq \mathcal{A}$.

$$E\left[\sup_{\mathcal{A}_n}\sqrt{n}\left|\sum_{i=1}^n \alpha(x_i)\right|\right] \lesssim E\left[\int_0^{D_n}\left[\log N(\epsilon,\mathcal{A}_n,\rho_n)\right]^{\frac{1}{2}}d\epsilon\right] \lesssim E\left[\int_0^{D_n}\left[\log N_{[]}(\epsilon,\Lambda_C^\gamma(\mathcal{X}),||\cdot||_\infty)\right]^{\frac{1}{2}}d\epsilon\right]$$
$$\lesssim \left(E\left[\sup_{\mathcal{A}_n} n^{-1}\sum_{i=1}^n \alpha^2(x_i)\right]\right)^{\frac{1}{2}-\frac{d_x}{4\gamma}} \leq \left(E\left[\sup_{\mathcal{A}} n^{-1}\left|\sum_{i=1}^n \alpha^2(x_i)-E[\alpha^2(X)]\right|\right]+\sup_{\mathcal{A}_n}E[\alpha^2(X)]\right)^{\frac{1}{2}-\frac{d_x}{4\gamma}} \quad (51)$$

Let $\mathcal{A}^2 = \{\omega(x) = \alpha^2(x) - E[\alpha^2(X)] : \alpha(x) \in \mathcal{A}\}$. Since $m(x,\theta) \in \mathcal{M}$ are uniformly bounded, it follows that $N_{[]}(\epsilon,\mathcal{A}^2,||\cdot||_\infty) \lesssim N_{[]}(\epsilon,\mathcal{M},||\cdot||_\infty)$. Hence, Theorems 2.7.1 and 2.5.6 in van der Vaart & Wellner imply $\mathcal{A}^2$ is a Donsker class, and therefore $E\left[\sup_{\mathcal{A}} n^{-1}\left|\sum_i \alpha^2(x_i) - E[\alpha^2(X)]\right|\right] = O_p(n^{-\frac{1}{2}})$. In addition, $m(x,\theta) \in \mathcal{M}$ being uniformly bounded implies $\sup_{\mathcal{A}_n} E[\alpha^2(X)] \lesssim \sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}} E\left[\left(E[v^*(X)/f_X(X) - \theta(Z)|X]\right)^2 f_X^2(X)\right] = \epsilon_n^2$. Therefore, it follows from Markov's inequality and (51) that $\sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}} n^{-1}\left|\sum_i m^2(x_i,\theta) - E[m^2(X,\theta)]\right| = O_p(n^{-\frac{1}{2}}(n^{-\frac{1}{2}} + \epsilon_n^2)^{\frac{1}{2}-\frac{d_x}{4\gamma}})$. Together with $\epsilon_n = o(n^{-\frac{1}{8}})$ and $\delta_0 < (2\gamma - d_x)/16\gamma$ this concludes showing part b) of the Lemma.

I now establish part c). In (52), the first inequality follows from $\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w} \subseteq \Theta_0^{\bar{\epsilon}_n,||\cdot||_w}$ for $\Theta_0^{\bar{\epsilon}_n,||\cdot||_w} = \{\theta \in \Theta : \inf_{\Theta_0}||\theta - \theta_0||_w \leq \epsilon_n\}$. Since $E[\theta(Z)|x] = v^*(x)/f_X(x)$ for all $\theta \in \Theta_0$, it follows that $E[m^2(X,\theta)] = \inf_{\Theta_0}||\theta-\theta_0||_w^2 \leq \epsilon_n^2$. Therefore, part b) of the lemma and $\epsilon_n = o(\eta_n)$, imply the final result in (52).

$$\sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}} n^{-1}\sum_{i=1}^n m^2(x_i,\theta) \leq \sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}} n^{-1}\left|\sum_{i=1}^n m^2(x_i,\theta) - E[m^2(x_i,\theta)]\right| + \sup_{\Theta_0^{\bar{\epsilon}_n,||\cdot||_w}} E[m^2(X,\theta)] = o_p(\eta_n^2) \quad (52)$$

For the second claim of part c), note that the first inequality in (53) is implied by $\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w} \subseteq \Theta_n$. Part a) and (52) establish the final equality in (53).

$$\sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}} n^{-1}\sum_{i=1}^n \hat{m}^2(x_i,\theta) \leq 2\sup_{\Theta_n} n^{-1}\sum_{i=1}^n \left(\hat{m}(x_i,\theta) - m(x_i,\theta)\right)^2 + 2\sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}} n^{-1}\sum_{i=1}^n m^2(x_i,\theta) = o_p(\eta_n^2) \quad (53)$$

Results (52) and (53) verify claim c) and hence conclude the proof of the Lemma. ■

Corollary .1. *Let $\bar{Q}_n = n^{-1}\sum_i m^2(x_i,\theta)$. Under Assumptions 1, 2(i)-(ii), 3(i)-(iv), 4(i)-(iii) and 5(i)-(iv), a) $\sup_{\Theta_n}|Q_n(\theta) - Q(\theta)| = o_p(n^{-\frac{1}{4}-\frac{\delta_0}{2}})$. Furthermore, if in addition $\epsilon_n = o(\eta_n)$ for $\eta_n = n^{-\tau}$ with $1/8 \leq \tau \leq 1/4$ and $\epsilon_n n^{\frac{1}{4}+\frac{\delta_0}{2}} \to +\infty$, then it follows that b) $\sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}}|Q_n(\theta) - Q(\theta)| = o_p(\eta_n n^{-\frac{1}{4}-\frac{\delta_0}{2}})$*

Proof of Corollary .1: I first show part a) and begin by examining $\sup_{\Theta_n}|Q_n(\theta) - \bar{Q}_n(\theta)|$. In (54), the second equality follows from the Cauchy-Schwarz inequality, $m(x,\theta)$ being uniformly bounded because $m(x,\theta) \in \Lambda_C^\gamma(\mathcal{X})$ by Assumption 3(iv) and part a) of Lemma (.1).

$$\sup_{\Theta_n}|Q_n(\theta) - \bar{Q}_n(\theta)| \leq \sup_{\Theta_n}\frac{1}{n}\sum_{i=1}^n |(\hat{m}(x_i,\theta) - m(x_i,\theta))(\hat{m}(x_i,\theta) + m(x_i,\theta))| = o_p(n^{-\frac{1}{4}-\frac{\delta_0}{2}}) \quad (54)$$

Let $\mathcal{M}^2 = \{m^2(x,\theta) : \theta \in \Theta\}$. Since the $m(x,\theta)$ are uniformly bounded, $\mathcal{M}^2$ is Lipschitz in $\mathcal{M} \subseteq \Lambda_C^\gamma(\mathcal{X})$. Therefore, by Theorem 2.10.6 in van der Vaart & Wellner $\mathcal{M}^2$ is Donsker and $\sup_\Theta|\bar{Q}_n(\theta) - Q(\theta)| = O_p(n^{-\frac{1}{2}})$. Hence, (54) implies $\sup_{\Theta_n}|Q_n(\theta) - Q(\theta)| \leq \sup_{\Theta_n}|Q_n(\theta) - \bar{Q}_n(\theta)| + \sup_\Theta|\bar{Q}_n(\theta) - Q(\theta)| = o_p(n^{-\frac{1}{4}-\frac{\delta_0}{2}})$, establishing part a).

I now show part b). First note that since the Cauchy Schwarz inequality implies that $\sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}}\left|Q_n(\theta) - \bar{Q}_n(\theta)\right| \leq \left[\sup_{\Theta_n}\sum_i(\hat{m}(x_i,\theta) - m(x_i,\theta))^2\right]^{\frac{1}{2}}\left[\sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}}\sum_i(\hat{m}(x_i,\theta) + m(x_i,\theta))^2\right]^{\frac{1}{2}}$, parts a) and c) of Lemma .1 imply that $\sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}}\left|Q_n(\theta) - \bar{Q}_n(\theta)\right| = o_p(\eta_n n^{-\frac{1}{4}-\frac{\delta_0}{2}})$. Together with part b) of Lemma .1 and $\eta_n = n^{-\tau}$ for $1/8 \leq \tau \leq 1/4$, this implies the final result in (55)

$$\sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}}|Q_n(\theta) - Q(\theta)| \leq \sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}}|Q_n(\theta) - \bar{Q}_n(\theta)| + \sup_{\Theta_{0n}^{\bar{\epsilon}_n,||\cdot||_w}}|\bar{Q}_n(\theta) - Q(\theta)| = o_p(\eta_n n^{-\frac{1}{4}-\frac{\delta_0}{2}}) \quad (55)$$

which concludes the proof of the Corollary. ■

Proof of Theorem 3.1: To establish the Theorem, I first show $||\hat{\theta}-\theta_0||_w = o_p(n^{-\frac{1}{8}-\frac{\delta_0}{8}})$ and then use this result to refine the rate of convergence. In (56), the first inequality holds for any $\epsilon > 0$ and $\theta_{0n} = \inf_{\Theta_n} ||\theta_0 - \theta_n||_\infty$ because by definition $\hat{\theta} \in \arg\min_{\Theta_n} Q_n(\theta)$. In turn, since $\sup_{\Theta_n} |Q_n(\theta) - Q(\theta)| = o_p(n^{-\frac{1}{4}-\frac{\delta_0}{2}})$ by part a) of Corollary .1, the second inequality in (56) follows. The final result in (56) is implied by noting that $Q(\theta) = E[(f_X(x)E[\theta(Z)|X] - v^*(X))^2] = ||\theta-\theta_0||_w^2$ because $E[\theta_0(Z)|x] = v^*(x)/f_X(x)$ and that $||\theta_0 - \theta_{0n}||_w^2 \leq ||\theta_0-\theta_{0n}||_\infty^2 = o(n^{-\frac{1}{2}-\delta_0})$ by Assumption 3(iii).

$$
\begin{aligned}
P\left(||\hat{\theta}-\theta_0||_w \geq n^{-\frac{1}{8}-\frac{\delta_0}{8}}\epsilon\right) &\leq P\left(\inf_{\theta\in\Theta_n:||\theta-\theta_0||_w \geq n^{-\frac{1}{8}-\frac{\delta_0}{8}}\epsilon} Q_n(\theta) \leq Q_n(\theta_{0n})\right) \\
&\leq P\left(\inf_{\theta\in\Theta_n:||\theta-\theta_0||_w \geq n^{-\frac{1}{8}-\frac{\delta_0}{8}}\epsilon} Q_n(\theta) \leq Q_n(\theta_{0n}) \bigcap \sup_{\Theta_n} |Q_n(\theta)-Q(\theta)| \leq n^{-\frac{1}{4}-\frac{\delta_0}{2}}\right) + o(1) \\
&\leq P\left(\inf_{\theta\in\Theta_n:||\theta-\theta_0||_w \geq n^{-\frac{1}{8}-\frac{\delta_0}{8}}\epsilon} Q(\theta) \leq Q(\theta_{0n}) + 2n^{-\frac{1}{4}-\frac{\delta_0}{2}}\right) + o(1) = o(1) \quad (56)
\end{aligned}
$$

The derivations in (56) imply that $||\hat{\theta}-\theta_0||_w = o_p(n^{-\frac{1}{8}-\frac{\delta_0}{8}})$. To improve on this rate, let $\delta_{0n} = n^{-\frac{1}{8}-\frac{\delta_0}{8}}\epsilon$ and $\delta_{1n} = n^{-\frac{1}{8}-\frac{1}{16}-\frac{\delta_0}{8}}\epsilon$ for any $\epsilon > 0$. Since we have already shown $||\hat{\theta}-\theta||_w = o_p(n^{-\frac{1}{8}-\frac{\delta_0}{8}}) = o_p(\delta_{0n})$ we can use arguments similar to (56) to derive the first inequality in (57). Using part b) of Corollary .1 gives us the second inequality in (57). The final result is follows by $Q(\theta) = ||\theta-\theta_0||_w^2$, Assumption 3(iii) implying $||\theta_0-\theta_{0n}||_w^2 \leq ||\theta_0-\theta_{0n}||_\infty^2 = o(n^{-\frac{1}{2}-\delta_0})$ and recalling that $\delta_{1n}^2 = n^{-\frac{1}{4}-\frac{1}{8}-\frac{\delta_0}{4}}$.

$$
\begin{aligned}
P\left(||\hat{\theta}-\theta_0||_w \geq \delta_{1n}\right) &\leq P\left(\inf_{\theta\in\Theta_n:\delta_{0n}\geq||\theta-\theta_0||_w\geq\delta_{1n}} Q_n(\theta) \leq Q_n(\theta_{0n})\right) + o(1) \\
&\leq P\left(\inf_{\theta\in\Theta_n:\delta_{0n}\geq||\theta-\theta_0||_w\geq\delta_{1n}} Q_n(\theta) \leq Q_n(\theta_{0n}) \bigcap \sup_{\Theta_{0n}^{\bar{\delta}_{0n},||\cdot||_w}} |Q_n(\theta)-Q(\theta)| \leq n^{-\frac{1}{4}-\frac{1}{8}-\frac{\delta_0}{2}}\right) + o(1) \\
&\leq P\left(\inf_{\theta\in\Theta_n:\delta_{0n}\geq||\theta-\theta_0||_w\geq\delta_{1n}} Q(\theta) \leq Q(\theta_{0n}) + 2n^{-\frac{1}{4}-\frac{1}{8}-\frac{\delta_0}{2}}\right) + o(1) = o(1) \quad (57)
\end{aligned}
$$

Hence, (57) implies $||\hat{\theta}-\theta_0||_w = o_p(n^{-\frac{1}{8}-\frac{1}{16}-\frac{\delta_0}{8}})$. It is now possible to improve on this rate again by letting $\delta_{0n} = n^{-\frac{1}{8}-\frac{1}{16}-\frac{\delta_0}{8}}\epsilon$, $\delta_{1n} = n^{-\frac{1}{8}-\frac{1}{16}-\frac{1}{32}-\frac{\delta_0}{8}}\epsilon$ and repeating the arguments in (57), which shows $||\hat{\theta}-\theta||_w = o_p(n^{-\frac{1}{8}-\frac{1}{16}-\frac{1}{32}-\frac{\delta_0}{2}})$. By repeating this argument a large, but finite, number of times we can establish that $||\hat{\theta}-\theta_0||_w = o_p(n^{-\frac{1}{4}})$. ■

Lemma .2. *Let $u(z) = \pm\tilde{v}(z)$ and $u_n(z) = \arg\min_{\Theta_n} ||\theta_n - \tilde{v}||_\infty$. Under Assumptions 1, 2(i)-(ii), 3(i)-(vi), 4(i)-(iv) and 5(i)-(v), if $\hat{\theta} \in \Theta_n$ satisfies $||\hat{\theta}-\theta_0||_w = o_p(n^{-\frac{1}{4}})$, then:*

a) $n^{-1}\sum_i \hat{E}[u_n(Z)\hat{f}_X(X)|x_i]\hat{m}(x_i,\hat{\theta}) = n^{-1}\sum_i f_X(x_i)E[u(Z)|x_i]\hat{m}(x_i,\hat{\theta}) + o_p(n^{-\frac{1}{2}})$.

b) If in addition $||\hat{\theta}-\theta_0||_\infty = o_p(1)$, *then* $n^{-1}\sum_i E[u(Z)f_X(X)|x_i](\hat{m}(x_i,\hat{\theta})-\hat{m}(x_i,\theta_0)) = \langle u, \theta_0-\hat{\theta}\rangle_w + o_p(n^{-\frac{1}{2}})$.

c) $n^{-1}\sum_i E[u(Z)f_X(X)|x_i]\hat{m}(x_i,\theta_0) = n^{-1}\sum_i E[u(Z)f_X(X)|x_i](v^*(x_i) - \theta_0(z_i)\hat{f}_X(x_i)) + o_p(n^{-\frac{1}{2}})$.

d) $n^{-1}\sum_i E[\tilde{v}(Z)f_X(X)|x_i]\theta_0(z_i)(f_X(x_i)-\hat{f}_X(x_i)) = n^{-1}\sum_i E[\tilde{v}(Z)v^*(X)f_X(X)] - E[\tilde{v}(Z)|x_i]v^*(x_i)f_X(x_i) + o_p(n^{-\frac{1}{2}})$

Proof of Lemma .2: For a), we first examine $n^{-1}\sum_i \left(\hat{E}[\hat{f}_X(X)u_n(Z)|x_i] - \hat{E}[f_X(X)u_n(Z)|x_i]\right)\hat{m}(x_i,\hat{\theta})$. The first result in (58) follows from the Cauchy-Schwarz inequality. Since by assumption $\hat{\theta} \in \Theta_{0n}^{\bar{\epsilon},||\cdot||_w}$ for $\epsilon_n = o(n^{-\frac{1}{4}})$ with

probability tending to one, part c) of Lemma .1 and (50) imply the second result in (58).

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{E}[\hat{f}_X(x_i)u_n(Z)|x_i]-\hat{E}[f_X(X)u_n(Z)|x_i]\right)\hat{m}(x_i,\hat{\theta})\right| \\ \leq\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{E}[u_n(Z)(\hat{f}_X(X)-f_X(X))|x_i]\right)^2\right]^{\frac{1}{2}}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{m}^2(x_i,\hat{\theta})\right]^{\frac{1}{2}}=o_p(n^{-\frac{1}{2}}) \quad (58)$$

Now we examine $n^{-1}\sum_i\left(\hat{E}[f_X(X)u_n(Z)|x_i]-E[f_X(X)u_n(Z)|x_i]\right)\hat{m}(x_i,\hat{\theta})$. In (59) the first result follows by the Cauchy-Schwarz inequality. The same arguments as in (48) and (49) but setting $v^*(x)=0$ together with Assumption 3(vi) imply $n^{-1}\sum_i\left(\hat{E}[u_n(Z)f_X(X)|x_i]-E[u_n(Z)f_X(X)|x_i]\right)^2=o_p(n^{-\frac{1}{2}})$. Together with part c) of Lemma .1, this implies the final equality in (59).

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{E}[f_X(X)u_n(Z)|x_i]-E[f_X(X)u_n(Z)|x_i]\right)\hat{m}(x_i,\hat{\theta})\right| \\ \leq\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{E}[f_X(X)u_n(Z)|x_i]-E[f_X(X)u_n(Z)|x_i]\right)^2\right]^{\frac{1}{2}}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{m}^2(x_i,\hat{\theta})\right]^{\frac{1}{2}}=o_p(n^{-\frac{1}{2}}) \quad (59)$$

Next, we examine the term $n^{-1}\sum_i f_X(x_i)\left(E[u_n(Z)|x_i]-E[u(Z)|x_i]\right)\hat{m}(x_i,\hat{\theta})$. In (60), we apply the Cauchy-Schwarz inequality to derive the first result. Since $u\in\Theta$, it follows by Assumption 3(iii) that $||u_n-u||_\infty=o(n^{-\frac{1}{4}})$. Therefore, the inequality $E[u_n(Z)-u(Z)|x_i]^2\leq||u_n-u||_\infty^2$ implies $n^{-1}\sum_i\left(E[u_n(Z)|x_i]-E[u(Z)|x_i]\right)^2=o(n^{-\frac{1}{2}})$. The final result in (60) then follows by $f_X(x)$ bounded and $n^{-1}\sum_i\hat{m}^2(x_i,\hat{\theta})=o_p(n^{-\frac{1}{2}})$

$$\left|\frac{1}{n}\sum_{i=1}^{n}f_X(X)\left(E[u_n(Z)|x_i]-E[u(Z)|x_i]\right)\hat{m}(x_i,\hat{\theta})\right| \\ \leq\left[\frac{1}{n}\sum_{i=1}^{n}\left(E[u_n(Z)|x_i]-E[u(Z)|x_i]\right)^2\right]^{\frac{1}{2}}\left[\frac{1}{n}\sum_{i=1}^{n}f_X^2(x_i)\hat{m}^2(x_i,\hat{\theta})\right]^{\frac{1}{2}}=o_p(n^{-\frac{1}{2}}) \quad (60)$$

Combining (58), (59) and (60), concludes the proof of part a) of the Lemma.

We now establish part b). The first result in (61) is obtained by rearranging terms. The final equality in (61) follows by exchanging the order of summation, where $\hat{E}[E[u(Z)f_X(X)|X]|x_j]=p^{k_n'}(x_j)(P'P)^{-1}\sum_i p^{k_n}(x_i)E[u(Z)f_X(X)|x_i]$.

$$\frac{1}{n}\sum_{i=1}^{n}E[u(Z)f_X(X)|x_i](\hat{m}(x_i,\hat{\theta})-\hat{m}(x_i,\theta_0)) \\ =\frac{1}{n}\sum_{i=1}^{n}E[u(Z)f_X(X)|x_i]p^{k_n'}(x_i)(P'P)^{-1}\sum_{j=1}^{n}p^{k_n}(x_j)(\theta_0(z_j)-\hat{\theta}(z_j))\hat{f}_X(x_j) \\ =\frac{1}{n}\sum_{j=1}^{n}(\theta_0(z_j)-\hat{\theta}(z_j))\hat{E}[E[u(Z)f_X(X)|X]|x_j]\hat{f}_X(x_j) \quad (61)$$

Since $E[u(Z)f_X(X)|x]\in\Lambda_C^\gamma(\mathcal{X})$ by Assumption 3(vi), Theorem 1 in Newey (1997) implies that under Assumptions 1-2, 3(vi) and 4(iv) $||\hat{E}[E[u(Z)f_X(X)|X]|x]-E[u(Z)f_X(X)|x]||_\infty=o_p(1)$. A straightforward extension of Theorem 2.8 in Pagan & Ullah (1999) to X multidimensional, shows that Assumptions 1 and 5(i)-(v) imply $||\hat{f}_X-f_X||_\infty=o_p(1)$. Therefore, $E[u(Z)|x]f_X(x)$ and $f_X(x)$ being bounded implies that for M large enough, $P(||\hat{E}[E[u(Z)f_X(X)|x]]\hat{f}_X(x)||_\infty>M)\to 0$. Hence, for $\mathcal{F}_n=\{(\theta_0(z)-\theta(z))\hat{E}[E[u(Z)f_X(X)|X]|x]\hat{f}_X(x):\theta\in\Theta\}$ and M large enough we have with probability arbitrarily close to one $N_{[]}(\epsilon,\mathcal{F}_n,||\cdot||_\infty)\leq N_{[]}(\frac{\epsilon}{M},\Theta,||\cdot||_\infty)\lesssim e^{\epsilon^{-\frac{d_x}{\gamma}}}$. Thus, $n^{-\frac{1}{2}}\sum_i\left((\theta_0(z_i)-\theta(z_i))\hat{E}[E[u(Z)f_X(X)|X]|x_i]\hat{f}_X(x_i)-E[(\theta_0(Z)-\theta(Z))\hat{E}[E[u(Z)f_X(X)|X]|X]\hat{f}_X(X)]\right)$ is

asymptotically tight in $l^\infty(\Theta)$ by Theorem 2.11.23 in van der Vaart & Wellner. Since in addition we have that $\sup_{x,z}|(\theta_0(z)-\hat{\theta}(z))\hat{E}[E[u(Z)f_X(X)|X]|x]\hat{f}_X(x)| = o_p(1)$ by virtue of $\sup_x|\hat{E}[E[u(Z)f_X(X)|X]|x]\hat{f}_X(x)| = O_p(1)$ as shown, and $||\hat{\theta}-\theta_0||_\infty = o_p(1)$ by assumption, we conclude using (61) that:

$$\frac{1}{n}\sum_{i=1}^{n} E[u(Z)f_X(X)|x_i](\hat{m}(x_i,\hat{\theta})-\hat{m}(x_i,\theta_0)) = E[(\theta_0(Z)-\hat{\theta}(Z))\hat{E}[E[u(Z)f_X(X)|X]|X]\hat{f}_X(X)] + o_p(n^{-\frac{1}{2}}) \quad (62)$$

We now analyze the right hand side of (62). Since $E[u(Z)f_X(X)|x] \in \Lambda_C^\gamma(\mathcal{X})$ by Assumption 3(vi), Theorem 1 in Newey (1997) implies under Assumption 1, 2(i)-(ii), 3(vi) and 4(iv), $E[(\hat{E}[E[u(Z)f_X(X)|X]|X]-E[u(Z)f_X(X)|X])^2] = O_p(k_n/n+k_n^{-2\gamma/d_x}) = o_p(n^{-\frac{1}{2}})$ by Assumptions 4(iv) and 2(ii). As argued, for M large enough $P(||\hat{f}_X||_\infty > M) \to 0$, so that with arbitrary large probability $|E[(\theta_0(Z)-\hat{\theta}(Z))(\hat{E}[E[u(Z)f_X(X)|X]|X]-E[u(Z)f_X(X)|X])\hat{f}_X(X)]| \lesssim ||\hat{\theta}-\theta||_w(E[(\hat{E}[E[u(Z)f_X(X)|X]]-E[u(Z)f_X(X)|X])^2])^{\frac{1}{2}} = o_p(n^{-\frac{1}{2}})$ as $||\hat{\theta}-\theta||_w = o_p(n^{-\frac{1}{4}})$ by assumption. Together with (62), this implies the first equality in (63). Furthermore, Assumption 1 and 5(i)-(iv) imply by Theorem 4.2 2) in Bosq (1998) that $\sup_x E[(\hat{f}_X(x)-f_X(x))^2] = o(n^{-\frac{1}{2}})$. The Cauchy-Schwarz inequality, $E[u(Z)f_X(X)|x]$ being bounded and $||\hat{\theta}-\theta_0||_w = o_p(n^{-\frac{1}{4}})$ in turn imply the second equality in (63).

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} E[u(Z)f_X(X)|x_i](\hat{m}(x_i,\hat{\theta})-\hat{m}(x_i,\theta_0)) &= E[(\theta_0(Z)-\hat{\theta}(Z))E[u(Z)f_X(X)|X]\hat{f}_X(X)] + o_p(n^{-\frac{1}{2}}) \\ &= E[(\theta_0(Z)-\hat{\theta}(Z))E[u(Z)|X]f_X^2(X)] + o_p(n^{-\frac{1}{2}}) = \langle\theta_0-\hat{\theta},u\rangle_w + o_p(n^{-\frac{1}{2}}) \quad (63)\end{aligned}$$

Hence, (63) concludes the proof of part b) of the Lemma.

I now establish part c) of the Lemma. In (64) we use arguments identical to those in (61) to obtain the first equality for $\hat{E}[E[u(Z)f_X(X)|X]|x_i] = p^{k_n'}(x_i)(P'P)^{-1}\sum_j p^{k_n}(x_j)E[u(Z)f_X(X)|x_j]$. As shown in the derivation of (63), $E[(\hat{E}[E[u(Z)f_X(X)|X]|X]-E[u(Z)f_X(X)|X])^2] = o_p(n^{-\frac{1}{2}})$, and hence since $E[\theta_0(Z)f_X(X)-v^*(X)|x_i] = 0$, it follows by Markov's inequality that $n^{-1}\sum_i\left(\hat{E}[E[u(Z)f_X(X)|X]|x_i]-E[u(Z)f_X(X)|x_i]\right)(v^*(x_i)-\theta_0(z_i)f_X(x_i)) = o_p(n^{-\frac{1}{2}})$. In addition, as argued in (63), $\sup_x E[(\hat{f}_X(X)-f_X(X))^2] = o_p(n^{-\frac{1}{2}})$. Hence, θ_0 bounded, the Cauchy Schwarz inequality and Markov's inequality imply $n^{-1}\sum_i\left(\hat{E}[E[u(Z)|X]|x_i]-E[u(Z)|x_i]\right)\left(f_X(x_i)-\hat{f}_X(x_i)\right)\theta_0(z_i) = o_p(n^{-\frac{1}{2}})$, and hence the final result in (64) follows.

$$\begin{aligned}n^{-1}\sum_{i=1}^{n} E[u(Z)f_X(X)|x_i]\left(\hat{m}(x_i,\theta_0)-(v^*(x_i)-\theta_0(z_i)\hat{f}_X(x_i))\right] \\ = n^{-1}\sum_{i=1}^{n}\left(\hat{E}[E[u(Z)f_X(X)|X]|x_i]-E[u(Z)f_X(X)|x_i]\right)(v^*(x_i)-\theta_0(z_i)f_X(x_i)) \\ + n^{-1}\sum_{i=1}^{n}\left(\hat{E}[E[u(Z)f_X(X)|X]|x_i]-E[u(Z)f_X(X)|x_i]\right)\left(f_X(x_i)-\hat{f}_X(x_i)\right)\theta_0(z_i) = o_p(n^{-\frac{1}{2}}) \quad (64)\end{aligned}$$

Thus, (64) establishes the third claim in the Lemma.

In order to show part d), let $g(x,z) = E[\tilde{v}(Z)f_X(X)|x]\theta_0(z)$ and define the following kernel for a U-Statistic:

$$\begin{aligned}H_n(x_i,z_i,x_j,z_j) = g(x_i,z_i)\left(\int K\left(\frac{x_i-x}{h}\right)f_X(x)dx - K\left(\frac{x_i-x_j}{h}\right)\right) \\ + g(x_j,z_j)\left(\int K\left(\frac{x_j-x}{h}\right)f_X(x)dx - K\left(\frac{x_j-x_i}{h}\right)\right) \quad (65)\end{aligned}$$

Note that since $f_X(x)$, $\tilde{v}(z)$, $\theta_0(z) \in \Theta$ they are bounded, which implies $|g(x,z)(E[\hat{f}_X(x)]-f_X(x))]| \lesssim \sup_x|E[\hat{f}_X(x)]-f_X(x)|$. Furthermore, under Assumption 5(i)-(iv) by Theorem 4.2 2) in Bosq (1998) we have $\sup_x|E[\hat{f}_X(x)]-f_X(x)| =$

$O(h^k) = o(n^{-\frac{1}{2}})$ by Assumption 5(i)-(ii), which implies the first equality in (66). The second equality in (66) can be obtained by rearranging terms, where $H_n(x_i, z_i, x_j, z_j)$ is the U-Statistic Kernel defined in (66).

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} g(x_i, z_i)(f_X(x_i) - \hat{f}_X(x_i)) = n^{-\frac{1}{2}} \sum_{i=1}^{n} g(x_i, z_i)(E[\hat{f}_X(x_i)|x_i] - \hat{f}_X(x_i)) + o_p(1)$$
$$= \frac{1}{n^{\frac{3}{2}} h^{d_x}} \sum_{i=1}^{n} \sum_{j>i} H_n(x_i, z_i, x_j, z_j) + o_p(1) \quad (66)$$

In (67) I show $(n^{\frac{3}{2}} h^{d_x})^{-1} \sum_i \sum_{j>i} H_n(x_i, z_i, x_j, z_j) = (n^{\frac{3}{2}} h^{d_x})^{-1} \sum_i \sum_{j>i} E[H_n(x_i, z_i, X, Z)|x_i, z_i] + o_p(1)$. The first equality in (67) follows by noting that all crossterms have zero expectation. Standard calculations can be used to show $E[H_n^2(X_i, Z_i, X_j, Z_j)] = O(h^{d_x})$ and $E[(E[H_n(X_i, Z_i, X_j, Z_j)|X_i, Z_i])^2] = O(h^{2d_x})$. Combining both results, implies the last equality in (67).

$$E\left[\left(\frac{1}{n^{\frac{3}{2}} h^{d_x}} \sum_{i=1}^{n} \sum_{j>i} H_n(x_i, z_i, x_j, z_j) - E[H_n(x_i, z_i, X, Z)|x_i, z_i]\right)^2\right]$$
$$= \frac{1}{n^3 h^{2d_x}} \sum_{i=1}^{n} \sum_{j>i} E\left[\left(H_n(x_i, z_i, x_j, z_j) - E[H_n(x_i, z_i, X, Z)|x_i, z_i]\right)^2\right]$$
$$= \frac{n-1}{n^2 h^{2d_x}} E\left[\left(H_n(x_i, z_i, x_j, z_j) - E[H_n(x_i, z_i, X, Z)|x_i, z_i]\right)^2\right] = O(n h^{d_x}) \quad (67)$$

Therefore, Markov's inequality and $nh^{d_x} \to \infty$ together with (66) imply the first equality in (68). The second result in (68) can be obtained by rearranging terms and noticing that the expectation of the first term in $H_n(x_i, z_i, x_j, z_j)$ conditional on (x_i, z_i) is zero.

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} g(x_i, z_i)(f_X(x_i) - \hat{f}_X(x_i)) = \frac{1}{n^{\frac{3}{2}} h^{d_x}} \sum_{i=1}^{n} \sum_{j>i} E[H_n(x_i, z_i, X, Z)|x_i, z_i] + o_p(1)$$
$$= \frac{1}{n^{\frac{1}{2}} h^{d_x}} \sum_{i=1}^{n} \int g(x_j, z_j) \left(\int K\left(\frac{x_j - x}{h}\right) f_X(x) dx - K\left(\frac{x_j - x_i}{h}\right) \right) f_{XZ}(x_j, z_j) dx_j dz_j + o_p(1) \quad (68)$$

Assumptions 5(i)-(iv), Theorem 4.2 2) in Bosq (1998) imply $\sup_x |E[\hat{f}_X(x)] - f_X(x)| = \sup_x |h^{-d_x} \int K\left(\frac{x-u}{h}\right) f_X(u) du - f_X(x)| = O(h^k) = o(n^{-\frac{1}{2}})$. Since $g(x, z) = E[\tilde{v}(Z)|x] f_X(x) \theta_0(z)$ and $f_{XZ}(x, z)$ have uniformly bounded derivatives up to order k with respect to x by Assumption 5(iii) and $\theta_0(z)$ bounded, so does $g(x, z) f_{XZ}(x, z)$. Therefore, a standard Taylor expansion argument, as in Theorem 4.2 2) in Bosq (1998), can be used to establish that $\sup_{xz} |h^{-d_x} \int g(u, z) f_{XZ}(u, z) K\left(\frac{u-x}{h}\right) du - g(x, z) f_{XZ}(x, z)| = O(h^k) = o(n^{-\frac{1}{2}})$. Combining these two results with (69) gives us the first equality in (69). The second result is implied by $E[\theta_0(Z)|x] = v^*(x)/f_X(x)$.

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} g(x_i, z_i)(f_X(x_i) - \hat{f}_X(x_i)) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int g(x_j, z_j) f_X(x_j) f_{XZ}(x_j, z_j) dx_j dz_j - \int g(x_i, z_j) f_{XZ}(x_i, z_j) dz_j + o_p(1)$$
$$n^{-\frac{1}{2}} \sum_{i=1}^{n} E[\tilde{v}(Z) v^*(X) f_X(X)] - E[\tilde{v}(Z)|x_i] v^*(x_i) f_X(x_i) + o_p(1) \quad (69)$$

which establishes the proof of part d) of the Lemma. ■

Proof of Theorem 3.2: In order to establish that $n^{-\frac{1}{2}} \sum_i (y_i \hat{\theta}(z_i) - \langle v^*, m_0 \rangle)$ is $O_p(1)$, we begin in (70) by decomposing the expression into 2 terms. We use that $\langle v^*, m_0 \rangle = E[Y \theta_0(Z)]$ for any $\theta_0 \in \Theta_0$ to obtain the first equality in (70). Define the class of functions $\mathcal{F} = \{y\theta(z) : \theta \in \Theta\}$. Since $E[Y^2(\theta_1(Z) - \theta_2(Z))^2] \leq E[Y^2] ||\theta_1 - \theta_2||_\infty^2$, it follows that $N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_{\mathcal{L}^2}) \leq N_{[]}(\epsilon / E[Y^2], \Theta, ||\cdot||_\infty) \leq N_{[]}(\epsilon / E[Y^2], \Lambda_C^\omega(\mathcal{Z}), ||\cdot||_\infty) \lesssim e^{\epsilon^{-\frac{d_z}{\omega}}}$, by Theorem 2.7.1 in

van der Vaart & Wellner (2000). Therefore, Theorem 2.5.6 in van der Vaart & Wellner (2000) implies $\mathcal{F}$ is Donsker, and hence $\sup_\Theta n^{-\frac{1}{2}} \sum_i y_i\theta(z_i) - E[Y\theta(Z)] = O_p(1)$. Together with $E[Y(\theta_1(Z) - \theta_2(Z))] = \langle \tilde{v}, \theta_1 - \theta_2\rangle_w$, this gives us the second equality in (70). The last equality in (70) follows by (62) and (63).

$$n^{-\frac{1}{2}} \sum_{i=1}^n y_i\hat{\theta}(z_i) - \langle v^*, m_0\rangle = n^{-\frac{1}{2}} \sum_{i=1}^n (y_i\hat{\theta}(z_i) - E[Y\hat{\theta}(Z)]) + n^{\frac{1}{2}} E[Y(\hat{\theta}(Z) - \theta_0(Z))]$$
$$= n^{\frac{1}{2}} \langle \tilde{v}, \hat{\theta} - \theta_0\rangle_w + O_p(1) = n^{\frac{1}{2}} E[(\hat{\theta}(Z) - \theta_0(Z))\hat{E}[E[\tilde{v}(Z)f_X(X)|X]|X]\hat{f}_X(X)] + O_p(1) \quad (70)$$

To conclude, we need to show $n^{\frac{1}{2}} E[(\theta_0(Z) - \hat{\theta}(Z))\hat{E}[E[\tilde{v}(Z)f_X(X)|X]|X]\hat{f}_X(X)]$ is bounded in probability. Define the class of functions $\mathcal{F}_n = \{(\theta_0(z) - \theta(z))\hat{E}[E[\tilde{v}(Z)f_X(X)|X]|x]\hat{f}_X(x) : \theta \in \Theta\}$. As shown in the derivation of (62), $n^{-\frac{1}{2}} \sum_i \Big((\theta_0(z_i) - \theta(z_i))\hat{E}[E[\tilde{v}(Z)f_X(X)|X]|x_i]\hat{f}_X(x_i) - E[(\theta_0(Z) - \theta(Z))\hat{E}[E[\tilde{v}(Z)f_X(X)|X]|X]\hat{f}_X(X)]\Big)$ is asymptotically tight in $l^\infty(\Theta)$, which implies the first equality in (71). The second equality in (71) follows from (61). It follows from parts c) and d) of Lemma .2, that $n^{-\frac{1}{2}} \sum_i f_X(x_i)E[\tilde{v}(Z)|x_i]\hat{m}(x_i, \theta_0) = n^{-\frac{1}{2}} \sum_i E[\tilde{v}(Z)v^*(X)f_X(X)] - E[\tilde{v}(Z)|x_i]\theta_0(z_i)f_X^2(x_i) + o_p(1)$. Therefore, the central limit theorem and part a) of Lemma .2 imply the final equality in (71) for $\tilde{v}_n(Z) = \arg\min_{\Theta_n} ||\tilde{v} - \theta_n||_\infty$.

$$n^{\frac{1}{2}} E[(\theta_0(Z) - \hat{\theta}(Z))\hat{E}[E[\tilde{v}(Z)f_X(X)|X]|X]\hat{f}_X(X)] = n^{-\frac{1}{2}} \sum_{i=1}^n (\theta_0(z_i) - \hat{\theta}(z_i))\hat{E}[E[\tilde{v}(Z)f_X(X)|X]|x_i]\hat{f}_X(x_i) + O_p(1)$$
$$= n^{-\frac{1}{2}} \sum_{i=1}^n f_X(x_i)E[\tilde{v}(Z)|x_i](\hat{m}(x_i, \hat{\theta}) - \hat{m}(x_i, \theta_0)) + O_p(1) = n^{-\frac{1}{2}} \sum_{i=1}^n \hat{E}[\hat{f}_X(X)\tilde{v}_n(Z)|x_i]\hat{m}(x_i, \hat{\theta}) + O_p(1) \quad (71)$$

To conclude the proof, I examine the term $n^{-\frac{1}{2}} \sum_i \hat{E}[\hat{f}_X(X)\tilde{v}_n(Z)|x_i]\hat{m}(x_i, \hat{\theta})$ in (72). Since $\hat{\theta}$ minimizes $Q_n(\theta)$ on Θ_n, the first equality in (72) holds with probability approaching 1 by Assumption 6(i) for some $\epsilon_n = o(n^{-\frac{1}{2}})$.

$$0 \geq \frac{1}{n} \sum_{i=1}^n (\hat{E}[v^*(X) - \hat{\theta}(Z)\hat{f}_X(X)|x_i])^2 - \frac{1}{n} \sum_{i=1}^n (\hat{E}[v^*(X) - (\hat{\theta}(Z) + \epsilon_n u_n(Z))\hat{f}_X(X)|x_i])^2$$
$$= 2\frac{\epsilon_n}{n} \sum_{i=1}^n \hat{m}(x_i, \hat{\theta})\hat{E}[u_n(Z)\hat{f}_X(X)|x_i] + \frac{\epsilon_n^2}{n} \sum_{i=1}^n (\hat{E}[u_n(Z)\hat{f}_X(X)|x_i])^2 \quad (72)$$

By (50), $n^{-1} \sum_i (\hat{E}[u_n(Z)\hat{f}_X(X)|x_i] - \hat{E}[u_n(Z)f_X(X)|x_i])^2 = o_p(n^{-\frac{1}{2}})$. As argued in showing (62), $\sup_{x,\Theta} \hat{E}[\theta(Z)|x] = O_p(1)$, and hence by $f_X(x)$ being bounded we get $\epsilon_n^2 n^{-1} \sum_i (\hat{E}[u_n(Z)\hat{f}_X(X)|x_i])^2 = O_p(\epsilon_n^2)$. Furthermore, since (72) holds for $u = \pm\tilde{v}$ and $\epsilon_n = o(n^{-\frac{1}{2}})$, it follows from (72) that $n^{-1} \sum_i \hat{m}(x_i, \hat{\theta})\hat{E}[u_n(Z)\hat{f}_X(X)|x_i] = o_p(n^{-\frac{1}{2}})$. Together with (71) and (70), this implies $n^{-\frac{1}{2}} \sum_i (y_i\hat{\theta}(z_i) - \langle v^*, m_0\rangle) = O_p(1)$, which establishes the Theorem. ■

APPENDIX C - Proof of Theorems 3.4, 3.5, 3.6, Lemma 3.1, Auxiliary Theorems .1, 3.3 and Auxiliary Lemma .3

Theorem .1. *Assume (i) $Q(\theta) \geq 0$ and $\Theta_0 = \{\theta \in \Theta : Q(\theta) = 0\}$ with Θ compact with respect to $||\cdot||$, (ii) $\Theta_n \subseteq \Theta$ are closed and $\sup_\Theta \inf_{\Theta_n} ||\theta - \theta_n|| = o(c_{1n})$ with $c_{1n} = o(1)$, (iii) $\sup_{\Theta_n} |Q_n(\theta_n) - Q(\theta_n)| = o_p(c_{2n})$ with $c_{2n} = o(1)$ and $\sup_{\Theta_{0n}^{\bar{\epsilon}_{1n}, ||\cdot||}} |Q_n(\theta_n) - Q(\theta_n)| = o_p(c_{3n})$ with $c_{3n} = o(1)$, (iv) $Q(\theta)$ is continuous in Θ with respect to $||\cdot||$ and $\sup_{\Theta_0^{\bar{\epsilon}, ||\cdot||}} Q(\theta) \leq C_1\epsilon^{\kappa_1}$ for some $\kappa_1 > 0$. Then for $a_n \to \infty$ with $a_n = O(\max[c_{1n}^{\min[\kappa_1, 1]}, c_{3n}]^{-1})$ and $b_n \to \infty$ 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 $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||) \xrightarrow{p} 0$. If in addition (v) $\inf_{(\Theta_0^{\bar{\epsilon}, ||\cdot||})^c} Q(\theta) \geq C_2\epsilon^{\kappa_2}$ for some $\kappa_2 > 0$, then $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||) = O_p(\max[b_n/a_n, c_{2n}]^{\frac{1}{\max[\kappa_2, 1]}})$.*

Proof of Theorem .1: Define Θ_{p_n} to be the pointwise projection of Θ_0 onto Θ_n under $||\cdot||$. In (73) we derive a bound for $\sup_{\Theta_{p_n}} Q_n(\theta)$. The first inequality follows from $\Theta_{p_n} \subseteq \Theta_n$, the definition of Θ_{p_n} and condition (ii). The

second inequality is implied by (iv), while the final result follows by (iii) and $a_n = O(\max[c_{1n}^{\min[\kappa_1,1]}, c_{2n}]^{-1})$.

$$\sup_{\Theta_{p_n}} Q_n(\theta) \leq \sup_{\Theta_{0n}^{\bar{c}_{1n},||\cdot||}} |Q_n(\theta) - Q(\theta)| + \sup_{\Theta_0^{\bar{c}_{1n},||\cdot||}} Q(\theta) \leq \sup_{\Theta_{0n}^{\bar{c}_{1n},||\cdot||}} |Q_n(\theta) - Q(\theta)| + C_1 c_{1n}^{\kappa_1} = O_p(a_n^{-1}) \tag{73}$$

Fix $\epsilon_n > 0$, which we will set constant for consistency and equal to $M \max[b_n/a_n, c_{2n}]^{\frac{1}{\max[\kappa_2,1]}}$ to obtain a rate of convergence. I now examine $h(\Theta_0, \hat{\Theta}_0)$. In (74), the final result follows by definition of Θ_{p_n} and condition (ii).

$$\begin{aligned} h(\Theta_0, \hat{\Theta}_0) = \sup_{\Theta_0} \inf_{\hat{\Theta}_0} ||\theta_0 - \hat{\theta}_0|| &\leq \sup_{\Theta_0} \inf_{\hat{\Theta}_0, \Theta_{p_n}} ||\theta_0 - \theta_{p_n}|| + ||\theta_{p_n} - \hat{\theta}_0|| \\ &\leq \sup_{\Theta_0} \inf_{\Theta_{p_n}} ||\theta_0 - \theta_{p_n}|| + \sup_{\Theta_{p_n}} \inf_{\hat{\Theta}_0} ||\theta_{p_n} - \hat{\theta}_0|| = \sup_{\Theta_{p_n}} \inf_{\hat{\Theta}_0} ||\theta_{p_n} - \hat{\theta}_0|| + o(c_{1n}) \end{aligned} \tag{74}$$

For n large enough, $c_{1n} < \epsilon_n/2$. When ϵ_n is constant this is clear, while if $\epsilon_n = M \max[b_n/a_n, c_{2n}]^{\frac{1}{\max[\kappa_2,1]}}$, then the result follows by $b_n \to \infty$ and $a_n = O(c_{1n}^{-1})$. Together with (74) this implies the first inequality in (75). The second inequality follows from $\Theta_{p_n} \subseteq \hat{\Theta}_0$ implying $\sup_{\Theta_{p_n}} \inf_{\hat{\Theta}_0} ||\theta_{p_n} - \hat{\theta}_0|| = 0$.

$$P\left(h(\Theta_0, \hat{\Theta}_0) < \epsilon_n\right) \geq P\left(\sup_{\Theta_{p_n}} \inf_{\hat{\Theta}_0} ||\theta_{p_n} - \hat{\theta}_0|| < \epsilon_n/2\right) \geq P\left(\Theta_{p_n} \subseteq \hat{\Theta}_0\right) \tag{75}$$

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, $P\left(h(\Theta_0, \hat{\Theta}_0) < \epsilon_n\right) \geq P\left(a_n \sup_{\Theta_{p_n}} Q_n(\theta) \leq b_n\right) \to 1$, since by (73) $a_n \sup_{\Theta_{p_n}} Q_n(\theta) = O_p(1)$ and $b_n \to \infty$. It therefore follows that $h(\Theta_0, \hat{\Theta}_0) = o_p(\epsilon_n)$.

To finish the proof it is necessary to examine $h(\hat{\Theta}_0, \Theta_0)$. The continuity of $Q(\theta)$ and the definition of Θ_0 imply:

$$\delta_n = \inf_{(\Theta_0^{\bar{c}_n,||\cdot||})^c} Q(\theta) > 0 \tag{76}$$

Note that $P\left(h(\hat{\Theta}_0, \Theta_0) > \epsilon_n\right) = P\left(h(\hat{\Theta}_0, \Theta_0) > \epsilon_n; \hat{\Theta}_0 = \emptyset\right) + P\left(h(\hat{\Theta}_0, \Theta_0) > \epsilon_n; \hat{\Theta}_0 \neq \emptyset\right)$. If $\hat{\Theta}_0 = \emptyset$, then $h(\hat{\Theta}_0, \Theta_0) = 0$, implying $P\left(h(\hat{\Theta}_0, \Theta_0) > \epsilon_n\right) = P\left(h(\hat{\Theta}_0, \Theta_0) > \epsilon_n; \hat{\Theta}_0 \neq \emptyset\right)$. To conclude, note the definition of $\hat{\Theta}_0$ and (76) imply the first inequality in (77). For n large enough, $b_n/a_n \leq \delta_n/2$. If ϵ_n is constant, then it is clear since $b_n = o(a_n)$; while if $\epsilon_n = M \max[b_n/a_n, c_{2n}]^{\frac{1}{\max[\kappa_2,1]}}$, then we use (v) to derive $\delta_n \geq M^{\kappa_2} b_n/a_n$ and the result holds for M large enough. This result implies the second inequality in (77). For ϵ_n a constant, the final result follows from $\sup_{\Theta_n} |Q_n(\theta_n) - Q(\theta_n)| = o_p(c_{2n})$, while if $\epsilon_n = M \max[b_n/a_n, c_{2n}]^{\frac{1}{\max[\kappa_2,1]}}$ the conclusion follows from (v), which implies $\delta_n \geq C_2 M^{\kappa_2} c_{2n}$.

$$\begin{aligned} P\left(h(\hat{\Theta}_0, \Theta_0) > \epsilon_n; \hat{\Theta}_0 \neq \emptyset\right) &\leq P\left(\exists \theta \in \Theta_n : Q_n(\theta) \leq b_n/a_n \ \ and \ \ Q(\theta) > \delta_n\right) \\ &\leq P\left(\sup_{\Theta_n} |Q(\theta) - Q_n(\theta)| > \delta_n/2\right) \to 0 \end{aligned} \tag{77}$$

Hence, $h(\hat{\Theta}_0, \Theta_0) = o_p(1)$ under conditions (i)-(iv), and if (v) holds, then $h(\hat{\Theta}_0, \Theta_0) = O_p(\max[b_n/a_n, c_{2n}]^{\frac{1}{\max[\kappa_2,1]}})$. Together with our discussion of $h(\Theta_0, \hat{\Theta}_0)$, this implies $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||) = o_p(1)$ under (i)-(iv) and $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||) = O_p(\max[b_n/a_n, c_{2n}]^{\frac{1}{\max[\kappa_2,1]}})$ under (i)-(v), which concludes the proof of the Theorem. ■

Proof of Theorem 3.3: By Example 3.1 in Stinchcombe & White (1992), the estimator $\hat{\theta}(z)$ is measurable and hence a well defined random variable. If the model is identified, so that $\Theta_0 = \{\theta_0\}$, then $P(||\hat{\theta} - \theta_0|| < \epsilon) \geq P(d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||) < \epsilon) \to 0$ by Assumption (ii). Therefore, without loss of generality we assume Θ_0 is not a singleton. Let $N_\epsilon(\theta_0)$ be an ϵ open neighborhood of θ_0 in Θ. Since Θ_0 is a closed subset of a compact set Θ, it follows

that it is also compact. Furthermore, since Θ_0 has more than one element, for ϵ small enough, the set $\Theta_0 \cap N^c_{\epsilon/2}(\theta_0)$ is not empty. Therefore, the continuity of $M(\theta)$, compactness of $\Theta_0 \cap N^c_{\epsilon/2}(\theta_0)$, and the fact that θ_0 is the unique minimum of $M(\theta)$ in Θ_0 imply that:

$$\delta = \min_{\theta \in \Theta_0 \cap N^c_{\epsilon/2}(\theta_0)} M(\theta) - M(\theta_0) > 0$$

Since $M(\theta)$ is continuous and Θ is compact, $M(\theta)$ is also uniformly continuous in Θ. It follows that there exists a ζ such that if $||\theta - \theta'|| < \zeta$ then $|M(\theta) - M(\theta')| < \delta/4$ for all θ, $\theta' \in \Theta$. Let $\gamma = \min\{\zeta, \epsilon/2\}$ and note that if $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||) < \gamma$ and $\hat{\Theta}_0 \cap N^c_\epsilon(\theta_0)$ is not empty, then it also follows that:

$$\min_{\theta \in \hat{\Theta}_0 \cap N^c_\epsilon(\theta_0)} M(\theta) - M(\theta_0) > \delta/2 \tag{78}$$

To see this, pick any $\theta \in \hat{\Theta}_0 \cap N^c_\epsilon(\theta_0)$ and let $\theta_P = \arg\inf_{\theta' \in \Theta_0} ||\theta - \theta'||$. Since $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||) < \epsilon/2$, it follows that $||\theta - \theta_P|| < \epsilon/2$. In addition, since $\theta \in N^c_\epsilon(\theta_0)$, $||\theta_0 - \theta_P|| \geq ||\theta_0 - \theta|| - ||\theta_P - \theta|| \geq \epsilon/2$, which implies that $\theta_P \in \Theta_0 \cap N^c_{\epsilon/2}(\theta_0)$, and hence $M(\theta_P) \geq M(\theta_0) + \delta$. On the other hand, $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||) < \zeta$, and thus $|M(\theta) - M(\theta_P)| < \delta/2$, which implies $M(\theta) - M(\theta_0) \geq M(\theta_P) - M(\theta_0) - |M(\theta) - M(\theta_P)| \geq \delta/2$, establishing (78). It thus follows that:

$$P\left(||\hat{\theta} - \theta_0|| < \epsilon\right) \geq P\left(M(\hat{\theta}) - M(\theta_0) \leq \delta/2; d_H(\hat{\Theta}_0, \Theta_0, ||\cdot|| < \gamma)\right) \tag{79}$$

Let $\tilde{\theta} = \arg\inf_{\theta' \in \hat{\Theta}_0} ||\theta_0 - \theta'||$. If $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||) < \gamma$, then $||\theta_0 - \tilde{\theta}|| < \zeta$, implying that $M(\tilde{\theta}) \leq M(\theta_0) + \delta/4$. Furthermore, by definition of $\hat{\theta}$, $M_n(\hat{\theta}) \leq M_n(\tilde{\theta})$ and hence, $M(\hat{\theta}) - M(\tilde{\theta}) \leq |M(\hat{\theta}) - M_n(\hat{\theta})| + |M(\tilde{\theta}) - M_n(\tilde{\theta})|$. Thus, $M(\tilde{\theta}) \leq M(\theta_0) + \delta/4$ and $|M(\hat{\theta}) - M_n(\hat{\theta})| + |M(\tilde{\theta}) - M_n(\tilde{\theta})| \leq \delta/4$ imply that $M(\hat{\theta}) - M(\theta_0) \leq \delta/2$. Therefore, using (79), we get that:

$$P\left(||\hat{\theta} - \theta_0|| < \epsilon\right) \geq P\left(|M(\hat{\theta}) - M_n(\hat{\theta})| + |M(\tilde{\theta}) - M_n(\tilde{\theta})| \leq \delta/4; d_H(\hat{\Theta}_0, \Theta_0, ||\cdot|| < \gamma)\right) \tag{80}$$

But $P\left(|M(\hat{\theta}) - M_n(\hat{\theta})| + |M(\tilde{\theta}) - M_n(\tilde{\theta})| \leq \delta/4; d_H(\hat{\Theta}_0, \Theta_0, ||\cdot|| < \gamma)\right) \to 1$ under Assumptions (iii) and (iv) and $\hat{\theta}$, $\tilde{\theta} \in \hat{\Theta}_0 \subseteq \Theta_n$, which using (80) establishes the theorem. ■

Proof of Theorem 3.4: I first show consistency under $d_H(\cdot, \cdot, ||\cdot||_\infty)$ by verifying the conditions for Theorem .1. By Newey & Powell (2003), Θ is compact under $||\cdot||_\infty$. Assumption 3(vii) implies condition (ii) is satisfied with $c_{1n} = n^{-\frac{1}{2}-\delta_0}$, while by Corollary .1, $||\cdot||_w \leq \sup_x f_X(x)||\cdot||_\infty$ and $f_X(x)$ bounded, condition (iii) is satisfied by $c_{2n} = n^{-\frac{1}{4}-\frac{\delta_0}{2}}$ and $c_{3n} = n^{-\frac{1}{2}-\frac{\delta_0}{2}}$. To verify condition (iv), in (81) we use the fact that $E[\theta_0(Z)|x] = v^*(x)/f_X(x)$ for all $\theta_0 \in \Theta_0$ to obtain the first equality and use $||\cdot||_w \leq \sup_x f_X(x)||\cdot||_\infty$ and $f_X(x)$ bounded for the second inequality.

$$\sup_{\Theta_0^{\bar{\epsilon}, ||\cdot||_\infty}} E[(E[v^*(X) - \theta(Z) f_X(X)|X])^2] = \inf_{\Theta_0} \sup_{\Theta_0^{\bar{\epsilon}, ||\cdot||_\infty}} E[(E[\theta_0(Z) - \theta(Z)|X])^2 f_X^2(X)] \lesssim \epsilon^2 \tag{81}$$

Hence, (81) implies (iv) is satisfied for some $C_1 > 0$ and $\kappa_1 = 2$. Theorem .1 requires $a_n = O(\max[c_{1n}^{\min[\kappa_1, 1]}, c_{3n}]^{-1})$, which given our parameter values simplifies to $a_n = O(n^{\frac{1}{2}+\delta_0})$. Hence, conditions (i)-(iv) of Theorem .1 are satisfied, which implies $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_\infty) \xrightarrow{p} 0$.

I now proceed to show the second claim of the Theorem. Under Assumption 3(vii), Corollary .1 and since $||\cdot||_w \lesssim ||\cdot||_\infty$, conditions (i)-(iii) of Theorem .1 are still satisfied with $c_{1n} = c_{3n} = n^{-\frac{1}{2}-\frac{\delta_0}{2}}$ and $c_{2n} = n^{-\frac{1}{4}-\frac{\delta_0}{2}}$. Furthermore, the same arguments as in (81) immediately imply $\sup_{\Theta_0^{\bar{\epsilon}, ||\cdot||_w}} Q(\theta) \leq \epsilon^2$. Hence, condition (iv) is satisfied for $\kappa_1 = 2$

and $C_1 = 1$. To verify condition (v), we use $E[\theta_0(Z)|x]f_X(x) = v^*(x)$ for all $\theta_0 \in \Theta_0$ to derive the first equality in (82), while the second inequality follows from the definition of $(\Theta_0^{\bar{\epsilon},||\cdot||_w})^c$.

$$\inf_{\left(\Theta_0^{\bar{\epsilon},||\cdot||_w}\right)^c} E[(E[v^*(X) - \theta(Z)f_X(X)|X])^2] = \sup_{\Theta_0} \inf_{\left(\Theta_0^{\bar{\epsilon},||\cdot||_w}\right)^c} E[(E[\theta_0(Z) - \theta(Z)|X])^2 f_X^2(X)] \geq \epsilon^2 \tag{82}$$

Therefore, (82) implies condition (v) holds for $C_2 = 1$ and $\kappa_2 = 2$. Hence, conditions (i)-(v) of Theorem .1 are verified, which implies $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_w) = O_p(\max[b_n/a_n, c_{2n}]^{\frac{1}{\max[\kappa_2,1]}}) = O_p(n^{-\frac{1}{8}-\frac{\delta_0}{4}})$. We can now exploit the local behavior of the objective function to improve on this rate by using arguments similar to those in the proof of Theorem .1. I start by showing $h(\Theta_0, \hat{\Theta}_0) = o_p(n^{-\frac{1}{4}-\frac{\delta_0}{2}})$. Let Θ_{p_n} denote the projection of Θ_0 onto Θ_n under $||\cdot||_\infty$ and set $\delta_{0n} = n^{-\frac{1}{4}-\frac{\delta_0}{2}}$. With these definitions we can derive (83), where the last inequality follows from $||\cdot||_w \lesssim ||\cdot||_\infty$ and Assumption 3(vii).

$$\begin{aligned} h(\Theta_0, \hat{\Theta}_0) = \sup_{\Theta_0} \inf_{\hat{\Theta}_0} ||\theta_0 - \hat{\theta}_0||_w &\leq \sup_{\Theta_0} \inf_{\hat{\Theta}_0, \Theta_{p_n}} ||\theta_0 - \theta_{p_n}||_w + ||\theta_{p_n} - \hat{\theta}_0||_w \\ &\leq \sup_{\Theta_0} \inf_{\Theta_{p_n}} ||\theta_0 - \theta_{p_n}||_w + \sup_{\Theta_{p_n}} \inf_{\hat{\Theta}_0} ||\theta_{p_n} - \hat{\theta}_0||_w = \sup_{\Theta_{p_n}} \inf_{\hat{\Theta}_0} ||\theta_{p_n} - \hat{\theta}_0||_w + o(\delta_{0n}) \end{aligned} \tag{83}$$

Derivation (83) implies that the first inequality in (84) holds for n large enough, while the second inequality simply follows from $\Theta_{p_n} \subseteq \hat{\Theta}_0$ implying $\sup_{\Theta_{p_n}} \inf_{\hat{\Theta}_0} ||\theta_{p_n} - \hat{\theta}_0||_w = 0$.

$$P\left(h(\Theta_0, \hat{\Theta}_0) < \delta_{0n}\right) \geq P\left(\sup_{\Theta_{p_n}} \inf_{\hat{\Theta}_0} ||\theta_{p_n} - \hat{\theta}_0||_w < \delta_{0n}/2\right) \geq P\left(\Theta_{p_n} \subseteq \hat{\Theta}_0\right) \tag{84}$$

By definition of $\hat{\Theta}_0$, the event $\Theta_{p_n} \subseteq \hat{\Theta}_0$ is equivalent to $Q_n(\theta) \leq b_n/a_n$ for all $\theta \in \Theta_{p_n}$. Note, however, that $\sup_{\Theta_{p_n}} Q_n(\theta) \leq \sup_{\Theta_{0n}^{\bar{\delta}_{0n},||\cdot||_w}} |Q_n(\theta) - Q(\theta)| + \sup_{\Theta_0^{\bar{\delta}_{0n},||\cdot||_\infty}} Q(\theta) = o_p(n^{-\frac{1}{2}-\frac{\delta_0}{2}}) + O(\delta_{0n}^2)$ by Corollary .1 and (81). Therefore, since $a_n = n^{\frac{1}{2}+\frac{\delta_0}{2}}$, it follows that $a_n \sup_{\Theta_{p_n}} = O_p(1)$. Together with (84) and $b_n \to \infty$, this gives us the conclusion in (85)

$$P\left(h(\Theta_0, \hat{\Theta}_0) < \delta_{0n}\right) \geq P\left(a_n \sup_{\Theta_{p_n}} Q_n(\theta) \leq b_n\right) \to 1 \tag{85}$$

which shows $h(\Theta_0, \hat{\Theta}_0) = o_p(n^{-\frac{1}{4}-\frac{\delta_0}{2}})$. Let $\epsilon_{1n} = n^{-\frac{1}{8}-\frac{\delta_0}{5}}$ and $\delta_{1n} = n^{-\left(\frac{1}{8}+\frac{1}{16}\right)-\frac{\delta_0}{4}}$. The first equality in (86) is implied by $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_w) = O_p(n^{-\frac{1}{8}-\frac{\delta_0}{4}})$ and $h(\Theta_0, \hat{\Theta}_0) = o_p(n^{-\frac{1}{4}-\frac{\delta_0}{2}})$. By (82), the second event in (86) implies there exists a $\theta \in \Theta_{0n}^{\bar{\epsilon}_{1n},||\cdot||_w}$ with $Q_n(\theta) \leq b_n/a_n$ and $Q(\theta) \geq \delta_{1n}^2$. Since $b_n/a_n = (C_b/C_a)n^{-\frac{1}{2}-\frac{\delta_0}{2}} \log n$ and $\delta_{1n}^2 = n^{-\left(\frac{1}{4}+\frac{1}{8}\right)-\frac{\delta_0}{2}}$ we have that for n large enough $b_n/a_n \leq \delta_{1n}^2/2$, which establishes the second inequality in (86). By Corollary .1, and since $\epsilon_{1n} = o(n^{-\frac{1}{8}})$, it follows that $\sup_{\Theta_{0n}^{\bar{\epsilon}_{1n},||\cdot||_w}} |Q_n(\theta) - Q(\theta)| = o_p(n^{-\left(\frac{1}{4}+\frac{1}{8}\right)-\frac{\delta_0}{2}})$, which implies the final result in (86).

$$\begin{aligned} P\left(d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_w) > \delta_{1n}\right) &= P\left(h(\hat{\Theta}_0, \Theta_0) > \delta_{1n}; \hat{\Theta}_0 \subseteq \Theta_{0n}^{\bar{\epsilon}_{1n},||\cdot||_w}\right) + o(1) \\ &\leq P\left(\sup_{\Theta_{0n}^{\bar{\epsilon}_{1n},||\cdot||_w}} |Q_n(\theta) - Q(\theta)| > \delta_{1n}^2/2\right) + o(1) \to 0 \end{aligned} \tag{86}$$

Hence, (86) establishes that $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_w) = o_p(n^{-\left(\frac{1}{8}+\frac{1}{16}\right)-\frac{\delta_0}{4}})$. Now we can improve on this rate of convergence by letting $\epsilon_{2n} = n^{-\left(\frac{1}{8}+\frac{1}{16}\right)-\frac{\delta_0}{4}}$ and $\delta_{2n} = n^{-\left(\frac{1}{8}+\frac{1}{16}+\frac{1}{32}\right)-\frac{\delta_0}{4}}$ and repeating the same arguments as in (86) to conclude that $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_w) = o_p(n^{-\left(\frac{1}{8}+\frac{1}{16}+\frac{1}{32}\right)-\frac{\delta_0}{4}})$. Repeating this argument a large but finite number of times, we can eventually establish $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_w) = o_p(n^{-\frac{1}{4}})$, which concludes the proof of the Theorem. ■

Proof of Theorem 3.5: I begin by establishing $||\tilde{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$. Since Θ_0 is an equivalence class under $||\cdot||_w$ and $\tilde{\theta} \in \hat{\Theta}_0 + \lambda_n \mathcal{E}_n$, it is sufficient to show that $d_H(\hat{\Theta}_0 + \lambda_n \mathcal{E}_n, \Theta_0, ||\cdot||_w) = o_p(n^{-\frac{1}{4}})$. In (87) I first show

$h(\Theta_0, \hat{\Theta}_0 + \lambda_n \mathcal{E}_n) = o_p(n^{-\frac{1}{4}})$. The second inequality in (87) follows from $\hat{\Theta}_0 \subseteq \hat{\Theta}_0 + \lambda_n \mathcal{E}_n$, while the final result in (87) follows from Theorem 3.4, which implies $d_H(\Theta_0, \hat{\Theta}_0, ||\cdot||_w) = o_p(n^{-\frac{1}{4}})$.

$$h(\Theta_0, \hat{\Theta}_0 + \lambda_n \mathcal{E}_n) = \sup_{\Theta_0} \inf_{\hat{\Theta}_0 + \lambda_n \mathcal{E}_n} ||\theta_0 - \tilde{\theta}||_w \leq \sup_{\Theta_0} \inf_{\hat{\Theta}_0} ||\theta_0 - \tilde{\theta}||_w = o_p(n^{-\frac{1}{4}}) \tag{87}$$

In order to establish $h(\hat{\Theta}_0 + \lambda_n \mathcal{E}_n, \Theta_0) = o_p(n^{-\frac{1}{4}})$, in (88) use the fact that if $\tilde{\theta} \in \hat{\Theta}_0 + \lambda_n \mathcal{E}_n$, then it can be written in the form $\tilde{\theta} = \bar{\theta} + \lambda_n \hat{E}[\theta_n(Z)|x]$ for some $\bar{\theta} \in \hat{\Theta}_0$ and $\theta_n \in \Theta_n$ to derive the first inequality. The final result in (88) follows from Theorem 3.4, which implies $d_H(\Theta_0, \hat{\Theta}_0, ||\cdot||_w) = o_p(n^{-\frac{1}{4}})$.

$$\begin{aligned} h(\hat{\Theta}_0 + \lambda_n \mathcal{E}_n, \Theta_0) = \sup_{\hat{\Theta}_0 + \lambda_n \mathcal{E}_n} \inf_{\Theta_0} ||\theta_0 - \theta||_w &\leq \sup_{\hat{\Theta}_0} \inf_{\Theta_0} ||\theta_0 - \bar{\theta}||_w + \lambda_n \sup_{\Theta_n} ||\hat{E}[\theta_n(Z)|x]||_w \\ &= o_p(n^{-\frac{1}{4}}) + \lambda_n \sup_{\Theta_n} ||\hat{E}[\theta_n(Z)|x]||_w \end{aligned} \tag{88}$$

In order to control the term $\sup_{\Theta_n} ||\hat{E}[\theta_n(Z)|X]||_w$ in (88), first note that $E[\hat{E}[\theta_n(Z)|X]|x] = \hat{E}[\theta_n(Z)|x]$ and use $f_X(x)$ bounded to obtain the first equality in (89). The second inequality in (89) follows from $\Theta_n \subseteq \Theta$ and the triangle inequality. Theorem 1 in Newey (1997) implies that under Assumptions 1, 2(i)-(ii), 3(v) and 4(i), pointwise in $\theta \in \Theta$, $||\hat{E}[\theta|X] - E[\theta|X]||_{\mathcal{L}^2} = O_p(k_n^{\frac{1}{2}}/n^{\frac{1}{2}} + k_n^{-\frac{\gamma}{d_x}})$. Theorem 1 in Newey (1997), however, actually implies the result holds uniformly in Θ under our additional assumptions since the approximation error in Assumption 2(ii) is uniform in Θ and $\theta \in \Theta$ are uniformly bounded. Together with $\theta \in \Theta$ being uniformly bounded, this implies the last result in (89).

$$\begin{aligned} \sup_{\Theta_n} ||\hat{E}[\theta_n(Z)|X]||_w &\lesssim \sup_{\Theta_n} ||\hat{E}[\theta_n(Z)|X]||_{\mathcal{L}^2} \\ &\leq \sup_{\Theta} ||\hat{E}[\theta_n(Z)|X] - E[\theta_n(Z)|X]||_{\mathcal{L}^2} + \sup_{\Theta} ||E[\theta_n(Z)|X]||_{\mathcal{L}^2} = O_p(k_n^{\frac{1}{2}}/n^{\frac{1}{2}} + k_n^{-\frac{\gamma}{d_x}}) + O(1) \end{aligned} \tag{89}$$

Since $k_n^{\frac{1}{2}}/n^{\frac{1}{2}} + k_n^{-\frac{\gamma}{d_x}} \to 0$ under Assumptions 4(i) (89) implies that $\lambda_n \sup_{\Theta_n} ||\hat{E}[\theta_n|X]||_w = O_p(\lambda_n) = o_p(n^{-\frac{1}{4}})$ due to Assumption 7(i). Together with (88) and (87) this implies $d_H(\hat{\Theta}_0 + \lambda_n \mathcal{E}_n, \Theta_0, ||\cdot||_w) = o_p(n^{-\frac{1}{4}})$, and hence $||\tilde{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$, which establishes the first claim of the Theorem.

I will show that $||\tilde{\theta} - \theta_0||_\infty = o_p(1)$ by verifying the Assumptions of Theorem 3.3. To do so, we will first need to imbed $\{\hat{\Theta}_0 + \lambda_n \mathcal{E}_n\}$ within a compact parameter space $\tilde{\Theta}$. Let $C_B^1(\mathcal{X})$ be the set of continuously differentiable functions on $\mathcal{X}$ satisfying $\sup_x |D^\alpha g(x)| \leq B$ for $|\alpha| \leq 1$. Define $\tilde{\Theta} = \Theta + C_B^1(\mathcal{X})$. Newey & Powell (2003) show that Θ is compact under $||\cdot||_\infty$. In addition, since $\mathcal{X}$ is convex and compact by Assumption 1, Theorem 1.34 in Adams & Fournier (2003) implies that $C^1(\mathcal{X})$ is compactly imbedded in $C^0(\mathcal{X})$. Hence, as $C_B^1(\mathcal{X})$ is a closed bounded subset of $C^1(\mathcal{X})$, it follows that $C_B^1(\mathcal{X})$ is compact under $||\cdot||_\infty$ as well. We conclude that $\tilde{\Theta}$ is compact under $||\cdot||_\infty$, and hence verify the compactness requirement on $\tilde{\Theta}$ in Assumption 1 of Theorem 3.3. Recall $M(\theta) = E[(v^*(X) - \theta(Z) f_X(X))^2]$. In (90) I show $M(\theta)$ is strictly convex on Θ_0. The first equality is derived by expanding the square. If $v^*(x)/f_X(x) = c$ is constant, then the minimizer is clearly unique for $\theta_0(z) = c$. If on the other hand $v^*(x)/f_X(x)$ is not constant, then for $\theta_1(x) \in \Theta_0$ and $(a, b) \neq (0, 1)$ we have $a + b\theta_1(x) \notin \Theta_0$, because $E[a + b\theta_1(Z)|X] = a + bv^*(x)/f_X(x) \neq v^*(x)/f_X(x)$ by X continuously distributed. Thus, if $\theta_1, \theta_2 \in \Theta_0$, then $\theta_1 \neq a + b\theta_2$, which implies the Cauchy Schwarz inequality holds strictly, giving the inequality in (90).

$$\begin{aligned} M(\lambda\theta_1 + (1-\lambda)\theta_2) &= \lambda^2 M(\theta_1) + (1-\lambda)^2 M(\theta_2) + 2\lambda(1-\lambda) E[(v^*(X) - \theta_1(Z) f_X(X))(v^*(X) - \theta_2(Z) f_X(X))] \\ &< \lambda^2 M(\theta_1) + (1-\lambda)^2 M(\theta_2) + 2\lambda(1-\lambda) M^{\frac{1}{2}}(\theta_1) M^{\frac{1}{2}}(\theta_2) \leq \min\{M(\theta_1), M(\theta_2)\} \end{aligned} \tag{90}$$

Since $M(\theta)$ is strictly a convex continuous functional on Θ_0, and Θ_0 is a closed convex subset of compact set, it follows that $M(\theta)$ attains a unique minimum on Θ_0, which concludes verifying Assumption (i) of Theorem 3.3. In order to verify Assumption (ii), we need to establish $\{\hat{\Theta}_0 + \lambda_n \mathcal{E}_n\} \subset \tilde{\Theta}$ with probability approaching 1. Since $\hat{\Theta}_0 \subseteq \Theta_n \subseteq \Theta$, the result will follow if $\lambda_n \mathcal{E}_n \subseteq C_B^1(\mathcal{X})$ with arbitrarily large probability. In (91), the first equality follows by the definition of $\hat{E}[\theta_n(Z)|x]$ and the Cauchy-Schwarz inequality. The second result follows from the definition of ξ_{jn}. Note that the largest eigenvalue of $P(P'P)^{-2}P'$ is equal to $\tilde{\lambda}_n^{-1}$, for $\tilde{\lambda}_n$ smallest eigenvalue of $P'P$. Therefore, since $\theta_n \in \Theta_n$ are uniformly bounded we derive the final result in (91).

$$\begin{aligned}\sup_{\Theta_n} \sup_{|\alpha|=j,\mathcal{X}} \lambda_n |D^\alpha \hat{E}[\theta_n(Z)|x]| &\leq \sup_{\Theta_n} \sup_{|\alpha|=j,\mathcal{X}} \lambda_n ||D^\alpha p^{k_n'}(x)|||(P'P)^{-1} \sum_{i=1}^n p^{k_n}(x_i)\theta_n(z_i)|| \\ &\leq \sup_{\Theta_n} \lambda_n \xi_{jn} \left[\sum_{i=1}^n p^{k_n'}(x_i)\theta_n(z_i)(P'P)^{-2} \sum_{i=1}^n p^{k_n}(x_i)\theta_n(z_i)\right]^{\frac{1}{2}} = O_p(\lambda_n \xi_{jn} \tilde{\lambda}_n^{-1}) \quad (91)\end{aligned}$$

In the proof of Theorem 1 in Newey (1997), it is shown that $\tilde{\lambda}_n^{-1} = O_p(1)$, and hence since under Assumption 7(ii) $\lambda_n \xi_{jn} \to 0$ for $j \in \{0, 1\}$, (91) implies that $\sup_{\Theta_n} \sup_{|\alpha|\leq 1,\mathcal{X}} \lambda_n |D^\alpha \hat{E}[\theta_n|X]| = o_p(1)$. It follows that with probability tending to one $\lambda_n \mathcal{E}_n \subseteq C_B^1(\mathcal{X})$ and hence $\{\hat{\Theta}_0 + \lambda_n \mathcal{E}_n\}$ are subsets of $\tilde{\Theta}$. To complete verifying Assumption (ii) of Theorem 3.3 we still need to establish $d_H(\hat{\Theta}_0 + \lambda_n \mathcal{E}_n, \Theta_0, ||\cdot||_\infty) \xrightarrow{p} 0$. In (92) I begin by examining $h(\Theta_0, \hat{\Theta}_0 + \lambda_n \mathcal{E}_n)$ and using $\hat{\Theta}_0 \subseteq \hat{\Theta}_0 + \lambda_n \mathcal{E}_n$ to obtain the first inequality. The final result in (92) follows from Theorem 3.4 which shows $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_\infty) \xrightarrow{p} 0$.

$$h(\Theta_0, \hat{\Theta}_0 + \lambda_n \mathcal{E}_n) = \sup_{\Theta_0} \inf_{\hat{\Theta}_0 + \lambda_n \mathcal{E}_n} ||\theta_0 - \tilde{\theta}||_\infty \leq \sup_{\Theta_0} \inf_{\hat{\Theta}_0} ||\theta_0 - \hat{\theta}||_\infty = o_p(1) \quad (92)$$

In (93), I examine $h(\hat{\Theta}_0 + \lambda_n \mathcal{E}_n, \Theta_0)$. To establish the first equality in (93) we use the triangle inequality and the fact that if $\tilde{\theta} \in \hat{\Theta}_0 + \lambda_n \mathcal{E}_n$, then it can be written in the form $\tilde{\theta} = \bar{\theta} + \lambda_n \hat{E}[\theta_n(Z)|x]$ for some $\bar{\theta} \in \hat{\Theta}_0$ and $\theta_n \in \Theta_n$. Since $d_H(\hat{\Theta}_0, \Theta_0, ||\cdot||_\infty) = o_p(1)$ by Theorem 3.4, it follows that $h(\hat{\Theta}_0, \Theta_0) = o_p(1)$. In addition, (91), $\tilde{\lambda}_n^{-1} = O_p(1)$ by Theorem 1 in Newey (1997) and $\xi_{0n}\lambda_n \to 0$ by Assumption 7(ii) implies $\sup_{\Theta_n} \lambda_n ||\hat{E}[\theta_n(Z)|x]||_\infty = o_p(1)$, which establishes the final result in (93).

$$h(\hat{\Theta}_0 + \lambda_n \mathcal{E}_n, \Theta_0) = \sup_{\hat{\Theta}_0 + \lambda_n \mathcal{E}_n} \inf_{\Theta_0} ||\theta_0 - \tilde{\theta}||_\infty \leq \sup_{\hat{\Theta}_0} \inf_{\Theta_0} ||\hat{\theta} - \theta_0||_\infty + \sup_{\Theta_n} \lambda_n ||\hat{E}[\theta_n(Z)|x]||_\infty = o_p(1) \quad (93)$$

Thus, (92) and (93) imply $d_H(\hat{\Theta}_0 + \lambda_n \mathcal{E}_n, \Theta_0, ||\cdot||_\infty) \xrightarrow{p} 0$, verifying Assumption (ii) in Theorem 3.3. Let $M_n(\theta) = n^{-1} \sum_i (v^*(x_i) - \theta(z_i, x_i)\hat{f}_X(x_i)^2$. The continuity of $M(\theta)$ and $M_n(\theta)$ on $\tilde{\Theta}$ under $||\cdot||_\infty$ is immediate, which verifies Assumption (iii). To verify Assumption (iv) in Theorem 3.3, let $\tilde{\mathcal{F}}_n = \{v^*(x) - \theta(z, x)\hat{f}_X(x) : \theta(z, x) \in \tilde{\Theta}\}$. As shown in (62), $||\hat{f}_X - f_X||_\infty = o_p(1)$ and therefore $f_X(x)$ bounded implies for M large enough $P(||\hat{f}_X||_\infty > M) \to 0$. Therefore, with probability approaching one, we have $N_{[]}(\epsilon, \tilde{\mathcal{F}}, ||\cdot||_\infty) \leq N_{[]}(\epsilon/M, \tilde{\Theta}, ||\cdot||_\infty) \leq N_{[]}(\epsilon/2, \Theta, ||\cdot||_\infty) \times N_{[]}(\epsilon/2, C_B^1(\mathcal{X}), ||\cdot||_\infty) < \infty$ by Theorem 2.7.1 in van der Vaart & Wellner (2000). Let $\tilde{\mathcal{F}}^2 = \{f^2(z, x) : f(z, x) \in \tilde{\mathcal{F}}\}$. Since with probability tending to one $f(z, x) \in \tilde{\mathcal{F}}$ are uniformly bounded, it follows that for some $C > 0$, $N_{[]}(\epsilon, \tilde{\mathcal{F}}^2, ||\cdot||_\infty) \leq N_{[]}(\epsilon/C, \tilde{\mathcal{F}}, ||\cdot||_\infty) < \infty$. Hence, by Theorem 2.4.1 in van der Vaart & Wellner (2000), the class $\tilde{\mathcal{F}}^2$ is Glivenko-Cantelli, establishing (94).

$$\sup_{\theta \in \tilde{\Theta}} |M_n(\theta) - M(\theta)| = \sup_{f \in \tilde{\mathcal{F}}^2} \left| \frac{1}{n} \sum_{i=1}^n f(x_i, z_i) - E[fX, Z)] \right| \xrightarrow{p} 0 \quad (94)$$

Result (94)verifying Assumption (iv) in Theorem 3.3. Therefore, Theorem 3.3 implies $||\tilde{\theta} - \theta_0||_\infty \xrightarrow{p} 0$, concluding the proof of the second claim of the Theorem. ■

Lemma .3. *Let* $\tilde{v}_n = \arg\min_{\Theta_n} ||\tilde{v} - \theta_n||_\infty$. *Under Assumptions 1, 2(i)-(ii) and 3-5, if* $\hat{\theta} \in \Theta_n$ *and* $||\hat{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$, *then* $n^{-1}\sum_i E[\tilde{v}(Z)f_X(X)|x_i]\hat{m}(x_i,\hat{\theta}) = n^{-1}\sum_i \hat{E}[\tilde{v}_n(Z)|x_i]\hat{f}_X(x_i)(v^*(x_i) - \hat{\theta}(z_i)\hat{f}_X(x_i)) + o_p(n^{-\frac{1}{2}})$

Proof of Lemma .3: To begin the proof, note that manipulations identical to those in deriving (61) imply the first equality in (95). The second result in (95) follows for $\varepsilon(x_i,z_i) = (v^*(x_i) - \hat{\theta}(z_i)\hat{f}_X(x_i)) - E[v^*(X) - \hat{\theta}(Z)\hat{f}_X(X)|x_i]$.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n E[\tilde{v}(Z)f_X(X)|x_i]\hat{m}(x_i,\hat{\theta}) &= \frac{1}{n}\sum_{i=1}^n \hat{E}[E[\tilde{v}_n(Z)|X]f_X(X)|x_i](v^*(x_i) - \hat{\theta}(z_i)\hat{f}_X(x_i)) \\ &= \frac{1}{n}\sum_{i=1}^n \hat{E}[E[\tilde{v}(Z)|X]f_X(X)|x_i]E[v^*(X) - \hat{\theta}(Z)\hat{f}_X(X)|x_i] + \frac{1}{n}\sum_{i=1}^n \hat{E}[E[\tilde{v}(Z)|X]f_X(X)|x_i]\varepsilon(x_i,z_i) \quad (95)\end{aligned}$$

Let $g(x) = E[\tilde{v}(Z)f_X(X)|x]$. In (96) we examine the term $n^{-1}\sum_i \hat{E}[g(X)|x_i]E[v^*(X) - \hat{\theta}(Z)\hat{f}_X(X)|x_i]$. The first result in (96) follows by the Cauchy-Schwarz inequality and recalling that $m(x_i,\hat{\theta}) = E[v^*(X) - \hat{\theta}(Z)f_X(X)|x_i]$. By Assumption 3(v), we can set $v^*(x) = 0$ in (49) to obtain $n^{-1}\sum_i(\hat{E}[g(X)|x_i] - g(x_i))^2 = o_p(n^{-\frac{1}{2}})$. In addition, since $||\hat{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$, part c) of Lemma .1 implies $n^{-1}\sum_i m^2(x_i,\hat{\theta}) = o_p(n^{-\frac{1}{2}})$. Furthermore, since $\hat{\theta} \in \Theta$, it is uniformly bounded and as shown in deriving (50), $n^{-1}\sum_i(E[\hat{\theta}(Z)|x_i])^2(\hat{f}_X(x_i) - f_X(x_i))^2 = o_p(n^{-\frac{1}{2}})$. Combining these results we obtain the last equality in (96).

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \left(\hat{E}[g(X)|x_i] - g(x_i)\right)E[v^*(X) - \hat{\theta}(Z)\hat{f}_X(X)|x_i]\right| &\le \left[\frac{1}{n}\sum_{i=1}^n \left(\hat{E}[g(X)|x_i] - g(x_i)\right)^2\right]^{\frac{1}{2}}\left[\frac{1}{n}\sum_{i=1}^n m^2(x_i,\hat{\theta})\right]^{\frac{1}{2}} \\ &+ \left[\frac{1}{n}\sum_{i=1}^n \left(\hat{E}[g(X)|x_i] - g(x_i)\right)^2\right]^{\frac{1}{2}}\left[\frac{1}{n}\sum_{i=1}^n \left(E[\hat{\theta}(Z)|x_i]\right)^2(\hat{f}_X(x_i) - f_X(x_i))^2\right]^{\frac{1}{2}} = o_p(n^{-\frac{1}{2}}) \quad (96)\end{aligned}$$

In (61) I show that $n^{-1}\sum_i g(x_i)E[v^*(X) - \hat{\theta}(Z)\hat{f}_X(X)|x_i] = n^{-1}\sum_i \hat{E}[\tilde{v}_n(Z)|x_i]\hat{f}_X(x_i)E[v^*(X) - \hat{\theta}(Z)\hat{f}_X(X)|x_i] + o_p(n^{-\frac{1}{2}})$. The first result in (97) follows by the Cauchy-Schwarz inequality and the same arguments used in (96). The same arguments as in (49) but with $v^*(x) = 0$ and $f_X(x) = 1$ and Assumption 3(v) imply $\sup_{\Theta_n} n^{-1}\sum_i(\hat{E}[\theta_n(Z)|x_i] - E[\theta_n(Z)|x_i])^2 = o_p(n^{-\frac{1}{2}})$. In addition, as shown in (50), $n^{-1}\sum_i(\hat{f}_X(x_i) - f_X(x_i))^2 = o_p(n^{-\frac{1}{2}})$. Therefore, it follows that $n^{-1}\sum_i(g(x_i) - \hat{E}[\tilde{v}_n(Z)|x_i]\hat{f}_X(x_i))^2 = o_p(n^{-\frac{1}{2}})$, which implies the final result in (97).

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n \left(\hat{E}[\tilde{v}_n(Z)|x_i]\hat{f}_X(x_i) - E[\tilde{v}(Z)|x_i]f_X(x_i)\right)E[v^*(X) - \hat{\theta}(Z)\hat{f}_X(X)|x_i]\right| \\ &\qquad \le \left[\frac{1}{n}\sum_{i=1}^n \left(\hat{E}[\tilde{v}_n(Z)|x_i]\hat{f}_X(x_i) - E[\tilde{v}(Z)|x_i]f_X(x_i)\right)^2\right]^{\frac{1}{2}} \times o_p(n^{-\frac{1}{4}}) = o_p(n^{-\frac{1}{2}}) \quad (97)\end{aligned}$$

The arguments in (96) and (97) imply that $n^{-1}\sum_i(\hat{E}[E[\tilde{v}(Z)|X]f_X(X)|x_i] - \hat{E}[\tilde{v}_n(Z)|x_i]\hat{f}_X(x_i))^2 = o_p(n^{-\frac{1}{2}})$. Since in addition $E[\varepsilon(x_i,z_i)|x_i] = 0$, Markov's inequality implies (98).

$$\frac{1}{n}\sum_{i=1}^n \hat{E}[E[\tilde{v}(Z)|X]f_X(X)|x_i]\varepsilon(x_i,z_i) = \frac{1}{n}\sum_{i=1}^n \hat{E}[\tilde{v}_n(Z)|x_i]\hat{f}_X(x_i)\varepsilon(x_i,z_i) + o_p(n^{-\frac{1}{2}}) \quad (98)$$

Combining (95), (96), (97) and (98) establishes the claim of the Lemma. ■.

Proof of Theorem 3.6: In order to establish the asymptotic normality of $n^{-\frac{1}{2}}(\sum_i y_i\tilde{\theta}(z_i,x_i) - \langle v^*, m_0\rangle)$, we use that since $\theta_0 \in \Theta_0$ we have $E[\theta_0(Z)|x] = v^*(x)/f_X(x)$ to obtain (99).

$$\begin{aligned}n^{-\frac{1}{2}}\sum_{i=1}^n y_i\tilde{\theta}(z_i,x_i) - \langle v^*, m_0\rangle &= n^{-\frac{1}{2}}\sum_{i=1}^n y_i\theta_0(z_i) - E[Y\theta_0(Z)] + n^{\frac{1}{2}}E[Y(\tilde{\theta}(X,Z) - \theta_0(Z))] \\ &+ n^{-\frac{1}{2}}\sum_{i=1}^n y_i(\tilde{\theta}(z_i,x_i) - \theta_0(z_i)) - E[Y(\tilde{\theta}(Z,X) - \theta_0(Z))] \quad (99)\end{aligned}$$

I first show the third term in (99) converges in probability to zero. Notice that if $\tilde{\theta} \in \hat{\Theta}_0 + \lambda_n \mathcal{E}_n$, then it can be written as $\tilde{\theta} = \bar{\theta} + \lambda_n \hat{E}[\theta_n(Z)|x]$ for some $\bar{\theta} \in \hat{\Theta}_0$ and $\theta_n \in \Theta_n$. Define the class of functions $\mathcal{F} = \{y\theta(z) : \theta \in \Theta\}$. Since $E[Y^2(\theta_1(Z) - \theta_2(Z))^2] \leq E[Y^2]||\theta_1 - \theta_2||_\infty^2$, it follows that $N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_{\mathcal{L}^2}) \leq N_{[]}(\epsilon/E[Y^2], \Theta, ||\cdot||_\infty) \leq N_{[]}(\epsilon/E[Y^2], \Lambda_C^\omega(\mathcal{Z}), ||\cdot||_\infty) \lesssim e^{\epsilon^{-\frac{d_z}{\omega}}}$, by Theorem 2.7.1 in van der Vaart & Wellner (2000). Therefore, Theorem 2.5.6 in van der Vaart & Wellner (2000) implies $\mathcal{F}$ is Donsker. Furthermore, Theorem 3.5 establishes $||\tilde{\theta} - \theta_0||_\infty = o_p(1)$, while by (91) $\lambda_n \sup_{\Theta_n} ||\hat{E}[\theta_n(Z)|x]||_\infty = o_p(1)$ and hence we have that $||\bar{\theta} - \theta_0||_\infty = o_p(1)$. Together with $\mathcal{F}$ being Donsker, this implies the second result in (100).

$$n^{-\frac{1}{2}} \sum_{i=1}^n y_i(\tilde{\theta}(z_i, x_i) - \theta_0(z_n)) - E[Y(\tilde{\theta}(Z, X) - \theta_0(Z))] = n^{-\frac{1}{2}} \sum_{i=1}^n y_i(\bar{\theta}(z_i) - \theta_0(z_i)) - E[Y(\bar{\theta}(Z) - \theta_0(Z))]$$
$$+ \lambda_n n^{-\frac{1}{2}} \sum_{i=1}^n \hat{E}[\theta_n(Z)|x_n] - E[\hat{E}[\theta_n(Z)|X]] = \lambda_n n^{-\frac{1}{2}} \sum_{i=1}^n \hat{E}[\theta_n(Z)|x_n] - E[\hat{E}[\theta_n(Z)|X]] + o_p(1) \quad (100)$$

Furthermore, by (91) we also have that $\sup_{\Theta_n} \lambda_n ||\hat{E}[\theta_n(Z)|x]||_\infty = O_p(\lambda_n \xi_{0n} \tilde{\lambda}_n^{-1})$, for $\tilde{\lambda}_n$ the smallest eigenvalue of $P'P$. Hence, using (100) we derive the first result in (101). In Theorem 1 of Newey (1997) it is shown that $\tilde{\lambda}_n^{-1} = O_p(1)$, and therefore Assumption 7(iii) implies the final result in (101).

$$n^{-\frac{1}{2}} \sum_{i=1}^n y_i(\tilde{\theta}(z_i, x_i) - \theta_0(z_i)) - E[Y(\tilde{\theta}(Z, X) - \theta_0(Z))] = O_p(n^{\frac{1}{2}} \lambda_n \xi_{0n} \tilde{\lambda}_n^{-1}) + o_p(1) = o_p(1) \quad (101)$$

I now examine the second term in (99). Note that since $\tilde{\theta} = \bar{\theta} + \lambda_n \hat{E}[\theta_n(Z)|x]$ for some $\bar{\theta} \in \hat{\Theta}_0$ and $\theta_n \in \Theta_n$, and in addition $E[Y\theta(Z)] = \langle \theta, \tilde{v} \rangle_w$ for all $\theta \in \overline{V}$ (but not necessarily for $\tilde{\theta}$), we can obtain the first equality in (102) by rearranging terms. In addition, $n^{\frac{1}{2}} \lambda_n |E[\hat{E}[\theta_n(Z)|X](Y - E[\tilde{v}(Z)|X])]| \leq n^{\frac{1}{2}} \lambda_n \sup_{\Theta_n} ||\hat{E}[\theta_n|X]||_\infty E[|Y - E[\tilde{v}(Z)|X]|] = o_p(1)$ by the same arguments as in (101), which in turn implies the last result in (102).

$$n^{\frac{1}{2}} E[Y(\tilde{\theta}(X, Z) - \theta_0(Z))] = n^{\frac{1}{2}} \langle \tilde{\theta} - \theta_0, \tilde{v} \rangle_w + n^{\frac{1}{2}} \lambda_n E[\hat{E}[\theta_n|X](Y - E[\tilde{v}(Z)|X])] = n^{\frac{1}{2}} \langle \tilde{\theta} - \theta_0, \tilde{v} \rangle_w + o_p(1) \quad (102)$$

Combining (102) with part b) of Lemma .2 we can derive the first equality in (103). In turn, parts c) and d) of Lemma .2 implies the second equality in (103).

$$n^{\frac{1}{2}} E[Y(\tilde{\theta}(X, Z) - \theta_0(Z))] = n^{-\frac{1}{2}} \sum_{i=1}^n E[\tilde{v}(Z) f_X(X)|x_i](\hat{m}(x_i, \theta_0) - \hat{m}(x_i, \tilde{\theta})) + o_p(1)$$
$$= -n^{-\frac{1}{2}} \sum_{i=1}^n E[\tilde{v}(Z) f_X(X)|x_i] \hat{m}(x_i, \tilde{\theta}) - n^{-\frac{1}{2}} \sum_{i=1}^n E[\tilde{v}(Z)|x_i] \theta_0(z_i) f_X^2(x_i) - E[\tilde{v}(Z) v^*(X) f_X(X)] + o_p(1) \quad (103)$$

To conclude, we need to show $n^{-\frac{1}{2}} \sum_i E[\tilde{v}(Z) f_X(X)|x_i] \hat{m}(x_i, \tilde{\theta}) = o_p(1)$. In (104), we use the fact that $\tilde{\theta} = \bar{\theta} + \lambda_n \theta_n$ for some $\bar{\theta} \in \hat{\Theta}_0$ and $\theta_n \in \Theta_n$, and use Assumption 6(ii) to derive the first inequality with probability tending to one for some $\epsilon_n = o(n^{-\frac{1}{2}})$. Expanding the square and collecting terms gives the second equality in (104).

$$0 \geq n^{-1} \sum_{i=1}^n (v^*(x_i) - (\bar{\theta}(z_i) - \lambda_n \hat{E}[\theta_n(Z)|x_i]) \hat{f}_X(x_i))^2 - n^{-1} \sum_{i=1}^n (v^*(x_i) - (\bar{\theta}(z_i) - \lambda_n \hat{E}[\theta_n(Z) + u_n(Z)\epsilon_n/\lambda_n|x_i]) \hat{f}_X(x_i))^2$$
$$= 2\epsilon_n n^{-1} \sum_{i=1}^n (v^*(x_i) - \tilde{\theta}(x_i, z_i) \hat{f}_X(x_i)) \hat{E}[u_n(Z)|x_i] \hat{f}_X(x_i) - \epsilon_n^2 n^{-1} \sum_{i=1}^n (\hat{E}[u_n(Z)|x_i] \hat{f}_X(x_i))^2 \quad (104)$$

The arguments in (62) imply $\sup_x |\hat{E}[u_n(Z)|x] \hat{f}_X(x)| = O_p(1)$, so that $\epsilon_n^2 n^{-1} \sum_i (\hat{E}[u_n(Z)|x_i] \hat{f}_X(x_i))^2 = O_p(\epsilon_n^2)$. Since (104) holds for $u = \pm \tilde{v}$ and $\epsilon_n = o(n^{-\frac{1}{2}})$, (104) implies $n^{-1} \sum_i (v^*(x_i) - \tilde{\theta}(x_i, z_i) \hat{f}_X(x_i)) \hat{E}[u_n(Z)|x_i] \hat{f}_X(x_i) = o_p(n^{-\frac{1}{2}})$. We now use this result to show $n^{-\frac{1}{2}} \sum_i E[\tilde{v}(Z) f_X(X)|x_i] \hat{m}(x_i, \tilde{\theta}) = o_p(1)$. In (105), the first equality follows by $\tilde{\theta}(x_i, z_i) = \bar{\theta}(z_i) + \lambda_n \hat{E}[\theta_n(Z)|x_i]$ for some $\bar{\theta} \in \hat{\Theta}_0$ and $\theta_n \in \Theta_n$. As argued in (62), however,

$\sup_{x,\Theta_n} |\hat{E}[\theta(Z)|x]\hat{f}_X(x)| = O_p(1)$. In addition, since $||\tilde{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$ by Theorem 3.5, it must be that $||\bar{\theta} - \theta_0||_w = o_p(n^{-\frac{1}{4}})$. Hence, because $\bar{\theta} \in \Theta_n$ Lemma .3 implies the second equality in (105). The final two results in (105) then follow by $\lambda_n n^{\frac{1}{2}} \to 0$, $\sup_{x,\Theta_n} |\hat{E}[\theta_n(Z)|x]\hat{f}_X(x)| = O_p(1)$ and $|E[\tilde{v}(Z)|x]f_X(x)|$ being bounded.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(v^*(x_i) - \tilde{\theta}(x_i,z_i)\hat{f}_X(x_i))\hat{E}[\tilde{v}_n(Z)|x_i]\hat{f}_X(x_i) &= \frac{1}{n}\sum_{i=1}^{n}(v^*(x_i) - \bar{\theta}(z_i)\hat{f}_X(x_i))\hat{E}[\tilde{v}_n(Z)|x_i]\hat{f}_X(x_i) \\ &- \frac{\lambda_n}{n}\sum_{i=1}^{n}\hat{E}[\tilde{v}_n(Z)|x_i]\hat{f}_X^2(x_i)\hat{E}[\theta_n(Z)|x_i] = \frac{1}{n}\sum_i E[\tilde{v}(Z)f_X(X)|x_i]\hat{m}(x_i,\bar{\theta}) + o_p(n^{-\frac{1}{2}}) + O_p(\lambda_n) \\ &= \frac{1}{n}\sum_i E[\tilde{v}(Z)f_X(X)|x_i]\hat{m}(x_i,\tilde{\theta}) + \frac{\lambda_n}{n}\sum_i E[\tilde{v}(Z)f_X(X)|x_i]\hat{E}[\theta_n(Z)|x_i]\hat{f}_X(x_i) + o_p(n^{-\frac{1}{2}}) = o_p(n^{-\frac{1}{2}}) \quad (105)\end{aligned}$$

Combining (99), (101), (103) and (105) together with the Central Limit Theorem concludes the proof. ■

Proof of Lemma 3.1: I begin by establishing that $||\hat{v} - \tilde{v}||_w = o_p(1)$. Let $C(\theta) = E[(E[\theta(Z)|X])^2 f_X^2(X)]/2 - E[Y\theta(Z)]$ and $C_n(\theta) = n^{-1}\sum_i (\hat{E}[\theta(Z)|x_i])^2 \hat{f}_X^2(x_i)/2 - y_i\theta(z_i)$. In (106) I show $\sup_{\Theta_n}\sum_i |C_n(\theta) - C(\theta)| = o_p(1)$. The first inequality in (106) follows by $\Theta_n \subseteq \Theta$. Let $\mathcal{A}_1 = \{f_X^2(x)(E[\theta(Z)|x])^2 : \theta \in \Theta\}$. Since $\theta \in \Theta$ and $f_X(x)$ are uniformly bounded, $\sup_x |f_X^2(x)(E[\theta_1(z)|x])^2 - f_X^2(x)(E[\theta_2(z)|x])^2| \lesssim ||\theta_1 - \theta_2||_\infty$. Therefore, $N_{[]}(\epsilon, \mathcal{A}_1, ||\cdot||_\infty) \lesssim N_{[]}(\epsilon, \Theta, ||\cdot||_\infty) < \infty$ for all $\epsilon > 0$ by Theorem 2.7.1 in van der Vaart & Wellner (1996). Hence, the class $\mathcal{A}_1$ is Glivenko Cantelli by Theorem 2.4.1 in van der Vaart & Wellner (1996) and therefore it follows that $\sup_\Theta |n^{-1}\sum_i f_X^2(x_i)(E[\theta(Z)|x_i])^2 - E[f_X^2(X)(E[\theta(Z)|X])^2]| \xrightarrow{p} 0$. Similarly, let $\mathcal{A}_2 = \{y\theta(z) : \theta(z) \in \Theta\}$. Because $E[|Y(\theta_1(Z) - \theta_2(Z)|] \leq E[|Y|]||\theta_1 - \theta_2||_\infty$, it follows that $N_{[]}(\epsilon, \mathcal{A}_2, ||\cdot||_{\mathcal{L}^1}) \lesssim N_{[]}(\epsilon, \Theta, ||\cdot||_\infty) < \infty$ for all $\epsilon > 0$. Hence, $\sup_\Theta |n^{-1}\sum_i y_i\theta(z_i) - E[Y\theta(Z)]| \xrightarrow{p} 0$. As discussed in (62), $||\hat{f}_X - f_X||_\infty = o_p(1)$. Theorem 1 in Newey (1997) implies that under Assumption 1, 2(i)-(ii), 3(v) and 4(i) pointwise in θ, $||\hat{E}[\theta|x] - E[\theta|x]||_\infty = o_p(1)$. Theorem 1 in Newey 1997 actually implies the result holds uniformly in Θ under our uniformity conditions in Assumption 2(ii). Combining these results implies the final result in (106).

$$\begin{aligned}\sup_{\Theta_n}|C_n(\theta) - C(\theta)| &\leq \sup_\Theta \left|\frac{1}{2n}\sum_{i=1}^{n}\hat{f}_X^2(x_i)(\hat{E}[\theta(Z)|x_i])^2 - f_X^2(x_i)(E[\theta(Z)|x_i])^2\right| \\ &+ \sup_\Theta \left|\frac{1}{2n}\sum_{i=1}^{n} f_X^2(x_i)(E[\theta(Z)|x_i])^2 - \frac{1}{2}E[f_X^2(X)(E[\theta(Z)|X])^2]\right| + \sup_\Theta \left|\frac{1}{n}\sum_{i=1}^{n} y_i\theta(z_i) - E[Y\theta(Z)]\right| \xrightarrow{p} 0 \quad (106)\end{aligned}$$

In (107) we conclude showing $||\hat{v} - \tilde{v}||_w = o_p(1)$. As argued in Section 3.6, the set of minimizers to $C(\theta)$ form an equivalence class under $||\cdot||_w$ which includes $\tilde{v}$. The first inequality in (107) follows for $\tilde{v}_{0n} = \arg\inf_{\Theta_n} ||\tilde{v} - \theta_n||_\infty$ and $||\tilde{v}_{0n} - \tilde{v}||_\infty = o(1)$. By Newey & Powell (2003), Θ is compact under $||\cdot||_\infty$ and hence also under $||\cdot||_w$. Since $C(\theta)$ is continuous under $||\cdot||_w$, it follows that $\min_{\Theta \cap ||\theta - \tilde{v}||_w > \epsilon} C(\theta) - C(\tilde{v}) > \delta$ for some $\delta > 0$. Using this δ we obtain the second inequality in (107). In turn (106) gives us the third inequality in (107). The final result is implied by $||\tilde{v}_{0n} - \tilde{v}||_\infty = o(1)$ and the continuity of $C(\theta)$ under $||\cdot||_\infty$.

$$\begin{aligned}P\left(||\hat{v} - \tilde{v}||_w > \epsilon\right) \leq P\left(\inf_{||\theta - \tilde{v}||_w > \epsilon} C_n(\theta) \leq C_n(\tilde{v}_{0n})\right) &\leq P\left(\inf_{||\theta - \tilde{v}||_w > \epsilon} C_n(\theta) \leq C_n(\tilde{v}_{0n}) \bigcap \sup_{\Theta_n}|C_n(\theta) - C(\theta)| < \frac{\delta}{2}\right) \\ &+ P\left(\sup_{\Theta_n}|C_n(\theta) - C(\theta)| > \frac{\delta}{2}\right) \leq P\left(\inf_{||\theta - \tilde{v}||_w > \epsilon} C(\theta) \leq C(\tilde{v}_{0n}) + \delta\right) + o(1) \to 0 \quad (107)\end{aligned}$$

I now proceed to establish $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$. In (108) the first equality follows by $||\tilde{\theta} - \theta_0||_\infty = o_p(1)$ by Theorem 3.5, $||\hat{f}_X - f_X||_\infty = o_p(1)$ and $\sup_\Theta ||\hat{E}[\theta|x] - E[\theta|x]||_\infty = o_p(1)$ as discussed in the derivation of (106). Markov's inequality, $||\tilde{v} - \hat{v}||_w = o_p(1)$ and θ_0 and f_X bounded in turn imply the second equality in (108). A law of large

numbers gives us the last result in (108).

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{E}[\hat{v}(Z)|x_i]\hat{f}_X^2(x_i))\tilde{\theta}(z_i,x_i) = \frac{1}{n}\sum_{i=1}^{n}(y_i - E[\hat{v}(Z)|x_i]f_X^2(x_i))\theta_0(z_i) + o_p(1)$$
$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - E[\tilde{v}(Z)|x_i]f_X^2(x_i))\theta_0(z_i) + o_p(1) \xrightarrow{p} E\left[Y(-E[\tilde{v}(Z)|X]f_X^2(X))\theta_0(Z)\right] \quad (108)$$

Identical arguments show $n^{-1}\sum_i \left(y_i - \hat{E}[\hat{v}(Z)|x_i]\hat{f}_X^2(x_i)\right)^2 \tilde{\theta}^2(z_i,x_i) \xrightarrow{p} E\left[(Y - E[\tilde{v}(Z)|X]f_X^2(X))^2\theta_0^2(Z)\right]$. The continuous mapping theorem completes the proof of the Lemma. ■

APPENDIX D - Proof of Lemmas 2.1, 2.2 and 2.3

Proof of Lemma 2.1: In order to analyze the sign of $\langle v_i^*, m_0\rangle$, we begin in (109) by examining the integral at different intervals. The second equality in (109) follows by doing the change of variables $u = x + \pi/i$ in the first component of each summand and noting that $\sin i(u - \pi/i) = \sin(iu - \pi) = -\sin iu$.

$$\langle v^*, m_0\rangle = \int_{-\pi}^{\pi}\sin(ix)m_0(x)dx = \sum_{k=1}^{i}\left(\int_{-\pi+\frac{(2k-1)\pi}{i}}^{-\pi+\frac{2k\pi}{i}}\sin(ix)m_0(x)dx + \int_{-\pi+\frac{2k\pi}{i}}^{-\pi+\frac{(2k-1)\pi}{i}}\sin(ix)m_0(x)dx\right)$$
$$= \sum_{k=1}^{i}\left(\int_{-\pi+\frac{(2k-1)\pi}{i}}^{-\pi+\frac{2k\pi}{i}}[m_0(u) - m_0(u-\pi/i)]\sin(iu)du\right) \quad (109)$$

By assumption, however, $m_0(x)$ is weakly increasing, and therefore $m_0(u) - m_0(u-\pi/i) \geq 0$. In addition, $\sin(iu) > 0$ on intervals of the form $\left(-\pi + \frac{(2k-1)\pi}{i}, -\pi + \frac{2k\pi}{i}\right)$ if i is odd and $\sin(iu) < 0$ on these intervals if i is even. It therefore follows that $sign\{\langle v_i^*, m_0\rangle\} = (-1)^{i+1}$. ■

Proof of Lemma 2.2: In order to see how concavity determines the sign of $\langle v_i^*, m_0\rangle$, in (110) we integrate by parts to obtain the first equality. The second equality follows from $\sin(i\pi) = \sin(-i\pi) = 0$ for all i.

$$\langle v_i^*, m_0\rangle = \int_{-\pi}^{\pi}\cos(ix)m_0(x)dx = \frac{\sin(ix)}{i}m_0(x)\Big|_{-\pi}^{\pi} - \int_{-\pi}^{\pi}\frac{\sin(ix)}{i}\frac{\partial m_0(x)}{\partial x}dx = -\int_{-\pi}^{\pi}\frac{\sin(ix)}{i}\frac{\partial m_0(x)}{\partial x}dx \quad (110)$$

Since $m_0(x)$ is concave, however, it follows that $\frac{\partial m_0(x)}{\partial x}$ is weakly decreasing. Hence, by Lemma 2.1 we have that $sign\left\{-\int_{-\pi}^{\pi}\sin(ix)\frac{\partial m_0(x)}{\partial x}dx\right\} = (-1)^{i+1}$, which establishes the Lemma. ■

Proof of Lemma 2.3: To establish the Lemma, simply use the assumed additive separability and Fubini's Theorem to attain the first equality in (111).

$$\langle v_{ij}^*, m_0\rangle = \int_{-\pi}^{\pi}\sin(ix_1)m_{01}(x_1)dx_1\int_{-\pi}^{\pi}\sin(jx_2)dx_2 + \int_{-\pi}^{\pi}\sin(jx_2)m_{02}(x_2)dx_2\int_{-\pi}^{\pi}\sin(ix_1)dx_1 = 0 \quad (111)$$

The final result in (111) follows from $\int_{-\pi}^{\pi}\sin(ix)dx = 0$ for all i. ■

APPENDIX E - TABLES

In this appendix we report results to help understand what the levels of the parameters γ and λ_n mean in the Monte Carlos of Section 4. Recall that the first stage estimator $\tilde{\theta}(z,x)$ is of the form $\tilde{\theta}(z,x) = q^{k_{1n}'}(z)\hat{\beta}_1 + \lambda_n\hat{E}[q_{k_{1n}'}(Z)|x]\hat{\beta}_2$. To understand the magnitude of λ_n, we examine how much $\lambda_n\hat{E}[q^{k_{1n}'}(Z)|x]\hat{\beta}_2$ contributes to the variability of $\tilde{\theta}(z,x)$. For each replication and specification of (γ,λ), we calculated $||\tilde{\theta}(z,x)||_n^2 = \frac{1}{n}\sum_i\tilde{\theta}^2(z_i,x_i)$ and

Table 2: $\text{MEAN}\Big(||\lambda_n \hat{E}[q^{k'_{1n}}(Z)|x]\hat{\beta}_2||_n \Big/ ||\tilde{\theta}(z,x)||_n\Big)$

	$\lambda = 0$	$\lambda = 0.001$	$\lambda = 0.01$	$\lambda = 0.1$
$\gamma = 0.01$	0	0.0034	0.0106	0.0188
$\gamma = 0.02$	0	0.0034	0.0144	0.0248
$\gamma = 0.03$	0	0.0036	0.0186	0.0311
$\gamma = 0.04$	0	0.0042	0.0231	0.0372
$\gamma = 0.05$	0	0.0045	0.0283	0.0432
$\gamma = 0.06$	0	0.0061	0.0333	0.0490
$\gamma = 0.07$	0	0.0070	0.0389	0.0551
$\gamma = 0.08$	0	0.0089	0.0442	0.0609
$\gamma = 0.09$	0	0.0117	0.0495	0.0668
$\gamma = 0.10$	0	0.0148	0.0551	0.0728

$||\lambda_n \hat{E}[q^{k'_{1n}}(Z)|x]\hat{\beta}_2||_n^2 = \frac{1}{n}\sum_i \lambda_n^2 (\hat{E}[q^{k'_{1n}}(Z)|x_i]\hat{\beta}_2)^2$. Table 2 reports the mean of $||\lambda_n \hat{E}[q^{k'_{1n}}(Z)|x]\hat{\beta}_2||_n / ||\tilde{\theta}(z,x)||_n$ across replications. For all choices of λ_n, the endogenous component $\lambda_n \hat{E}[q^{k'_{1n}}(Z)|x]\hat{\beta}_2$ is a small part of the total variability of $\tilde{\theta}(z,x)$, with the ratio never exceeding 10%.

Let $\epsilon_n(\gamma) = \gamma Q_n(\bar{\theta}) + (1-\gamma) Q_n(\hat{\theta})$ denote the rule used in the Monte Carlos of Section 4 to select the bandwidth ϵ_n. To understand how γ translates into different values for $\epsilon_n(\gamma)$, we calculated the value of $\epsilon_n(\gamma)/Q_n(\hat{\theta})$ for each replication and choice of γ. Table 3 reports the mean across replications of these series. The different values of γ span a wide range of selections of $\epsilon_n(\gamma)$ from a modest increment of $Q_n(\hat{\theta})$ by 18% for $\gamma = 0.01$ to a more considerable increment of 180% for $\gamma = 0.1$.

Table 3: $\text{MEAN}\Big(\epsilon_n(\gamma)/Q_n(\hat{\theta})\Big)$

	$\gamma = 0.01$	$\gamma = 0.02$	$\gamma = 0.03$	$\gamma = 0.04$	$\gamma = 0.05$	$\gamma = 0.06$	$\gamma = 0.07$	$\gamma = 0.08$	$\gamma = 0.09$	$\gamma = 0.1$
$\epsilon_n(\gamma)/Q_n(\hat{\theta})$	1.1796	1.3591	1.5387	1.7183	1.8979	2.0774	2.2570	2.4366	2.6162	2.7957

References

[1] Ai, C. and Chen, X., "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions". *Econometrica*, 71, pp. 1795-1844, (2003).

[2] Blundell, R., Chen, X. and Kristensen, D., "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves". Working paper, Department of Economics, University College London, (2004).

[3] Bosq, D., "Nonparametric Statistics for Stochastic Processes". Springer-Verlag, New York (1998).

[4] Chalak, K. and White, H., "An Extended Class of Instrumental Variables for the Estimation of Causal Effects". Working Paper, University of California at San Diego (2006)

[5] Chen, X., "Large Sample Sieve Estimation of Semi-Nonparametric Models". Working paper, Department of Economics, New York Unversity (2006).

[6] Chernozhukov, V., Hong, H. and Tamer, E., "Estimation and Confidence Regions for Parameter Sets in Econometric Models". *Econometrica*, 75, pp. 1243-1284 (2007).

[7] Chesher, A., "Identification in Nonseparable Models". *Econometrica*, 71, pp. 1405-1441 (2003).

[8] Chesher, A., "Nonparametric Identification Under Discrete Variation". *Econometrica*, 73, pp. 1525-1550 (2005).

[9] Chesher, A., "Instrumental Values". *Journal of Econometrics*, 139, pp.15-34 (2007)

[10] Darolles, S., Florens, J. and Renault, E., "Nonparametric Instrumental Regression". Working paper, GREMAQ, University of Tolouse, (2003).

[11] Hall, P. and Horowitz, J., "Nonparametric Methods for Inference in the Presense of Instrumental Variables". *Annals of Statistics*, 33, pp. 2904-2929, (2005).

[12] Horowitz, J. "Asymptotic Normality of a Nonparametric Instrumental Variables Estimator". Working Paper, Department of Economics, Northwestern University (2007).

[13] Imbens, G. W. and Newey W. K., "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity". Working paper, Department of Economics, Harvard, (2006).

[14] Manski, C. F., "Nonparametric Bounds of Treatment Effects". *American Economic Review, Paper and Proceedings*, 80, pp. 319-323 (1990).

[15] Manski, C. F., "Partial Identification of Probability Distributions". Springer, New York, (2003).

[16] Newey, W. K., "Convergence Rates and Asymptotic Normality for Series Estimators". *Journal of Econometrics*, 79, pp. 147-168 (1997).

[17] Newey, W. K. and McFadden, D., "Large Sample Estimation and Hypothesis Testing". in *Handbook of Econometrics, vol IV*, ed. by R. Engle and D. McFadden, Elsevier Science B.V., 2111-2245.

[18] Newey, W. K. and Powell, J., "Instrumental Variables Estimation of Nonparametric Models". *Econometrica*, 71, pp. 1565-1578, (2003).

[19] Newey, W. K., Powell, J. L. and Vella F., "Nonparametric Estimation of Triangular Simultaneous Equation Models". *Econometrica*, 67, pp. 565-603, (1999).

[20] Pagan, A. and Ullah, A., "Nonparametric Econometrics". Cambridge University Press, Cambridge, (1999).

[21] Politis, D. N., Romano, J. P. and Wolf, M., "Subsampling". Springer, New York (1999).

[22] Polyanin A. D. and Manzhirov A. V., "Handbook of Integral Equation". CRC Press, Boca Raton (1998).

[23] Romano, J. P. and Shaikh A. M, "Inference for the Identified Set in Partially Identified Models", Working paper, Department of Economics, University of Chicago (2006b).

[24] Santos, A., "Inference in Nonparametric Instrumental Variables with Partial Identification". Working paper, Department of Economics, University of California at San Diego, (2007).

[25] Schennach, S., Chalak, K. and White, H. "Estimating Average Marginal Effects in Nonseparable Structural Systems". Working paper, Department of Economics, University of California at San Diego, (2007).

[26] Severini, T. A. and Tripathi, G., "Some Identification Issues in Nonparametric Linear Models with Endogenous Regressors". *Econometric Theory*, 22, pp. 258-278, (2006).

[27] Severini, T. A. and Tripathi, G., "Efficiency Bounds for Estimating Linear Functionals of Nonparametric Regression Models with Endogenous Regressors". Working Paper, University of Connecticut, (2007).

[28] Stinchcombe, M. B. and White, H. "Some Measurability Results for Extrema of Random Functions over Random Sets". *Review of Economic Studies*, 59, pp. 495-512, (1992).

[29] van der Vaart, A. W., "Asymptotic Statistics". Cambridge University Press, Cambridge, (1998).

[30] van der Vaart, A. W. and Wellner, J., "Weak Convergence and Empirical Processes - With Applications to Statistics". Springer, New York, (1996).

[31] White, H. and Chalak, K., "A Unified Framework for Defining and Identifying Causal Effects". Working Paper, Univeristy of California at San Diego (2006).