Estimating Misspecified Moment Inequality Models

Hiroaki Kaido and Halbert White

Abstract This chapter studies partially identified structures defined by a finite number of moment inequalities. When the moment function is misspecified, it becomes difficult to interpret the conventional identified set. Even more seriously, this can be an empty set. We define a pseudo-true identified set whose elements can be interpreted as the least-squares projections of the moment functions that are observationally equivalent to the true moment function. We then construct a set estimator for the pseudo-true identified set and establish its $O_p(n^{-1/2})$ rate of convergence.

1 Introduction

This chapter develops a new approach to estimating structures defined by moment inequalities. Moment inequalities often arise as optimality conditions in discrete choice problems or in structures where economic variables are subject to some type of censoring. Typically, parametric models are used to estimate such structures. For example, in their analysis of an entry game in the airline markets, Ciliberto and Tamer (2009) use a linear specification for airlines' profit functions and assume that unobserved heterogeneity in the profit functions can be captured by independent normal random variables. In asset pricing theory with short sales prohibited, Luttmer (1996) specifies the functional form of the pricing kernel as a power function of

H. Kaido (✉)
Department of Economics, Boston University, 270 Bay State Rd., Boston,
MA 02215, USA
e-mail: hkaido@bu.edu

H. White
Department of Economics (0508), University of California, San Diego,
9500 Gilman Dr., La Jolla, CA 92093-0508, USA
e-mail: hwhite@ucsd.edu

X. Chen and N. R. Swanson (eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, DOI: 10.1007/978-1-4614-1653-1_13,

consumption growth, based on the assumption that the investor's utility function is additively separable and isoelastic.

Any conclusions drawn from such methods rely on the validity of the model specification. Although commonly used estimation and inference methods for moment inequality models are robust to potential lack of identification, typically they are not robust to misspecification. Compared to cases where the parameter of interest is point identified, much less is known about the consequences of misspecified moment inequalities. As we will discuss, these can be serious. In general, misspecification makes it hard to interpret the estimated set of parameter values; an even more serious possibility is that the identified set could be an empty set. If the identified set is empty, every nonempty estimator sequence is inconsistent. Furthermore, it is often hard to see if the estimator is converging to some object that can be given any meaningful interpretation. An exception is the estimation method developed by Ponomareva and Tamer (2010), which focuses on estimating a regression function with interval censored outcome variables.

This chapter develops a new estimation method that is robust to potential parametric misspecification in general moment inequality models. Our contributions are three-fold. First, we define a pseudo-true identified set that is nonempty under mild assumptions and that can be interpreted as the projection of the set of function-valued parameters identified by the moment inequalities. Second, we construct a set estimator using a two-stage estimation procedure, and we show that the estimator is consistent for the pseudo-true identified set in Hausdorff metric. Third, we give conditions under which the proposed estimator converges to the pseudo-true identified set at the $n^{-1/2}$-rate.

The first stage is a nonparametric estimator of the true moment function. Given this, why perform a parametric second-stage estimation? After all, the nonparametric first stage estimates the same object of interest, without the possibility of parametric misspecification. There are a variety of reasons a researcher may nevertheless prefer to implement the parametric second stage: first is the undeniably appealing interpretability of the parametric specification; second is the much more precise estimation and inference afforded by using a parametric specification; and third, the second term of the second-stage objective function may offer a potentially useful model specification diagnostic. Future research may permit deriving the asymptotic distribution of this term under the null of correct parametric specification to provide a formal test. The two-stage procedure proposed here delivers these benefits, while avoiding the more serious adverse consequences of potential misspecification.

The chapter is organized as follows. Section 2 describes the data generating process and gives examples that fall within the scope of this chapter. We also introduce our definition of the pseudo-true identified Sect. 3 defines our estimator and presents our main results. We conclude in Sect. 4. We collect all proofs into the appendix.

2 The Data Generating Process and the Model

Our first assumption describes the data generating process (DGP).

Assumption 2.1 Let $(\Omega, \mathfrak{F}, \mathbb{P}_0)$ be a complete probability space. Let $k, \ell \in \mathbb{N}$. Let $X : \Omega \to \mathbb{R}^k$ be a Borel measurable map, let $\mathcal{X} \subseteq \mathbb{R}^k$ be the support of X, and let P_0 be the probability measure induced by X on $\mathcal{X}$. Let $\rho_0 : \mathcal{X} \to \mathbb{R}^\ell$ be an unknown measurable function such that $E[\rho_0(X)]$ exists and

$$E[\rho_0(X)] \leq 0, \tag{1}$$

where the expectation is taken with respect to P_0.

In what follows, we call ρ_0 the *true moment function*. The moment inequalities (1) often arise as an optimality condition in game-theoretic models (Bajari et al. 2007; Ciliberto and Tamer 2009) or models that involve variables that are subject to some kind of censoring (Manski and Tamer 2002). In empirical studies of such models, it is common to specify a parametric model for ρ_0.

Assumption 2.2 Let $p \in \mathbb{N}$ and let Θ be a subset of $\mathbb{R}^p$ with nonempty interior. Let $m : \mathcal{X} \times \Theta \to \mathbb{R}^\ell$ be such that $m(\cdot, \theta)$ is measurable for each $\theta \in \Theta$ and $m(x, \cdot)$ is continuous on Θ, a.e. $- P_0$. For each $\theta \in \Theta$, $m(\cdot, \theta) \in L^2_\ell := \{f : \mathcal{X} \to \mathbb{R}^\ell : E[f(X)'f(X)] < \infty\}$.

Throughout, we call $m(\cdot, \cdot)$ the *parametric moment function*.

Definition 2.1 Let $m_\theta(\cdot) := m(\cdot, \theta)$. Define $\mathcal{M}_\Theta := \{m_\theta \in L^2_\ell : \theta \in \Theta\}$. $\mathcal{M}_\Theta$ is *correctly specified* $(-P_0)$ if there exists $\theta_0 \in \Theta$ such that

$$P_0[\rho_0(X) = m(X, \theta_0)] = 1.$$

Otherwise, the model is misspecified.

If the model is correctly specified, we may define the set of parameter values that can be identified by the inequalities in (1):

$$\Theta_I := \{\theta \in \Theta : E[m(X, \theta)] \leq 0\}.$$

We call Θ_I the *conventional identified set*. This set collects all parameter values that yield parametric moment functions that are observationally equivalent to ρ_0.

It becomes difficult to interpret Θ_I when the model is misspecified, as pointed out by Ponomareva and Tamer (2010) for a regression model with an interval-valued outcome variable. Suppose first that the model is misspecified but Θ_I is nonempty. The set is still a collection of parameter values that are observationally equivalent to each other, but since there is no θ in Θ_I that corresponds to the true moment function, further structure is required to unambiguously interpret Θ_I as a collection of "pseudo-true parameter(s)". Further, Θ_I may be empty, especially if $\mathcal{M}_\Theta$ is a small class of functions. This makes the interpretation of Θ_I even more difficult. In fact, interpretation is impossible, as there is nothing to interpret.

Often, the economics of a given problem impose further structure on the DGP. To specify this, we let $0 < L \leq \ell$, and for measurable $s : \mathcal{X} \to \mathbb{R}^L$, let $\|s\|_L := E[s(X)'s(X)]^{1/2}$. Let $L_L^2 := \{s : \mathcal{X} \to \mathbb{R}^L, \|s\|_L < \infty\}$, and let $\mathcal{S} \subseteq L_L^2$.

Assumption 2.3 There exists $\varphi : \mathcal{X} \times \mathcal{S} \to \mathbb{R}^\ell$ such that for each $x \in \mathcal{X}$, $\varphi(x, \cdot)$ is continuous on $\mathcal{S}$ and for each $s \in \mathcal{S}$, $\varphi(\cdot, s)$ is measurable. Further, there exists $s_0 \in \mathcal{S}$ such that

$$\rho_0(x) = \varphi(x, s_0), \quad \forall x \in \mathcal{X}.$$

When $\rho_0 \in L_\ell^2$ and there is no further structure on ρ_0 available, we let $L = \ell$, $\mathcal{S} = L_\ell^2$, and take φ to be the evaluation functional $e : \mathcal{X} \times \mathcal{S} \to \mathbb{R}^\ell$:

$$\varphi(x, s) = e(x, s) \equiv s(x),$$

as then $\varphi(x, \rho_0) = e(x, \rho_0) \equiv \rho_0(x)$ and $s_0 = \rho_0$. In this case, it is not necessary to explicitly introduce φ. Often, however, further structure on the form of ρ_0 is available. Typically, this is reflected in s depending non-trivially only on a strict subvector of X, say X_1. In such cases, we may write $\mathcal{S} \subseteq L_{\mathcal{X}_1}^2$ for clarity. We give several examples below.

When Assumption 2.3 holds, we typically parametrize the unknown function s_0. For example, it is common to specify s_0 as a linear function of some of the components of x. As we will see in the examples, a common modeling assumption is

Assumption 2.4 There exists $r : \mathcal{X} \times \Theta \to \mathbb{R}^L$ such that with $r_\theta := r(\cdot, \theta)$,

$$m(x, \theta) = \varphi(x, r_\theta), \quad \forall (x, \theta) \in \mathcal{X} \times \Theta.$$

Thus, misspecification occurs when there is no θ_0 in Θ such that $s_0 = r_{\theta_0}$.

More generally, misspecification can occur because the researcher mistakenly imposes Assumption 2.3, in which case s_0 fails to exist and there is again no θ_0 in Θ such that $\rho_0(x) = \varphi(x, r_{\theta_0})$. As s_0 is an element of an infinite-dimensional space, we may refer to this as "nonparametric" misspecification. To proceed, we assume that, as is often plausible, the researcher is sufficiently able to specify the structure of interest that nonparametric misspecification is not an issue, either because correct φ restrictions are imposed or no φ restrictions are imposed. We thus focus on the case of parametric misspecification, where s_0 exists but there is no θ_0 in Θ such that $s_0 = r_{\theta_0}$.

2.1 Examples

In this section, we present several motivating examples and also give commonly used parametric specifications in these examples. For any vector x, we use $x^{(j)}$ to denote the jth component of the vector. Similarly, for a vector valued function $f(x)$, we use $f^{(j)}(x)$ to denote the jth component of $f(x)$.

Example 2.1 (Interval censored outcome) Let $Z : \Omega \to \mathbb{R}^{dz}$ be a regressor with support $\mathcal{Z}$. Let $Y : \Omega \to \mathbb{R}$ be an outcome variable that is generated as:

$$Y = s_0(Z) + \epsilon, \tag{2}$$

where $s_0 \in \mathcal{S} := L^2_{\mathcal{Z}}$, say, and ϵ satisfies $E[\epsilon|Z] = 0$. We let $\mathcal{Y}$ denote the support of Y. Suppose Y is unobservable, but there exist $(Y_L, Y_U)' : \Omega \to \mathcal{Y} \times \mathcal{Y}$ such that $Y_L \leq Y \leq Y_U$ almost surely. Then, $(Y_L, Y_U, Z)'$ satisfies the following inequalities almost surely:

$$E[Y_L|Z] - s_0(Z) \leq 0 \tag{3}$$

$$s_0(Z) - E[Y_U|Z] \leq 0. \tag{4}$$

Let $x = (y_L, y_U, z)' \in \mathcal{X} := \mathcal{Y} \times \mathcal{Y} \times \mathcal{Z}$. Given a collection $\{A_1, \ldots, A_K\}$ of Borel subsets of $\mathcal{Z}$, the inequalities in (3), (4) imply that the moment inequalities in (1) hold with

$$\rho_0(x) = \varphi(x, s_0) := \begin{bmatrix} y_L - s_0(z) \\ s_0(z) - y_U \end{bmatrix} \otimes 1_A(z), \tag{5}$$

where $1_A(z) := (1\{z \in A_1\}, \ldots, 1\{z \in A_K\})'$.[1] For each $x \in \mathcal{X}$ and $s \in \mathcal{S}$, the functional φ evaluates vertical distances of $r(z)$ from y_L and y_U and multiplies them by the indicator function evaluated at z. Additional information on ρ_0 available in this example is that the moment functions are based on the vertical distances.

A common specification for s_0 is $s_0(z) = r_{\theta_0}(z) = z'\theta_0$ for some $\theta_0 \in \Theta \subseteq \mathbb{R}^{dz}$. The parametric moment function is then given for each $x \in \mathcal{X}$ by $m(x, \theta) = \varphi(x, r_\theta)$. Therefore, this example satisfies Assumption 2.4.

Example 2.2 Tamer (2003) considers a simultaneous game of complete information. For each $j = 1, 2$, let $Z_j : \Omega \to \mathbb{R}^{dz}$ and $\epsilon_j : \Omega \to \mathbb{R}$ be firm j's characteristics that are observable to the firms. The econometrician observes the Z's but not the ϵ's. For each j, let $g_j : \mathcal{Z} \times \{0, 1\} \to \mathbb{R}$. These functions are known to the firms but not to the econometrician. Suppose that each firm's payoff is given by

$$\pi_j(Z_j, Y_j, Y_{-j}) = (\epsilon_j - g_j(Z_j, Y_{-j}))Y_j, \quad j = 1, 2,$$

[1] Here, we take the indicators (or instruments) $1_A(z)$ as given. The indicators $1_A(z)$ could be replaced by any finite vector of measurable non-negative functions of z. Andrews and Shi (2011) give examples of such functions.

where $Y_j \in \mathcal{Y} := \{0, 1\}$ is firm j's entry decision, and $Y_{-j} \in \mathcal{Y}$ is the other firm's entry decision. The econometrician observes these decisions. Given (z_1, z_2), the firms' payoffs can be summarized in Table 1.

Suppose the firms and the econometrician know that $g(z, 1) \geq g(z, 0)$ for any value of z. This means that, other things equal, the opponent's entry would reduce the firm's own profit. In this setting, there are several possible equilibrium outcomes depending on the realization of (ϵ_1, ϵ_2). If $\epsilon_1 > g_1(z_1, 1)$ and $\epsilon_2 > g_2(z_2, 1)$, then $(1, 1)$ is the unique Nash equilibrium (NE) outcome. Similarly, if $\epsilon_1 > g_1(z_1, 1)$ and $\epsilon_2 < g_2(z_2, 1)$, $(1, 0)$ is the unique NE outcome, and if $\epsilon_1 < g_1(z_1, 1)$ and $\epsilon_2 > g_2(z_2, 1)$, $(0, 1)$ is the unique NE outcome. Now, if $\epsilon_1 < g_1(z_1, 1)$ and $\epsilon_2 < g_2(z_2, 1)$, there are two Nash equilibria, and they give the outcomes $(1, 0)$ and $(0, 1)$. Let $F_j, j = 1, 2$ be the unknown CDFs of ϵ_1 and ϵ_2.[2] Without any assumptions on the equilibrium selection mechanism, the model predicts the following set of inequalities:

$$P(Y_1 = 1, Y_2 = 1|Z_1 = z_1, Z_2 = z_2) = (1 - F_1(g_1(z_1, 1)))(1 - F_2(g_2(z_2, 1))) \tag{6}$$

$$P(Y_1 = 1, Y_2 = 0|Z_1 = z_1, Z_2 = z_2) \geq (1 - F_1(g_1(z_1, 1)))F_2(g_2(z_2, 1)) \tag{7}$$

$$P(Y_1 = 1, Y_2 = 0|Z_1 = z_1, Z_2 = z_2) \leq F_2(g_2(z_2, 1)). \tag{8}$$

Let $x := (y_1, y_2, z_1, z_2)' \in \mathcal{X} := \mathcal{Y} \times \mathcal{Y} \times \mathcal{Z} \times \mathcal{Z}$. Let $s_0 \in \mathcal{S} := \{s \in L^2_{\mathcal{Z}\times\mathcal{Z}} : s(z_1, z_2) \in [0, 1]^2, \forall(z_1, z_2) \in \mathcal{Z} \times \mathcal{Z}\}$ be defined by

$$\begin{aligned} s_0^{(1)}(z_1, z_2) &:= F_1(g_1(z_1, 1)) \\ s_0^{(2)}(z_1, z_2) &:= F_2(g_2(z_2, 1)). \end{aligned}$$

Here, $s_0^{(j)}(z_1, z_2)$ is the conditional probability that firm j's profit upon entry is negative given z_1 and z_2. Given a collection $\{A_j, j = 1, \ldots, K\}$ of Borel subsets of $\mathcal{Z} \times \mathcal{Z}$, let $1_A(z) := (1\{(z_1, z_2) \in A_1\}, \ldots, 1\{(z_1, z_2) \in A_K\})'$. The inequalities (6)–(8) imply the moment inequalities in (1) hold with

$$\begin{aligned} \rho_0(x) &= \varphi(x, s_0) \\ &= \begin{pmatrix} 1\{y_1 = 1, y_2 = 1\} - (1 - s_0^{(1)}(z_1, z_2))(1 - s_0^{(2)}(z_1, z_2)) \\ (1 - s_0^{(1)}(z_1, z_2))(1 - s_0^{(2)}(z_1, z_2)) - 1\{y_1 = 1, y_2 = 1\} \\ (1 - s_0^{(1)}(z_1, z_2))s_0^{(2)}(z_1, z_2) - 1\{y_1 = 1, y_2 = 0\} \\ 1\{y_1 = 1, y_2 = 0\} - s_0^{(2)}(z_1, z_2) \end{pmatrix} \otimes 1_A(z). \end{aligned}$$

The additional information on ρ_0 is that it is based on the differences between some combinations of the conditional probabilities $s_0(z_1, z_2)$ and indicators for specific events.

[2] The players do not need to know the F's, but these are important to the econometrician.

Table 1 The entry game payoff matrix

$Y_1\backslash Y_2$	0	1
0	$(0,0)$	$(0,\epsilon_2-g_2(z_2,0))$
1	$(\epsilon_1-g_1(z_1,0),0)$	$(\epsilon_1-g_1(z_1,1),\epsilon_2-g_2(z_2,1))$

A common parametric specification for g_j is $g_j(z_j,y_{-j})=z_j'\gamma_0-y_{-j}\beta_{j,0}$ for some $\beta_{j,0}\in B\subseteq\mathbb{R}_+$ and $\gamma_0\in\Gamma\subseteq\mathbb{R}^{dz}$. It is also common to assume that F_j, $j=1,2$ belong to a known parametric class $\{F(\cdot;\alpha),\alpha\in\mathcal{A}\}$ of distributions. Then the parametric moment function can be defined for each x by $m(x,\theta):=\varphi(x,r_\theta)$, where $\theta:=(\alpha_1,\alpha_2,\beta_1,\beta_2,\gamma)'$ and

$$r_\theta^{(1)}(z_1,z_2)=F(z_1'\gamma-\beta_1;\alpha_1) \tag{9}$$

$$r_\theta^{(2)}(z_1,z_2)=F(z_2'\gamma-\beta_2;\alpha_2). \tag{10}$$

This example also satisfies Assumption 2.4.

Example 2.3 (Discrete choice) Suppose an agent chooses $Z\in\mathbb{R}^{dz}$ from a set $\mathcal{Z}:=\{z_1,\ldots,z_K\}$ in order to maximize her expected payoff $E[s_0(Y,Z)\mid\mathcal{I}]$, where Y is a vector of observable random variables, $s_0\in\mathcal{R}:=L^2_{\mathcal{Y}\times\mathcal{Z}}$ is the payoff function, and $\mathcal{I}$ is the agent's information set. The optimality condition for the agent's choice is given by:

$$E[s_0(Y,z_j)-s_0(Y,Z)\mid\mathcal{I}]\le 0,\quad j=1,\ldots,K. \tag{11}$$

Let $x:=(y,z)'\in\mathcal{X}:=\mathcal{Y}\times\mathcal{Z}$. The optimality conditions in (11) imply that the unconditional moment inequalities in (1) hold with

$$\rho_0(x)=\varphi(x,s_0)=\begin{pmatrix}\begin{bmatrix}s_0(y,z_1)-s_0(y,z_1)\\ \vdots\\ s_0(y,z_K)-s_0(y,z_1)\end{bmatrix}\times 1\{z=z_1\}\\ \vdots\\ \begin{bmatrix}s_0(y,z_1)-s_0(y,z_K)\\ \vdots\\ s_0(y,z_K)-s_0(y,z_K)\end{bmatrix}\times 1\{z=z_K\}\end{pmatrix}.$$

For given y, the functional φ evaluates the profit differences between a given choice z (e.g., z_1) and every other possible choice. The additional information on ρ_0 is that it is based on the profit differences.

A common specification for s_0 is $s_0(y,z)=r_{\theta_0}(y,z)=\psi(y,z;\alpha_0)+z'\beta_0+\epsilon_z$ for some known function ψ, unknown $(\alpha_0,\beta_0)\in\Theta\subset\mathbb{R}^{d_\alpha+d_\beta}$, and an unobservable choice-dependent error ϵ_z. For simplicity, we assume that ϵ_z satisfies $E[\epsilon_{z_i}-\epsilon_{z_j}\mid\mathcal{I}]=0$ for any i,j; see Pakes et al (2006) and Pakes (2010) for detailed discussions.

The parametric moment function is then given for each $x \in \mathcal{X}$ by $m(x,\theta) = \varphi(x, r_\theta)$. This example satisfies Assumption 2.4.

Example 2.4 (Pricing kernel) Let $Z : \Omega \to \mathbb{R}^{d_Z}$ be the payoffs of d_Z securities that are traded at a price of $P \in \mathcal{P} \subseteq \mathbb{R}^{d_Z}$. If short sales are not allowed for any securities, then the feasible set of portfolio weights is restricted to $\mathbb{R}_+^{d_Z}$ and the standard Euler equation does not hold. Instead, the following Euler inequalities hold (see Luttmer 1996):

$$E[s_0(Y)Z - P] \leq 0,$$

where $Y : \Omega \to \mathcal{Y}$ is a state variable, e.g. consumption growth, and $s_0 \in \mathcal{S} := \{s \in L^2_{\mathcal{Y}} : s(y) \geq 0, \forall y \in \mathcal{Y}\}$ is the pricing kernel function. The moment inequalities thus hold with the true moment function:

$$\rho_0(x) = \varphi(x, s_0) = s_0(y)z - p,$$

where $x := (y, z, p)' \in \mathcal{Y} \times \mathcal{Z} \times \mathcal{P}$. This function evaluates the pricing kernel r at y and computes a vector of pricing errors. The additional information on ρ_0 is that it is based on the pricing errors.

A common specification for s_0 is $s_0(y) = r_{\theta_0}(y) = \beta_0 y^{-\gamma_0}$, where $\beta_0 \in B \subseteq [0, 1]$ is the investor's subjective discount factor and $\gamma_0 \in \Gamma \subseteq \mathbb{R}_+$ is the relative risk aversion coefficient. Let $\theta := (\beta, \gamma)'$. The parametric moment function is then given for each $x \in \mathcal{X}$ by $m(x, \theta) = \varphi(x, r_\theta)$, satisfying Assumption 2.4.

2.2 Projection

The inequality restrictions $E[\varphi(X, s_0)] \leq 0$ may not uniquely identify s_0. Define

$$\mathcal{S}_0 := \{s \in \mathcal{S} : E[\varphi(X, s)] \leq 0\}.$$

We define a pseudo-true identified set of parameters as a collection of projections of elements in $\mathcal{S}_0$. Let W be a given non-random finite $L \times L$ symmetric positive-definite matrix. For each $s \in \mathcal{S}$, define the norm $\|s\|_W := E[s(X)'Ws(X)]^{1/2}$. For each $s \in \mathcal{S}$ and $A \subseteq \mathcal{S}$, the projection map $\Pi_A : \mathcal{S} \to A$ is the map such that

$$\|s - \Pi_A s\|_W = \inf_{a \in A} \|s - a\|_W.$$

Let $\mathcal{R}_\Theta := \{r_\theta \in \mathcal{S} : \theta \in \Theta\}$. Given Assumption 2.4, we can define

$$\Theta_* := \{\theta \in \Theta : r_\theta = \Pi_{\mathcal{R}_\Theta} s, s \in \mathcal{S}_0\}.$$

When φ is the evaluation map e, Θ_* is simply $\Theta_* := \{\theta \in \Theta : m_\theta = \Pi_{\mathcal{M}_\Theta} s, s \in \mathcal{S}_0\}$.

Θ_* can be interpreted as the set of parameters that correspond to the elements m_θ in the $\mathcal{R}_\Theta$ -projection of $\mathcal{S}_0$. This set is nonempty (under some regularity conditions), and each element can be interpreted as a projection of s inducing a functional $\varphi(\cdot, s)$ that is observationally equivalent to ρ_0. In this sense, each element in Θ_* has an interpretation as a pseudo-true value. Thus, we call Θ_* the *pseudo-true identified set*. [White (1982) uses θ_* to denote the unique pseudo-true value in the fully identified case.]

We illustrate the relationship between Θ_I and Θ_* with an example. Consider Example 2.1. Let $\Theta \subseteq \mathbb{R}^{d_Z}$. The conventional identified set is given by

$$\Theta_I = \{\theta \in \Theta : E[(Y_L - Z'\theta)1\{Z \in A_j\}] \leq 0, \\ \text{and } E[(Z'\theta - Y_U)1\{Z \in A_j\}] \leq 0, \quad j = 1, \ldots, K\}. \tag{12}$$

The pseudo-true identified set is given by

$$\Theta_* = \{\theta \in \Theta : \theta = E[ZZ']^{-1}E[Zs(Z)], s \in \mathcal{S}_0\}. \tag{13}$$

Let D be a $d_Z \times K$ matrix whose jth column is $E[Z\,1\{Z \in A_j\}]$. For this example, the following result holds:

Proposition 2. 1 *Let the conditions of Example 2.1 hold, and let Θ_* be given as in (13). Let Θ_I be given as in (12). Then $\Theta_I \subseteq \Theta_*$. Suppose further that $\mathcal{M}_\Theta$ is correctly specified, that $E[Y_U|Z] = E[Y_L|Z] = Z'\theta_0$ a.s, and that $d_Z \leq rank(D)$. Then $\Theta_I = \Theta_* = \{\theta_0\}$.*

As this example shows, unless there is some information that helps restrict $\mathcal{S}_0$ very tightly, Θ_I is often a proper subset of Θ_*. This is because without such information, $\mathcal{S}_0$ is typically a much richer class of functions than $\mathcal{R}_\Theta$. Another important point to note is that, although Θ_* is well-defined generally, Θ_I can be empty quite easily. In particular, for any $x, x' \in \mathcal{X}$, let $x_\lambda := \lambda x + (1-\lambda)x', 0 \leq \lambda \leq 1$. Θ_I is empty if there exists (x, x') and $\lambda \in [0, 1]$ such that (i) $x_\lambda \in \mathcal{X}$ and $(E[Y_L|x_\lambda] - E[Y_U|x])/\|x_\lambda - x\| > (E[Y_U|x'] - E[Y_U|x])/\|x' - x\|$ or (ii) $x_\lambda \in \mathcal{X}$ and $(E[Y_U|x_\lambda] - E[Y_L|x])/\|x_\lambda - x\| < (E[Y_L|x'] - E[Y_L|x])/\|x' - x\|$.[3] Figure 1, which is similar to Fig. 1 in Ponomareva and Tamer (2010), illustrates an example that satisfies condition (i) for the one-dimensional case.

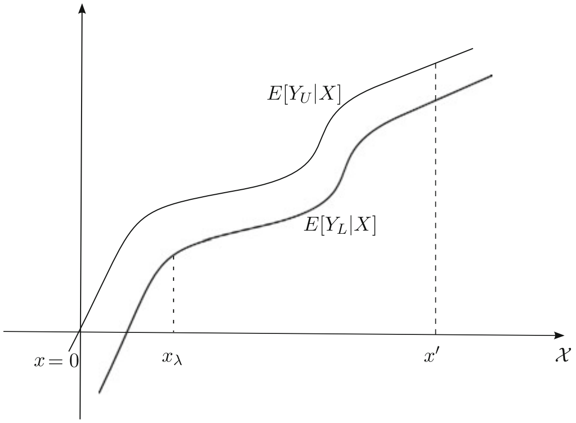
In this example, each element in Θ_* solves the following moment restrictions:

$$E[Z(Z'\theta - Y)] = E[Zu(X)], \tag{14}$$

with $u(x) = s(z) - y$ for some $s \in \mathcal{S}_0$. This can be viewed as a special case of *incomplete linear moment restrictions* studied in Bontemps, Magnac, and Maurin (2011)

[3] For this example, Θ_I is never empty as long as the number $(2K)$ of moment inequalities equals the number of parameters (ℓ).

Fig. 1 An example with an empty conventional identified set



(BMM, henceforth).[4] BMM shows that the set of parameters that solves incomplete linear moment restrictions is necessarily convex and develops an inference method that exploits this property.

We note here that this connection to BMM's work only occurs when the parametric class is of the form: $\mathcal{R}_\Theta = \{r_\theta : r_\theta(z) = z'\theta,\ \theta \in \Theta\}$. The elements of Θ_*, however, do not generally solve incomplete linear moment restrictions when $\mathcal{R}_\Theta$ includes nonlinear functions of θ. Therefore, BMM's inference method is only applicable when r_θ is linear. Our estimation procedure is more flexible than theirs in the following two respects. First, one may allow projection to a more general class of parametric functions that includes nonlinear functions of θ. Second, as a consequence of the first point, we do not require Θ_* to be convex. We, however, pay a price for achieving this generality. We require s to satisfy suitable smoothness conditions, which are not required by BMM. We discuss these conditions in detail in the following section.

3 Estimation

3.1 Set Estimator

For W as above and each $(\theta, s) \in \Theta \times \mathcal{S}$, let the *population criterion function* be defined by

$$\begin{aligned} Q(\theta, s) = {} & E[(s(X_i) - r_\theta(X_i))' W(s(X_i) - r_\theta(X_i))] \\ & - \inf_{\vartheta \in \Theta} E[(s(X_i) - r_\vartheta(X_i))' W(s(X_i) - r_\vartheta(X_i))]. \end{aligned} \tag{15}$$

[4] We are indebted to an anonymous referee for pointing out a relationship between BMM's framework and ours. General incomplete linear moment restrictions are given by $E[V(Z'\theta - Y)] = E[Vu(V)]$, where V is a vector of random variables, and u is an unknown bounded function. See BMM for details.

Using the population criterion function, the "pseudo-true" identified set Θ_* can be equivalently written as

$$\Theta_* = \{\theta : Q(\theta, s) = 0, \quad s \in \mathcal{S}_0\}.$$

Given a sample $\{X_1, \ldots, X_n\}$ of observations, let the *sample criterion function* be defined for each $(\theta, s) \in \Theta \times \mathcal{S}$ by

$$\begin{aligned} Q_n(\theta, s) := & \frac{1}{n}\sum_{i=1}^{n}(s(X_i) - r_\theta(X_i))' W(s(X_i) - r_\theta(X_i)) \\ & - \inf_{\vartheta \in \Theta} \frac{1}{n}\sum_{i=1}^{n}(s(X_i) - r_\vartheta(X_i))' W(s(X_i) - r_\vartheta(X_i)). \end{aligned} \tag{16}$$

Ideally, we would like to estimate Θ_* by $\tilde{\Theta}_n$, say, where $\tilde{\Theta}_n := \{\theta : Q_n(\theta, s) \leq c_n, s \in \mathcal{S}_0\}$. But $\mathcal{S}_0$ is unknown, so we must estimate it. Thus, we employ a two-stage procedure, similar to that studied in Kaido and White (2010). Section 3.3 discusses how to construct a first-stage estimator of $\mathcal{S}_0$. For now, we suppose that such an estimator exists. For this, let $\mathcal{F}(A)$ be the set of closed subsets of a set A. See Kaido and White (2010) for background, including discussion of Effros measurability.

Assumption 3.1 (First-stage estimator) For each n, let $\mathcal{S}_n \subseteq \mathcal{S}$. $\hat{\mathcal{S}}_n : \Omega \to \mathcal{F}(\mathcal{S}_n)$ is (Effros-) measurable.

Given a first-stage estimator, we define a set estimator for the pseudo-true identified set. Let $\{c_n\}$ be a sequence of non-negative constants. The set estimator for Θ_* is defined by

$$\hat{\Theta}_n := \{\theta \in \Theta : Q_n(\theta, s) \leq c_n, s \in \hat{\mathcal{S}}_n\}. \tag{17}$$

We establish our consistency results using the Hausdorff metric. Let $||\cdot||$ denote the Euclidean norm, and for any closed subsets A and B of a finite-dimensional Euclidean space (e.g., containing θ), let

$$d_H(A, B) := \max\{\vec{d}_H(A, B), \vec{d}_H(B, A)\}, \quad \vec{d}_H(A, B) := \sup_{a \in A} \inf_{b \in B} \|a - b\|, \tag{18}$$

where d_H and $\vec{d}_H$ are the Hausdorff metric and directed Hausdorff distance respectively.

Before stating our assumptions, we introduce some additional notation. Let D_θ^α denote the differential operator $\partial^\alpha / \partial\theta_1^{\alpha_1} \cdots \partial\theta_p^{\alpha_p}$ with $|\alpha| := \sum_{j=1}^{p} \alpha_j$. Similarly, we let D_x^β denote the differential operator $\partial^\beta / \partial x_1^{\beta_1} \cdots \partial x_k^{\beta_k}$ with $|\beta| := \sum_{j=1}^{k} \beta_j$. For a function $f : \mathcal{X} \to \mathbb{R}$ and $\gamma > 0$, let $\underline{\gamma}$ be the smallest integer smaller than γ and define

$$\|f\|_\gamma := \max_{|\beta|\le \underline{\gamma}} \sup_{x\in\mathcal{X}} \left|D_x^\beta f(x)\right| + \max_{|\beta|=\underline{\gamma}} \sup_{x,y\in\mathcal{X}} \frac{\left|D_x^\beta f(x) - D_x^\beta f(x)\right|}{\|x - -y\|^{\gamma-\underline{\gamma}}}.$$

Let $\mathcal{C}_M^\gamma(\mathcal{X})$ be the set of all continuous functions $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_\gamma \le M$. Let $\mathcal{C}_{M,L}^\gamma(\mathcal{X}) := \{f : \mathcal{X} \to \mathbb{R}^L : f^{(j)} \in \mathcal{C}_M^\gamma(\mathcal{X}), j = 1, \ldots, L\}$. Finally, for any $\eta > 0$, let $\mathcal{S}_0^\eta := \{s \in \mathcal{S} : \inf_{s'\in\mathcal{S}_0} \|s - s'\|_W < \eta\}$.

Our first assumption places conditions on the parameter spaces Θ and $\mathcal{S}$. We let $int(\Theta)$ denote the interior of Θ.

Assumption 3.2 (i) Θ is compact; (ii) $\mathcal{S}$ is a compact convex set with nonempty interior; (iii) there exists $\gamma > k/2$ such that $\mathcal{S} \subseteq \mathcal{C}_{M,L}^\gamma(\mathcal{X})$; (iv) $\mathcal{R}_\Theta$ is a convex subset of $\mathcal{S}$; (v) $\Theta_* \subseteq int(\Theta)$.

Assumption 3.2 (i) is standard in the literature of extremum estimation and also ensures the compactness of the pseudo-true identified set. Assumption 3.2 (iii) imposes a smoothness requirement on each component of $s \in \mathcal{S}$. Together with Assumption (ii), this implies that $\mathcal{S}$ is compact under the uniform norm, which will be also used for establishing the Hausdorff consistency of $\hat{\mathcal{S}}_n$ in the following section. For the Hausdorff consistency of $\hat{\Theta}_n$, the requirement $\gamma > k/2$ can be relaxed to $\gamma > 0$, and it also suffices that the smoothness requirement holds for functions in neighborhoods of $\mathcal{S}_0$. The stronger requirement given here, however, will be useful for deriving the rates of convergence of $\hat{\Theta}_n$ and $\hat{\mathcal{S}}_n$.

For ease of analysis, we assume below that the observations are from a sample of IID random vectors.

Assumption 3.3 The observations $\{X_i, i = 1, \ldots, n\}$ are independently and identically distributed.

The following two assumptions impose regularity conditions on r_θ.

Assumption 3.4 (i) $r(x, \cdot)$ is twice continuously differentiable on the interior of Θ a.e. $- P_0$, and for any j, x, and $|\alpha| \le 2$, there exists a measurable bounded function $C : \mathcal{X} \to \mathbb{R}$ such that $|D_\theta^\alpha r_\theta^{(j)}(x) - D_\theta^\alpha r_{\theta'}^{(j)}(x)| \le C(x)\|\theta - \theta'\|$; (ii) there exists a measurable bounded function $R : \mathcal{X} \to \mathbb{R}$ such that

$$\max_{\substack{j=1,\ldots,l\\ |\alpha|\le 2}} \sup_{\theta\in\Theta} \left|D_\theta^\alpha r_\theta^{(j)}(x)\right| \le R(x).$$

For each x, let $\nabla_\theta r_\theta(x)$ be a $L \times p$ matrix whose jth row is the gradient vector of $r_\theta^{(j)}$ with respect to θ. For each $x \in \mathcal{X}$ and $i, j \in \{1, \ldots, L\}$, let $\partial^2/\partial\theta_i\partial\theta_j r_\theta(x)$ be a $L \times 1$ vector whose kth component is given by $\partial^2/\partial\theta_i\partial\theta_j r_\theta^{(k)}(x)$. For each $\theta \in \Theta$, $s \in \mathcal{S}$, and $x \in \mathcal{X}$, let $H_W(\theta, s, x)$ be a $p \times p$ matrix whose (i, j)th component is given by

$$H_W^{(i,j)}(\theta, s, x) = 2\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j} r_\theta(x)\right)' W(r_\theta(x) - s(x)). \qquad (19)$$

Let $\eta > 0$. For each $s \in \mathcal{S}_0^\eta$ and $\epsilon > 0$, let $V^\epsilon(s)$ be the neighborhood of $\theta_*(s)$ defined by

$$V^\epsilon(s) := \{\theta \in \Theta : \|\theta - \theta_*(s)\| \le \epsilon\}.$$

Let $\mathcal{N}_{\epsilon,\eta} := \{(\theta, s) : \theta \in V^\epsilon(s), s \in \mathcal{S}_0^\eta\}$ be the graph of the correspondence V^ϵ on $\mathcal{S}_0^\eta$.

Assumption 3.5 There exist $\bar{\epsilon} > 0$ and $\bar{\eta} > 0$ such that the Hessian matrix $\nabla_\theta^2 Q(\theta, s) := E[H_W(\theta, s, X_i) + 2\nabla_\theta r_\theta(X_i)' W \nabla_\theta r_\theta(X_i)]$ is positive definite uniformly over $\mathcal{N}_{\bar{\epsilon},\bar{\eta}}$.

Assumption 3.4 imposes a smoothness requirement on r_θ as a function of θ, enabling us to expand the first order condition for minimization, as is standard in the literature. Assumption 3.5 requires that Hessian of $Q(\theta, s)$ with respect to θ to be positive definite uniformly on a suitable neighborhood of $\Theta_* \times \mathcal{S}_0$. For the consistency of $\hat{\Theta}_n$, it suffices to assume that the Hessian is uniformly non-singular over $\mathcal{N}_{\bar{\epsilon},\bar{\eta}}$, but a stronger condition given here will be useful to ensure a quadratic approximation of the criterion function, which is crucial for the $\sqrt{n}$-consistency of $\hat{\Theta}_n$.

Further, we assume that $\hat{\mathcal{S}}_n$ is consistent for $\mathcal{S}_0$ in a suitable Hausdorff metric. Specifically, for subsets A, B of $\mathcal{S}$, let

$$d_{H,W}(A, B) := \max\left\{\sup_{a\in A}\inf_{b\in B} \|a - b\|_W, \sup_{b\in B}\inf_{a\in A} \|a - b\|_W\right\}.$$

Assumption 3.6 $d_{H,W}(\hat{\mathcal{S}}_n, \mathcal{S}_0) = o_p(1)$.

Theorem 3.1 is our first main result, which establishes the consistency of the set estimator defined in (17) with c_n set to 0. This result is established by extending the standard consistency proof for extremum estimators to the current setting. Note that, under Assumption 3.2 (iv), the projection $\theta_*(s) := \Pi_{\mathcal{R}_\Theta} s$ of each point $s \in \mathcal{S}$ to $\mathcal{R}_\Theta$ exists and is uniquely determined. In other words, for each $s \in \mathcal{S}$, $\theta_*(s)$ is point identified. By setting $c_n = 0$, the set estimator is then asymptotically equivalent to the collection of minimizers $\hat{\theta}_n(s) := \text{argmin}_{\theta' \in \Theta} Q_n(\theta, s)$ of the sample criterion function. The main challenge for establishing Hausdorff consistency is to show that $\hat{\theta}_n(s) - \theta_*(s)$ vanishes in probability over a sufficiently large neighborhood of $\mathcal{S}_0$. The proof of the theorem in the appendix formally establishes this and gives the desired result.

Theorem 3.1 *Suppose Assumptions 2.1–2.4 and 3.1–3.6 hold. Let $\hat{\Theta}_n$ be defined as in (17) with $c_n = 0$ for all n. Then $d_H(\hat{\Theta}_n, \Theta_*) = o_p(1)$.*

The result of Theorem 3.1 is similar to that of Theorem 3.2 in Chernozhukov et al. (2007), who establish the Hausdorff consistency of a level-set estimator with $c_n = 0$

when Q_n degenerates on a neighborhood of the identified set.[5] When Assumption 3.2 (iv) fails to hold, this estimator may not be consistent. We, however, conjecture that it would be possible to construct a Hausdorff consistent estimator of Θ_* even in such a setting by choosing a positive sequence $\{c_n\}$ of levels that tends to 0 as $n \to \infty$ and by exploiting the fact that $\hat{\mathcal{S}}_n$ converges to $\mathcal{S}_0$ in a suitable Hausdorff metric. In fact, Kaido and White (2010) establish the Hausdorff consistency of their two-stage set estimator using this argument, but in their analysis, the first-stage parameter (s in our setting) must be finite dimensional. Extending Theorem 3.1 to a more general one that allows non-convex parametric classes is definitely of interest, but to keep our tight focus here, we leave this as a future work.

3.2 The Rate of Convergence

Theorem 3.1 uses the fact that $d_H(\hat{\Theta}_n, \Theta_*)$ can be bounded by $d_{H,W}(\hat{\mathcal{S}}_n, \mathcal{S}_0)$. Although $\hat{\mathcal{S}}_n$ does not converge at a parametric rate generally, the convergence rate of $\hat{\Theta}_n$ can be improved when $\hat{\mathcal{S}}_n$ converges to $\mathcal{S}_0$ at a rate $o_p(n^{-1/4})$. This is analogous to the results obtained for the point identified case; see, for example, Newey (1994), Ai and Chen (2003), and Ichimura and Lee (2010).

Assumption 3.7 $d_{H,W}(\hat{\mathcal{S}}_n, \mathcal{S}_0) = o_p(n^{-1/4})$.

Theorem 3.2 *Suppose the conditions of Theorem 3.1 hold. Suppose in addition Assumption 3.7 holds. Let $\hat{\Theta}_n$ be defined as in (17) with $c_n = 0$ for all n. Then, $d_H(\hat{\Theta}_n, \Theta_*) = O_p(n^{-1/2})$.*

For this, setting c_n to 0 is crucial for achieving the $O_p(n^{-1/2})$ rate. We here note that Theorem 3.2 builds on Lemma A.2 in the appendix, which establishes the convergence rate (in directed Hausdorff distance) of $\hat{\Theta}_n$ in (17) with a possibly nonzero level c_n. This lemma does not require Assumption 3.2 (iv) but assumes the Hausdorff consistency of $\hat{\Theta}_n$ as a high-level condition. This is why Theorem 3.2 is stated for $\hat{\Theta}_n$ with $c_n = 0$. As previously discussed, however, if Theorem 3.1 is extended to allow non-convex parametric classes, this lemma can be used to characterize the estimator's convergence rate under a more general setting.

3.3 The First-Stage Estimator

This section discusses how to construct a first-stage set estimator. A challenge is that the object of interest $\mathcal{S}_0$ is a subset of an infinite-dimensional space. This requires us to use a nonparametric estimation technique for estimating $\mathcal{S}_0$. This type of estimation

[5] Their framework does not consider misspecification. Their object of interest is therefore the conventional identified set Θ_I. In our setting, the sample criterion function degenerates, i.e., $Q_n(\theta, s) = 0$, on a neighborhood of $\Theta_* \times \mathcal{S}_0$ under Assumption 3.2 (iv).

problem was recently analyzed in Santos (2011), who studies estimation of linear functionals of function-valued parameters in nonparametric instrumental variable problems. We rely on his results on consistency and the rate of convergence, which extend Chernozhukov et al. (2007) analysis to a nonparametric setting. Specifically, for each $s \in \mathcal{S}$, let

$$\mathcal{Q}_n(s) := \sum_{j=1}^{l} \Big(\frac{1}{n}\sum_{i=1}^{n} \varphi^{(j)}(X_i, s)\Big)_+^2. \tag{20}$$

This is a sample criterion function defined on $\mathcal{S}$. For instance, $\mathcal{Q}_n$ for Example 2.1 is given by

$$\mathcal{Q}_n(s) = \sum_{j=1}^{K} \Big(\frac{1}{n}\sum_{i=1}^{n}(Y_{L,i} - s(Z_i))1_{A_j}(Z_i)\Big)_+^2 + \sum_{j=1}^{K} \Big(\frac{1}{n}\sum_{i=1}^{n}(s(Z_i) - Y_{U,i})1_{A_j}(Z_i)\Big)_+^2.$$

Our first-stage set estimator is a level set of $\mathcal{Q}_n$ over a sieve $\mathcal{S}_n \subseteq \mathcal{S}$. Given a sequence of non-negative constants $\{a_n\}$ and $\{b_n\}$, define

$$\hat{\mathcal{S}}_n := \{s \in \mathcal{S}_n : \mathcal{Q}_n(s) \le b_n/a_n\}. \tag{21}$$

We add regularity conditions on φ, $\{\mathcal{S}_n\}$, and $\{(a_n, b_n)\}$ to ensure the Hausdorff consistency of $\hat{\mathcal{S}}_n$ and derive its convergence rate. The following two assumptions impose smoothness requirements on the map φ.

Assumption 3.8 For each j, there is a function $B_j : \mathcal{X} \to \mathbb{R}_+$ such that

$$|\varphi^{(j)}(x, s) - \varphi^{(j)}(x, s')| \le B_j(x)\rho(s, s'), \quad \forall s, s' \in \mathcal{S},$$

where $\rho(s, s') := \sup_{x \in \mathcal{S}} \max_{j=1,\dots,l} |s^{(j)}(x) - s'^{(j)}(x)|$.

For each $s \in \mathcal{S}$, let $\mathcal{I}(s) := \{j \in \{1, \dots, l\} : E[\varphi^{(j)}(X_i, s)] > 0\}$. $\mathcal{I}(s)$ is the set of indexes whose associated moments violate the inequality restrictions. For each j, let $\bar{\varphi}^{(j)} := E[\varphi^{(j)}(X_i, s)]$.

Assumption 3.9 (i) For each $s \in \mathcal{S}$ and j, $\bar{\varphi}^{(j)} : \mathcal{S} \to \mathbb{R}$ is continuously Fréchet differentiable with the Fréchet derivative $\dot{\varphi}_s^{(j)} : \mathcal{S} \to \mathbb{R}$, and for each $s \in \mathcal{S}$, the operator norm $\|\dot{\varphi}_s^{(j)}\|_{op}$ of $\dot{\varphi}_s^{(j)}$ is bounded away from 0 for some $j \in \{1, \dots, l\}$; (ii) for each $s \notin \mathcal{S}_0$, there exist $j \in \mathcal{I}(s)$ and $C_j > 0$ such that $E[\varphi^{(j)}(X_i, s)] \ge C_j \|s - s_0\|_W$ for some $s_0 \in \mathcal{S}_0$.

We also add regularity conditions on $\mathcal{S}_n$, which can be satisfied by commonly used sieves including polynomials, splines, wavelets, and certain artificial neural network sieves.

Assumption 3.10 (i) For each n, $\mathcal{S}_n \subseteq \mathcal{S}$, and both $\mathcal{S}_n$ and $\mathcal{S}$ are closed with respect to ρ; (ii) for every $s \in \mathcal{S}$, there is $\Pi_n s \in \mathcal{S}_n$ such that $\sup_{s\in\mathcal{S}} \|s - \Pi_n s\|_W = O(\delta_n)$ for some sequence $\{\delta_n\}$ of non-negative constants such that $\delta_n \to 0$.

Theorem 3.3 *Suppose Assumptions 2.1–2.3, 3.2 (i)–(iii), 3.3, 3.8, 3.9 (i), and 3.10 hold. Let $a_n = O(\max\{n^{-1}, \delta_n^2\}^{-1})$ and $b_n \to \infty$ with $b_n = o(a_n)$. Then*

$$d_{H,W}(\hat{\mathcal{S}}_n, \mathcal{S}_0) = o_p(1).$$

In addition, suppose that Assumption 3.9 (ii) holds. Then

$$d_{H,W}(\hat{\mathcal{S}}_n, \mathcal{S}_0) = O_p\big(\sqrt{b_n/a_n}\big).$$

Theorem 3.3 can be used to establish Assumptions 3.6 and 3.7, which are imposed in Theorems 3.1 and 3.2. These conditions are satisfied for Example 2.1 with a single regressor.

In what follows, for any two sequences of positive constants $\{c_n\}$, $\{d_n\}$, let $c_n \asymp d_n$ mean there exist constants $0 < C_1 < C_2 < \infty$ such that $C_1 \leq |c_n/d_n| \leq C_2$ for all n.

Corollary 3.1 *In Example 2.1, suppose that $\mathcal{Z}$ is a compact convex subset of the real line and $r_\theta(z) = \theta^{(1)} + \theta^{(2)}z$, where $\theta \in \Theta \subseteq \mathbb{R}^2$. Suppose that Θ is compact and convex. Suppose further that $\{(Y_{L,i}, Y_{U,i}, Z_i)\}_{i=1,\dots,n}$ is a random sample from P_0 and that $P_0(Z \in A_k) > 0$ for all k and $Var(Z) > 0$. Let $\mathcal{S} := \{s \in L^2_{\mathcal{Z},1} : \mathcal{Z} \to \mathbb{R} : \|s\|_\infty \leq M, |s(z) - s(z')| \leq M\|z - z'\|, \forall z, z' \in \mathcal{Z}\}$ for some $M > 0$. Let $\{r_q(\cdot)\}_{q=1}^{J_n}$ be splines of order two with J_n knots on $\mathcal{Z}$. Define $\mathcal{S}_n := \{s : s(z) = \sum_{q=1}^{J_n} \beta_q r_q(z)\}$ with $J_n \asymp n^{c_1}, c_1 > 1/3$. Let $\hat{\mathcal{S}}_n$ be defined as in (21) with $a_n \asymp n^{c_2}$, where $2/3 < c_2 < 1$ and $b_n \asymp \ln n$. Then: (i) $\hat{\mathcal{S}}_n$ is (Effros-) measurable;*
(ii) $d_{H,W}(\hat{\mathcal{S}}_n, \mathcal{S}_0) = o_p(1)$; (iii) $d_{H,W}(\hat{\mathcal{S}}_n, \mathcal{S}_0) = o_p(n^{-1/4})$.

Given these results, we further show that the estimator of the pseudo-true identified set is consistent and converges at a $n^{-1/2}$-rate.

Corollary 3.2 *Suppose that the conditions of Corollary 3.1 hold. Let Q be defined as in (15) with $W = 1$. Let Q_n be defined as in (16) and $\hat{\Theta}_n$ be defined as in (17) with $c_n = 0$ and $\hat{\mathcal{S}}_n$ as in Corollary 3.1. Then $d_H(\hat{\Theta}_n, \Theta_*) = O_p(n^{-1/2})$.*

4 Concluding Remarks

Moment inequalities are widely used to estimate discrete choice problems and structures that involve censored variables. In many empirical applications, potentially misspecified parametric models are used to estimate such structures. This chapter studies a novel estimation procedure that is robust to misspecification of moment inequalities. To overcome the challenge that the conventional identified set may be empty under misspecification, we defined a pseudo-true identified set as the least squares projection of the set of functions at which the moment inequalities are satisfied. This set is nonempty under mild assumptions. We also proposed a two-stage set estimator for estimating the pseudo-true identified set. Our estimator first estimates the identified set of function-valued parameters by a level-set estimator over a suitable sieve. The pseudo-true identified set can then be estimated by projecting the first-stage estimator to a finite-dimensional parameter space. We give conditions, under which the estimator is consistent for the pseudo-true identified set in the Hausdorff metric and converges at a rate $O_p(n^{-1/2})$. Developing inference procedures based on the proposed estimator would be an interesting future work. Another interesting extension would be to study the optimal choice of the weighting matrix. In this chapter, we maintained the assumption that W is fixed and does not depend on (θ, s). Given the form of the criterion function, the most natural choice of W would be the inverse matrix of the variance covariance matrix of $s(X_i) - r_\theta(X_i)$. This matrix is generally unknown but can be consistently estimated by its sample analog: $\hat{W}_n(\theta, s) := (\frac{1}{n}\sum_{i=1}^n (s(X_i) - r_\theta(X_i))(s(X_i) - r_\theta(X_i))')^{-1}$. Defining a sample criterion function using $\hat{W}_n(\theta, s)$ as a weighting matrix would lead to a three-step procedure. Such a procedure may result in more efficient estimation of Θ_*.[6] Yet, another interesting direction would be to develop a specification test for the moment inequality models based on the current framework. This direction would extend the results of Guggenberger et al. (2008), which studies a testing procedure that tests the nonemptiness of the identified set.

A Mathematical Proofs

A.1 Notation

Throughout the appendix, let $\|\cdot\|$ denote the usual Euclidean norm. For each $s, s' \in \mathcal{S}$, let $\rho(s, s') := \sup_{x\in\mathcal{S}} \max_{j=1,\dots,l} |s^{(j)}(x) - s'^{(j)}(x)|$. For each $a \times b$ matrix A, let $\|A\|_{op} := \min\{c : \|Av\| \le c\|v\|, v \in \mathbb{R}^b\}$ be the operator norm. For any symmetric matrix A, let $\xi(A)$ denote the smallest eigenvalue of A.

For a given pseudometric space (T, ρ), let $N(\epsilon, T, \rho)$ be the *covering number*, i.e., the minimal number of ϵ-balls needed to cover T. For each measurable function $f : \mathcal{X} \to \mathbb{R}$ and $1 \le p < \infty$, let $\|f\|_{L^p} := E[|f(X)|^p]^{1/p}$ provided that the integral exists. Similarly, let $\|f\|_\infty := \inf\{c : P(|f(X)| > c) = 0\}$. For a given function space $\mathcal{G}$ equipped with a norm $\|\cdot\|_{\mathcal{G}}$ and $l, u \in \mathcal{G}$, let $[l, u] := \{f \in \mathcal{G} : l \le f \le u\}$. For each $f \in \mathcal{G}$, let $B_{\epsilon,f} := \{[l, u] : l \le f \le u, \|l - u\|_{\mathcal{G}} < \epsilon\}$ be the ϵ-bracket of f. The *bracketing number* $N_{[\,]}(\epsilon, \mathcal{G}, \|\cdot\|_{\mathcal{G}})$ is the minimum number of ϵ-brackets needed to cover $\mathcal{G}$. An *envelope function* G of a function

[6] We are indebted to an anonymous referee for this point.

class $\mathcal{G}$ is a measurable function such that $g(x) \leq G(x)$ for all $g \in \mathcal{G}$. For each $\delta > 0$, the *bracketing integral* of $\mathcal{G}$ with an envelope function G is defined as $J_{[]}(\delta, \mathcal{G}, \|\cdot\|_{\mathcal{G}}) := \int_0^\delta \sqrt{1 + \ln N_{[]}(\epsilon\|G\|_{\mathcal{G}}, \mathcal{G}, \|\cdot\|_{\mathcal{G}})}d\epsilon$.

A.2 Projection

Proof of Proposition 2.1 Note that under the conditions of Example 2.1, Assumption 2.3 holds. This ensures $\mathcal{S}_0$ is nonempty. By Eq. (13), Θ_* is nonempty. Furthermore, let $\theta \in \Theta_I$, and for each $z \in \mathcal{Z}$, let $r_\theta(z) := z'\theta$. Note that $r_\theta \in \mathcal{S}_0$. Thus, (13) holds with $s = r_\theta$, which ensures the first claim.

For the second claim, note that the condition $E[Y_U|Z] = E[Y_L|Z] = Z'\theta_0$ a.s implies that any $\theta \in \Theta_I$ must satisfy

$$E[Z1\{Z \in A_j\}]'(\theta_0 - \theta) = 0, \quad j = 1, 2, \ldots, K. \tag{A.1}$$

By the rank condition on D, the unique solution to (A.1) is $\theta_0 - \theta = 0$. Thus, $\{\theta_0\} = \Theta_I$. Since $\{\theta_0\} \subseteq \Theta_*$ by the first claim, it suffices to show that θ_0 is the unique element of Θ_*. For this, note that under our assumptions, $\mathcal{S}_0 = \{s_0\}$ with $s_0(z) = z'\theta_0$. Thus, $\Theta_* = \{\theta_0\}$. This completes the proof. □

A.3 Consistency of the Parametric Part

For each $s \in \mathcal{S}$, let $\theta_*(s) := \arg\min_{\theta\in\Theta} Q(\theta, s)$ and $\hat{\theta}_n(s) := \arg\min_{\theta\in\Theta} Q_n(\theta, s)$.

Lemma A.1 *Suppose that Assumptions 3.4 and 3.2 (iv) hold. Then, (i) for each $x \in \mathcal{X}$ and any $s, s' \in \mathcal{S}$, there exists a function $C_1 : \mathcal{X} \to \mathbb{R}_+$ such that*

$$\left\| r_{\theta_*(s)}(x) - r_{\theta_*(s')}(x) \right\| \leq C_1(x)\rho(s, s'); \tag{A.2}$$

(ii) For each $x \in \mathcal{X}$, $j = 1, \ldots, L$, and any $s, s' \in \mathcal{S}$, there exists a function $C_2 : \mathcal{X} \to \mathbb{R}_+$ such that

$$\left\| \nabla_\theta^{(j)} r_{\theta_*(s)}(x) - \nabla_\theta^{(j)} r_{\theta_*(s')}(x) \right\| \leq C_2(x)\rho(s, s'). \tag{A.3}$$

Proof of Lemma A.1 Assumption 3.4 ensures that

$$\left\| r_{\theta_*(s)}(x) - r_{\theta_*(s')}(x) \right\| \leq L^{1/2} C(x) \left\| \theta_*(s) - \theta_*(s') \right\|. \tag{A.4}$$

Assumption 3.2 (iv) ensures that for each $s \in L^2_{\mathcal{S},L}$, $\theta_*(s) = \Pi_{\mathcal{R}_\Theta} s$ is uniquely determined, where $\Pi_{\mathcal{R}_\Theta}$ is the projection mapping from the Hilbert space $L^2_{\mathcal{S},L}$ to the closed convex subset $\mathcal{R}_\Theta$. Furthermore, Lemma 6.54 (d) in Aliprantis and Border (2006) and the fact that ρ is stronger than $\|\cdot\|_W$ imply

$$\left\|\theta_*(s) - \theta_*(s')\right\| \leq \left\|s - s'\right\|_W \leq c\rho(s, s'), \tag{A.5}$$

for some $c > 0$. Combining (A.4) and (A.5) ensures (i). Similarly, Assumption 3.4 ensures that for each $x \in \mathcal{X}$

$$\left\|\nabla_\theta^{(j)} r_{\theta_*(s)}(x) - \nabla_\theta^{(j)} r_{\theta_*(s')}(x)\right\| \leq J^{1/2} C(x) \left\|\theta_*(s) - \theta_*(s')\right\|. \tag{A.6}$$

Combining (A.5) and (A.6) ensures (ii). □

Proof of Theorem 3.1 Step 1: Let $s \in \mathcal{S}$ be given. For each $\theta \in \Theta$, let $Q_s(\theta) := Q(\theta, s)$ and $Q_{n,s}(\theta) := Q_n(\theta, s)$. By Assumption 3.2 (iv) and Theorem 6.53 in Aliprantis and Border (2006), Q_s is uniquely minimized at $\theta_*(s)$. By Assumption 3.2 (i), Θ is compact. By Assumption 3.2, $Q(\theta)$ is continuous. Furthermore, Assumption 3.4 ensures the applicability of the uniform law of large numbers. Thus, $\sup_{\theta\in\Theta} |Q_{n,s}(\theta) - Q_s(\theta)| = o_p(1)$. Hence, by Theorem 2.1 in Newey and McFadden (1994), $\hat{\theta}_n(s) - \theta_*(s) = o_p(1)$.

By Assumptions 3.2 (v), 3.4 (ii), and the fact that $\hat{\theta}_n(s)$ is consistent for $\theta_*(s)$, $\hat{\theta}_n(s)$ solves the first order condition:

$$\nabla_\theta Q_n(\theta, s) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta r_\theta(X_i)' W(s(X_i) - r_\theta(X_i)) = 0, \tag{A.7}$$

with probability approaching one. Expanding this condition at $\theta_*(s)$ using the mean-value theorem applied to each element of $\nabla_\theta Q_n(\theta, s)$ yields

$$\nabla_\theta^2 Q_n(\bar{\theta}_n(s), s)(\hat{\theta}_n(s) - \theta_*(s)) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta r_{\theta_*(s)}(X_i)' W(s(X_i) - r_{\theta_*(s)}(X_i)), \tag{A.8}$$

where $\bar{\theta}_n(s)$ lies on the line segment that connects $\hat{\theta}_n(s)$ and $\theta_*(s)$.[7] For each $s \in \mathcal{S}_0^{\bar{\eta}}$, let

$$\psi_s(x) := \nabla_\theta r_{\theta_*(s)}(x)' W(s(x) - r_{\theta_*(s)}(x)). \tag{A.9}$$

Below, we show that the function class $\Psi := \{f_s : f_s = \psi_s^{(j)}, s \in \mathcal{S}_0^{\bar{\eta}}, j = 1, 2, \ldots, J\}$ is a Glivenko–Cantelli class.

[7] Since the mean value theorem only applies element by element to the vector in (A.8), the mean value $\bar{\theta}_n$ differs across the elements. For notational simplicity, we use $\bar{\theta}_n$ in what follows, but the fact that they differ element to element should be understood implicitly. For the measurability of these mean values, see Jennrich (1969) for example.

By Assumption 3.4 (ii), Lemma A.1, the triangle inequality, and the Cauchy–Schwarz inequality, for any $s, s' \in \mathcal{S}$,

$$\begin{aligned}
|\psi_s^{(j)}(x) - \psi_{s'}^{(j)}(x)| &\le \left\|(\nabla_\theta^{(j)} r_{\theta_*(s)}(x) - \nabla_\theta^{(j)} r_{\theta_*(s')}(x))'W\right\| \\
&\quad \times \left\|s(x) - r_{\theta_*(s)}(x)\right\| + \left\|\nabla_\theta^{(j)} r_{\theta_*(s')}(x)'W\right\| \\
&\quad \times \left\|[s(x) - s'(x)] + [r_{\theta_*(s')}(x) - r_{\theta_*(s)}(x)]\right\| \\
&\le (C_2(x)\|W\|_{op}(M + R(x)) + (1 + C_1(x))\|W\|_{op} R(x)) \\
&\quad \times \sup_{x \in \mathcal{S}} \left\|s(x) - s'(x)\right\| \\
&\le F(x)\rho(s, s'),
\end{aligned} \tag{A.10}$$

where $F(x) := (C_2(x)\|W\|_{op}(M + R(x)) + (1 + C_1(x))\|W\|_{op}R(x)) \times \sqrt{L}$. For any $\epsilon > 0$, let $u := \epsilon/2\|F\|_{L^1}$. By, Theorem 2.7.11 in van der Vaart and Wellner (1996) and Assumption 3.2 (ii), we obtain

$$\begin{aligned}
N_{[]}(\epsilon, \Psi, \|\cdot\|_{L^1}) &= N_{[]}(2u\|F\|_{L^1}, \Psi, \|\cdot\|_{L^1}) \\
&\le N(u, \mathcal{S}_0^{\bar{\eta}}, \rho).
\end{aligned} \tag{A.11}$$

For each $j = 1, \dots, L$, let $\mathcal{S}_0^{\bar{\eta},(j)} := \{s^{(j)} : s \in \mathcal{S}_0^{\bar{\eta}}\}$. For each $j, g \in \mathcal{S}_0^{\bar{\eta},(j)}$, and $\epsilon > 0$, let $B_\epsilon^{(j)}(g) := \{f \in \mathcal{S}_0^{\bar{\eta},(j)} : \|f - g\|_\infty < \epsilon\}$. Similarly, for each $s \in \mathcal{S}_0^{\bar{\eta}}$, let $B_{u,\rho}(s) := \{f \in \mathcal{S}_0^{\bar{\eta},(j)} : \rho(f, s) < \epsilon\}$. As we will show below, $N_j := N(u, \mathcal{S}_0^{\bar{\eta},(j)}, \|\cdot\|_\infty)$ is finite for all j. Thus, for each j there exist $f_{1,j}, \dots, f_{N_j,j} \in \mathcal{S}_0^{\bar{\eta},(j)}$ such that $\mathcal{S}_0^{\bar{\eta},(j)} \subseteq \bigcup_{l=1}^{N_j} B_u^{(j)}(f_{l,j})$. We can then obtain a grid of distinct points $f_1, \dots, f_N \in \mathcal{S}_0^{\bar{\eta}}$ such that $f_i^{(j)} = f_{l,j}$ for some $1 \le l \le N_j$, where $N = \prod_{j=1}^L N_j$. Then, by the definition of ρ, $\mathcal{S}_0^{\bar{\eta}} \subseteq \bigcup_{i=1}^N B_{u,\rho}(f_i)$. Thus,

$$N\big(u, \mathcal{S}_0^{\bar{\eta}}, \rho\big) \le \prod_{j=1}^L N\big(u, \mathcal{S}_0^{\bar{\eta},(j)}, \|\cdot\|_\infty\big) \le N\big(u, \mathcal{C}_M^\gamma(\mathcal{X}), \|\cdot\|_\infty\big)^L < \infty, \tag{A.12}$$

where the last inequality follows from Assumption 3.2 (ii)–(iii) and Theorem 2.7.1 in van der Vaart and Wellner (1996). By Theorem 2.4.1 in van der Vaart and Wellner (1996), Ψ is a Glivenko–Cantelli class.

Note that, by Assumptions 3.2 (v) and 3.4, $\theta^*(s)$ solves the population analog of (A.7). Thus,

$$E[\nabla_\theta r_{\theta_*(s)}(X_i)'W(s(X_i) - r_{\theta_*(s)}(X_i))] = E[\psi_s(x)] = 0. \tag{A.13}$$

These results together with the strong law of large numbers whose applicability is ensured by Assumptions 3.3 and 3.4 (ii) imply

$$\sup_{s\in\mathcal{S}_0^{\bar\eta}}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_s^{(j)}(X_i)\right| = o_p(1),\quad j=1,\ldots,J. \tag{A.14}$$

Step 2: In this step, we show that the Hessian $\nabla_\theta^2 Q_n(\theta,s)$ is invertible with probability approaching 1 uniformly over $\mathcal{N}_{\bar\epsilon,\bar\eta}$. Let $\mathcal{H} := \{h_{\theta,s} : \mathcal{X}\to\mathbb{R} : h_{\theta,s}(x) = H_W^{(i,j)}(\theta,s,x) + 2\nabla_\theta r_\theta^{(i)}(x)'W\nabla_\theta r_\theta^{(j)}(x), 1\le i,j\le p, \theta\in\Theta, s\in\mathcal{S}_0^{\bar\eta}\}$. Note that $h_{\theta,s}$ takes the form:

$$\begin{aligned} h_{\theta,s}(x) &= 2\sum_{k=1}^{L}\sum_{h=1}^{L}\frac{\partial^2 r_\theta^{(h)}(x)}{\partial\theta_i\partial\theta_j}W^{(h,k)}\big(s^{(k)}(x)-r_\theta^{(k)}(x)\big)\\ &\quad+\sum_{k=1}^{L}\sum_{h=1}^{L}\frac{\partial r_\theta^{(h)}(x)}{\partial\theta_i}W^{(h,k)}\frac{\partial r_\theta^{(k)}(x)}{\partial\theta_j}\end{aligned}$$

for some $1\le i,j\le p, \theta\in\Theta$, and $s\in\mathcal{S}_0^{\bar\eta}$. Consider the function classes $\mathcal{F}_1 := \{D_\theta^\alpha r_\theta^{(k)} : \theta\in\Theta, |\alpha|\le 2, k=1,\ldots,L\}$ and $\mathcal{F}_2 := \{s^{(k)} : s\in\mathcal{S}_0^{\bar\eta}, k=1,\ldots,L\}$. Assumptions 3.2 (i), 3.4, and Theorem 2.7.11 in van der Vaart and Wellner (1996) ensure $N_{[]}(\epsilon,\mathcal{F}_1,\|\cdot\|_{L^2}) \le N(u,\Theta,\|\cdot\|) < \infty$ with $u := \epsilon/2\|C\|_{L^2}$. Assumption 3.2 (ii)–(iii) and Corollary 2.7.2 in van der Vaart and Wellner (1996) ensure $N_{[]}(\epsilon,\mathcal{F}_2,\|\cdot\|_{L^2}) \le N_{[]}(\epsilon,\mathcal{C}_M^\gamma(\mathcal{X}),\|\cdot\|_{L^2}) < \infty$. Since $\mathcal{H}$ can be obtained by combining functions in $\mathcal{F}_1$ and $\mathcal{F}_2$ by additions and pointwise multiplications, Theorem 6 in Andrews (1994) implies $N_{[]}(\epsilon,\mathcal{H},\|\cdot\|_{L^2}) < \infty$. This bracketing number is given in terms of the L^2-norm, but we can also obtain a bracketing number in terms of the L^1-norm. For this, let $h_1,\ldots,h_p$ be the centers of $\|\cdot\|_{L^2}$-balls that cover $\mathcal{H}$. Then, the brackets $[h_i-\epsilon, h_i+\epsilon], i=1,\ldots,p$ cover $\mathcal{H}$, and each bracket has length at most 2ϵ in $\|\cdot\|_{L^1}$. Thus, $N_{[]}(\epsilon,\mathcal{H},\|\cdot\|_{L^1}) < \infty$. By Theorem 2.7.1 in van der Vaart and Wellner (1996), $\mathcal{H}$ is a Glivenko–Cantelli class. Hence, uniformly over $\Theta\times\mathcal{S}_0^{\bar\eta}$,

$$\begin{aligned}\nabla_\theta^2 Q_n(\theta,s) &= \frac{1}{n}\sum_{i=1}^{n}H_W(\theta,s,X_i) + 2\nabla_\theta r_\theta(X_i)'W\nabla_\theta r_\theta(X_i)\\ &\xrightarrow{p} E[H_W(\theta,s,X_i)+2\nabla_\theta r_\theta(X_i)'W\nabla_\theta r_\theta(X_i)].\end{aligned} \tag{A.15}$$

Note that $d_{H,W}(\hat{\mathcal{S}}_n,\mathcal{S}_0) = o_p(1)$ by Assumption 3.6. Thus, $(\bar\theta_n(s),s)\in\mathcal{N}_{\epsilon,\bar\eta}$ with probability approaching one. By Assumption 3.5 and (A.15), there exists $\delta>0$ such that $\nabla_\theta^2 Q_n(\bar\theta_n(s),s)$'s smallest eigenvalue is above δ uniformly over $\mathcal{N}_{\bar\epsilon,\bar\eta}$. Thus, the Hessian $\nabla_\theta^2 Q_n(\bar\theta_n(s),s)$ in (A.8) is invertible with probability approaching 1.

Step 3: Steps 1–2 imply that, uniformly over $\mathcal{S}_0^{\bar{\eta}}$,

$$\begin{aligned}\|\theta_*(s) - \hat{\theta}_n(s')\| &= \|\theta_*(s) - \theta_*(s') + \theta_*(s') - \hat{\theta}_n(s')\| \\ &\leq \|\theta_*(s) - \theta_*(s')\| + 2\delta^{-1} \sup_{s \in \mathcal{S}_0^{\bar{\eta}}} \left\| \frac{1}{n} \sum_{i=1}^{n} \psi_s(X_i) \right\| \\ &\leq \|s - s'\|_W + o_p(1), \qquad \text{(A.16)}\end{aligned}$$

where we used the fact that $\|\theta_*(s) - \theta_*(s')\| \leq \|s - s'\|_W$ by Lemma 6.54 (d) in Aliprantis and Border (2006).

Step 4: Finally, note that by Step 3,

$$\begin{aligned}\vec{d}_H(\Theta_*, \hat{\Theta}_n) &= \sup_{\theta \in \Theta_*} \inf_{\theta' \in \hat{\Theta}_n} \|\theta - \theta'\| = \sup_{s \in \mathcal{S}_0} \inf_{s' \in \hat{\mathcal{S}}_n} \|\theta_*(s) - \hat{\theta}_n(s')\| \\ &\leq \sup_{s \in \mathcal{S}_0} \inf_{s' \in \hat{\mathcal{S}}_n} \|s - s'\|_W + o_p(1) \qquad \text{(A.17)} \\ \vec{d}_H(\hat{\Theta}_n, \Theta_*) &= \sup_{\theta' \in \hat{\Theta}_n} \inf_{\theta \in \Theta_*} \|\theta - \theta'\| = \sup_{s' \in \hat{\mathcal{S}}_n} \inf_{s \in \mathcal{S}_0} \|\theta_*(s) - \hat{\theta}_n(s')\| \\ &\leq \sup_{s' \in \hat{\mathcal{S}}_n} \inf_{s \in \mathcal{S}_0} \|s - s'\|_W + o_p(1). \qquad \text{(A.18)}\end{aligned}$$

Equation (18) and Assumption 3.6 then ensure the desired result. □

A.4 Convergence Rate

The following lemma controls the rate at which $\hat{\Theta}_n$ covers Θ_*. Given a sequence $\{\eta_n\}$ such that $\eta_n \to 0$, we let $V^{\delta_{1n}}(s) := \{\theta' : \|\theta' - \theta(s)\| \leq e_n, \quad e_n = O_p(\eta_n)\}$ and let $\mathcal{N}_{\eta_n,0} := \{(\theta, s) : \theta \in V^{\eta_n}(s), s \in \mathcal{S}_0\}$.

Lemma A.2 *Suppose Assumptions 2.1–2.3, 3.1–3.2, and 3.6 hold. Let $\{\delta_{1n}\}$ and $\{\epsilon_n\}$ be sequences of non-negative numbers converging to 0 as $n \to \infty$. Let $G : \Theta \times \mathcal{S} \to \mathbb{R}_+$ be a function such that G is jointly measurable and lower semicontinuous. For each n, let $G_n : \Omega \times \Theta \times \mathcal{S} \to \mathbb{R}$ be a function such that for each $\omega \in \Omega$, $G_n(\omega, \cdot, \cdot)$ is jointly measurable and lower semicontinuous, and for each $(\theta, s) \in \Theta \times \mathcal{S}$, $G_n(\cdot, \theta, s)$ is measurable. Let $\Theta_* := \{G(\theta, s) = 0, s \in \mathcal{S}_0\}$ and $\hat{\Theta}_n := \{\theta \in \Theta : G_n(\theta, s) \leq \inf_{\theta \in \Theta} G_n(\theta, s) + c_n, s \in \hat{\mathcal{S}}_n\}$. Suppose that $d_H(\hat{\Theta}_n, \Theta_*) = O_p(\delta_{1n})$. Suppose further that there exists a positive constant κ and a neighborhood $V(s)$ of $\theta_*(s)$ such that*

$$G(\theta, s) \geq \kappa \|\theta - \theta_*(s)\|^2 \qquad \text{(A.19)}$$

for all $\theta \in V(s)$, $s \in \mathcal{S}_0$. Suppose that uniformly over $\mathcal{N}_{\delta_{1n},0}$,

$$G_n(\theta, s) = G(\theta, s) + O_p(\|\theta - \theta_*(s)\|/\sqrt{n}) + o_p(\|\theta - \theta_*(s)\|^2) + O_p(\epsilon_n). \qquad \text{(A.20)}$$

Then

$$\vec{d}_H(\Theta_*, \hat{\Theta}_n) = O_p(\max\{c_n^{1/2}, \epsilon_n^{1/2}, 1/\sqrt{n}\}).$$

Proof of Lemma A.2 The proof of this Lemma is similar to Theorem 1 in Sherman (1993). By (A.19), (A.20), and the Hausdorff consistency of $\hat{\Theta}_n$, it follows that, uniformly over $\mathcal{N}_{\delta_{1n},0}$,

$$c_n \geq \kappa\|\theta - \theta_*(s)\|^2 + O_p(\|\theta - \theta_*(s)\|/\sqrt{n}) + o_p(\|\theta - \theta_*(s)\|^2) + O_p(\epsilon_n), \quad \text{(A.21)}$$

with probability approaching 1. As in Theorem 1 in Sherman (1993), write $K_n\|\theta - \theta(s)\|$ for the $O_p(\|\theta - \theta_*(s)\|/\sqrt{n})$ term, where $K_n = O_p(1/\sqrt{n})$ and also note that $o_p(\|\theta - \theta_*(s)\|^2)$ is bounded from below by $-\frac{\kappa}{2}\|\theta - \theta^*(s)\|^2$ with probability approaching 1. Thus, we obtain

$$\frac{\kappa}{2}\|\theta - \theta_*(s)\|^2 + K_n\|\theta - \theta_*(s)\| \leq c_n + O_p(\epsilon_n). \quad \text{(A.22)}$$

Completing the square, we obtain

$$\frac{1}{2}\kappa(\|\theta - \theta_*(s)\| - K_n/\kappa)^2 \leq c_n + O_p(\epsilon_n) + \frac{1}{2}K_n^2/\kappa = c_n + O_p(\epsilon_n) + O_p(1/n). \quad \text{(A.23)}$$

Taking square roots gives

$$\begin{aligned}\|\theta - \theta_*(s)\| &\leq (2/\kappa)^{1/2}c_n^{1/2} + K_n/\kappa + O_p(\epsilon_n^{1/2}) + O_p(1/\sqrt{n}) && \text{(A.24)}\\ &= O_p(c_n^{1/2}) + O_p(\epsilon_n^{1/2}) + O_p(1/\sqrt{n}). && \text{(A.25)}\end{aligned}$$

Thus,

$$\begin{aligned}\vec{d}_H(\Theta_*, \hat{\Theta}_n) &= \sup_{s\in\mathcal{S}_0}\inf_{\theta\in\hat{\Theta}_n}\|\theta - \theta_*(s)\| && \text{(A.26)}\\ &\leq \sup_{s\in\mathcal{S}_0}\inf_{\theta\in V^{\delta_{1n}}(s)}\|\theta - \theta_*(s)\|\\ &\leq O_p(c_n^{1/2}) + O_p(\epsilon_n^{1/2}) + O_p(1/\sqrt{n}). && \text{(A.27)}\end{aligned}$$

This completes the proof. □

The following lemma controls the rate at which $\hat{\Theta}_n$ is contracted into a neighborhood of Θ_*. Given $s \in \mathcal{S}$ and a sequence $\{\delta_n\}$ such that $\delta_n \to 0$, let $U^{\delta_n}(s) := \{\theta \in \Theta : \|\theta - \theta_*(s)\| \geq \delta_n\}$.

Lemma A.3 *Suppose Assumptions 2.1–2.3, 3.1–3.2, and 3.6 hold. Let* G_n *be defined as in Lemma A.2. Suppose that there exist positive constants* (k, κ_2) *and a sequence* $\{\delta_{1n}\}$ *such that*

$$G_n(\theta, s) \geq \kappa_2\|\theta - \theta_*(s)\|^2 \quad \text{(A.28)}$$

with probability approaching 1 for all $\theta \in U^{\delta_n}(s)$ with $\delta_n := (k\delta_{1n}/\sqrt{n})^{1/2}$ and $s \in \mathcal{S}_0^{\bar{\eta}}$. Then,

$$\vec{d}_H(\hat{\Theta}_n, \Theta_*) = O_p(\delta_{1n}^{1/2}/n^{1/4}) + O_p(c_n^{1/2}).$$

Proof of Lemma A.3 Note first that $\hat{\mathcal{S}}_n$ is in $\mathcal{S}_0^{\bar{\eta}}$ with probability approaching 1 by Assumption 3.6. Let $\tilde{c}_n := \sqrt{n}c_n$ and $\bar{c}_n := \max\{\kappa_2 k\delta_{1n}, \tilde{c}_n\}$. Let $\epsilon_n := (\bar{c}_n/\kappa_2\sqrt{n})^{1/2}$. Then, uniformly over $\mathcal{S}_0^{\bar{\eta}}$,

$$\inf_{\Theta\cap U^{\epsilon_n}(s)} \sqrt{n}G_n(\theta, s) \geq \kappa_2\sqrt{n}\epsilon_n^2 \geq \bar{c}_n. \tag{A.29}$$

Since $\sqrt{n}G_n(\hat{\theta}_n(s), s) \leq \tilde{c}_n$ for all $s \in \hat{\mathcal{S}}_n$, the results above ensure

$$\begin{aligned}\vec{d}_H(\hat{\Theta}_n, \Theta_*) &= \sup_{s\in\hat{\mathcal{S}}_n} \inf_{\theta\in\Theta_*} \|\hat{\theta}_n(s) - \theta\| \\ &\leq \sup_{s\in\hat{\mathcal{S}}_n} \|\hat{\theta}_n(s) - \theta_*(s)\| \leq \epsilon_n = O_p(\delta_{1n}^{1/2}/n^{1/4}) + O_p(\tilde{c}_n^{1/2}/n^{1/4}).\end{aligned}$$

This ensures the claim of the Lemma. □

Proof of Theorem 3.2 We first show (A.19) holds with $G(\theta, s) = Q(\theta, s)$. For this, we use the second-order Taylor expansion of $Q(\theta, s)$. For $\theta \in V^{\delta_{1n}}(s)$, it holds by Assumptions 3.2 (v) and 3.4 that

$$\begin{aligned}Q(\theta, s) = Q(\theta_*(s), s) &+ \nabla_\theta Q(\theta_*(s), s)'(\theta - \theta_*(s)) \\ &+ \frac{1}{2}(\theta - \theta_*(s))'\nabla_\theta^2 Q(\bar{\theta}(s), s)(\theta - \theta_*(s)),\end{aligned} \tag{A.30}$$

where $\bar{\theta}(s)$ is on the line segment that connects θ and $\theta_*(s)$. By (15), $Q(\theta_*(s), s) = 0$, and by the first order condition of the optimality, $\nabla_\theta Q(\theta_*(s), s) = 0$. Thus, it follows that

$$Q(\theta, s) = \frac{1}{2}(\theta - \theta_*(s))'\nabla_\theta^2 Q(\bar{\theta}(s), s)(\theta - \theta_*(s)) \geq \kappa\|\theta - \theta_*(s)\|^2, \tag{A.31}$$

where $\kappa := \inf_{\theta\in\Theta, s\in\mathcal{S}_0} \xi(\nabla_\theta^2 Q(\theta, s))/2$, and $\kappa > 0$ by Assumption 3.5.

We next show that (A.20) holds for

$$\begin{aligned}G_n(\theta, s) = \frac{1}{n}\sum_{i=1}^n &(s(X_i) - r_\theta(X_i))'W(s(X_i) - r_\theta(X_i)) \\ &- \frac{1}{n}\sum_{i=1}^n (s(X_i) - r_{\theta_*(s)}(X_i))'W(s(X_i) - r_{\theta_*(s)}(X_i)).\end{aligned} \tag{A.32}$$

In what follows, let $\hat{E}_n$ denote the expectation with respect to the empirical distribution. Using the Taylor expansion of G_n and G with respect to θ at $\theta_*(s)$, we may write

$$G_n(\theta, s) - G(\theta, s) = S_{1,n}(\theta, s) + S_{2,n}(\theta, s), \tag{A.33}$$

where

$$\begin{aligned} S_{1n}(\theta, s) &:= -2(\theta - \theta_*(s))'(\hat{E}_n - E)[\nabla_\theta r_{\theta_*(s)}(x)' W(s(x) - r_{\theta_*}(x))] \\ &\quad + o_p(\|\theta - \theta_*(s)\|^2) \qquad (\text{A.34}) \\ S_{2n}(\theta, s) &:= (\theta - \theta_*(s))'(\hat{E}_n - E)[\nabla_\theta r_{\theta_*(s)}(x)' W \nabla_\theta r_{\theta_*(s)}(x)](\theta - \theta_*(s)). \qquad (\text{A.35}) \end{aligned}$$

Thus, for (A.20) to hold, it suffices to show that $S_{1n}(\theta, s) = O_p(\|\theta - \theta_*(s)\|/\sqrt{n}) + o_p(\|\theta - \theta_*(s)\|^2)$ and $S_{2n}(\theta, s) = O_p(\epsilon_n)$ for some $\epsilon_n \to 0$. For S_{1n}, note that our assumptions suffice for the conditions of Lemma A.4. Thus, Φ is a P_0-Donsker class. This ensures $S_{1n}(\theta, s) = O_p(\|\theta - \theta_*(s)\|/\sqrt{n}) + o_p(\|\theta - \theta_*(s)\|^2)$. We now consider S_{2n}. For each $s \in \mathcal{S}_0$ and $x \in \mathcal{X}$, let $\phi_s(x) := \nabla_\theta r_{\theta_*(s)}(x)' W \nabla_\theta r_{\theta_*(s)}(x)$. Note that

$$\begin{aligned} E\left[\sup_{(\theta,s)\in\mathcal{N}_{\delta_{1n},0}} |S_{2n}(\theta, s)|\right] &\leq \delta_{1n}^2 n^{-1/2} E\left[\sup_{s\in\mathcal{S}_0} |\mathbb{G}_n \phi_s|\right] \\ &\leq n^{-1/2}\delta_{1n}^2 C J_{[]}(1, \mathcal{S}_0, \|\cdot\|_{L^2}) \left\|\sup_{s\in\mathcal{S}_0} |\phi_s|\right\|_{L^2}, \end{aligned} \tag{A.36}$$

where the last inequality follows from Lemma B.1 of Ichimura and Lee (2010). Now, Markov's inequality, Lemma A.4, and Assumption 3.4 (ii) ensure that $S_{2n} = O_p(\epsilon_n)$, where $\epsilon_n = n^{-1/2}\delta_{1n}^2$.

We further set $c_n = 0$. Note that the estimator defined in (17) with $c_n = 0$ equals the set estimator $\hat{\Theta}_n = \{\theta : G_n(\theta, s) \leq \inf_{\theta\in\Theta} G_n(\theta, s)\}$. By Assumption 3.7 and Step 4 of the proof of Theorem 3.1, we may take $\delta_{1n} = O_p(n^{-1/4})$ as an initial rate. Lemma A.2 then implies that $\vec{d}_H(\Theta_*, \hat{\Theta}_n) = O_p(\epsilon_n^{1/2})$, where $\epsilon_n = O_p(n^{-1/2}\delta_{1n}^2) = O_p(n^{-1})$. Thus, $\vec{d}_H(\Theta_*, \hat{\Theta}_n) = O_p(n^{-1/2})$.

Now we consider $\vec{d}_H(\hat{\Theta}_n, \Theta_*)$. We show that (A.28) holds for G_n. For each θ and s, let $L_n(\theta, s) := \frac{1}{n}\sum_{i=1}^n (s(X_i) - r_\theta(X_i))' W(s(X_i) - r_\theta(X_i))$. Let $s \in \mathcal{S}_0^{\bar{\eta}}$ and $\theta \in U^{\delta_{1n}}(s)$. A second-order Taylor expansion of $G_n(\theta, s) = L_n(\theta, s) - L_n(\theta_*(s), s)$ with respect to θ at $\theta_*(s)$ gives

$$\begin{aligned} G_n(\theta, s) &= \nabla_\theta L_n(\theta_*(s), s)'(\theta - \theta_*(s)) + \frac{1}{2}(\theta - \theta_*(s))' \nabla_\theta^2 L_n(\bar{\theta}_n(s), s)(\theta - \theta_*(s)) \\ &= o_p(1) + \frac{1}{2}(\theta - \theta_*(s))' \nabla_\theta^2 L_n(\bar{\theta}_n(s), s)(\theta - \theta_*(s)) \\ &\geq \kappa_2 \|\theta - \theta_*(s)\|^2, \end{aligned} \tag{A.37}$$

with probability approaching 1 for some $\kappa_2 > 0$, where $\bar{\theta}_n(s)$ is a point on the line segment that connects θ and $\theta_*(s)$. The last inequality follows from Step 3 of the proof of Theorem 3.1 and Assumption 3.5.

Set $\tilde{c}_n = 0$. Then, Lemma A.3 implies $\vec{d}_H(\hat{\Theta}_n, \Theta_*) = O_p(\delta_{1n}^{1/2}/n^{1/4})$. Setting $\delta_{1n} = O_p(n^{-1/4})$ refines this rate to $O_p(n^{-3/8})$. Repeated applications of Lemma A.3 then implies $\vec{d}_H(\hat{\Theta}_n, \Theta_*) = O_p(n^{-1/2})$. As both of the directed Hausdorff distances converge to 0 at the stochastic order of $n^{-1/2}$, the claim of the theorem follows. □

Lemma A.4 *Suppose Assumptions 3.2 and 3.4 hold. Then Φ is a P_0-Donsker class.*

Proof of Lemma A.4 The proof of Theorem 3.1 shows that each $f_s \in \Phi$ is Lipschitz in s. For any $\epsilon > 0$, Assumption 3.2 (ii)–(iii), Theorems 2.7.11 and 2.7.2 in van der Vaart and Wellner (1996), and (A.12) imply

$$\ln N_{[]}(\epsilon\|F\|_{L^2}, \Psi, \|\cdot\|_{L^2}) \leq \ln N(\epsilon/2, \mathcal{S}_0^{\delta_2}, \rho)^L \leq C(1/\epsilon)^{k/\gamma}, \tag{A.38}$$

where C is a constant that depends only on k, γ, L, and $\mathrm{diam}(\mathcal{X})$. Thus, for any $\delta > 0$,

$$J_{[]}(\delta, \Phi, \|\cdot\|_{L^2}) \leq \int_0^\delta \sqrt{1 + C(1/\epsilon)^{k/\gamma}}d\epsilon < \infty. \tag{A.39}$$

Example 2.14.4 in van der Vaart and Wellner (1996) ensures that Ψ is P_0-Donsker. □

A.5 First Stage Estimation

In the following, we work with the following population criterion function. For each $s \in \mathcal{S}$, let $\mathcal{Q}$ be defined by

$$\mathcal{Q}(s) := \sum_{j=1}^{l} E[\varphi^{(j)}(X_i, s)]_+^2. \tag{A.40}$$

Lemma A.5 *Suppose that Assumption 3.9 (i) holds. Let the criterion function be given as in (A.40). Then, there exists a positive constant C_2 such that*

$$\mathcal{Q}(s) \leq \inf_{s_0 \in \mathcal{S}_0} C_2\|s - s_0\|_W^2.$$

Proof of Lemma A.5 Let $s \in \mathcal{S}$ be arbitrary. For any $s_0 \in \mathcal{S}$, $E[\varphi^{(j)}(X, s_0)] \leq 0$ for $j = 1, \ldots, l$. Let V be an open set that contains s and s_0. Assumption 3.9 (i) and Theorem 1.7 in Lindenstrauss et al. (2007), it holds that

$$\begin{aligned}\mathcal{Q}(s) &\le \sum_{j=1}^{l}\Big(E[\varphi^{(j)}(X_i,s)] - E[\varphi^{(j)}(X_i,s_0)]\Big)_+^2\\ &\le \left(\sum_{j=1}^{l}\|\sup_{g\in\tilde{V}_j}\dot{\varphi}_g^{(j)}\|_{op}^2\right)\|s-s_0\|_W^2, \end{aligned}\tag{A.41}$$

where $\tilde{V}_j := \{g \in V : \dot{\varphi}_g^{(j)}$ exists$\}$. Let $C_2 := \sum_{j=1}^{l}\|\sup_{g\in\mathcal{S}}\dot{\varphi}_g^{(j)}\|_{op}^2$. It holds that $0 < C_2 < \infty$ by the hypothesis. We thus obtain

$$\mathcal{Q}(s) \le C_2\|s-s_0\|_W^2 \tag{A.42}$$

for all $s_0 \in \mathcal{S}_0$. Note that $s_0 \mapsto \|s-s_0\|_W$ is continuous and $\mathcal{S}_0$ is compact by Assumption 3.2 (ii)–(iii) and Assumption 3.10 (i). Taking infimum over $\mathcal{S}_0$ then ensures the desired result. □

Lemma A.6 *Suppose Assumption 3.9 (ii) holds. Let the criterion function be given as in (A.40). Then there exists a positive constant C such that*

$$\mathcal{Q}(s) \ge \inf_{s_0\in\mathcal{S}_0} C_3\|s-s_0\|_W^2.$$

Proof of Lemma A.6 If $s \in \mathcal{S}_0$, the conclusion is immediate. Suppose that $s \notin \mathcal{S}_0$. By Assumption 3.9 (ii), there exists $s_0 \in \mathcal{S}_0$

$$\mathcal{Q}(s) = \sum_{j\in\mathcal{I}(s)}(E[\varphi^{(j)}(X_i,s)])^2 \ge C_j\|s-s_0\|_W^2. \tag{A.43}$$

Let $C_3 := C_j$. Thus, the claim of the lemma follows. □

In the following, let $\mathcal{G} := \{g : g(x) = \varphi_s^{(j)}(x), s\in\mathcal{S}, j = 1,\dots,l\}$.

Lemma A.7 *Suppose Assumptions 3.2, 3.4 , and 3.8 hold. Then $\mathcal{G}$ is a P_0-Donsker class.*

Proof of Lemma A.7 By Assumption 3.8, $\varphi_s^{(j)}$ is Lipschitz in s. The rest of the proof is the same as that of Lemma A.4. □

Proof of Theorem 3.3 We establish the claims of the theorem by applying Theorem B.1 in Santos (2011). Note first that Assumption 3.2 (ii)–(iii) and Assumption 3.10 (i) ensure that $\mathcal{S}$ is compact. This ensures condition (i) of Theorem B.1 in Santos (2011). Condition (ii) of Theorem B.1 in Santos (2011) is ensured by Assumption 3.10. Lemma A.7 ensures that uniformly over Θ_n

$$\mathcal{Q}_n(s) = \mathcal{Q}(s) + O_p(n^{-1}). \tag{A.44}$$

Thus, condition (iii) of Theorem B.1 in Santos (2011) hold with $C_1 = 1$ and $c_{2n} = n^{-1}$. Lemma A.5 ensures that $\mathcal{Q}(s) \leq \inf_{s_0 \in \mathcal{S}_0} C_2 \|s - s_0\|_W^2$ for some $C_2 > 0$. Thus, condition (iv) of Theorem B.1 in Santos (2011) hold with $\kappa_1 = 2$. Now, the first claim of Theorem B.1. in Santos (2011) establishes

$$d_{H,W}(\hat{\mathcal{S}}_n, \mathcal{S}_0) = o_p(1). \tag{A.45}$$

Furthermore, Lemma A.6 ensures $\mathcal{Q}(s) \geq \inf_{s_0 \in \mathcal{S}_0} C_3 \|s - s_0\|^2$ for some $C_3 > 0$. This ensures condition (v) of Theorem B.1 in Santos (2011) with $\kappa_2 = 2$. Now, the second claim of Theorem B.1. in Santos (2011) ensures

$$d_{H,W}(\hat{\mathcal{S}}_n, \mathcal{S}_0) = O_p(\max\{(b_n/a_n)^{1/2}, \delta_n\}). \tag{A.46}$$

Since $(b_n/a_n)^{1/2}/\delta_n \to \infty$, the claim of the theorem follows. □

Proof of Corollary 3.1 In what follows, we explicitly show $\mathcal{Q}_n$'s dependence on $\omega \in \Omega$. Let $\mathcal{Q}_n : \Omega \times \mathcal{S} \to \mathbb{R}$ be defined by $\mathcal{Q}_n(\omega, s) = \sum_{j=1}^{l} (\frac{1}{n} \sum_{i=1}^{n} \varphi(X_i(\omega), s))_+^2$. By Assumption 2.3, φ is continuous in s for every x and measurable for every s. Also note that X_i is measurable for every i. Thus, by Lemma 4.51 in Aliprantis and Border (2006), $\mathcal{Q}_n$ is jointly measurable in (ω, s) and lower semicontinuous in s for every ω. Note that $\mathcal{S}$ is compact by Assumptions 3.2 (ii)–(iii) and 3.10 (i), which implies $\mathcal{S}$ is locally compact. Since $\mathcal{S}$ is a metric space, it is a Hausdorff space. Thus, by Proposition 5.3.6 in Molchanov (2005), $\mathcal{Q}_n$ is a normal integrand defined on a locally compact Hausdorff space. Proposition 5.3.10 in Molchanov (2005) then ensures the first claim.

Now we show the second claim using Theorem 3.3 (i). Assumptions 2.1–2.3 hold with φ defined in (5). Assumption 3.2 holds by our hypothesis with $\gamma = 1$. Assumption 3.3 is also satisfied by the hypothesis. Note that for each j, $\varphi^{(j)}(x, s) = (y_L - s(z))1_{A_k}(z)$ or $\varphi^{(j)}(x, s) = (s(z) - y_U)1_{A_k}(z)$ for some $k \in \{1, \ldots, K\}$. Without loss of generality, let j be an index for which $\varphi^{(j)}(x, s) = (y_L - s(z))1_{A_k}(z)$ for some Borel set A_k. For any $s, s' \in \mathcal{S}$,

$$|\varphi^{(j)}(x, s) - \varphi^{(j)}(x, s')| = |(s'(z) - s(z))1_{A_k}(z)| \leq \rho(s, s'). \tag{A.47}$$

It is straightforward to show the same result for other indexes. Thus, Assumption 3.8 is satisfied.

Now for j such that $\varphi^{(j)}(x, s) = (y_L - s(z))1_{A_k}(z)$, note that

$$|\bar{\varphi}^{(j)}(s + h) - \bar{\varphi}^{(j)}(s) - E[h(Z)(-1_{A_k}(Z))]| = 0. \tag{A.48}$$

Thus, the Fréchet derivative is given by $\dot{\varphi}_s^{(j)}(h) = E[h(Z)(-1_{A_k}(Z))]$. By Proposition 6.13 in Folland (1999), the norm of the operator is given by $\|\dot{\varphi}_s^{(j)}\|_{op} = E[|-1_{A_k}(Z)|^2]^{1/2} = P_0(Z \in A_k) > 0$, which ensures the boundedness (continuity) of the operator. It is straightforward to show the same result for other indexes. Hence, Assumption 3.9 (i) is satisfied. By construction, Assumption 3.10 (i) is satisfied, and Assumption 3.10 (ii) holds with $\delta_n \asymp J_n^{-1}$ (See Chen 2007). These ensure the conditions of Theorem 3.3 (i). Thus, the second claim follows.

For the third claim, let $s \in \mathcal{S} \setminus \mathcal{S}_0$. Then, there exists j such that $E[\varphi^{(j)}(X_i, s)] > 0$. Without loss of generality, suppose that $E[\varphi^{(j)}(X_i, s)] = E[(Y_{L,i} - s(Z_i))1_{A_k}(Z_i)] \geq \delta > 0$. Let $s_0 \in \mathcal{S}_0$ be such that

$$E[(Y_{L,i} - s_0(Z_i))1_{A_k}(Z_i)] = 0. \tag{A.49}$$

Such s_0 always exists by the intermediate value theorem. Then, for j with which $\varphi^{(j)}(x, s) = (y_L - s(z))1_{A_k}(z)$, it follows that

$$\begin{aligned} E[\varphi^{(j)}(X_i, s)] &= E[(Y_{L,i} - s(Z_i))1_{A_k}(Z_i)] - E[(Y_{L,i} - s_0(Z_i))1_{A_k}(Z_i)] \\ &= E[(s_0(Z_i) - s(Z_i))1_{A_k}(Z_i)] > 0 \end{aligned} \tag{A.50}$$

Thus, we have

$$E[\varphi^{(j)}(X_i, s)] \geq C\|s_0 - s\|_W, \tag{A.51}$$

where $C := \inf_{q \in E} E[q(Z_i)1_{A_k}(Z_i)]$ and $E := \{q \in \mathcal{S} : \|q\|_W = 1, E[q(Z_i)1_{A_k}(Z_i)] > 0\}$. Since C is the minimum value of a linear function over a convex set, it is finite. Furthermore, by the construction of E, it holds that $C > 0$. Thus, Assumption 3.9 (ii) holds. Thus, by Theorem 3.3 (ii), the third claim follows. □

Proof of Corollary 3.2 We show the claim of the corollary using Theorem 3.2. Note that we have shown, in the proof of Corollary 3.1, that Assumptions 2.1–2.3, 3.2 (i)–(iii), and 3.3 hold. Thus, to apply Theorem 3.2, it remains to show Assumptions 2.4, 3.2 (iv), and 3.4–3.7.

Assumption 2.4 is satisfied by the parameterization $r_\theta(z) = \theta^{(1)} + \theta^{(2)}z$. For Assumption 3.2 (iv), note that $\mathcal{R}_\Theta$ is given by

$$\mathcal{R}_\Theta = \left\{r_\theta : r_\theta = \theta^{(1)} + \theta^{(2)}z, \quad \theta \in \Theta\right\}.$$

Since Θ is convex, for any $\lambda \in [0, 1]$, it holds that $\lambda r_\theta + (1 - \lambda)r_{\theta'} = r_{\lambda\theta+(1-\lambda)\theta'} \in \mathcal{R}_\Theta$. Thus, Assumption 3.2 (iv) is satisfied. For Assumption 3.4, note first that r_θ is twice continuously differentiable on the interior of Θ. Because r_θ is linear, $\max_{|\alpha|\leq 2} |D^\alpha_\theta r_\theta(z) - D^\alpha_\theta r_{\theta'}(z)| = (1 + z^2)^{1/2}\|\theta - \theta'\|$ by the Cauchy–Schwarz inequality. By the compactness of $\mathcal{Z}$, $C(z) := (1+z^2)^{1/2}$ is bounded. Thus, Assumption 3.4 (i) is satisfied. Similarly, $\max_{|\alpha|\leq 2} \sup_{\theta\in\Theta} |D^\alpha_\theta r_\theta| \leq \max\{1, |z|, C(1 + z^2)^{1/2}\} =: R(z)$, where $C := \sup_{\theta\in\Theta} \|\theta\|$. By the compactness of $\mathcal{Z}$ and Θ, R is bounded. Thus, Assumption 3.4 (ii) is satisfied. Note that the Hessian of $Q(\theta, s)$ with respect to θ is given by $2E[(1, z)(1, z)']$, which does not depend on θ nor s and is positive definite by the assumption that $Var(Z) > 0$. Thus, Assumption 3.5 is satisfied. Assumptions 3.6 and 3.7 are ensured by Corollary 3.1. Now the conditions of Theorem 3.2 are satisfied. Thus, the claim of the Corollary follows. □

References

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

Aliprantis, C. D., and K. C. Border (2006): Infinite Dimensional Analysis-A Hitchhiker's Guide. Springer, Berlin.

Andrews, D. W. K. (1994): "Chapter 37: Empirical Process Methods in Econometrics", vol. 4 of Handbook of Econometrics, pp. 2247–2294. Elsevier, Amsterdam.

Andrews, D. W. K., and X. Shi (2011): "Inference for Parameters Defined by Conditional Moment Inequalities", Discussion Paper, Yale University.

Bajari, P., C. L. Benkard, and J. Levin (2007): "Estimating Dynamic Models of Imperfect Competition", Econometrica, 75(5), 1331–1370.

Bontemps, C., T. Magnac, and E. Maurin (2011): "Set Identified Linear Models", CeMMAP Working Paper.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models", Handbook of Econometrics, 6, 5549–5632.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models1", Econometrica, 75(5), 1243–1284.

Ciliberto, F., and E. Tamer (2009): "Market Structure and Multiple Equilibria in Airline Markets", Econometrica, 77(6), 1791–1828.

Folland, G. (1999): Real Analysis: Modern Techniques and Their Applications, vol. 40. Wiley-Interscience, New York.

Guggenberger, P., J. Hahn, and K. Kim (2008): "Specification Testing under Moment Inequalities", Economics Letters, 99(2), 375–378.

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

Jennrich, R. I. (1969): "Asymptotic Properties of Nonlinear Least Squares Estimators", Annals of Mathematical Statistics, 40(2), 633–643.

Kaido, H., and H. White (2010): "A Two-Stage Approach for Partially Identified Models", Discussion Paper, University of California San Diego.

Lindenstrauss, J., D. Preiss, and J. Tiser (2007): "Differentiability of Lipschitz Maps", in Banach Spaces and Their Applications in, Analysis, pp. 111–123.

Luttmer, E. G. J. (1996): "Asset Pricing in Economies with Frictions", Econometrica, 64(6), 1439–1467.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome", Econometrica, 70(2), 519–546.

Molchanov, I. S. (2005): Theory of Random Sets. Springer, Berlin.

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

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing", Handbook of Econometrics, 4, 2111–2245.

Pakes, A. (2010): "Alternative Models for Moment Inequalities", Econometrica, 78(6), 1783–1822.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2006): "Moment Inequalities and Their Application", Working Paper, Harvard University.

Ponomareva, M., and E. Tamer (2010): "Misspecification in Moment Inequality Models: Back to Moment Equalities?" Econometrics Journal, 10, 1–21.

Santos, A. (2011): "Instrumental Variables Methods for Recovering Continuous Linear Functionals", Journal of Econometrics, 161, 129–146.

Sherman, R. P. (1993): “The Limiting Distribution of the Maximum Rank Correlation Estimator”, Econometrica, 61(1), 123–137.

Tamer, E. (2003): “Incomplete Simultaneous Discrete Response Model with Multiple Equilibria”, The Review of Economic Studies, 70(1), 147–165.

van der Vaart, A. W., and J. A. Wellner (1996): Weak Convergence and Empirical Processes: with Applications to Statistics. Springer, New York.

White, H. (1982): “Maximum Likelihood Estimation of Misspecified Models”, Econometrica, 50(1), 1–25.