

**WEIGHTED MINIMUM MEAN-SQUARE DISTANCE
FROM INDEPENDENCE ESTIMATION**

BY

DONALD J. BROWN and MARTEN H. WEGKAMP

COWLES FOUNDATION PAPER NO. 1042



**COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281**

2002

<http://cowles.econ.yale.edu/>

WEIGHTED MINIMUM MEAN-SQUARE DISTANCE
FROM INDEPENDENCE ESTIMATION

BY DONALD J. BROWN AND MARTEN H. WEGKAMP¹

1. INTRODUCTION

MANSKI (1983) INTRODUCED minimum mean-square distance from independence estimation of semiparametric econometric models separable in the unobserved exogenous variables ε , i.e., $\varepsilon = \rho(X, Y, \theta)$ where X is a random vector of observed exogenous variables, Y is a random vector of observed endogenous variables, θ is a vector of unknown parameters, ε is drawn from a fixed but unknown distribution, and ε is stochastically independent of X . An important special case is the implicit nonlinear simultaneous equations model, where a reduced form function $Y = \rho^{-1}(X, \varepsilon, \theta)$ exists. This model is a central topic of this paper. Manski proved strong consistency of his estimator, but was unable to derive the first-order asymptotic distribution. The criterion function for Manski's estimator is the mean-square distance between the joint empirical cumulative distribution function of ε and X and the product of its marginal cumulative distribution functions. The criterion function for our estimator is the mean-square distance between the joint empirical cumulative distribution of ε and X and the product of its marginal cumulative distribution functions, weighted by a probability measure on the product space of ε , and x .

These weighted minimum mean-square distance from independence estimators offer tractable procedures for those applications where the econometrician assumes that ε is stochastically independent of X and ε is drawn from a fixed but unknown distribution. Such is the case for the continuous random utility model proposed by Brown and Matzkin (1998).

Comparing minimum mean-square distance from independence estimation to maximum likelihood estimation, Manski observes that the joint maximization of the likelihood over the product of the parameter space Θ and the family of possible distributions of ε is often computationally intractable. In contrast, weighted minimum mean-square distance from independence estimation only requires that we minimize over the finite dimensional parameter space Θ . If $\hat{\theta}$ is a consistent estimate of θ_0 , the true parameter value, then we give sufficient conditions on the mapping $\theta \mapsto \rho(x, y, \theta)$ for consistent estimation of the true distribution of ε .

Since weighted minimum mean-square distance from independence is an extremum estimator, identification means that the asymptotic criterion function has a unique minimum at the truth—see Newey and McFadden (1994). In particular, if $\rho(X, Y, \theta)$ is nonlinear in θ , then as they point out “primitive conditions for identification (existence of a unique global minimum) become quite difficult.” In practice, global GMM identification for nonlinear simultaneous equations models is simply assumed by the econometrician.

A striking and important feature of weighted minimum mean-square distance from independence estimation is the existence of primitive conditions for identification. For implicit nonlinear simultaneous equations models, Brown (1983) and more generally

¹ The authors thank Don Andrews, Steve Berry, and David Pollard for their helpful remarks. We also appreciate the comments of the referees and co-editor.

Roehrig (1988) have given sufficient conditions on the primitive $\rho(X, Y, \theta)$ for identification, if ε is assumed to be stochastically independent of X .

Unfortunately, Manski's regularity conditions for strong consistency of minimum mean-square distance from independence estimation—see the Corollary of Manski (1983, p. 314)—are unattractive in at least two respects. First, he simply assumes the existence of a unique minimum, a high-level assumption for which he provides no sufficient conditions on the model's primitives. Second and more importantly—given our prior discussion of the Brown and Roehrig results—is his assumption that the sets $S(\nu, \eta, \theta) = \{(x, y) : (x, y) \in S, x < \nu, \rho(x, y, \theta) < \eta\}$, where S is a compact convex subset of Euclidean space and $\rho(x, y, \theta)$ is continuous on $S \times \Theta$, are convex with boundaries having measure zero with respect to the true fixed but unknown distribution of ε . These technical assumptions are difficult to verify in practice. The latter assumption is crucial for his consistency argument, since it allows him to invoke a uniform law of large numbers due to Ranga Rao (1962).

In this paper, we introduce the family of weighted minimum mean-distance from independence estimators that are computationally tractable and identified. Moreover our regularity conditions for consistency and asymptotic normality are satisfied in many applications. That is, we show that if $\rho(x, y, \theta)$ is sufficiently smooth in (x, y, θ) and the possible distributions of ε have sufficiently smooth densities, then our estimators are strongly consistent and asymptotically normal. Also, we prove under these assumptions that bootstrap estimates of the sampling distribution and the asymptotic variance are also consistent.

As conjectured by Manski, the main tools of our analysis are techniques derived from the theory of empirical processes, necessitated by our nonsmooth criterion function. For instance, see Pakes and Pollard (1989) for a lucid discussion and econometric application of empirical process theory. Their paper and this paper are related both in method and economic motivation. An application of their results is the estimation of a discrete random choice model and an intended application of our results is the estimation of the continuous random choice model of Brown and Matzkin.

Two significant differences between our paper and the paper of Pakes and Pollard are that our estimator is an extremum estimator, i.e., we minimize a nonsmooth random criterion function and their estimator is a Z -estimator,² i.e., they approximately solve a family of possibly nonsmooth random equations. More importantly, Theorem 3.2 in Wegkamp (1999, p. 40) employed here subsumes as special cases: M -estimation, Cramer-von Mises estimation, regression and weighted minimum mean-square distance from independence estimation. See Wegkamp (1995), Andrews (1999), and Pollard (forthcoming) for similar results. Additional references on empirical process theory and their statistical applications can be found in Dudley (1999), Pollard (1984, 1985), and van der Vaart and Wellner (1996). Econometric applications can be found in Andrews (1994).

This paper is organized as follows. We discuss in turn identification, consistency, asymptotic normality, and resampling. A treatment of asymptotic efficiency is beyond the scope of this paper. For an application of our estimation procedure, see Brown and Wegkamp (2000).

2. IDENTIFICATION OF MINIMUM DISTANCE FROM INDEPENDENCE ESTIMATORS

Compactness of the parameter space Θ and the continuity of the asymptotic criterion function imply that the optimum is well-separated, provided the extremum estimator has a unique global minimum (maximum), i.e., the model is identified.

² See Section 3.3 in van der Vaart and Wellner (1996) for a discussion.

We assume that Θ is a compact subset of a Euclidean space. Moreover, let \mathcal{X} be a subset of \mathbb{R}^L , \mathcal{Y} be a subset of \mathbb{R}^K , and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be a subset of \mathbb{R}^{L+K} . Then $Z = (X, Y)$ is a random vector taking values in \mathcal{Z} . For all $\theta \in \Theta$, $\rho(\cdot, \theta)$ is a mapping from \mathcal{Z} into \mathbb{R}^K .

We define minimum distance from independence estimators, as extremum estimators where the asymptotic function $d(\cdot, \cdot)$ is a metric on the space of joint cumulative distribution functions (c.d.f.'s) of (X, ε) , where ε takes values in \mathbb{R}^K . If $H(x, \varepsilon)$ is the joint c.d.f. of (X, ε) and $F(x), G(\varepsilon)$ the associated marginal c.d.f.'s, then $d(H(x, \varepsilon), F(x)G(\varepsilon)) = 0$ iff X and ε are stochastically independent. In Manski (1983) d is the mean-square distance, in Brown and Matzkin (1998) d is the metric on the space of c.d.f.'s induced by the Prohorov metric on the space of measures, and in this paper d is the weighted mean-square distance. Our discussion of implicit nonlinear simultaneous equations models follows the expositions of Brown (1983), Roehrig (1988), and Brown-Matzkin (1998).

A structure S is an ordered pair $\langle \rho(X, Y, \theta), H(x, \varepsilon) \rangle$. The structural equations are defined as $\varepsilon = \rho(X, Y, \theta)$. Our model consists of all structures S that satisfy the following assumptions:

ASSUMPTION I.1: $\exists!$ reduced form $Y = \gamma(X, \varepsilon, \theta)$ such that $\varepsilon \equiv \rho(X, \gamma(X, \varepsilon, \theta), \theta)$.

ASSUMPTION I.2: The matrix $\partial \rho / \partial y$ has full rank a.e.

ASSUMPTION I.3: $H(x, \varepsilon) = F(x)G(\varepsilon)$ for all $(x, \varepsilon) \in \mathcal{X} \times \mathcal{Y}$, i.e., X and ε are stochastically independent.

ASSUMPTION I.4: $H(x, \varepsilon)$ is absolutely continuous, with respect to Lebesgue measure, with positive density.

The following notation will prove useful:

- (i) $H_\theta(s, t) \equiv P\{X \leq s, \rho(X, \gamma(X, \varepsilon, \theta_0), \theta) \leq t\}$.
- (ii) $F(x)$ and $G_\theta(\varepsilon)$ are the associated marginal c.d.f.'s of $H_\theta(x, \varepsilon)$.
- (iii) $M(\theta) \equiv d(H_\theta(x, \varepsilon), F(x)G_\theta(\varepsilon))$.

Given Assumption I.1 each structure generates a joint c.d.f. of (X, Y) . Our maintained assumption is that the observed c.d.f. of $Z = (X, Y)$ is generated by some structure $S_0 = \langle \rho(X, Y, \theta_0), H_{\theta_0}(x, \varepsilon) \rangle$ in our model.

Brown (1983, pp. 180–181) in his seminal paper on identification proved the fundamental result, Theorem 1, for the special case of semiparametric implicit simultaneous equations models that are only nonlinear in the variables. Subsequently, Roehrig (1988) extended Brown's analysis to nonparametric and semiparametric implicit nonlinear simultaneous equations models.

THEOREM 1 (Roehrig (1988, Lemma 3.3, p. 437)):

$$H_\theta(x, \varepsilon) = F(x)G_\theta(\varepsilon) \text{ a.e.} \Leftrightarrow \frac{\partial \rho(x, \gamma(x, \varepsilon, \theta_0), \theta)}{\partial x} = 0 \text{ a.e.}$$

The identification condition for minimum distance from independence estimators is an immediate consequence of Theorem 1.

THEOREM 2: θ_0 is the unique global minimum of $M(\theta)$ iff $\forall \theta \neq \theta_0, \exists$ a set of pairs $(\bar{x}, \bar{\varepsilon})$ with positive probability such that

$$\left. \frac{\partial \rho(x, \gamma(x, \varepsilon, \theta_0), \theta)}{\partial x} \right|_{(\bar{x}, \bar{\varepsilon})} \neq 0.$$

The following result is well known, but necessary for our proof of consistency. First we recall the definition of a well-separated minimum.

DEFINITION 1: θ_0 is a well-separated minimum of $M(\theta)$ if $\inf_{\{\theta \in \Theta: m(\theta, \theta_0) \geq \varepsilon\}} M(\theta) > M(\theta_0)$, where m is a metric on Θ .

THEOREM 3 (Newey-McFadden (1994, Theorem 2.1, p. 2121)): *If Θ is compact, $M(\theta)$ is continuous on Θ and θ_0 is the unique global minimum of $M(\theta)$, then θ_0 is a well-separated minimum of M .*

3. CONSISTENCY AND ASYMPTOTIC NORMALITY OF WEIGHTED MINIMUM MEAN-SQUARE DISTANCE FROM INDEPENDENCE ESTIMATORS

Our main model assumption in this and the next section is that $\rho(X, Y, \theta)$ is independent of X if and only if $\theta = \theta_0$. Based on independent observations Z_1, \dots, Z_n , we will now construct an estimate of θ_0 , and establish its limiting sampling distribution under a set of regularity conditions. The independence assumption between X and $\rho(X, Y, \theta_0)$ is equivalent with

$$H_\theta(x, \varepsilon) = F(x)G_\theta(\varepsilon) \forall (x, \varepsilon) \in \mathcal{X} \times \mathcal{Y} \Leftrightarrow \theta = \theta_0.$$

As a consequence, for any bounded measure μ on \mathcal{X} , the criterion function³

$$M(\theta) = \int [H_\theta(x, \varepsilon) - F(x)G_\theta(\varepsilon)]^2 d\mu(x, \varepsilon)$$

is minimized at $\theta = \theta_0$. Motivated by this observation, we propose to minimize the empirical counterpart of $M(\theta)$,

$$M_n(\theta) = \int [H_{n\theta}(x, \varepsilon) - F_n(x)G_{n\theta}(\varepsilon)]^2 d\mu(x, \varepsilon)$$

over $\theta \in \Theta$. Here $F_n, G_{n\theta}$, and $H_{n\theta}$ are the empirical c.d.f.'s associated with F, G_θ , and H_θ , respectively, based on the observed data Z_1, \dots, Z_n . For instance,

$$G_{n\theta}(\varepsilon) = \frac{1}{n} \sum_{i=1}^n I\{\rho(Z_i, \theta) \leq \varepsilon\}.$$

³ Our criterion function differs from the one proposed by Manski, who computes the mean-square distance with respect to the empirical measure. A rigorous proof of asymptotic normality of Manski's estimator requires the use of the less familiar theory of U -processes; see de la Peña and Giné (1999, Chapter 5).

The use of a fixed measure allows us to establish the asymptotic properties of our estimator within the theory of empirical processes and to smooth the empirical criterion function $M_n(\theta)$, facilitating the computation of $\hat{\theta}_n = \arg \min_\theta M_n(\theta)$.

The resulting minimum is denoted by $\hat{\theta}$,⁴ which satisfies

$$M_n(\hat{\theta}) \leq M_n(\theta) \quad \text{for all } \theta \in \Theta.$$

Next we describe the set of regularity assumptions, followed by a brief discussion.

ASSUMPTION A.1: *The parameter space Θ is compact, and θ_0 is an interior point.*

ASSUMPTION A.2: *The model is identified.*

ASSUMPTION A.3: *The collection of functions $\{\rho(\cdot, \cdot, \theta) : \theta \in \Theta\}$ is either*

- *a subset of a finite dimensional space or*
- *each coordinate of the mapping $(x, y) \mapsto \rho(x, y, \theta)$ is an element of $C_K^\alpha[\mathcal{X} \times \mathcal{Y}]$,⁵ $K > 0$ and \mathcal{X} and \mathcal{Y} compact, for all $\theta \in \Theta$. In this case we require that $H(x, \varepsilon)$ has a bounded density.*

ASSUMPTION A.4: *The random vector ε has a continuous c.d.f. G .*

ASSUMPTION A.5: *The mapping $\theta \mapsto \rho(x, y, \theta)$ is Lipschitz at θ_0 uniformly in $x \in \mathcal{X}$, $y \in \mathcal{Y}$.*

ASSUMPTION A.6: *The mapping $\theta \mapsto D_\theta(x, \varepsilon) \equiv H_\theta(x, \varepsilon) - F(x)G_\theta(\varepsilon)$ is differentiable at θ_0 in $L_2(\mu)$, that is, there exists $\Delta \in L_2(\mu)$ such that*

$$\lim_{\|\theta - \theta_0\| \rightarrow 0} \int \left(\frac{D_\theta(x, \varepsilon) - (\theta - \theta_0)' \Delta(x, \varepsilon)}{\|\theta - \theta_0\|} \right)^2 d\mu(x, \varepsilon) = 0.$$

ASSUMPTION A.7: *The mapping $\theta \mapsto M(\theta)$ has a positive definite second derivative matrix V at θ_0 .*

The first Assumption A.1 is a standard condition in the literature. The second Assumption A.2 was the main issue in the previous section, where we derived a necessary and sufficient condition under the Assumptions I.1–I.4, i.e., Theorem 2. Concerning the third Assumption A.3, we observe that the standard compactness conditions for spaces of smooth functions used in economic theory (e.g., see Mas-Colell (1985, Section K in Chapter 1)) are sufficient to guarantee the third assumption. We note in passing that only certain metric entropy properties of $\{\rho(\cdot, \cdot, \theta) : \theta \in \Theta\}$ are needed to conduct our proof; conditions on these spaces other than Assumption A.3 may also give the desired metric entropy property.

Assumptions A.5 and A.6 are implied by pointwise smoothness of the mapping $\theta \mapsto \rho(\cdot, \cdot, \theta)$. It should be noted that Assumption A.6 is weaker than pointwise differentiability (cf. van der Vaart (1998, Lemma 7.6, p. 95), and Pollard (forthcoming, Chapter 4)).

We are now in the position to state our main results.

⁴ We will assume without loss of generality that a minimum exists, since otherwise we can always take any $\theta \in \Theta$ that minimizes M_n within a constant $1/n^2$ without affecting the results.

⁵ Each coordinate mapping must have uniformly bounded (by K) partial derivatives through order $\beta = \lfloor \alpha \rfloor$, and the derivatives of order β will satisfy a uniform Hölder condition of order $\alpha - \beta$, and with Lipschitz constant bounded by K . For a complete description of the space $C_K^\alpha[X \times \mathcal{Y}]$, we refer to Dudley (1999, p. 252) or van der Vaart and Wellner (1996, p. 154).

THEOREM 4 (Consistency): *Under Assumptions A1, A2, A3, and continuity of M at θ_0 (which is implied by A.7), $\hat{\theta}$ is strongly consistent, i.e., $\hat{\theta} \rightarrow_{\text{a.s.}} \theta_0$.*

COROLLARY 1: *Under the assumptions of Theorem 4 and A.5, $H_{n\hat{\theta}}(x, \varepsilon) \rightarrow_{\text{a.s.}} H(x, \varepsilon)$ for all $(x, \varepsilon) \in \mathcal{X} \times \mathcal{Y}$.*

THEOREM 5 (Asymptotic Normality): *Under the regularity Assumptions A.1–A.3 described above, $\sqrt{n}(\hat{\theta} - \theta_0)$ converges to a mean zero, nondegenerate multivariate normal distribution. The limiting covariance matrix Σ is $4V^{-1}WV^{-1}$, where V is defined in Assumption A.7 and*

$$W = \iint \Delta(x, \varepsilon)\Delta'(\bar{x}, \bar{\varepsilon})[F(x)F(\bar{x})G(\min(\varepsilon, \bar{\varepsilon})) + F(\min(x, \bar{x}))G(\varepsilon)G(\bar{\varepsilon}) + H(\min(x, \bar{x}), \min(\varepsilon, \bar{\varepsilon})) - 3H(x, \varepsilon)H(\bar{x}, \bar{\varepsilon})]d\mu(x, \varepsilon)d\mu(\bar{x}, \bar{\varepsilon}).$$

The minimum between two vectors x and \bar{x} should be understood coordinatewise.

4. RESAMPLING ESTIMATES OF THE SAMPLING DISTRIBUTION AND ASYMPTOTIC VARIANCE

In this section we provide an alternative to the normal approximation of the sampling distribution of $\hat{\theta}$ by means of resampling. We show that the ordinary nonparametric bootstrap is consistent. To formulate our result, let the pairs Z_1^*, \dots, Z_n^* be the (bootstrap) sample drawn from the data Z_1, \dots, Z_n with replacement. We denote the bootstrap counterpart of M_n based on the bootstrap sample by M_n^* , and let $\hat{\theta}^*$ be its minimum over Θ .

THEOREM 6: *Under the regularity Assumptions A.1–A.7 described above, the conditional distribution of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ consistently estimates the distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ (in probability).*

We end with a discussion of the asymptotic covariance matrix of $\sqrt{n}(\hat{\theta} - \theta_0)$. In principle, under sufficient smoothness assumptions, we could plug in $\hat{\theta}$ and P_n in the covariance matrix $\Sigma = \Sigma(\theta_0, P)$. Here P is the probability measure of Z , and P_n is the empirical measure, putting mass $1/n$ at each observation Z_i . However, $\Sigma(\theta_0, P)$ has a complicated structure, and the bootstrap estimator of the variance provides an attractive alternative. Second, we show that the delete $-d$ jackknife estimator⁶ of the variance of linear combinations $c'\hat{\theta}$ is consistent for d satisfying

$$(1) \quad d/n \geq \varepsilon \quad \text{for some } \varepsilon > 0 \quad \text{and} \quad n - d \rightarrow \infty.$$

We were not able to show that the ordinary jackknife ($d = 1$) works due to the lack of smoothness of the map $\theta \mapsto M_n(\theta)$. For the same reason, the jackknife estimator of the variance of the sample median is inconsistent (cf. Shao and Tu (1995)).

⁶ Let $\hat{\theta}_{d,s}$ be the estimate based on the data set $Z_i, i \in s$, where s is a subset of $\{1, 2, \dots, n\}$ with size $n - d$. Let \mathcal{S} be the collection of all possible subsets of $\{1, 2, \dots, n\}$ of size $n - d$, and let $N = \binom{n}{d}$ be its cardinality. The delete $-d$ jackknife of $c'\hat{\theta}$ is defined as

$$\frac{n-d}{dN} \sum_{s \in \mathcal{S}} \left(c'\hat{\theta}_{d,s} - \frac{1}{N} \sum_s c'\hat{\theta}_{d,s} \right)^2.$$

THEOREM 7: *Under the regularity conditions A.1–A.7, the nonparametric bootstrap and delete $-d$ jackknife estimators of the variance of $c'\sqrt{n}(\hat{\theta} - \theta_0)$, where d satisfies (1), are consistent, for all $c \in \mathbb{R}^{\dim(\theta)}$.*

5. PROOFS

First we introduce some additional notation and results. Define the sets

$$A_{\theta,y} = \{z \in \mathcal{X} : \rho(z, \theta) \leq y\} \quad \text{and} \quad B_x = \{t \in \mathcal{X} : t \leq x\},$$

and the associated collections

$$\begin{aligned} \mathcal{A} &= \{A_{\theta,y} : \theta \in \Theta, y \in \mathcal{Y}\}, \quad \mathcal{B} = \{B_x : x \in \mathcal{X}\}, \quad \text{and} \\ \mathcal{C} &= \{A \cap (B \times \mathcal{Y}) : A \in \mathcal{A}, B \in \mathcal{B}\}. \end{aligned}$$

Let P be the probability measure of $Z = (X, Y)$, and let P_n be its empirical measure based on Z_1, \dots, Z_n , which puts mass $1/n$ at each observation. Recall the definition of

$$D_\theta(x, \varepsilon) = H_\theta(x, \varepsilon) - F(x)G_\theta(\varepsilon),$$

and define further

$$D_{n\theta}(x, \varepsilon) = H_{n\theta}(x, \varepsilon) - F_n(x)G_{n\theta}(\varepsilon).$$

Observe that

$$\begin{aligned} D_{n\theta}(x, \varepsilon) - D_\theta(x, \varepsilon) &= \{H_{n\theta}(x, \varepsilon) - H_\theta(x, \varepsilon)\} + F_n(x)\{G_\theta(\varepsilon) - G_{n\theta}(\varepsilon)\} \\ &\quad + G_\theta(\varepsilon)\{F(x) - F_n(x)\} \end{aligned}$$

and consequently,

$$\begin{aligned} &\sup_{\theta \in \Theta, x \in \mathcal{X}, \varepsilon \in \mathcal{Y}} |D_{n\theta}(x, \varepsilon) - D_\theta(x, \varepsilon)| \\ &\leq \sup_{A \in \mathcal{A}} |(P_n - P)(A)| + \sup_{C \in \mathcal{C}} |(P_n - P)(C)| + \sup_{B \in \mathcal{B}} |(P_n - P)(B)|. \end{aligned}$$

For any measure Q on \mathcal{X} , any class of functions $\mathcal{F} \subset L_2(Q)$, and any positive number δ , let $N(\delta, \mathcal{F}, L_2(Q))$ be the δ -covering number (possibly infinite) of the class \mathcal{F} with respect to the $L_2(Q)$ metric, that is, the number of closed balls with radius δ in $L_2(Q)$ needed to cover \mathcal{F} . The δ -bracketing number is denoted by $N_B(\delta, \mathcal{F}, L_2(Q))$, i.e., the number of δ -brackets needed to cover \mathcal{F} . A δ -bracket of a function $f \in \mathcal{F}$ is the pair (f_L, f_U) such that $f_L \leq f \leq f_U$ and $\int |f_U - f_L|^2 dQ \leq \delta$; see, e.g., van der Vaart and Wellner (1996).

LEMMA 1: *Suppose that $\{\rho(\cdot, \theta) : \theta \in \Theta\}$ is a subset of a finite dimensional vector space. Then*

$$\sup_{Q \text{ discrete}} N(\delta, \mathcal{F}, L_2(Q)) \leq C\delta^{-V}$$

for $\mathcal{F} = \mathcal{A}, \mathcal{B}, \mathcal{C}$ and $V > 1$.

PROOF: The statement is well known for the class of sets \mathcal{B} (cf. van der Vaart and Wellner (1996)). For the class \mathcal{A} we argue as follows. Let d be the dimension of \mathcal{Y} . Since $\varepsilon = (\varepsilon_1, \dots, \varepsilon_d)'$ and $\rho(z, \theta) = (\rho^{(1)}(z, \theta), \dots, \rho^{(d)}(z, \theta))'$, we can write $A_{\theta, \varepsilon}$ as an intersection, viz.,

$$A_{\theta, \varepsilon} = A_{\theta, \varepsilon}^{(1)} \cap \dots \cap A_{\theta, \varepsilon}^{(d)},$$

where

$$A_{\theta, \varepsilon}^{(i)} = \{z \in \mathcal{X} : \rho^{(i)}(z, \theta) \leq \varepsilon_i\}.$$

It is well known that $\{A_{\theta, \varepsilon}^{(i)} : \theta \in \Theta, \varepsilon \in \mathcal{Y}\}$ is a VC-class of sets if $\{\rho^{(i)}(z, \theta) : \theta \in \Theta\}$ is a subset of a finite dimensional vector space; see for instance van der Vaart and Wellner (1996, Lemma 2.6.15, p. 146). Hence all $\{A_{\theta, \varepsilon}^{(i)} : \theta \in \Theta, \varepsilon \in \mathcal{Y}\}$ are VC classes for $i = 1, \dots, d$. The VC property is closed under taking intersections (cf. van der Vaart and Wellner (1996, Lemma 2.6.17, p. 147), so that \mathcal{A} forms a VC-class of sets, and hence the claim for \mathcal{A} follows by Theorem 2.6.4 in van der Vaart and Wellner (1996, p. 136).

The claim for \mathcal{C} follows since \mathcal{A} and $\mathcal{B} \times \mathcal{Y}$ are VC, and hence $\mathcal{C} = \{A \cap (B \times \mathcal{Y}), A \in \mathcal{A}, B \in \mathcal{B}\}$ is VC as shown in van der Vaart and Wellner (1996, Lemma 2.6.17, p. 147). Q.E.D.

LEMMA 2: *Suppose that \mathcal{X} is compact with nonempty interior and that each coordinate mapping of $\rho(z, \theta) \in C_K^\alpha(\mathcal{X})$ for all $\theta \in \Theta$. Then the collections $\mathcal{F} = \mathcal{A}, \mathcal{B}, \mathcal{C}$ all satisfy*

$$\log N_B(\delta, \mathcal{F}, L_2(P)) \leq C\delta^{-V}$$

for some $V = 2D/\alpha < 2$ and for all probability measures P with a uniformly bounded density and $\alpha > D$.

PROOF: Corollary 2.7.3 in van der Vaart and Wellner (1996, p. 157) bounds the entropy of bracketing of the collection of subgraphs of $C_K^\alpha(\mathcal{X})$, and the result for \mathcal{A} is immediate. The condition $\alpha > D$ is needed to ensure that $V < 2$. It is easy to show that the δ bracketing number of \mathcal{C} is the product of the δ bracketing numbers of \mathcal{A} and \mathcal{B} as they are bounded classes. Taking the logarithm entails the desired result. Q.E.D.

In particular, the entropy bounds above show that the classes \mathcal{A}, \mathcal{B} , and \mathcal{C} are P -Donsker classes under Assumption A.3.

PROOF OF THEOREM 4: First, observe that

$$\begin{aligned} & \sup_{\theta \in \Theta} |(M_n - M)(\theta)| \\ & \leq 4 \sup_{\theta \in \Theta} \int |D_{n\theta} - D_\theta| d\mu \\ & \leq 4 \left(\int d\mu \right) \sup_{\theta \in \Theta, x \in \mathcal{X}, \varepsilon \in \mathcal{Y}} |(D_{n\theta} - D_\theta)(x, \varepsilon)| \\ & \leq 4 \left(\int d\mu \right) \left\{ \sup_{A \in \mathcal{A}} |(P_n - P)A| + \sup_{B \in \mathcal{B}} |(P_n - P)B| + \sup_{C \in \mathcal{C}} |(P_n - P)C| \right\} \\ & \xrightarrow{\text{a.s.}} 0 \end{aligned}$$

since \mathcal{A} , \mathcal{B} , and \mathcal{C} are Glivenko-Cantelli classes and μ is a bounded measure. Hence with probability one,

$$M_n(\hat{\theta}) \leq M_n(\theta_0) + o(1/n) = M(\theta_0) + o(1/n) \leq M(\hat{\theta}) + o(1/n).$$

The compactness assumption on Θ and the identifiability assumption yield that $M(\theta)$ has a unique, well-separated minimum (at θ_0) (cf. Theorem 3). Q.E.D.

LEMMA 3: *Under Assumptions A.3, A.4, A.5, the process $\sqrt{n}(D_{n\theta} - D_\theta)(x, \varepsilon)$ is stochastically equicontinuous at θ_0 with respect to the Euclidean metric on Θ , for all $x \in \mathcal{X}$ and $\varepsilon \in \mathcal{Y}$, i.e.,*

$$\sqrt{n}(D_{n\theta} - D_\theta)(x, \varepsilon) - \sqrt{n}(D_{n\theta_0} - D_{\theta_0})(x, \varepsilon) \xrightarrow{P} 0 \quad \text{as} \quad \theta \xrightarrow{P} \theta_0.$$

PROOF: The decomposition

$$\begin{aligned} \sqrt{n}(D_{n\theta} - D_\theta)(x, \varepsilon) &= \sqrt{n}(H_{n\theta} - H_\theta)(x, \varepsilon) - F_n(x)\sqrt{n}(G_{n\theta} - G_\theta)(\varepsilon) \\ &\quad - G_\theta(\varepsilon)\sqrt{n}(F_n - F)(x) \end{aligned}$$

forms a sum of three terms, each stochastically equicontinuous at θ_0 . This is a consequence of the already mentioned Donsker property of \mathcal{A} , \mathcal{B} , and \mathcal{C} and the fact that the mapping $\theta \mapsto G_\theta$ is continuous at θ_0 , since the Lipschitz condition on $\theta \mapsto \rho(\cdot, \theta)$ yields both

$$\begin{aligned} \mathbb{P}\{\rho(Z, \theta) \leq \lambda\} &\geq G(\lambda - C\|\theta - \theta_0\|) \quad \text{and} \\ \mathbb{P}\{\rho(Z, \theta) \leq \lambda\} &\leq G(\lambda + C\|\theta - \theta_0\|) \end{aligned}$$

and the continuity follows.

It should be stressed that the Donsker property implies that the process is stochastically equicontinuous with respect to the $L_2(P)$ metric on the sets $A \in \mathcal{A}$ and $C \in \mathcal{C}$, not necessarily the Euclidean distance on Θ .⁷ However, the Lipschitz condition on $\theta \mapsto \rho(\cdot, \theta)$ yields that

$$\begin{aligned} \mathbb{P}\{\rho(Z, \theta) \leq \lambda, \rho(Z, \theta_0) \leq \lambda\} &\geq \mathbb{P}\{\rho(Z, \theta_0) \leq \lambda - C\|\theta - \theta_0\|\} \\ &= G_{\theta_0}(\lambda - C\|\theta - \theta_0\|) \\ &\rightarrow G_{\theta_0}(\lambda) \quad \text{for} \quad \theta \rightarrow \theta_0 \end{aligned}$$

as $G_{\theta_0} \equiv G$ is continuous. On the other hand,

$$\begin{aligned} \mathbb{P}\{\rho(Z, \theta) \leq \lambda, \rho(Z, \theta_0) \leq \lambda\} &\leq \mathbb{P}\{\rho(Z, \theta_0) \leq \lambda\} \\ &= G_{\theta_0}(\lambda). \end{aligned}$$

We have shown that $\mathbb{P}\{A_{\theta, \lambda} \cap A_{\theta_0, \lambda}\} \rightarrow \mathbb{P}\{A_{\theta_0, \lambda}\}$ as $\theta \rightarrow \theta_0$. By a similar argument we see that $\mathbb{P}\{A_{\theta, \lambda}\} \rightarrow \mathbb{P}\{A_{\theta_0, \lambda}\}$ as $\theta \rightarrow \theta_0$, so that

$$P(A_{\theta, \lambda} - A_{\theta_0, \lambda})^2 = P\{A_{\theta, \lambda}\} + P\{A_{\theta_0, \lambda}\} - 2P\{A_{\theta, \lambda} \cap A_{\theta_0, \lambda}\} \rightarrow 0,$$

⁷ Recall that the empirical process $\sqrt{n}(P_n - P)$, indexed by (indicator functions of) sets $A \in \mathcal{A}$, is stochastically equicontinuous at I_{A_0} , iff for all $\varepsilon, \eta > 0$ there exists a $\delta > 0$ such that $\limsup_{n \rightarrow \infty} \mathbb{P}\{\sup_{P|I_{A_0} - I_{A_0}| \leq \delta^2} |\sqrt{n}(P_n - P)(A)| > \varepsilon\} < \eta$.

as $\theta \rightarrow \theta_0$. In other words, the parameterization $\theta \mapsto I_{A_{\theta,y}}$ is continuous at θ_0 in the $L_2(P)$ sense. Hence in view of the stochastic equicontinuity with respect to the $L_2(P)$ distance,

$$G_{n\theta}(\varepsilon) - G_\theta(\varepsilon) - G_{n,\theta_0}(\varepsilon) + G_{\theta_0}(\varepsilon) \xrightarrow{P} 0,$$

for all $\theta \rightarrow_P \theta_0$ and all $\varepsilon \in \mathcal{Y}$. A similar argument applies to the first term and the claim follows. *Q.E.D.*

PROOF OF COROLLARY 1: The calculation in the proof of Lemma 3 using Assumption A.5 shows that $\theta \mapsto H_\theta$ is continuous. The consistency of $\hat{\theta}$ above implies that $H_{\hat{\theta}}(x, \varepsilon) \rightarrow H_{\theta_0}(x, \varepsilon) \equiv H(x, \varepsilon)$ a.s. by the continuous mapping theorem. The proof of Theorem 4 implies further that $\sup_{\theta \in \Theta} |H_{n\theta}(x, \varepsilon) - H_\theta(x, \varepsilon)| \rightarrow_{a.s.} 0$. In particular, $|H_{n\hat{\theta}}(x, \varepsilon) - H_{\hat{\theta}}(x, \varepsilon)| \rightarrow_{a.s.} 0$, and Corollary 1 follows after an application of the triangle inequality. *Q.E.D.*

PROOF OF THEOREM 5: The result follows after an application of Theorem 3.2 in Wegkamp (1999, p. 48). We need to check the following conditions:

- (i) $\hat{\theta} \rightarrow_P \theta_0$;
- (ii) $M(\theta)$ has a nonsingular second derivative V at θ_0 ;
- (iii) $\sqrt{n}(M_n - M)(\theta)$ is stochastically differentiable at θ_0 , that is, there exists a W_n that converges weakly to a tight Gaussian distribution such that

$$\sqrt{n}(M_n - M)(\theta) = \sqrt{n}(M_n - M)(\theta_0) + (\theta - \theta_0)' W_n + o_P(1 + \sqrt{n}\|\theta - \theta_0\|)$$

for $\theta \rightarrow \theta_0$.

We have already established consistency, and we assumed the second requirement (ii). It remains to verify the stochastic differentiability requirement (iii). Recall that

$$D_\theta(x, \varepsilon) = (\theta - \theta_0)' \Delta(x, \varepsilon) + r_\theta(x, \varepsilon),$$

where $\int \Delta^2 d\mu < \infty$, and $\int r_\theta^2(x, \varepsilon) d\mu = o(\|\theta - \theta_0\|^2)$ for $\theta \rightarrow \theta_0$. Next, observe that for $\theta \rightarrow_P \theta_0$

$$\begin{aligned} M_n(\theta) - M(\theta) &= \int (D_{n\theta} - D_\theta + D_\theta)^2 d\mu - \int D_\theta^2 d\mu \\ &= \int (D_{n\theta} - D_\theta)^2 d\mu + 2 \int D_\theta (D_{n\theta} - D_\theta) d\mu \\ &= \int D_{n\theta_0}^2 d\mu + o_P(1/n) + 2(\theta - \theta_0)' \int \Delta (D_{n\theta} - D_\theta) d\mu \\ &\quad + o_P(n^{-1/2} \|\theta - \theta_0\|) \\ &= M_n(\theta_0) + 2(\theta - \theta_0)' \int \Delta D_{n\theta_0} d\mu + o_P(n^{-1/2} \|\theta - \theta_0\| + 1/n). \end{aligned}$$

In the above calculations we used that

$$\begin{aligned} \int (D_{n\theta} - D_\theta)^2 d\mu &= \int D_{n\theta_0}^2 d\mu + \int (D_{n\theta} - D_\theta - D_{n\theta_0})^2 d\mu \\ &\quad + 2 \int D_{n\theta_0} (D_{n\theta} - D_\theta - D_{n\theta_0}) d\mu \\ &\equiv I + II + III, \end{aligned}$$

where $I = M_n(\theta_0)$ by definition, $II = o_P(1/n)$ for $\theta \rightarrow_P \theta_0$ by Lemma 3 above and the continuous mapping theorem (cf. van der Vaart and Wellner (1996, Theorem 1.3.6, p. 20)), and finally

$$\begin{aligned} III &\leq 2 \left(\int D_{n\theta_0}^2 d\mu \right)^{1/2} \left(\int (D_{n\theta} - D_\theta - D_{n\theta_0})^2 d\mu \right)^{1/2} \\ &= 2o_P(n^{-1/2}) \cdot o_P(n^{-1/2}) = o_P(1/n). \end{aligned}$$

Also,

$$\begin{aligned} \int D_\theta(D_{n\theta} - D_\theta) d\mu &= (\theta - \theta_0)' \int \Delta(D_{n\theta} - D_\theta) d\mu + \int r_\theta(D_{n\theta} - D_\theta) d\mu \\ &= (\theta - \theta_0)' \int \Delta(D_{n\theta} - D_\theta) d\mu + o_P(n^{-1/2} \|\theta - \theta_0\|) \end{aligned}$$

as for all $\theta \rightarrow_P \theta_0$,

$$\begin{aligned} \left| \int r_\theta(D_{n\theta} - D_\theta) d\mu \right| &\leq \left(\int r_\theta^2 d\mu \right)^{1/2} \cdot \left(\int (D_{n\theta} - D_\theta)^2 d\mu \right)^{1/2} \\ &= o_P(n^{-1/2} \|\theta - \theta_0\|). \end{aligned}$$

The other calculations are quite similar and have been omitted for this reason. Thus the conditions of Theorem 3.2 in Wegkamp (1999) are met, and consequently

$$\begin{aligned} \hat{\theta} &= \theta_0 - 2V^{-1} \cdot \left(\int \Delta(z) D_{n, \theta_0}(z) d\mu(z) \right) + o_P(n^{-1/2}) \\ &\equiv \theta_0 - 2V^{-1} \Gamma_n + o_P(n^{-1/2}) \end{aligned}$$

holds true. The independence between $\varepsilon = \rho(Z, \theta_0)$ and X and Fubini's theorem imply that $\mathbb{E}\Gamma_n = 0$. Writing $H \equiv H_{\theta_0}$, $G \equiv G_{\theta_0}$, $H_n \equiv H_{n\theta_0}$, and $G_n \equiv G_{n\theta_0}$, the covariance term of the leading linear term equals

$$\begin{aligned} D(\Gamma_n) &= \mathbb{E}\Gamma_n \Gamma_n' = \mathbb{E} \left(\int \Delta(z) D_{n\theta_0}(z) \right) \left(\int \Delta(\bar{z}) D_{n\theta_0}(\bar{z}) \right)' \\ &= \iint \Delta(z) \Delta'(\bar{z}) \mathbb{E} D_n(Z) D_n(\bar{z}) d\mu(z) d\mu(\bar{z}) \\ &= \iint \Delta(z) \Delta'(\bar{z}) \cdot \mathbb{E}[(H_n - H)(z)(H_n - H)(\bar{z}) \\ &\quad + (H - F_n G_n)(z)(H - F_n G_n)(\bar{z}) + (H_n - H)(z)(H - F_n G_n)(\bar{z}) \\ &\quad + (H_n - H)(\bar{z})(H - F_n G_n)(z)] d\mu(z) d\mu(\bar{z}). \end{aligned}$$

A tedious, but straightforward calculation further reveals that

$$\begin{aligned} D(\Gamma_n) &= \frac{1}{n} \iint \Delta(x, \varepsilon) \Delta'(\bar{x}, \bar{\varepsilon}) [F(x)F(\bar{x})G(\min(\varepsilon, \bar{\varepsilon})) \\ &\quad + F(\min(x, \bar{x}))G(\varepsilon)G(\bar{\varepsilon}) + H(\min(x, \bar{x}), \min(\varepsilon, \bar{\varepsilon})) \\ &\quad - 3H(x, \varepsilon)H(\bar{x}, \bar{\varepsilon})] d\mu(x, \varepsilon) d\mu(\bar{x}, \bar{\varepsilon}) + o\left(\frac{1}{n}\right) \\ &\equiv \frac{1}{n} W + o\left(\frac{1}{n}\right). \end{aligned}$$

In view of the preceding stochastic expansion of $\hat{\theta}$ and since $\sqrt{n}(H_n - H)(x, \varepsilon)$, $\sqrt{n}(G_n - G)(\varepsilon)$, and $\sqrt{n}(F_n - F)(x)$ all converge to Gaussian processes, $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution to $\mathcal{N}(0, 4V^{-1}WV^{-1})$, by an application of Donsker's theorem, the continuous mapping theorem, and Slutsky's lemma. Q.E.D.

Before proving Theorem 6, we first establish some auxiliary results, to wit, the bootstrap counterparts of Theorem 4 and Lemma 3.

LEMMA 4: *Under the same assumptions as in Theorem 4, $\hat{\theta}^* \rightarrow_{\text{a.s.}} \theta_0$ for almost all samples Z_1, \dots, Z_n .*

PROOF: By the triangle inequality, for almost all samples Z_1, \dots, Z_n , we have

$$\begin{aligned} & \sup_{\theta} |(M_n^* - M)(\theta)| \\ & \leq \sup_{\theta} |(M_n^* - M_n)(\theta)| + \sup_{\theta} |(M_n - M)(\theta)| \\ & \leq 4\mu(\mathcal{X}) \sup_{\theta, x, \varepsilon} |(D_{n, \theta}^* - D_{n, \theta})(x, \varepsilon)| + 4\mu(\mathcal{X}) \sup_{\theta, x, \varepsilon} |(D_{n, \theta} - D_{\theta})(x, \varepsilon)| \\ & \xrightarrow{\text{a.s.}} 0, \end{aligned}$$

since \mathcal{A} , \mathcal{B} , and \mathcal{C} are P -Donsker classes. The remainder of the proof goes as the one for Theorem 4 and has therefore been omitted. Q.E.D.

LEMMA 5: *Assume Assumptions A.3, A.4, and A.5. Then the process $\sqrt{n}(D_{n\theta}^* - D_{n\theta})(x, \varepsilon)$, is stochastically equicontinuous at θ_0 with respect to the Euclidean distance on Θ for all $x \in \mathcal{X}$ and $\varepsilon \in \mathcal{Y}$, conditionally given Z_1, \dots, Z_n .*

PROOF: Giné and Zinn (1990) proved that the empirical process $\sqrt{n}(P_n - P)$ can be bootstrapped if and only if the class of functions that index the process is P -Donsker. Therefore, as a consequence of the Donsker property of \mathcal{A} , \mathcal{B} , and \mathcal{C} ,

$$\sqrt{n}(D_{n\theta}^* - D_{n\theta})(x, \varepsilon) - \sqrt{n}(D_{n\theta} - D_{\theta})(x, \varepsilon) \xrightarrow{P} 0,$$

and the desired result follows from Lemma 3.

Q.E.D.

PROOF OF THEOREM 6: The proof closely follows the arguments for M -estimators obtained by Arcones and Giné (1992). Observe that by similar arguments given in the proof of Theorem 5, for all $\theta \rightarrow_P \theta_0$

$$\begin{aligned} & M_n^*(\theta) - M_n(\theta) \\ & = \int (D_{n\theta}^* - D_{n\theta})^2 d\mu + 2 \int D_{n\theta} (D_{n\theta}^* - D_{n\theta}) d\mu \\ & = \int (D_{n\theta_0}^* - D_{n\theta_0})^2 d\mu + o_P(1/n) + 2 \int D_{\theta} (D_{n\theta}^* - D_{n\theta}) d\mu \\ & \quad + 2 \int (D_{n\theta} - D_{\theta})(D_{n\theta}^* - D_{n\theta}) d\mu \end{aligned}$$

$$\begin{aligned}
 &= \int (D_{n\theta_0}^* - D_{n\theta_0})^2 d\mu + 2 \int D_\theta (D_{n\theta}^* - D_{n\theta}) d\mu \\
 &\quad + 2 \int D_{n\theta_0} (D_{n\theta_0}^* - D_{n\theta_0}) d\mu + o_P(1/n) \\
 &= \int (D_{n\theta_0}^* - D_{n\theta_0})^2 d\mu + 2 \int D_{n\theta_0} (D_{n\theta_0}^* - D_{n\theta_0}) d\mu \\
 &\quad + 2(\theta - \theta_0)' \int \Delta (D_{n\theta}^* - D_{n\theta}) d\mu + o_P(n^{-1/2} \|\theta - \theta_0\| + n^{-1}).
 \end{aligned}$$

Consequently, for $\theta \rightarrow_P \theta_0$ and $\eta \rightarrow_P \theta_0$,

$$\begin{aligned}
 &M_n^*(\theta) - M_n^*(\eta) \\
 &= [(M_n^* - M_n)(\theta) - (M_n^* - M_n)(\eta)] + [(M_n - M)(\theta) - (M_n - M)(\eta)] \\
 &\quad + [M(\theta) - M(\eta)] \\
 &= 2(\theta - \eta)' \int \Delta [(D_{n\theta_0}^* - D_{n\theta_0}) + (D_{n\theta_0} - D_{\theta_0})] d\mu \\
 &\quad + \frac{1}{2}(\theta - \theta_0)' V(\theta - \theta_0) - \frac{1}{2}(\eta - \theta_0)' V(\eta - \theta_0) \\
 &\quad + o_P(\|\theta - \theta_0\|^2 + \|\eta - \theta_0\|^2 + n^{-1/2} \|\theta - \theta_0\| + n^{-1/2} \|\eta - \theta_0\| + n^{-1}).
 \end{aligned}$$

We define

$$\Delta_n = 2 \int \Delta (D_{n\theta_0} - D_{\theta_0}) d\mu \quad \text{and} \quad \Delta_n^* = 2 \int \Delta (D_{n\theta_0}^* - D_{n\theta_0}) d\mu$$

and we take $\theta = \hat{\theta}$ and $\eta = \theta_0 - (\Delta_n + \Delta_n^*)$. Observe that $\eta \in \Theta$ for n sufficiently large, as θ_0 is an interior point of Θ . To simplify matters, we assume without loss of generality that $V = I$. Hence

$$\begin{aligned}
 &M_n^*(\theta) - M_n^*(\eta) \\
 &= (\theta - \eta)' (\Delta_n^* + \Delta) + \frac{1}{2} \|\theta - \theta_0\|^2 - \frac{1}{2} \|\eta - \theta_0\|^2 \\
 &\quad + o_P(\|\theta - \theta_0\|^2 + \|\eta - \theta_0\|^2 + n^{-1/2} \|\theta - \theta_0\| + n^{-1/2} \|\eta - \theta_0\| + n^{-1})
 \end{aligned}$$

and

$$\begin{aligned}
 0 &\geq M_n^*(\hat{\theta}^*) - M_n^*(\theta_0 - (\Delta_n + \Delta_n^*)) \\
 &= (\hat{\theta}^* - \theta_0)' (\Delta_n^* + \Delta_n) + \frac{1}{2} \|\Delta_n + \Delta_n^*\|^2 + \frac{1}{2} \|\hat{\theta}^* - \theta_0\|^2 - \frac{1}{2} \|\Delta_n^* + \Delta_n\|^2 \\
 &\quad + o_P(\|\hat{\theta}^* - \theta_0\|^2 + \|\Delta_n + \Delta_n^*\|^2 + n^{-1/2} \|\hat{\theta}^* - \theta_0\| + n^{-1/2} \|\Delta_n + \Delta_n^*\| + n^{-1}) \\
 &= \frac{1}{2} \|\hat{\theta}^* - \theta_0 + (\Delta_n^* + \Delta_n)\|^2 \\
 &\quad + o_P(\|\hat{\theta}^* - \theta_0\|^2 + n^{-1/2} \|\hat{\theta}^* - \theta_0\| + \|\Delta_n + \Delta_n^*\|^2 + n^{-1/2} \|\Delta_n + \Delta_n^*\| + n^{-1}),
 \end{aligned}$$

whence

$$n\|\hat{\theta}^* - \theta_0 + (\Delta_n^* + \Delta_n)\|^2 \rightarrow 0$$

in P_n -probability. By the preceding theorem,

$$\hat{\theta} - \theta_0 = -\Delta_n + o_P(n^{-1/2}),$$

so that combination yields $\hat{\theta}^* - \hat{\theta} = -\Delta_n^* + o_P(n^{-1/2})$. The term Δ_n^* has the same limiting distribution as Δ_n by the bootstrap theorem for the mean in \mathbb{R}^d . This concludes the proof. *Q.E.D.*

We now turn to the proof of Theorem 7. Again we set out with the technical lemma's first, concerning uniform integrability of $\|\sqrt{n}(\hat{\theta} - \theta_0)\|^2$.

LEMMA 6: *If Assumption A.3 holds, we have for all $k > 0$*

$$\mathbb{E}\left(\sup_{x, \varepsilon, \theta} |\sqrt{n}(D_{n\theta}(x, \varepsilon) - D_\theta(x, \varepsilon))|\right)^k < \infty,$$

and

$$\mathbb{E}^*\left(\sup_{x, \varepsilon, \theta} |\sqrt{n}(D_{n\theta}^*(x, \varepsilon) - D_{n\theta}(x, \varepsilon))|\right)^k < \infty \quad \text{a.s.}$$

PROOF: First notice that

$$\begin{aligned} &\mathbb{E}\left(\sup_{x, \varepsilon, \theta} |\sqrt{n}(D_{n\theta}(x, \varepsilon) - D_\theta(x, \varepsilon))|\right)^k \\ &\leq C_k \left\{ \mathbb{E} \sup_{A \in \mathcal{A}} |\sqrt{n}(P_n - P)A|^k + \mathbb{E} \sup_{B \in \mathcal{B}} |\sqrt{n}(P_n - P)B|^k \right. \\ &\quad \left. + \mathbb{E} \sup_{C \in \mathcal{C}} |\sqrt{n}(P_n - P)C|^k \right\}. \end{aligned}$$

If $\{\rho(\cdot, \theta) : \theta \in \Theta\}$ is a subset of a finite dimensional vector space, an application of Theorem 2.14.1 in van der Vaart and Wellner (1996, p. 237) yields

$$\left(\mathbb{E} \sup_{A \in \mathcal{A}} |\sqrt{n}(P_n - P)A|^k\right)^{1/k} \leq C \sup_{Q \text{ discrete}} \int_0^1 \sqrt{1 + \log N(\delta, \mathcal{A}, L_2(Q))} d\delta.$$

The right-hand side is finite by the bounds obtained in Lemma 1. The same applies for \mathcal{B} and \mathcal{C} , and combination of the previous two displays establishes the first part for finite dimensional spaces. The bootstrap counterpart follows by the same argument. For the case of smooth functions we do not have a uniform bound for the covering numbers, but a bound on the bracketing numbers instead. Another difference is that we needed to assume the existence of a bounded probability density for $H(x, \varepsilon)$. For this case, Theorem 2.14.5 and Theorem 2.12.2 in van der Vaart and Wellner (1996, pp. 244, 240)

yield respectively

$$\begin{aligned} & \left\{ \mathbb{E} \left(\sup_{x, \varepsilon, \theta} |\sqrt{n}(D_{n\theta}(x, \varepsilon) - D_\theta(x, \varepsilon))| \right)^k \right\}^{1/k} \\ & \leq C \mathbb{E} \sup_{A \in \mathcal{A}} |\sqrt{n}(P_n - P)A| + Cn^{-\frac{1}{2} + \frac{1}{k}} \quad \text{for } k \geq 2 \\ & \leq C \int_0^1 \sqrt{1 + \log N_B(\delta, \mathcal{A}, L_2(P))} d\delta + Cn^{-\frac{1}{2} + \frac{1}{k}}. \end{aligned}$$

The bound on the bracketing numbers in Lemma 2 shows that the right-hand side is finite. The same is true of course for the classes \mathcal{B} and \mathcal{C} , and the first claim follows for the case of smooth functions. Also, by the same reasoning, for $k \geq 2$,

$$\begin{aligned} & \left\{ \mathbb{E} \left(\sup_{x, \varepsilon, \theta} |\sqrt{n}(D_{n\theta}^*(x, \varepsilon) - D_{n\theta}(x, \varepsilon))| \right)^k \right\}^{1/k} \\ & \leq C \int_0^1 \sqrt{1 + \log N_B(\delta, \mathcal{A}, L_2(P_n))} d\delta + Cn^{-\frac{1}{2} + \frac{1}{k}} \\ & \leq C \int_0^1 \sqrt{1 + \log N_B(\delta/2, \mathcal{A}, L_2(P))} d\delta + Cn^{-\frac{1}{2} + \frac{1}{k}} \quad \text{a.s.,} \end{aligned}$$

where we used the uniform law of large numbers in the last line. This completes our proof. *Q.E.D.*

LEMMA 7: *Under Assumptions A.1, A.2, A.3, A.6, and A.7, $\|\sqrt{n}(\hat{\theta} - \theta)\|$ is uniformly square integrable.*

PROOF: It suffices to show that $\mathbb{E}\|\sqrt{n}(\hat{\theta} - \theta_0)\|^3 < \infty$. First, observe that by Assumption A.7, there exist $\delta > 0$ and $c > 0$ such that for all $\|\theta - \theta_0\| < \delta$,

$$c\|\theta - \theta_0\|^2 \leq M(\theta) - M(\theta_0) = M(\theta).$$

We find

$$\begin{aligned} \mathbb{E}\|\hat{\theta} - \theta_0\|^3 &= \mathbb{E}\|\hat{\theta} - \theta_0\|^3 \{ \|\hat{\theta} - \theta_0\| < \delta \} + \mathbb{E}\|\hat{\theta} - \theta_0\|^3 \{ \|\hat{\theta} - \theta_0\| \geq \delta \} \\ &\leq C(\mathbb{E}M^{3/2}(\hat{\theta}) + \mathbb{E}\{ \|\hat{\theta} - \theta_0\| \geq \delta \}). \end{aligned}$$

The constant $C > 0$ is a generic constant independent of n . In the last line we invoked Assumption A.1 as well. We will bound the two terms on the right-hand side separately. Observe that for any fixed $\delta > 0$ (not depending on n), there exists an $\eta > 0$ such that $\|\theta - \theta_0\| \geq \delta$ implies that $M(\theta) - M(\theta_0) > \eta$. This is a consequence of θ_0 being a well separated minimum of M , which follows from Assumptions A1, A2, and A7 (cf. the proof of Theorem 4). Therefore

$$\begin{aligned} \mathbb{P}\{ \|\hat{\theta} - \theta_0\| \geq \delta \} &\leq \mathbb{P}\{ M(\hat{\theta}) - M(\theta_0) \geq \eta \} \\ &= \mathbb{P}\{ M(\hat{\theta}) - M_n(\hat{\theta}) + M_n(\hat{\theta}) - M(\theta_0) \geq \eta \} \\ &\leq \mathbb{P}\left\{ 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \geq \eta \right\} \\ &\leq (\eta/2)^{-\alpha} \mathbb{E} \sup_{x, \varepsilon, \theta} |D_\theta(x, \varepsilon) - D_{n\theta}(x, \varepsilon)|^\alpha = \mathcal{O}(n^{-\alpha/2}) \end{aligned}$$

for any $\alpha > 2$ by Lemma 6. The other term can be handled as follows:

$$\begin{aligned} \mathbb{E}M^{3/2}(\hat{\theta}) &= \mathbb{E}((M - M_n)(\hat{\theta}) + M_n(\hat{\theta}))^{3/2} \\ &\leq \mathbb{E}(|(M_n - M)(\hat{\theta})| + |M_n(\theta_0)|)^{3/2} \\ &\leq \mathbb{E}\left(\int (D_{n\hat{\theta}} - D_{\hat{\theta}})^2 d\mu + 2 \int |D_{\hat{\theta}}\|D_{n\hat{\theta}} - D_{\hat{\theta}}| d\mu + M_n(\theta_0)\right)^{3/2} \\ &\leq \mathbb{E}\left(2C \sup_{x, \varepsilon, \theta} (D_{n\theta} - D_{\theta})^2(x, \varepsilon) + C\|\hat{\theta} - \theta_0\| \sup_{x, \varepsilon, \theta} |(D_{n\theta} - D_{\theta})(x, \varepsilon)|\right)^{3/2} \\ &\leq C\mathbb{E} \sup_{x, \varepsilon, \theta} |(D_{n\theta} - D_{\theta})(x, \varepsilon)|^3 \\ &\quad + C(\mathbb{E}\|\hat{\theta} - \theta_0\|^3)^{1/2} \left(\mathbb{E} \sup_{x, \varepsilon, \theta} |D_{n\theta}(x, \varepsilon) - D_{\theta}(x, \varepsilon)|^3\right)^{1/2} \\ &= \mathcal{O}(n^{-3/2}) + \mathcal{O}(n^{-3/4})(\mathbb{E}\|\hat{\theta} - \theta_0\|^3)^{1/2}. \end{aligned}$$

Combining all three preceding displays, we obtain that

$$\mathbb{E}\|\hat{\theta} - \theta_0\|^3 \leq \mathcal{O}(n^{-3/2}) + \mathcal{O}(n^{-3/4})(\mathbb{E}\|\hat{\theta} - \theta_0\|^3)^{1/2}.$$

Since $\mathbb{E}\|\hat{\theta} - \theta_0\|^3 < \infty$ by Assumption A.1, the conclusion follows. *Q.E.D.*

PROOF OF THEOREM 7: It is shown in Shao and Tu (1995, Theorem 2.10, p. 52), that the stochastic expansion

$$c'\hat{\theta} = c'\theta_0 - 2V^{-1} \int c'\Delta D_{n\theta_0} d\mu + o_p(1/\sqrt{n})$$

and the uniform integrability of $\|\sqrt{nc'}(\hat{\theta} - \theta_0)\|^2$ imply the consistency of J_{-d}^2 , provided the tuning parameter d satisfies

$$d/n \geq \varepsilon \quad \text{for some } \varepsilon > 0 \quad \text{and} \quad n - d \rightarrow \infty.$$

The weak convergence result Theorem 6 and uniform square integrability of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ imply that the bootstrap provides a consistent alternative for estimating the variance of a linear combination of θ_0 . It remains to prove the uniform square integrability, which is immediate from Lemma 8 below. *Q.E.D.*

LEMMA 8: *Under Assumptions A.1, A.2, A.3, A.6, and A.7,*

$$\mathbb{E}^*\|\sqrt{n}(\hat{\theta}^* - \hat{\theta})\|^3 < \infty \quad \text{a.s.}$$

PROOF: As in the proof of Lemma 7, there exists $c = c(\theta_0, \delta) > 0$ such that

$$\begin{aligned} c\|\hat{\theta}^* - \theta_0\|^2 \{ \|\hat{\theta}^* - \theta_0\| \leq \delta \} \\ &\leq M(\hat{\theta}^*) - M_n(\hat{\theta}^*) + M_n(\hat{\theta}^*) - M_n^*(\hat{\theta}^*) + M_n^*(\hat{\theta}^*) \\ &\leq \{M(\hat{\theta}^*) - M_n(\hat{\theta}^*)\} + \{M_n(\hat{\theta}^*) - M_n^*(\hat{\theta}^*)\} \\ &\quad + \{M_n^*(\theta_0) - M_n(\theta_0)\} + M_n(\theta_0). \end{aligned}$$

Most terms can be handled as before in the proof of Lemma 7, and for the additional terms we invoke Lemma 6. Q.E.D.

Department of Economics, Yale University, 30 Hillhouse Avenue, New Haven, CT 06511, U.S.A.; donald.brown@yale.edu,

and

Department of Statistics, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511, U.S.A.; marten.wegkamp@yale.edu.

Manuscript received December, 2000; final revision received December, 2001.

REFERENCES

- ANDREWS, D. W. K. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics, Volume 4*, ed. by R. F. Engle and D. L. McFadden. Amsterdam: Elsevier, 2247–2294.
- (1999): "Estimation when a Parameter Is on the Boundary," *Econometrica*, 67, 1341–1384.
- ARCONES, M., AND E. GINÉ (1992): "On the Bootstrap of M -estimators and Other Statistical Functionals," in *Exploring the Limits of the Bootstrap*, ed. by R. LePage and L. Billard. New York: John Wiley and Son, 13–48.
- BROWN, B. W. (1983): "The Identification Problem in Systems Nonlinear in the Variables," *Econometrica*, 51, 175–196.
- BROWN, D. J., AND R. MATZKIN (1998): "Estimation of Nonparametric Functions in Simultaneous Equations Models, with an Application to Consumer Demand," Working paper, Yale University.
- BROWN, D. J., AND M. H. WEGKAMP (2000): "Asymptotics in Minimum Distance from Independence Estimation," Cowles Foundation Discussion Paper No. 1252.
- DE LA PEÑA, V., AND E. GINÉ (1999): *Decoupling: From Dependence to Independence*. New York: Springer-Verlag.
- DUDLEY, R. (1999): *Uniform Central Limit Theorems*. Cambridge: Cambridge University Press.
- GINÉ, E., AND J. ZINN (1990): "Bootstrapping General Empirical Measures," *Annals of Probability*, 18, 851–869.
- MANSKI, C. F. (1983): "Closest Empirical Distribution Estimation," *Econometrica*, 51, 305–320.
- MAS-COLELL, A. (1985): *The Theory of General Economic Equilibrium: A Differential Approach*. Cambridge: Cambridge University Press.
- NEWBY, W., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Volume 4*, ed. by R. F. Engle and D. L. McFadden. Amsterdam: Elsevier, 2111–2245.
- PAKES, A., AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.
- POLLARD, D. (1984): *Convergence of Empirical Processes*. New York: Springer-Verlag.
- (1985): "New Ways to Prove Central Limit Theorems," *Econometric Theory*, 1, 295–314.
- (forthcoming): *Asymptopia*.
- RANGA RAO, R. (1962): "Relations between Weak and Uniform Convergence of Measures with Applications," *Annals of Mathematical Statistics*, 33, 659–680.
- ROEHRIG, C. S. (1988): "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 56, 433–447.
- SHAO, J., AND D. TU (1995): *The Jackknife and Bootstrap*. New York: Springer.
- VAN DER VAART, A. (1998): *Asymptotic Statistics*. Cambridge: Cambridge University Press.
- VAN DER VAART, A., AND J. WELLNER (1996): *Weak Convergence and Empirical Processes*. New York: Springer.
- WEGKAMP, M. (1995): "Asymptotic Results for Parameter Estimation in General Empirical Processes," Technical Report TW9504, University of Leiden.
- (1999): *Entropy Methods in Statistical Estimation*. Amsterdam: CWI-tract 125.