Ann. Inst. Statist. Math.
Vol. 57, No. 2, 291–313 (2005)
©2005 The Institute of Statistical Mathematics

STRONG CONVERGENCE OF ESTIMATORS AS ε_n-MINIMISERS OF OPTIMISATION PROBLEMS*

PETR LACHOUT[1], ECKHARD LIEBSCHER[2] AND SILVIA VOGEL[2]

[1] *Department of Probability and Statistics, Charles University, Sokolovska 83, 18600 Prague, Czech Republic*
[2] *Institute of Mathematics, Technical University Ilmenau, 98684 Ilmenau/Thür, Germany*

(Received February 3, 2003; revised February 20, 2004)

Abstract. In the paper we prove strong consistency of estimators as solution of optimisation problems. The approach of the paper covers non-identifiable models, and models for dependent samples. We provide statements about consistency of M-estimators in regression models with random and with non-random design.

Key words and phrases: Strong convergence, M-estimators, epi-convergence, stochastic optimisation.

1. Introduction

The main goal of this paper is to provide a concept for proving strong consistency of estimators which are obtained as a solution of an optimisation problem or as an approximate solution of it. This concept uses ideas from stochastic optimisation and is based on the idea of epi-convergence. The application of the concept is demonstrated in the case of M-estimation.

In his pioneering work Wald (1949) proved consistency of maximum-likelihood estimators and had so an impact on later consistency proofs for parametric estimators. In the last decades three main techniques have been developed for proving strong consistency of estimators as solution of optimisation problems. The first one goes via uniform laws of large numbers (cf. Shorack and Wellner (1986), Pötscher and Prucha (1997)). The second technique uses a statement on convergence of convex functions (Andersen and Gill (1982)). The third one applies the idea of epi-convergence. Epi-convergence almost surely was considered by Salinetti and Wets (1986) and since then, it has been used in several further papers, for example Dupačová and Wets (1988), Artstein and Wets (1994), Korf and Wets (2001) etc. Strong consistency of estimators arising from optimisation problems was studied in papers by Pfanzagl (1969), and Dupačová and Wets (1988). Hess (1996) proved strong consistency of approximate maximum-likelihood estimators. Dudley (1998) showed strong consistency results under a bracketing condition.

Asymptotic normality of estimators coming from optimisation problems is examined in papers by Dupačova and Wets (1988), Shapiro (1989, 1991), Geyer (1994) and Pflug (1995). In settings which are different from ours, King and Rockafellar (1993) derived consistency results and statements on asymptotic normality.

*The research was partially supported by the Deutsche Forschungsgemeinschaft (project number 436TSE113/40) and by the Grant Agency of the Czech Republic under Grant 201/03/1027.

In this paper we employ the approach using epi-convergence. So we avoid deriving uniform strong laws of large numbers. Providing such laws could be a problem in situations of complicated dependence structures of the sample. We work with convergence properties of the objective function which are similar to but slightly weaker than epi-convergence almost surely. A general theorem on the strong convergence of minimizers of stochastic minimisation problems is the starting point of our concept. Subsequently, we provide sufficient conditions for the conditions of this general theorem. The advantages of our approach are the following:

(i) Our approach covers discontinuous and non-identifiable models, and cases where a unique optimum of the underlying limit problem does not exist.

(ii) The results can be applied to samples of independent random variables as well as to samples with complicated dependence structure.

(iii) The space of parameters is not restricted to $\mathbb{R}^d$, e.g. it could be a separable metric space of real functions.

So we obtain generalisations of statements of several earlier papers. Furthermore we consider approximate estimators which are defined similarly to Hess (1996) and Dudley (1998). Our definition includes the situation where an approximate value of the optimiser is supplied by a certain numerical estimation algorithm. This situation often occurs in applications.

The paper is organised as follows: In the first part of Section 2 we introduce some notions connected with epi-convergence almost surely and follow the presentations by Vogel (1994), and by Vogel and Lachout (2003*a*, 2003*b*). The main results of the present paper are given in the second part of Section 2.

In Sections 3 to 5 we consider several applications. Section 3 deals with M-estimation in random design regression models where the sample is a part of an ergodic sequence. In Section 4 it is discussed how we can get consistency even when the aspect of model selection is incorporated. The consistency of M-estimators in fixed design regression models is studied in Section 5.

2. General theory

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be the probability space and $\{f_n\}$ be a sequence of functions $f_n : \Xi \times \Omega \to \mathbb{R} \cup \{+\infty\}$ which can be regarded as a sequence of random functions $\{f_n(\cdot)\}$ where $f_n(x)$ stands for $f_n(x, \cdot)$. We equip the space Ξ with a metric d. In the sequel we give some definitions for the convergence of $\{f_n\}$ to a deterministic function. These definitions were taken from Vogel (1994) and Vogel and Lachout (2003*a*, 2003*b*). Moreover, we provide two useful lemmas.

Definition 2.1. The sequence $\{f_n\}$ of random functions is a lower semicontinuous approximation almost surely to $f : \Xi \to \mathbb{R} \cup \{+\infty\}$ on $\Theta \subset \Xi$ (symbol: $f_n \xrightarrow[\Theta]{l-a.s.} f$) if and only if for $\mathbb{P}$-almost all ω, all $x_0 \in \Theta$ and for all sequences $\{x_n\}$ tending to x_0,

$$\liminf_{n\to\infty} f_n(x_n, \omega) \geq f(x_0).$$

Lemma 2.1. *$f_n \xrightarrow[\Theta]{l-a.s.} f$ is equivalent to*

$$\mathbb{P}\left\{ \sup_{V \in \mathcal{N}(\theta_0)} \liminf_{n\to\infty} \inf_{t \in V} f_n(t) \geq f(\theta_0) \forall \theta_0 \in \Theta \right\} = 1. \tag{2.1}$$

$\mathcal{N}(\theta_0)$ is the system of neighbourhoods of θ_0.

DEFINITION 2.2. The sequence $\{f_n\}$ of random functions is an epi-upper approximation almost surely to $f : \Xi \to \mathbb{R} \cup \{+\infty\}$ on $\Theta \subset \Xi$ (symbol: $f_n \xrightarrow[\Theta]{epi-u-a.s.} f$) if and only if for $\mathbb{P}$-almost all ω and all $y_0 \in \Theta$, there is a sequence $\{y_n\}$ such that

$$y_n \to y_0 \quad \text{and} \quad \limsup_{n\to\infty} f_n(y_n, \omega) \le f(y_0).$$

LEMMA 2.2. *$f_n \xrightarrow[\Theta]{epi-u-a.s.} f$ is equivalent to*

$$\mathbb{P}\left\{\sup_{V\in\mathcal{N}(\theta_0)} \limsup_{n\to\infty} \inf_{t\in V} f_n(t) \le f(\theta_0) \forall \theta_0 \in \Theta\right\} = 1.$$

It is straightforward to establish an "upper" version of the first definition and an "epi-lower" version of the second one, but we need only the above definitions in our paper. Lemmas 2.1 and 2.2 can be found in similar versions in several papers (see Rockafellar and Wets (1998), p. 242). In the subsequent theorem, we employ the two above definitions for different sets. If the sequence $\{f_n\}$ satisfies both $f_n \xrightarrow[\Xi]{l-a.s.} f$ and $f_n \xrightarrow[\Xi]{epi-u-a.s.} f$ then it epi-converges almost surely (cf. Salinetti and Wets (1986), Hess (1996)), i.e. for all $\omega \in \bar{\Omega}$, the sequence of deterministic functions $\{f_n(\cdot, \omega)\}$ epi-converges where $\mathbb{P}(\bar{\Omega}) = 1$. Thus, to explain the relationship between epi-convergence almost surely and several kinds of convergence of sequences of random functions, it suffices to consider the relationship between the corresponding types of convergence for sequences of deterministic functions. Uniform convergence implies epi-convergence provided that the limit function is lower semicontinuous. Thus our concept uses weaker assumptions in comparison to the concept of uniform convergence. For a sequence of convex functions on $D \subset \mathbb{R}^n$, epi-convergence to a function f is equivalent to uniform convergence to f on compact sets in the interior of D (cf. Proposition 7.17 in Rockafellar and Wets (1998)). Concerning epi-convergence of sequences of deterministic functions, we refer to the monograph by Rockafellar and Wets ((1998), Chapter 7B). Several stochastic versions of epi-convergence are discussed in Salinetti and Wets (1986) and Pflug (2003).

Let $\hat{\theta}_n$ be an estimator for the unknown parameter $\theta_0 \in \Theta$ where Θ is the parameter set of some model. Assume that $\hat{\theta}_n$ is an ε_n-minimiser of f_n, i.e.

$$f_n(\hat{\theta}_n) \le m_n + \varepsilon_n \quad \text{for} \quad n \in \mathbb{N} \tag{2.2}$$

where $\{\varepsilon_n\}$ is a sequence of positive random numbers tending to zero almost surely and $m_n = \inf_{\theta\in\Theta} f_n(\theta)$. This definition includes many cases in applications where a numerical approximate minimiser $\hat{\theta}_n$ is computed by a numerical algorithm. An other argument for using approximate minimisers is that the infimum m_n need not be measurable. The following theorem provides some sufficient conditions for the consistency of the estimator. Let

$$d(x, A) = \inf\{d(x, y) : y \in A\} \quad \text{for} \quad x \in \mathbb{R}, \ A \subset \mathbb{R}.$$

THEOREM 2.1. *Suppose that (2.2) is satisfied, and either Θ is compact or there is a compact set $K \subset \Theta$, an $\alpha \in \mathbb{R}$, such that with probability one,*

$$\emptyset \neq \{x : f_n(x) \leq \alpha\} \subset K \quad \textit{for all} \quad n \geq n_0(\omega). \tag{2.3}$$

Furthermore, assume that there is a function $f : \Theta \longrightarrow \mathbb{R} \cup \{+\infty\}$ such that

$$f_n \xrightarrow[\Theta\setminus\Psi]{l-a.s.} f, \tag{2.4}$$

$$f_n \xrightarrow[\{\bar{\theta}\}]{epi-u-a.s.} f \quad \textit{for some} \quad \bar{\theta} \in \Psi := \operatorname*{argmin}_{\theta\in\Theta} f(\theta), \tag{2.5}$$

and $\Psi \neq \emptyset$.

(a) *Then*

$$\limsup_{n\to\infty} m_n \leq \min_{\theta\in\Theta} f(\theta) \quad a.s. \quad \textit{and} \tag{2.6}$$

$$\lim_{n\to\infty} d(\hat{\theta}_n, \Psi) = 0 \quad a.s. \tag{2.7}$$

(b) *Moreover, if in addition, $\Psi = \{\theta_0\}$ holds, then*

$$\lim_{n\to\infty} \hat{\theta}_n = \theta_0 \qquad a.s.$$

Remark 1. This theorem is closely related to Theorems 4.1 and 4.2 proved by Vogel (1994). We require the convergence of f_n on a smaller set here. A similar result where (2.4) and (2.5) are replaced by the more restrictive assumption of epi-convergence a.s., follows from Theorem 7.33 of Rockafellar and Wets (1998) as discussed in Korf and Wets ((2001), Sections 7, 8). Other related results are due to Robinson (1987).

Remark 2. We need the compactness assumption on Θ or the validity of (2.3) to ensure the existence of a convergent subsequence of $\{\hat{\theta}_n\}$. The existence of such a sequence is often explicitly assumed instead of compactness. Alternatively, one can work with other more complicated (but weaker) assumptions ensuring the existence of such a subsequence (see also King and Rockafellar (1993), p. 151).

Remark 3. Assumptions (2.4) and (2.5) concern a lower approximation outside the set of minimisers of f and an upper approximation on the set of minimisers of f for the sequence $\{f_n\}$, respectively.

Remark 4. From the proof below, it can be recognized that the following assertion is true: If $f_n \xrightarrow[\Theta]{l-a.s.} f$ is assumed instead of (2.4) and the other assumptions of Theorem 2.1 are satisfied, then

$$\lim_{n\to\infty} m_n = \min_{\theta\in\Theta} f(\theta) \qquad \text{a.s.}$$

PROOF. We assume Θ to be compact, since the case where (2.3) is satisfied can easily be transferred into the compact case. By virtue of (2.4) and (2.5), there is a set $\Omega^* \subset \Omega$, $\mathbb{P}(\Omega^*) = 1$ such that the following three conditions are fulfilled:

(i) $\varepsilon_n(\omega) \to 0$ for all $\omega \in \Omega^*$,

(ii) for all $\omega \in \Omega^*$, $x_0 \in \Theta\backslash\Psi$, and for all sequences $\{x_n\}$ tending to x_0,

$$\liminf_{n\to\infty} f_n(x_n,\omega) \geq f(x_0) \tag{2.8}$$

and

(iii) for every $\omega \in \Omega^*$, there is some sequence $\{y_n(\omega)\}$ with $y_n(\omega) \to \bar{\theta}$ and

$$\limsup_{n\to\infty} f_n(y_n(\omega),\omega) \leq f(\bar{\theta}). \tag{2.9}$$

Let us fix $\omega \in \Omega^*$ and write $f_n(\cdot)$ instead of $f_n(\cdot,\omega)$, and θ_n instead of $\hat{\theta}_n(\omega)$. Now we assume that there is a subsequence $\{\theta_{n_k}\}_{k=1,2,\ldots}$ of $\{\theta_n\}$ with $d(\theta_{n_k},\Psi) \to D > 0$ as $k \to \infty$. Since Θ is compact, there is a subsequence $\{\theta_{l_k}\}_{k=1,2,\ldots}$ of $\{\theta_{n_k}\}$ such that $\theta_{l_k} \to \theta^* \notin \Psi$ as $k \to \infty$ and by (2.2),

$$\liminf_{k\to\infty} \inf_{\theta\in\Theta} f_{n_k}(\theta) = \lim_{k\to\infty} f_{l_k}(\theta_{l_k}).$$

Therefore, by (2.2) and (2.8),

$$\liminf_{k\to\infty} m_{n_k} = \lim_{k\to\infty} f_{l_k}(\theta_{l_k}) \geq \liminf_{n\to\infty} f_n(\tilde{\theta}_n) \geq f(\theta^*) > m \tag{2.10}$$

where $m = \min_{\theta\in\Theta} f(\theta)$, $\tilde{\theta}_n = \theta_{l_k}$ if $n = l_k$ for some $k \in \mathbb{N}$, and $\tilde{\theta}_n = \theta^*$ otherwise. On the other hand, by (2.9),

$$\limsup_{n\to\infty} m_n \leq \limsup_{n\to\infty} f_n(y_n) \leq f(\bar{\theta}) = m$$

which contradicts (2.10). Hence the claim $d(\theta_{n_k},\Psi) \to D > 0$ is not true and (2.7) is satisfied. □

Obviously, condition $\limsup_{n\to\infty} f_n(\bar{\theta}) \leq f(\bar{\theta})$ a.s. is sufficient for (2.5). In her paper (1994) Vogel established Theorem 5.1(i) which gives sufficient conditions for the lower semicontinuous approximation almost surely in $\mathbb{R}^n$. We prove now a similar theorem for separable metric spaces Θ. Let $B(\theta,\rho) := \{t \in \Theta : d(t,\theta) < \rho\}$ denote the open ball around θ with radius ρ.

LEMMA 2.3. *Suppose that Θ is a separable metric space and f is lower semicontinuous on Θ. If for all $\theta \in \Theta$, $\varepsilon > 0$, there is some $\rho = \rho(\theta,\varepsilon) > 0$ such that*

$$\liminf_{n\to\infty} \inf_{t\in B(\theta,\rho)} f_n(t) \geq f(\theta) - \varepsilon \qquad a.s., \tag{2.11}$$

then $f_n \xrightarrow[\Theta]{l-a.s.} f$ holds true.

PROOF. In view of Lemma 2.1, we have to show that (2.1) holds true.

(i) First we construct an appropriate $\Omega^* \subset \Omega$. Let $\varepsilon = \frac{1}{m}$, $m \in \mathbb{N}$. According to the assumptions, for each $\theta \in \Theta$, there is an open ball $B(\theta,\rho(\theta,\varepsilon))$ such that (2.11) holds for $\omega \in \Omega_{\theta,m}$ and $\rho(\theta,\varepsilon) < \varepsilon$ with some set $\Omega_{\theta,m} \subset \Omega$, $\mathbb{P}(\Omega_{\theta,m}) = 1$. These balls

$B(\theta, \rho(\theta,\varepsilon))$ form an open cover of Θ which is a separable metric space. Therefore there is a countable open subcover of Θ consisting of sets $B(\theta, \rho(\theta,\varepsilon))$, $\theta \in \Gamma_m$. Now

$$\Omega^* = \bigcap_{m\in\mathbb{N}, \theta\in\Gamma_m} \Omega_{\theta,m}, \qquad \mathbb{P}(\Omega^*) = 1.$$

(ii) Now let $\theta_0 \in \Theta$ and $\varepsilon = \frac{1}{m}$, $m \in \mathbb{N}$ be arbitrary but fixed. Since f is lower semicontinuous, there exists a $\mu \in \mathbb{N}$, $\mu \geq m$ such that

$$f(t) \geq f(\theta_0) - \varepsilon \quad \text{for all} \quad t \in \Theta : d(t, \theta_0) \leq \frac{1}{\mu}.$$

There is some $\bar\theta \in \Gamma_\mu$ and some $\bar\rho \in (0, 1/\mu)$ such that $\theta_0 \in B(\bar\theta, \bar\rho)$ and by (2.11),

$$\liminf_{n\to\infty} \inf_{t\in B(\bar\theta,\bar\rho)} f_n(t,\omega) \geq f(\bar\theta) - \frac{1}{\mu} \quad \text{for all} \quad \omega \in \Omega^*.$$

Moreover, there is some $\tilde\rho > 0$ with $B(\theta_0, \tilde\rho) \subset B(\bar\theta, \bar\rho)$. Further

$$\begin{aligned}\sup_{V\in\mathcal{N}(\theta_0)} \liminf_{n\to\infty} \inf_{t\in V} f_n(t,\omega) &= \sup_{r>0} \liminf_{n\to\infty} \inf_{t\in B(\theta_0,r)} f_n(t,\omega) \\ &= \sup_{r:\tilde\rho>r>0} \liminf_{n\to\infty} \inf_{t\in B(\theta_0,r)} f_n(t,\omega) \\ &\geq \liminf_{n\to\infty} \inf_{t\in B(\bar\theta,\bar\rho)} f_n(t,\omega) \\ &\geq f(\bar\theta) - \frac{1}{\mu} \geq f(\theta_0) - \frac{2}{m}\end{aligned}$$

for all $\omega \in \Omega^*$. Consequently by $m \to \infty$, (2.1) is fulfilled. □

Let P_n and P be a random and a nonrandom measure on E, respectively. Now we turn to prove strong consistency in the special case where

$$f_n(t) = \int_E \varphi(t,x)\mathrm{d}P_n(x), \qquad f(t) = \int_E \varphi(t,x)\mathrm{d}P(x)$$

for $t \in \Theta$, and $\varphi : \Theta \times E \to \mathbb{R}$ is a measurable function. We assume that these Lebesgue integrals exist. The case where P_n is the empirical measure of a sample $X_1, \ldots, X_n$ is important for applications.

CONDITION $\mathcal{S}$. For every θ, function $\varphi(\cdot, x)$ is lower semicontinuous at θ for all $x \in E\backslash V_\theta$ where V_θ has P-measure zero. Further for all $\theta \in \Theta$,

$$\int_E \varphi(\theta,x)\mathrm{d}P(x) < +\infty \qquad \text{and}$$

for all $\theta \in \Theta$, there is some $\rho > 0$ such that

$$\int_E \inf_{t\in B(\theta,\rho)} \varphi(t,x)\mathrm{d}P(x) > -\infty. \tag{2.12}$$

The following Theorem 2.2 on strong consistency of $\hat{\theta}_n$ gives a generalisation of Theorem 3.9 in Dupačová and Wets (1988):

THEOREM 2.2. *Suppose that Θ is a compact metric space, and Conditions $\mathcal{S}$ and* (2.2) *are satisfied. Moreover, assume that for any $\theta \in \Theta\backslash\Psi$, $\rho > 0$, and some $\bar{\theta} \in \Psi$,*

$$\liminf_{n\to\infty} \int_E \inf_{t\in B(\theta,\rho)} \varphi(t,x)\mathrm{d}P_n(x) \geq \int_E \inf_{t\in B(\theta,\rho)} \varphi(t,x)\mathrm{d}P(x) \qquad a.s., \tag{2.13}$$

$$\limsup_{n\to\infty} \int_E \varphi(\bar{\theta},x)\mathrm{d}P_n(x) \leq \int_E \varphi(\bar{\theta},x)\mathrm{d}P(x) \qquad a.s., \tag{2.14}$$

Ψ as above. Then conclusions (a) *and* (b) *of Theorem* 2.1 *hold true.*

PROOF. Let $\theta^* \in \Theta\backslash\Psi, \varepsilon > 0$ be arbitrary but fixed. By (2.13),

$$\begin{aligned} \liminf_{n\to\infty} \inf_{t\in U_k} f_n(t) &\geq \liminf_{n\to\infty} \int_E \inf_{t\in U_k} \varphi(t,x)\mathrm{d}P_n(x) \\ &\geq \int_E \inf_{t\in U_k} \varphi(t,x)\mathrm{d}P(x) \qquad \text{a.s.} \end{aligned} \tag{2.15}$$

for $k \in \mathbb{N}$, $U_k = B(\theta^*, 1/k)$. By Condition $\mathcal{S}$, for any $x \in E\backslash V_{\theta^*}$,

$$\liminf_{n\to\infty} \inf_{t\in U_n} \varphi(t,x) \geq \varphi(\theta^*,x).$$

Hence by (2.12) and Fatou's lemma, there is some m such that

$$\int_E \inf_{t\in U_m} \varphi(t,x)\mathrm{d}P(x) \geq \int_E \varphi(\theta^*,x)\mathrm{d}P(x) - \varepsilon. \tag{2.16}$$

(2.15) and (2.16) imply (2.11). By Condition $\mathcal{S}$, f is lower semicontinuous, and $\Psi \neq \emptyset$ since Θ is compact. Therefore applying Lemma 2.3, it follows that (2.4) is fulfilled. Furthermore, (2.14) implies (2.5). Now the theorem is a consequence of Theorem 2.1. □

The following example shows that the described technique works even in such cases where P_n is not the usual empirical measure. Moreover, an uniform convergence technique is not available in this case.

Example 1. We consider the model of right censoring. Let $X_1, X_2, \dots$ and Y_1, $Y_2, \dots$ be two independent sequences of i.i.d. random variables. X_i and Y_i are the lifetime and the censoring time of the i-th sample item. We denote the distribution functions of X_i and Y_i by F and G, respectively. The variables X_i and Y_i are not explicitly given. We only observe pairs $(Z_1,\delta_1),\dots,(Z_n,\delta_n)$ where $Z_i = \min\{X_i,Y_i\}$, $\delta_i = I(X_i \leq Y_i)$. $Z_{(1)} \leq Z_{(2)} \leq \dots \leq Z_{(n)}$ denote the order statistics of $Z_1,\dots,Z_n$, and $\delta_{(i)}$ is the concomitant of $Z_{(i)}$. We introduce the Kaplan-Meier distribution function F_n:

$$\hat{F}_n(x) = 1 - \prod_{i=1}^{n} \left(1 - \frac{\delta_{(i)}}{n-i+1}\right)^{I(Z_{(i)}\leq x)}.$$

P_n is the corresponding measure on $\mathbb{R}$. We assume that the function F belongs to a parametric family $\{F(\cdot \mid \theta)\}_{\theta\in\Theta}$ of distributions and $F(\cdot \mid \theta_0) = F$. Let us study the M-estimation problem

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta\in\Theta} f_n(\theta), \qquad f_n(\theta) := \int_{-\infty}^{\infty} \varphi(\theta, x)\mathrm{d}\hat{F}_n(x)$$

with a function $\varphi : \Theta \times \mathbb{R} \to [0, +\infty)$ satisfying

$$\int_{-\infty}^{\infty} \inf_{t\in B(\theta,\rho)} \varphi(t, x)\mathrm{d}F(x) < \infty$$

for some ball $B(\theta, \rho)$ and every $\theta \in \Theta$. Further assume that for every θ, the function $\rho(x, \cdot)$ is lower semicontinuous at θ for all x except for a set V_θ of P-measure zero, P is the corresponding measure to F. Suppose that (a) $\tau_F < \tau_G$ or (b) $\tau_F = \tau_G$ and F is continuous at τ_F where $\tau_L = \inf\{x : L(x) = 1\}$. Hence by Lemma 1 of Wang (1995),

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \inf_{t\in B(\theta,\rho)} \varphi(t, x)\mathrm{d}\hat{F}_n(x) = \int_{-\infty}^{\infty} \inf_{t\in B(\theta,\rho)} \varphi(t, x)\mathrm{d}F(x) \qquad \text{a.s.}$$

Thus (2.13) is satisfied. In the same way one shows the validity of (2.14). In view of Theorem 2.2, we obtain

$$\lim_{n\to\infty} d(\hat{\theta}_n, \Psi) = 0 \quad \text{a.s.}, \qquad \text{where} \quad \Psi := \operatorname*{argmin}_{\theta\in\Theta} \int_{-\infty}^{\infty} \varphi(\theta, x)\mathrm{d}F(x).$$

This statement can be regarded as a generalized version of Theorem 1 in Wang (1995).

Let $\{X_n\}$ be a stationary sequence of $\mathbb{R}^m$-valued random variables, and P_n be the empirical measure of the sample $X_1, \ldots, X_n$ such that

$$f_n(t) = \frac{1}{n} \sum_{i=1}^{n} \varphi(t, X_i).$$

Using the notion of ergodicity for stationary sequences and applying the strong law of large numbers, we obtain the following corollary. Here we take into account that P is the stationary measure and $f(t) = \int \varphi(t, x)\mathrm{d}P(x)$ $(t \in \Theta)$.

COROLLARY 2.1. *Assume that $\{X_n\}$ is ergodic (in the sense of ergodicity of stationary processes), Θ is a compact metric space, and Conditions $\mathcal{S}$ and (2.2) are satisfied.*

(a) *Then*

$$\lim_{n\to\infty} d\left(\hat{\theta}_n, \operatorname*{argmin}_{\theta\in\Theta} f(\theta)\right) = 0 \qquad a.s.$$

(b) *Moreover, if in addition, $f(\theta) > f(\theta_0)$ holds for all $\theta \in \Theta\backslash\{\theta_0\}$, then*

$$\lim_{n\to\infty} \hat{\theta}_n = \theta_0 \qquad a.s.$$

In her paper Kaňková (1978) treated the special case where φ is continuous and $\varphi(\theta, \cdot)$ is concave. Considering the case of samples of i.i.d. random variables, part (b) of this corollary is essentially the same as Theorem 1 of Wang (1995) and similar to Theorem 1.12 by Pfanzagl (1969).

3. M-estimators in regression models with random design

Here we consider the regression model

$$Y_i = g(X_i \mid \theta_0) + Z_i \qquad (i = 1, 2, \dots) \tag{3.1}$$

where $\{Z_k\}_{k=1,2,\dots}$ and $\{X_k\}_{k=1,2,\dots}$ are two sequences of real respective $\mathbb{R}^m$-valued random variables such that X_i and Z_i are independent for each i. Let $g : \mathbb{R}^m \times \Theta \to \mathbb{R}$, $\Theta \subset \mathbb{R}^p$ be a measurable function. $\theta_0 \in \Theta$ is the true parameter of the model. Let

$$f_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i - g(X_i \mid \theta))$$

with a nonnegative continuous function ρ, and let $\hat{\theta}_n$ be an estimator for θ_0 satisfying

$$f_n(\hat{\theta}_n) \le \inf_{\theta \in \Theta} f_n(\theta) + \varepsilon_n \quad \text{and} \quad \varepsilon_n \ge 0,\ \varepsilon_n \to 0 \quad \text{a.s.} \tag{3.2}$$

This estimator $\hat{\theta}_n$ is called an *ε_n-approximate M-estimator* for θ_0. One special case is given by the ε_n-approximate maximum likelihood estimator which was examined in Hess (1996). In the sequel we use the notation

$$\mathbb{E}_P H(Y, X) = \int_{\mathbb{R}^{m+1}} H(y, x) \mathrm{d}P(y, x),$$

where X and Y are random variables, and P is the distribution of (Y, X). Let P_X denote the distribution of X. An application of Theorem 2.2 leads to the following statement.

THEOREM 3.1. *Suppose that there is a distribution measure P on $\mathbb{R}^{m+1}$ such that*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} h(Y_i, X_i) = \int_{\mathbb{R}^{m+1}} h(y, x) \mathrm{d}P(y, x) \qquad a.s. \tag{3.3}$$

for any function $h : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$ with $\int_{\mathbb{R}^{m+1}} |h(y, x)| \mathrm{d}P(y, x) < \infty$. Let X and Y be random variables having joint distribution P, and $Z := Y - g(X \mid \theta_0)$. Moreover, assume that Θ is a compact metric space, $\mathbb{E}_P \rho(Y - g(X \mid \theta)) < +\infty$ for all $\theta \in \Theta$, and for every $\theta \in \Theta$ and P_X-almost all ξ, $g(\xi \mid \cdot)$ is continuous at θ. Suppose that $\theta_0 \in \Psi$ where $\Psi = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}_P \rho(Y - g(X \mid \theta))$.

(a) *Then*

$$\lim_{n \to \infty} d(\hat{\theta}_n, \Psi) = 0 \qquad a.s. \tag{3.4}$$

(b) *If in addition,*

$$\mathbb{E}_P \rho(Z + a) > \mathbb{E}_P \rho(Z) \qquad \text{for all} \quad a \ne 0, \tag{3.5}$$

and

$$\mathbb{P}_P\{g(X \mid \theta_0) \ne g(X \mid \theta)\} > 0 \qquad \text{for all} \quad \theta \in \Theta,\ \theta \ne \theta_0 \tag{3.6}$$

are satisfied, then

$$\lim_{n \to \infty} \hat{\theta}_n = \theta_0 \qquad a.s.$$

PROOF. Let

$$\varphi(\theta,(x_1,x_2)^T)=\rho(x_1-g(x_2\mid\theta)) \quad \text{and} \quad f(\theta)=\mathbb{E}_P\rho(Y-g(X\mid\theta)).$$

It can easily be shown that (3.5) and (3.6) imply $f(\theta)>f(\theta_0)$ for all $\theta\in\Theta\backslash\{\theta_0\}$ and thus $\Psi=\{\theta_0\}$ (cf. Lemma 5 of Berlinet *et al.* (2000)). Now Theorem 3.1 is a consequence of Theorem 2.2. □

Condition (3.3) represents a type of ergodicity assumption on $\{(Y_k,X_k)\}$. If $\{(Y_k,X_k)\}$ is stationary and ergodic then (3.3) is fulfilled. Condition (3.5) can be regarded as a contrast condition and (3.6) ensures the identifiability of the parameter θ_0. Theorem 3.1 may be immediately applied to nonlinear autoregressive models which are covered by model (3.1) (cf. Liebscher (2003)). In the i.i.d. case convergence rates a.s. of M-estimates are examined in the paper by Arcones (1994).

Let us consider a couple of examples which illustrate the applicability of the Theorem 3.1 in different settings, especially in cases where common approaches fail. Example 2 deals with the situation where the sample is not necessarily stationary, g has a discontinuity point and the set Ψ of minimisers consists of more than one point (model is not identifiable). Example 3 describes a model function which is not identifiable. Example 4 shows that our concept works even in cases where uniform convergence technique does not work. Here let $\rho(x)=x^2$, and $\hat{\theta}_n=(\hat{\theta}_{n1},\ldots,\hat{\theta}_{nl})^T$. In the Examples 2 to 4 we assume that $\mathbb{E}_PZ=0$, $\mathbb{E}_PZ^2<+\infty$ and $\mathbb{E}_P(g(X\mid\theta_0)-g(X\mid\theta))^2<+\infty$.

Example 2. Here we deal with a threshold regression model, more precisely, with model (3.1) where

$$g(x\mid\theta)=\begin{cases} a_1x+a_2 & \text{for} \quad x\le r,\\ a_3x+a_4 & \text{for} \quad x>r.\end{cases}$$

Here $\{X_k\}$, $\{Z_k\}$ are two sequences of random variables such that X_k, Z_k are independent for each k, the variables Z_k are independent and $\{X_k\}$ forms a Markov chain. $\theta=(a_1,a_2,a_3,a_4,r)^T$ is the parameter vector. We assume that $\{X_k\}$ is a Markov ergodic sequence with stationary distribution π_X satisfying $\pi_X([-\kappa,\kappa])=0$, $\pi_X((-\kappa-\delta,-\kappa))>0$, $\pi_X((\kappa,\kappa+\delta))>0$ for all $\delta>0$ where $\kappa>0$. Let $\theta_0=(\tilde{a}_1,\tilde{a}_2,\tilde{a}_3,\tilde{a}_4,0)^T$ with $\tilde{a}_1\ne\tilde{a}_3$ or $\tilde{a}_2\ne\tilde{a}_4$ and $\Theta=\{\theta\in\mathbb{R}^5:|a_i|\le\bar{a}_i,|r|\le\bar{r}\}$ with given $\bar{a}_1,\ldots,\bar{a}_4,\bar{r}>\kappa$. This model is not identifiable and

$$\begin{aligned}\mathbb{E}_P\rho(Y-g(X\mid\theta)) &= \mathbb{E}_P(g(X\mid\theta_0)-g(X\mid\theta))^2+\mathbb{E}_PZ^2\\ &= \mathbb{E}_PZ^2 \quad \text{for all} \quad \theta\in\Psi=\{(\tilde{a}_1,\tilde{a}_2,\tilde{a}_3,\tilde{a}_4,r)^T:r\in[-\kappa,\kappa]\},\\ \mathbb{E}_P\rho(Y-g(X\mid\theta)) &> \mathbb{E}_PZ^2 \quad \text{for all} \quad \theta\notin\Psi.\end{aligned}$$

Then by Theorem 3.1, identity (3.4) and

$$\lim_{n\to\infty}\hat{\theta}_{nj}=\tilde{a}_j \quad (j=1,\ldots,4) \quad \text{a.s. hold true.}$$

Example 3. Let $\Theta=[a_1,a_2]\times[b_1,b_2]\times[c_1,c_2]$ with $b_1<0<b_2$. In this example, we consider the regression model (3.1) with regression function

$$g(x\mid\theta)=a+be^{cx},$$

where $\theta = (a, b, c)^T \in \Theta$. Let $\theta_0 = (a_0, b_0, c_0)^T \in \Theta$ be the true parameter vector. In the case $b_0 = 0$ we have $\Psi = \{\theta : b = 0, a = a_0, c \in [c_1, c_2]\}$. Then by Theorem 3.1, we obtain (3.4) and

$$\lim_{n\to\infty} \hat{\theta}_{n1} = a_0, \qquad \lim_{n\to\infty} \hat{\theta}_{n2} = 0.$$

Example 4. Here let $\Theta = [0, 1]$ and let X have a uniform distribution on $[-1, 1]$. For $\theta > 0$, we define the regression function of model (3.1) by

$$g(x \mid \theta) = \begin{cases} \frac{1}{\theta} - \frac{|x|}{\theta^2} & \text{for} \quad |x| \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\theta_0 = 0$. Now a natural definition of $g(x \mid 0)$ is given by 0 since it is the pointwise limit of $g(x \mid \theta)$ for $\theta \to 0$ except for $x = 0$. For $x \neq 0$, $g(x \mid \cdot)$ is continuous. Moreover,

$$\begin{aligned} \lim_{\theta\downarrow 0} \mathbb{E}_P(Y - g(X \mid \theta))^2 &= \lim_{\theta\downarrow 0} \mathbb{E}(g(X \mid 0) - g(X \mid \theta))^2 + \mathbb{E}_P Z^2 \\ &= \frac{1}{2} \lim_{\theta\downarrow 0} \int_{-\theta}^{\theta} \left(\frac{1}{\theta} - \frac{|x|}{\theta^2}\right)^2 \mathrm{d}x + \mathbb{E}_P Z^2 \\ &= \lim_{\theta\downarrow 0} \frac{1}{3\theta} + \mathbb{E}_P Z^2 = +\infty \end{aligned}$$

and

$$\mathbb{E}_P(Y - g(X \mid 0))^2 = \mathbb{E}_P Z^2 = \min_{\theta} \mathbb{E}_P(Y - g(X \mid \theta))^2.$$

The function f_n is continuous almost surely for each n and cannot uniformly converge to the discontinuous f. Thus the technique of uniform convergence does not apply in this case. On the other hand, an application of Theorem 3.1 implies

$$\lim_{n\to\infty} \hat{\theta}_n = \theta_0 \qquad \text{a.s.}$$

In the following we derive two lemmas stating sufficient conditions for (3.5) in the case of a convex function ρ. An other way to find sufficient conditions for (3.5) is described in Liese and Vajda ((1994), Lemma 2b). Every convex function ρ has derivatives from the right $\partial^+\rho$ and from the left $\partial^-\rho$ which are nondecreasing and continuous from the right and from the left, respectively. Moreover we have

$$\rho(z + a) - \rho(z) = \int_z^{z+a} \partial^+\rho(t)\mathrm{d}t = \int_z^{z+a} \partial^-\rho(t)\mathrm{d}t \qquad \text{for all} \quad z, a \in \mathbb{R} \tag{3.7}$$

and

$$\rho(z + a) \ge \rho(z) + a\partial^+\rho(z) \qquad \text{for} \quad z, a \in \mathbb{R}. \tag{3.8}$$

Here F is the distribution function of Z.

LEMMA 3.1. *Assume that* $\mathbb{E}\partial^+\rho(Z) = 0$, *and there are real numbers* $b_1 < b_3 < b_4 < b_2$ *such that*

$$\begin{aligned} &\partial^+\rho(z_2) > \partial^+\rho(z_1) \quad \textit{for all} \quad z_2 > z_1, \ z_1, z_2 \in [b_1, b_2] \quad \textit{and} \\ &\mathbb{P}\{b_3 < Z < b_4\} > 0. \end{aligned}$$

Then (3.5) *is satisfied.*

PROOF. By (3.7), we obtain

$$\rho(z+a) > \rho(z) + a\partial^+\rho(z) \quad \text{for} \quad z \in [b_3, b_4],\ a \neq 0.$$

Further by (3.8),

$$\int_{-\infty}^{\infty} (\rho(z+a) - \rho(z))\mathrm{d}F(z) > a \int_{-\infty}^{\infty} \partial^+\rho(z)\mathrm{d}F(z) = 0$$

which is (3.5). □

In Lemma 3.1 one can work with $\partial^-\rho$ instead of $\partial^+\rho$. If ρ is twice differentiable on $\mathbb{R}$, $\mathbb{E}\rho'(Z) = 0$ and $\rho''(z) > 0$ for all $z \in \mathbb{R}$, then the assumptions of Lemma 3.1 are satisfied which in turn implies the validity of (3.5).

LEMMA 3.2. *The conditions*

$$\begin{aligned} 0 < \mathbb{E}\partial^+\rho(Z+a) < +\infty \quad &\textit{for} \quad a > 0 \quad \textit{and} \\ -\infty < \mathbb{E}\partial^+\rho(Z+a) < 0 \quad &\textit{for} \quad a < 0, \end{aligned}$$

are sufficient for (3.5).

PROOF. The function

$$a \mapsto M(a) = \int_{-\infty}^{\infty} (\rho(z+a) - \rho(z))\mathrm{d}F(z)$$

is convex and $M(0) = 0$. By Lebesgue's theorem on dominated convergence,

$$\begin{aligned} \partial^+ M(a) &= \int_{-\infty}^{\infty} \lim_{h \downarrow 0} h^{-1}(\rho(z+a+h) - \rho(z+a))\mathrm{d}F(z) \\ &= \int_{-\infty}^{\infty} \partial^+\rho(z+a)\mathrm{d}F(z) \qquad (a \in \mathbb{R}). \end{aligned}$$

Therefore $\partial^+ M(a) > 0$ for $a > 0$ and $\partial^+ M(a) < 0$ for $a < 0$ which completes the proof. □

At the end of this section we study several examples of functions ρ.

Example 5. Let $\rho(z) = z^2$. Condition (3.5) follows from $\mathbb{E}Z = 0$ in view of Lemma 3.1.

Example 6. Let $\rho(z) = |z|^p$, $p > 1$. Here by Lemma 3.1,

$$\mathbb{E}|Z|^{p-1}\operatorname{sgn}(Z) = 0 \tag{3.9}$$

implies (3.5). For example, identity (3.9) is satisfied for symmetric F.

Example 7. Let $\rho(z) = |z|$, and F be the distribution function of Z. Assume that the median of Z is unique and $\text{med}(Z) = 0$, i.e. $F(t) < 0.5$ for $t < 0$, $F(t) > 0.5$ for $t > 0$. In this case we apply Lemma 3.2. Since

$$\begin{aligned}\mathbb{E}\partial^+\rho(Z+a) &= \mathbb{E}(I(Z+a \geq 0) - I(Z+a<0))\\ &= 1 - 2\mathbb{P}\{Z+a<0\} = 1 - 2F(-a-0),\end{aligned}$$

the assumptions of Lemma 3.2 are fulfilled and (3.5) holds true.

Let us now consider the case where the density λ of Z is given by

$$\lambda(z) = \begin{cases} 1 & \text{for} \quad z \in [-1, -\frac{1}{2}) \cup [\frac{1}{2}, 1),\\ 0 & \text{otherwise.}\end{cases}$$

In this case the median is not unique and we have

$$\int_{-\infty}^{\infty} \rho(z+a)\mathrm{d}F(z) = \begin{cases} \frac{3}{4} & \text{for} \quad a \in [-\frac{1}{2}, \frac{1}{2}],\\ 1 - |a| + a^2 & \text{for} \quad |a| \in (\frac{1}{2}, 1),\\ |a| & \text{for} \quad |a| \geq 1.\end{cases}$$

Therefore (3.5) is not satisfied. In the regression model (3.1) with independent random variables X and Z, and the above distribution of Z, the minimising properties of f depend heavily on the shape of g. Let $g(x \mid (a,b)^T) = ax + b$, and a_0, b_0 be the true parameters of the model (3.1). Assume that X has a density on $\mathbb{R}$ which is everywhere positive. Then we have $\Psi = \{(a_0, b) : b \in [b_0 - \frac{1}{2}, b_0 + \frac{1}{2}]\}$. According to Theorem 3.1, we cannot expect that the estimator for b is consistent, but (3.4) holds true.

4. M-estimation and model selection

In this section we consider the model selection problem in connection with parameter estimation for the model (3.1). Among the papers dealing with model selection criteria in connection with M-estimation, we refer to papers by Burman and Nolan (1995) and Rao and Wu (1989) were strong consistency was proved. The effects of model selection on consistency and the asymptotic distribution of the estimator was studied in Pötscher (1991). Here for simplicity, we assume that $\{X_k\}$ and $\{Z_k\}$ are two independent sequences of i.i.d. random variables. We incorporate various model functions $g_1, \ldots, g_\kappa$ in one model with function g such that $g_j : \mathbb{R}^m \times \Theta_j \to \mathbb{R}$, and $\Theta_j \subset \mathbb{R}^{\mu_j}$ is compact. The dimension of Θ_j can differ from model to model. The first component of the parameter vector θ gives the number of the model function such that

$$g(x \mid \theta) = g_j(x \mid \tilde{\theta}) \quad \text{for} \quad x \in \mathbb{R}^m \quad \text{if} \quad \theta = (j, \tilde{\theta})^T, \ \ \tilde{\theta} \in \Theta_j.$$

Moreover, $\Theta = \bigcup_{j=1}^{\kappa} \{j\} \times \Theta_j$. $\theta_0 = (j_0, \tilde{\theta}_0)^T$ is the true parameter, and j_0 is the number of the true model. The distance $d(\cdot,\cdot)$ is introduced by

$$d(\theta_1, \theta_2) = \begin{cases} \|\theta_1 - \theta_2\| & \text{if} \quad \theta_{11} = \theta_{21},\\ a & \text{otherwise,}\end{cases}$$

where $a > 0$ is a given quantity. Let the estimator $\hat{\theta}_n = (\hat{\theta}_{n1}, \tilde{\theta}_n^T)^T$ satisfies (3.2). Thus the model selection is realised via minimising f_n. Now one can apply Theorem 3.1(b) to obtain

$$\lim_{n\to\infty} \hat{\theta}_n = \theta_0 \quad \text{a.s.}$$

Corollary 4.1. *Under the assumptions of Theorem* 3.1(b), *the model is selected asymptotically correct:*

$$\lim_{n\to\infty} \hat{\theta}_{n1} = j_0 \qquad a.s.$$

In applications the situation very often occur where the regression function of the true model belongs to several partial models. In this case Theorem 3.1(a) can be applied instead of part (b) of this theorem. In this way we obtain Corollary 4.2.

Corollary 4.2. *Let the assumptions of Theorem* 3.1(a) *and* (3.5) *be satisfied. Assume that* $g_1, \dots, g_\kappa$ *are continuous and* X *has a density on* $\mathbb{R}^m$ *which is everywhere positive. Moreover*

$$g(x \mid \theta_0) = g(x \mid \theta_1) \qquad \textit{for all} \quad x \in \mathbb{R}^m, \;\; \theta_1 \in \Theta_0 \subset \Theta$$

and for all $\theta_1 \in \Theta \backslash \Theta_0$, *there is an* $x \in \mathbb{R}^m$ *such that* $g(x \mid \theta_0) \neq g(x \mid \theta_1)$. *Then*

$$\lim_{n\to\infty} \sum_{j=1}^{\kappa} \inf_{\tilde{\theta}:(j,\tilde{\theta})\in\Theta_0} \|\tilde{\theta}_n - \tilde{\theta}\| I(\hat{\theta}_{n1} = j) = 0 \qquad a.s.$$

We demonstrate the application of these corollaries in the following example.

Example 8. Let

$$g(x \mid \theta) = \begin{cases} a + be^{cx} & \text{for} \quad \theta_1 = 1, \;\; \theta = (1, a, b, c)^T, \\ a + dx & \text{for} \quad \theta_1 = 2, \;\; \theta = (2, a, d)^T \end{cases}$$

be the regression function of (3.1) including two models. Assume that X has a density which is everywhere positive. Here $\Theta_1 \subset \mathbb{R}^3$, and $\Theta_2 \subset \mathbb{R}^2$. In the case $\theta_0 \in \Theta_1$, let $\theta_0 = (j_0, a_0, b_0, c_0)^T$, $j_0 = 1$ and $\theta_0 = (j_0, a_0, d_0)^T$, $j_0 = 2$ otherwise.

Case 1. Either $j_0 = 1$, $b_0 \neq 0$, $c_0 \neq 0$ or $j_0 = 2$, $d_0 \neq 0$.

Theorem 3.1(b) yields that

$$\lim_{n\to\infty} \hat{\theta}_n = \theta_0 \quad \text{a.s.} \qquad \text{and} \qquad \lim_{n\to\infty} \hat{\theta}_{n1} = j_0 \quad \text{a.s.}$$

Case 2. Either $j_0 = 1$, $b_0 = 0$ or $\theta_{10} = 2$, $d_0 = 0$.

Both situations lead to the same model function. From Corollary 4.2, it follows that

$$\lim_{n\to\infty} \hat{a}_n = a_0 \qquad \text{a.s.}$$

$$\lim_{n\to\infty} D_n = 0 \qquad \text{a.s.,} \qquad D_n = \begin{cases} \sqrt{(\hat{a}_n - a_0)^2 + \hat{b}_n^2} & \text{for} \quad \hat{\theta}_{n1} = 1, \\ \sqrt{(\hat{a}_n - a_0)^2 + \hat{d}_n^2} & \text{for} \quad \hat{\theta}_{n1} = 2. \end{cases}$$

5. Estimators in regression models with fixed design

The fixed-design regression model reads as follows:

$$Y_k = g(x_k \mid \theta_0) + Z_k \qquad (k = 1, 2, \ldots)$$

where $\{Z_k\}_{k=1,2,\ldots}$ is a sequence of independent random variables. $x_1, x_2, \ldots$ is the sequence of deterministic design points. Let $g : \mathbb{R}^m \times \Theta \to \mathbb{R}$, $\Theta \subset \mathbb{R}^p$ be a measurable function. Suppose that $\hat{\theta}_n$ is an estimator for θ_0 with property

$$f_n(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} f_n(\theta) + \varepsilon_n \tag{5.1}$$

where $\varepsilon_n \geq 0$, $\varepsilon_n \to 0$ a.s.,

$$f_n(\theta) = a_n^{-1} \sum_{i=1}^{n} \rho(Y_i - g(x_i \mid \theta)) \tag{5.2}$$

and $\{a_n\}$ is a suitable sequence of positive real numbers tending to ∞. Let $\rho : \mathbb{R} \to \mathbb{R}$ be a nonnegative function and $\Delta_i(t) := g(x_i \mid \theta_0) - g(x_i \mid t)$. First we provide a rather general theorem about convergence of $\hat{\theta}_n$.

THEOREM 5.1. *Assume that Θ is compact and the function f defined by*

$$f(\theta) := \liminf_{n \to \infty} \mathbb{E} f_n(\theta)$$

is not equal to ∞ on Θ. Suppose that $\inf_{t \in B(\theta,R)} \rho(Z_i + \Delta_i(t))$ is a random variable for each $\theta \in \Theta$, $i \in \{1, \ldots, n\}$, and for every $\theta \in \Theta$ and $\varepsilon > 0$, there is some $R > 0$ such that

$$\liminf_{n \to \infty} a_n^{-1} \sum_{i=1}^{n} \left(\mathbb{E} \inf_{t \in B(\theta,R)} \rho(Z_i + \Delta_i(t)) - \mathbb{E}\rho(Z_i + \Delta_i(\theta)) \right) > -\varepsilon, \tag{5.3}$$

$$\lim_{n \to \infty} a_n^{-1} \left(\sum_{i=1}^{n} \inf_{t \in B(\theta,R)} \rho(Z_i + \Delta_i(t)) - \tilde{b}_n(\theta) \right) = 0 \quad a.s., \tag{5.4}$$

$$\lim_{n \to \infty} \left(f_n(\theta) - \frac{b_n(\theta)}{a_n} \right) = 0 \quad a.s. \tag{5.5}$$

where

$$b_n(\theta) := \sum_{i=1}^{n} \mathbb{E}\rho(Z_i + \Delta_i(\theta)), \qquad \tilde{b}_n(\theta) := \sum_{i=1}^{n} \mathbb{E} \inf_{t \in B(\theta,R)} \rho(Z_i + \Delta_i(t)).$$

Let $\Psi := \operatorname{argmin}_{\theta \in \Theta} f(\theta) \neq \emptyset$. *Then*

$$\lim_{n \to \infty} d(\hat{\theta}_n, \Psi) = 0 \quad a.s.$$

PROOF. Assumption (5.5) implies $f_n \xrightarrow[\{\theta_0\}]{epi-u-a.s.} f$. Let $\theta \in \Theta$ and $\varepsilon > 0$. By (5.3) and (5.4),

$$\begin{aligned}&\liminf_{n\to\infty} \inf_{t\in B(\theta,R)} f_n(t) - f(\theta)\\ &\geq \liminf_{n\to\infty} a_n^{-1} \sum_{i=1}^{n} \left(\inf_{t\in B(\theta,R)} \rho(Z_i + \Delta_i(t)) - \mathbb{E} \inf_{t\in B(\theta,R)} \rho(Z_i + \Delta_i(t)) \right) - \varepsilon\\ &\geq -\varepsilon \qquad \text{a.s.}\end{aligned}$$

By virtue of Lemma 2.3, we obtain $f_n \xrightarrow[\Theta]{l-a.s.} f$. An application of Theorem 2.2 leads to Theorem 5.1. □

In several applications sufficient conditions for (5.3) and (5.4) can be obtained by applying strong laws of large numbers (cf. Petrov (1995), Chapter 6).

Next we study strong consistency of approximate M-estimators in the case of convex functions ρ. We assume that the convex function ρ fulfils

$$\rho(z+a) \geq \rho(z) + a\partial^+\rho(z) + \eta(a)\psi(z) \tag{5.6}$$

with nonnegative measurable functions η, ψ. Considering $\rho(z) = z^2$, inequality (5.6) is valid with $\eta(a) = a^2$, $\psi(z) = 1$. For other special cases see below. Condition (5.6) is fulfilled if

$$\int_z^{z+a} (\partial^+\rho(t) - \partial^+\rho(z))\mathrm{d}t \geq \eta(a)\psi(z).$$

If the second derivative of ρ is continuous on $\mathbb{R}$ and $\inf_z \rho''(z) \geq d > 0$, then Taylor expansion leads to

$$\rho(z+a) \geq \rho(z) + a\rho'(z) + \frac{a^2}{2}d$$

which shows the validity of (5.6) with $\eta(a) = a^2/2$, $\psi(z) = d$, $s(z) = \rho'(z)$. The result for convex functions ρ is given by the following theorem:

THEOREM 5.2. *Suppose that Θ is compact, the convex function ρ satisfies* (5.6) *and $\{Z_i\}$ is a sequence of i.i.d. random variables with $\mathbb{E}\psi(Z_1) > 0$, $\mathbb{E}s(Z_1) = 0$, $\mathbb{E}s^2(Z_1) < +\infty$, $\mathbb{E}\psi^2(Z_1) < +\infty$. Moreover, assume that for any $\theta \neq \theta_0$, $\theta \in \Theta$, for any $\varepsilon > 0$, there is some $R > 0$ such that*

$$\sum_{k=1}^{\infty} \sup_{t\in B(\theta,R)} \Delta_k(t)^2 \frac{1}{T_k^2(\theta)} < +\infty, \tag{5.7}$$

$$\sum_{k=1}^{\infty} \sup_{t\in B(\theta,R)} \eta(\Delta_k(t))^2 \frac{1}{T_k^2(\theta)} < +\infty, \tag{5.8}$$

$$\limsup_{n\to\infty} a_n^{-1} \sum_{k=1}^{n} \sup_{t\in B(\theta,R)} |\Delta_k(t) - \Delta_k(\theta)| < \varepsilon \tag{5.9}$$

and

$$\liminf_{n\to\infty} a_n^{-1} T_n(\theta) > 0 \tag{5.10}$$

where

$$T_n(\theta) := \sum_{k=1}^{n} \inf_{t \in B(\theta,R)} \eta(\Delta_k(t)).$$

Then

$$\lim_{n\to\infty} \hat{\theta}_n = \theta_0 \qquad a.s.$$

Remark 5. Wu (1981) proved a similar statement for least squares estimators (Theorem 3). In contrast to that theorem, no Lipschitz conditions are required in Theorem 5.2. In the papers by Liese and Vajda (1994, 1995) and by Berlinet *et al.* (2000), the authors derived necessary and sufficient conditions for weak consistency of M-estimators.

PROOF. Here we prove that for all $\theta \in \Theta$, $\theta \neq \theta_0$ and all $\bar{\varepsilon} > 0$, there is some $R > 0$ such that

$$\liminf_{n\to\infty} \inf_{t\in B(\theta,R)} \tilde{f}_n(t) \geq f(\theta) - \bar{\varepsilon}$$

(compare with condition (3.1) of Wu (1981)) where

$$\begin{aligned} \tilde{f}_n(\theta) &= a_n^{-1} \sum_{k=1}^{n} (\rho(Y_k - g(x_k \mid \theta)) - \rho(Y_k - g(x_k \mid \theta_0))) \\ &= a_n^{-1} \sum_{k=1}^{n} (\rho(Z_k + \Delta_k(\theta)) - \rho(Z_k)), \\ f(\theta) &:= \begin{cases} \mathbb{E}\psi(Z_1) \cdot \liminf_{n\to\infty} a_n^{-1} T_n(\theta) & \text{for} \quad \theta \neq \theta_0, \\ 0 & \text{for} \quad \theta = \theta_0. \end{cases} \end{aligned}$$

Inequality (5.1) is also fulfilled if f_n is replaced by $\tilde{f}_n$. We have

$$\begin{aligned} (5.11) \qquad & \inf_{t\in B(\theta,R)} \tilde{f}_n(t) \\ & \geq a_n^{-1} \left(\sum_{k=1}^{n} \inf_{t\in B(\theta,R)} (\Delta_k(t) s(Z_k)) + \sum_{k=1}^{n} \inf_{t\in B(\theta,R)} \eta(\Delta_k(t))\psi(Z_k) \right) \\ & \geq a_n^{-1} T_n(\theta)(A_n + B_n + \mathbb{E}\psi(Z_1)) + D_n \end{aligned}$$

where

$$\begin{aligned} A_n &:= \frac{1}{T_n(\theta)} \sum_{k=1}^{n} \left(\inf_{t\in B(\theta,R)} (\Delta_k(t)s(Z_k)) - \mathbb{E} \inf_{t\in B(\theta,R)} (\Delta_k(t)s(Z_k)) \right), \\ B_n &:= \frac{1}{T_n(\theta)} \sum_{k=1}^{n} \inf_{t\in B(\theta,R)} \eta(\Delta_k(t))(\psi(Z_k) - \mathbb{E}\psi(Z_k)), \\ D_n &:= \frac{1}{a_n} \sum_{k=1}^{n} \mathbb{E} \inf_{t\in B(\theta,R)} (\Delta_k(t)s(Z_k)). \end{aligned}$$

Further

$$\mathbb{E} \inf_{t\in B(\theta,R)} (\Delta_k(t)s(Z_k))$$

$$= \mathbb{E}\inf_{t\in B(\theta,R)}(\Delta_k(t)s(Z_k))I(s(Z_k)\geq 0) + \mathbb{E}\inf_{t\in B(\theta,R)}(\Delta_k(t)s(Z_k))I(s(Z_k)<0)$$
$$= \mathbb{E}s(Z_1)I(s(Z_1)\geq 0)\inf_{t\in B(\theta,R)}\Delta_k(t) + \mathbb{E}s(Z_1)I(s(Z_1)<0)\sup_{t\in B(\theta,R)}\Delta_k(t)$$
$$= \mathbb{E}s(Z_1)I(s(Z_1)\geq 0)\left(\inf_{t\in B(\theta,R)}\Delta_k(t) - \sup_{t\in B(\theta,R)}\Delta_k(t)\right).$$

Therefore (5.9) implies

$$D_n \geq -\varepsilon\mathbb{E}s(Z_1)I(s(Z_1)\geq 0) =: -\bar{\varepsilon}.$$

Since by (5.7) and (5.8),

$$\sum_{k=1}^{\infty}\sup_{t\in B(\theta,R)}\Delta_k^2(t)\mathbb{E}s^2(Z_1)\frac{1}{T_k^2(\theta)} < +\infty, \qquad \text{and}$$
$$\sum_{k=1}^{\infty}\sup_{t\in B(\theta,R)}\eta(\Delta_k(t))^2\mathbb{E}\psi^2(Z_1)\frac{1}{T_k^2(\theta)} < +\infty$$

hold, an application of Theorem 6.7 of Petrov (1995) leads to

$$A_n = o(1) \quad \text{and} \quad B_n = o(1) \quad \text{a.s.}$$

Consequently, (5.10) and (5.11) yield

$$\liminf_{n\to\infty}\inf_{t\in B(\theta,R)}\tilde{f}_n(t) \geq \mathbb{E}\psi(Z_1)\cdot\liminf_{n\to\infty}a_n^{-1}T_n(\theta) - \bar{\varepsilon} \quad \text{a.s.} \qquad \square$$

In the remainder of the section we study one example of functions ρ and the power curve model.

Power functions ρ: Let $\rho(x) = |x|^p$.

(i) If $p\in(1,2)$, then for $\delta>0$, we obtain

$$|x+a|^p \geq |x|^p + pa|x|^{p-1}\operatorname{sgn}(x) + \min\{a^2,|a|^p\}\psi(x) \qquad (a,x\in\mathbb{R}) \tag{5.12}$$

where $\psi(x) = 0\ \forall x\in[-\delta,\delta]$, $\psi(x) = C_1|x|^{p-2}\ \forall x: |x|>\delta>0$ and $C_1>0$ is a constant not depending on a or x.

(ii) Case $p>2$: It can be proven that

$$|x+a|^p \geq |x|^p + pa|x|^{p-1}\operatorname{sgn}(x) + a^2\psi(x) \qquad (a,x\in\mathbb{R}) \tag{5.13}$$

with $\psi(x) = p(p-1)2^{-p-3}|x|^{p-2}$.

The proofs for (5.12) and (5.13) can be found in the Appendix. Therefore the power function ρ satisfies the inequality (5.6).

Power curve model (cf. Wu (1981), Example 3): We consider the model

$$Y_k = (k+\theta_0)^d + Z_k \qquad (k=1,2,\ldots)$$

with $d \geq 1$, i.e. $x_k = k$, $g(x \mid \theta) = (x+\theta)^d$ and $\Delta_k(t) = (k+\theta_0)^d - (k+t)^d$. Let $\Theta = [\theta_1, \theta_2]$, $\theta_1 \geq 0$ and $\theta_0 \in \Theta$. The parameter θ_0 is estimated by $\hat{\theta}_n$ according to (5.1) and (5.2) with $\rho(x) = |x|^p$. Now we introduce $a_n = n^{pd-p+1}$ for $p \in (1,2)$, and $a_n = n^{2d-1}$ for $p \geq 2$. In the following $\bar{C}_1, \ldots, \bar{C}_8$ denote positive constants which do not depend on k and n, but on θ (we drop the dependence on θ in the notation). We have

$$\begin{aligned}(5.14)\qquad & \limsup_{n\to\infty} a_n^{-1} \sum_{k=1}^{n} \sup_{t\in B(\theta,R)} |\Delta_k(t) - \Delta_k(\theta)| \\ & \leq \bar{C}_1 \limsup_{n\to\infty} a_n^{-1} \sum_{k=1}^{n} k^{d-1} \sup_{t\in B(\theta,R)} |t-\theta|.\end{aligned}$$

For $d > 1$, the right hand side of (5.14) is equal to zero and we choose $R \leq |\theta_0 - \theta|/2$. In the case $d = 1$, we can choose R such that (5.9) is fulfilled for some $\varepsilon > 0$.

Case (i) $p \geq 2$. By (5.13), $\eta(a) = a^2$. Obviously, $T_n(\theta) \geq \bar{C}_3 n^{2d-1}$ such that (5.10) is satisfied (see Lemma A.3 in the Appendix). Observe that

$$\sup_{t\in B(\theta,R)} \Delta_k(t)^2 T_k^{-2}(\theta) \leq \bar{C}_4 k^{-2d}, \qquad \sup_{t\in B(\theta,R)} \eta(\Delta_k(t))^2 T_k^{-2}(\theta) \leq \bar{C}_5 k^{-2}.$$

Thus (5.7) and (5.8) are satisfied and the M-estimator $\hat{\theta}_n$ is strongly consistent.

Case (ii) $p \in (1,2)$. Here we have $\eta(a) = \min\{a^2, |a|^p\}$ by (5.12) and $T_n(\theta) \geq \bar{C}_6 n^{pd-p+1}$ (see Lemma A.3). Moreover,

$$\sup_{t\in B(\theta,R)} \Delta_k(t)^2 T_k^{-2}(\theta) \leq \bar{C}_7 k^{q}, \qquad \sup_{t\in B(\theta,R)} \eta(\Delta_k(t))^2 T_k^{-2}(\theta) \leq \bar{C}_8 k^{-2}$$

($q = -2pd + 2d + 2p - 4$) which implies (5.7) and (5.8). Hence the assumptions of Theorem 5.2 are fulfilled and the M-estimator $\hat{\theta}_n$ is strongly consistent. Therefore these considerations generalise Wu's Example 3 where $\rho(x) = x^2$.

Acknowledgements

The authors are grateful to the referees for helpful hints.

Appendix

LEMMA A.1. *For $p \in (1,2)$ and $\delta > 0$, we have*

$$|x+a|^p \geq |x|^p + pa|x|^{p-1}\operatorname{sgn}(x) + \min\{a^2, |a|^p\}\psi(x)$$

where $\psi(x) = 0$ $\forall x \in [-\delta, \delta]$, $\psi(x) = \bar{C}x^{p-2}$ $\forall x : |x| > \delta$, and $\bar{C}$, δ are positive constants not depending on a or x.

PROOF. Since the function $a \mapsto |x+a|^p$ is convex for each x, we obtain

$$|x+a|^p \geq |x|^p + pa|x|^{p-1}\operatorname{sgn}(x) \qquad \forall x \in [-\delta, \delta].$$

Therefore, it remains to prove the lemma in the case $|x| > \delta$. Without loss of generality, let $x > 0$. We have

$$\frac{|x+a|^p - x^p - apx^{p-1}}{a^2x^{p-2}} = \frac{\left|1+\frac{a}{x}\right|^p - 1 - \frac{a}{x}p}{(a/x)^2} = \Phi(a/x),$$

$$\Phi(a) = a^{-2}(|1+a|^p - 1 - ap).$$

Case $|a| < x$: Φ is bounded away from zero on compact intervals such that

$$|x+a|^p \geq x^p + pax^{p-1} + C_1a^2x^{p-2}.$$

Case $a \geq x$:

$$\begin{aligned} |x+a|^p - x^p - pax^{p-1} &= a^p\left(\left(1+\frac{x}{a}\right)^p - \left(\frac{x}{a}\right)^p - p\left(\frac{x}{a}\right)^{p-1}\right) \\ &\geq C_2a^p \geq C_2\delta^{2-p}a^px^{p-2} \end{aligned}$$

with an appropriate constant $C_2 > 0$.

Case $a \leq -x$: Similarly to the previous case,

$$\begin{aligned} |x+a|^p - x^p - pax^{p-1} &= |a|^p\left(\left(1+\frac{x}{a}\right)^p - \left(-\frac{x}{a}\right)^p + p\left(-\frac{x}{a}\right)^{p-1}\right) \\ &\geq C_3|a|^p \geq C_3\delta^{2-p}|a|^px^{p-2} \end{aligned}$$

with a suitable constant $C_3 > 0$. Thus the proof is complete. □

LEMMA A.2. *For* $p > 2$, *we have*

$$|x+a|^p \geq |x|^p + pa|x|^{p-1}\operatorname{sgn}(x) + a^2\psi(x)$$

with $\psi(x) = p(p-1)2^{-p-3}|x|^{p-2}$.

PROOF. Without loss of generality, let $x > 0$ and $a \neq 0$. We obtain

$$|x+a|^p = |x|^p + pa|x|^{p-1}\operatorname{sgn}(x) + p(p-1)\int_0^a (a-t)|x+t|^{p-2}\mathrm{d}t.$$

Case $a > 0$: Obviously,

$$|x+a|^p \geq |x|^p + pa|x|^{p-1}\operatorname{sgn}(x) + \frac{1}{2}p(p-1)a^2|x|^{p-2}.$$

Case $-2x < a < 0$: Since $x + t \geq x + \frac{a}{4} > \frac{x}{2}$ for $t \in [\frac{a}{4}, 0]$, we have

$$\begin{aligned} |x+a|^p &\geq |x|^p + pa|x|^{p-1}\operatorname{sgn}(x) + p(p-1)\int_{a/4}^0 (t-a)\mathrm{d}t\left|\frac{x}{2}\right|^{p-2} \\ &= |x|^p + pa|x|^{p-1}\operatorname{sgn}(x) + 7p(p-1)2^{-p-3}a^2|x|^{p-2}. \end{aligned}$$

Case $a \leq -2x$: Then

$$\begin{aligned} |x+a|^p &\geq |x|^p + pa|x|^{p-1}\operatorname{sgn}(x) + p(p-1)\int_a^{3a/4} (t-a)\mathrm{d}t\left|\frac{x}{2}\right|^{p-2} \\ &= |x|^p + pa|x|^{p-1}\operatorname{sgn}(x) + p(p-1)2^{-p-3}a^2|x|^{p-2} \end{aligned}$$

since $x + t \leq x + \frac{3a}{4} \leq -\frac{x}{2}$ for $t \in [a, \frac{3a}{4}]$. □

LEMMA A.3. *Let $\theta_0 \in \Theta$ and the settings of the power curve model are valid. Then for each $\theta \neq \theta_0$, there is an $R > 0$ and a $\tilde{C} > 0$ such that*

$$T_n(\theta) \geq \tilde{C} n^{\nu} \quad \textit{for all} \quad n \geq 1,$$

where $\nu = 2d - 1$ if $p \geq 2$, $\nu = pd - p + 1$ if $p \in (1, 2)$. R and $\tilde{C}$ may depend on θ.

PROOF. Here $\Delta_k(t) = (k + \theta_0)^d - (k + t)^d$. Let $R \in (0, \min\{|\theta_0 - \theta|/2, 1\})$.

Case (i) $p \geq 2$ and $\eta(a) = a^2$: Let $t_1 = \min\{\theta_0, \theta - R\}$ and $t_2 = \max\{\theta_0, \theta + R\}$. Then

$$\begin{aligned} T_n(\theta) &= \sum_{k=1}^{n} \inf_{t \in B(\theta,R)} (\Delta_k(t))^2 \\ &\geq \sum_{k=1}^{n} \inf_{t \in [t_1,t_2]} (d(k+t)^{d-1})^2 \inf_{t \in B(\theta,R)} (\theta_0 - t)^2 \\ &\geq \frac{d^2}{4} (\theta_0 - \theta)^2 \sum_{k=1}^{n-1} k^{2d-2} \end{aligned}$$

which implies the lemma in Case (i).

Case (ii) $p \in (1, 2)$ and $\eta(a) = \min\{a^2, |a|^p\}$: Let $R < |\theta_0 - \theta|/2$. There is a $k_0 = k_0(\theta, \theta_0)$ such that $|\Delta_k(t)| \geq 1$ for $t \in B(\theta, R)$. Analogously to above,

$$\begin{aligned} T_n(\theta) &= \sum_{k=1}^{n} \inf_{t \in B(\theta,R)} \eta(\Delta_k(t)) \\ &\geq \sum_{k=k_0}^{n} \inf_{t \in [t_1,t_2]} |d(k+t)^{d-1}|^p \inf_{t \in B(\theta,R)} |\theta_0 - t|^p \\ &\geq d^p \left| \frac{\theta_0 - \theta}{2} \right|^p \sum_{k=k_0-1}^{n-1} k^{pd-p} \end{aligned}$$

which implies the lemma in Case (ii). □

REFERENCES

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study, *The Annals of Statistics*, **10**, 1100–1120.

Arcones, M. A. (1994). Some strong limit theorems for M-estimators, *Stochastic Processes and Their Applications*, **53**, 241–268.

Artstein, Z. and Wets, R. J.-B. (1994). Stability results for stochastic programs and sensors, allowing for discontinuous objective functions, *SIAM Journal on Optimization*, **4**, 537–550.

Berlinet, A., Liese, F. and Vajda, I. (2000). Necessary and sufficient conditions for consistency of M-estimates in regression models with general errors, *Journal of Statistical Planning and Inference*, **89**(1–2), 243–267.

Burman, P. and Nolan, D. (1995). A general Akaike-type criterion for model selection in robust regression, *Biometrika*, **82**, 877–886.

Dudley, R. M. (1998). Consistency of M-estimators and one-sided bracketing, *High Dimensional Probability*, Proceedings of the conference in Oberwolfach (eds. E. Eberlein *et al.*), 33–58, Birkhäuser, Basel.

Dupačová, J. and Wets, R. J.-B. (1988). Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems, *The Annals of Statistics*, **16**, 1517–1549.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation, *The Annals of Statistics*, **22**, 1993–2010.

Hess, C. (1996). Epi-convergence of sequences of normal integrands and strong consistency of the maximum likelihood estimator, *The Annals of Statistics*, **24**, 1298–1315.

Kaňková, V. (1978). Optimum solution of a stochastic optimization problem with unknown parameters, *Information Theory, Statistical Decision Functions, Random Processes*, Transactions of the 7th Prague conference, Vol. B, Prague 1974, 239–244.

King, A. J. and Rockafellar, T. (1993). Asymptotic theory for solutions in statistical estimation and stochastic programming, *Mathematics of Operations Research*, **18**, 148–162.

Korf, L. A. and Wets, R. J.-B. (2001). Random lsc functions: An ergodic theorem, *Mathematics of Operations Research*, **26**, 421–445.

Liebscher, E. (2003). Estimation in nonlinear autoregressive models, *Journal of Multivariate Analysis*, **84**, 247–261.

Liese, F. and Vajda, I. (1994). Consistency of M-estimates in general regression models, *Journal of Multivariate Analysis*, **50**, 93–114.

Liese, F. and Vajda, I. (1995). Necessary and sufficient conditions for consistency of generalized M-estimates, *Metrika*, **42**, 291–324.

Petrov, V. V. (1995). *Limit Theorems of Probability Theory. Sequences of Independent Random Variables*, Oxford Studies in Probability, Clarendon Press, Oxford.

Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates, *Metrika*, **14**, 249–272.

Pflug, G. Ch. (1995). Asymptotic stochastic programs, *Mathematics of Operations Research*, **20**, 769–789.

Pflug, G. Ch. (2003). Stochastic optimization and statistical inference, *Handbooks in Operations Research and Management Science* (eds. A. Ruszczinski and A. Shapiro), **10**, 427–482.

Pötscher, B. M. (1991). Effects of model selection on inference, *Econometric Theory*, **7**, 163–185.

Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models, Asymptotic Theory*, Springer, Berlin.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem, *Biometrika*, **76**, 369–374.

Robinson, S. M. (1987). Local epi-continuity and local optimization, *Mathematical Programming*, **37**, 208–222.

Rockafeller, R. T. and Wets, R. J.-B. (1998). *Variational Analysis*, Springer, Berlin.

Salinetti, G. and Wets, R. J.-B. (1986). On the convergence in distribution of measurable multifunctions (random sets), normal integrands, stochastic processes and stochastic infima, *Mathematics of Operations Research*, **11**, 385–419.

Shapiro, A. (1989). Asymptotic properties of statistical estimators in stochastic programming, *The Annals of Statistics*, **17**, 841–858.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs, *Annals of Operations Research*, **30**, 169–186.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*, J. Wiley, New York.

Vogel, S. (1994). A stochastic approach to stability in stochastic programming, *Journal of Computational and Applied Mathematics*, **56**, 65–96.

Vogel, S. and Lachout, P. (2003*a*). On continuous convergence and epi-convergence of random functions. Part I: Theory and relations, *Kybernetika*, **39**, 75–98.

Vogel, S. and Lachout, P. (2003*b*). On continuous convergence and epi-convergence of random functions. Part II: Sufficient conditions and applications, *Kybernetika*, **39**, 99–118.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate, *The Annals of Mathematical Statistics*, **20**, 595–601.

Wang, J.-L. (1995). M-estimators for censored data: Strong consistency, *Scandinavian Journal of Statistics*, **22**, 197–205.

Wu, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation, *The Annals of Statistics*, **9**, 501–513.