International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

Asymptotic Behavior of Statistical Estimators and Optimal Solutions for Stochastic Optimization Problems

Dupacova, J. & Wets, R.J.-B.

IIASA Working Paper

WP-86-041

August 1986

WORKING PAPER

ASYMPTOTIC BEHAVIOR OF STATISTICAL ESTIMATORS AND OPTIMAL SOLUTIONS FOR STOCHASTIC OPTIMIZATION PROBLEMS

Jitka Dupačová
Roger Wets

August 1986
WP-86-041

IIASA
International Institute
for Applied Systems Analysis

ASYMPTOTIC BEHAVIOR OF STATISTICAL ESTIMATORS AND OPTIMAL SOLUTIONS FOR STOCHASTIC OPTIMIZATION PROBLEMS

Jitka Dupačová
Roger Wets

August 1986
WP-86-41

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
2361 Laxenburg, Austria

FOREWORD

This paper presents the first results on a new statistical approach to the problem of incomplete information in stochastic programming. The tools of nondifferentiable optimization used here help to prove the consistency of (approximate) optimal solutions based on an increasing information on the true probability distribution without unnatural smoothness assumptions. They also allow to take fully into account the presence of constraints.

Alexander B. Kurzhanski
Chairman
System and Decision Sciences Program

CONTENTS

ASYMPTOTIC BEHAVIOR OF STATISTICAL ESTIMATORS AND OPTIMAL SOLUTIONS FOR STOCHASTIC OPTIMIZATION PROBLEMS

Jitka Dupačová and Roger Wets

1. INTRODUCTION

The calculation of estimates for various statistical parameters has been one of the main concerns of Statistics since its inception, and a number of elegant formulas have been developed to obtain such estimates in a number of particular instances. Typically such cases correspond to a situation when the random phenomenon is univariate in nature, and there are no "active" restrictions on the estimate of the unknown statistical parameter. However, that is not the case in general, many estimation problems are multivariate in nature and there are restrictions on the choice of the parameters. These could be simple nonnegativity constraints, but also much more complex restrictions involving certain mathematical relations between the parameters that need to be estimated. Classical techniques, that can still be used to handle least square estimation with linear equality constraints on the parameters for example, break down if there are inequality constraints or a nondifferentiable criterion function. In such cases one cannot expect that a simple formula will yield the relationship between the samples and the best estimates. Usually, the latter must be found by solving an optimization problem. Naturally the solution of such a problem depends on the collected samples and one is confronted with the questions of the consistency and of the asymptotic behavior of such estimators. This is the subject of this article.

To overcome the technical problems caused by the intrinsic lack of smoothness, we rely on the guidelines and the tools provided by theory of nondifferentiable optimization. In fact, the problem of proving consistency of the estimators, and the study of their asymptotic behavior is closely related to that of obtaining confidence intervals for the solution of stochastic optimization problems when there is only partial information about the probability distribution of the random coefficients of the problem. In fact it was the need to deal with this class of prob-

lems that originally motivated this study. We shall see in Section 2 that stochastic optimization problems as well as the problem of finding statistical estimators are two instances of the following general class of problems:

$$\text{find } x \in R^n \quad \text{that minimizes } E\{f(x, \underset{\sim}{\xi})\} ,$$

where $f : R^n x\ \Xi \rightarrow R \cup \{+\infty\}$ is an extended real valued function and $\underset{\sim}{\xi}$ is a random variable with values in Ξ; for more details see Section 3. It is implicit in this formulation that the expectation is calculated with respect to the true probability distribution P of the random variable $\underset{\sim}{\xi}$, whereas in fact all that is known is a certain approximate P^ν. Our objective is to study the behavior of the optimal solution (estimate) x^ν, obtained by solving the optimization problem using P^ν instead of P to calculate the expectation, when the $\{P^\nu, \nu = 1,\ldots\}$ is a sequence of probability measures converging to P. In Section 3 we give conditions under which consistency can be proved. Constraints on the choice of the optimal x are incorporated in the formulation of the problem by allowing the function f to take on the value $+\infty$. The results are obtained without explicit reference to the form of these constraints.

There is of course a substantial statistical literature dealing with the questions broached here, beginning with the seminal article of Wald (1949) and the work of Huber (1967) on maximum likelihood estimators. Of more direct parentage, at least as far as formulation and use of mathematical techniques, is the work on stochastic programming problems with partial information. Wets (1979) reports some preliminary results, further developments were presented at the 1980 meeting on stochastic optimization at IIASA (Laxenburg, Austria) and recorded in Solis and Wets (1981), see also Dupačová (1983a, b) and (1984b) for a special case. In a projected paper we shall deal with estimates of the convergence rates, as well as with the convergence of the associated Lagrangian function.

2. EXAMPLES

The results apply equally well to estimation or stochastic optimization problems with or without constraints, with differentiable or nondifferentiable criterion function. However, the examples that we detail here are those that fall outside the classical mold, viz. unconstrained smooth problems.

Restrictions on the statistical estimates or the optimal decisions of stochastic optimization problems, follow from technical and modeling considerations as well as natural statistical assumptions. The least square estimation problem with linear equality constraints, a basic statistical method, see e.g. Rao (1965), can be solved by a usual tools of differential calculus. The inequality constraints however introduce a lack of smoothness that does not allow us to fall back on the old stand-bys. In Judge and Takayama (1966), Liew (1976) the theory of quadratic programming is used to exhibit and discuss the statistical properties of least square estimates subject to inequality constraints for the case of large and small samples.

In connection with the maximum likelihood estimation, the case of parameter restrictions in the form of smooth nonlinear equations was studied by Aitchinson and Silvey (1958) including results on asymptotic normality of the estimates. The Lagrangian approach was further developed by Silvey (1959), extended to the case of a multisample situation by Sen (1979) including analysis of the situation when the true parameter value does not fulfill the constraints (the nonnull case).

Typically one must take into account in the estimation of variances and variance components nonnegativity restrictions. Unconstrained maximum likelihood estimation in factor analysis and in more complicated structural analysis models, see e.g. Lee (1980), may lead to negative estimates of the variances. Replacing these unappropriate estimates by zeros gives estimates which are no longer optimal with respect to the chosen fitting function. Similarly, there is a problem of getting negative estimates of variance components, see Example 2.3. In statistical practice, these nonpositive variance estimates are usually fixed at zero and the data is eventually reanalyzed. In general, such an approach may lead to plausible results in case of estimating one restricted parameter only and it is mostly unappropriate in multi-dimensional situations; see e.g. the evidence given by Lee (1980).

The possibility of using mathematical programming techniques to get constrained estimates was explored by Arthanari and Dodge (1981). As mentioned in the introduction we use mathematical programming theory not only to get inequality constrained estimates but to get asymptotic results for a large class of decision and estimation problems which contains, inter alia, restricted M-estimates and stochastic programming with incomplete information. In comparison with the results of ad hoc approaches valid mostly for one-dimensional restricted estimation our method can be used for high-dimensional cases and without unnatural smoothness assumptions, in spite of the fact that the violation of differentiability assumptions

cannot be easily bypassed by the use of directional derivatives (in contrast to the one-dimensional case).

EXAMPLE 2.1 *Inequality constrained least squares estimation of regression coefficients.* Assume that the dependent variable y can be explained or predicted on the base of information provided by independent variables $x_1, \ldots, x_p$. In the simplest case of linear model, the observations y_j on y are supposed to be generated according to

$$y_j = \sum_{i=1}^{p} x_{ij} \beta_i + \varepsilon_j, \quad j = 1, \ldots, \nu ,$$

where $\beta_1, \ldots, \beta_p$ are unknown parameters to be estimated, ε_j, $j = 1, \ldots, \nu$, denote the observed values of residual and $X = (x_{ij})$ is a (p, ν) matrix whose rows consist of the observed values of the independent variables.

In the practical implementation of this model, there may be in addition some a priori constraints imposed on the parameters such as nonnegativity constraints on the elasticities, see Liew (1976), a required presigned positive difference between input and output tonnage due to the meeting loss, Arthanari and Dodge (1981). Assume that these constraints are of the form

$$A\beta \leq c$$

where A(m, p), c(m, 1) are given matrices. The use of the least squares method leads to the optimization problem:

$$\begin{aligned} &\text{minimize} \sum_{j=1}^{\nu} \left[y_j - \sum_{i=1}^{p} x_{ij} \beta_i \right]^2 \\ &\text{subject to} \sum_{i=1}^{p} a_{ki} \beta_i \leq c_k, \; k = 1, \ldots, m , \end{aligned} \tag{2.1}$$

which can be solved by quadratic programming techniques.

In our general framework, problem (2.1) corresponds to the case of objective function:

$$\begin{aligned} f(x, \xi) &= \left[\xi_0 - \sum_{i=1}^{p} \xi_i x_i \right]^2 \quad \text{if} \quad x \in S = \{x \mid Ax \leq c\} , \\ &= +\infty \quad \text{otherwise} \end{aligned} \tag{2.2}$$

with the P^ν the empirical distributions.

Alternatively, minimizing the sum of absolute errors corresponds to the optimization problem

$$\begin{aligned} &\text{minimize} \sum_{j=1}^{\nu} |y_j - \sum_{i=1}^{p} x_{ij}\beta_i| \\ &\text{subject to} \sum_{i=1}^{p} a_{ki}\beta_i \le c_k \ , \quad 1 \le k \le m \ , \end{aligned} \tag{2.3}$$

which can be solved by means of the simplex method for linear programming, see e.g. Arthanari and Dodge (1981). The formulation of (2.3) is again based on the empirical distribution function P^ν, the objective functions is:

$$\begin{aligned} f(x, \xi) &= |\xi_0 - \sum_{i=1}^{p} \xi_i x_i| \quad \text{if } x \in S \\ &= +\infty \quad \text{otherwise} \end{aligned} \tag{2.4}$$

Note, that this function f is not differentiable on S.

Finally, when robustizing the least squares approach, instead of minimizing a sum of squares a sum of less rapidly increasing functions of residuals is minimized, see e.g. Huber (1973):

$$\begin{aligned} &\text{minimize} \sum_{j=1}^{\nu} \rho\left[y_j - \sum_{i=1}^{p} x_{ij}\beta_i\right] \\ &\text{subject to} \sum_{i=1}^{p} a_{ki}\beta_i \le c_k \ , \quad 1 \le k \le m \ . \end{aligned} \tag{2.5}$$

The function ρ is assumed to be convex, non-monotone and to possess bounded derivatives of sufficiently high order, e.g.

$$\begin{aligned} \rho(u) &= \frac{1}{2}u^2 \qquad \text{for } |u| < c \\ &= c|u| - \frac{1}{2}c^2 \quad \text{for } |u| \ge c \ . \end{aligned}$$

This also fits the general framework; the objective function is:

$$f(x, \xi) = \rho\left[\xi_0 - \sum_{i=1}^{p} \xi_i x_i\right] \quad \text{if } x \in S$$

(2.6)

$= +\infty$ otherwise

and the empirical distribution function P^{ν} is again used to obtain (2.5).

EXAMPLE 2.2 *Heywood cases in factor analysis.* The model for confirmative factor analysis (Jöreskog (1969)) is

$$x = \Lambda f + e$$

where x(n, 1) is a column vector containing the observed variables, f is a column vector containing the k common factors, e(n, 1) is a column vector containing the individual parts of the observables components and Λ(n, k) is the matrix of factor loadings. It is assumed that f and e are normally distributed with mean zero, var f = Θ and var e = Ψ, which is diagonal. Consequently, x is normally distributed with mean zero and with the variance matrix

$$\Sigma = \Lambda \Psi \Lambda^T + \Phi \ . \tag{2.7}$$

The parameter vector consists of the free elements of Λ, Ψ and Φ and it should be estimated using the sample variance matrix S of observables x. This is done by minimizing a suitable fitting function, such as

$$f_1(\Sigma, S) = \log|\Sigma| + \mathrm{tr}(S\,\Sigma^{-1}) - \log|S| - n \tag{2.8}$$

(the maximum likelihood method), or

$$f_2(\Sigma, S) = \frac{1}{2}\mathrm{tr}((S - \Sigma)V)^2 \ , \tag{2.9}$$

where V is a matrix of weights (the weighted least squares method). Evidently, both (2.8) and (2.9) with (2.7) substituted for Σ, are objective functions of non-trivial unconstrained optimization problems, which can be solved by different methods such as the method of Davidon–Fletcher–Powell (see Fletcher and Powell (1963) or by the Gauss–Newton algorithm. In practice, however, about one third of the data yield one or more nonpositive estimates of the diagonal elements Ψ_{ii} of the matrix Ψ, which are individual variances. These solutions are called Heywood cases and to deal with them, (2.8) or (2.9) should be minimized under conditions $\Psi_{ii} \geq 0$, i = 1, . . . , n. Thus the appropriate formulation defines f as follows:

$$f(\Sigma, S) = f_1(\Sigma, S) \quad \text{if} \quad \Psi_{ii} \geq 0 \ , \quad i = 1, \ldots, n$$

$= +\infty$ otherwise

and similarly for f_2.

EXAMPLE 2.3 *Negative estimates of variance components.* Consider a general linear model with random effects

$$y = Z\gamma + \sum_{i=1}^{p} X_i \beta_i + \varepsilon \tag{2.10}$$

where $y(\nu, 1)$ is the vector of observations on the variable y, $Z(\nu, r)$, $X_i(\nu, r_i)$, $i = 1, \ldots, p$, are mutually uncorrelated random vectors with $E\beta_i = 0$, var $\beta_i = \sigma_i^2 I_{r_i}$, $i = 1, \ldots, p$ and $E\varepsilon = 0$, var $\varepsilon = \sigma_0^2 I_\nu$, and $\gamma_1, \ldots, \gamma_r$, $\sigma_0^2, \ldots, \sigma_p^2$ are unknown parameters to be estimated.

One of the simplest examples is the following variance analysis model for random effect one-way classification: Consider k populations where the j-th measurement (observation) in the i-th population is given by

$$y_{ij} = \mu + a_i + e_{ij} \ , \quad j = 1, \ldots, n \ , \quad i = 1, \ldots, k \ . \tag{2.11}$$

In (2.11), μ is the fixed effect, a_i, $i = 1, \ldots, k$, is the random effect of the i-th population and e_{ij} is residual. Random variables $a_1, \ldots, a_k$ and $e_{11}, \ldots, e_{kn}$ are independent with distributions $N(0, \sigma_a^2)$ and $N(0, \sigma_e^2)$, respectively. The parameters μ, σ_a^2, σ_e^2 are to be estimated. The traditional estimates of the variance components σ_a^2, σ_e^2 in model (2.11) are obtained by a simple procedure: one equates the mean squares

$$\frac{1}{k(n-1)} S_e = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$$

and

$$\frac{1}{k-1} S_a = \frac{1}{k-1} \sum_{i=1}^{k} n(\bar{y}_{i.} - \bar{y}_{..})^2 \ ,$$

where $\bar{y}_{i.} = \frac{1}{n} \sum_{j=1}^{n} y_{ij}$, $i = 1, \ldots, k$, and $\bar{y}_{..} = \frac{1}{nk} \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}$, with their expectations σ_e^2 and $\sigma_a^2 n + \sigma_e^2$ that give the estimates

$$s_e^2 = \frac{1}{k(n-1)} S_e \tag{2.12}$$

$$s_a^2 = \frac{1}{n} \left[\frac{1}{k-1} S_a - s_e^2 \right] \ . \tag{2.13}$$

Whereas s_e^2 is evidently nonnegative, this need not be the case of s_a^2, so that the problem of negative estimate of the variance component s_a^2 comes to the fore.

The resulting estimates (2.12), (2.13) of the variance components in (2.11) follow also as a special result of the MIVQUE and MINQUE estimation developed for the general model (2.10): Unbiased estimates of a linear parametric function $\sum_{l=0}^{p} \sigma_l^2 q_l$ are sought in the form $y^T A y$ where

$$AZ = 0,\ A(\nu,\ \nu) \text{ is symmetric matrix} \tag{2.14}$$

and which are optimal in some sense. The MIVQUE estimates correspond to a matrix A that minimizes the variance of $y^T A y$ subject to the conditions (2.14) and the MINQUE estimates correspond to a matrix A that minimizes $\mathrm{tr}(A(I + \sum_{l=1}^{p} X_l X_l^T))^2$ subject to conditions (2.14). In none of the mentioned approaches, however, the natural nonnegativity constraints on the estimates of the variances σ_l^2, $i = 1, \ldots, p$, are introduced explicitly.

Again, there are two possible explanations of negative estimates of variance components: the model may be incorrect or a statistical noise obscured the underlaying situation. Among others, Herbach (1959) and Thompson (1962) studied variance analysis models with random effects by means of different variants of the maximum likelihood method under nonnegativity constraints. Correspondingly, in terms of the general model, we have for instance

$$f(\sigma_a^2,\ \sigma_e^2,\ \mu,\ y) = (\pi)^{-\frac{nk}{2}} (\sigma_e^2 + n\sigma_a^2)^{-\frac{k}{2}} (\sigma_e^2)^{-\frac{k(n-1)}{2}} \cdot$$

$$\cdot \exp\left\{ -\frac{1}{2\sigma_e^2} \left[\sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \mu)^2 - \frac{\sigma_a^2}{\sigma_e^2 + n\sigma_a^2} \sum_{i=1}^{k} \left[\sum_{j=1}^{n} y_{ij} - u\mu \right]^2 \right] \right\}$$

$$\text{if } \sigma_a^2 \geq 0,\ \sigma_e^2 \geq 0$$

$$= -\infty \quad \text{otherwise} ,$$

Similarly, nonnegative MINQUE and MIVQUE estimates are of interest.

EXAMPLE 2.4 *M–estimates.* Let Θ be a given locally compact parameter set, (Ξ, A, P) a probability space and $f: \Theta \times \Xi \to R$ a given function. For a sample $\{\xi_1, \ldots, \xi_\nu\}$ from the considered distribution, any estimate $T^\nu = T^\nu(\xi_1, \ldots, \xi_\nu) \in \Theta$ defined by condition

$$T^\nu \in \operatorname{argmin} \sum_{j=1}^{\nu} f(T\,\xi_j) \tag{2.15}$$

is called an M-estimate. In the pioneering paper by Huber (1967) (see also Huber (1981)), nonstandard sufficient conditions were given under which $\{T^\nu\}$ converges a.s. (or in probability) to a constant $\vartheta_0 \in \Theta$ and asymptotic normality of $\sqrt{\nu}(T^\nu - \vartheta_0)$ was proved under assumption that Θ is an open set.

The problem (2.15) is evidently a special case of our general framework; the P^ν again correspond to the empirical distribution functions and we have unconstrained criterion function. We shall aim to remove both of these assumptions to get results valid for a whole class of probability measures P^ν estimating P, which contains the empirical probability measure connected with the original definition (2.15) of M-estimates, and for constrained estimates.

EXAMPLE 2.5 *Stochastic optimization with incomplete information.* Consider the following decision model of stochastic optimization:

Given a probability space (Ξ, A, P), a random element ξ on Ξ, a measurable function $f: R^n \times \Xi \to R$ and a set $S \subset R^n$

$$\text{minimize } E\{f(x, \xi)\} = \int_{\Xi} f(x, \xi) P(d\xi) \text{ on the set } S \subset R^n \ . \tag{2.16}$$

A wide variety of stochastic optimization problems, e.g., stochastic programs with recourse or probability constrained models (see e.g. Dempster (1980), Ermoliev et al. (1985), Kall (1976), Prékopa (1973), Wets (1983)) fit into this abstract framework.

In many practical situations, however, the probability measure P need not be known completely. One possibility how to deal with such a situation is to estimate the optimal solution x^* of (2.16) by an optimal solution of the problem

$$\text{minimize } \int_{\Xi} f(x, \xi)\, P^\nu(d\xi) \text{ on the set } S \subset R^n$$

where P^ν is a suitable estimate of P based on the observed dates. In this context, there are different possibilities to estimate or approximate P and the use of empirical distribution is only one of them. The case of P belonging to a given parametric family of probability measures but with an unknown parameter vector was studied e.g. in Dupačová (1984a, b).

For problem (2.16), large dimensionality of the decision vector x is typical. This circumstance together with nondifferentiability (or even with noncontinuity) of f and with the presence of constraints raises qualitatively new problems.

3. CONSISTENCY: CONVERGENCE OF OPTIMAL SOLUTIONS

From a conceptual viewpoint or for theoretical purposes, it is convenient as well as expedient to study problems of statistical estimation as well as stochastic optimization problems with partial information, in the following general framework. Let (Ξ, A, P) be a probability space, with Ξ – the support of P – a closed subset of a Polish space X, and A the Borel sigma-field relative to Ξ; we may think of Ξ as the set of possible values of the random element ξ defined on the probability space of events $(\Omega, A', \tilde{P}')$. If P is known, the problem is to:

$$\text{find } x^* \in R^n \text{ that minimizes } Ef(x) \ , \tag{3.1}$$

where

$$Ef(x) := \int_{\Xi} f(x, \xi) P(d\xi)) = E\{f(x, \xi)\} \tag{3.2}$$

and

$$f : R^n \times \Xi \longrightarrow R \cup \{\infty\} = (-\infty, \infty)$$

is a random lower semicontinuous function; we set

$$(Ef)(x) = \infty \ ,$$

whenever $\xi \mapsto f(x, \xi)$ is not bounded above by a summable (extended real-valued) function. We refer to

$$\text{dom } Ef := \{x \mid Ef(x) < \infty\}$$

as the *effective domain* of Ef. Points that do not belong to dom Ef cannot minimize Ef and thus are effectively excluded from the optimization problem (3.1). Hence, the model makes specific provisions for the presence of constraints that may limit the choice of x. Note that by definition of the integral, we always have

$$\text{dom } Ef \subset \{x \mid f(x, \xi) < \infty \text{ a.s.}\} \ .$$

An extended real-valued function $h : R^n \longrightarrow \bar{R} = [-\infty, \infty]$ is said to be *proper* if

$h > -\infty$ and not identically $+\infty$; it is *lower semicontinuous* (*l.sc.*) at x if for any sequence $(x^k)_{k=1}^{\infty}$, converging to x

$$\liminf_{k \to \infty} h(x^k) \geq h(x) \ ,$$

where the quantities involved could be ∞ or $-\infty$. The extend real-valued function f defined on $R^n \times \Xi$ is a *random lower semicontinuous function* if

$$\text{for all } \xi \in \Xi,\ f(\cdot, \xi) \text{ is l.sc.} \tag{3.3i}$$

$$f \text{ is } B^n \otimes A \text{ - measurable} \tag{3.3ii}$$

where B^n is the Borel sigma-field on R^n. This concept, under the name of "normal integrand", was introduced by Rockafellar (1976), as a generalization of Caratheodory integrands, to handle problems in the Calculus of Variations and Optimal Control Theory. When dealing with problems of that type, as well as stochastic optimization problems such as (3.1), the traditional tools of functional analysis are no longer quite appropriate. The classical geometrical approach that associates functions with their graph must be abandoned in favor of a new geometrical viewpoint that associates functions with their "epigraphs" (or hypographs), for more about the motivation and the underlying principles of the epigraphical approach consult Rockafellar and Wets (1984). The *epigraph* of a function $h : R^n \to \bar{R}$ is the set

$$\text{epi } h = \{(x, \alpha) \in R^n \times R \mid h(x) \leq \alpha\} \ .$$

Rockafellar (1976) shows that $f : R^n \times \Xi \to \bar{R}$ is a random l.sc. function if and only if

$$\text{the multifunction } \xi \mapsto \text{ epi } f(\cdot, \xi) \text{ is nonempty, closed-valued} \ , \tag{3.4i}$$

$$\text{the multifunction } \xi \mapsto \text{ epi } f(\cdot, \xi) \text{ is measurable} \ ; \tag{3.4ii}$$

recall that a multifunction $\xi \mapsto \Gamma(\xi) : \Xi \to R^{n+1}$ is measurable if for all closed sets $F \subset R^{n+1}$

$$\Gamma^{-1}(F) := \{\xi \in \Xi \mid \Gamma(\xi) \cap F \neq \phi\} \in A$$

for further details about measurable multifunctions see Rockafellar (1976), Castaing and Valadier (1976), and the bibliography of Wagner (1977) supplemented by Ioffe (1978). We shall use repeatedly the following result due to Yankov, von Neuman, and Kuratowski and Ryll Nardzewski.

PROPOSITION 3.1 Theorem of Measurable Selections. *If* $\Gamma: \Xi \rightrightarrows R^n$ *is a closed-valued measurable multifunction, then there exists a least one measurable selector, i.e. a measurable function* $x: \text{dom } \Gamma \to R^n$ *such that for all* $\xi \in \text{dom } \Gamma$, $x(\xi) \in \Gamma(\xi)$, *where* $\text{dom } \Gamma := \{\xi \in \Xi | \Gamma(\xi) \neq \phi\} = \Gamma^{-1}(R^N) \in A$.

For a proof see Rockafellar (1976), for example. As immediate consequences of the definition (3.3) of random l.sc. functions, the equivalence with the conditions (3.4) and the preceding proposition, we have:

PROPOSITION 3.2 *Let* $f: R^n \times \Xi \to \bar{R}$ *be a random l.sc. function. Then for any A measurable function* $x: \Xi \to R^n$, *the function*

$$\xi \mapsto f(x(\xi), \xi) \text{ is } A\text{-measurable} .$$

Moreover, the infimal function

$$\xi \mapsto \inf f(\cdot, \xi) := \inf_{x \in R^n} f(x, \xi)$$

is A-measurable, and the set of optimal solution

$$\xi \mapsto \text{argmin } f(\cdot, \xi) := \{x | f(x, \xi) = \inf f(\cdot, \xi)\}$$

is a closed-valued measurable multifunction from Ξ *into* R^n, *and this implies that there exists a measurable function*

$$\xi \mapsto x^*(\xi): \text{dom } (\text{argmin } f(\cdot, \xi)) \rightrightarrows R^n$$

such that $x^*(\xi)$ *minimizes* $f(\cdot, \xi)$ *whenever* $\text{argmin } f(\cdot, \xi) \neq \phi$.

For a succinct proof, see Section 3 of Rockafellar and Wets (1984).

If instead of P, we only have limited information available about P – e.g. some knowledge about the shape of the distribution and a finite sample of values of $\underset{\sim}{\xi}$ or of a function of $\underset{\sim}{\xi}$ – then to estimate x^* we usually have to rely on the solution of an optimization problem that "approximates" (3.1), viz.

$$\text{find } x^\nu \in R^n \text{ that minimizes } E^\nu f(x) \tag{3.5}$$

where

$$E^\nu f(x) := E^\nu\{f(x, \xi)\} = \int_\Xi f(x, \xi) P^\nu(d\xi) . \tag{3.6}$$

The measure P^ν is not necessarily the empirical measure, but more generally the

"best" (in terms of a given criterion) approximate to P on the basis of the information available. As more information is collected, we could refine the approximation to P and hopefully find a better estimate of x^*. To model this process, we rely on the following set-up: let (Z, F, μ) be a sample space with $(F^\nu)_{\nu=1}^\infty$ an increasing sequence of sigma-field contained in F. A sample ζ -- e.g. $\zeta = \{\xi^1, \xi^2, \dots\}$ obtained by independent sampling of the values of $\underset{\sim}{\xi}$ -- leads us to a sequence $\{P^\nu(\cdot, \zeta), \nu = 1, \dots\}$ of probability measures defined on (Ξ, A). Since only the information collected up to stage ν can be used in the choice of P^ν, we must also require that for all $A \in A$

$$\zeta \mapsto P^\nu(A, \zeta) \quad \text{is } F^\nu\text{-measurable} .$$

Since P^ν depends on ζ, so does the approximate problem (3.5), in particular its solution x^ν. A sequence of estimators

$$\{x^\nu : Z \to R^n, \nu = 1, \dots\}$$

is (strongly) *consistent* if μ-almost surely they converge to x^*, this, of course, implies weak consistency (convergence in probability).

The following results extend the classical Consistency Theorem of Wald (1940) and the extensions by Huber (1967), to the more general setting laid out here above. Consistency is obtained by relying on assumptions that are weaker than those of Huber (1967) even in the unconstrained case. To do so, we rely on the theory of epi-convergence in conjunction with the theory of random sets (measurable multifunctions) and random l.sc. functions.

A sequence of functions $\{g^\nu : R^n \to \bar{R}, \nu = 1, \dots\}$ is said to *epi-converge* to $g : R^n \to \bar{R}$ if for all x in R^n, we have

$$\liminf_{\nu \to \infty} g^\nu(x^\nu) \geq g(x) \quad \textit{for all} \quad \{x^\nu\}_{\nu=1}^\infty \quad \textit{converging to} \quad x , \tag{3.7}$$

and

$$\textit{for some} \quad \{x^\nu\}_{\nu=1}^\infty \quad \textit{converging to} \quad x, \limsup_{\nu \to \infty} g^\nu(x^\nu) \leq g(x) . \tag{3.8}$$

Note that any one of these conditions imply that g is lower semicontinuous. We then say that g is the *epi-limit* of the g^ν, and write $g = \text{epi}-\lim_{\nu \to \infty} g^\nu$. We refer to this type of convergence as epi-convergence, since it is equivalent ot the set-convergence of the epigraphs. For more about epi-convergence and its properties, consult Attouch (1984). Our interest in epi-convergence stems from the fact that

from a variational viewpoint it is the weakest type of convergence that possesses the following properties:

PROPOSITION 3.3 [Attouch and Wets (1981), Salinetti and Wets (1986)]. *Suppose* $\{g; g^\nu: R^n \to \bar{R}, \nu = 1,\dots\}$ *is a collection of functions such that* $g = \text{epi}-\lim_{\nu\to\infty} g^\nu$. *Then*

$$\limsup_{\nu\to\infty} (\inf g^\nu) \leq \inf g \ , \tag{3.9}$$

and, if

$$x^k \in \operatorname{argmin} g^{\nu_k} \quad \textit{for some subsequence } \{\nu_k, k = 1,\dots\}$$

and $x = \lim_{k\to\infty} x^k$, *it follows that*

$$x \in \operatorname{argmin} g \ ,$$

and

$$\lim_{k\to\infty} (\inf g^{\nu_k}) = \inf g \ ;$$

so in particular if there exists a bounded set $D \subset R^n$ *such that for some subsequence* $\{\nu_k, k = 1,\dots\}$,

$$\operatorname{argmin} g^{\nu_k} \cap D \neq \phi \ ,$$

then the minimum of g *is attained at some point in the closure of* D.

Moreover, if $\operatorname{argmin} g \neq \phi$, *then* $\lim_{\nu\to\infty}(\inf g^\nu) = \inf g$ *if and only if* $x \in \operatorname{argmin} g$ *implies the existence of sequences* $\{\varepsilon_\nu \geq 0, \nu = 1,\dots\}$ *and* $\{x^\nu \in R^n, \nu = 1,\dots\}$ *with*

$$\lim_{\nu\to\infty} \varepsilon_\nu = 0, \quad \text{and} \quad \lim_{\nu\to\infty} x^\nu = x$$

such that for all $\nu = 1,\dots$

$$x^\nu \in \varepsilon_\nu - \operatorname{argmin} g^\nu := \{x \mid g^\nu(x) \leq \varepsilon_\nu + \inf g^\nu\} \ .$$

The next theorem that proves the μ-almost sure epi-convergence of expectation functionals, is build upon approximation results for stochastic optimization problems, first derived in the case $f(\cdot, \xi)$ convex (Theorem 3.3, Wets (1984)), and later for the locally Lipschitz case (Theorem 2.8, Birge and Wets (1986)). We work with the following assumptions.

ASSUMPTION 3.4 "Continuities" of f. *The function*

$$f : R^n \times \Xi \to (-\infty, \infty]$$

with

$$\operatorname{dom} f := \{(x, \xi) | f(x, \xi) < \infty\} = S \times \Xi,\ S \subset R^n \text{ closed and nonempty} ,$$

is such that for all $x \in S$,

$$\xi \mapsto f(x, \xi) \quad \text{is continuous on } \Xi ,$$

and for all $\xi \in \Xi$

$$x \mapsto f(x, \xi) \quad \text{is l.sc. on } R^n ,$$

and locally lower Lipschitz on S, *in the following sense: to any* x *in* S, *there corresponds a neighborhood* V *of* x *and a bounded continuous function* $\beta : \Xi \to R$ *such that for all* $x' \in V \cap S$ *and* $\xi \in \Xi$,

$$f(x, \xi) - f(x', \xi) \leq \beta(\xi) \cdot \|x - x'\| . \tag{3.10}$$

ASSUMPTION 3.5 Convergence in distribution. *Given the sample space* (Z, F, μ) *and an increasing sequence of sigma-fields* $(F^\nu)_{\nu=1}^\infty$ *contained in* F, *let*

$$P^\nu : A \times Z \to [0, 1], \quad \nu = 1, \ldots$$

be such that for all $\zeta \in Z$

$$P^\nu(\cdot\ , \zeta) \text{ is a probability measure on } (\Xi, A) ,$$

and for all $A \in A$

$$\zeta \mapsto P^\nu(A, \zeta) \quad \text{is } F^\nu\text{-measurable} .$$

For μ*-almost all* ζ *in* Z, *the sequence*

$$\{P^\nu(\cdot\ , \zeta),\ \nu = 1, \ldots\} \quad \text{converges in distribution to } P ,$$

and with $P =: P^0(\cdot\ , \zeta)$, *for all* $x \in S$, *the sequence* $\{P^\nu(\cdot\ , \zeta)\}_{\nu=0}^\infty$ *is* $f(x, \cdot)$*-tight (asymptotic negligibility), i.e. to every* $x \in S$ *and* $\varepsilon > 0$ *there corresponds a compact set* $K_\varepsilon \subset \Xi$ *such that for* $\nu = 0, 1, \ldots$

$$\int_{\Xi \setminus K_\varepsilon} |f(x, \xi)| P^\nu(d\xi, \zeta) < \varepsilon , \tag{3.11}$$

and

$$\int_{\Xi} \inf_{x \in R^n} f(x, \xi) P^{\nu}(d\xi, \xi) > -\infty \ . \tag{3.12}$$

The assumption that

$$\xi \mapsto \text{dom } f(\cdot, \xi) := \{x \mid f(x, \xi) < \infty\} = S$$

is constant, which is satisfied by all the examples in Section 2, may appear more restrictive than it actually is. Indeed, it is easy to see that

$$\text{dom } Ef = \bigcap_{\xi \in \Xi} \text{dom } f(\cdot, \xi) \ ,$$

if Ξ is the support of the measure P and for all $x \in \bigcap_{\xi \in \Xi} \text{dom } f(\cdot, \xi)$, the function $f(x, \cdot)$ is bounded above by a summable function. Then, with $S = \bigcap_{\xi \in \Xi} \text{dom } f(\cdot, \xi)$ and

$$f^+(x, \xi) = \begin{cases} f(x, \xi) & \text{if } x \in S \\ +\infty & \text{otherwise} \end{cases} \ ,$$

we may as well work with f^+ instead of f, since

$$Ef(x) = Ef^+(x) = E\{f^+(x, \xi)\} \ ,$$

and now $\xi \mapsto \text{dom } f^+(\cdot, \xi) = S$ is constant.

Assumption 3.4 implies that f is a random lower semicontinuous function (normal integrand). Indeed, for all $\xi \in \Xi$, $f(\cdot, \xi)$ is proper and lower semicontinuous (3.3.i) and $(x, \xi) \mapsto f(x, \xi)$ is $B^n \otimes A$-measurable (3.3.ii) since for all $\alpha \in R$,

$$\text{lev}_{\alpha} f := \{(x, \xi) \mid f(x, \xi) \leq \alpha\} \quad \text{is closed} \ .$$

To see this, suppose $\{(x^k, \xi^k)\}_{k=1}^{\infty} \subset \text{lev}_{\alpha} f$ is a sequence converging to (x, ξ); then from Assumption 3.4 we have that for k sufficiently large, and all ξ

$$f(x, \xi) \leq f(x^k, \xi) + \beta(\xi) \|x - x^k\| \ ,$$

in particular

$$f(x, \xi^k) \leq f(x^k, \xi^k) + \beta \|x - x^k\| \leq \alpha + \beta \|x - x^k\|$$

where $\beta = \max_{\xi \in \Xi} \beta(\xi)$ is finite, since $\beta(\cdot)$ is bounded. Now $\xi \mapsto f(x, \xi)$ is continuous on Ξ, thus taking limits as k goes to ∞, we obtain

$$f(x, \xi) \leq \alpha + \beta \lim_{k \to \infty} \|x - x^k\| = \alpha \ ,$$

i.e. $(x, \xi) \in \mathrm{lev}_\alpha f$. Since f is a random l.sc. function if follows from Proposition 3.2 that

$$\xi \mapsto \inf_{x \in R} f(x, \xi) =: \gamma(\xi)$$

is measurable. Thus condition (3.12) does not sneak in another measurability condition, it requires simply that the measurable function γ be quasi-integrable.

Huber (1967), as well as others see e.g. Ibragimov and Has'minski (1981), assumes that S is open. Since constraints usually do not involve strict inequalities, this is an unnatural restriction, except when there are no constraints, i.e. $S = R^n$ in which case S is also closed. In any case, whatever be the optimality results one may be able to prove with S open, they remain valid when S is replaced by its closure, assuming minimal continuity properties for the expectation functionals, but the converse does not hold.

To simplify notations we shall, whenever it is convenient, drop the explicit reference of the dependence on ζ of the probability measures P^ν and the resulting expectation functionals $E^\nu f$, nonetheless the reader should always be aware that all μ-a.s. statements refer to the underlying probability space (Z, F, μ). We begin by showing that Ef, as well as the $E^\nu f$, are well-defined functions.

LEMMA 3.6 *Under Assumptions* 3.4 *and* 3.5, *there exists* $Z_0 \in F$, $\mu(Z_0) = 1$ *such that for all* $\zeta \in Z_0$, Ef *and* $\{E^\nu f,\ \nu = 1, \dots\}$ *are proper lower semicontinuous functions such that*

$$S = \mathrm{dom}\ Ef = \mathrm{dom}\ E^\nu f(\cdot, \zeta)$$

on which the expectation functionals are finite.

PROOF Let us first fix ζ, and assume that for this ζ all the conditions of Assumption 3.5 are satisfied. If $x \notin S$, then $f(x, \xi) = \infty$ for all ξ in Ξ and hence $Ef = E^\nu f = \infty$, i.e.,

$$S \supset \mathrm{dom}\ Ef, \quad S \supset \mathrm{dom}\ E^\nu f\ .$$

With $P^0 = P$, for $x \in S$ and any $\varepsilon > 0$, there is a compact set K_ε (Assumption 3.5) such that

$$\int_\Xi f(x, \xi) P^\nu(d\xi) \le (\max_{\xi \in K_\varepsilon} |f(x, \xi)|) \cdot P^\nu(K_\varepsilon)$$

$$+ \int_{\Xi \setminus K_\varepsilon} |f(x, \xi)| P^\nu(d\xi) < \infty\ ,$$

as follows from (3.11) and the fact that $f(x, \cdot)$ is continuous and finite on $K_\varepsilon \subset \Xi$. Thus $E^\nu f(x) < \infty$.

The fact that $Ef > -\infty$, and $E^\nu f > -\infty$ follows directly from condition (3.12). It is also this condition that we use to show that the expectation functionals are lower semicontinuous since it allows us to appeal to Fatou's Lemma to obtain: given $\{x^\nu\}_{\nu=1}^\infty$ a sequence converging to x;

$$\liminf_{\nu \to \infty} Ef(x^\nu) \geq \int \liminf_{\nu \to \infty} f(x^\nu, \xi) P(d\xi)$$

$$\geq \int f(x, \xi) P(d\xi) = Ef(x)$$

where the last inequality follows from the lower semicontinuity of $f(\cdot, \xi)$ at x. Of course, the same string of inequalities holds for all $\{P^\nu, \nu = 1, \dots\}$.

Since the above holds for every ν μ-almost surely on Z, the set

$$Z_0 = \{\zeta \in Z | E^\nu f(\cdot, \zeta) \quad \text{is finite, l.sc. on } S, \quad \text{for } \nu = 0, 1, \dots\}$$

is of measure 1.□

THEOREM 3.7 *Suppose* $\{E^\nu f, \nu = 1, \dots\}$ *is a sequence of expectation functionals defined by*

$$E^\nu f(x) = \int_\Xi f(x, \xi) P^\nu(d\xi) = E^\nu\{f(x, \xi)\}$$

and $Ef(x) = E\{f(x, \xi)\}$ *such that* f *and the collection* $\{P; P^\nu, \nu = 1, \dots\}$ *satisfy Assumptions 3.4 and 3.5. Then,* μ*-almost surely*

$$Ef = \operatorname{epi}\text{-}\lim_{\nu \to \infty} E^\nu f = \operatorname{ptwse}\text{-}\lim_{\nu \to \infty} E^\nu f$$

where $\operatorname{ptwse}\text{-}\lim_{\nu \to \infty} E^\nu f$ *denotes the pointwise limit.*

PROOF The argument essentially follows that of Theorem 2.8 Birge and Wets (1986), with minor modifications to take care of the slightly weaker assumptions and the fact that the expectation functionals depend on ζ. We begin by showing that μ-almost surely Ef is the pointwise limit of the $E^\nu f$. We fix $\zeta \in Z$, and assume that the conditions of Assumption 3.5 are satisfied for this particular ζ. Suppose $x \in S$, and set

$$h(\xi) := f(x, \xi) \ .$$

From condition (3.11), it follows that for all $\varepsilon > 0$, there is a compact set K_ε such that for all ν

$$\int_{\Xi \setminus K_\varepsilon} |h(\xi)| P^\nu(d\xi) < \varepsilon \ .$$

Let $\gamma_\varepsilon := \max_{\xi \in K_\varepsilon} |h(\xi)|$. We know that γ_ε is finite since K_ε is compact and h is continuous on Ξ (Assumption 3.4). Let h^ε be a truncation of h, defined by

$$h^\varepsilon(\xi) = \begin{cases} h(\xi) & \text{if } |h(\xi)| \le \gamma_\varepsilon \\ \gamma_\varepsilon & \text{if } h(\xi) > \gamma_\varepsilon \\ -\gamma_\varepsilon & \text{if } h(\xi) < \gamma_\varepsilon \end{cases}$$

The function h^ε is bounded and continuous, and for all ξ in Ξ

$$|h^\varepsilon(\xi)| \le |h(\xi)| \ .$$

Now, from the convergence in distribution of the P^ν,

$$\lim_{\nu \to \infty} \left[\alpha_\nu^\varepsilon := \int_\Xi h^\varepsilon(\xi) P^\nu(d\xi) \right] = \int_\Xi h^\varepsilon(\xi) P(d\xi) := \alpha^\varepsilon \ . \tag{3.13}$$

Moreover, for all ν

$$\int_{\Xi \setminus K_\varepsilon} h^\varepsilon(\xi) P^\nu(d\xi) < \varepsilon \ .$$

Now, let

$$\alpha_\nu := E^\nu f(x) = \int_{K_\varepsilon} h(\xi) P^\nu(d\xi) + \int_{\Xi \setminus K_\varepsilon} h(\xi) P^\nu(d\xi) \ .$$

We have that for all ν

$$|\alpha_\nu - \alpha_\nu^\varepsilon| = \left| \int_{\Xi \setminus K_\varepsilon} (h(\xi) - h^\varepsilon(\xi)) P^\nu(d\xi) \right| < 2\varepsilon \ ,$$

and also

$$|Ef(x) - \alpha^\varepsilon| < 2\varepsilon \ .$$

These two last estimates, when used in conjunction with (3.13) yield: for all $\varepsilon > 0$

$$|Ef(x) - \alpha_\nu| < 6\varepsilon \ .$$

Thus for all x in S

$$Ef(x) = \lim_{\nu \to \infty} E^\nu f(x) = \lim_{\nu \to \infty} \alpha_\nu \ ,$$

and since, by Lemma 3.6,

$$S = \text{dom } Ef = \text{dom } E^\nu f \ ,$$

it means that $Ef = \text{ptwse}-\lim_{\nu\to\infty} E^\nu f$, and that condition (3.8) of epi-convergence is satisfied, since we can choose $\{x^\nu = x\}_{\nu=1}^\infty$ for the sequence converging to x.

There remains to verify condition (3.7) of epi-convergence. If $x \notin S$, then for every sequence $\{x^\nu\}_{\nu=1}^\infty$ converging to x, since S is closed we have that $x^\nu \notin S$ for ν sufficiently large and hence $E^\nu f(x^\nu) = \infty$, which implies that

$$\liminf_{\nu\to\infty} E^\nu f(x^\nu) = \infty \geq Ef(x) = \infty \ .$$

If $x \in S$, and $\{x^\nu\}_{\nu=1}^\infty$ is a sequence converging to x, unless x^ν is in S infinitely often, $\liminf_{\nu\to\infty} E^\nu f(x^\nu) = \infty$, and then condition (3.7) is trivially satisfied. So let us assume that $\{x^\nu\}_{\nu=1}^\infty \subset S$. For ν sufficiently large, from (3.10) it follows that there is a bounded continuous function β such that

$$f(x, \xi) - \beta(\xi) \cdot \|x - x^\nu\| \leq f(x^\nu, \xi) \ .$$

Integrating both sides with respect to P^ν, and taking $\liminf_{\nu\to\infty}$, we obtain

$$\lim_{\nu\to\infty} E^\nu f(x) - \lim_{\nu\to\infty} \beta^\nu \cdot \|x - x^\nu\| \leq \liminf_{\nu\to\infty} E^\nu f(x^\nu)$$

where $\beta^\nu = \int \beta(\xi) P^\nu(d\xi)$ converge to a finite limit since the P^ν converge in distribution to P, and by pointwise convergence of the $E^\nu f$ this yields

$$Ef(x) \leq \liminf_{\nu\to\infty} E^\nu f(x^\nu) \ . \ \square$$

To apply in this context, Propositions 3.2 and 3.3, we must show that the expectation functionals $\{E^\nu f, \ \nu = 1, \dots\}$ are random l.sc. functions.

THEOREM 3.8 *Under* Assumptions 3.4 *and* 3.5, *the expectation functionals*

$$E^\nu f : R^n \times Z \to \overline{R}, \quad \text{for } \nu = 1, \dots,$$

are μ-almost surely random lower semicontinuous functions, such the $\zeta \mapsto$ epi $E^\nu f(\cdot, \zeta)$ *is F^ν-measurable.*

PROOF Lemma 3.6 shows that there exists a set $Z_0 \subset Z$ of μ-measure 1 such that for all $\zeta \in Z_0$, the multifunction

$$\zeta \mapsto \text{epi } E^\nu f(\cdot, \zeta) : Z_0 \rightrightarrows R^{n+1} \text{ is nonempty, closed-valued} \ .$$

This is condition (3.4.i), thus there remains only to establish (3.4.ii), i.e.

$$\zeta \mapsto \text{epi } E^\nu f(\cdot, \zeta) \text{ is } F^\nu\text{-measurable} \ .$$

for $\nu = 1,\ldots$. Theorem 3.7 proves that with respect to the topology of convergence in distribution, the map

$$P^\nu \mapsto \text{epi } E^\nu f \text{ is continuous .}$$

Moreover, since $\zeta \mapsto P^\nu(A, \zeta)$ is F^ν-measurable for all $A \in \mathcal{A}$, it means that given any finite collection of closed sets $\{F_i \subset \Xi\}_{i=1}^q$ and scalars $\{\beta_i\}_{i=1}^q \subset [0, 1]$, the set

$$\{\zeta \in Z | P^\nu(F_i, \zeta) < \beta_i, i = 1, \ldots, q\} \in F^\nu$$

which means that the function

$$\zeta \mapsto P^\nu(\cdot, \zeta): Z \to \mathcal{P} := \{\text{probability measures on } (\Xi, \mathcal{A})\}$$

is F^ν-measurable. To see this, observe that the "convergence in distribution"-topology can be obtained from the base of open sets

$$\{Q \in \mathcal{P} | Q(F_i) < \beta_i, i = 1, \ldots, k\} ,$$

see Billingsley (1968), that also generate the Borel field on $\mathcal{P}$. Thus

$$\zeta \mapsto \text{epi } E^\nu f(\cdot, \zeta)$$

is the composition of a continuous function, and a F^ν-measurable function, and hence is F^ν-measurable.□

In the proof of Theorem 3.8, we have used the continuity of the map $P^\nu \mapsto$ epi $E^\nu f$, in fact Theorem 3.7 only proves epi-convergence, without introducing explicitly the epi-topology for the space of lower semicontinuous functions. The fact that epi-convergence induces a topology on the space of l.sc. functions is well-established, see for example Dolecki, Salinetti and Wets (1983) and Attouch (1984), and thus with this proviso, Theorem 3.7 proves the epi-continuity of the map $P^\nu \mapsto$ epi $E^\nu f$.

THEOREM 3.9 Consistency. *Under* Assumptions 3.4 *and* 3.5 *we have that* μ-*almost surely*

$$\limsup_{\nu \to \infty} (\inf E^\nu f) \leq \inf Ef \tag{3.14}$$

Moreover, there exists $Z_0 \in F$ *with* $\mu(Z \setminus Z_0) = 0$, *such that*

(i) *for all* $\zeta \in Z_0$, *any cluster point* $\hat{x}$ *of any sequence* $\{x^\nu, \nu = 1,\ldots\}$ *with* $x^\nu \in$ argmin $E^\nu f^\nu(\cdot, \zeta)$ *belongs to* argmin Ef *(i.e. is an optimal estimate),*

(ii) *for* $\nu = 1,\ldots$

$$\zeta \mapsto \operatorname{argmin} E^{\nu} f(\cdot, \zeta) : Z_0 \rightrightarrows R^n \ ,$$

is a closed-valued F^{ν}*-measurable multifunction.*

In particular, if there is a compact set $D \subset R^n$ *such that for* $\nu = 1,\ldots$

$$(\operatorname{argmin} E^{\nu} f) \cap D \ \textit{is nonempty } \mu\textit{-a.s.} \ ,$$

and

$$\{x^*\} = \operatorname{argmin} Ef \cap D \ ,$$

then there exist $\{x^{\nu} : Z_0 \to R^n\}_{\nu=1}^{\infty}$ F^{ν}*-measurable selections of* $\{\operatorname{argmin} E^{\nu} f\}_{\nu=1}^{\infty}$ *such that*

$$x^* = \lim_{\nu \to \infty} x^{\nu}(\zeta) \quad \textit{for} \quad \mu\textit{-almost all } \zeta \ ,$$

and also

$$\inf Ef = \lim_{\nu \to \infty} (\inf E^{\nu} f) \quad \mu\textit{-a.s.} \ .$$

PROOF The inequality (3.14) immediately follows from (3.9) and the epi-convergence μ-almost surely of the expectation functionals $E^{\nu} f$ to Ef (Theorem 3.7) as does the assertion (i) about cluster points of optimal solutions (Proposition 3.2). The fact that $(\operatorname{argmin} E^{\nu} f)$ is a closed-valued F^{ν}-measurable multifunction follows from Theorem 3.8 and Proposition 3.2.

Now suppose $Z_0 \subset Z$ be such that $\mu(Z_0) = 1$, for all $\zeta \in Z_0$, $Ef = \operatorname{epi-lim}_{\nu \to \infty} E^{\nu} f$, and for all $\nu = 1,\ldots$, $(\operatorname{argmin} E^{\nu} f) \cap D$ is nonempty. For all ν, the multifunction

$$\zeta \mapsto (\operatorname{argmin} E^{\nu} f(\cdot, \zeta) \cap D) : Z_0 \rightrightarrows R^n$$

is nonempty compact-valued, and F^{ν}-measurable; it is the intersection of two closed-valued measurable multifunctions, see Rockafellar (1976). Now for any $\zeta \in Z_0$, let $\{\tilde{x}^{\nu}\}_{\nu=1}^{\infty}$ be any sequence in R^n such that for all ν,

$$\tilde{x}^{\nu}(\zeta) \in \operatorname{argmin} E^{\nu} f(\cdot, \zeta) \cap D \ .$$

Then, any cluster point of the sequence is in D, since it is compact, and in argmin Ef as follows from Proposition 3.2. Actually, $x^* = \lim_{\nu \to \infty} x^{\nu}$. To see this

note that, if x^* is not the limit point of the sequence there exists a subsequence $\{\nu_k\}_{k=1}^{\infty}$ such that for some $\delta > 0$, and all $k = 1, \ldots,$

$$\tilde{x}^k \in \operatorname{argmin} E^{\nu_k} f \cap D, \quad \text{and} \quad \|x^* - \tilde{x}^k\| > \delta \ ,$$

but this is contradicted by the fact that this subsequence included in D contains a further subsequence that is convergent.

Now, for $\nu = 1,\ldots,$ let $x^\nu : Z \to R^n$ be an F^ν-measurable selection of the F^ν-measurable multifunction $\zeta \mapsto (\operatorname{argmin} E^\nu f(\cdot, \zeta) \cap D)$, cf. Proposition 3.1. By the preceding argument for all $\zeta \in Z_0$, where $\mu(Z_0) = 1$,

$$x^* = \lim_{\nu \to \infty} x^\nu(\zeta)$$

and from Proposition 3.3, it then also follows that

$$\lim_{\nu \to \infty} (\inf E^\nu f(\cdot, \zeta)) = \inf Ef = Ef(x^*)$$

for all $\zeta \in Z_0$.□

It should be noted that contrary to earlier work – see Wald (1940), Huber (1967) – we do not assume the uniqueness of the optimal solutions, at least in the case of the stochastic programming model, introduced in section 2, this would not be a natural assumption. Also, let us observe that we have not given here the most general possible version of the Consistency Theorem that could be obtained by relying on the tools introduced here. There are conditions that are necessary *and* sufficient for the convergence of infima – see Salinetti and Wets (1986), Robinson (1985) – that could be used here in conjunction with convergence results for measurable selections (Salinetti and Wets (1981)) to yield a slightly sharper theorem, but the conditions would be much harder to verify, and would be of very limited interest in this context. Also, since epi-convergence is of local character, we could reward our statements to obtain "local" consistency by restricting our attention to a neighborhood of some x^* in argmin Ef.

We conclude by an existence result. A function $h : R^n \to \bar{R}$ is *inf-compact* if for all $\alpha \in R$

$$\operatorname{lev}_\alpha h := \{x \in R^n \mid h(x) \leq \alpha\} \text{ is compact} \ .$$

If h is proper ($h > -\infty$, $\operatorname{dom} h \neq 0$) and inf-compact, then ($\inf h$) is finite and attained for some $x \in R^n$. For example, if $h = g + \Psi_S$, where g is continuous and Ψ_S is

the indicator function of the nonempty compact set $S(\Psi_S(x) = 0$ if $x \in S$, and ∞ otherwise), then h is inf-compact. Another sufficient condition is to have g coercive. Inf-compactness is the most general condition that is verifiable under which existence can be established. The next proposition generalizes results of Wets (1973) and Hiriart-Unruty (1976). Essentially, we assume that $f(\cdot, \xi)$ is inf-compact with positive probability.

PROPOSITION 3.10 *Under* Assumptions 3.4 and 3.5, *the condition: there exists* $A \in \mathcal{A}$ *with* $P(A) > 0$ *(resp.* $P_\nu(A) > 0$*) such that for all* $\alpha \in R$, *the set*

$$\text{lev}_\alpha f \cap (R^n \times A) \text{ is bounded} .$$

Then Ef *is inf-compact (resp.* $E^\nu f$ *is* μ*-a.s. inf-compact).*

PROOF It clearly suffices to prove the proposition for P, the same argument applies for all P^ν μ-a.s.. Let

$$\gamma(\xi) := \inf\{0, \inf_{x \in R^n} f(x, \xi)\} .$$

The function is measurable (Proposition 3.2) and P-summable, see (3.12). The function f', defined by

$$f'(x, \xi) := f(x, \xi) - \gamma(\xi)$$

is then nonnegative. Moreover $f' \geq f$ and thus

$$\text{lev}_\alpha f' \cap (R^n \times A) \subset \text{lev}_\alpha f \cap (R^n \times A) .$$

Set $\alpha_1 := \alpha / P(A)$ and let A_1 be the projection on R^n of $\text{lev}_{\alpha_1} f' \cap (R^n \times A)$. Then if $x \notin A_1$ and $\xi \in A$

$$f'(x, \xi) > \alpha_1$$

and since f' is nonnegative, with $\bar{\gamma} = E\{\gamma(\xi)\}$,

$$Ef(x) = Ef'(x) + \bar{\gamma} \geq \int_A f'(x, \xi) P(d\xi) + \bar{\gamma}$$

$$> \alpha_1 \cdot P(A) + \bar{\gamma} = \alpha + \bar{\gamma}$$

Hence $\text{lev}_{\alpha + \bar{\gamma}} Ef \subset A_1$, a bounded set. To complete the proof it suffices to observe that from Lemma 3.6 we know that $\text{lev}_\alpha Ef$ is closed since Ef is lower semicontinuous, and this with the above implies that $\text{lev}_{\alpha + \bar{\gamma}} Ef$ is compact for all $\alpha \in R$.□

REFERENCES

Aitchinson, J. and S.V. Silvey (1958): Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Stat.* **29**, 813–828.

Arthanari, T.S. and Y. Dodge (1981): *Mathematical Programming in Statistics*. Wiley, New York.

Attouch, H. (1984): *Variational Convergence for Functions and Operators*, Research Notes in Mathematics, Pitman, London.

Attouch, H. and R. Wets (1981): Approximation and convergence in nonlinear optimization, in *Nonlinear Programming 4*, eds. O. Mangasarian, R. Meyer and S. Robinson, Academic Press, New York, 367–394.

Billingsley, P. (1968): *Convergence of probability measures*. J. Wiley, New YOrk.

Birge, J. and R. Wets (1986): Designing approximation schemes for stochastic problems, in particular for stochastic programs with recourse. *Math. Programming Study* **27**.

Castaing, C. and M. Valadier (1976): *Convex Analysis and Measurable Multifunctions*, Springer Verlag Lecture Notes in Mathematics 560, Berlin.

Dempster, M. ed. (1980): *Stochastic Programming*. Academic Press, London.

Dolecki, S., G. Salinetti and R. Wets (1983): Convergence of functions: equisemicontinuity. *Trans. Amer. Math. Soc.* **276**, 409–429.

Dupačová, J. (1983a): Stability in stochastic programming with recourse. *Acta Univ. Carol.-Math et Phys.* **24**, 23–34.

Dupačová, J. (1983b): The problem of stability in stochastic programming (in Czech). Dissertation for the Doctor of Sciences degree, Faculty of Mathematics and Phyics, Charles University, Prague.

Dupačová, J. (1984a): Stability in stochastic programming with recourse – Estimated parameters. *Math. Progr.* **28**, 72–83.

Dupačová, J. (1984b): On asymptotic normality of inequality constrained optimal decisions. In: *Proc. 3-rd Prague conference on asymptotic statistics*, eds. P. Mandl and M. Hušková. Elsevier, Amsterdam, p. 249–257.

Ermoliev, Yu., A. Gaivoronski and C. Nedeva (1985): Stochastic optimization problems with incomplete information on distribution function. *SIAM J. Control and Optimization* **23**, 696–716.

Fletcher, R. and M.J.D. Powell (1963): A rapidly convergent descent method for minimization. *The Computer Journal* **6**, 163–168.

Herback, L.H. (1959): Properties of model II – type analysis of variance tests, A: Optimum nature of the F–test for model II in the balanced case. *Ann. Math. Stat.* **30**, 939–959.

Hiniart-Urruty, J-B. (1976): About properties of the mean functional and of the continuous infimal convolution in stochastic convex analysis. In: *Optimization Techniques II.* Springer Lecture Notes in Computer Sciences 41, Berlin, p. 763–789.

Huber, P. (1967): The behavior of maximum likelihood estimates under nonstandard conditions, *Proc. Fifth Berkeley Symp. Math. Stat. Prob.* **1**, 221–233.

Huber, P.J. (1973): Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Stat.* **1**, 799–821.

Huber, P.J. (1981): *Robust statistics*. Wiley, New York.

Ibragimov, I.A. and R.Z. Haśminskii (1981): *Statistical estimation. Asymptotic theory*. Springer, New York.

Ioffe, A. (1978): Survey of measurable selection theorems: Russian literature supplement, *SIAM J. on Control and Optimization* **16**, 729–731.

Jöreskog, K.G. (1969): A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34**, 183–202.

Judge, G.G. and T. Takayama (1966): Inequality restrictions in regression analysis. *J. ASA* **61**, 166–181.

Kall, P. (1976): *Stochastic Linear Programming*. Springer, Berlin.

Lee, S.-Y. (1980): Estimation of covariance structure models with parameters subject to functional restraints. *Psychometrika* **48**, 309–324.

Liew, Chong Kiew (1976): Inequality constrained least squares estimation. *J. ASA* **71**, 746–751.

Prékopa, A. (1973): Contributions to the theory of stochastic programming. *Mathematical Programming* **4**, 202–221.

Rao, C.R. (1965): *Linear statistical inference and its applications*. Wiley, New York.

Rao, C.R. (1971): Estimation of variance and covariance components MINQUE theory. *J. Multivar. Anal.* **1**, 257–275.

Robinson, S. (1985): Local epi-continuity and local optimization, Tech. Report University of Wisconsin.

Rockafellar, R.T. (1976): Integral functionals, normal integrands and measurable multifunctions, in *Nonlinear Operators and the Calculus of Variations*, eds. J. Gossez and L. Waelbroeck, Springer Verlag Lecture Notes in Mathematics 543, Berlin.

Rockafellar, T.T. and R. Wets (1984): Variational systems, an introduction, in *Multifunctions and Integrands*, ed. G. Salinetti, Springer Verlag Lecture Notes in Mathematics 1091, Berlin. 1–54.

Salinetti, G. and R. Wets (1981): On the convergence of closed valued measurable multifunctions. *Trans. Amer. Math. Soc.* **266**, 275–289.

Salinetti, G. and R. Wets (1986): Convergence of infima, especially stochastic infima, Tech. Report Univ. Roma "La Sapienza".

Sen, P.K. (1979): Asymptotic properties of maximum likelihood estimators based on conditional specification. *Ann. Statist.* **7**, 1019–1033.

Silvey, S.D. (1959): The Lagrangian multiplier test. *Ann. Math. Stat.* **30**, 389–407.

Solis, R. and R. Wets (1981): A statistical view of stochastic programming. Tech. report, Univ. Kentucky.

Thompson, W.A. Jr. (1962): The problem of negative estimates of variance components. *Ann. Math. Stat.* **33**, 273–289.

Wagner, D. (1977): Survey of measurable selection theorems, *SIAM J. Control and Optimization* **15**, 859–903.

Wald, A. (1949): Note on the consistency of the maximum likelihood estimate, *Ann. Math. Stat.* **20**, 595–601.

Wets, R. (1973): On inf-compact mathematical programs. *Fifth Conference on Optimization Techniques*, Part I: Springer Verlag Lecture Notes in Computer Sciences 3, p. 426–436.

Wets, R. (1979): A statistical approach to the solution of stochastic programs with (convex) simple recourse. Working Paper, Univ. of Kentucky.

Wets, R. (1983): Stochastic programming: solution techniques and approximation schemes. In: Bachem, A., M. Grötschel and B. Korte (eds). *Mathematical Programming: The State of the Art*. Springer, Berlin, p. 566–603.