The Annals of Statistics
1988, Vol. 16, No. 4, 1517–1549

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

BY JITKA DUPAČOVÁ AND ROGER WETS[1]

Charles University and IIASA, and University of California, Davis and IIASA

We study the asymptotic behavior of the statistical estimators that maximize a not necessarily differentiable criterion function, possibly subject to side constraints (equalities and inequalities). The consistency results generalize those of Wald and Huber. Conditions are also given under which one is still able to obtain asymptotic normality. The analysis brings to the fore the relationship between the problem of finding statistical estimators and that of finding the optimal solutions of stochastic optimization problems with partial information. The last section is devoted to the properties of the saddle points of the associated Lagrangians.

1. Introduction. Deriving 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 pheonomenon 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, which can still be used to handle least-squares 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 nonsmooth analysis. The problem of proving consistency of the estimators and the study of their asymptotic behavior is closely related to that of obtaining confidence

Received May 1987; revised March 1988.
[1]Supported in part by a National Science Foundation grant.
AMS 1980 *subject classifications*. 62F12, 62A10, 90C15.
Key words and phrases. Statistical estimators, consistency, stochastic programming, epi-convergence, asymptotically normal, subdifferentiability.

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 problems 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:

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

where f: $R^n \times \Xi \to R \cup \{+\infty\}$ is an extended real-valued function and ξ 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 ξ, 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 $\{P^\nu, \nu = 1, \dots\}$ 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 preliminary results, and 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; 1984b) for a special case and Dupačová and Wets (1986).

Section 4 is devoted to asymptotic analysis. Since we are confronted with a very general class of problems, one should not expect to obtain, in general, the standard limiting results. Our purpose here has been to relax the conditions under which one can prove asymptotic normality for the optimal estimators (or solutions). To do so, we rely on subdifferential calculus, which allows us to derive asymptotic normality for a class of criterion functions that lack the usual differentiability properties. We also give the conditions under which the presence of constraints will not prevent asymptotic normality. We extend the earlier results of Huber (1967) in a number of directions: (i) we allow for constraints, (ii) the probability measures converging to P are not necessarily the empirical measures and (iii) there are no differentiability assumptions on the likelihood (criterion) function. Finally, in Section 5, we sketch out another approach to these questions, by relying on the Lagrangian associated to constrained optimization problems. Consistency is now derived for the optimal estimators as well as for the associated Lagrangian multipliers. The technique is very similar to that used in Section 3, except that now we rely on the epi/hypo-convergence of the Lagrangian functions. Again it is shown that under certain conditions, that in

some sense are weaker than those of Section 4 (there are fewer constraints), one can obtain asymptotic normality, for the vector function of estimators and associated Lagrangian multipliers.

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-squares 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) and Liew (1976) the theory of quadratic programming is used to exhibit and discuss the statistical properties of least-squares 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 Aitchison 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 multidimensional 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). A recent paper by Shapiro (1988) has brought to our attention a number of additional examples and many more references.

EXAMPLE 2.1 (Inequality constrained least-squares estimation of regression coefficients). Assume that the dependent variable y can be explained or predicted on the basis 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

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

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:

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

which can be solved by quadratic programming techniques.

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

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

with the P^{ν} the empirical distributions.

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

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

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 function is

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

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)]:

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

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

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

This also fits the general framework; the objective function is

$$\begin{aligned} f(x, \xi) &= \rho\left(\xi_0 - \sum_{i=1}^{p} \xi_i x_i \right) \quad \text{if } x \in S, \\ &= +\infty \qquad\qquad\qquad\quad \text{otherwise,} \end{aligned} \tag{2.6}$$

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 $\Lambda(n, k)$ is the matrix of factor loadings. It is assumed that f and e are normally distributed with mean 0, $\operatorname{var} f = \Phi$ and $\operatorname{var} e = \Psi$, which is diagonal. Consequently, x is normally distributed with mean 0 and with the variance matrix

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

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

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

(the maximum likelihood method), or

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

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 nontrivial unconstrained optimization problems, which can be solved by different methods such as the method of Davidon, Fletcher and 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 $\Psi_{ii} \geq 0$, $i = 1, \ldots, n$. Thus, the appropriate formulation defines f as

$$\begin{aligned} f(\Sigma, S) &= f_1(\Sigma, S) \quad \text{if } \Psi_{ii} \geq 0,\ i = 1, \ldots, n, \\ &= +\infty \qquad\quad \text{otherwise,} \end{aligned}$$

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 matrices of observed values of the independent variables and β_i, $i = 1, \ldots, p$, are mutually uncorrelated random vectors with $E\beta_i = 0$, $\operatorname{var}\beta_i = \sigma_i^2 I_{r_i}$, $i = 1, \ldots, p$, and $E\varepsilon = 0$, $\operatorname{var}\varepsilon = \sigma_0^2 I_\nu$, 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 jth measurement (observation) in the ith population is given by

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

In (2.11), μ is the fixed effect, a_i, $i = 1, \ldots, k$, is the random effect of the ith population and e_{ij} is the 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 $\mu, \sigma_a^2, \sigma_e^2$ 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

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

and

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

where $\bar{y}_{i\cdot} = (1/n)\Sigma_{j=1}^{n} y_{ij}$, $i = 1, \ldots, k$, and $\bar{y}_{\cdot\cdot} = (1/nk)\Sigma_{i=1}^{k}\Sigma_{j=1}^{n} y_{ij}$, 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 [see Rao (1971)] developed for the general model (2.10): Unbiased estimates of a linear parameteric function $\Sigma_{i=0}^{p}\sigma_i^2 q_i$ are sought in the form $y^T A y$, where

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

and which are optimal in some sense. The MIVQUE estimates correspond to a matrix A that minimizes the variance of $y^{\mathrm{T}}Ay$ subject to the conditions (2.14) and the MINQUE estimates correspond to a matrix A that minimizes $\mathrm{tr}(A(I + \Sigma_{i=1}^{p} X_i X_i^{\mathrm{T}}))^2$ subject to conditions (2.14). In none of the mentioned approaches, however, the natural nonnegativity constraints on the estimates of the variances σ_i^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 underlying 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, for the analysis of variance model (2.11)

$$f(\sigma_a^2, \sigma_e^2, \mu, \xi) = (2\pi)^{-n/2}(\sigma_e^2 + n\sigma_a^2)^{-1/2}(\sigma_e^2)^{-(n-1)/2}$$
$$\times \exp\left\{-\frac{1}{2\sigma_e^2}\left[\sum_{j=1}^{n}(\xi_j - \mu)^2 - \frac{\sigma_a^2}{\sigma_e^2 + n\sigma_a^2}\left(\sum_{j=1}^{n}\xi_j - n\mu\right)^2\right]\right\}$$
$$\text{if } \sigma_a^2 \geq 0, \ \sigma_e^2 \geq 0,$$
$$= -\infty \quad \text{otherwise},$$

where ξ is the n-vector of random outcomes of the measurements in a fixed population.

Similarly, nonnegative MINQUE and MIVQUE estimates are of interest.

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

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

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 $\theta_0 \in \Theta$ and asymptotic normality of $\sqrt{\nu}(T^\nu - \theta_0)$ 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 $(\Xi, \mathcal{A}, P)$, a random element ξ on Ξ, a measurable function $f\colon R^n \times \Xi \to R$ and a set $S \subset R^n$,

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

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

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

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

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 parameteric family of probability measures but with an unknown parameter vector was studied, e.g., in Dupačová (1984a).

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 $(\Xi, \mathscr{A}, P)$ be a probability space, with Ξ—the support of P—a closed subset of a Polish space X, and $\mathscr{A}$ the Borel σ-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 $(\Omega, \mathscr{A}', P')$. 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

$$f\colon R^n \times \Xi \to R \cup \{\infty\} = (-\infty, \infty]$$

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

$$\operatorname{dom} Ef := \{x | Ef(x) < \infty\}$$

as the *effective domain* of Ef. Points that do not belong to $\operatorname{dom} Ef$ 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

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

An extended real-valued function $h\colon R^n \to \overline{R} = [-\infty, \infty]$ is said to be *proper* if $h > -\infty$ and not identically $+\infty$; it is *lower semicontinuous* (l.s.c.) 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 extended real-valued function f defined on $R^n \times \Xi$ is a *random lower semicontinuous function* if

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

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

where $\mathscr{B}^n$ is the Borel σ-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: $R^n \to \overline{R}$ is the set

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

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

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

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

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

$$\Gamma^{-1}(F) := \{\xi \in \Xi | \Gamma(\xi) \cap F \neq \varnothing\} \in \mathscr{A}.$$

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* Γ: $\Xi \rightrightarrows R^n$ *is a closed-valued measurable multifunction, then there exists at least one measurable selector, i.e., a measurable function* x: $\operatorname{dom} \Gamma \to R^n$ *such that for all* $\xi \in \operatorname{dom} \Gamma$, $x(\xi) \in \Gamma(\xi)$, *where* $\operatorname{dom} \Gamma := \{\xi \in \Xi | \Gamma(\xi) \neq \varnothing\} = \Gamma^{-1}(R^n) \in \mathscr{A}$.

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

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

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

Moreover, the infimal function

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

is $\mathscr{A}$*-measurable, and the set of optimal solution*

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

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

$$\xi \mapsto x^*(\xi): \operatorname{dom}(\arg\min f(\cdot, \xi)) \rightrightarrows R^n$$

such that $x^*(\xi)$ *minimizes* $f(\cdot, \xi)$ *whenever* $\arg\min f(\cdot, \xi) \neq \varnothing$.

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 ξ or of a function of ξ—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 $(Z, \mathscr{F}, \mu)$ be a sample space with $(\mathscr{F}^\nu)_{\nu=1}^\infty$ an increasing sequence of σ-field contained in $\mathscr{F}$. A sample ζ—e.g., $\zeta = \{\xi^1, \xi^2, \dots\}$ obtained by independent sampling of the values of ξ—leads us to a sequence $\{P^\nu(\cdot, \zeta), \nu = 1, \dots\}$ of probability measures defined on $(\Xi, \mathscr{A})$. 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 $A \in \mathscr{A}$,

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

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 (1949) and the extensions by Huber (1967) to the more general setting laid out previously. Consistency is obtained by relying on assumptions that are weaker than those of Huber (1967) even in the unconstained 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.s.c. functions.

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

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

and

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

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 to 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) and Salinetti and Wets (1986)]. *Suppose* $\{g; g^\nu\colon R^n \to \overline{R},\ \nu = 1,\dots\}$ *is a collection of functions such that* $g = \text{epi-lim}_{\nu\to\infty} g^\nu$. *Then*

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

and, if

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

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

$$x \in \arg\min g$$

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* $\{\nu_k,\ k = 1,\dots\}$,

$$\arg\min g^{\nu_k} \cap D \neq \varnothing,$$

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

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

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

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

$$x^\nu \in \varepsilon_\nu - \arg\min g^\nu := \{x \mid g^\nu(x) \le \varepsilon_\nu + \inf g^\nu\}.$$

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

ASSUMPTION 3.4 ("Continuities" of f). The function

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

with

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

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

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

and for all $\xi \in \Xi$,

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

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 β: $\Xi \to R$ such that for all $x' \in V \cap S$ and $\xi \in \Xi$,

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

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

$$P^\nu\colon \mathscr{A} \times Z \to [0,1], \qquad \nu = 1, \ldots,$$

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

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

and for all $A \in \mathscr{A}$,

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

For μ-almost all ζ in Z, the sequence

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

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

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

The assumption that

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

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

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

if Ξ is the support of the measure P and for all $x \in \bigcap_{\xi \in \Xi} \operatorname{dom} f(\cdot, \xi)$, the function $f(x, \cdot)$ is bounded above by a summable function. Then, with $S = \bigcap_{\xi \in \Xi} \operatorname{dom} f(\cdot, \xi)$ 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 $\xi \mapsto \operatorname{dom} f^+(\cdot, \xi) = S$ 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 (3.3i) and $(x, \xi) \mapsto f(x, \xi)$ is $\mathscr{B}^n \otimes \mathscr{A}$-measurable (3.3ii) since for all $\alpha \in R$,

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

To see this, suppose $\{(x^k, \xi^k)\}_{k=1}^\infty \subset \operatorname{lev}_\alpha f$ 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 $k \to \infty$, we obtain

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

i.e., $(x, \xi) \in \operatorname{lev}_\alpha f$. Since f is a random l.s.c. function it 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)], assume 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 the notation, 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 $(Z, \mathscr{F}, \mu)$. 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* $Z_0 \in \mathscr{F}$, $\mu(Z_0) = 1$ *such that for all* $\zeta \in Z_0$, Ef *and* $\{E^\nu f, \nu = 1, \dots\}$ *are proper lower semicontinuous functions such that*

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

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.,

$$S \supset \operatorname{dom} Ef, \qquad S \supset \operatorname{dom} E^\nu f.$$

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

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

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,

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

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 preceding holds for every ν, μ-a.s. on Z, the set

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

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-*a.s.*

$$Ef = \operatorname*{epi\text{-}lim}_{\nu\to\infty} E^\nu f = \operatorname*{ptwse\text{-}lim}_{\nu\to\infty} E^\nu f,$$

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

PROOF. The argument essentially follows that of Birge and Wets (1986), Theorem 2.8, 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-a.s. 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,

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

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

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

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

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

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

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

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.s.c. functions.

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

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

are μ*-a.s. random lower semicontinuous functions, such that* $\zeta \to \operatorname{epi} E^\nu f(\cdot,\zeta)$ *is* $\mathscr{F}^\nu$*-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

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

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

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

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

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

Moreover, since $\zeta \mapsto P^\nu(A,\zeta)$ is $\mathscr{F}^\nu$-measurable for all $A \in \mathscr{A}$, 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

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

which means that the function

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

is $\mathscr{F}^\nu$-measurable. To see this, observe that the "convergence in distribution" topology can be obtained from the base of open sets

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

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

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

is the composition of a continuous function, and a $\mathscr{F}^\nu$-measurable function, and hence is $\mathscr{F}^\nu$-measurable. □

In the proof of Theorem 3.8, we have used the continuity of the map $P^\nu \mapsto \operatorname{epi} E^\nu f$; 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.s.c. 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 $P^\nu \mapsto \operatorname{epi} E^\nu f$.

THEOREM 3.9 (Consistency). *Under Assumptions* 3.4 *and* 3.5 *we have that* μ-*a.s.*

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

Moreover, there exists $Z_0 \in \mathscr{F}$ *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* $x^\nu \in \arg\min E^\nu f^\nu(\cdot, \zeta)$ *belongs to* $\arg\min Ef$ (*i.e., is an optimal estimate*);

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

$$\zeta \mapsto \arg\min E^\nu f(\cdot, \zeta)\colon Z_0 \rightrightarrows R^n$$

is a closed-valued $\mathscr{F}^\nu$-*measurable multifunction.*

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

$$(\arg\min E^\nu f) \cap D \text{ is nonempty } \mu\text{-a.s.},$$

and

$$\{x^*\} = \arg\min Ef \cap D,$$

then there exist $\{x^\nu\colon Z_0 \to R^n\}_{\nu=1}^\infty$ $\mathscr{F}^\nu$-*measurable selections of* $\{\arg\min E^\nu f\}_{\nu=1}^\infty$ *such that*

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

and also

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

PROOF. The inequality (3.14) immediately follows from (3.9) and the epi-convergence μ-a.s. 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 ($\arg\min E^\nu f$) is a closed-valued $\mathcal{F}^\nu$-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 $\zeta \in Z_0$, $Ef = \text{epi-lim}_{\nu\to\infty} E^\nu f$, and for all $\nu = 1,\ldots,$ ($\arg\min E^\nu f$) $\cap D$ is nonempty. For all ν, the multifunction

$$\zeta \mapsto (\arg\min E^\nu f(\cdot,\zeta) \cap D)\colon Z_0 \rightrightarrows R^n$$

is nonempty compact-valued and $\mathcal{F}^\nu$-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,

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

Then any cluster point of the sequence is in D, since it is compact, and in $\arg\min Ef$ 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 $k = 1,\ldots,$

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

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 $x^\nu\colon Z \to R^n$ be an $\mathcal{F}^\nu$-measurable selection of the $\mathcal{F}^\nu$-measurable multifunction $\zeta \mapsto (\arg\min E^\nu f(\cdot,\zeta) \cap D)$, 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 (1949) and 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) and Robinson (1987)] 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 $\arg\min Ef$.

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

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

If h is proper ($h > -\infty$, 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 the results of Wets (1973) and Hiriart-Urruty (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* $A \in \mathscr{A}$ *with* $P(A) > 0$ [*resp.* $P_\nu(A) > 0$] *such that for all* $\alpha \in R$, *the set*

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

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

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

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

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

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

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

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

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

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

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

4. Asymptotics, convergence rates. In Section 3 we exhibited sufficient conditions for the convergence with probability 1 of the estimators $\{x^\nu\colon Z \to R^n,\ \nu = 1, \dots\}$ to x^*, the optimal solution of the limit problem. Here we go one step further and analyze the rate of convergence in probabilistic terms. The argumentation is related to that of Huber (1967), adapted to fit the more general class of problems under consideration; this was already the pattern

followed by Solis and Wets (1981) in the unconstrained case and by Dupačová (1983a, 1983b, 1984b) for stochastic programs with recourse under special assumptions. As already indicated in Section 1, we extend the results of Huber (1967) in a number of directions: (i) we allow for constraints, (ii) the probability measures converging to P are not necessarily the empirical measures and (iii) there are no differentiability assumptions on the likelihood (criterion) function [in terms of Huber's set-up, this would correspond to the case when his function Ψ is not uniquely determined; see Section 3 of Huber (1967)].

One way to look at the results of this section is to view them as providing limiting conditions under which one may be able to obtain asymptotic normality. (Note that when there are constraints, one should usually not expect the asymptotic distribution to be Gaussian.) This, in turn, allows us to obtain certain probabilistic estimates for the convergence "rates." To approximate the distribution of x^ν, to obtain confidence intervals for example, we need an assertion that a suitably normalized sequence converges in distribution to a *nondegenerate* random vector. The normalizing coefficients need not be unique but they suggest a rate of convergence. Following Lehmann (1983), we shall say that the sequence $x^\nu - x^*$ *goes to* 0 *with the rate of convergence* $1/k_\nu$ if $k_\nu \to \infty$ as $\nu \to \infty$ and if there is a continuous distribution function H such that

$$P\{k_\nu \|x^\nu - x^*\| \le a\} \to H(a) \quad \text{as } \nu \to \infty.$$

We begin by a quick review of the main definitions and results that provides us with a good notion for the subgradients of not necessarily differentiable functions. Any assumption of differentiability of $f(\cdot, \xi)$, would be inappropriate and would for one reason or another eliminate from the domain of applicability all the examples mentioned in Section 2. To handle the lack of differentiability, we rely on the theory of subdifferentiability developed to handle nonsmooth functions [see Clarke (1983), Rockafellar (1983) and Aubin and Ekeland (1984)].

The *contingent derivative* of a lower semicontinuous function h: $R^n \to (-\infty, +\infty]$ at x, a point at which h is finite, with respect to the direction y is

$$h'(x; y) := \operatorname*{epi\text{-}lim\,inf}_{t \downarrow 0} \frac{h(x + ty) - h(x)}{t},$$

using the convention $\infty - \infty = \infty$. It is not difficult to see that h' is always well defined with values in the extended reals. If $x \notin \operatorname{dom} h$, then $h'(x; \cdot) = \infty$, otherwise

$$h'(x; y) = \liminf_{\substack{y' \to y \\ t \downarrow 0}} \frac{h(x + ty') - h(x)}{t}.$$

The (*upper*) *epi derivative of* h *at* x, where h is finite, *in direction* y, is the epi-limit superior of the collection $\{h'(x'; \cdot),\ x' \in R^n\}$ at x, i.e.,

$$h^\uparrow(x; \cdot) := \operatorname*{epi\text{-}lim\,sup}_{x' \to x} h'(x'; \cdot),$$

$$h^\uparrow(x; y) = \inf_{\substack{\{x' \to x\} \\ \{y' \to y\}}} \limsup h'(x'; y'),$$

where by writing $\{x' \to x\}$ and $\{y' \to y\}$ we mean that the infimum must be taken with respect to all nets—or equivalently here sequences—converging to x and y [see Aubin and Ekeland (1984), Chapter 7, Section 3].

It is remarkable that if h is proper and $x \in \operatorname{dom} h$, the function $y \mapsto h^{\uparrow}(x; \cdot)$ is sublinear and l.s.c. [Rockafellar (1980), Theorems 1 and 2]. Moreover, if h is Lipschitzian around x, then $h^{\uparrow}(x; \cdot)$ is everywhere finite (and hence continuous); in particular, if h is continuously differentiable at x, then $h^{\uparrow}(x; y)$ is the directional derivative of h in direction y, and if h is convex in a neighborhood of x, then

$$h^{\uparrow}(x; y) = \lim_{t \downarrow 0} \frac{h(x + ty) - h(x)}{t}$$

is the one-sided directional derivative in direction y. The sublinearity and lower semicontinuity of $h^{\uparrow}(x; \cdot)$ make it possible to define the notion of a subgradient of h at x, by exploiting the fact that there is a one-to-one correspondence between the proper lower semicontinuous, sublinear functions g and the nonempty closed convex subsets C of R^n, given by

$$g(y) = \sup_{v \in C} v \cdot y \quad \text{for all } y \in R^n,$$

and

$$C = \{v \in R^n | v \cdot y \leq g(y) \text{ for all } y \in R^n\}$$

[see Rockafellar (1970)]. Assuming that $h^{\uparrow}(x; \cdot)$ is proper, let $\partial h(x)$ be the nonempty closed convex set such that for all y,

$$h^{\uparrow}(x; y) = \sup_{v \in \partial h(x)} v \cdot y.$$

Every vector v in $\partial h(x)$ is a *subgradient* of h at x. If h is smooth (continuously differentiable), then

$$\partial h(x) = \{\nabla h(x), \text{ the gradient of } h \text{ at } x\};$$

if h is convex, then

$$\partial h(x) = \{v | h(x + y) \geq h(x) + v \cdot y \text{ for all } y \in R^n\}$$

is the usual definition of the subgradients of a convex function. More generally, if h is locally Lipschitz at x, then

$$\partial h(x) = \operatorname{co}\left\{v = \lim_{x' \to x} \nabla h(x') | h \text{ is smooth at } x'\right\}.$$

For the proofs of these preceding assertions and further details, consult Rockafellar (1981) and Aubin and Ekeland (1984).

Before we return to the problem at hand, we state the results about the additivity of subgradients that are relevant to our analysis, we begin with a general result that shows that the derivatives and subgradient functions of the random l.s.c. function f and the expectation functionals $E^{\nu} f$ and Ef have the appropriate measurability properties.

THEOREM 4.1. *Suppose* h: $R^n \times \Xi \to \overline{R}$ *is a random lower semicontinuous function. Then so are its contingent derivative and its (upper) epi-derivative. Moreover, for all* $x \in R^n$, $\xi \mapsto \partial h(x, \xi)$ *is a random closed convex set.*

PROOF. Theorem 3.1 of Salinetti and Wets (1981) tells us that the lim sup and lim inf of sequences of random closed sets (closed-valued measurable multifunctions) are random closed sets. Since the epigraphs of the epi-lim sup and epi-lim inf are, respectively, the lim inf and lim sup of the corresponding sequence of epigraphs [see, for example, Dolecki, Salinetti and Wets (1983), Section 2], the assertion about the derivatives follows from their definitions and property (3.4) of random lower semicontinuous functions. Since $h^{\uparrow}(x; \cdot, \xi)$ is sublinear, it follows that its conjugate—another random l.s.c. function [Rockafellar (1976)]—is the indicator of the random closed convex set $\xi \mapsto \partial f(x, \xi)$. □

Our interest in subdifferential theory is conditioned by the fact that for a very large class of functions (with values in the extended reals), we can characterize optimality in terms of a differential inclusion, a point x^0 that minimizes the proper l.s.c. function h on R^n, must necessarily satisfy

$$0 \in \partial h(x^0);$$

if h is convex this is also a sufficient condition. There is a subdifferential calculus, but for our purposes the following results about the subdifferentials on sums of l.s.c. functions is all we need. We say that a l.s.c. function is *subdifferentially regular* at x if $h'(x; \cdot) = h^{\uparrow}(x; \cdot)$. If h is convex or subsmooth on a neighborhood of x, thus in particular if h is $\mathscr{C}^1$ at x, it is subdifferentially regular at x; h is *subsmooth* on a neighborhood V of x, if for all $y \in V$,

$$h(y) = \max_{t \in T} \phi_t(y),$$

where T is a compact topological space, each ϕ_t is of class $\mathscr{C}^1$, and both $\phi_t(x)$ and $\nabla_x \phi_t(x)$ are continuous with respect to (t, x). If h is subsmooth on an open set U, it is also locally Lipschitz on U [Clarke (1975)].

LEMMA 4.2 [Rockafellar (1979)]. *Suppose* h_1 *and* h_2 *are l.s.c. functions on* R^n *and* x *a point at which both* h_1 *and* h_2 *are finite. Suppose that* $\operatorname{dom} h_1(x; \cdot)$ *is nonempty and* h_2 *is locally Lipschitz at* x. *Then*

$$\partial(h_1 + h_2)(x) \subset \partial h_1(x) + \partial h_2(x).$$

Moreover, equality holds if h_1 *and* h_2 *are subdifferentially regular at* x.

LEMMA 4.3 [Clarke (1983)]. *Let* U *be an open subset of* R^n, *and suppose* h: $U \times \Xi \to R$ *is measurable with respect to* ξ *and there exist a summable function* β *such that for all* x^0, x^1 *in* U *and* $\xi \in \Xi$,

$$|h(x^0, \xi) - h(x^1, \xi)| \leq \beta(\xi)\|x^0 - x^1\|.$$

Suppose, moreover, that for some $\bar{x} \in U$, $Eh(\bar{x})$ *is finite. Then* Eh *is finite and Lipschitz on* U, *and for all* x *in* U,

$$\partial Eh(x) \subset E\{\partial h(x, \xi)\} = \int \partial h(x, \xi) P(d\xi).$$

Moreover, equality holds whenever $h(\cdot, \xi)$ is a.s. subdifferentially regular at x, in which case also Eh is subdifferentially regular at x.

Theorem 4.1 shows that $\xi \mapsto \partial h(x, \xi)$ is a random (nonempty) closed set; it is easy to verify that under the assumptions of Lemma 4.3, h is a random l.s.c. function on $U \times \Xi$. In fact, for all ξ, $\partial h(x, \xi)$ is a compact subset of R^n [see Clarke (1983), Proposition 2.1.2]. The integral of a random closed set Γ defined on Ξ (with values in the closed subsets of R^n) is

$$\int \Gamma(\xi) P(d\xi) := \left\{ x = \int s(\xi) P(d\xi) \mid s(\xi) \in \Gamma(\xi) \text{ a.s., } s \in L^1 \right\}$$

[see Aumann (1965)]. If P is absolutely continuous and Γ is integrably bounded [the function $\xi \mapsto \sup\{\|x\| \mid x \| \in \Gamma(\xi)\}$ is summable], then $\int \Gamma(\xi) P(d\xi) = \int \operatorname{co} \Gamma(\xi) P(d\xi)$ is convex, where $\operatorname{co} \Gamma(\xi)$ is the convex hull of ξ. If Γ is uniformly bounded, then $\int \Gamma(\xi) P(d\xi)$ is a compact subset of R^n.

We shall be working with the same set-up as in Section 3, but with a somewhat more restricted class of random l.s.c. functions. Instead of Assumption 3.4, we shall be using the following one.

ASSUMPTION 4.4. The function $f\colon R^n \times \Xi \to (-\infty, \infty]$ is of the following type:

$$f(x, \xi) = f_0(x, \xi) + \Psi_S(x),$$

where Ψ_S is the indicator function of the closed nonempty set $S \subset R^n$, i.e.,

$$\Psi_S(x) = 0 \quad \text{if } x \in S,$$
$$= \infty \quad \text{otherwise,}$$

and f_0 is a finite-valued function on $R^n \times \Xi$, with

$$\xi \mapsto f_0(x, \xi) \text{ relatively continuous on } \Xi,$$

for all $x \in S$, and on any open set U that contains S, the function

$$x \to f_0(x, \xi) \text{ is locally Lipschitz}$$

for all $\xi \in \Xi$, and such that to any bounded open set V there corresponds a function β uniformly integrable with respect to P^ν, $\nu = 1, 2, \ldots$, such that for any pair x^0, x^1 in V,

$$|f_0(x^0, \xi) - f_0(x^1, \xi)| \leq \beta(\xi) \|x^0 - x^1\|. \tag{4.1}$$

The only condition of Assumption 3.4 that does not appear explicitly in Assumption 4.4, either in exactly the same form or in a stronger form, is the lower semicontinuity of $f(\cdot, \xi)$ on R^n for all ξ in Ξ. But that is an immediate consequence of the fact that $f_0(\cdot, \xi)$ is locally Lipschitz and S is closed. Thus, f is a proper random lower semicontinuous function, and so is also f_0. Moreover, all the results and the observations of Section 3 are immediately applicable to both f and f_0, as well as to the corresponding expectation functionals. Of course,

these functions will now have Lipschitz properties that we shall exploit in our analysis. In the convex case it might be possible to work with weaker restrictions on the function f by relying on finer results about the additivity of subgradients [see Rockafellar and Wets (1982)]. Combining the results of Section 3 with those about subgradients of random l.s.c. functions, in particular Lemma 4.3, we can show that:

LEMMA 4.5. *Under Assumptions* 4.4 *and* 3.5, *we have that* μ-*a.s.* Ef *and* $\{E^\nu f,\ \nu = 1,\dots\}$ *are proper lower semicontinuous functions that are locally Lipschitz on* S. *Moreover, we always have*

$$\partial Ef_0(x) \subset E\{\partial f_0(x,\xi)\} = \int_\Xi \partial f_0(x,\xi)P(d\xi)$$

and for $\nu = 1,\dots,$

$$\partial E^\nu f_0(x,\zeta) \subset \int_\Xi \partial f_0(x,\xi)P^\nu(d\xi,\zeta), \quad \mu\text{-}a.s.,$$

with equality if for all ξ, $f_0(\cdot,\xi)$ *is subdifferentially regular at* x. *Moreover, if* $x \in S$,

$$\partial Ef(x) \subset \partial Ef_0(x) + \partial\Psi_S(x)$$

and for $\nu = 1,\dots,$

$$\partial E^\nu f(x,\zeta) \subset \partial E^\nu f_0(x,\zeta) + \partial\Psi_S(x), \quad \mu\text{-}a.s.,$$

with equality if Ψ_S *and for all* ξ, $f_0(\cdot,\xi)$ *are subdifferentially regular at* x.

REMARK 4.6. If $x \in S$, $\partial\Psi_S(x)$ is the polar of the tangent cone $T_S(x)$ to S at x [Clarke (1975)]. If S is a differentiable manifold, then $\partial\Psi_S(x)$ is the orthogonal complement of the tangent space at x and, of course, Ψ_S is differentially regular at x. This is also the case when S is locally convex at x, or if x belongs to the boundary of S and this boundary is locally a differentiable manifold. More generally, Ψ_S is subdifferentially regular at x, if the tangent cone to S at x, has the representation

$$T_S(x) = \{y \mid \exists \lambda_k \downarrow 0,\ y^k \to y \text{ with } x + \lambda_k y^k \in S\}.$$

So far, we have limited our assumptions to certain continuity properties of the function f with respect to x and ξ. In order to derive the asymptotic behavior we need to impose some additional conditions about the way the information collected from the samples is included in the approximating probability measures P^ν, in particular on how it affects the subgradients of the functions $E^\nu f$. Let us introduce the following notation: $u_0(x,\xi)$ will always denote an element of $\partial f_0(x,\xi)$ and $v_S(x)$ an element of $\partial\Psi_S(x)$. In view of Theorem 4.1 and Lemma 4.5 if $x \in S$, we always have that $v(x) \in \partial Ef(x)$ implies the existence of $v_S(x) \in \partial\Psi_S(x)$ and $u_0(x,\cdot)$ measurable with $u_0(x,\xi) \in \partial f_0(x,\xi)$, P-a.s. such that

$$v(x) = v_0(x) + v_S(x) = E\{u_0(x,\xi)\} + v_S(x).$$

Moreover, similar formulas hold μ-a.s. if the integration is with respect to $P^\nu(\cdot, \zeta)$ instead of P. If the functions $f_0(\cdot, \xi)$, as well as Ψ_S, are a.s. subdifferentially regular, then a type of converse statement also holds. We have that

$$0 \in \partial Ef(x^*)$$

implies the existence of $v_S(x^*) \in \partial \Psi_S(x^*)$ and of a random function $u_0(x^*, \cdot)$ from Ξ to R^n with $u_0(x^*, \cdot) \in \partial f_0(x^*, \xi)$, P-a.s. such that

$$0 = E\{u_0(x^*, \xi)\} + v_S(x^*). \tag{4.2}$$

Similarly,

$$0 \in \partial E^\nu f(x^\nu)$$

means that there exist $v_S(x^\nu) \in \partial \Psi_S(x^\nu)$ and a random function $u_0(x^\nu, \cdot)$ from Ξ to R^n with $u_0(x^\nu, \cdot) \in \partial f_0(x^\nu, \cdot)$ P^ν-a.s. such that

$$\begin{aligned} 0 &= v_0^\nu(x^\nu) + v_S(x^\nu) \\ &= E^\nu\{u_0(x^\nu, \xi)\} + v_S(x^\nu). \end{aligned} \tag{4.3}$$

ASSUMPTION 4.7 (Statistical information). The probability measures $\{P^\nu, \nu = 1, \dots\}$ are such that for some $v^\nu \in \partial E^\nu f(x^*, \zeta)$ and $v \in \partial Ef(x^\nu(\zeta))$:

(i) $\sqrt{\nu}[v^\nu(x^*, \zeta) + v(x^\nu(\zeta))]$ converges to 0 in probability;

(ii) $\sqrt{\nu}[v_S(x^\nu(\zeta)) - v_S(x^*)]$ converges to 0 in probability;

(iii) $\sqrt{\nu}\, v^\nu(x^*, \zeta)$ is asymptotically Gaussian with distribution function $N(0, \Sigma_1)$, where Σ_1 is the covariance matrix.

Moreover,

(iv) Ef_0 is twice continuously differentiable at x^* with nonsingular Hessian H.

Before we prove the main result of this section, let us examine some of the implications of these assumptions. The assumption that Ef_0 is of class C^2 is, of course, rather restrictive, but without it it maybe hard to obtain asymptotic normality; a more general class of limiting distributions (piecewise normal) for constrained problems has recently been identified by King and Rockafellar (1986). Note that this does not require that f_0 be of class C^2.

The assumption that $\sqrt{\nu}[v_S(x^\nu(\zeta)) - v_S(x^*)]$ converges in probability to 0, essentially means that the convergence of x^ν to x^* is "smooth." Of course, it will be satisfied if x^* belongs to the interior of the set S of constraints, in which case $v_S(x^*)$ and μ-a.s. $v_S(x^\nu(\zeta))$ are 0 for ν sufficiently large. It will also be trivially satisfied if the binding constraints are linear and, x^* and μ-a.s. $x^\nu(\zeta)$, belong to the linear variety spanned by these constraints. In fact, we can expect this condition to be satisfied unless the vector x^* is a boundary point at which the boundary has high curvature, in particular at point at which the boundary is not smooth, e.g., a vertex.

The condition about asymptotic normality of the subgradients $\sqrt{\nu}\, v^\nu(x^*)$ is best understood in the following context. Suppose condition (ii) is satisfied, in

fact let us assume that $v_S(x^*) = v_S(x^\nu(\zeta))$ a.s. And suppose also that P^ν is the empirical distribution. Then $\|v^\nu(x^*, \zeta)\|$ records the error of the estimate of the subgradients of Ef at x^*; note that $0 \in \partial Ef(x^*)$.

The first condition yields an estimate for the errors of the subgradients of $E^\nu f$ at x^* and Ef at $x^\nu(\zeta)$. The assumption is that enough information is collected so as to guarantee a certain convergence rate to 0. This is a crucial assumption and after the statement of the theorem we will return to this condition and give sufficient conditions that imply it.

THEOREM 4.8. *Under Assumptions* 4.4, 3.5 *and* 4.7, $\sqrt{\nu}(x^\nu(\cdot) - x^*)$ *is asymptotically normal with distribution* $N(0, \Sigma)$, *where* $\Sigma = H^{-1}\Sigma_1(H^{-1})^T$.

PROOF. Since Ef_0 is assumed to be C^2 and $x^\nu(\cdot)$ converges to x^*, for ν sufficiently large,

$$\nabla Ef_0(x^\nu) - \nabla Ef_0(x^*) = H(x^\nu - x^*) + o(\|x^\nu - x^*\|), \quad \mu\text{-a.s.}$$

Now, since $v(x^*) = 0$,

$$\sqrt{\nu}\,(\nabla Ef_0(x^\nu) - \nabla Ef_0(x^*)) = \sqrt{\nu}\,[v(x^\nu) + v^\nu(x^*)] - \sqrt{\nu}\,v^\nu(x^*) + \sqrt{\nu}\,[v_S(x^*) - v_S(x^\nu)].$$

By Assumption 4.7 the first term converges to 0 in probability, the second one converges in distribution to $N(0, \Sigma_1)$ and the third one converges in probability to 0. Hence, $\sqrt{\nu}[\nabla Ef_0(x^\nu) - \nabla Ef_0(x^*)]$ converges in distribution to $N(0, \Sigma_1)$ (Slutsky's theorem). This is then also the asymptotic distribution of $\sqrt{\nu}\,H(x^\nu - x^*)$. The result now follows by the nonsingularity of the matrix H. □

The remainder of this section, is devoted to recording certain conditions that will yield condition (i) of Assumption 4.7. In view of Markov's inequality it would suffice to control the variance of $\|v^\nu(x^*) + v(x^\nu)\|$ to obtain the desired convergence. More generally, we have the following:

LEMMA 4.9. *Suppose that* $E_\mu\{v^\nu(x^*, \zeta)\} = 0$, *that*

$$E_\mu\{\|v_0^\nu(x^*, \zeta) - v_0(x^*)\|^2\} \le \beta^2/\nu$$

and that

$$\frac{\|v^\nu(x^*, \zeta) + v(x^\nu(\zeta))\|}{\nu^{-1/2} + \|v(x^\nu(\zeta))\|} \text{ converges to } 0 \text{ in probability } (\mu).$$

Then, under Assumptions 4.4 *and* 3.5, *for any* (*measurable*) *selections* $v^\nu(x^*, \cdot)$, *with*

$$v^\nu(x^*, \zeta) \in \partial E^\nu f(x^*, \zeta), \quad \mu\text{-a.s.},$$

such that μ-*a.s.* $v(x^*) = 0$, *the random vector*

$$\sqrt{\nu}\,[v^\nu(x^*, \zeta) + v(x^\nu(\zeta))]$$

converges to 0 *in probability as* $\nu \to \infty$.

PROOF. We need to show that to any $\varepsilon > 0$, there corresponds ν_ε such that for all $\nu \geq \nu_\varepsilon$,

$$\mu\left[\|v^\nu(x^*) + v(x^\nu)\| \geq \nu^{-1/2}\delta_\varepsilon\right] \leq \varepsilon,$$

where $\delta_\varepsilon \to 0$ as $\varepsilon \to 0$.

Chebychev's inequality and the assumptions of the theorem imply that for all α,

$$\mu\left[\|v^\nu(x^*, \zeta)\| > \alpha\nu^{-1/2}\right] \leq \nu\alpha^{-2}E_\mu\|v^\nu(x^*, \zeta)\|^2 \leq \nu(\beta/\alpha)^2.$$

And hence with $\alpha^2 = 2\beta^2/\varepsilon$, we have

$$\mu\left[\|v^\nu(x^*)\| > \nu^{-1/2}\beta\sqrt{2}/\sqrt{\varepsilon}\right] \leq \varepsilon/2.$$

This, in conjunction with the last one of our assumptions, i.e.,

$$\mu\left[\|v^\nu(x^*) + v(x^\nu)\| \geq \varepsilon\left(\nu^{-1/2} + \|v(x^\nu)\|\right)\right] \leq \varepsilon/2, \tag{4.4}$$

implies that the events

$$\|v^\nu(x^*) + v(x^\nu)\| < \varepsilon\left(\nu^{-1/2} + \|v(x^\nu)\|\right)$$

and

$$\|v^\nu(x^*)\| \leq \nu^{-1/2}\beta\sqrt{2/\varepsilon}$$

have probability (μ) at least $1 - \varepsilon$. Thus, for ε small,

$$\mu\left[\|v(x^\nu)\| \leq \nu^{-1/2}(\beta + \varepsilon)/(1 - \alpha)\right] > 1 - \varepsilon,$$

since $\|v(x^\nu)\| \leq \|v^\nu(x^*) + v(x^\nu)\| + \|v^\nu(x^*)\|$. This, together with (4.4), gives

$$\mu\left[\|v^\nu(x^*) + v(x^\nu)\| < \nu^{-1/2}\varepsilon(1 + (\beta + \varepsilon)/(1 - \varepsilon))\right] > 1 - \varepsilon$$

and this yields the desired expression with $\delta_\varepsilon = \varepsilon(1 + (\beta + \varepsilon)/(1 - \varepsilon))$. □

It is easy to see why the condition $E_\mu\{v^\nu(x^*, \zeta)\} = 0$ would be satisfied when the P^ν are providing moment estimates that are at least as good as the empirical distributions. The same holds for the second assumption in Lemma 4.9, there is a reduction in the variance estimate that is at least as significant as that which would be attained by using the empirical distribution. Finally, the last assumption of Lemma 4.9 means that we can allow for a certain slack in the convergence in probability of $\sqrt{\nu}\|v^\nu(x^*) + v(x^\nu)\|$ to 0. In the Appendix of Dupačová and Wets (1987) we gave a derivation of this condition by using assumptions that are related to those used by Huber (1967). The differences are due to the fact that the probability measures $P^\nu(\cdot, \xi)$ are not necessarily the empirical ones and that subgradients are used instead of gradients.

5. Asymptotic Lagrangians. The results of Sections 3 and 4 can be extended to Lagrangians by relying on the theory of epi/hypo-convergence for saddle functions, Attouch and Wets (1983a). This gives us not just asymptotic properties for the sequence $\{x^\nu, \nu = 1, \dots\}$ of optimal solutions but also for the associated Lagrange multipliers.

We now introduce an explicit representation of the constraints in the formulation of the problem:

$$\begin{aligned} &\text{minimize} && z = E\{f_0(x, \xi)\} \\ (5.1)\qquad &\text{subject to} && f_i(x) \le 0, \qquad i = 1, \dots, s, \\ &&& f_i(x) = 0, \qquad i = s+1, \dots, m, \end{aligned}$$

where for $i = 1, \dots, m$, the f_i are finite-valued continuous functions and f_0 is a finite-valued random l.s.c. function. When instead of P, we use P^ν then the objective function is modified and becomes

$$E^\nu f_0(x) = \int_\Xi f_0(x, \xi) P^\nu(d\xi).$$

The (standard) associated Lagrangians are

$$L(x, y) = \begin{cases} Ef_0(x) + \sum_{i=1}^m y_i f_i(x) & \text{if } y_i \ge 0, \text{ for } i = 1, \dots, s, \\ -\infty & \text{otherwise} \end{cases}$$

and

$$L^\nu(x, y) = \begin{cases} E^\nu f_0(x) + \sum_{i=1}^m y_i f_i(x) & \text{if } y_i \ge 0, \text{ for } i = 1, \dots, s, \\ -\infty & \text{otherwise.} \end{cases}$$

Consistency can be studied in the same framework as that described at the beginning of Section 3. The Lagrangians L^ν also depend on ζ. Suppose that f_0 satisfies the conditions of Assumption 3.4. Note that some of these conditions are automatically satisfied since f_0 is a finite-valued random l.s.c. function. Suppose also that the $\{P^\nu, \nu = 1, \dots\}$ satisfy Assumption 3.5 with f_0 replacing f (in the asymptotic negligibility condition), then it follows from Lemma 3.6 that μ-a.s. the Lagrangians L^ν are finite-valued random l.s.c. functions on $(R^n \times (R^s_+ \times R^{m-s})) \times Z$; on the complement all functions L^ν are $-\infty$. This is all we need to guarantee the required measurability properties, in particular we have that

$$((x, y), \zeta) \mapsto L^\nu(x, y, \zeta) \text{ is } \mathscr{B}^{n+m} \otimes \mathscr{A}\text{-measurable.}$$

DEFINITION 5.1. The sequence of functions $\{h^\nu\colon R^n \times R^m \to [-\infty, \infty], \nu = 1, \dots\}$ epi/hypo-converges to $h\colon R^n \times R^m \to [-\infty, \infty]$ if for all (x, y) we have

(i) for every subsequence $\{h^{\nu_k}, k = 1, \dots\}$ and sequence $\{x^k\}_{k=1}^\infty$ converging to x, there exists a sequence $\{y^k\}_{k=1}^\infty$ converging to y such that

$$h(x, y) \le \liminf_{k \to \infty} h^{\nu_k}(x^k, y^k)$$

and

(ii) for every subsequence $\{h^{\nu_k}, k = 1, \dots\}$ and sequence $\{\hat{y}^k\}_{k=1}^\infty$ converging to y, there exists a sequence $\{\hat{x}^k\}_{k=1}^\infty$ converging to x such that

$$h(x, y) \ge \limsup_{k \to \infty} h^{\nu_k}(\hat{x}^k, \hat{y}^k).$$

This type of convergence of bivariate functions was introduced by Attouch and Wets (1983a) in order to study the convergence of saddle points; in Attouch and Wets (1983b) it is argued that it actually is the weakest type of convergence that will guarantee the convergence of saddle points.

THEOREM 5.2 (Consistency). *From Assumptions* 3.4 *and* 3.5, *with f replaced by* f_0, *it follows that there exists* $Z_0 \in \mathcal{F}$ *with* $\mu(Z \setminus Z_0) = 0$ *such that*

$$L = \operatorname*{epi/hypo-lim}_{\nu \to \infty} L^\nu, \quad \mu\text{-a.s.}$$

and hence

(i) *for all* $\zeta \in Z_0$, *any cluster point* $(\hat{x}, \hat{y})$ *of any sequence* $\{(x^\nu, y^\nu), \nu = 1, \dots\}$, *with* (x^ν, y^ν) *a saddle point of* $L^\nu(\cdot, \cdot, \zeta)$, *is a saddle point of* L;

(ii) *if* D *is a compact subset of* $R^n \times R^m$ *that meets for all* ν, *or at least for some subsequence, the set of saddle points of* $L^\nu(\cdot, \cdot, \zeta)$ *for some* $\zeta \in Z_0$, *then there exist* (x^ν, y^ν) *saddle points of* $L^\nu(\cdot, \cdot, \zeta)$ *for* $\nu = 1, \dots$ *that have at least one cluster point*;

(iii) *moreover, if the preceding condition is satisfied for all* $\zeta \in Z_0$ *and* L *has a unique saddle point, then there exists a sequence*

$$\{(x^\nu, y^\nu) \colon Z_0 \to R^n \times R^m, \nu = 1, \dots\}$$

of $\mathcal{F}^\nu$*-measurable functions that for all* $\zeta \in Z_0$, *determine saddle point of the* L^ν, *and converge to the saddle point of* L.

We note that sufficient condition for the existence of saddle points are provided by the condition introduced in Proposition 3.10 [with f the essential objective function of problem (5.1)], in conjunction with the Mangasarian–Fromovitz constraint qualification.

5.3 ASYMPTOTIC NORMALITY. The techniques of Section 4 can also be used to obtain asymptotic normality results. However, there is not yet a good concept of subdifferentiability for bivariate functions, except in the convex case [Rockafellar (1964)], and in the differentiable case, of course. With ∂L (∂L^ν resp.) the set of subgradients of the Lagrangians in the convex or differentiable case, the condition that (x^*, y^*) is a saddle point of L can be expressed as

$$0 \in \partial L(x^*, y^*),$$

and $0 \in \partial L^\nu(x^\nu, y^\nu, \zeta)$ in the case of L^ν. For example, in the convex case when all the functions $\{f_i, i = 0, 1, \dots, m\}$ are differentiable, this condition is equivalent to

$$\begin{aligned}
0 &= E\{\nabla f_0(x^*, \xi)\} + \sum_{i=1}^{m} y_i^* f_i(x^*), \\
0 &\geq f_i(x^*), && i = 1, \dots, s, \\
0 &= f_i(x^*), && i = s+1, \dots, m, \\
0 &= y_i^* f_i(x^*), && y_i^* \geq 0,\ i = 1, \dots, s,
\end{aligned}$$

and similarly for L^ν.

It is easy to see that when Assumptions 4.4 and 3.5 hold (with f_0 instead of f), as well as Assumption 4.7, but this time with v^ν and v subgradients of L^ν and L, respectively, and $S = R^n \times (R^s_+ \times R^{m-s})$, then by the same argument as in the proof of Theorem 4.8, we obtain

$$\sqrt{\nu}\,(x^\nu(\cdot) - x^*,\ y^\nu(\cdot) - y^*) \text{ is asymptotically normal.}$$

For an application to the preceding results to the case of linearly restricted L_1-regression (2.3), see Dupačová (1987).

REFERENCES

AITCHISON, J. and SILVEY, S. V. (1958). Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29** 813–828.

ARTHANARI, T. S. and DODGE, Y. (1981). *Mathematical Programming in Statistics*. Wiley, New York.

ATTOUCH, H. (1984). *Variational Convergence for Functions and Operators*. Pitman, London.

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

ATTOUCH, H. and WETS, R. (1983a). A convergence theory for saddle functions. *Trans. Amer. Math. Soc.* **280** 1–41.

ATTOUCH, H. and WETS, R. (1983b). A convergence for bivariate functions aimed at the convergence of saddle values. *Mathematical Theories of Optimization. Lecture Notes in Math.* **979** 1–42. Springer, Berlin.

AUBIN, J.-P. and EKELAND, I. (1984). *Applied Nonlinear Analysis*. Wiley, New York.

AUMANN, R. J. (1965). Integrals of set-valued functions. *J. Math. Anal. Appl.* **12** 1–12.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

BIRGE, J. and WETS, R. (1986). Designing approximation schemes for stochastic problems, in particular for stochastic programs with recourse. *Math. Programming Stud.* **27** 54–102.

CASTAING, C and VALADIER, M. (1976). *Convex Analysis and Measurable Multifunctions. Lecture Notes in Math.* **560**. Springer, Berlin.

CLARKE, R. (1975). Generalized gradients and applications. *Trans. Amer. Math. Soc.* **205** 247–262.

CLARKE, R. (1983). *Optimization and Nonsmooth Analysis*. Wiley, New York.

DEMPSTER, M., ED. (1980). *Stochastic Programming*. Academic, London.

DOLECKI, S., SALINETTI, G. and WETS, R. (1983). Convergence of functions: Equisemicontinuity. *Trans. Amer. Math. Soc.* **276** 409–429.

DUPAČOVÁ J. (1983a). Stability in stochastic programming with recourse. *Acta Univ. Carolina—Math. Phys.* **24** 23–34.

DUPAČOVÁ J. (1983b). The problem of stability in stochastic programming (in Czech). Dissertation, Faculty of Mathematics and Physics, Charles Univ.

DUPAČOVÁ J. (1984a). Stability in stochastic programming with recourse—Estimated parameters. *Math. Programming* **28** 72–83.

DUPAČOVÁ J. (1984b). On asymptotic normality of inequality constrained optimal decisions. In *Proc. Third Prague Conference on Asymptotic Statistics* (P. Mandl and M. Hušková, eds.) 249–257. North-Holland, Amsterdam.

DUPAČOVÁ J. (1987). Asymptotic properties of restricted L_1-estimates of regression. In *Statistical Data Analysis Based on the L_1 Norm and Related Methods* (Y. Dodge, ed.) 263–274. North-Holland, Amsterdam.

DUPAČOVÁ J. and WETS, R. (1986). Asymptotic behavior of statistical estimators and of optimal solutions for stochastic optimization problems. IIASA WP-86-41, Laxenburg, Austria.

DUPAČOVÁ J. and WETS, R. (1987). Asymptotic behavior of statistical estimators and of optimal solutions for stochastic optimization problems. II. IIASA WP-87-9, Laxenburg, Austria.

ERMOLIEV, Y., GAIVORONSKI, A. and NEDEVA, C. (1985). Stochastic optimization problems with incomplete information on distribution function. *SIAM J. Control Optim.* **23** 696–716.

FLETCHER, R. and POWELL, M. J. D. (1963). A rapidly convergent descent method for minimization. *Comput. J.* **6** 163–168.

HERBACH, 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. Statist.* **30** 939–959.

HIRIART-URRUTY, J.-B. (1976). About properties of the mean functional and of the continuous infimal convolution in stochastic convex analysis. *Optimization Techniques II. Lecture Notes in Computer Science* **41** 763–789. Springer, Berlin.

HUBER, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. Univ. California Press.

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

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

IBRAGIMOV, I. A. and HAŚMINSKII, R. Z. (1981). *Statistical Estimation. Asymptotic Theory*. Springer, New York.

IOFFE, A. (1978). Survey of measurable selection theorems: Russian literature supplement. *SIAM J. Control Optim.* **16** 729–731.

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

JUDGE, G. G. and TAKAYAMA, T. (1966). Inequality restrictions in regression analysis. *J. Amer. Statist. Assoc.* **61** 166–181.

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

KING, A. and ROCKAFELLAR, R. T. (1986). Non-normal asymptotic behavior of solution estimates in linear-quadratic stochastic optimization. Manuscript, Univ. Washington.

LEE, S.-Y. (1980). Estimation of covariance structure models with parameters subject to functional restraints. *Psychometrika* **45** 309–324.

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

LIEW, C. K. (1976). Inequality constrained least squares estimation. *J. Amer. Statist. Assoc.* **71** 746–751.

PRÉKOPA, A. (1973). Contributions to the theory of stochastic programming. *Math. 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. Multivariate Anal.* **1** 257–275.

ROBINSON, S. (1987). Local epi-continuity and local optimization. *Math. Programming* **37** 208–22.

ROCKAFELLAR, R. T. (1964). Minimax theorems and conjugate saddle functions. *Math. Scand.* **14** 151–173.

ROCKAFELLAR, R. T. (1970). *Convex Analysis*. Princeton Univ. Press, Princeton, N.J.

ROCKAFELLAR, R. T. (1976). Integral functionals, normal integrands and measurable multifunctions. *Nonlinear Operators and the Calculus of Variations. Lecture Notes in Math.* **543** 157–207. Springer, Berlin.

ROCKAFELLAR, R. T. (1979). Directionally Lipschitzian functions and subdifferential calculus. *Proc. London Math. Soc.* **39** 331–355.

ROCKAFELLAR, R. T. (1980). Generalized directional derivatives and subgradients of nonconvex functions. *Canad. J. Math.* **32** 257–280.

ROCKAFELLAR, R. T. (1981). *The Theory of Subgradients and Its Applications to Problems of Optimization. Convex and Nonconvex Functions*. Halderman, Berlin.

ROCKAFELLAR, R. T. (1983). Generalized subgradients in mathematical programming. In *Mathematical Programming: The State-of-the-Art* (A. Bachem, M. Grötschel and B. Korte, eds.) 368–390. Springer, Berlin.

ROCKAFELLAR, R. T. and WETS, R. (1982). On the interchange of subdifferentiation and conditional expectation for convex functionals. *Stochastics* **7** 173–182.

ROCKAFELLAR, R. T. and WETS, R. (1984). Variational systems, an introduction. *Multifunctions and Integrands. Lecture Notes in Math.* **1091** 1–54. Springer, Berlin.

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

SALINETTI, G. and WETS, R. (1986). Convergence of infima, especially stochastic infima. Technical Report, Univ. Roma "La Sapienza."

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

SHAPIRO, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *Internat. Statist. Rev.* To appear.

SILVEY, S. D. (1959). The Lagrangian multiplier test. *Ann. Math. Statist.* **30** 389–407.

SOLIS, R. and WETS, R. (1981). A statistical view of stochastic programming. Technical Report, Univ. Kentucky.

THOMPSON, W. A., JR. (1962). The problem of negative estimates of variance components. *Ann. Math. Statist.* **33** 273–289.

WAGNER, D. (1977). Survey of measurable selection theorems. *SIAM J. Control Optim.* **15** 859–903.

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

WETS, R. (1973). On inf-compact mathematical programs. *Fifth Conference on Optimization Techniques. I. Lecture Notes in Computer Science* **3** 426–436. Springer, Berlin.

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

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

WETS, R. (1984). Modeling and solution strategies for unconstrained stochastic optimisation problems. *Ann. Oper. Res.* **1** 3–22.

DEPARTMENT OF MATHEMATICAL STATISTICS
CHARLES UNIVERSITY
SOKOLOVSKA 83
18600 PRAGUE 8
CZECHOSLOVAKIA

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
DAVIS, CALIFORNIA 95616