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

Constrained Estimation: Consistency and Asymptotics

Wets, R.J.-B.

IIASA Working Paper

WP-90-075

December 1990

Working Paper

CONSTRAINED ESTIMATION: CONSISTENCY AND ASYMPTOTICS

Roger J-B Wets

WP-90-075
December 1990

IIASA

International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria

Telephone: (0 22 36) 715 21 *0 □ Telex: 079137 iiasa a □ Telefax: (0 22 36) 71313

CONSTRAINED ESTIMATION: CONSISTENCY AND ASYMPTOTICS

Roger J-B Wets

WP-90-075
December 1990

International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria

Telephone: (0 22 36) 715 21 *0 □ Telex: 079 137 iiasa a □ Telefax: (0 22 36) 71313

Foreword

We review some of the recent results obtained for constrained estimation, involving possibly nondifferentiable criterion functions. New tools are required to push consistency and asymptotic results beyond those that can be reached by classical means.

Alexander B. Kurzhanski
Chairman
System and Decision Sciences Program

1. Introduction

The choice of a statistic is limited by our capability of using it in a practical environment. This used to mean that the estimator had to be a relatively simple function. With the advent of the computer this no longer needs to be the case. We can build estimators that involve conditional "switches", nondifferentiable functions, and may not be expressible in terms "standard" functions. For example, the estimator could be the optimal solution of a certain optimization problem (possibly involving constraints). This is certainly going to be the case if there are restrictions on the choice of the estimator and more than a single parameter needs to be estimated. These restrictions could be simple nonnegativity constraints (as in the estimation of a variance), but much more complex restrictions could also be present, such as certain relations between the parameters that need to be estimated. Classical techniques, that can still be used to handle, for example, least square estimation with linear equality constraints on the parameters, break down if there are inequality constraints or we replace least squares by a nondifferentiable criterion function.

The following example, involving least square estimation of regression coefficients, will help make more concrete the issues that need to be addressed.

Example 1.1. *Estimating regression coefficients.* Assume that the dependent y variables can be predicted on the basis of information provided by independent x variables. In the linear model, the observations y_j would be generated according to

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

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

In practice, 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, 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:

$$\text{minimize} \sum_{j=1}^{\nu} \left(y_j - \sum_{i=1}^{p} x_{ij}\beta_i \right)^2$$

$$\text{subject to} \sum_{i=1}^{p} a_{ki}\beta_i \leq c_k,\ k = 1, \ldots, m.$$

The (best) estimate for the unknown coefficient is then the optimal solution of a quadratic program

A robust version would lead to a problem of the type:

$$\text{minimize} \sum_{j=1}^{\nu} \rho(y_j - \sum_{i=1}^{p} x_{ij}\beta_i)$$

$$\text{subject to} \sum_{i=1}^{p} a_{ki}\beta_i \leq c_k, 1 \leq k \leq m.$$

with ρ a convex function (with bounded derivatives of sufficiently high order), for example,

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & \text{for } |u| < c, \\ c|u| - \frac{1}{2}c^2 & \text{for } |u| \geq c. \end{cases}$$

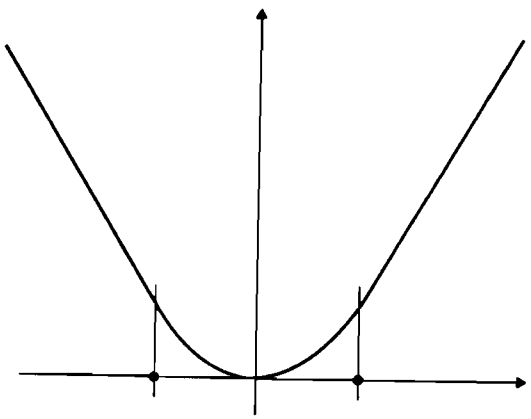

Fig. 1. "Robust" least squares

The robust estimator is then the optimal solution of a (generalized) linear-quadratic optimization problem.

Minimizing the sum of absolute errors yields the optimization problem

$$\text{minimize} \sum_{j=1}^{\nu} |y_j - \sum_{i=i}^{p} x_{ij}\beta_i|$$

$$\text{subject to } \sum_{i=1}^{p} a_{ki}\beta_i \leq c_k, 1 \leq k \leq m.$$

The estimator is the optimal solution of a problem featuring a nondifferentiable objective. (This simple case can be reformulated as a linear programming problem.) □

We have to accept the view that finding best estimates may require solving a complex (involving constraints) optimization problem. And, thus, there is a need to come up with a mathematical framework that will back up such an approach, i.e., a mathematical theory that will provide the tools to study consistency and to analyze the asymptotics of this this class of problems. We are going to be concerned with a review of a number of results that propose such a theory. The approach that is detailed here is not the only one possible, related results have been obtained by Shapiro(1989), Vogel (1988) Ermoliev and Norkin (1989) by relying on different tools.

2. The model

We take for general model, the following optimization problem:

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

More precisely: Let $(\Xi, \mathcal{A}, P)$ be a probability space, with Ξ – the support of P – a closed subset of a Polish space X, and $\mathcal{A}$ the Borel σ-field relative to Ξ; we may think of Ξ as the set of possible values of the random element ξ. With P known, the problem is to:

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

where

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

and

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

is a random lsc function (see below); $(Ef)(x) = \infty$, if there is no summable function that majorizes $\xi \mapsto f(x,\xi)$. We allow for the function f to take on the value ∞, in order to allow for constraints; $f(x,\xi) = \infty$ correspond to the situation of an infeasible x (that does not satisfy the constraints) if ξ is observed.

In terms of example 1.1 (regression model), we have that

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

when the criterion is least squares;

$$f(x,\xi) = \begin{cases} \rho(\xi_0 - \sum_{i=1}^{p} \xi_i x_i) & \text{if } x \in S, \\ +\infty & \text{otherwise.} \end{cases}$$

when the criterion is the robust version of least squares; and

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

in the case of minimizing the sum of absolute errors.

When dealing with problems of this type, the traditional tools of analysis are no longer quite appropriate. The classical geometrical approach that associates functions with their graphs must be abandoned in favor of a new geometrical viewpoint that associates functions with their epigraphs (or hypographs). The *epigraph* of a function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is the set

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

The function $f : \mathbb{R}^n \times \Xi \to \overline{\mathbb{R}}$ is a *random lsc (lower semicontinuous) function* if and only if

$$\text{the set-valued mapping } \xi \mapsto \operatorname{epi} f(\cdot,\xi) \text{ is closed-valued ,}$$

$$\text{the set-valued mapping } \xi \mapsto \operatorname{epi} f(\cdot,\xi) \text{ is a random set ;}$$

recall that set-valued mapping $\xi \mapsto \Gamma(\xi) : \Xi \to \mathbb{R}^{n+1}$ is a random set if for all closed sets $F \subset \mathbb{R}^{n+1}$

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

Proposition 2.1. *(*Yankov-von Neuman's Measurable Selections' Theorem). *If $\Gamma : \Xi \rightrightarrows \mathbb{R}^n$ is a random closed set, then there exists a least one measurable selector, i.e., a random vector $x : \operatorname{dom}\Gamma \to \mathbb{R}^n$ such that for all $\xi \in \operatorname{dom}\Gamma$, $x(\xi) \in \Gamma(\xi)$, where*

$$\operatorname{dom}\Gamma := \{\xi \in \Xi | \Gamma(\xi) \neq \emptyset\} = \Gamma^{-1}(\mathbb{R}^n) \in \mathcal{A}.$$

As an immediate consequences of this proposition and the definition of a random lsc functions, we have:

Proposition 2.2. *Let $f : \mathbb{R}^n \times \Xi \to \overline{\mathbb{R}}$ be a random lsc function. Then for any $\mathcal{A}$-measurable function $x : \Xi \to \mathbb{R}^n$, the function $\xi \mapsto f(x(\xi),\xi)$ is $\mathcal{A}$-measurable . Moreover, the infimal function*

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

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

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

is a random closed set (from Ξ into the subsets of $\mathbb{R}^n$). This implies that there exists a measurable selector

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

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

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 number of samples of ξ — then to estimate x^* we usually have to rely on the solution of an optimization problem that ***approximates*** (hopefully) the one constructed with the P-measure:

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

where

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

The measure P^ν could be the empirical measure, but more generally it is the *best* approximation of P given 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 are going to think of the measures P^ν as random measures that depend on a random vector ζ whose realizations determine the information available.

Let $(Z,\mathcal{F},\mu)$ be a sample space with $(\mathcal{F}^\nu)_{\nu=1}^\infty$ an increasing sequence of σ-field contained in $\mathcal{F}$. A sample ζ — e.g. $\zeta = \{\xi^1,\xi^2,\ldots\}$ obtained in this case by independent sampling of the variables ξ — leads us to a sequence $\{P^\nu(\cdot,\zeta),\ \nu = 1,\ldots\}$ of probability measures defined on $(\Xi,\mathcal{A})$. Only the information collected up to stage ν can be used in the choice of P^ν, thus for all $A \in \mathcal{A}$, $\zeta \mapsto P^\nu(A,\zeta)$ is $\mathcal{F}^\nu$-measurable . Since P^ν depends on ζ, so does the approximating problem, and in particular so does its solution set, and so does every selection x^ν of this solution set. A sequence of estimators $\{x^\nu : Z \to \mathbb{R}^n, \nu = 1,\ldots\}$ is *consistent* if μ-almost surely they converge to x^*.

The ultimate goal is to show that under rather benign assumptions, the solutions of the approximating problems are consistent if the measures P^ν converge narrowly (weakly) to P μ-almost surely. Our approach will actually derive a stronger result: we are going to show that the sequence of approximating problem are (epi-)consistent! To do this, we need a notion of convergence for "optimization" problems.

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

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

and

$$\text{for some } \{x^\nu\}_{\nu=1}^\infty \text{ converging to } x, \limsup_{\nu\to\infty} g^\nu(x^\nu) \leq g(x);$$

we then say that $g = \operatorname{epi-lim}_{\nu\to\infty} g^\nu$. Note that these conditions imply that g is lower semicontinuous.

Proposition 2.3. [Attouch and Wets (1981), Salinetti and Wets (1986)]. *Suppose* $\{g; g^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}, \nu = 1, \ldots\}$ *is a collection of functions such that* $g = \operatorname{epi-lim}_{\nu\to\infty} g^\nu$. *Then*

$$\limsup_{\nu\to\infty} \inf g^\nu \leq \inf g,$$

and, if $x^k \in \operatorname{argmin} g^{\nu_k}$ *for some subsequence* $\{\nu_k\}$ *and* $x = \lim_{k\to\infty} x^k$, *it follows that* $x \in \operatorname{argmin} g$, *and* $\lim_{k\to\infty} \inf g^{\nu_k} = \inf g$.

Thus, if $g = \operatorname{epi-lim}_{\nu\to\infty} g^\nu$ and if there exists a bounded set $D \subset \mathbb{R}^n$ such that for some subsequence $\{\nu_k\}$,

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

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

To prove consistency of the estimators it will certainly be sufficient to prove that μ-almost surely the functions $E^\nu f$ epi-converge to Ef.

3. Consistency

Three different results are going to be featured. The first two are general in nature, whereas the last one is restricted to the case when the P^ν are empirical measures. The first one of these results relies on the following assumptions.

Assumption A1. *For the random lsc function* $f : \mathbb{R}^n \times \Xi \to (-\infty, \infty]$,

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

for all $x \in S$, $\xi \mapsto f(x,\xi)$ *is continuous on* Ξ,

Assumption B1. *For all* $\xi \in \Xi$, *the function* $x \mapsto f(x,\xi)$ *is locally lower Lipschitz on* S: *to every* x *in* S, *one can associate a neighborhood* V *of* x *and a bounded continuous function* $\beta : \Xi \to \mathbb{R}$ *such that for all* $x' \in V \cap S$ *and* $\xi \in \Xi$,

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

Assumption C1. *For all ν, $P^\nu : \mathcal{A} \times Z \to [0,1]$ are "conditional" probability measures, i.e., for all $\zeta \in Z$ $P^\nu(\cdot,\zeta)$ is a probability measure on $(\Xi,\mathcal{A})$, and for all $A \in \mathcal{A}$, $\zeta \mapsto P^\nu(A,\zeta)$ is $\mathcal{F}^\nu$-measurable.*

Assumption D1. *For μ-almost all $\zeta \in Z$, $P^\nu(\cdot,\zeta)$ converge narrowly (weakly) to P.*

Assumption E1. *For all $x \in S$, μ-almost all $\zeta \in Z$, the collection of measures $\{P; P^\nu(\cdot,\zeta)\}$ is $f(x,\cdot)$-tight (asymptotic negligibility), i.e., to every $x \in S$ and $\varepsilon > 0$ there correspond a compact set $K_\varepsilon \subset \Xi$ such that for $\nu = 0,1,\ldots$*

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

and

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

One important distinction between this approach and the "classical" approach — consult, for example, Huber (1967) and Ibragimov and Has'minski (1981) — is that they assume that the set S is open. This might seem to be just an innocuous difference, but it leads to a very different collection of asymptotic results! With S open one may hope for asymptotic normality, but, as we are going to see, that is not the case if S is closed.

To simplify notations we follow, henceforth, the accepted practice of dropping reference of the dependence on ζ, this applies to the measures P^ν, the resulting functions $E^\nu f$ and, the argmin sets associated with those functions. Nonetheless one should always keep in mind that all μ-a.s. statements refer to the underlying probability space $(Z,\mathcal{F},\mu)$.

Theorem 3.1. [Dupačová and Wets (1988), theorems 3.7 and 3.8] *Let f be a random lsc function. Suppose $\{E^\nu f, \nu = 1,\ldots\}$ 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 $Ef(x) = E\{f(x,\xi)\} = \int_\Xi f(x,\xi)\, P(d\xi)$ such that f and the collection $\{P; P^\nu, \nu = 1,\ldots\}$ satisfy assumptions A1-E1. Then, μ-almost surely,

(i) the expectation functionals $E^\nu f : \mathbb{R}^n \times Z \to \overline{\mathbb{R}}$ are random lsc functions,

(ii) with $\underset{\nu\to\infty}{\text{ptwse-lim}}\, E^\nu f$ the pointwise limit of the sequence $E^\nu f$,

$$Ef = \underset{\nu\to\infty}{\text{epi-lim}}\, E^\nu f = \underset{\nu}{\text{ptwse-lim}}\, E^\nu f.$$

With the help of the results and the observations in the previous section, we come to the following (important) corollary.

Corollary 3.2. (Consistency). *Let f be a random lsc function. Suppose $\{Ef; E^\nu f, \nu = 1,\ldots\}$ is the collection of expectation functionals defined in theorem 3.1. Then, under assumptions A1-E1, we have*

$$\limsup_{\nu\to\infty}(\inf E^\nu f) \le \inf Ef \quad \mu\text{-almost surely .}$$

Moreover, there exists $Z_0 \in \mathcal{F}$ of measure 1 such that

(i) for all $\zeta \in Z_0$, if $\hat{x}$ is a cluster point of any sequence $\{x^\nu\}$ with $x^\nu \in \operatorname{argmin} E^\nu f^\nu(\cdot,\zeta)$ then $\hat{x} \in \operatorname{argmin} Ef$ (i.e., $\hat{x}$ is the actual value of the quantities being estimated),

(ii) for all ν, $\zeta \mapsto \operatorname{argmin} E^\nu f(\cdot,\zeta) : Z_0 \rightrightarrows \mathbb{R}^n$, is a random closed set that is $\mathcal{F}^\nu$-measurable.

Proof. The first inequality immediately follows from the corresponding one in theorem 2.3. The epi-convergence μ-almost surely of the expectation functionals $E^\nu f$ to Ef (theorem 3.1) in combination with theorem 2.3 yield the assertion about the cluster points of sequences that lie in $\operatorname{argmin} E^\nu f$. That $\zeta \mapsto \operatorname{argmin} E^\nu f(\cdot,\zeta)$ is a random closed set is also a consequence of the theorem and proposition 2.2. □

Let also observe that we do not assume the uniqueness of the solution; the "classical" analysis relies fundamentally on such an assumption, Wald(1940), Huber(1967).

Similar results can be obtained under a somewhat different set of assumptions. Assumption B1 is very much "finite dimensional" in nature. It is difficult to find a "reasonable" parametric estimation problem whose criterion function would fail to satisfy assumption B1. On the other hand, if the decision variable x belongs to an infinite dimensional space (as would be the case in nonparametric statistics), it is difficult to think of a "reasonable" criterion function that would satisfy assumption B1. This has led a study of the consistency problem from a different angle. We are going to exploit the following fact: if P^ν is a sequence of probability measure converging narrowly to a probability measure P, then they converge uniformly on certain sets. We restrict our attention to the finite dimensional case to facilitate comparison. Let $\mathbb{B}$ denote the unit ball in $\mathbb{R}^n$.

Assumption A2. *For the random lsc function $f : \mathbb{R}^n \times \Xi \to (-\infty, \infty]$,*

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

for all $\alpha \in \mathbb{R}$, $\operatorname{lev}_\alpha f = \{(x,\xi) \,|\, f(x,\xi) \le \alpha\} \subset (S \times \Xi)$ are closed, and f is bounded on bounded subsets of $S \times \Xi$. Moreover, for $\rho > 0$ let $w_\rho(\xi) := \inf\{f(x,\xi) \,|\, x \in \rho\mathbb{B}\}$ and assume there exist usc (upper semicontinuous) functions $\{u_\rho : X \to \overline{\mathbb{R}} \,|\, u \text{ usc }, u_\rho \le w_\rho, \rho \in \mathbb{R}_+\}$, and a measurable function h such that $h \le u_\rho$ and $\int_\Xi h\, d\mu > -\infty$

Assumption B2. *The strict (inf-)level sets of the functions* $\{f(x,\cdot), x \in S\}$,

$$\operatorname{lev}_\alpha^< f(x,\cdot) := \{\xi \in \Xi \,|\, f(x,\xi) < \alpha\}$$

are P-continuity sets, i.e., their boundaries are P-null sets.

Theorem 3.3. [Lucchetti and Wets (1990), theorems 12 and 13] *Let* f *be a random lsc function. Suppose* $\{E^\nu f, \nu = 1,\ldots\}$ *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 $Ef(x) = E\{f(x,\xi)\} = \int_\Xi f(x,\xi)\, P(d\xi)$ *such that* f *and the collection* $\{P; P^\nu, \nu = 1,\ldots\}$ *satisfy assumptions A2, B2, C1-E1. Then,* μ*-almost surely, the expectation functionals* $E^\nu f : \mathbb{R}^n \times Z \to \overline{\mathbb{R}}$ *are random lsc functions, and*

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

Corollary 3.2 is also a corollary of this theorem, and thus consistency of the set of estimators $\operatorname{argmin} E^\nu f$ is guaranteed. As already pointed out earlier, the major difference lies in the use made of assumption B2. The conditions in assumption A2 are mostly technical in nature and do not seem to exclude any potential application.

The third epi-consistency results is of a very different nature. It relies on the law of large numbers for random sets, and applies only in the convex case. The result can easily be extended to the case when the decision variables x lie in a reflexive Banach space. Again we state the assumptions in a manner that facilitates comparison.

Assumption A3. $f : \mathbb{R}^n \times \Xi \to (-\infty, \infty]$, *is a random convex lsc function, i.e., for all* ξ, *the function* $x \mapsto f(x,\xi)$ *is convex. Moreover,* Ef *is finite somewhere, say at* $\bar{x}$.

Assumption B3. *There exists a measurable function* $u : \Xi \to \mathbb{R}^n$ *such that*

$$f(x,\xi) - f(\bar{x},\xi) \geq \langle u(\xi), x - \bar{x}\rangle \quad \text{for all } x \in \mathbb{R}^n$$

and $\int \|u(\xi)\|\, P(d\xi)$ *is finite.*

Assumption C3. *For all* ν, $P^\nu : \mathcal{A} \times Z \to [0,1]$ *is the empirical measure (process) obtained from the first* ν *observations of the random vector* ξ.

Theorem 3.4. [King and Wets (1990), theorem 2.2] *Let* f *be a random convex lsc function. Suppose* $\{E^\nu f, \nu = 1,\ldots\}$ *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 $Ef(x) = E\{f(x,\xi)\} = \int_\Xi f(x,\xi)\,P(d\xi)$ *such that* f *and the collection* $\{P; P^\nu, \nu = 1,\ldots\}$ *satisfy assumptions A3-C3. Then,* μ*-almost surely, the expectation functionals* $E^\nu f : \mathbb{R}^n \times Z \to \overline{\mathbb{R}}$ *are random lsc functions, and*

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

Again, Corollary 3.2 is a corollary of this theorem. We note that assumption B3 refers to a function u that must determine for all ξ a ***subgradient*** of $f(\cdot,\xi)$ at $\bar{x}$. The condition that this function must be summable is thus a growth-type condition of the same nature as that found in assumption B1.

The argument used in the proof of theorem 3.4 is based on (1) the fact that it will suffice to prove (in this convex case) that the conjugate functions of $E^\nu f$ epi-converge to the conjugate of Ef (a sequence of convex functions epi-converge to a limit function g if and only if their conjugates converge to the conjugate of g), and (2) that functions epi-converge if and only if their epigraphs converge as sets (Painlevé-Kuratowski convergence), and (3) the epigraphs of the conjugates of the $E^\nu f$ are the (normalized) sum of the random sets epi $f^*(\cdot,\xi^k)$ for $k = 1,\ldots,\nu$ where f^* is the conjugate of f and ξ^k is the random variable to be observed at stage k (the ξ^k are iid). This allows for the use of the law of large numbers for random set of Artstein and Hart (1981).

4. Asymptotics: Convergence rates

Let us now proceed with the assumption that the sequence of problems

$$\text{minimize } E^\nu f(\cdot,\zeta), \qquad \nu = 1,\ldots,$$

are epi-consistent with

$$\text{minimize } Ef.$$

The next question on our agenda is to know at which rate the solutions of the approximating problems will actually converge to true parameter x^*, i.e., the solution of the limit problem. Of course, because the solutions of the approximating problems depend on the random variable ζ (modeling the sampling process), we can only achieve "probabilistic" convergence rates. Typically, we are going to be interested in the distribution of the error, i.e., the quantity

$$\|x^* - x^\nu(\zeta)\|, \qquad \text{or more generally } x^* - x^\nu,$$

as ν tends to ∞. If the problem is without constraints, then one may hope for asymptotic normality, if the functions $x \mapsto f(x,\xi)$ are sufficiently smooth (at least of class C^2) in

a neighborhood of x^*. And this is actually the case as demonstrated first by Huber (1967). Dupačova and Wets (1988) extend Huber's result to the case when the $f(\cdot,\xi)$ are not necessarily differentiable and there is some provision for equality constraints. But in general, one cannot expect the asymptotic distribution of the error to be normal. Indeed let us consider the following very simple example. Let $\{\zeta^\nu \mid \nu = 1,\ldots\}$ be iid (independent, identically distributed) normal random variables with known mean 0 and variance σ^2. Let

$$x^\nu \in \operatorname{argmin}\left\{\frac{1}{\nu}\sum_{k=1}^{\nu}|x-\zeta^k|^2 \quad \text{on } x \geq 0\right\}.$$

The asymptotic distribution of x^ν is easy to derive from the expression

$$x^\nu = \max\left\{0, \frac{1}{\nu}\sum_{k=1}^{\nu}\zeta^k\right\} := M(\frac{1}{\nu}\sum_{k=1}^{\nu}\zeta^k);$$

half of the probability mass is concentrated at $\{0\}$, the other half of the mass is spread on $\mathbb{R}_+$ like the right half of a normal distribution with variance σ^2. The mapping M is non differentiable at $\{0\}$. However it is directionally differentiable at $\{0\}$, with

$$M'(0;h) = \begin{cases} 0 & \text{if } h \leq 0, \\ h & \text{if } h > 0, \end{cases}$$

and the asymptotic distribution is given by M': $M'(0,\tilde{\zeta})$ with $\tilde{\zeta}$ a random variable with 0 mean and variance σ^2. As can be surmised from this example, most mappings that arise in constrained optimization are nondifferentiable, and the asymptotic normality for the distribution of the error is bound to be the exception rather than the rule.

However, if normality will not be at hand, as the preceding example confirms, the example also suggest that the asymptotic distribution of the error will have to be of the following form: $M'(x^*;\tilde{\zeta})$ where $\tilde{\zeta}$ is a normally distributed random variable and $M'(x^*;\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ is a (nondifferentiable) Lipschitz mapping that roughly speaking corresponds to the directional derivative of map argmin mapping.

We are going to feature two results. The first one is mostly concerned with conditions under which one can still prove asymptotic normality. The second one gives a general answer in the convex case.

To state these results we shall be in need of some "standard" subdifferentiability results that are reviewed in the Appendix. For the sake of this discussion let us simply deal with problems of the following kind:

$$\text{minimize } Ef_0(x) \text{ subject to } x \in X$$

where $f_0(\cdot,\xi)$ is a locally Lipschitz function and X is a polyhedral set. When $f_0(\cdot,\xi)$ is locally Lipschitz, the set of subgradients $\partial f_0(\bar{x},\xi)$ at $\bar{x}$ can characterized in the following terms:

$$\partial f_0(\bar{x},\xi) = \text{co}\{v = \lim_{x'\to\bar{x}} \nabla f_0(x',\xi) \,\big|\, f_0(\cdot,\xi) \text{ is smooth at } x'\};$$

see the Appendix for a brief justification of the preceding definition.

We shall also assume that $\xi \mapsto f_0(x,\xi)$ is continuous (for all $x \in X$), and that to any bounded open subset V of $\mathbb{R}^n$ there corresponds a function β uniformly integrable with respect to $P^\nu, \nu = 1,\ldots$, such that for any pair x^0, x^1 in V:

$$|f_0(x^0,\xi) - f_0(x^1,\xi)| \le \beta(\xi)\|x^0 - x^1\|.$$

Together these conditions are those that appear in assumption A1, except, of course, for the ("upper") locally Lipschitz property.

It is easy to show that under these conditions $\xi \mapsto \partial f_0(\bar{x},\xi)$ is a random closed convex set and

$$\partial E f_0(\bar{x}) \subset E\{\partial f_0(\bar{x},\xi)\} = \int \partial f_0(\bar{x},\xi),$$

as follows from a result of Clarke (1983); the integral of set-valued mappings is to be interpreted in the sense of Aumann (1965). More to the point, with δ_X the indicator function of X, we have that

$$\partial(E f_0(\bar{x}) + \delta_X(\bar{x})) \subset E\{\partial(f_0(\bar{x},\xi + \delta_X(\bar{x}))\}$$

with equality if Ef_0 is subdifferentiably regular at $\bar{x}$ (consult the Appendix); $\delta_X(\bar{x})$ is the polar of the tangent cone of X at $\bar{x}$ if $\bar{x} \in X$.

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 $\partial\delta_X(x)$. In view of the above and the measurable selection theorem 2.1, if $x \in X$, we always have that $v(x) \in \partial E f(x)$ (with $f = f_0 + \delta_X$) implies the existence of $v_s(x) \in \partial\delta_X(x)$ 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. Because the functions $f_0(\cdot,\xi)$, as well as δ_X, are μ-a.s. subdifferentially regular, then a type of converse statement also holds. We have that

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

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

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

Similarly,

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

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

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

Assumptions C1-E1 that deal with the "probabilistic" structure of the problem will be supplemented with the following one:

Assumption F. Statistical Information. *The probability measures* $\{P^\nu, \nu = 1,\ldots\}$ *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.

Theorem 4.1. Dupačová and Wets (1988). *Under assumptions A1, C1-E1, F and for* $x \mapsto f_0(x,\xi)$ *locally Lipschitz for all* $\xi \in \Xi$, *X polyhedral,*

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

with distribution $N(0,\Sigma)$ *where* $\Sigma = H^{-1}\Sigma_1(H^{-1})^T$.

Before we move on to a more interesting result, let us examine condition (ii):

$$\sqrt{\nu}[v_s(x^\nu(\zeta)) - v_s(x^*)] \text{ converges to 0 in probability .}$$

It basically means that convergence of x^ν to x^* must be smooth. Of course, this will be the case if x^* belongs to the interior of the set X of constraints, in which case $v_s(x^*)$ and μ-a.s. $v_s(x^\nu(\zeta))$ are zero for ν sufficiently large. It will also be satisfied in a few other rare

situations. This is the condition that imposes strict restrictions on the use of this theorem. We already knew from the example given at the beginning of this section that in fact this theorem has a limited range of applicability.

To capture the asymptotic properties of $x^\nu - x^*$ it will be necessary to enlarge the class of acceptable limit distribution. A good class, that will certainly take care of all the examples mentioned in section 1 as well as many other situations, is that of random vector that are *conically normal*; by this we mean that their distribution is the projection (with respect to a certain norm) of a normal distribution on a convex cone. More specifically, let K be a convex cone in $\mathbb{R}^N$ and θ a random N-vector normally distributed with mean 0 and covariance matrix Σ, then $\tilde{\theta}$, the "projection" of θ on K, will have a conically normal distribution. The distribution of $\tilde{\theta}$ (on $\mathbb{R}^N$) is given by:

$$P_K(A) = P(\{x \mid \text{prj}_K(x) \in A\}) \quad \forall \text{ measurable sets } A.$$

The following theorem only applies to the convex case, but captures the essence of the type of results one may hope to derive. It is direct application of general results obtained by King (1988) for Lipschitz mappings.

Theorem 4.2. King (1988). *Let $f_0 : \mathbb{R}^n \times \Xi \to \overline{\mathbb{R}}$ be a random lsc function, convex in x and let $X \subset \mathbb{R}^n$ be a polyhedral set. Suppose that*

$$\textit{minimize } Ef_0(x) \quad \textit{on } X$$

has an optimal solution x^*. Suppose also

(a) $x \mapsto f_0(x, \xi)$ is $\mathcal{C}^1$ for all $\xi \in \Xi$,

(b) Ef_0 is $\mathcal{C}^2$ with $H := \nabla^2 Ef_0(x^*)$ positive definite,

(c) $E\|\nabla f(x^*, \xi)\|^2 < \infty$, $\quad \xi \mapsto \sup_{x_1, x_2 \in X} \frac{|\nabla f(x_1, \xi) - \nabla f(x_2, \xi)|}{|x_1 - x_2|}$ is $\mathcal{L}^2$,

(d) P^ν are the empirical measures.

Let θ be a normally distributed random vector with mean 0 and variance

$$E\{\nabla f_0(x^*, \xi) \nabla f_0(x^*, \xi)^T\}.$$

Then

$$\sqrt{\nu}(x^\nu - x^*) \textit{ is asymptotically conically normal .}$$

The asymptotic distribution is the projection, with respect to the metric induced by the matrix H, on the cone

$$X' = \{u \in T_X(x^*) \mid E\nabla f_0(x^*, \xi)u = 0\}$$

where $T_X(x^*)$ the tangent cone to X at x^*. The asymptotic distribution is the distribution of the vector:

$$\tilde{\theta} = \underset{u \in X'}{\operatorname{argmin}} \|u - H^{-1}\theta\|_H^2.$$

In the case of a (generalized) linear-quadratic problem

$$\begin{aligned} &\text{minimize } cx + \tfrac{1}{2}x \cdot Cx + E\{\rho_{V,Q}(T(\xi)x - h(\xi))\} \\ &\text{subject to } x \in X \end{aligned}$$

where C, Q are positive definite matrices, X, V are polyhedral sets, and $T(\xi), h(\xi)$ are random, and

$$\rho_{V,Q}x) = \sup_{v \in V} \left\{vx - \tfrac{1}{2}v \cdot Qv\right\}.$$

The theorem takes on the following form (it was the seminal result of this type).

Corollary 4.3. King and Rockafellar (1986). *Let x^* be optimal solution of the stochastic (generalized) linear-quadratic problem. Let x^ν be the solutions of the (generalized) linear-quadratic stochastic optimization problem with E replaced by E^ν and the measures P^ν are the empirical measures. Then*

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

The asymptotic distribution is the distribution of

$$\tilde{\theta} = \underset{X'}{\operatorname{argmin}}\{u \cdot \theta + \tfrac{1}{2}u \cdot Cu + E[\rho_{V'(\xi),Q}(T(\xi)u)\}$$

where

θ *is normally distributed: mean* 0, *covariance* $\operatorname{cov}(T^*(\xi)\nabla\rho_{V,Q}(T(\xi)x^* - h(\xi)))$,

$X' = \{u \in T_X(x^*) \,|\, u \cdot [c + Cx^* + E\{\nabla\rho_{V,Q}(T(\xi)x^* - h(\xi))\}] = 0\}$

$V'(\xi) = \{v \in T_V(w(\xi)) \,|\, v \cdot [T(\xi)x^* - h(\xi) - Q \cdot w(\xi]) = 0\}$

and

$$w(\xi) = \nabla\rho_{V,Q}(T(\xi)x^* - h(\xi)).$$

Appendix

We provide here a brief review of the main concepts that enter in the building of a subdifferentiability theory. It provides an entry to the literature on the subject, see Clarke (1983), Rockafellar (1983), Aubin and Ekeland (1984), Rockafellar and Wets (forthcoming).

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

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

relying on 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

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

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

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

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

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 lsc [Theorems 1 and 2, Rockafellar (1980)]. 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 $h^{\uparrow}(x;\cdot)$ makes 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 $\mathbb{R}^n$, given by

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

and

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

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,

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

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

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

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

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

For the proofs of these identities and for further details, consult Rockafellar (1981), Aubin and Ekeland (1984) and Rockafellar and Wets (forthcoming).

References

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

Artstein, Z. and S. Hart, Law of large numbers for random sets and allocation processes, *Mathematics of Operations Research*, **6** (1981), 482-492.

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

Aubin, J.-P. and I. Ekeland (1984). *Applied Nonlinear Analysis.* Wiley Interscience, New York.

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

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

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

Ermoliev, Yu.M. and V.I. Norkin (1989). Normalized Convergence in Stochastic Optimization. IIASA working paper WP-89-091. International Institute for Applied Systems Analysis, Laxenburg, Austria.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions, *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 221-233. University of California Press.

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

King, A.J. (1987). Central Limit Theory for Lipschitz Mappings. IIASA working paper WP-87-127. International Institute for Applied Systems Analysis, Laxenburg, Austria.

King, A.J. (1988). Asymptotic Distributions for Solutions in Stochastic Optimization and Generalized *M*–Estimation. IIASA working paper WP-88-58. International Institute for Applied Systems Analysis, Laxenburg, Austria.

King, A. and R.T. Rockafellar (1986). Non-normal asymptotic behaviour of solution estimates in linear-quadratic stochastic optimization, Manuscript, University of Washington.

King, A. and R.T. Rockafellar (1990). Non-normal asymptotic behavior of solution estimates in linear-quadratic stochastic optimization. IIASA working paper (forthcoming).

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

Lucchetti, R. and R.J-B Wets (1990). Minimization of integral functionals with application to optimal control and stochastic optimization. Manuscript, University of California.

Liew, Chong Kiev (1976). Inequality constrained least squares estimation. *J. Amer. Statist. Assoc.* **71** 746-751.

Rockafellar, R.T. (1976). Integral functionals, normal integrands and measurable multifunctions. In *Nonlinear Operators and the Calculus of Variations.* (J. Gossez and L. Waelbroeck, eds.) Lecture Notes in Math. **543** 157-207. Springer, Berlin.

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 Verlag, Berlin.

Rockafellar, R.T. and R. Wets (1992). *Variational Analysis*, Springer-Verlag, Berlin. (forthcoming)

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.

Shapiro, A. (1989). Asymptotic Analysis of Stochastic Programs. Research Report Navorsingsverslag 90/89(12). University of South Africa.

Vogel, S. (1988). Stability results for stochastic programming problems. *Optimization* **19** pp. 269–288.

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