NOT FOR QUOTATION
WITHOUT PERMISSION
OF THE AUTHOR

MODELING AND SOLUTION STRATEGIES FOR UNCONSTRAINED STOCHASTIC OPTIMIZATION PROBLEMS

Roger J.-B. Wets

March 1983
WP-83-36

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

MODELING AND SOLUTION STRATEGIES FOR UNCONSTRAINED STOCHASTIC OPTIMIZATION PROBLEMS*

Roger J-B. Wets*

University of Kentucky and I.I.A.S.A.
Lexington, KY, USA Laxenburg, Austria

ABSTRACT

We review some modeling alternatives for handling risk in decision making processes for unconstrained stochastic optimization problems. Solution strategies are discussed and compared.

* Invited lecture at International Institute on Stochastics and Optimization, Gargnano, Italy. September 1 - September 10, 1982.

Supported in part by a Guggenheim Fellowship and a grant of the National Science Foundation.

1. INTRODUCTION

Finding the optimal decision x* when the objective function (cost function, performance criterion,...) is available explicitly, say given by a function f(x), boils down to minimizing (or maximizing) f on R^n, i.e. x* must satisfy the relation

$$(1.1) \qquad x^* \in \operatorname{argmin} f = \{x \mid f(x) \le \inf f\} .$$

There are no real conceptual difficulties here. The only question is to find a procedure that yields x*. In fact there is a rich collection of methods available for doing exactly that, depending only on the possibility of calculating at small cost either the function values and/or its gradient, or still better second order type information.

On the other hand when optimal decisions must be reached in an environment beset with uncertainties, not only does the formulation of the decision model demand a deeper probing of the aspirations criteria in order to give to the optimization model its appropriate form, but usually significant computational obstacles must be overcome to calculate optimal decisions. Let us suppose we have a cost function given by

$$(x,\xi) \mapsto f(x,\xi): R^n \times \Xi \to R$$

where ξ are some random parameters whose real values will only be revealed after a decision x has been selected. By Ξ (usually a subset of a finite dimensional space R^s) we denote the sample space of ξ, by which is meant the set of possible values of the random variables. The function f is finite valued which implicitly implies that it has been possible to attach to each combination of x and ξ a precise cost $f(x,\xi)$. If the decision maker is indifferent to risk, then the optimal decision is reached by minimizing the function

$$x \mapsto F(x) = E[f(x,\xi)] = \int f(x,\xi)P(d\xi)$$

on R^n, where P is a given probability measure. It is assumed that $E\{f(x,\xi)\}$ is finite for all x, a harmless restriction in practice. Thus, in this case we have to find x* that satisfies

$$(1.2) \qquad x^* \in \operatorname{argmin}(F = E[f(\cdot,\xi)]) .$$

In theory every procedure developed to solve (1.1) could be employed to solve (1.2). However, the implementation of these methods demands easy access to function values and gradients, and even to Hessians. Given the limitations of multivariate calculus, these quantities can very often only be obtained numerically and if a high level of accuracy is required, the cost of calculating $F(x) = E[f(x,\xi)]$

for any fixed x may by far exceed the gain one may derive from knowing the optimal solution! It is thus imperative to develop solution methods that do rely on approximates, even on very rough estimates.

Solution strategies for solving (1.2) are studied later on. Before we do so however, let us return first to the premises that led us to (1.2). In order to accept F as the objective function of our stochastic optimization problem, we had to assume that the decision maker was indifferent to risk, i.e. his preference relationship between two risky events is totally and uniquely determined by the expectation of their cost or return. In particular x_1 is preferred to x_2 only if

$$E[f(x_1,\xi)] < E[f(x_2,\xi)] \ ;$$

for example, he equates the events:

cost \$10 with probability 1

and

cost \$1,000 with probability 1/100 .

In general decision makers do not exhibit such indifference to risk and one should take into account their attitude toward risk. Depending on the context he may be risk averse or risk seeking. This is usually dealt with through a utility function which is used to rescale the cost functional so as to make it conform to the decision maker's attitude. Let $u: R \to R$ be such a utility function. The problem becomes then: find x* that satisfies

$$x^* \in \operatorname{argmin} E[u(f(\cdot,\xi))] \ .$$

Although, for decision purposes there may be significant differences between (1.2) and this problem, as far as the development of solution procedures are concerned, these two problems are of the same type and henceforth we shall simply assume that if the utility function is anything else than linear (risk indifference) it has been incorporated in f, i.e. $f = u(\text{cost})$, which allows us to think of (1.2) as the prototype for this whole class of problems.

Sometimes it may be necessary to rely on a formulation of an objective for the stochastic optimization that does not easily fit in the framework provided by (1.2). Two examples of this type are:

find $x \in R^n$ such that $E\{f(x,\omega)\} \leq \alpha$
and variance $[f(x,\omega)]$ is minimized,

or

find $x \in R^n$ such that prob. $[f(x,\omega) \geq M]$ is minimized.

Both formulations reflect also a certain attitude towards risk of the decision maker. In the first one he is indifferent to all risky events that do not exceed an average cost of α, and among those he prefers those that deviate as little as possible from their expectations, in a least square sense. In the second problem, he exhibits a high level of risk aversion in the sense that only the tail of the distribution is of real concern to him; he seeks to avoid costs exceeding M as often as possible. Actually both objectives can be modeled through a utility function but this utility function is discontinuous. Formulating a stochastic optimization problem in this fashion might be appropriate in some very specific instances, but generally such an approach is fraught with pitfalls and should only be used when there are no alternatives available, typically only when more detailed description of the attitude towards risk is beyond our modeling capabilities. In such cases a careful analysis of the sensitivity of the solution to modeling parameters should accompany any assertion about "solutions." In the rest we shall only be concerned with problems of the type (1.2). This paper specifically deals with F finite valued; stochastic constraints present conceptual challenges that we shall not face here.

2. SOME PROPERTIES OF F

Let us refer to $F = E\{f(\cdot,\xi)\}$ as an *expectation functional.* We shall study some of its properties, when it is viewed as a map defined on R^n, i.e.

$$x \mapsto F(x) : R^n \to R .$$

We limit ourselves to those properties that may be useful when designing algorithmic procedures for solving (1.2). Let $\Xi \subset R^s$ be the support of the distribution P of the random vector , i.e. the smallest closed subset of R^p of measure 1. We think of Ξ as the set of "possible" values of ξ.

2.1. PROPOSITION. *Suppose* $x \mapsto f(x,\xi)$ *is convex for all* ξ *in* Ξ*, (or for* P*-almost all* ξ *in* Ξ*). Then* F *is convex, in which case we have*

$$\partial F(x) = \mathrm{cl} \int_{\Xi} \partial f(x,\xi) P(d\xi) \tag{2.2}$$

where ∂ *denotes the subgradient set,* cl *is closure and the* $\int$ *in* (2.2) *is to be understood as the set-valued integral of the multifunction* $\xi \mapsto \partial f(x,\xi)$.

The proof of formula (2.2) is quite technical and may be skipped without the risk of losing the continuity of the arguments that are to follow. In general the integral of a measurable multifunction is not a closed set, even if the multifunction is a subgradient multifunction; that is why closure should appear in the right-hand side of (2.2). However, the proof actually shows that the hypotheses on F, in particular F finite, imply that $\int \partial f(x,\xi)P(d\xi)$ is automatically closed. In fact it is compact.

PROOF. Convexity simply follows from integrating both sides of the inequality

$$f(x^\lambda,\xi) \le (1-\lambda)f(x^0,\xi) + \lambda f(x^1,\xi) .$$

We have assumed that F is well-defined, in fact finite on R^n, which implies, among other things, that f is finite valued on $R^n \times \Xi$, and $\xi \mapsto f(x,\xi)$ is measurable while $x \mapsto f(x,\xi)$ is continuous (convexity). All this implies that the subgradient set $\partial f(x,\xi)$ is a compact convex set given by

$$\partial f(x,\xi) = \{v \mid vx - f(x,\xi) \ge \sup_y [vy - f(y,\xi)]\} .$$

The function $\xi \mapsto g(y,\xi) = \sup_y [vy-f(y,\xi)]$ is measurable since it is the supremum of a collection of measurable functions. From this it follows that the multifunction

$$\xi \mapsto \partial f(x,\xi) = \{v \mid vx - g(v,\xi) \ge f(x,\xi)\}$$

is a compact-valued measurable multifunction. Let $v\colon \Xi \to R^n$ be any summable selection of $\partial f(x,\cdot)$. In particular we have that for every $y \in R^n$,

$$f(y,\xi) - f(x,\xi) \ge v(\xi)(y-x) ,$$

which after integration on both sides yields

$$F(y) - F(x) \ge v(y-x)$$

where $v = \int v(\xi)P(d\xi)$. This shows that

$$\partial F(x) \subset \int \partial f(x,\xi)P(\xi) .$$

(Since ∂F is closed, the inclusion would remain valid if we close the term on the right.)

To complete the proof of (2.2), we now take $\hat{v}$ in $\partial F(x)$ and show that there exists v a selection of $\partial f(x,\cdot)$ such that $\hat{v} = \int v(\xi)P(d\xi)$. To do this we must follow a circuitous route. Let us consider the following optimization problem:

find $x \in L_n^\infty = L^\infty(\Xi,P;R^n)$ such that $x(\cdot)$ is constant
and $I_f(x) = \int f(x(\xi),\xi)P(d\xi)$ is minimized.

The problem consists of minimizing the integral functional I_f over the subspace of constant functions of L_n^∞. A constant function $x_0(\cdot)$ optimal if and only if there exists an L_n^1 function $u\colon \Xi \to R^n$ such that $\int u(\xi)P(d\xi) = 0$ and

$$x_0(\xi) \in \operatorname{argmin}[f(x,\xi) - u(\xi)\cdot x \mid x \in R^n] \text{ , a.s. .}$$

This follows from Fenchel's Duality Theorem [1], we note that the "constraint qualification" is satisfied since f is finite on $R^n \times \Xi$ and summable for all $x \in R^n$. We can naturally identify the constant functions with points in R^n. Thus from what preceeds it follows that a vector x_0 minimizes F on R^n if and only if there exists a summable function $u(\cdot)$ such that

$$\int u(\xi)P(d\xi) = E[u(\xi)] = 0$$

and a.s. (almost surely)

$$x_0 \in \operatorname{argmin}[f(x,\xi) - u(\xi)x \mid x \in R^n]$$

or equivalently,

$$u(\xi) \in \partial f(x_0,\xi) \quad \text{a.s. .}$$

Now $\hat{v} \in \partial F(x)$ if and only if $0 \in \partial[F(x)-\hat{v}(x)]$, i.e. x minimizes $[F-\hat{v}\cdot]$. In view of the above this can occur if and only if there exists a summable $u(\cdot)$ with $E[u(\xi)] = 0$ such that a.s.

$$u(\xi) \in \partial[f(x,\xi) - \hat{v}x] \text{ .}$$

Let $v(\xi) = u(\xi)-\hat{v}$. The function $v(\cdot)$ is a measurable selection of $\partial f(x,\cdot)$. This completes the proof of (2.2) since it implies that to every $v \in \partial F(x)$ there corresponds a summable selection of $\partial f(x,\cdot)$ whose integral is v. □

2.3. COROLLARY. *Suppose* $x \mapsto f(x,\xi)$ *is convex for almost all* ξ. *Then* x_0 *minimizes* F *on* R^n *if and only if there exists a summable function* $v\colon \Xi \to R^n$ *such that* $E[v(\xi)] = 0$ *and a.s.*

$$v(\xi) \in \partial f(x_0,\xi)$$

or equivalently

$$0 \in \operatorname{argmin}[f(\cdot,\xi) - v(\xi)\cdot] \text{ .}$$

2.4. COROLLARY. *Suppose* $x \mapsto f(x,\xi)$ *is convex for almost all* ξ, *and either* $f(\cdot,\xi)$ *is differentiable at* $\bar{x}$ *for almost all* ξ *or* $\{\xi \mid \partial f(\bar{x},\xi) \neq \text{singleton}\}$ *has measure* 0. *Then* F *is differentiable at* $\bar{x}$. *In particular if* P *is absolutely continuous and* f *is of the form*

$$f(x,\xi) = f^\circ(L(\xi)x + \ell(\xi)) \tag{2.5}$$

where $L(\cdot)$ *is a random matrix and* $\ell(\cdot)$ *a random vector, then* F *is differentiable.*

PROOF. The first assertion follows directly from formula (2.2) and the fact that a finite convex function is differentiable whenever its subgradient set is a singleton. If f is given by (2.5), then actually we have that

$$\partial f(x,\xi) = \partial f^{\circ}(L(\xi)x + \ell(\xi)) \cdot L(\xi) .$$

The map

$$\chi \mapsto \partial f^{\circ}(\chi)$$

is not a singleton at most on a set of Lebesgue measure 0. Since P is absolutely continuous (with respect to the Lebesgue measure) it follows that for every x, $\partial f(x,\cdot)$ is not a singleton at most on a set of P-measure zero. The above then in turn implies that for all x, $\partial F(x)$ is a singleton from which follows the differentiability of the convex function F. □

Corollary 2.4 can easily be generalized. Rather than having f given by (2.5) consider the case when $f(x,\xi) = f^{\circ}(G(x,\xi))$ with $G(\cdot,\xi)$ sufficiently smooth.

3. APPROXIMATIONS AND ERROR BOUNDS

Except when f possesses separability properties, it is usually quite difficult to compute very accurately the value of F, or a subgradient of F. One must very often content oneself with a numerical scheme that approximates F(x), or any other quantity required to calculate the iterates of the minimization algorithm. Given that in order to solve (1.2), approximations are a fact of life, one is naturally led to two main strategies. The first one is to develop approximation schemes that yield upper and/or lower bounds through the careful choice of approximates for the function f or the measure P. The second one is to accept at each iterative step, approximates with statistically independent errors that, in a probabilistic sense, will cancel each other out when the number of iterations is sufficiently large. The methods described in this section rely on the first type of approximations, the subsequent section exemplifies the second strategy. We shall not consider the case when f is approximated, this is best done in a context when f is further specialized and exhibits specific structural properties.

We begin with a general approximation result.

3.1. THEOREM. *Suppose* $\{P_\nu, \nu=1,\ldots\}$ *is a sequence of probability measures converging (in distribution) weakly to a probability measure* P, *such that for* $\nu=1,\ldots$

$$F_\nu(x) = \int f(x,\xi)P_\nu(d\xi)$$

and

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

are finite for all $x \in R^n$*, where* $f\colon R^n \times O \to R$ *is convex (and thus continuous) in* x *and continuous and bounded in* ξ *on an open set* O *with* $\Xi \subset O \subset R^s$*. Then the functions* $\{F_\nu,\ \nu=1,\ldots\}$ *converge pointwise to* F*. Moreover, if there exists a bounded set* D *such that*

$$\operatorname{argmin} F_\nu \cap D \neq \emptyset$$

for all ν *sufficiently large, where*

$$\operatorname{argmin} F_\nu = \{x \in R^n \mid F_\nu(x) \leq \inf F_\nu\} ,$$

then

$$\lim_{\nu\to\infty} \inf F_\nu = \inf F$$

and the minimum of F *is attained at some point in the closure of* D.

PROOF. The pointwise convergence of the F_ν to F follows from the assumption of weak convergence (in distribution) of the P_ν to P since the function $\xi \mapsto f(x,\xi)$ is bounded and continuous for all $x \in R^n$ [2, Portemanteau Theorem]. Since the F_ν and F are finite convex functions as follows from Proposition 2.1, we have that the functions F_ν also epi-converge to F [3, Corollary 2A], this means that [4, Theorem 9]

$$\operatorname*{Lim\,sup}_{\nu\to\infty} \operatorname{argmin} F_\nu \subset \operatorname{argmin} F ,$$

i.e. if $\{x^k,\ k=1,\ldots\}$ is a sequence such that

$$x^k \in \operatorname{argmin} F_{\nu_k}$$

for some subsequence $\{\nu_k,\ k=1,\ldots\}$. The assumptions, viz. $\operatorname{argmin} F_\nu \cap D \neq \emptyset$ for some bounded D for all ν, imply that there exists a bounded sequence $\{x^\nu,\ \nu=1,\ldots\}$ with

$$x^\nu \in \operatorname{argmin} F_\nu \cap D .$$

Thus there exists a convergent subsequence $\{x^{\nu_k},\ k=1,\ldots\}$ whose limit point $\bar{x} = \lim_{k\to\infty} x^{\nu_k}$ is such that

$$\bar{x} \in \operatorname{argmin} F ,$$

and then

$$\lim_{k\to\infty} F_{\nu_k}(x^{\nu_k}) = F(\bar{x}) = \inf F .$$

There remains only to argue that the entire sequence $\{\inf F_\nu,\ \nu=1,\ldots\}$ converges to $\inf F$. But this simply follows from the observation that the preceding argument applied to any subsequence yields a further subsequence converging to $\inf F$. □

This theorem takes care of the case when the function $\omega \mapsto f(x,\omega)$ is bounded or when O is bounded, the continuity of $f(x,\cdot)$ on O implying automatically boundedness in such a case. However many applications do not satisfy such a set-up. A significant number of applications have $f(x,\cdot)$ unbounded. It is often a positively homogeneous function that tends to ∞ as $\|\omega\|$ tends to $+\infty$, i.e. it is an inf-compact function, see e.g. [5] and the examples mentioned there. And the probability measure P does not necessarily have bounded support, for example P could be a multivariate normal. In such case, some care must be given to the choice of the approximating sequence $\{P_\nu, \nu=1,\dots\}$. The situation to avoid is typified by the following example.

3.2. EXAMPLE. Let $g(t) = t^2$ and

$$\begin{aligned} P_\nu(A) &= 1-\nu^{-1} && \text{if} \quad 0 \in A \quad \text{but} \quad \nu \notin A\ , \\ &= \nu^{-1} && \text{if} \quad \nu^{-1} \in A \quad \text{but} \quad 0 \notin A\ . \end{aligned}$$

Then $\{P_\nu, \nu=1,\dots\}$ converge in distribution to P with

$$\begin{aligned} P(A) &= 1 && \text{if} \quad 0 \in A\ , \\ &= 0 && \text{otherwise}\ . \end{aligned}$$

However $\int g(t)P_\nu(dt) = \nu$ does not converge to $\int g(t)P(dt) = 0$.

3.3. THEOREM. *Suppose* $\{P_\nu, \nu=1,\dots\}$ *is a sequence of probability measures converging weakly to a probability measure* P *all defined on* $\Omega \subset R^s$, *such that for any fixed* $x \in R^n$ *and* $\varepsilon > 0$ *there exists a bounded set* $S \subset R^s$ *and* ν_ε *such that for all* $\nu \geq \nu_\varepsilon$

$$\int_{\Omega\setminus S} |f(x,\omega)|P_\nu(d\omega) < \varepsilon\ . \tag{3.4}$$

and the expectation functionals, for $\nu=1,\dots$

$$F_\nu(x) = \int f(x,\omega)P_\nu(d\omega) \qquad \textit{and} \qquad F(x) = \int f(x,\omega)P(d\omega)$$

are finite on R^n, *with the function* f *convex in* x *and continuous in* ω. *Moreover if there exists a bounded set* $D \subset R^n$ *such that*

$$\operatorname{argmin} F_\nu \cap D = \emptyset$$

for all ν *sufficiently large, then*

$$\lim_{\nu\to\infty} \inf F_\nu = \inf F$$

and the minimum of F *is attained at some point in the closure of* D.

PROOF. The arguments are the same as those used to prove Theorem 3.1, except that we can no longer deduce the pointwise convergence of the $\{F_\nu, \nu=1,\ldots\}$ to F from the weak convergence of the probability measures. For a fixed $x \in R^n$, let

$$g(\omega) = f(x,\omega) .$$

From (3.4) it follows that there exists a bounded set $S \subset R^s$ such that for all $\nu \geq \nu_\varepsilon$

$$\int_{\Omega\setminus S} |g(\omega)| P_\nu(d\omega) < \varepsilon$$

Let $M_\varepsilon = \sup_{\omega\in S} |g(\omega)|$. We know that M_ε is finite since S is bounded and g is continuous. Let g^ε be a truncation of g defined by

$$g^\varepsilon(\omega) = \begin{cases} g(\omega) & \text{if} \quad g(\omega) \leq M_\varepsilon , \\ M_\varepsilon & \text{if} \quad g(\omega) > M_\varepsilon , \\ -M_\varepsilon & \text{if} \quad g(\omega) < M_\varepsilon . \end{cases}$$

The function g^ε is bounded and continuous and we have that for all $\omega \in \Omega$

$$|g^\varepsilon(\omega)| \leq |g(\omega)|$$

and thus we have that

$$(3.5) \qquad \lim_{\nu\to\infty} \left[\beta_\nu^\varepsilon = \int_\Omega g^\varepsilon(\omega) P_\nu(d\omega)\right] = \int_\Omega g^\varepsilon(\omega) P(d\omega) = \beta^\varepsilon ,$$

and also

$$\int_{\Omega\setminus S} |g^\varepsilon(\omega)| P_\nu(d\omega) < \varepsilon ,$$

for all $\nu \geq \nu_\varepsilon$. Now let

$$\beta_\nu = F_\nu(x) = \int_S g(\omega) P_\nu(d\omega) + \int_{\Omega\setminus S} g(\omega) P_\nu(d\omega) .$$

We have that for all $\nu \geq \nu^\varepsilon$

$$|\beta_\nu - \beta_\nu^\varepsilon| = \left|\int_{\Omega\setminus S} (g(\omega) - g^\varepsilon(\omega)) P_\nu(d\omega)\right| < 2\varepsilon$$

and also that

$$|\beta - \beta^\varepsilon| < 2\varepsilon \qquad \text{where} \quad \beta = F(x) .$$

Combining the two preceding inequalities with (3.5) shows that for every $\varepsilon > 0$ we can find ν_ε such that for all $\nu \geq \nu_\varepsilon$, $|\beta_\nu - \beta| < 6\varepsilon$, i.e.

$$\lim F_\nu(x) = F(x) . \ \Box$$

The two preceding theorems essentially imply that any reasonable type of approximation scheme will yield the sought for convergence of the infima and of the solutions. The appropriate choice of the approximating sequence however can provide upper bounds and lower bounds on the solution as we show next. In [6, Section 3] we have reviewed the bounds that can be derived when the function $\omega \mapsto f(x,\omega)$ is convex (or concave). Here we study other types of approximating schemes based on stochastic ordering [7, Chapter 17].

Let $\precsim_C$ be a *partial ordering induced by* (the closed convex cone) C on R^s, i.e. we refer to a vector $t_1 \in R^s$ as *preceding* $t_2 \in R^s$ with respect to $\precsim_C$ if

$$t_2 - t_1 \in C ,$$

in which case we write

$$t_1 \precsim_C t_2 .$$

A random vector $\xi_1 \colon \Omega \to R^s$ is said to *stochastically precede* the random vector ξ_2 with respect to $\precsim_C$ if

$$P\{\omega \mid \xi_1(\omega) \precsim_C \xi_2(\omega)\} = 1 .$$

Note that this condition is stronger than the possibly more natural condition that for all $t \in R^s$

$$P[\xi_1 \precsim_C t] \leq P[\xi_2 \precsim_C t] .$$

In fact the above implies this last inequality, as can easily be verified.

We say that $\phi \colon R^s \to R$ is an *order preserving function* with respect to $\precsim_C$ if

$$t_1 \precsim_C t_2 \quad \text{implies} \quad \phi(t_1) \leq \phi(t_2) .$$

With ϕ such an order preserving function and $\xi_1(\cdot) \precsim_C \xi_2(\cdot)$ we have

$$P[\phi(\xi_1) \leq \phi(\xi_2)] = 1 ,$$

and thus

$$E\{\phi(\xi_1)\} \leq E\{\phi(\xi_2)\} .$$

This means that order relation is preserved by taking expectation (of order preserving functions).

Now returning to the problem at hand, let us suppose that for all fixed $x \in R^n$

$$\xi \mapsto f(x,\xi) \colon \Xi \subset R^s \to R$$

is order preserving with respect to a partial ordering $\precsim_C$ on R^s. Then an appropriate choice of probability distributions to approximate P is to create a sequence

$\{P_\nu, \nu=1,\ldots\}$ such that the $\{F_\nu, \nu=1,\ldots\}$ converge monotonically from above, or from below, to F. In the process, when solving the approximating problems, we obtain upper and lower bounds on the infima of F.

Again let $\Xi \subset R^s$ denote the support of the measure P. Suppose $\precsim_C$ is a partial ordering on R^s and α_ℓ is a lower bound of Ξ with respect to $\precsim_C$, by which we mean

$$\alpha_\ell \precsim_C \zeta \quad \text{for all} \quad \zeta \in \Xi \ ,$$

and let β_u be an upper bound of Ξ, i.e.

$$\zeta \precsim_C \beta_u \quad \text{for all} \quad \zeta \in \Xi \ .$$

Let us define

$$\begin{aligned} P_\ell(A) = 1 &\quad \text{if } \alpha_\ell \in A \ , \\ 0 &\quad \text{otherwise} \ , \end{aligned}$$

and

$$\begin{aligned} P_u(A) = 1 &\quad \text{if } \beta_u \in A \ , \\ 0 &\quad \text{otherwise} \ . \end{aligned}$$

Then with the above, we have

$$F_\ell(x) = \int f(x,\xi)P_\ell(d\xi) \le F(x) \le \int f(x,\xi)P_u(d\xi) = F_u(x)$$

and in particular

$$\inf F_\ell \le \inf F \le \inf F_u \ .$$

It is easy to see how to sharpen these inequalities. Suppose C is such that

$$D := \{\zeta \in R^s \mid \alpha_\ell \precsim_C \zeta \precsim_C \beta_u\}$$

is bounded. Let γ be any point in D and let us define $P_{\ell 1}$, a discrete probability measure, as follows:

$$P_{\ell 1}(\gamma) = P\{\zeta \mid \gamma \precsim_C \zeta\} \ ,$$

$$P_{\ell 1}(\alpha_\ell) = 1 - P_{\ell 1}(\gamma) \ .$$

On the other hand let

$$P_{u1}(\gamma) = P\{\zeta \mid \zeta \precsim_C \gamma\} \ ,$$

$$P_{u1}(\beta_u) = 1 - P_{u1}(\gamma) \ .$$

With

$$F_i(x) = \int f(x,\xi)P_i(d\xi) \quad \text{for} \quad i=\ell,u,\ell 1,u1 \ ,$$

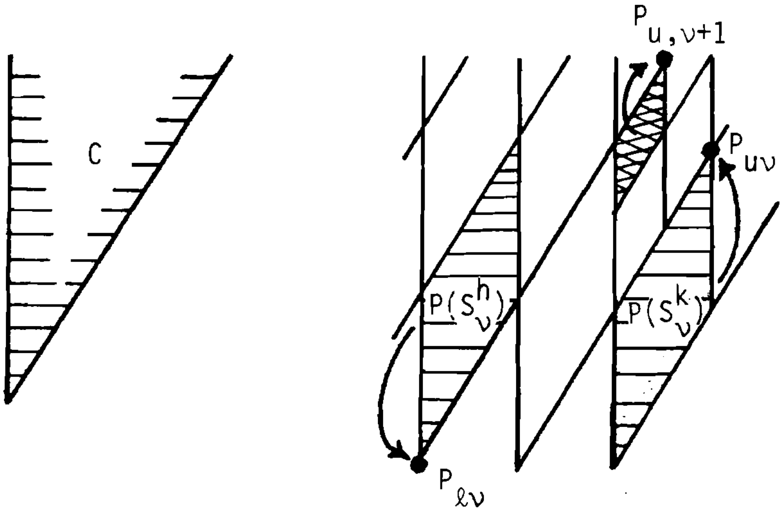
we get

$$F_\ell(x) \le F_{\ell 1}(x) \le F(x) \le F_{u1}(x) \le F_u(x)$$

and thus in particular

$$\inf F_\ell \le \inf F_{\ell 1} \le \inf F \le \inf F_{u1} \le \inf F_u \ .$$

It is now easy to see how to construct sequences of probability measures $\{P_{\ell\nu},\ \nu=1,\dots\}$ and $\{P_{u\nu},\ \nu=1,\dots\}$ that in turn yield sequences $\{F_{\ell\nu},\ \nu=1,\dots\}$ and $\{F_{\ell u},\ \nu=1,\dots\}$ converging monotonically to F. In particular, if C is a simplicial cone, then it is possible to tessellate $R^s \supset \Xi$, and assign the probability of each cell to its lower vertex (preceded by all others) to determine $F_{\ell\nu}$ and to its upper vertex (preceding all others) to determine $F_{u\nu}$. If C is not a simplicial cone, we can still proceed in this fashion, replacing $\precsim_C$ by the ordering $\precsim'$ induced by a simplicial cone C' contained in C, since $t_1 \precsim' t_2$ then implies $t_1 \precsim_C t_2$. The figure below illustrates such a construction.



3.6 Figure: Construction of $P_{\ell\nu}$, $P_{u\nu}$.

It might appear that the existence of a partial ordering $\precsim_C$ with respect to which $f(x,\cdot)$ is order preserving, is a somewhat artificial hypothesis that only will be satisfied in a very limited number of cases. To dispel this impression we consider one class of function f that is common in stochastic programming [6].

Let W,T and q be (fixed) matrices of appropriate dimensions, and define

$$f(x,\xi) = c(x) + \inf_{y\in R^p}[qy \mid Wy = \xi - Tx,\ y \ge 0] \ . \tag{3.7}$$

Without loss of generality we may assume that $q \ge 0$. If not, we know from the theory of linear programming that, unless the linear program is unbounded, there exixsts a vector π such that

$$q' = q - \pi W \ge 0 \ .$$

The original problem is then equivalent to

$$F(x) = \pi E\{\xi\} + c'(x) + E\{\inf[q'y \mid Wy = \xi - Tx,\ y \geq 0]\}$$

where

$$c'(x) = c(x) - \pi Tx\ .$$

Thus we may assume that $q \geq 0$. By pos W we denote the convex cone generated by the columns of W, i.e.

$$\text{pos } W = \left\{t \mid t = \sum_{j=1}^{n'} W^j y_j,\ y_j \geq 0\right\}\ .$$

Let $\{t^\ell \in R^{m'} \mid \ell=1,\dots,k\}$ be a subcollection of the column-vectors that determine a *frame* for pos W, i.e.

$$\text{pos } W = \left\{t = \sum_{\ell=1}^{k} \alpha_\ell t^\ell,\ \alpha_\ell \geq 0\right\}$$

and none of the t^ℓ can be obtained as a positive combination of the others.

3.8. PROPOSITION. *Suppose* f *is given through* (3.7). *Suppose moreover that for all vectors* $\{t^\ell,\ \ell=1,\dots,k\}$ *that belong to a frame of* pos W, *the function*

$$\lambda \mapsto f(x,\xi+\lambda t^\ell) : R_+ \to \bar{R}$$

is monotone decreasing. Then $\xi \mapsto f(x,\xi)$ *is order preserving with respect to the partial ordering* $\precsim$ *induced by the closed convex cone* pos W.

PROOF. It suffices to prove that if $\xi^1 \precsim \xi^2$ then

$$f(x,\xi^1) \leq f(x,\xi^2)\ .$$

Since by assumption $f(x,\xi^1)$ is finite we have that

$$f(x,\xi^1) = c(x) + qy^1$$

where

$$Wy^1 = \xi^1 - Tx\ ,\qquad y^1 \geq 0\ .$$

Since $\xi^1 \precsim \xi^2$ we have that

$$\xi^2 - Tx = Wy^1 + \sum_{\ell=1}^{k} \alpha_\ell t^\ell$$

for some $\alpha_\ell \geq 0$. Proceeding now one t^k direction at the time we get

$$f(x,\xi^1) \leq f(x,\xi^1+\alpha_1 t^1) \leq f(x,\xi^1 + \sum_{\ell=1}^{2} \alpha_\ell t^\ell) \leq \cdots \leq f(x,\xi^1 + \sum_{\ell=1}^{k} \alpha_\ell t^\ell)$$

which yields the desired inequality. □

4. A STOCHASTIC QUASI-NEWTON METHOD

Rather than first designing a careful approximation to P, that will be appropriately refined when the need arises (i.e. when the calculated error bounds exceed a certain level) we could take a "stochastic" approach to solving (1.2). By this we mean that each step is calculated by relying on a stochastic approximate of the quantities involved. More specifically, let $\xi^1, \xi^2, \ldots, \xi^k$ be a finite (unbiased) sample of the random variable ξ. Let

$$\phi(x \mid \xi^1, \xi^2, \ldots, \xi^k) = k^{-1} \sum_{\ell=1}^{k} f(x, \xi^\ell) \ .$$

This quantity can be viewed as a *stochastic approximate* of the value F(x). It has two basic properties

(i) if the size of the sample is increased, the value of $\phi(x \mid \xi^1, \ldots, \xi^k)$ tends in probability to F(x), and

(ii) if we have an algorithmic procedure that relies at each step on an independent set of samples then the errors caused by this approximation tend to cancel each other out.

The method of *stochastic quasi-gradients* [8,9], a generalization of the method of stochastic approximations, relies on (ii) combined with an appropriate choice of step-size to obtain the convergence of the method. When errors, e.g. $F(x) - \phi(x \mid \xi^1, \ldots, \xi^k)$, are viewed as noise then increasing the sample size will decrease the effect of noise in the calculations and if the steps of the algorithm are such that the effect of noise goes to 0 (in probability) with the iteration count, then property (i) is used to give us the convergence (in probability) of the method [10]. We only sketch out here a second-order method (of the Quasi-Newton type), details and further developments will appear elsewhere. We assume that the functions $f(\cdot, \xi)$ are differentiable and we rely on Proposition 2.1 for the calculation of the gradients of F.

Step 0. Select $x^1 \in R^n$, $H^1 = I_{(n\times n)}$ and set $\nu=1$. Choose a sample $(\xi^1, \ldots, \xi^{k(1)})$ and set

$$g^1 = \frac{1}{k(1)} \sum_{\ell=1}^{k(1)} \nabla f(x^1, \xi^\ell) \ ,$$

Step 1. Set

$$x^{\nu+1} = x^\nu - \rho_\nu H^\nu g^\nu \ .$$

(Here the ρ_ν are nonnegative scalars.)
Choose a sample $(\xi^1, \ldots, \xi^{k(\nu)})$ and set

$$g^{\nu+1} = \frac{1}{k(\nu)} \sum_{\ell=1}^{k(\nu)} \nabla f(x^{\nu+1}, \xi^\ell) \ .$$

Step 2. Update H: set

$$H^{\nu+1} = H^{\nu} + \frac{(v-H^{\nu}d)(v-H^{\nu}d)^T}{(v-Hd)^T \cdot d}$$

where

$$v = x^{\nu+1} - x^{\nu}$$

$$d = g^{\nu+1} - g^{\nu}$$

and ^{T} denotes transposition.

This is a rank one update of H; other updates such as that corresponding to the BFGS update for example, can also be used. There are two basic differences with the standard Quasi-Newton procedure as it would apply to F. First the gradient of F at x^{ν} is estimated rather than actually calculated, g^{ν} is only a stochastic approximate of $\nabla F(x^{\nu})$. And second, rather than performing a line search to find the minimum of $\lambda \mapsto F(x^{\nu} - \lambda H^{\nu} g^{\nu})$, we rely on a step size determined by a scalar ρ_{ν}. The basic reason being that a line search would be prohibitively expensive (if not impossible).

To obtain convergence in probability results one relies on one hand on having $k(\nu) \nearrow \infty$ as ν goes to ∞, and on choosing the step size ρ_{ν} so as to guarantee sufficiently small changes in the x^{ν}. Repetitious sampling guarantees the cancellation of estimation errors.

REFERENCES

[1] R.T. Rockafellar, An extension of Fenchel's duality theorem for convex functions, *Dual Math. J.* 33(1966), 81-90.

[2] P. Billingsley, *Convergence of Probability Measures*, J. Wiley, New York, 1975.

[3] G. Salinetti and R. Wets, On the relation between two types of convergence for convex functions, *J. Math. Anal. Appl.* 60(1977), 211-226.

[4] R. Wets, Convergence of convex functions, variational inequalities and convex optimization problems, in *Variational Inequalities and Complementarity Problems*, eds. R. Cottle, F. Gianessi and J-L. Lions, J. Wiley & Sons, New York, 1980. 375-403.

[5] K. Marti, Computation of descent direction in stochastic optimization problems with invariant distribution, Tech. Report, Hochschule Bundeswehr München, 1982.

[6] R. Wets, Stochastic programming: solution techniques and approximation schemes, in *Mathematical Programming: The State-of-the-Art 1982*, Springer-Verlag, Berlin, 1983.

[7] A. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

[8] Y. Ermoliev, Stochastic quasigradient methods and their applications to systems optimization, *Stochastics*, to appear.

[9] H. Kushner, Stochastic approximation algorithms for constrained optimization problems, *Annals of Statistics* 2(1974).

[10] B. Poljak, Nonlinear programming methods in the presence of noise, *Mathematical Programming* 14(1978), 87-97.