NUMERICAL TECHNIQUES FOR STOCHASTIC OPTIMIZATION PROBLEMS

Yuri Ermoliev
Roger J-B Wets

December 1984
PP-84-04

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
2361 Laxenburg, Austria

PREFACE

Rapid changes in today's environment emphasize the need for models and methods capable of dealing with the uncertainty inherent in virtually all systems related to economics, meteorology, demography, ecology, etc. Systems involving interactions between man, nature and technology are subject to disturbances which may be unlike anything which has been experienced in the past. In particular, the technological revolution increases uncertainty as each new stage perturbs existing knowledge of structures, limitations and constraints. At the same time, many systems are often too complex to allow for precise measurement of the parameters or the state of the system. Uncertainty, nonstationarity, disequilibrium are pervasive characteristics of most modern systems.

In order to manage such situations (or to survive in such an environment) we must develop systems which can facilitate our response to uncertainty and changing conditions. In our individual behavior we

often follow guidelines that are conditioned by the need to be prepared for all (likely) eventualities: insurance, wearing seat-belts, savings versus investments, annual medical check-ups, even keeping an umbrella at the office, etc. One can identify two major types of mechanisms: the short term *adaptive* adjustments (defensive driving, marketing, inventory control, etc.) that are made after making some observations of the system's parameters, and the long term *anticipative* actions (engineering design, policy setting, allocation of resources, investment strategies, etc.) The main challenge to the system analyst is to develop a modeling approach that combines both mechanisms (adaptive and anticipative) in the presence of a large number of uncertainties, and this in such a way that it is computationally tractable.

The technique most commonly used, *scenario analysis*, to deal with long term planning under uncertainty is seriously flawed. Although it can identify "optimal" solutions for each scenario (that specifies some values for the unknown parameters), it does not provide any clue as to how these "optimal" solutions should be combined to produce merely a reasonable decision.

As uncertainty is a broad concept, it is possible – and often useful -- to approach it in many different ways. One rather general approach, which has been successfully applied to a wide variety of problems, is to assign explicitly or implicitly, a probabilistic measure -- which can also be interpreted as a measure of confidence, possibly of subjective nature -- to the various unknown parameters. This leads us to a class of stochastic optimization problems, conceivably with only partially known distribution functions (and incomplete observations of the unknown

parameters), called *stochastic programming problems*. They can be viewed as extensions of the linear and nonlinear programming models to decision problems that involve random parameters.

Stochastic programming models were first introduced in the mid 50's by Dantzig, Beale, Tintner, and Charnes and Cooper for linear programs with random coefficients for decision making under uncertainty; Dantzig even used the name "linear programming under uncertainty". Nowadays, the term "stochastic programming" refers to the whole field – models, theoretical underpinnings, and in particular, solution procedures -- that deals with optimization problems involving random quantities (i.e., with stochastic optimization problems), the accent being placed on the computational aspects; in the USSR the term "stochastic programming" has been used to designate not only various types of stochastic optimization problems but also stochastic procedures that can be used to solve deterministic nonlinear programming problems but which play a particularly important role as solution procedures for stochastic optimization problems.

Although stochastic programming models were first formulated in the mid 50's, rather general formulations of stochastic optimization problems appeared much earlier in the literature of mathematical statistics, in particular in the theory of sequential analysis and in statistical decision theory. All statistical problems such as estimation, prediction, filtering, regression analysis, testing of statistical hypotheses, etc., contain elements of stochastic optimization; even Bayesian statistical procedures involve loss functions that must be minimized. Nevertheless, there are differences between the typical

formulation of the optimization problems that come from statistics and those from decision making under uncertainty.

Stochastic programming models are mostly motivated by problems arising in so-called "here-and-now" situations, when decisions must be made on the basis of, existing or assumed, *a priori* information about the random (relevant) quantities, without making additional observations. This situation is typical for problems of long term planning that arise in operations research and systems analysis. In mathematical statistics we are mostly dealing with "wait-and-see" situations when we are allowed to make additional observations "during" the decision making process. In addition, the accent is often on closed form solutions, or on ad hoc procedures that can be applied when there are only a few decision variables (statistical parameters that need to be estimated). In stochastic programming, which arose as an extension of linear programming, with its sophisticated computational techniques, the accent is on solving problems involving a large number of decision variables and random parameters, and consequently a much larger place is occupied by the search for efficient solutions procedures.

Unfortunately, stochastic optimization problems can very rarely be solved by using the standard algorithmic procedures developed for deterministic optimization problems. To apply these directly would presuppose the availability of efficient subroutines for evaluating the multiple integrals of rather involved (nondifferentiable) integrands that characterize the system as functions of the decision variables (objective and constraint functions), and such subroutines are neither available nor will they become available short of a small upheaval in (numerical)

mathematics. And that is why there is presently not software available which is capable of handling general stochastic optimization problems, very much for the same reason that there is no universal package for solving partial differential equations where one is also confronted by multidimensional integrations. A number of computer codes have been written to solve certain specific applications, but it is only now that we can reasonably hope to develop generally applicable software; generally applicable that is within well-defined classes of stochastic optimization problems. This means that we should be able to pass from the artisanal to the production level. There are two basic reasons for this. First maybe, the available technology (computer technology, numerically stable subroutines) has only recently reached a point where the computing capabilities match the size of the numerical problems faced in this area. Second, the underlying mathematical theory needed to justify the computational shortcuts making the solution of such problems feasible has only recently been developed to an implementable level.

The purpose of this paper is to discuss the way to deal with uncertainties in a stochastic optimization framework and to develop this theme in a general discussion of modeling alternatives and solution strategies. We shall be concerned with motivation and general conceptual questions rather than by technical details. Most everything is supposed to happen in finite dimensional Euclidean space (decision variables, values of the random elements) and we shall assume that all probabilities and expectations, possibly in an extended real-valued sense, are well defined.

CONTENTS

NUMERICAL TECHNIQUES FOR STOCHASTIC OPTIMIZATION PROBLEMS

Yuri Ermoliev and Roger J-B Wets

1. OPTIMIZATION UNDER UNCERTAINTY

Many practical problems can be formulated as optimization problems or can be reduce to them. Mathematical modeling is concerned with a description of different type of relations between the quantities involved in a given situation. Sometimes this leads to a unique solution, but more generally it identifies a set of possible states, a further criterion being used to choose among them a more, or most, desirable state. For example the "states" could be all possible structural outlays of a physical system, and the preferred state being the one that guarantees the highest level of reliability, or an "extremal" state that is chosen in terms of certain desired physical property: dielectric conductivity, sonic resonance, etc. Applications in operations research, engineering, economics have focussed attention on situations where the system can

be affected or controlled by outside decisions that should be selected in the best possible manner. To this end, the notion of an optimization problem has proved very useful. We think of it in terms of a set S whose elements, called the feasible solutions, represent the alternatives open to a decision maker. The aim is to optimize, which we take here to be to minimize, over S a certain function g_0, the objective function. The exact definition of S in a particular case depends on various circumstances, but it typically involves a number of functional relationships among the variables identifying the possible "states". As prototype for the set S we take the following description

$$S: =\{x \in R^n \mid x \in X, g_i(x) \leq 0, \quad i=1,\ldots,m\}$$

where X is a given subset of R^n (usually of rather simple character, say R^n_+ or possibly R^n itself), and for $i=1,\cdots,m$, g_i is a real-valued function on R^n. The optimization problem is then formulated as:

$$\begin{aligned} &\text{find } x \in X \subset R^n \text{ such that} \\ &g_i(x) \leq 0, \qquad i=1,\cdots,m, \\ &\text{and } z = g_0(x) \text{ is minimized.} \end{aligned} \tag{1.1}$$

When dealing with conventional deterministic optimization problems (linear or nonlinear programs), it is assumed that one has precise information about the objective function g_0 and the constraints g_i. In other words, one knows all the relevant quantities that are necessary for having well-defined functions g_i, $i=1,\cdots,m$. For example, if this is a production model, enough information is available about future demands and prices, available inputs and the coefficients of the input-output relationships, in order to define the cost function g_0 as well as give a

sufficiently accurate description of the balance equations, i.e., the functions g_i, $i=1,\cdots,m$. In practice, however, for many optimization problems the functions g_i, $i=0,\cdots m$ are not known very accurately and in those cases, it is fruitful to think of the functions g_i as depending on a pair of variables (x,ω) with ω as vector that takes its values in a set $\Omega \subset R^q$. We may think of ω as the *environment*-determining variable that conditions the system under investigation. A decision x results in different outcomes

$$\left[g_0(x,\omega), g_1(x,\omega), \cdots, g_m(x,\omega)\right]$$

depending on the uncontrollable factors, i.e. the environment (state of nature, parameters, exogenous factors, etc.). In this setting, we face the following "optimization" problem:

$$\begin{aligned} &\text{find } x \in X \subset R^n \text{ such that} \\ &g_i(x,\omega) \leq 0, \qquad i=1,\cdots,m, \\ &\text{and } z(\omega) = g_0(x,\omega) \text{ is minimized.} \end{aligned} \tag{1.2}$$

This may suggest a parametric study of the optimal solution as a function of the environment ω and this may actually be may be useful in some cases, but what we really seek is some x that is "feasible" and that minimizes the objective for all or for nearly all possible values of ω in Ω, or is some other sense that needs to be specified. Any fixed $x \in X$, may be feasible for some $\omega' \in \Omega$, i.e. satisfy the constraints $g_i(x,\omega') \leq 0$ for $i=1,\cdots,m$, but infeasible for some other $\omega \in \Omega$. The notion of feasibility needs to be made precise, and depends very much on the problem at hand, in particular whether or not we are able to obtain some information about the environment, the value of ω, before choosing the decision

x. Similarly, what must be understood by optimality depends on the uncertainties involved as well as on the view one may have of the overall objective(s), e.g. avoid a disastrous situation, do well in nearly all cases, etc. We cannot "solve" (1.2) by finding the optimal solution for every possible value of ω in Ω, i.e. for every possible environment, aided possibly in this by parametric analysis. This is the approach preconized by *scenario analysis*. If the problem is not insensitive to its environment, then knowing that $x^1 = x^*(\omega^1)$ is the best decision in environment ω^1 and $x^2 = x^*(\omega^2)$ is the best decision in environment ω^2 does not really tell us how to choose some x that will be a reasonably good decision whatever be the environment ω^1 or ω^2; taking a (convex) combination of x^1 and x^2 may lead to an infeasible decision for both possibilities: problem (1.2) with $\omega = \omega^1$ or $\omega = \omega^2$.

In the simplest case of complete information, i.e. when the environment ω will be completely known before we have to choose x, we should, of course, simply select the optimal solution of (1.2) by assigning to the variables ω the known values of these parameters. However, there may be some additional restrictions on this choice of x in certain practical situations. For example, if the problem is highly nonlinear or/and quite large, the search for an optimal solution may be impractical (too expensive, for example) or even physically impossible in the available time, the required response-time being too short. Then, even in this case, there arises -- in addition to all the usual questions of optimality, design of solutions procedures, convergence, etc. -- the question of implementability. Namely, how to design a practical (implementable) decision rule (function)

$$\omega \mapsto x(\omega)$$

which is viable, i.e. $x(\omega)$ is feasible for (1.2) for all $\omega \in \Omega$, and that is "optimal" in some sense, ideally such that for all $\omega \in \Omega$, $x(\omega)$ minimizes $g_0(\cdot,\omega)$ on the corresponding set of feasible solutions. However, since such an ideal decision rule is only rarely simple enough to be implementable, the notion of optimality must be redefined so as to make the search for such a decision rule meaningful.

A more typical case is when each observation (information gathering) will only yield a partial description of the environment ω : it only identifies a particular collection of possible environments, or a particular probability distribution on Ω. In such situations, when the value of ω is not known in advance, for any choice of x the values assumed by the functions $g_i(x,\cdot)$, $i=1,\cdots,m$, cannot be known with certainty. Returning to the production model mentioned earlier, as long as there is uncertainty about the demand for the coming month, then for any fixed production level x, there will be uncertainty about the cost (or profit). Suppose, we have the very simple relation between x (production level) and ω (demand):

$$g_0(x,\omega) = \begin{cases} \alpha(x-\omega) & \text{if} \quad \omega \geq x \\ \beta(\omega-x) & \text{if} \quad x \geq \omega \end{cases} \tag{1.3}$$

where α is the unit surplus-cost (holding cost) and β is the unit shortage-cost. The problem would be to find an x that is "optimal" for all foreseeable demands ω in Ω rather than a function $\omega \mapsto x(\omega)$ which would tell us what the optimal production level should have been once ω is actually observed.

When no information is available about the environment ω, except that $\omega \in \Omega$ (or to some subset of Ω), it is possible to analyze problem (1.2) in terms of the values assumed by the vector

$$\left[g_0(x,\omega), g_1(x,\omega),\ldots,g_m(x,\omega)\right]$$

as ω varies in Ω. Let us consider the case when the functions $g_1, \cdots, g_m$ do not depend on ω. Then we could view (1.2) as a multiple objective optimization problem. Indeed, we could formulate (1.2) as follows:

$$\begin{aligned} &\text{find } x \in X \subset R^n \text{ such that} \\ &g_i(x) \leq 0, \quad i=1, \cdots, m \\ &\text{and for each } \omega \in \Omega,\ z_\omega = g_0(x,\omega) \text{ is minimized.} \end{aligned} \tag{1.4}$$

At least if Ω is a finite set, we may hope that this approach would provide us with the appropriate concepts of feasibility and optimality. But, in fact such a reformulation does not help much. The most commonly accepted point of view of optimality in multiple objective optimization is that of Pareto-optimality, i.e. the solution is such that any change would mean a strictly less desirable state in terms of at least one of the objectives, here for some ω in Ω. Typically, of course, there will be many Pareto-optimal points with no equivalence between any such solutions. There still remains the question of how to choose a (unique) decision among the Pareto-optimal points. For instance, in the case of the objective function defined by (1.3), with $\Omega = [\underline{\omega},\bar{\omega}] \subset (0,\infty)$ and $\alpha > 0$, $\beta > 0$, each $x = \omega$ is Pareto-optimal, see Figure 1,

$$\begin{aligned} &g_0(x,\omega) = g_0(\omega,\omega) = 0 \\ &g_0(\omega,\omega') > 0 \qquad \text{for all} \quad \omega' \neq \omega\,. \end{aligned}$$

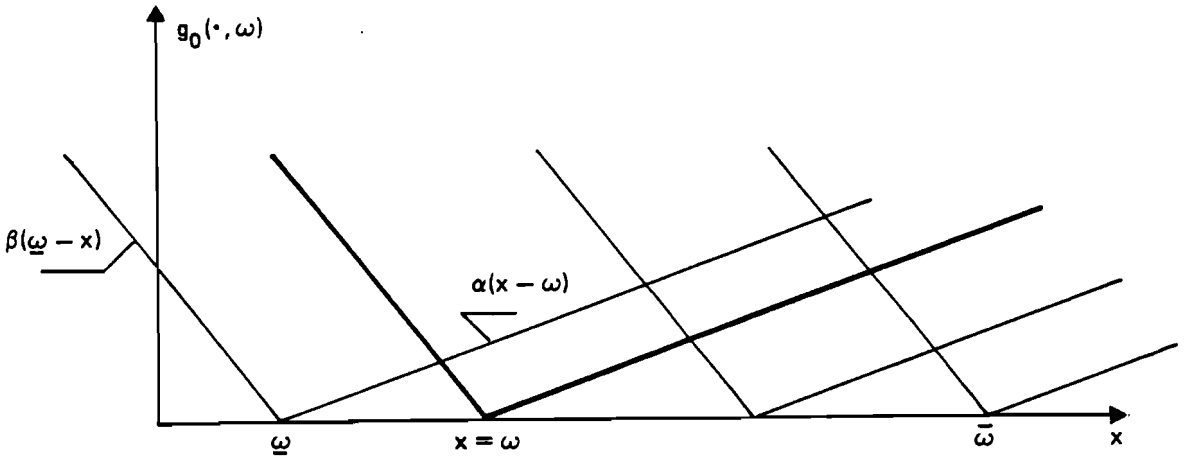

Figure 1. Pareto-optimality

One popular approach to selecting among the Pareto-optimal solutions is to proceed by "worst-case analysis". For a given x, one calculates the worst that could happen -- in terms of all the objectives -- and then choose a solution that minimizes the value of the worst-case loss; scenario analysis also relies on a similar approach. This should single out some point that is optimal in a pessimistic minimax sense. In the case of the example (1.3), it yields $x^* = \bar{\omega}$ which suggests a production level sufficiently high to meet every foreseeable demand. This may turn out to be a quite expensive solution in the long run!

2. STOCHASTIC OPTIMIZATION: ANTICIPATIVE MODELS

The formulation of problem (1.2) as a stochastic optimization problem presuppose that in addition to the knowledge of Ω, one can rank the future alternative environments ω according to their comparative fre-

quency of occurrence. In other words, it corresponds to the case when weights -- *an a priori probability measure, objective or subjective* -- can be assigned to all possible $\omega \in \Omega$, and this is done in a way that is consistent with the calculus rules for probabilities. Every possible environment ω becomes an element of a probability space, and the meaning to assign to feasibility and optimality in (1.2) can be arrived at by reasonings or statements of a probabilistic nature. Let us consider the here-and-now situation, when a solution must be chosen that does not depend on future observations of the environment. In terms of problem (1.2) it may be some $x \in X$ that satisfies the constraints

$$g_i(x,\omega) \leq 0, \qquad i=1,\cdots,m, \tag{1.2}$$

with a certain level of reliability:

$$\text{prob. } \{\omega \mid g_i(x,\omega) \leq 0, \quad i=1,\cdots,m\} \geq \alpha \tag{2.1}$$

where $\alpha \in (0,1)$, not excluding the possibility $\alpha = 1$, or in the average:

$$E\{g_i(x,\omega)\} \leq 0, \qquad i=1\cdots,m. \tag{2.2}$$

There are many other possible probabilistic definitions of feasibility involving not only the mean but also the variance of the random variable $g_i(x,\bullet)$.

$$\text{Var } g_i(x,\bullet) := E\left[g_i(x,\omega) - E\{g_i(x,\omega)\}\right]^2,$$

such as

$$E\{g_i(x,\omega)\} + \beta\left[\text{Var } g_i(x,\bullet)\right]^{1/2} \leq 0 \tag{2.3}$$

for β some positive constant, or even higher moments or other nonlinear functions of the $g_i(x,\bullet)$ may be involved. The same possibilities are avail-

able in definiting optimality. Optimality could be expressed in terms of the (feasible) x that minimizes

$$\text{prob. } \{\omega \mid g_0(x,\omega) \geq \alpha_0\} \tag{2.4}$$

for a prescribed level α_0, or the expected value of future cost

$$E\{g_0(x,\omega)\}\,, \tag{2.5}$$

and so on.

Despite the wide variety of concrete formulations of stochastic optimization problems, generated by problems of the type (1.2) all of them may finally be reduced to the following rather general version given below, and for conceptual and theoretical purposes it is useful to study stochastic optimization problems in those general terms: Given a probability space (Ω,A,P), that gives us a description of the possible environments Ω with associated probability measure P, a ***stochastic programming problem*** is:

$$\begin{aligned} &\text{find } x \in X \subset R^n \text{ such that} \\ &F_i(x) = E\{f_i(x,\omega)\} = \int f_i(x,\omega)\, P(d\omega) \leq 0, \quad \text{for } i=1,\cdots,m, \\ &\text{and } z = F_0(x) = E\{f_0(x,\omega)\} = \int f_0(x,\omega)\, P(d\omega) \text{ is minimized,} \end{aligned} \tag{2.6}$$

where X is a (usually closed) fixed subset of R^n, and the functions

$$f_i\colon\ R^n \times \Omega \rightarrow R\,, \qquad i=1,\cdots,m,$$

and

$$f_0\colon\ R^n \times \Omega \rightarrow \bar{R} := R \cup \{-\infty, +\infty\},$$

are such that, at least for every x in X, the expectations that appear in (2.6) are well-defined.

For example, the constraints (2.1) that are called *probabilistic or chance constraints*, will be of the above type if we set:

$$f_i(x,\omega) = \begin{cases} \alpha - 1 & \text{if } g_l(x,\omega) \leq 0 \quad \text{for } l=1,\ldots,m, \\ \alpha & \text{otherwise} \end{cases} \tag{2.7}$$

The variance, which appears in (2.3) and other moments, are also mathematical expectations of some nonlinear functions of the $g_i(x,\cdot)$.

How one actually passes from (1.2) to (2.6) depends very much on the concrete situation at hand. For example, the criterion (2.4) and the constraints (2.1) are obtained if one classifies the possible outcomes

$$g_0(x,\omega), g_1(x,\omega), \cdots, g_m(x,\omega),$$

as ω varies on Ω, into "bad" and "good" (or acceptable and nonacceptable). To minimize (2.4) is equivalent to minimizing the probability of a "bad" event. The choice of the level α as it appears in (2.1), is a problem in itself, unless such a constraint is introduced to satisfy contractually specified reliability levels. The natural tendency is to choose the reliability level α as high as possible, but this may result in a rapid increase in the overall cost. Figure 2 illustrates a typical situation where increasing the reliability level beyond a certain level $\bar{\alpha}$ may result in enormous additional costs.

To analyze how high one should go in the setting of reliability levels, one should, ideally, introduce the loss that would be incurred if the constraints were violated, to be balanced against the value of the objective function. Suppose the objective function is of type (2.5), and in the simple case when violating the constraint $g_i(x,\omega) \leq 0$, it generates a cost

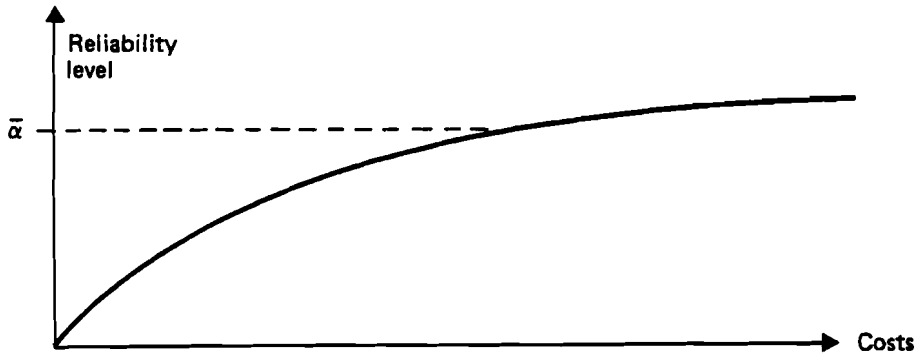

Figure 2. Reliability versus cost.

$(q_i \geq 0)$:

$$q_i \cdot g_i(x,\omega),$$

proportional to the amount by which we violate the constraint, we are led to the objective function:

$$f_0(x,\omega) = g_0(x,\omega) + \sum_{i=1}^{m} q_i \Big[\max[0, g_i(x,\omega)]\Big], \tag{2.8}$$

for the stochastic optimization problem (2.6). For the production (inventory) model with cost function given by (1.3), it would be natural to minimize the expected loss function

$$F_0(x) = \alpha \int_{\omega \leq x} (x-\omega) P(d\omega) + \beta \int_{x \leq \omega} (\omega - x)\, P(d\omega) = E\{g_0(x,\omega)\}$$

which we can also write as

$$F_0(x) = E\left\{\max\left[\alpha(x-\omega),\ \beta(\omega-x)\right]\right\}. \tag{2.9}$$

A more general class of problems of this latter type comes with the objective function:

$$F_0(x) = E\left\{\max_{y\in Y} \rho(x,y,\omega)\right\} \tag{2.10}$$

where $Y \subset R^P$. Such a problem can be viewed as a model for decision making under uncertainty, where the x are the decision variables themselves, the ω variables correspond to the states of nature with given probability measure P, and the y variables are there to take into account the worst case.

3. ABOUT SOLUTION PROCEDURES

In the design of solution procedures for stochastic optimization problems of type (2.6), one must come to grips with two major difficulties that are usually brushed aside in the design of solution procedures for the more conventional nonlinear optimization problems (1.1): in general, the exact evaluation of the functions F_i, $i=1,\cdots,m$, (or of their gradients, etc.) is out of question, and moreover, these functions are quite often nondifferentiable. In principle, any nonlinear programming technique developed for solving problems of type (1.1) could used for solving stochastic optimization problems. Problems of type (2.6) are after all just special case of (1.1), and this does also work well in practice if it is possible to obtain explicit expressions for the functions F_i, $i=1,\cdots,m$, through the analytical evaluation of the corresponding integrals

$$F_i(x) = E\{f_i(x,\omega)\} = \int f_i(x,\omega)\, P(d\omega) .$$

Unfortunately, the exact evaluation of these integrals, either analytically or numerically by relying on existing software for quadratures, is only possible in exceptional cases, for every special types of probability measures P and integrands $f_i(x,\cdot)$. For example, to calculate the values of the constraint function (2.1) even for $m = 1$, and

$$g_1(x,\omega) = h(\omega) - \sum_{j=1}^{n} t_j(\omega)\, x_j \tag{3.1}$$

with random parameters $h(\cdot)$ and $t_j(\cdot)$, it is necessary to find the probability of the event

$$\left\{\omega \mid \sum_{j=1}^{n} t_j(\omega)\, x_j \geq h(\omega)\right\}$$

as a function of $x = (x_1, \cdots, x_n)$. Finding an analytical expression for this function is only possible in a few rare cases, the distribution of the random variable

$$\omega \mapsto h(\omega) - \sum_{j=1}^{n} t_j(\omega) x_j$$

may depend dramatically on x; compare $x = (0, \cdots 0)$ and $x = (1,\ldots,1)$.

Of course, the exact evaluation of the functions F_i is certainly not possible if only partial information is available about P, or if information will only become available while the problem is being solved, as is the case in optimization systems in which the values of the outputs $\{f_i(x,\omega),\ i=0, \cdots, m\}$ are obtained through actual measurements or Monte Carlo simulations.

In order to bypass some of the numerical difficulties encountered with multiples integrals in the stochastic optimization problem (2.6),

one may be tempted to solve a substitute problem obtained from (1.2) by replacing the parameters by their expected values, i.e. in (2.6) we replace

$$E\{f_i(x,\omega)\} \quad \text{by} \quad f_i(x,\bar{\omega}),$$

where $\bar{\omega} = E\{\omega\}$. This is relatively often done in practice, sometimes the optimal solution might only be slightly affected by such a crude approximation, but unfortunately, this supposedly harmless simplification, may suggest decisions that not only are far from being optimal, but may even "validate" a course of action that is contrary to the best interests of the decision maker. As a simple example of the errors that may derive from such a substitution let us consider:

$$f_0(x,\omega) = (\omega x)^2, \; x \in R, \; P[\omega = +1] = P[\omega = -1] = \tfrac{1}{2},$$

then

$$f_0(x,\bar{\omega}) \equiv 0, \quad \text{but} \quad E\{f_0(x,\omega)\} = x^2.$$

Not having access to precise evaluation of the function values, or the gradients of the F_i, $i=0,\cdots,m$, is the main obstacle to be overcome in the design of algorithmic procedures for stochastic optimization problems. Another peculiarity of this type of problems is that the functions

$$x \mapsto F_i(x), \quad i=0,\cdots,m,$$

are quite often nondifferentiable -- see for example (2.1), (2.3), (2.4), (2.9) an (2.10) -- they may even be discontinuous as indicated by the simple example in Figure 3.

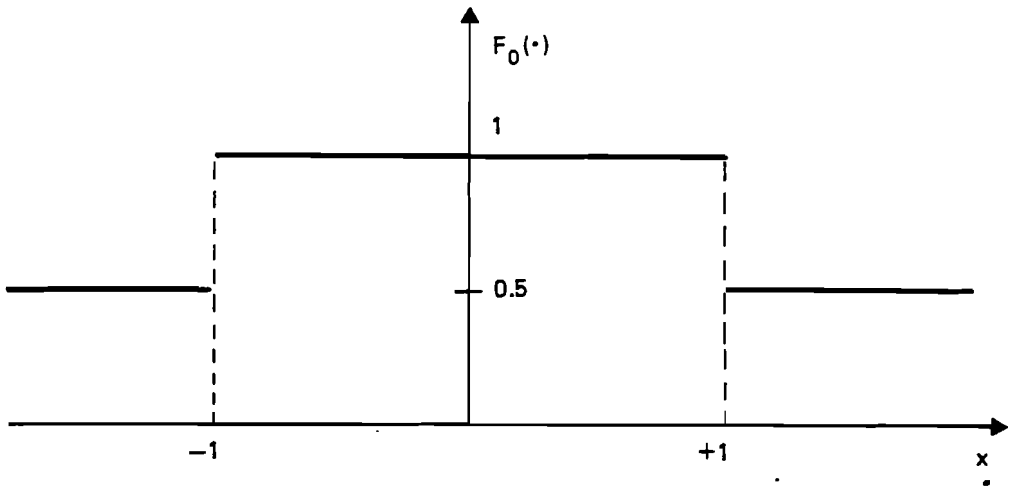

Figure 3. $F_0(x) = P\{\omega \mid \omega x \leq 1\}$, $P[\omega = +1] = P[\omega = -1] = \frac{1}{2}$.

The stochastic version of even the simplest linear problem may lead to nondifferential problem as vividly demonstrated by Figure 3. It is now easy to imagine how complicated similar functions defined by linear inequalities in R^n might become. As another example of this type, let us consider a constraint of the type (1.2), i.e. a probabilistic constraint, where the $g_i(\cdot,\omega)$ are linear, and involve only one 1-dimensional random variable $h(\cdot)$. The set S of feasible solutions are those x that satisfy

$$P\{\omega \mid x+3 \geq h(\omega),\ x \leq h(\omega)\} \geq 2/3\,,$$

where $h(\cdot)$ is equal to 0,2, or 4 each with probability 1/3. Then

$$S = [-1,0] \cup [1,2]$$

is disconnected.

The situation is not always that hopeless, in fact for well-formulated stochastic optimization problem, we may expect a lot of regularity, such as convexity of the feasibility region, convexity and/or Lipschitz properties of the objective function, and so on. This is well documented in the literature.

In the next two sections, we introduce some of the most important formulations of stochastic programming problems and show that for the development of conceptual algorithms, problem (2.6) may serve as a guide, in that the difficulties to be encountered in solving very specific problems are of the same nature as those one would have when dealing with the quite general model (2.6).

4. STOCHASTIC OPTIMIZATION: ADAPTIVE MODELS

In the stochastic optimization model (2.6), the decision x has to be chosen by using an *a priori* probabilistic measure P without having the opportunity of making additional observations. As discussed already earlier, this corresponds to the idea of an optimization model as a tool for planning for possible future environments, that is why we used the term: anticipative optimization. Consider now the situation when we are allowed to make an observation before choosing x, this now corresponds to the idea of optimization in a learning environment, let us call it *adaptive optimization.*

Typically, observations will only give a partial description of the environment ω. Suppose B contains all the relevant information that could become available after making an observation; we think of B as a subset of A. The decision x must be determined on the basis of the

information available in B, i.e. it must be a function of ω that is "B-measurable". The statement of the corresponding optimization is similar to (2.6), except that now we allow a larger class of solutions -- the B-measurable functions -- instead of just points in R^n (which in this setting would just correspond to the constant functions on Ω). The problem is to find a B-measurable function

$$\omega \mapsto x(\omega)$$

that satisfies: $x(\omega) \in X$ for all ω,

$$E\left\{f_i(x(\omega),\omega) \mid B\right\}(\omega) \leq , \quad i=1,\cdots,m,$$

and

$$z = E\left\{f_0(x(\omega),\omega)\right\} \text{ is minimized.} \tag{4.1}$$

where $E\{\cdot \mid B\}$ denotes the *conditional expectation* given B. Since x is to be a B-measurable function, the search for the optimal x, can be reduced to finding for each $\omega \in \Omega$ the solution of

$$\begin{aligned} &\text{find } x \in X \subset R^n \text{ such that} \\ &E\{f_i(x,\cdot) \mid B\}(\omega) \leq 0, \quad i=1,\ldots,m \\ &\text{and } z_\omega = E\{f_0(x,\cdot) \mid B\}(\omega) \text{ is minimized.} \end{aligned} \tag{4.2}$$

Each problem of this type has exactly the same features as problem (2.6) except that expectation has been replaced by conditional expectation; note that problem (4.1) will be the same for all ω that belong to the same elementary event of B. In the case when ω becomes completely known, i.e. when $B = A$, then the optimal $\omega \mapsto x(\omega)$ is obtained by solving for all

ω, the optimization problem:

$$\begin{aligned} &\text{find } x \in X \subset R^n \text{ such that} \\ &f_i(x,\omega) \leq 0, \quad i=1,\cdots,m, \\ &\text{and } z_\omega = f_0(x,\omega) \text{ is minimized,} \end{aligned} \tag{4.3}$$

i.e. we need to make a parametric analysis of the optimal solution as a function of ω.

If the optimal decision rule $\omega \mapsto x^*(\omega)$ obtained by solving (4.1), is implementable in a real-life setting it may be important to know the distribution function of the optimal value

$$\omega \mapsto E\{f_0(x^*(\cdot),\cdot) \mid B\}(\omega)$$

This is known as the *distribution problem* for random mathematical programs which has received a lot of attention in the literature, in particularly in the case when the functions f_i, $i=0,\cdots,m$, are linear and $B = A$.

Unfortunately in general, the decision rule $x^*(\cdot)$ obtained by solving (4.2), and in particular (4.3), is much too complicate for practical use. For example, in our production model with uncertain demand, the resulting output may lead to highly irregular transportation requirements, etc. In inventory control, one has recourse to "simple", (s,S)-policies in order to avoid the possible chaotic behavior of more "optimal" procedures; an (s,S)-policy is one in which an order is placed as soon as the stock falls below a buffer level s and the quantity ordered will restore to a level S the stock available. In this case, we are restricted to a specific family of decision rules, defined by two parameters s and S which have to be defined before any observation is made.

More generally, we very often require the decision rules $\omega \mapsto x(\omega)$ to belong to prescribed family

$$\{x(\lambda,\cdot),\ \lambda \in \Lambda\}$$

of decision rules parametrized by a vector λ, and it is this λ that must be chosen here-and-now before any observations are made. Assuming that the members of this family are B-measurable, and substituting $x(\lambda,\cdot)$ in (4.1), we are led to the following optimization problem

$$\begin{aligned} &\text{find } \lambda \in \Lambda \text{ such that} \\ &x(\lambda,\omega) \in X \quad \text{for all} \quad \omega \in \Omega \\ &H_i(\lambda) = E\left\{f_i(x(\lambda,\omega),\omega)\right\} \leq 0, \quad i=1,\cdots,m \\ &\text{and } H_0(\lambda) = E\left\{f_0(x(\lambda,\omega),\omega)\right\} \text{ is minimized.} \end{aligned} \tag{4.4}$$

This again is a problem of type (2.6), except that now the minimization is with respect to λ. Therefore, by introducing the family of decision rules $\{x(\lambda,\cdot),\ \lambda \in \Lambda\}$ we have reduced the problem of adaptive optimization to a problem of anticipatory optimization, no observations are made before fixing the values of the parameters λ.

It should be noticed that the family $\{x(\lambda,\cdot),\ \lambda \in \Lambda\}$ may be given implicitly. To illustrate this let us consider a problem studied by Tintner. We start with the linear programming problem (4.5), a version of (1.2):

$$\begin{aligned} &\text{find } x \in R_+^n \text{ such that} \\ &\textstyle\sum_{j=1}^n a_{ij}(\omega)x_j \geq b_i(\omega), \quad i=1,\cdots,m \\ &\text{and } z = \textstyle\sum_{j=1}^n c_j(\omega)\, x_j \text{ is minimized,} \end{aligned} \tag{4.5}$$

where the $a_{ij}(\cdot), b_i(\cdot)$ and $c_j(\cdot)$ are positive random variables. Consider the family of decision rules: let λ_{ij} be the portion of the i-th resource to

be assigned to activity j, thus

$$\sum_{j=1}^{n} \lambda_{ij} = 1, \ \lambda_{ij} \geq 0 \quad \text{fo} \quad i=1, \cdots, m; \ j=1, \cdots, n, \tag{4.6}$$

and for j=1, ... n ,

$$x_j(\lambda,\omega) \in \operatorname{argmin}_{x \in R_+} \left\{ c_j(\omega)x \mid a_{ij}(\omega)x \geq \lambda_{ij} b_i(\omega), \ i=1, \cdots, m \right\}$$

i.e.

$$x_j(\lambda,\omega) = \max_{1 \leq i \leq m} \lambda_{ij} \ b_i(\omega) / a_{ij}(\omega) \ .$$

This decision rule is only as good as the λ_{ij} that determine it. The optimal λ's are found by minimizing

$$\sum_{j=1}^{n} E \left\{ c_j(\omega) \max_{1 \leq i \leq m} (\lambda_{ij} \ b_j(\omega) / a_{ij}(\omega)) \right\} \tag{4.7}$$

subject to (4.6), again a problem of type (2.6).

5. ANTICIPATION AND ADAPTATION: RECOURSE MODELS

The (two-stage) recourse problem can be viewed as an attempt to incorporate both fundamental mechanisms of anticipation and adaptation within a single mathematical model. In other words, this model reflects a trade-off between long-term anticipatory strategies and the associated short-term adaptive adjustments. For example, there might be a trade-off between a road investment's program and the running costs for the transportation fleet, investments in facilities location and the profit from its day-to-day operation. The linear version of the

recourse problem is formulated as follows:

$$\begin{aligned} &\text{find } x \in R^n_+ \text{ such that} \\ &F_i(x) = b_i - A_i x \leq 0\,, \quad i=1,\cdots,m\,, \\ &\text{and } F_0(x) = c\,x + E\{Q(x,\omega)\} \text{ is minimized} \end{aligned} \tag{5.1}$$

where

$$Q(x,\omega) = \inf_{y \in R^{n'}_+} \{q(\omega)y \mid W(\omega)y = h(\omega) - T(\omega)x\}\ ; \tag{5.2}$$

some or all of the coefficients of matrices and vectors $q(\cdot)$, $W(\cdot)$, $h(\cdot)$ and $T(\cdot)$ may be random variables. In this problem, the long-term decision is made before any observation of $\omega \sim \left[q(\omega), W(\omega), h(\omega), T(\omega)\right]$. After the true environment is observed, the discrepancies that may exist between $h(\omega)$ and $T(\omega)x$ (for fixed x and observed $h(\omega)$ and $T(\omega)$) are corrected by choosing a recourse action y, so that

$$W(\omega)y = h(\omega) - T(\omega)x\,, \quad y \geq 0\,, \tag{5.3}$$

that minimizes the loss

$$q(\omega)y\ .$$

Therefore, an optimal decision x should minimize the total cost of carrying out the overall plan: direct costs as well as the costs generated by the need of taking correct (adaptive) action.

A more general model is formulated as follows. A long-term decision x must be made before the observation of ω is available. For given $x \in X$ and observed ω, the recourse (feedback) action $y(x,\omega)$ is chosen so as to solve the problem

$$\begin{aligned} &\text{find } y \in Y \subset R^{n'} \text{ such that} \\ &f_{2i}(x,y,\omega) \leq 0\,, \quad i=1,\cdots,m'\,, \\ &\text{and } z_2 = f_{20}(x,y,\omega) \text{ is minimized}\,, \end{aligned} \tag{5.4}$$

assuming that for each $x \in X$ and $\omega \in \Omega$ the set of feasible solutions of this problem is nonempty (in technical terms, this is known as relatively complete recourse). Then to find the optimal x, one would solve a problem of the type:

$$\begin{aligned} &\text{find } x \in X \subset R^n, \text{ such that} \\ &F_0(x) = E\{f_{20}(x, y(x,\omega), \omega)\} \text{ is minimized.} \end{aligned} \tag{5.5}$$

If the state of the environment ω remains unknown or partially unknown after observation, then

$$\omega \mapsto y(x,\omega)$$

is defined as the solution of an adaptive model of the type discussed in Section 4. Give B the field of possible observations, the problem to be solved for finding $y(x,\omega)$ becomes: for each $\omega \in \Omega$

$$\begin{aligned} &\text{find } y \in Y \subset R^{n'} \quad \text{such that} \\ &E\{f_{2i}(x,y,\cdot) \mid B\}(\omega) \leq 0, \quad i=1,\cdots,m' \\ &\text{and } z_{2\omega} = E\{f_{20}(x,y,\cdot) \mid B\}\ (\omega) \text{ is minimized} \end{aligned} \tag{5.6}$$

If $\omega \mapsto y(x,\omega)$ yields the optimal solution of this collection of problems, then to find an optimal x we again have to solve a problem of type (5.5).

Let us notice that if

$$f_{20}(x,y,\omega) = cx + q(\omega)y$$

and for $i=1,\cdots,m'$,

$$f_{2i}(x,y,\omega) = \begin{cases} 1-\alpha & \text{if } T_i(\omega)x + W_i(\omega)y - h_i(\omega) \geq 0, \\ \alpha & \text{otherwise} \end{cases}$$

then (5.5), with the second stage problem as defined by (5.6), corresponds to the statement of the recourse problem in terms of conditional probabilistic (chance) constraints.

There are many variants of the basic recourse models (5.1) and (5.5). There may be in addition to the deterministic constraints on x some expectation constraints such as (2.3), or the recourse decision rule may be subject to various restrictions such as discussed in Section 4, etc. In any case as is clear from the formulation, these problems are of the general type (2.6), albeit with a rather complicated function $f_0(x,\omega)$.

6. DYNAMIC ASPECTS: MULTISTAGE RECOURSE PROBLEMS

It should be emphasized that the "stages" of a two-stage recourse problem do not necessarily refer to time units. They correspond to steps in the decision process, x may be a here-and-now decision whereas the y correspond to *all* future actions to be taken in different time period in response to the environment created by the chosen x and the observed ω in that specific time period. In another instance, the x,y solutions may represent sequences of control actions over a given time horizon,

$$x = (x(0), x(1), \cdots, x(T)),$$
$$y = (y(0), y(1), \cdots, y(T)),$$

the y-decisions being used to correct for the basic trend set by the x-control variables. As a special case we have

$$x = (x(0), x(1), \cdots, x(s)),$$
$$y = (y(s+1), \cdots, y(T)),$$

that corresponds to a mid-course maneuver at time s when some observations have become available to the controller. We speak of *two-stage dynamic* models. In what follows, we discuss in more detail the possible statements of such problems.

In the case of dynamical systems, in addition to the x,y solutions of problems (5.5)-(5.4), there may also be an additional group of variables

$$z = \left[z(0), z(1), \cdots, z(T)\right]$$

that record the ***state of the system*** at times $0,1,\cdots,T$. Usually, the variables x,y,z,ω are connected through a (differential) system of equations of the type:

$$\Delta z(t) = h\left[t, z(t), x(t), y(t), \omega\right], \quad t=0,\cdots,T-1, \tag{6.1}$$

where

$$\Delta z(t) = z(t+1) - z(t), \quad z(0)=z_0,$$

or they are related by an implicit function of the type:

$$h\left[t, z(t+1), z(t), x(t), y(t), \omega\right] = 0, \quad t=0,\cdots,T-1. \tag{6.2}$$

The latter one of these is the typical form one finds in operations research models, economics and system analysis, the first one (6.1) is the conventional one in the theory of optimal control and its applications in engineering, inventory control, etc. In the formulation (6.1) an additional computational problem arises from the fact that it is necessary to solve a large system of linear or nonlinear equations, in order to obtain a description of the evolution of the system.

The objective and constraints functions of stochastic dynamic problems are generally expressed in terms of mathematical expectations of functions that we take to be:

$$g_i\left[z(0), x(0), y(0), \cdots, z(T), x(T), y(T)\right], \quad i=0,1,\ldots,m. \tag{6.3}$$

If no observations are allowed, then equations (6.1), or (6.2), and (6.3) do

not depend on y, and we have the following one-stage problem

$$\begin{aligned}
&\text{find } x = \left[x(0), x(1), \ldots, x(T)\right] \text{ such that} \qquad (6.4)\\
&x(t) \in X(t) \subset R^n, \qquad t=0, \cdots, T,\\
&\Delta z(t) = h\left[t, z(t), x(t), \omega\right], \qquad t=0, \cdots, T-1,\\
&E\left[g_i(z(0), x(0), \cdots, z(T), x(T), \omega\right] \leq 0, \qquad i=1, \cdots, m\\
&\text{and } v = E\{g_0(z(0), x(0), \cdots, z(T), x(T), \omega)\} \text{ is minimized}
\end{aligned}$$

or with the dynamics given by (6.2). Since in (6.1) or (6.2), the variables $z(t)$ are functions of (x,ω), the functions g_i are also implicit functions of (x,ω), i.e. we can rewrite problem (6.4) in terms of functions

$$f_i(x,\omega) = g_i\left[z(x,\omega), x, \omega\right],$$

the stochastic dynamic problem (6.4) is then reduced to a stochastic optimization problem of type (2.6). The implicit form of the objective and the constraints of this problem requires a special calculus for evaluating these functions and their derivatives, but it does not alter the general solution strategies for stochastic programming problems.

The two-stage recourse model allows for a recourse decision y that is based on (the first stage decision x and) the result of observations. The following simple example should be useful in the development of a dynamical version of that model. Suppose we are interested in the design of an optimal trajectory to be followed, in the future, by a number of systems that have a variety of (dynamical) characteristics. For instance, we are interested in building a road between two fixed points (see Figure 4) at minimum total cost taking into account, however, certain safety requirements. To compute the total cost we take into account not just the construction costs, but also the cost of running the

vehicles on this road.

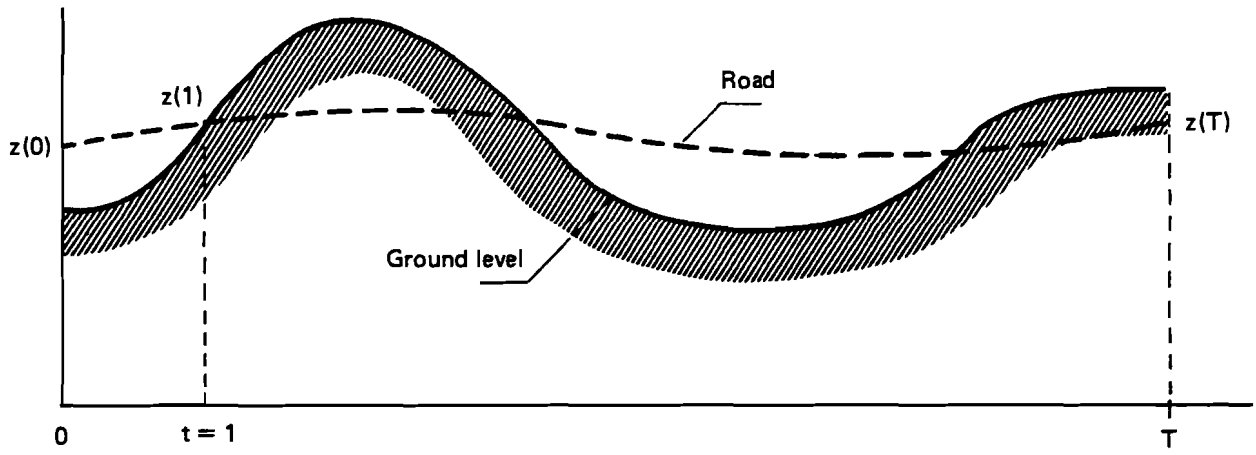

Figure 4.. Road design problem.

For a fixed feasible trajectory

$$z = \left[z,(0),\ z(1),\ldots,\ z(T)\right],$$

and a (dynamical) system whose characteristics are identified by a parameter $\omega \in \Omega$, the dynamics are given by the equations, for $t=0,\ldots,\ T-1$, and $\Delta z(t) = z(t+1) - z(t)$,

$$\Delta z(t) = h\left[t,\ z(t),\ y(t),\ \omega\right], \tag{6.5}$$

and

$$z(0) = z_0,\ z(T) = z_T\ .$$

The variables

$$y = \left[y(0),\ y(1),\ \cdots,\ y(T)\right]$$

are the control variables at times $t=0,1,\ldots,T$. The choice of the z-trajectory is subject to certain restrictions, that include safety considerations, such as

$$|\Delta z(t)| \leq d_1, \ |\Delta z(t) - \Delta z(t-1)| \leq d_2 ,$$

i.e. the first two derivatives cannot exceed certain prescribed levels.

For a specific system $\omega \in \Omega$, and a fixed trajectory z, the optimal control actions (recourse)

$$y(z,\omega) = \left[y(0,z,\omega), y(1,z,\omega), \cdots, y(T,z,\omega)\right]$$

is determined by minimizing the loss function

$$g_0\left[z(0), y(0), \cdots, z(T-1), y(T-1), z(T),\omega\right]$$

subject to the system's equations (6.5) and possibly some constraints on y. If P is the *a priori* distribution of the systems parameters, the problem is to find a trajectory (road design) z that minimizes in the average the loss function, i.e.

$$F_0(z) = E\left\{g_0\left[z(0), y(0,z,\omega), \cdots, z(T-1), y(T-1,z,\omega), z(T),\omega\right]\right\} \quad (6.7)$$

subject to some constraints of the type (6.6).

In this problem the observation takes place in one step only. We have amalgamated all future observations that will actually occur at different time periods in a single collection of possible environments (events). There are problems where ω has the structure

$$\omega = \left[\omega(0), \omega(1), \cdots, \omega(T)\right]$$

and the observations take place in T steps. As an important example of such a class, let us consider the following problem: the long term

decision $x = \left[x(0), x(1), \cdots, x(T)\right]$ and the corrective recourse actions $y = \left[y(0), y(1), \cdots, x(T)\right]$ must satisfy the linear system of equations:

$$\begin{aligned}
&A_{00}\,x(0) + B_0\,y(0) && \geq h(0)\\
&A_{10}\,x(0) + A_{11}\,x(1) + B_1\,y(1) && \geq h(1)\\
&\cdots\cdots \qquad \cdots\cdots \qquad \cdots\cdots\\
&A_{T0}\,x(0) + A_{T1}\,x(1) + \cdots + A_{TT}\,x(T) + B_T\,y(T) && \geq h(T),\\
&x(0) \geq 0, \cdots, x(T) \geq 0; \quad y(0) \geq 0, \cdots, y(T) \geq 0
\end{aligned}$$

where the matrices A_{tk}, B_t and the vectors $h(t)$ are random, i.e. depend on ω. The sequence $x = \left[x(0), \cdots, x(T)\right]$ must be chosen before any information about the values of the random coefficients can be collected. At time $t=0, \cdots, T$, the actual values of the matrices, and vectors,

$$A_{tk}, \quad k=0, \cdots, t;\ B_t,\ h(t),\ d(t)$$

are revealed, and we adapt to the existing situation by choosing a corrective action $y(t,x,\omega)$ such that

$$y(t,x,\omega) \in \operatorname{argmin}\left[d(t)y \mid B_t y \geq h(t) - \textstyle\sum_{k=0}^{t} A_{tk}\,x(k),\ y \geq 0\right].$$

The problem is to find $x = \left[x(0), \cdots, x(T)\right]$ that minimizes

$$F_0(x) = \textstyle\sum_{t=0}^{T}\left[c(t)x(t) + E\{d(t)y(t,x,\omega)\}\right] \tag{6.9}$$

subject to $x(0) \geq 0, \cdots, x(T) \geq 0$.

In the functional (6.9), or (6.7), the dependence of $y(t,x,\omega)$ on x is nonlinear, thus these functions do not possess the separability properties necessary to allow direct use of the conventional recursive equations of dynamic programming. For problem (6.4), these equations can be derived, provided the functions g_i, $i=0, \cdots, m$, have certain specific properties. There are, however, two major obstacles to the use of such

recursive equations in the stochastic case: the tremendous increase of the dimensionality, and again, the more serious problem created by the need of computing mathematical expectations.

For example, consider the dynamic system described by the system of equations (6.1). Let us ignore all constraints except $x(t) \in X(t)$, for $t = 0, 1, \cdots, T$. Suppose also that

$$\omega = \left[\omega(0), \omega(1), \cdots, \omega(T)\right]$$

where $\omega(t)$ only depends on the past, i.e. is independent of $\omega(t+1), \cdots, \omega(T)$. Since the minimization of

$$F_0(x) = E\{g_0(z(0), x(0), \cdots, z(T), x(T), \omega)\}$$

with respect to x can then be written as:

$$\min_{x(0)} \min_{x(1)} \cdots \min_{x(T)} E\{g_0\}$$

and if g_0 is separable, i.e. can be expressed as

$$g_0 := \sum_{t=0}^{T-1} g_{0t}\left[\Delta z(t), x(t), \omega(t)\right] + g_{0T}\left[z(t), \omega(T)\right]$$

then

$$\begin{aligned}\min_x F_0(x) = \min_{x(0)} E\left\{g_{00}\left[\Delta z(0), x(0), \omega(0)\right]\right\} + \min_{x(1)} E\left\{g_{01}\left[\Delta z(1), x(1), \omega(1)\right]\right\} \\ + \cdots + \min_{x(T-1)} E\left\{g_{0,T-1}\left[\Delta z(T-1), x(T-1), \omega(T-1)\right]\right\} + \\ + E\left\{g_{0T}\left[z(t), \omega(T)\right]\right\}\end{aligned}$$

Recall that here, notwithstanding its sequential structure, the vector ω is to be revealed in one global observation. Rewriting this in backward recursive form yields the Bellman equations:

$$v_t(z_t) = \min \left[E\left\{g_{0t}\left[h(t, z_t, x, \omega(t)), x, \omega(t)\right]\right.\right. \tag{6.10}$$

$$+ v_{t+1}\left[z_t + h(t, z_t, x, \omega(t))\right]\Big\} \mid x \in X(t)\Big]$$

for $t = 0, \cdots, T-1$, and

$$v_T(z_T) = E\left\{g_{0T}\left[z_T, \omega(T)\right]\right\}, \tag{6.11}$$

where v_t is the value function (optimal loss-to-go) from time t on, given state z_t at time t, that in turn depends on $x(0), x(1), \cdots, x(t-1)$.

To be able to utilize this recursion, reducing ultimately the problem to:

find $x \in X(0) \subset R^n$ such that v_0 is minimized, where

$$v_0 = E\left\{g_{00}\left[h(0, z_0, x, \omega(0)), x, \omega(0)\right] + v_1\left[z_0 + h(0, z_0, x, \omega(0))\right]\right\},$$

we must be able to compute the mathematical expectations

$$E\left\{g_{0t}\left[\Delta z(t), x, \omega(t)\right]\right\}$$

as a function of the intermediate solutions $x(0), \cdots, x(t-1)$, that determine $\Delta z(t)$, and this is only possible in special cases. The main goal in the development of solution procedures for stochastic programming problems is the development of appropriate computational tools that precisely overcome such difficulties.

A much more difficult situation may occur in the (full) multistage version of the recourse model where observation of some of the environment takes place at each stage of the decision process, at which time (taking into account the new information collected) a new recourse action is taken. The whole process looks like a sequence of alternating: decision-observation- ... -observation-decision.

Let x be the decision at stage $k = 0$, which may itself be split into a sequence $x(0), \cdots, x(N)$, each $x(k)$ corresponding to that component of x that enters into play at stage k, similar to the dynamical version of the two-stage model introduced earlier. Consider now a sequence

$$y = \left[y(0), y(1), \cdots, y(N)\right]$$

of recourse decisions (adaptive actions, corrections), $y(k)$ being associated specifically to stage k. Let

$$B_k : = \text{ information set at stage } k \ ,$$

consisting of past measurements and observations, thus $B_k \subset B_{k+1}$.

The *multistage recourse problem* is

$$\begin{aligned}
&\text{find } x \in X \subset R^n \text{ such that} \qquad (6.12)\\
&f_0 i(x) \le 0, \quad i=1, \cdots, m_0 \ ,\\
&E\{f_{1i}(x, y(1), \omega) \mid B_1\} \le 0, \quad i=1, \cdots, m_1 ,\\
&\cdots\cdots \qquad\qquad \cdots\cdots\\
&E\{f_{Ni}(x, y(1), \cdots, y(N), \omega) \mid B_N\} \le 0, \quad i=1, \cdots, m_N \ ,\\
&y(k) \in Y(k), \quad k=1, \cdots, N \ ,\\
&\text{and } F_0(x) \text{ is minimized}
\end{aligned}$$

where

$$F_0(x) = E^{B_0}\left\{\min_{y(1)} E^{B_1}\left\{\cdots \min_{y(N-1)} E^{B_{N-1}}\{f(x, y(1), \cdots, y(N), \omega)\}\right\}\right\}$$

If the decision x affects only the initial stage $k = 0$, we can obtain recursive equations similar to (6.10) - (6.11) except that expectation E must be replaced by the conditional expectations E^{B_i}, which in no way simplifies the numerical problem of finding a solution. In the more general case when $x = \left[x(0), x(1), \cdots, x(N)\right]$, one can still write down recursion formulas but of such (numerical) complexity that all hope of solving

this class of problems by means of these formulas must quickly be abandoned.

7. SOLVING THE DETERMINISTIC EQUIVALENT PROBLEM

All of the preceding discussion has suggested that the problem:

$$\begin{aligned} &\text{find } x \in R^n \text{ such that} \\ &F_i(x) = \int f_i(x,\omega)\, P(d\omega) \le 0, \quad i=1,\cdots,m, \\ &\text{and } z = F_0(x) = \int f_0(x,\omega)\, P(d\omega) \text{ is minimized,} \end{aligned} \tag{7.1}$$

exhibits all the peculiarities of stochastic programs, and that for exploring computational schemes, at least at the conceptual level, it can be used as the canonical problem.

Sometimes it is possible to find explicit analytical expressions for an acceptable approximation of the F_i. The randomness in problem (7.1) disappears and we can rely on conventional deterministic optimization methods for solving (7.1). Of course, such cases are highly cherished, and can be dealt with by relying on standard nonlinear programming techniques.

One extreme case is when $\bar{\omega} = E\{\omega\}$ is a *certainty equivalent* for the stochastic optimization problem, i.e. the solution to (7.1) can be found by solving:

$$\begin{aligned} &\text{find } x \in X \subset R^n \text{ such that} \\ &f_i(x,\bar{\omega}) \le 0, \quad i=1,\cdots,m, \\ &\text{and } z = f_0(x,\bar{\omega}) \text{ is minimized,} \end{aligned} \tag{7.2}$$

this would be the case if the f_i are linear functions of ω. In general, as already mentioned in Section 3, the solution of (7.2) may have little in

common with the initial problem (7.1). But if the f_i are convex functions, then according to Jensen's inequality

$$E\{f_i(x,\omega)\} \geq f_i(x,\bar{\omega}), \qquad i=1,\cdots,m,$$

This means that the set of feasible solutions in (7.2) is larger than in (7.1) and hence the solution of (7.2) could provide a lower bound for the solution of the original problem.

Another case is a stochastic optimization problem with simple probabilistic constraints. Suppose the constraints of (7.1) are of the type

$$P\left\{\omega \mid \sum_{j=1}^{n} t_{ij}\, x_j > h_i(\omega)\right\} \geq \alpha_i, \qquad i=1,\cdots,m,$$

with deterministic coefficients t_{ij} and random right-hand sides $h_i(\cdot)$. Then these constraints are equivalent to the linear system

$$\sum_{j=1}^{n} t_{ij}\, x_j \geq h_i^*, \qquad i=1,\cdots,m,$$

where

$$h_i^* = \inf\{t \mid P[\omega \mid h_i(\omega) < t] \geq \alpha_i\}$$

If all the parameters t_{ij} and h_i in (7.3) are jointly normally distributed (and $\alpha_i \geq .5$), then the constraints

$$x_0 = 1$$
$$\sum_{j=0}^{n} \bar{t}_{ij}\, x_j + \beta \left(\sum_{j=0}^{n} \sum_{k=0}^{n} \tau_{ijk}\, x_j\, x_k\right)^{1/2} \leq 0$$

can be substituted for (7.3), where

$$t_{i0}(\cdot) = -h_i(\cdot)$$
$$\bar{t}_{ij} := E\{t_{ij}(\omega)\}, \qquad j=0,1,\cdots,n,$$
$$\tau_{ijk} := \operatorname{cov}\left[t_{ij}(\cdot),\, t_{ik}(\cdot)\right], \quad j=0,\cdots,n;\ k=0,\cdots,n,$$

and β is a coefficient that identifies the α-fractile of the normalized

normal distribution.

Another important class are those problems classified as stochastic programs with simple recourse, or more generally recourse problems where the random coefficients have a discrete distribution with a relatively small number of density points (support points). For the linear model (5.1) introduced in Section 5, where

$$\Omega = \left\{(q^1, W^1, h^1, T^1), \cdots, (q^N, W^N, h^N, T^N)\right\}$$

where for $k=1, \cdots, N$, the point (q^k, W^k, h^k, T^k) is assigned probability p_k, one can find the solution of (5.1) by solving:

$$\begin{array}{llll} \text{find } x \in R^n_+, \left\{y^k \in R^{n'}_+, k=1, \cdots, N\right\} \text{ such that} & & & (7.4) \\ Ax & & & \geq b, \\ T^1x + W^1y^1 & & & = h^1, \\ T^2x & + W^2y^2 & & = h^2, \\ \vdots & & \ddots & \vdots \\ T^Nx & & + W^Ny^N & = h^N, \\ cx + p_1q^1y^1 & + p_2q^2y^2 \cdots & + p_Nq^Ny^N & = z, \\ \text{and } z \text{ is minimized} & & & \end{array}$$

This problem has a (dual) block-angular structure. It should be noticed that the number N could be astronomically large, if only the vector h is random and each component of the vector

$$h = (h_1, h_2, \cdots, h_{m'})$$

has two independent outcomes, then $N = 2^{m'}$. A direct attempt at solving (7.4) by conventional linear programming techniques will only yield at each iteration very small progress in the terms of the x variables. Therefore, a special large scale optimization technique is needed for solving

even this relatively simple stochastic programming problem.

8. APPROXIMATION SCHEMES

If a problem is too difficult to solve one may have to learn to live with approximate solutions. The question however, is to be able to recognize an approximate solution if one is around, and also to be able to assess how far away from an optimal solution one still might be. For this one needs a convergence theory complemented by (easily computable) error bounds, improvement schemes, etc. This is an area of very active research in stochastic optimization, both at the theoretical and the software-implementation level. Here we only want to highlight some of the questions that need to be raised and the main strategies available in the design of approximation schemes.

For purposes of discussion it will be useful to consider a simplified version of (7.1):

$$\text{find } x \in X \subset R^n \text{ that minimizes} \qquad (8.1)$$
$$F_0(x) = \int f_0(x,\omega)\, P(d\omega),$$

we suppose that the other constraints have been incorporated in the definition of the set X. We deal with a problem involving one expectation functional. Whatever applies to this case also applies to the more general situation (7.1), making the appropriate adjustments to take into account the fact that the functions

$$F_i(x) = \int f_i(x,\omega)\, P(d\omega), \qquad i=1,\cdots,m,$$

determine constraints.

Given a problem of type (8.1) that does not fall in one of the nice categories mentioned in Section 7, one solution strategy may be to replace it by an approximation,*. There are two possibilities to simplify the integration that appears in the objective function, replace f_0 by an integrand f_0^ν or replace P by an approximation P_ν, and of course, one could approximate both quantities at once.

The possibility of finding an acceptable approximate of f_0 that renders the calculation of

$$\int f_0{}^\nu(x,\omega)\, P(d\omega) =: F_0^\nu(x),$$

sufficiently simple so that it can be carried out analytically or numerically at low-cost, is very much problem dependent. Typically one should search for a separable function of the type

$$f_0{}^\nu(x,\omega) \;=\; \textstyle\sum_{j=1}^{q} \varphi_j(x,\omega_j),$$

recall that $\Omega \subset R^q$, so that

$$F_0^\nu(x) = \textstyle\sum_{j=1}^{q} \int \varphi_j(x,\omega_j)\, P(d\omega) = \sum_{j=1}^{q} \int \varphi_j(x,\omega_j)\, P_j(d\omega_j)$$

where the P_j are the marginal measures associated to the j-th component of ω. The multiple integral is then approximated by the sum of 1-dimensional integrals for which a well-developed calculus is available, (as well as excellent quadrature subroutines). Let us observe that we do not necessarily have to find approximates that lead to 1-dimensional integrals, it would be acceptable to end up with 2-dimensional integrals, even in some cases -- when P is of certain specific types -- with 3-dimensional integrals. In any case, this would mean that the structure

* Another approach will be discussed in Section 9.

of f_0 is such that the interactions between the various components of ω play only a very limited role in determining the cost associated to a pair (x,ω). Otherwise an approximation of this type could very well throw us very far off base. We shall not pursue this question any further since they are best handled on a problem by problem basis. If $\{f_0^\nu, \nu=1, \cdots\}$ is a sequence of such functions converging, in some sense, to f, we would want to know if the solutions of

$$x^\nu \in \operatorname{argmin} F^\nu = \int f_0^\nu(\cdot,\omega)\, P(d\omega), \qquad \nu=1, \cdots$$

converge to the optimal solution of (8.1) and if so, at what rate. These questions would be handled very much in the same way as when approximating the probability measure as well be discussed next.

Finding valid approximates for f_0 is only possible in a limited number of cases while approximating P is always possible in the following sense. Suppose P_ν is a probability measure (that approximates P), then

$$|F_0^\nu(x) - F_0(x)| \leq \int |f_0(x,\omega)|\, |P_\nu - P|\,(d\omega). \tag{8.2}$$

Thus if f_0 has Lipschitz properties, for example, then by choosing P_ν sufficiently close to P we can guarantee a maximal error bound when replacing (8.1) by:

$$\text{find } x \in X \subset R^n \text{ that minimizes } F_0^\nu(x) = \int f_0(x,\omega)\, P_\nu(d\omega). \tag{8.3}$$

Since it is the multidimensional integration with respect to P that was the source of the main difficulties, the natural choice -- although in a few concrete cases there are other possibilities -- for P_ν is a discrete distribution that assigns to a finite number of points

$$\omega^1, \omega^2, \cdots, \omega^L$$

the probabilities

$$p_1, p_2, \cdots, p_L ;$$

Problem (8.3) then becomes:

$$\text{find } x \in X \subset R^n \text{ that minimizes } F_0^\nu(x) = \sum_{l=1}^{L} p_l \, f_0(x, \omega^l) \qquad (8.4)$$

At first glance it may now appear that the optimization problem can be solved by any standard nonlinear programming, the sum $\sum_{l=1}^{L}$ involving only a "finite" number of terms, the only question being how "approximate" is the solution of (8.4). However, if inequality (8.2) is used to design this approximation, to obtain a relatively sharp bound from (8.2), the number L of discrete points required may be so large that problem (8.4) is in no way any easier than our original problem (8.1). To fix the ideas, if $\Omega \subset R^{10}$, and P is a continuous distribution, a good approximation - as guaranteed by (8.2) - may require having $10^{10} \leq L \leq 10^{11}$! This is jumping from the stove into the frying pan.

This clearly indicates the need for more sophisticated approximation schemes. As background, we have the following convergence results. Suppose $\{P_\nu, \nu=1, \cdots\}$ is a sequence of probability measures that converge in distribution to P, and suppose that for all $x \in X$, the function $f_0(x, \omega)$ is uniformly integrable with respect to all P_ν, and suppose there exists a bounded set D such that

$$D \cap \operatorname{argmin} \left[F_0^\nu(x) = \int f_0(x, \omega) \, P_\nu(d\omega) \mid x \in X \right] \neq !0$$

for almost all ν, then

$$\inf_X F_0 = \lim_{\nu \to \infty} (\inf_X F_0^\nu)$$

and

$$\text{if } x^{\nu} \in \operatorname{argmin}_X F_0^{\nu},\ x = \lim_{k\to\infty} x^{\nu_k}$$

then

$$x \in \operatorname{argmin}_X F_0.$$

The convergence result indicates that we are given a wide latitude in the choice of the approximating measures, the only real concern is to guarantee the convergence in distribution of the P_ν to P, the uniform integrability condition being from a practical viewpoint a pure technicality.

However, such a result does not provide us with error bounds, but since we can choose the P_ν in such a wide variety of ways, we could for example have P_ν such that

$$\inf_X F_0^{\nu} \le \inf_X F_0 \tag{8.5}$$

and $P_{\nu+1}$ such that

$$\inf_X F_0 \le \inf_X F_0^{\nu+1} \tag{8.6}$$

providing us with upper and lower bounds for the infimum and consequently error bounds for the approximate solutions:

$$x^{\nu} \in \operatorname{argmin}_X F_0^{\nu}, \text{and } x^{\nu+1} \in \operatorname{argmin}_X F_0^{\nu+1}.$$

This, combined with a sequential procedure for redesigning the approximations P_ν so as to improve the error bounds, is very attractive from a computational viewpoint since we may be able to get away with discrete measures that involve only a relatively small number of points (and this seems to be confirmed by computational experience).

The only question now is how to find these measures that guarantee (8.5) and (8.6). There are basically two approaches: the first one that exploits the properties of the function $\omega \mapsto f_0(x,\omega)$ so as to obtain inequalities when taking expectations, and the second one that chooses P_ν in a class of probability measures that have characteristics similar to P but so that P_ν dominates or is dominated by P and consequently yields the desired inequality (8.5) or (8.6). A typical example of this latter case is to choose P_ν so that it majorizes or is majorized by P, another one is to choose P_ν so that for at least for some $\hat{x} \in X$:

$$P_\nu \in \operatorname{argmax} \left[\int f_0(\hat{x},\omega)\, Q(d\omega) \mid Q \in P \right] \tag{8.7}$$

where P is a class of probability measures on Ω that contains P, for example

$$P = \left\{ Q \mid \int \omega\, Q(d\omega) = E\{\omega\} \right\}.$$

Then

$$F_0^\nu(\hat{x}) \geq F_0(x) \geq \inf_X F_0$$

yields an upper bound. If instead of P_ν in the argmax we take P_ν in the argmin we obtain a lower bound.

If $\omega \mapsto f_0(x,\omega)$ is convex (concave) or at least locally convex (locally concave) in the area of interest we may be able to use Jensen's inequality to construct probability measures that yield lower (upper) approximates for F_0 and probability measures concentrated on extreme points to obtain upper (lower) approximates of F_0. We have already seen such an example in Section 7 in connection with problem (7.2) where P is

replaced by P_ν that concentrate all the probability mass on $\bar{\omega} = E\{\omega\}$.

Once an approximate measure P_ν has been found, we also need a scheme to refine it so that we can improve, if necessary, the error bounds. One cannot hope to have a universal scheme since so much will depend on the problem at hand as well as the discretizations that have been used to build the upper and lower bounding problems. There is, however, one general rule that seems to work well, in fact surprisingly well, in practice: choose the region of refinement of the discretization in such a way as to capture as much of the nonlinearity of $f_0(x,\cdot)$ as possible.

It is, of course, not necessary to wait until the optimal solution of an approximate problem has been reached to refine the discretization of the probability measure. Conceivably, and ideally, the iterations of the solutions procedure should be intermixed with the sequential procedure for refining the approximations. Common sense dictates that as we approach the optimal solution we should seek better and better estimates of the function values and its gradients. How many iterations should one perform before a refinement of the approximation is introduced, or which tell-tale sign should trigger a further refinement, are questions that have only been scantily investigated, but are ripe for study at least for certain specific classes of stochastic optimization problems.

As to the rate of convergence this is a totally open question, in general and in particular, except on an experimental basis where the results have been much better than what could be expected from the theory.

One open challenge is to develop the theory that validates the convergence behavior observed in practice.

9. STOCHASTIC PROCEDURES

Let us again consider the general formulation (2.6) for stochastic programs:

$$\begin{aligned} &\text{find } x \in X \subset R^n \text{ such that} \\ &F_i(x) = \int f_i(x,\omega)\, P(d\omega) \le 0, \qquad i=1,\cdots,m, \\ &\text{and } F_0(x) = \int f_0(x,\omega)\, P(d\omega) \text{ is minimized.} \end{aligned} \tag{9.1}$$

We already know from the discussion in Sections 3 and 7 that the exact evaluation of the integrals is only possible in exceptional cases, for special types of probability measures P and integrands f_i. The rule in practice is that it is only possible to calculate random observations $f_i(x,\omega)$ of $F_i(x)$. Therefore in the design of universal solution procedures we should rely on no more than the random observations $f_i(x,\omega)$. Under these premises, finding the solution of (9.1) is a difficult problem at the border between mathematical statistics and optimization theory. For instance, even the calculation of the values $F_i(\bar{x})$, $i=0,\ldots,m$, for a fixed $\bar{x}$ requires statistical estimation procedures: on the basis of the observations

$$f_i(\bar{x},\omega^0),\ f_i(\bar{x},\omega^1),\ \cdots,\ f_i(\bar{x},\omega^s),\ \cdots$$

one has to estimate the mean value

$$E\{f_i(\bar{x},\omega)\}.$$

The answer to the simplest question, whether or not a given $\bar{x} \in X$ is feasible, requires verifying the statistical hypothesis that

$$E\{f_i(\bar{x},\omega)\} \leq 0, \quad \text{for} \quad i=1,\cdots,m.$$

Since we can only rely on random observations, it seems quite natural to think of stochastic solution procedures that do not make use of the exact values of the $F_i(x)$, $i=0,\cdots,m$. Of course, we cannot guarantee in such a situation a monotonic decrease (or increase) of the objective value as we move from one iterate to the next, thus these methods must, by the nature of things, be non-monotonic.

Deterministic processes are special cases of stochastic processes, thus stochastic optimization gives us an opportunity to build more flexible and effective solution methods for problems that cannot be solved within the standard framework of deterministic optimization techniquest. *Stochastic quasi-gradient methods* is a class of procedures of that type. Let us only sketch out their major features. We consider two examples in order to get a better grasp of the main ideas involved.

Example 1: *Optimization by simulation.* Let us imagine that the problem is so complicated that a computer based simulation model has been designed in order to indicate how the future might unfold in time for each choice of a decision x. Suppose that the stochastic elements have been incorporated in the simulation so that for a single choice x repeated simulation runs results in different outputs. We always can identify a simulation run as the observation of an event (environment) ω from a sample space Ω. To simplify matters, let us assume that only a single quantity

$$f_0(x,\omega)$$

summarizes the output of the simulation run ω for given x. The problem is to

$$\text{find } x \in R^n \text{ that minimizes } F_0(x) = E\{f_0(x,\omega)\}. \tag{9.2}$$

Let us also assume that F_0 is differentiable. Since we do not know with any level of accuracy the values or the gradients of F_0 at x, we cannot apply the standard gradient method, that generates iterates through the recursion:

$$x^{s+1} := x^s - \rho_s \sum_{j=1}^{n} \frac{F_0(x^s + \Delta_s e^j) - F_0(x^s)}{\Delta_s} e^j , \tag{9.3}$$

where ρ_s is the step-size, Δ_s determines the mesh for the finite difference approximation to the gradient, and e^j is the unit vector on the j-th axis. A well-known procedure to deal with the minimization of functions in this setting is the so-called *stochastic approximation method* that can be viewed as a recursive Monte-Carlo optimization method. The iterates are determined as follows:

$$x^{s+1} := x^s - \rho_s \sum_{j=1}^{n} \frac{f_0(x^s + \Delta_s e^j, \omega^{sj}) - f_0(x^s, \omega^{s0})}{\Delta_s} e^j , \tag{9.4}$$

where $\omega^{s0}, \omega^{s1}, \cdots, \omega^{sn}$ are observations, not necessarily mutually independent one possibility is $\omega^{s0} = \omega^{s1} = \cdots = \omega^{sn}$. The sequence $\{x^s, s=0,1,...\}$ generated by the recursion (9.4) converges with probability 1 to the optimal solution provided, roughly speaking, that the scalars $\{\rho_s, \Delta_s; s=1, \cdots\}$ are chosen so as to satisfy

$$\rho_s \geq 0, \ \sum_s \rho_s = \infty, \ \sum_s (\rho_s^2 + \rho_s \Delta_s) < \infty,$$

($\rho_s = \Delta_s = 1/s$ are such sequences), the function F_0 has bounded second

derivatives and for all $x \in R^n$,

$$E\{|\Delta f_0(x,\omega)|^2\} \leq d(1+|x|^2),\ d > 0. \tag{9.5}$$

This last condition is quite restrictive, it excludes polynomial functions $f_0(\cdot,\omega)$ of order greater than 3. Therefore, the methods that we shall consider next will avoid making such a requirement, at least on all of R^n.

Example 2: *Optimization by random search.* Let us consider the minimization of a convex function F_0 with bounded second derivatives and n a relatively large number of variables. Then the calculation of the exact gradient ∇F_0 at x requires calling up a large number of times the subroutines for computing all the partial derivatives and this might be quite expensive. The finite difference approximation of the gradient in (9.3) require $(n+1)$ function-evaluations per iteration and this also might be time-consuming if function-evaluations are difficult. Let us consider that following random search method: at each iteration $s=0,1...$, choose a direction h^s at random, see Figure 5.

If F_0 is differentiable, this direction h^s or its opposite $-h^s$ leads into the region

$$\{x \mid F_0(x) \leq F_0(x^s)\}$$

of lower values for F_0, unless x^s is already the point at which F_0 is minimized. This simple idea is at the basis of the following random search procedure:

$$x^{s+1} := x^s - \frac{3}{2}\rho_s \frac{F_0(x^s+\Delta_s h^s) - F_0(x^s)}{\Delta_s} h_s, \tag{9.6}$$

which requires only two function-evaluations per iteration. Numerical

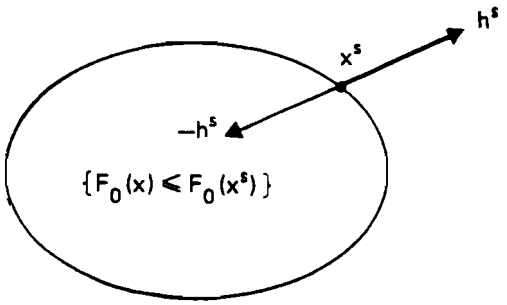

Figure 5. Random search directions $\pm h^s$.

experimentation shows that the number of function-evaluations needed to reach a good approximation of the optimal solution is substantially lower if we use (9.6) in place of (9.3). The vectors $h^0, h^1, \cdots, h^1, \cdots$ often are taken to be independent samples of vectors $h(\bullet)$ whose components are independent random variables uniformly distributed on $[-1, +1]$.

Convergence conditions for the random search method (9.6) are the same, up to some details, as those for the stochastic approximation method (9.4). They both have the following feature: the direction of movement from each $x^s, s=0,1, \cdots$ are statistic estimates of the gradient $\nabla F_0(x^s)$. If we rewrite the expressions (9.4) and (9.6) as :

$$x^{s+1} := x^s - \rho_s \xi^s, \; s=0,1, \cdots \tag{9.7}$$

where ξ^s is the direction of movement, then in both cases

$$E\{\xi^s \mid x^s\} = \nabla F_0(x^s) + o(\Delta_s) \tag{9.8}$$

A general scheme of type (9.7) that would satisfy (9.8) combines the ideas of both methods. There may, of course, be many other procedures that fit into this general scheme. For example consider the following iterative method:

$$x^{s+1} := x^s - \rho_s \frac{f_0(x^s + \Delta_s h^s, \omega^{s1}) - f_0(x^s, \omega^{s0})}{\Delta_s} h^s , \qquad (9.9)$$

which requires only two observations per iteration, in contrast to (9.4) that requires $(n+1)$ observations. The vector

$$\xi^s = \frac{3}{2} \frac{f_0(x^s + \Delta_s h^s, \omega^{s1}) - f_0(x^s, \omega^{s0})}{\Delta_s} h^s$$

also satisfies the condition (9.8),

$$\frac{2}{3} E\{\xi^s \mid x^s\} = E\left\{\frac{F_0(x^s + \Delta_s h^s) - F_0(x^s)}{\Delta_s} h^s\right\}$$

$$= E\left\{\sum_{j=1}^{n} \left[\frac{\partial}{\partial x_j} F_0(x^s)\right) h_i^s\right\} + O(\Delta_s) = \nabla F_0(x^s) + O(\Delta_s).$$

The convergence of all these particular procedures (9.4), (9.6), (9.9) follow from the convergence of the general scheme (9.7) - (9.8). The vector ξ^s satisfying (9.8) is called a *stochastic quasi-gradient* of F_0 at x_s, and the scheme (9.7) - (9.8) is an example of a stochastic quasi-gradient procedure.

Unfortunately this procedure cannot be applied, as such, to finding the solution of the stochastic optimization problem (9.1) since we are dealing with a constrained optimization problem, and the functions F_i, $i=0, \cdots, m$, are in general nondifferentiable. So, let us consider a

simple generalization of this procedure for solving the constrained optimization problem with nondifferentiable objective:

$$\text{find } x \in X \subset R^n \text{ that minimzes } F_0(x) \tag{9.10}$$

where X is closed convex set and F_0 is a real-valued (continuous) convex function. The new algorithm generates a sequence $x^0, x^1, \cdots, x^s, \cdots$ of points in X by the recursion:

$$x^{s+1} := prj_X [x^s - \rho_s \xi^s] \tag{9.11}$$

where prj_X means projection on X, and ξ^s satisfies

$$E\{\xi^s \mid x^0, x^1, \cdots, x^s\} \in \partial F_0(x^s) + \varepsilon^s \tag{9.12}$$

with

$$\partial F_0(x^s) := \text{the set of subgradients of } f_0 \text{ at } x^s ,$$

and ε^s is a vector, that may depend on $(x^0, \cdots, x^s)$, that goes to 0 (in a certain sense) as s goes to ∞. The sequence $\{x^s, s=0,1, \cdots\}$ converges with probability 1 to an optimal solution, when the following conditions are satisfied with probability 1:

$$\rho_s \geq 0, \ \sum_s \rho_s = \infty, \ \sum_s E\{\rho_s \| \varepsilon_s \| + \rho_s^2\} < \infty ,$$

and

$E\{\| \xi^s \|^2 \mid x^0, \cdots, x^s\}$ is bounded whenever $\{x^0, \cdots, x^s\}$ is bounded.

Convergence of this method, as well as its implementation, and different generalizations are considered in the literature.

To conclude let us suggest how the method could be implemented to solve the linear recourse problem (5.1). From the duality theory for linear programming, and the definition (5.2) of Q, one can show that

$$\partial\, Q(x,\omega) := \left\{ -u\ T(\omega) \mid u \in \operatorname{argmax}_v \left[v(h(\omega - T(\omega)x) \mid vW(\omega) \leq q(\omega) \right] \right\}.$$

Thus an estimate ξ^s of the gradient of F_0 at x^s is given by

$$\xi^s \ + \ c \ - u^s \ T(\omega^s)$$

where ω^s is obtained by random sampling from Ω (using the measure P), and

$$u^s \in \operatorname{argmax}_v \left[v(h(\omega^s) - T(\omega^s)x) \mid vW(\omega^s) \leq q(\omega^s) \right]$$

The iterates could then be obtained by

$$x^{s+1} := proj_X \left[x^s + \rho_s \ u^s T(\omega^s) - \rho_s c \right]$$

where

$$X \ = \ \{ x \in R^n_+ \mid Ax \leq b \}.$$

It is not difficult to show that under very weak regularity conditions (involving the dependence of $W(\omega)$ on ω),

$$E\{\xi^s \mid x^s\} \in \partial\, F_0(x^s).$$

10. CONCLUSION

In guise of conclusion, let us just raise the following possibility. The stochastic quasi-gradient method can operate by obtaining its stochastic quasi-gradient from 1 sample of the subgradients of $f_0(\cdot,\omega)$ at x^s, it could equally well use -- if this was viewed as advantageous -- obtain its stochastic quasi-gradient ξ^s by taking a finite sample of the subgradients of $f_0(\cdot,\omega)$ at x^s, say L of them. We would then set

$$\xi^s := \frac{1}{L} \sum_{l=1}^{L} \nu^l \ \text{ where } \ \nu^l \in \partial\, f_0(x^s,\omega^l) \qquad (10.1)$$

and $\omega^1, \cdots, \omega^L$ are random samples (using the measure P). The question of the efficiency of the method taking just 1 sample versus $L \geq 1$ should, and has been raised, cf. the implementation of the methods described in Chapter 16. But this is not the question we have in mind. Returning to Section 8, where we discussed approximation schemes, we nearly always ended up with an approximate problem that involves a discretization of the probability measures assigning probabilities $p_1, \cdots, p_L$ to points $\omega^1, \cdots, \omega^L$, and if a gradient-type procedure was used to solve the approximating problem, *the* gradient, or *a* subgradient of F_0 at x^s would be obtained as

$$\zeta^s := \sum_{l=1}^{L} p_l \, \nu^l \quad \text{where} \quad \nu^l \in \partial f_0(x^s, \omega^l). \tag{10.2}$$

The similarity between expressions (10.1) and (10.2) suggest possibly a new class of algorithms for solving stochastic optimization problems, one that relies on an approximate probability measure (to be refined as the algorithm progresses) to obtain its iterates, allowing for the possibility of a quasi-gradient at each step without losing some of the inherent adaptive possibilities of the quasi-gradient algorithm.

REFERENCES

Dempster, M. *Stochastic Programming*, Academic Press, New York.

Ermoliev, Y. Stochastic quasigradient methods and their applications to system optimization, *Stochastics* 9, 1-36, 1983.

Ermoliev, Y. *Numerical Techniques for Stochastic Optimization Problems*, IIASA Collaborative Volume, forthcoming (1985).

Kall, P. *Stochastic Lineare Programming*, Springer Verlag, Berlin, 1976.

Wets, R. Stochastic programming: solution techniques and approximation schemes, in *Mathematical Programming: The State of the Art*, eds., A. Bachem, M. Grötschel, and B. Korte. Springer Verlag, Berlin, 1983, pp.566-603.