SIMULTANEOUS NONSTATIONARY OPTIMIZATION, ESTIMATION AND APPROXIMATION PROCEDURES

Yuri Ermoliev and A.A. Gaivoronski

April 1982
CP-82-16

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

ABSTRACT

The main aim of this paper is to investigate those algorithmic procedures which solve optimization problems whilst either estimating the unknown parameters of these problems or approximating them by more simple problems. The problem of nonstationary optimization with time-varying functions and a set of optimal solutions (set of equilibriums) is considered. The proposed solution technique is based on the application of nonmonotonic optimization procedures. We derive the convergence of such procedures by studying the Hausdorf distance between a current approximate solution and the set of ε-optimal solutions. The Lipschitz continuity of the Hausdorf distance between sets of ε-optimal solutions upon the parameters of the problem is also discussed.

SIMULTANEOUS NONSTATIONARY OPTIMIZATION, ESTIMATION, AND APPROXIMATION PROCEDURES

Yuri Ermoliev and A.A. Gaivoronski

I. INTRODUCTION

Most mathematical programming applications require the estimation of unknown parameters in the objective function and constraints. In some cases, the tasks of optimization and estimation can be separated and optimization performed after estimation. However, it is often necessary to optimize and estimate *simultaneously*. For instance, optimization cannot be separated from estimation if the observation of unknown parameters depends on the current value of the control variables. In this situation we need algorithmic procedures which solve the optimization problem while estimating the unknown parameters. It will be shown that development of such procedures leads to nonstationary optimization problems, in particular to so-called limit extremal problems (Ermoliev and Gaivoronski 1979; Gaivoronski 1979; Ermoliev 1981).

The objective function $f(x,s)$ and the feasible set X_s in nonstationary problems (Ermoliev and Nurminski 1973; Nurminski 1977; Vertchenko 1977) depend on the iteration number $s = 0,1,\ldots$ It is necessary to create a sequence of approximate solutions $\{x^s\}_{s=0}^{\infty}$, that tends, in some sense, to follow the time-path of the optimal solutions: for $s \to \infty$

$$\lim[f(x^s,s) - \min\{f(x,s) \mid x \in X_s\}] = 0 \qquad (1)$$

The ideas behind the simultaneous optimization and approximation procedures are close to the idea of nonstationary optimization described above.

Many books and papers have been written on optimization and approximation problems. In some approaches the problem of approximation is examined using general optimization techniques; in this case the approximation problem is considered as a special optimization problem. In other approaches, optimization problems are characterized by using approximation ideas to simplify optimization methods. For instance, the methods of feasible directions solve nonlinear programming problems while approximating them by linear programming problems. Such methods approximate nonlinear objective functions $f^0(x)$ and constraints $f^i(x)$, $i=\overline{1,m}$, by linear functions at *every current point* x^s. The idea of optimization through approximation of "bad" functions $f^0(x)$, $f^i(x)$ by a sequence of "good" functions $f^0(x,s) \to f^0(x)$, $f^i(x,s) \to f^i(x)$ in the *entire feasible set* is discussed in Ermoliev and Nurminski (1973), Ermoliev (1976), and Katkovnik and Khejsin (1976).

This paper considers the case in which the approximation of the functions $f^0(x)$, $f^i(x)$ occurs in *the neighborhood* D_s of *every current point* x^s. At each iteration s, a certain domain D_s is determined within which the functions $f^0(x)$, $f^i(x)$ are approximated by the more simple functions $f^0(x,s)$, $f^i(x,s)$ (the latter may be linear, quadratic, convex, etc., depending on the context). A direction of search and a new point x^{s+1} are determined using $f^0(x,s)$, $f^i(x,s)$. A new domain D_{s+1} is then created and the iterations are continued.

The main feature of this method is that a precise approximation $f^0(x,s)$, $f^i(x,s)$ of the functions $f^0(x)$, $f^i(x)$ and a precise optimization of $f^0(x,s)$ are unnecessary. It is sufficient simply to iteratively improve the approximation during the optimization process. Moreover, every iteration is based on information regarding the behavior of the objective function within the neighborhood of x^s. This method is less likely to

stop at a local minimum of $f^0(x)$ than methods based on approximations at points x^s.

II. SIMULTANEOUS OPTIMIZATION AND ESTIMATION PROCEDURES

We shall first consider a simple example--minimization of the differentiable function

$$f(x) = \phi(x,u^*)$$

where $u^* \in R^k$ is a vector of *unknown parameters* and $x \in R^n$ is a vector of *control variables*. At each iteration $s = 0,1,\ldots$, an *observation* h^s is available which has the form of a direct observation of the parameter vector, i.e.,

$$Eh^s = u^* \quad .$$

The problem is to create a sequence of control variables $\{x^s\}_{s=0}^{\infty}$ which converges to the set of *optimal solutions*

$$X^* = \{x^* \mid f(x^*) = \min f(x) \ , \ x \in R^n\} \quad .$$

Note that $f(x)$ cannot be optimized directly because of the unknown parameters u^*. However, at iteration s we could obtain a statistical estimate u^s such that $u^s \to u^*$ with probability 1 and a sequence of functions $f(x,s) = \phi(x,u^s)$ such that

$$f(x,s) \to f(x)$$

with probability 1 for $s \to \infty$. The function $f(x,s)$ is available only at iteration s.

Consider the following procedure:

$$x^{s+1} = x^s - \rho_s f_x(x^s,s) \quad , \quad s = 0,1,\ldots \qquad (2)$$

This procedure, together with a procedure for calculating u^s, allows us to carry out the optimization while simultaneously estimating u^*. The principal difficulties associated with the convergence of procedure (1) are connected with the choice of

the *step-size* ρ_s. There is no guarantee that the new approximate solution x^{s+1} will belong to the domain of the smaller values of the functions $f(x,t)$ for $t \geq s+1$ (see Figure 1).

Convergence similar to that of (2) involving nondescent procedures has been studied within the framework of special nonstationary optimization problems in which it is assumed that the sequence of functions $\{f(x,s)\}_{s=0}^{\infty}$ and sets $\{X_s\}_{s=0}^{\infty}$ converges to some degree. It was shown in Ermoliev and Nurminski (1973) that under natural assumptions on the step-size sequence (such as $\rho_s \geq 0$, $\sum_{s=0}^{\infty} \rho_s = \infty$) for functions $f(x,s)$ convex with respect to x with the property $f(x,s) \to f(x)$, we have:

$$\lim f(x^s,s) = \min f(x) \quad .$$

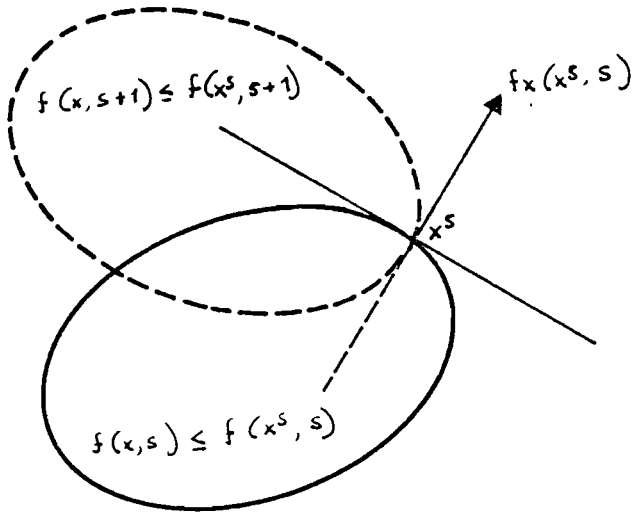

Figure 1.

III. GENERAL PROBLEM

Consider the problem of minimizing the function

$$f(x) = \phi(x,u^*) \tag{3}$$

subject to

$$x \in D(u^*) \tag{4}$$

where $x \in R^n$ is a vector of control (decision) variables and $u^* \in U \subseteq R^k$ is a vector of unknown parameters. Suppose that for an arbitrary given sequence of control variables $x^0,x^1,\ldots,x^s,\ldots$ it is possible to observe an ℓ-dimensional sequence $h^0,h^1,\ldots,h^s,\ldots$ such that

$$E\{h^s | x^0,x^1,\ldots,x^s\} = \psi(x^s,u^*) \quad , \tag{5}$$

where the function $\psi(x,u)$ is known. The problem is to create a sequence of control variables $\{x^s\}_{s=0}^{\infty}$ which minimizes the function $f(x)$ subject to given constraints. In more general cases the vector of unknown parameters may depend on time (i.e., on the iteration index s). We are therefore given a sequence of unknown k-dimensional parameters $u_s^* \in U \subseteq R^k$, $s=0,1,\ldots$ It is possible to observe an ℓ-dimensional sequence $h^0,h^1,\ldots,h^s,\ldots$ such that

$$E\{h^s | x^0,x^1,\ldots,x^s\} = \psi(x^s,u_s^*) \quad .$$

The required sequence $\{x^s\}_{s=0}^{\infty}$ has to minimize the functions $\psi(x,u_s^*)$ for $x \in X(u_s^*)$ in the sense that

$$\lim[\phi(x^s,u_s^*) - \min\{\phi(x,u_s^*) \,|\, x \in D(u_s^*)\}] = 0$$

for $s \to \infty$.

If a sequence of estimates u^s is found such that

$$\|u^s - u_s^*\| \to 0 \qquad (6)$$

for $s \to \infty$, then instead of functions $\phi(x,u_s^*)$ and sets $D(u_s^*)$ it is possible to consider the sequence of available functions $f(x,s) = \phi(x,u^s)$, sets $X_s = D(u^s)$ and the problem of finding a sequence $\{x^s\}$ such that

$$\lim[\phi(x^s,u^s) - \min\{\phi(x,u^s) \,|\, x \in D(u^s)\}] = 0 \quad . \qquad (7)$$

Before discussing a way of obtaining statistical estimates of u^s which satisfy (6), let us consider the iterative procedures for creating x^s such that it satisfies (7).

IV. THE SET OF ε-SOLUTIONS

The aim of $\{x^s\}$ is to track the set of optimal solutions

$$X_s^* = \{x^* \,|\, \phi(x^*,u^s) = \min\ \phi(x,u^s)\ ,\ x \in D(u^s)\} \quad .$$

Unfortunately the Hausdorf distance $d[X_s^*,X_{s+1}^*]$ between X_s^* and X_{s+1}^*, where

$$d[A,B] = \max\ \{\sup_{x\in A}\ \inf_{y\in B}\ \|x-y\|\ ,\ \sup_{x\in B}\ \inf_{y\in A}\ \|x-y\|\ \}\ ,$$

may be large even for small $\|u^{s+1} - u^s\|$. Therefore the distance between the current control variable x^s and X_{s+1}^* may increase rapidly compared with the distance between x^s and X_s^*. However, the Hausdorf distance between the sets of ε-solutions X_s^ε and X_{s+1}^ε,

$$X_s^\varepsilon = \{x^* \,|\, \phi(x^*,u^s) \le \min\ \phi(x,u^s) + \varepsilon\ ,\ x \in D(u^s)\}$$

satisfies the Lipschitz condition with respect to $\|u^{s+1} - u^s\|$ under reasonable assumptions when $\varepsilon > 0$. This fact was investigated and used in Ermoliev and Gaivoronski (1979) and Gaivoronski (1979) to study the convergence of procedures similar to (2).

To illustrate the basic idea more clearly, consider the case in which the feasible set $D(u^*)$ does not depend on the unknown parameters u^*. Denote the feasible set as X and let

$$X(u) = \{x(u) \mid \phi(x(u),u) = \min \phi(x,u) \ , \ x \in X\}$$

$$X^{\varepsilon}(u) = \{x \mid \phi(x,u) \leq \phi(x(u),u) + \varepsilon \ , \ x \in X\}$$

THEOREM 1. *Assume that*

(a) X *is a convex compact set;*

(b) $\phi(x,u)$ *is a convex continuous function with respect to* x *for all* $u \in U$ *and*

$$|\phi(x,u) - \phi(x,v)| \leq L \| u - v \| \tag{8}$$

for all $x \in X$, $u,v \in U$, *where* L *is a constant.*

Then

$$d[X^{\varepsilon}(u) \ , \ X^{\varepsilon}(v)] \leq \frac{2ML}{\varepsilon} \| u - v \|$$

where

$$M = \max \{ \| x - z \| \mid x \in X, z \in X\} \quad .$$

Proof. The set $X^{\varepsilon}(u)$ is compact. Therefore there are z', z" such that (see Figure 2):

$$d[X^{\varepsilon}(u), X^{\varepsilon}(v)] = \| z' - z" \| \quad . \quad .$$

Without loss of generality we could assume that:

$$\| z' - z" \| = \min \{ \| z" - x \| \mid x \in X^{\varepsilon}(u) \} \quad .$$

We have

$$\phi(z",u) - \phi(z',u) \geq \langle \hat{\phi}_x(z',u), z" - z' \rangle \quad ,$$

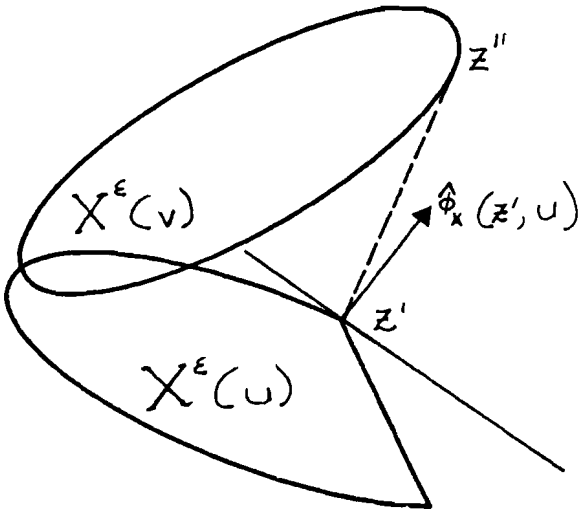

Figure 2.

where $\hat{\phi}_x(x,u)$ denotes a subgradient of the function $\phi(x,u)$ with respect to x. It is obvious that a $\hat{\phi}$ exists such that

$$\hat{\phi}_x(z',u) = \lambda(z'' - z') \quad ,$$

or

$$\phi(z'',u) - \phi(z',u) \geq \lambda \| z'' - z' \|^2 \quad ,$$

where $\lambda > 0$. Since

$$\varepsilon = \phi(z',u) - \phi(x(u'),u)$$

$$\leq \langle \lambda(z'' - z'), z' - x(u) \rangle$$

$$\leq \lambda M \| z'' - z' \| \quad ,$$

then

$$\lambda \geq \varepsilon / M \| z' - z'' \| \quad .$$

Therefore, for given $\hat{\phi}_x(z',u)$,

$$\langle \hat{\phi}_x(z',u) \mid z'' - z' \rangle = \lambda \| z'' - z' \|^2 \geq (\varepsilon/M) \| z'' - z' \|$$

or

$$\phi(z'',u) - \phi(z',u) \geq (\varepsilon/M) d[X^{\varepsilon}(u), X^{\varepsilon}(v)] \quad .$$

Since

$$\phi(z'',u) - \phi(z',u) = \phi(z'',u) - \phi(z'',v) + \phi(z'',v) - \phi(z',u)$$

$$\leq \phi(z'',v) - \phi(z',u) + L \| u - v \| \quad ,$$

then we will have

$$\phi(z'',v) - \phi(z',u) \geq (\varepsilon/M) d[X^{\varepsilon}(u), X^{\varepsilon}(v)] - L \| u - v \| \quad .$$

It is easy to see that

$$|\phi(z'',v) - \phi(z',u)|$$

$$= \left| \min_{x \in X} \phi(x,v) + \varepsilon - \min_{x \in X} \phi(x,u) - \varepsilon \right| \leq L \| v - u \| \quad .$$

Substituting this estimate into the previous inequality we obtain the desired result.

This theorem enables us to use many of the nondescent procedures discussed in Ermoliev (1976, 1981) to solve problem (7), and to prove the convergence of these procedures by studying the behavior of the distance between x^s and the set X_s^{ε}.

It should also be noted that this theorem clarifies the recently discovered Lipschitz continuity of the set of ε-subgradients for convex functions (Nurminski 1978; Hiriart-Urruty 1980). Indeed, suppose we have a convex function $q(u)$. The

subdifferential is

$$\partial q(u) = \operatorname{Arg} \min_x \phi(x,u)$$

where $\phi(x,u) = q(u) + q^*(x) - \langle x,u \rangle$, $q^*(x) = \min_u [q(u) - \langle x,u \rangle]$, and $\min_x \phi(x,u) = 0$.

On the other hand, from the definition of the ε-subdifferential $\partial_\varepsilon q(u)$ we have

$$\partial_\varepsilon q(u) = \{x \mid \phi(x,u) \leq \varepsilon\} = \{x \mid \phi(x,u) \leq \min_x \phi(x,u) + \varepsilon\} \quad .$$

V. NONSTATIONARY OPTIMIZATION PROCEDURES

Consider only the case in which the feasible set of the problem does not depend on unknown parameters and the operation of projection on the feasible set X is available. The nonstationary analog of the stochastic projection method has the form

$$x^{s+1} = \pi_X(x^s - \rho_s \xi^s) \quad , \quad s = 0,1,\ldots, \tag{9}$$

$$E\{\xi^s \mid x^0, x^1, \ldots, x^s\} = \hat{\phi}_x(x^s,u^s) + a^s \tag{10}$$

where the function $\phi(x,u)$ is considered to be convex continuous with respect to x; $\hat{\phi}_x(x^s,u^s)$ is a subgradient of $\phi(x,u)$ with respect to x; the step-size ρ_s may depend on the sequence of preceding approximations $(x^s,x^1,\ldots,x^s)$; and $u^s \in U$, where U is a compact set.

It should be noted that if $\phi(x,u)$ is differentiable with respect to x, $\xi^s = \phi_x(x^s,u^s)$, and $X = R^n$, then method (9) corresponds to method (2).

THEOREM 2. *Let the assumptions of Theorem 1 hold. Assume also that*

(a) $\| u^{s+1} - u^s \| \leq \delta_s$; $\delta_s/\rho_s \to 0$, $\| a^s \| \to 0$ *with probability* 1

for $s \to \infty$;

(b) $\rho_s > 0$, $\sum_{s=0}^{\infty} \rho_s = \infty$ *with probability 1,*

$$\sum_{s=0}^{\infty} E\rho_s^2 < \infty \ , \ E \| \xi^s \|^2 < \textit{Const.}$$

Then $$\lim[\phi(x^s,u^s) - \min \{\phi(x,u^s) \,|\, x \in X\}] = 0$$

with probability 1.

Proof. Let us set an $\varepsilon > 0$ and adopt the notation

$$w_s = d(x^s,X_s^\varepsilon) \ , \ X_s^\varepsilon = X^\varepsilon(u^s) \ , \ X_s^* = X(u^s)$$

All constants will be represented by the letter c.
In view of Theorem 1 and requirement (a), we have

$$w_{s+1} \leq (d[x^{s+1},X_s^\varepsilon] + d[X_s^\varepsilon,X_{s+1}^\varepsilon])^2 \leq d[x^{s+1},X_s^\varepsilon] + c(\delta_s+\delta_s^2) \ .$$

Let

$$\| \bar{x}^s - x^s \| = \min \{ \| x - x^s \| \,|\, x \in X_{s-1}^\varepsilon \} \ ,$$

$$\| \tilde{x}^s - x^s \| = \min \{ \| x - x^s \| \,|\, x \in X_s^\varepsilon \} \ ,$$

$$\Phi(u) = \min \{\phi(x,u) \,|\, x \in X\} \ .$$

The further evaluation of w_{s+1} yields:

$$
\begin{aligned}
w_{s+1} &\leq \| \bar{x}^{s+1} - x^{s+1} \|^2 + c(\delta_s + \delta_s^2) \leq \| \tilde{x}^s - x^{s+1} \|^2 \\
&+ c(\delta_s + \delta_s^2) \leq \| \tilde{x}^s - x^s + \rho_s \xi^s \|^2 + c(\delta_s + \delta_s^2) \\
&\leq w_s + c(\delta_s + \delta_s^2) + \rho_s^2 \| \xi^s \|^2 \\
&+ 2\rho_s \langle \xi^s - \hat{\phi}_x(x^s,u^s) - a^s \ , \ \tilde{x}^s - x^s \rangle + c\rho_s \| a^s \| \\
&- 2\rho_s [\phi(x^s,u^s) - \Phi(u^s) - \varepsilon] \quad ,
\end{aligned}
$$

where the inequality

$$
\langle \hat{\phi}_x(x^s,u^s) \ , \ \tilde{x}^s - x^s \rangle \leq \phi(x^s,u^s) - \Phi(u^s) - \varepsilon \quad ,
$$

was used. Therefore, we will also have

$$
\begin{aligned}
w_k \leq w_s &- 2 \sum_{r=s}^{k-1} \rho_r [\phi(x^r,u^r) - \Phi(u^r) - \varepsilon - c \| a^r \| - c\delta_r/\rho_r - c\delta_r^2/\rho_r] \\
&+ \sum_{r=s}^{k-1} \rho_r^2 \| \xi^r \|^2 + 2 \sum_{r=s}^{k-1} \rho_r \langle \xi^r - \hat{\phi}_x(x^r,u^r) - a^r \ , \ \tilde{x}^r - x^r \rangle \quad .
\end{aligned}
$$

From condition (10) and the martingale convergence theorems it follows that

$$
\sum_{r=s}^{k-1} \rho_r \langle \xi^r - \hat{\phi}_x(x^r,u^r) - a^r \ , \ \tilde{x}^r - x^r \rangle \to 0
$$

with probability 1 for $s \to \infty$. From condition (b) we have that

$$\sum_{r=s}^{k-1} \rho_r^2 \| \xi^r \|^2 \to 0 \quad ,$$

with probability 1 for $s \to \infty$. Therefore

$$w_k \leq w_s - 2 \sum_{r=s}^{k-1} \rho_r [\phi(x^r,u^r) - \phi(u^r) - \varepsilon - c(\| a^r \| + \delta_r/\rho_r + \delta_r^2/\rho_r)] + \gamma_s \tag{11}$$

where $\gamma_s \to 0$ with probability 1 for $s \to \infty$.

We shall now prove that $w_s \to 0$ with probability 1.

Suppose that there exist s' and $\Delta > 0$ such that $w_s > \Delta$ for $s > s'$. Then, from the continuous dependence of $X^\varepsilon(u)$ on $u \in U$ and the compactness of U, it follows that there exists an $\alpha > 0$ such that

$$\phi(x^s,u^s) - \Phi(u^s) - \varepsilon > \alpha \tag{12}$$

for $s > s'$. Substituting this into the previous inequality we obtain

$$w_k \leq w_s - 2\alpha \sum_{r=s}^{k-1} \rho_r [1 - \varepsilon_r] + \gamma_s \quad ,$$

where

$$\varepsilon_r = c(\| a^r \| + \delta_r/2\alpha\rho_r + \delta_r^2/2\alpha\rho_r) \quad .$$

From condition (b) the $\varepsilon_r \to 0$ with probability 1 for $r \to \infty$. Then, bearing in mind that $\sum_{s=0}^{\infty} \rho_s = \infty$, we obtain a contradiction when w_k is positive. Choose an arbitrary $\Delta > 0$ and suppose that $w_{s_r} < \Delta$, but that there is a number t_r, $s_r < t_r < s_{r+1}$, such that $w_{t_r} > 3\Delta$. From (11), it follows that for $k = s+1$

$$\max \{0 \ , \ w_{s+1} - w_s\} \to 0$$

with probability 1. Therefore, for sufficiently large r there is a number τ_r such that $s_r \leq \tau_r < t_r$, $w_{\tau_r} < 2\Delta$, and $w_s \geq \Delta$ for $\tau_r \leq s \leq t_r$. Since inequality (12) holds (for a certain number α) if $\tau_r \leq s \leq t_r$, then from (11) for $k = t_r$ and $s = \tau_r$, we obtain

$$w_{t_r} \leq w_{\tau_r} - 2\alpha \sum_{\ell=\tau_r}^{t_r-1} \rho_\ell (1 - \varepsilon_\ell) + \gamma_{\tau_r} \quad .$$

If we now choose a value of r large enough that $\gamma_{\tau_r} < \Delta$, $\varepsilon_\ell < 1$ for $\ell \geq \tau_r$, then $w_{t_r} \leq 3\Delta$, which contradicts the assumption that $w_{t_r} > 3\Delta$. Therefore, $w_s \to 0$ with probability 1 for $s \to \infty$. From this and from the inequality

$$\phi(x^s, u^s) - \min \{\phi(x, u^s) \mid x \in X\} \leq \varepsilon + c \min_{y \in X_s^\varepsilon} \| y - x^s \| \quad ,$$

the theorem is proved.

It should be noted that algorithm (9) is also applicable when $\| u^{s+1} - u^s \|$ and ρ_s do not approach zero.

THEOREM 3. *Assume that instead of requirements (a) and (b) of Theorem 2, the following conditions are satisfied:*

(a) $\| u^{s+1} - u^s \| \leq \delta_s$, $\overline{\lim}\ \delta_s = \delta > 0$ *for* $s \to \infty$;

(b) $\| a^s \| \to 0$ *with probability 1;*

(c) $\rho_s = \rho > 0$, $\varepsilon > 0$ *and*

$$0 < \gamma = 2(\rho\varepsilon - M\kappa)/M^2 \leq 1 \quad ,$$

where

$$\kappa = 2M\delta L/\varepsilon, \quad M = \max \{ \| x - y \| \,|\, x,y \in X \} \quad .$$

Then

$$\overline{\lim}\ E \min \{ \| x - x^{s+1} \| \,|\, x \in X_s^{\varepsilon} \} \leq q/\gamma \quad ,$$

$$q = \kappa^2 + 4\rho\delta L + c\rho^2 \quad ,$$

where c is a constant.

The above theorem demonstrates that the sequence $\{x^s\}_{s=0}^{\infty}$ will, on the average, be sufficiently close to the set of ε-solutions, provided that the choice of step-size ρ and the drifts of the u^s are reasonable. We should note that this condition may be satisfied by increasing the number of iterations taking place within unit time.

Gaivoronsky (1979) has given a number of other algorithms for solving nonstationary optimization problems with constraints of a general form. However, even the simple algorithms described above may serve as the basis for the numerical solution of many important classes of practical problems. Special classes of nonstationary optimization problems have been discussed by Dupač (1965), Tsypkin (1971), Fujitas and Fukao (1972), Uosaki (1974), and Eremin (1979).

VI. ESTIMATION PROCEDURES

Nonstationary optimization procedures similar to (9) allow us to carry out optimization and estimation simultaneously, if we have a simple iterative scheme for calculating the estimates u^s which satisfy (6). A useful method of creating an iterative estimation procedure is to rewrite the estimation problem as an optimization problem and then to use iterative optimization methods similar to (9).

For instance, in the simple case of Section II, if an observation h^s of the random vector h is available at iteration s, such that

$$Eh = u^* \quad ,$$

then the required vector u^* minimizes the function

$$r(u) = E \| u - h \|^2$$

because $u = Eh$ satisfies the optimality conditions

$$r_u(u) = 2(u - Eh) = 0 \quad . \tag{13}$$

If *a priori* knowledge about the unknown u^* is introduced as $u \in U$, then we could use the following stochastic projection method to minimize function (13) (see, for instance, Ermoliev 1976, 1981):

$$u^{s+1} = \pi_U[u^s - \delta_s(u^s - h^{s+1})], \quad s = 0,1,\ldots, \tag{14}$$

where δ_s is the step-size, which may depend on $(u^0,u^1,\ldots,u^s)$, and h^s is the observation of h. If $\delta_s \geq 0$, $\sum_{s=0}^{\infty} \delta_s = 0$ with probability 1, $\sum_{s=0}^{\infty} E\delta_s^2 < \infty$, and the set U is convex compact, then $u^s \to u^*$ with probability 1. Ermoliev and Gaivoronski (1979) noted a number of advantages of estimates obtained via iterative optimization procedures (in addition to the opportunity for on-line calculations). Firstly, δ_s may be chosen to be a function of

$(u^0,u^1,\ldots,u^s)$ in order to decrease the value of the objective function. Secondly, *a priori* knowledge about the unknown u^* may be taken into account in terms of constraints. In this case, a current estimate u^s would have the property $u^s \in U$ for all $s = 0,1,\ldots$, whereas a conventional estimate $\tilde{u}^s$ would normally only fulfill $\lim \tilde{u}^s \in U$ for $s \to \infty$. Therefore the estimates u^s are generally better for *small samples*.

In the more general case when the vector of observations h satisfies the condition

$$E\{h|x\} = \psi(x,u^*) \quad ,$$

the true vector u^* minimizes the function

$$r(x,u) = E \| \psi(x,u) - h \|^2$$

with respect to u for each feasible x. However, there may be unnecessary solutions. Since

$$r(x,u) - r(x,u^*) = \| \psi(x,u) - \psi(x,u^*) \|^2 \quad ,$$

then for the solution $u = u^*$ minimizing r(x,u) with respect to u to be unique it is necessary to assume that the equations

$$\psi(x,u) = \psi(x,u^*) \quad , \quad x \in X$$

represent the unique solution $u = u^*$.

This requirement can often be relaxed as follows. Consider the sequence of functions (for the given sequence of control variables $x^0,x^1,\ldots x^s,\ldots$):

$$g(u,s) = \int_{Y_s} r(x^s + y,u) \; P_s(dy) = E\{r(x^s + y^s,u) \,|\, x^s\} \quad ,$$

where the probabilistic measure $P_s(dy)$ is distributed on a domain Y_s and centered at the point 0 for $s \to \infty$. For instance, Y_s could be given by

$$Y_s = \{y = (y_1, y_2, \ldots, y_n) \mid -\Delta_s \leq y_j \leq \Delta_s \ , \ j = \overline{1,n}\} \ ,$$

and $P_s(dy)$ is used to generate the random vector y^s with independent components distributed uniformly over the interval $[-\Delta_s, \Delta_s]$, where Δ_s is a positive number, and $\Delta_s \to 0$ for $s \to \infty$.

The true vector u^* minimizes the function $g(u,s)$ for each $s = 0,1,\ldots$, such that

$$g(u,s) - g(u^*,s) = \int_{Y_s} \| \psi(x^s + y, u) - \psi(x^s + y, u^*) \| P_s(dy) \ .$$

Therefore there may be a unique solution to the problem of minimizing $g(u,s)$ even if the minimization of $r(x^s,u)$ with respect to u does not possess this property.

We could use a procedure similar to (9) to minimize $g(x,s)$. Assume that $g(u,s)$ is a convex continuous function with respect to u for all s; U is a convex compact set.

Consider the procedure

$$u^{s+1} = \pi_U(u^s - \delta_s \zeta^s) \qquad , \qquad s = 0,1,\ldots,$$

$$E\{\zeta^s \mid x^0, u^0, \ldots, x^s, u^s\} = \hat{g}_u(u^s, s) + D^s \ ,$$

where $\hat{g}_u$ is a subgradient of function $g(u,s)$.

For example, we can consider the function

$$\zeta^s = 2 \sum_{i=1}^{\ell} [\psi_i(x^s + y^s, u^s) - h_i^s] \psi_{iu}(x^s + y^s, u^s)$$

where $\psi_i(\cdot,u)$ are differentiable functions and $h_s = (h_1^s, h_2^s, \ldots, h_\ell^s)$ is an observation of the vector h at $x = x^s + y^s$ such that:

$$E\{h^s | x^0, x^1, \ldots, x^s\} = \psi(x^s + y^s, y^*) \quad .$$

It is easy to see that

$$E\{\zeta^s | x^s, u^s\} = g_u(u^s, s) \quad .$$

THEOREM 4. *(See Ermoliev and Gaivoronski, 1979). Assume that the above condition holds and that*

$$g(u,s) - g(u^*,s) \geq \lambda_s \gamma(u,u^*) \quad , \tag{15}$$

where $\lambda_s \geq 0$, $\gamma(u,u^*) \geq 0$ *and* $\gamma(u,u^*) = 0$ *only for* $u = u^*$; *the step-size* δ_s *may depend on* $(x^0, x^1, \ldots, x^s)$, *and* $\delta_s \geq 0$, $\sum_{s=0}^{\infty} \lambda_s \delta_s = \infty$ *with probability 1, and also*

$$\sum_{s=0}^{\infty} E\{\delta_s \| b^s \| + \delta_s^2\} < \infty \quad , \quad E \| \zeta^s \|^2 < \text{Const.}$$

Then $u^s \to u^*$ *with probability 1.*

The proof of this theorem is similar to the proof of Theorem 2. We shall now comment on condition (15).

Consider the important case

$$\Psi(x,u) = A(x)u \quad ,$$

where A(x) is a matrix. Then

$$g(u,s) - g(u^*,s) = \int_{Y_s} \| A(x^s + y)(u - u^*) \|^2 P_s(dy)$$

$$\geq c\lambda_s \| u - u^* \|^2 \quad ,$$

where λ_s is the smallest eigenvalue of the matrix

$$\int_{Y_s} A(x^s + y)A^T(x^s + y)\ P_s(dy) \quad .$$

Therefore, in this case, $\gamma(u,u^*) = \| u - u^* \|$.

VII. SIMULTANEOUS OPTIMIZATION AND APPROXIMATION PROCEDURES

Consider the problem of minimizing a differentiable function $f(x)$ in a set X. Suppose that a sequence of solution approximations $x^0, x^1, \ldots, x^s, \ldots$, is constructed according to the following rule:

Let

$$D_s = \{X = (x_1, x_2, \ldots, x_n) \mid x_i^s - \tau_s \leq x_i \leq x_i^s + \tau_s \quad , \quad i = \overline{1,n}\} \quad ,$$

where τ_s is a number. Let $\Psi(x,a)$ be convex functions with respect to x, parametrized by a parameter $a \in A$. Let these functions approximate the function $f(x)$ in the sense of minimizing the criteria

$$\Phi(a,s) = \int_{D_s} (f(x) - \Psi(x,a))^2\ P_s(dx)$$

where $P_s(dx)$ is a Borell measure. If we assume that $P_s(dx)$ is a probabilistic measure, we can then rewrite the above equation as

$$\Phi(a,s) = E(f(x_s + h) - \Psi(x_s + h, a))^2$$

where $h = (h_1, h_2, \ldots, h_n)$ is a random vector. For simplicity, we assume further that the components of h are independent and uniformly distributed over $[-\tau_s, \tau_s]$. Therefore

$$\Phi(a,s) = \frac{1}{(2\tau_s)^n} \int_{-\tau_s}^{\tau_s} \cdots \int_{-\tau_s}^{\tau_s} (f(x^s + h) - \Psi(x^s + h, a))^2 dh \quad .$$

The choice of the point x^{s+1} is based on the requirement that approximation $\Psi(x,a^s)$ should be minimized:

$$x^{s+1} = \pi_X(x^s - \rho_s \hat{\Psi}_x(x^s,a^s))$$

where π_X is the projection operator on X and $\hat{\Psi}_x(x^s,a^s)$ is a subgradient of $\Psi(x,a^s)$ with respect to x at x^s,a^s. Moreover, the next value of the parameter a is determined by

$$a^{s+1} = \Pi_A(a^s - \delta_s \zeta^s)$$

where ζ^s has the property that

$$E\{\zeta^s | x^0,a^0,\ldots,x^s,a^s\} = \Phi'_a(a^s,s) \quad .$$

For instance, we can choose ζ^s to be defined by:

$$\zeta^s = -2[f(\tilde{x}^s) - \Psi(\tilde{x}^s,a^s)]\Psi_a(\tilde{x}^s,a^s) \quad ,$$

where $\tilde{x}^s = x^s + h^s$ and $\{h^s\}$ are observations of $h = (h_1,h_2,\ldots,h_n)$, $h_i \in [-\tau_s,\tau_s]$.

Consider the following assumptions:

(a) $\Psi(x,a)$ is a convex continuous function with respect to x,a and differentiable with respect to a; f(x) is a differentiable function;

(b) X,A are convex compact sets;

(c) there exist a set $X^1 \subset X$ and an element $z \in X^1$ such that

$$f(x) > f(z) \ , \ \forall\, x \notin X^1 \quad ;$$

(d) for any $\varepsilon > 0$ there exist $\Delta > 0$ and $\tau > 0$ such that

$$\| f_x(x) - \hat{\Psi}_x(x,a) \| < \Delta$$

for all a such that

$$a \in A_\varepsilon(s) = \{a \in A \mid \Phi(a,s) - \min_{a \in A} \Phi(a,s) < \varepsilon\} \ , \ \tau_s < \tau \ ,$$

$$x \in X_\varepsilon(s) = \{x \in X^1 \mid \| f_x(x) \| > \varepsilon\} \ .$$

THEOREM 5. *Let the above assumptions hold and let the nonnegative parameters* ρ_s, δ_s, τ_s *satisfy the conditions:*

(i) ρ_s, δ_s, τ_s *are* $(x^0, a^0, \ldots, x^s, a^s)$*-measurable functions;*

(ii) $\tau_s \to 0$, $\rho_s/\delta_s \to 0$, $|\tau_{s+1} - \tau_s|/\tau_s\delta_s \to 0$ *with probability 1;*

(iii) $\sum_{s=0}^{\infty} \rho_s = \infty$, $\sum_{s=0}^{\infty} \delta_s = \infty$, $\sum_{s=0}^{\infty} (\rho_s^2 + \delta_s^2) < \infty$;

(iv) $\| \xi^s \| < C < \infty$.

Then with probability 1:

$$\lim_{s \to \infty} \min_{y \in X} \langle f_x(x^s) \ , \ x^s - y \rangle = 0 \ . \tag{16}$$

Let $X = R^n$, A be a convex compact set, and let the assumptions (a), (c), (i)-(iv) hold. Assume instead of (d) that:

(d') for any $\varepsilon > 0$ there exist $\Delta > 0$ and $\tau > 0$ such that:

$$\langle f_x(x) \ , \ \hat{\Psi}_x(x,a) \rangle > \Delta$$

for $a \in A_\varepsilon(s)$, $x \in X_\varepsilon(s)$, $\tau_s < \tau$.

THEOREM 6. *Let the assumptions (a)-(c), (d'), (i)-(iv) hold and assume also that there is a compact set K such that:*

$$x^s \in K \quad , \qquad s = 0,1,\ldots, \tag{17}$$

Then

$$\lim_{s\to\infty} \| f_x(x^s) \| = 0 \qquad \text{a.c} \tag{18}$$

Note that requirement (17) is not too stringent for the applications. It would be satisfied if, for example, we chose appropriate functions $\Psi(x,a)$. The requirement (iv) is satisfied if, for example, (17) holds and the random variables are bounded.

The proof can be outlined as follows:

1. First, as in Theorem 2, we prove that with probability 1:

$$\lim [\Phi(a^s,s) - \min_{a\in A} \Phi(a,s)] = 0 \qquad \text{for } s \to \infty \tag{19}$$

This can be done in the following way. Consider

$$w(a^s) = \min_{a\in A_\varepsilon(s)} \| a^s - a \|^2 = \| a^s - a^*(s) \|^2 \quad .$$

It can be proved that

$$w(a^{s+1}) \leq \| a^{s+1} - a^*(s) \|^2 + c\gamma_s + \gamma_s^2$$

where $c < \infty$, $\varepsilon > 0$, and γ_s is the Hausdorf distance between $A_\varepsilon(s)$ and $A_\varepsilon(s+1)$. The quantity γ_s can be estimated as follows:

$$\gamma_s \leq c^* \cdot \max \{\rho_s, |\tau_{s+1} - \tau_s|/\tau_s\}$$

where $c^* < \infty$ if $\varepsilon > 0$. Thus, according to (ii), $\gamma_s/\delta_s \to 0$. Then

$$w(a^{s+1}) \leq \| -\delta_s \zeta^s + a^s - a^*(s) \|^2 + c\gamma_s + \gamma_s^2 \leq w(a^s)$$

$$- 2\delta_s \langle \zeta^s, a^s - a^*(s) \rangle + \delta_s^2 \| \zeta^s \|^2 + c\gamma_s + \gamma_s^2$$

$$= w(a^s) - 2\delta_s \langle \phi_a'(a^s,s) , a^s - a^*(s) \rangle$$

$$+ c\gamma_s + \gamma_s^2 + \delta_s^2 \| \zeta^s \|^2 - 2\delta_s \langle \zeta^s - \Phi_a'(a^s,s) , a^s - a^*(s) \rangle .$$

It can be shown that

$$\langle \Phi_a'(a^s,s) , a^s - a^*(s) \rangle \geq c(\sqrt{w(a^s)} + \upsilon^s) \; ; \; \upsilon^s \to 0 \qquad \text{a.c.}$$

where c depends on ε and $c > 0$ if $\varepsilon > 0$, and that

$$\left| \sum_{s=0}^{\infty} \delta_s \langle \zeta^s - \Phi_a'(a^s,s) , a^s - a^*(s) \rangle \right| < \infty$$

with probability 1.

Therefore:

$$w(a^s) \leq w(a^\ell) - c_1 \sum_{k=\ell}^{s-1} \delta_k [\sqrt{w(a^k)} - \upsilon^k - c_2(\gamma_s/\delta_s + \gamma_s^2/\delta_s)] + \beta_\ell$$

where $\beta_\ell \to 0$ a.c. for $\ell \to \infty$, $c_1 < \infty$, $c_2 < \infty$. Since $\gamma_s/\delta_s \to 0$, we obtain

$$w(a^s) \leq w(a^\ell) - c_1 \sum_{k=\ell}^{s+1} \delta_k (\sqrt{w(a^k)} - \upsilon^k) + \beta_\ell \quad ,$$

where $\upsilon_k \to 0$ a.c. for $k \to \infty$.

Hence we can show that $w(a^s) \to 0$ a.c. for $s \to \infty$ which implies that (19) is true.

2. Now we can prove the convergence results (16) and (18). Consider, for instance, the result (18). We have

$$f(x^{s+1}) - f(x^s) = \langle f_x(x^s + \lambda(x^{s+1} - x^s)) \ , \ x^{s+1} - x^s \rangle$$

$$\leq -\rho_s \langle f_x(x^s) \ , \ \Psi_x(x^s, a^s) \rangle + \nu(s)\rho_s \quad , \quad (20)$$

where $\nu(s) \to 0$ a.c., $\lambda \in [0,1]$.

Suppose that there is an $\varepsilon > 0$ and a number m such that $\| f_x(x^s) \| > \varepsilon$ for $s > m$. Then from (d') and (19) it follows that

$$f(x^{s+1}) - f(x^s) < -\rho_s \Delta + \nu(s)\rho_s$$

and for $s > m$

$$f(x^s) < f(x^m) - \Delta \sum_{i=m}^{s-1} \rho_i \left(1 - \frac{\nu(s)}{\Delta}\right) \quad .$$

From (c) and (iii), this contradicts the assumption that $\| f_x(x^s) \| > \varepsilon$.

There therefore exist subsequences $\{x^{s_k}\}$ such that

$$\| f_x(x^{s_k}) \| \to 0 \quad .$$

It is now easy to obtain the result (18) from $\| x^s - x^{s+1} \| \to 0$ and (20).

These results can be generalized for problems with non-differentiable objective functions and constraints of a general form.

ACKNOWLEDGMENTS

The authors would like to thank Andrzej Wierzbicki, Roger Wets, and Michael Dempster for useful comments.

REFERENCES

Dupač, V. (1965) A dynamic stochastic approximation method. Annals of Mathematical Statistics, 6.

Eremin, I.T. (1979) Standard iterative processes of nonsmooth optimization for nonstationary problems of convex programming. USSR Computational Mathematics and Mathematical Physics, 1.

Ermoliev, Yu. M. (1976) Stochastic Programming Methods. Moscow: Nauka.

Ermoliev, Yu. M. (1981) Stochastic quasigradient methods and their application in systems optimization. WP-81-02, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Ermoliev, Yu. M. and E.A. Nurminski (1973) Limit extremal problems. Kibernetika, 4.

Ermoliev, Yu. M. and A.A. Gaivoronski (1979) Stochastic optimization and simultaneous parameter estimation. Izvestia Academii Nauka SSSR, Technischeskaj Kibernetika, 4.

Fujitas, S. and T. Fukao (1972) Convergence conditions of dynamic stochastic approximation method for nonlinear stochastic discrete-time dynamic systems. IEEE Transactions on Automatic Control, 17.

Gaivoronski, A.A. (1979) Study on nonstationary problems of stochastic programming. Abstract of dissertation. Institute of Cybernetics, Kiev, USSR.

Hiriart-Urruty, J.-B. (1980) ε-Subdifferential calculus: Proceedings of the Colloquium "Convex Analysis and Optimization". Imperial College, London, 28-29 February 1980.

Katkovnik, V.Ja. and V.E. Khejsin (1976) Iterative optimization algorithms for tracking extremum drift. USSR Automatika i Vychislitelnaja Teknika, 6.

Nurminski, E.A. (1977) The problem of nonstationary optimization. Kibernetika, 2.

Nurminski, E.A. (1978) Nondifferentiable optimization with ε-subgradient methods. WP-78-55, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Tsypkin, Ya.Z. (1971) Adaptation and Learning in Automatic Systems. New York: Academic Press.

Uosaki, K. (1974) Some generalizations of dynamic stochastic approximation procedures. Annals of Statistics, 2.

Vertchenko, P.I. (1977) Limit extremum problems of stochastic optimization. Abstract of dissertation. Institute of Cybernetics, Kiev, USSR.