ESTIMATION AND VARIATIONAL METHODS FOR GRADIENT ALGORITHM GENERATION

by

Paul Michel Toldalagi

Ingenieur Civil des Telecommunications

Paris - 1975.

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 1977

Signature of Author . . . . . .
Department of Electrical Engineering
and Computer Science, May 12, 1977.

Certified by . . . . . .
Thesis Supervisor

Accepted by . . . . . .
Chairman, Departmental Commitee on Graduate Students

ESTIMATION AND VARIATIONAL METHODS FOR GRADIENT ALGORITHM GENERATION

by

Paul Michel Toldalagi

Submitted to the Department of Electrical Engineering and Computer Science on May 16, 1977 in partial fulfillment of the requirements for the Degree of Master of Science.

ABSTRACT

This study contains a new approach to the unconstrained minimization of a multivariable function f(.) using a quasi-Newton step computation procedure. The whole problem is reformulated as the control problem of a linear system described by its state-space equations and having unknown dynamical properties. First of all, an adaptive identification problem arises and is solved by using set estimation concepts. The resulting dynamics contain in particular an estimate of the Hessian matrix of f(x), matrix which is used to regulate the system to zero. Some matrix symmetrization methods are also studied and finally used for generating a sequence of steps $x_{k+1} - x_k$ by the classical Newton method.

THESIS SUPERVISOR: Sanjoy K. Mitter

TITLE: Professor of Electrical Engineering and ComputerScience

ACKNOWLEDGEMENT

I wish to express my gratitude to Professor Sanjoy K. Mitter for his supervision and stimulating criticisms, which were crucial for the success of the present thesis.

Thanks are due also to Professor Louis F. Pau from the Ecole Nationale Superieure des Telecommunications for his friendly encouragements, as well as to Mr. Alain Bousquet, one of his students, for his help in carrying out some of the early computational tests.

I would also like to thank my friend and colleague, Khaled Yared, for his constant support.

Finally, I am grateful to my wife, Marianne, for her help.

TABLE OF CONTENTS

CHAPTER I

Introduction

The gradient methods in Optimization have been of considerable theoretical as well as practical interest for well over a decade.All of them consist primarily of building a sequence of points (x_k) using the gradient of the function f(x), which is to be minimized. Different procedures arise in the literature, among them the oldest and most interesting ones are the class of quasi-Newton methods. A particularly critical discussion of the evolution of the concept of quasi-Newton algorithms can be found in the introduction to S.W. Thomas' thesis [28]. In particular, he points out how loosely this terminology has been used for a large variety of different algorithms. Thus, in the present thesis, we shall understand it in the following sense. Let f(.) be a differentiable function from R^n to the real line also referred to as the objective function, and let $\nabla f(x) = g(x)$ be its gradient. We shall say that the iteration procedure for computing a sequence (x_k) converging to the minimum x^* of f(.), is a quasi-Newton procedure , if

$$x_{k+1} - x_k = - B_k^{-1} g(x_k) \qquad \text{for all } k = 0,1,\ldots .$$

The sequence (B_k) is a sequence of matrices of order n having generally an interpretation in terms of the actual Hessian matrix of f(x).

Curiously, it appears that throughout the technical literature on this subject within the last decade, the common feature of all such gradient algorithms was to determine the sequence of matrices (B_k , k=0,1,..) by construction, after having defined some desirable algebraic property. The case of Huang's work is typical in this sense - see [10] - .

In the present thesis, a new point of view is introduced. The idea is basically the following one: in most of the quasi-Newton algorithms, the matrix B_k is given an interpretation in terms of the local Hessian matrix of the objective function. Thus, the quasi-Newton methods become the original Newton method as soon as the B_k reaches the value of the actual Hessian matrix of the objective. This gives rise to a very fast convergence in a neighborhood of the optimum. Consequently, our aim is to estimate the local Hessian from the last points and gradients of the function, by introducing some random or unknown but bounded quantities. In fact, this observation appears in S.W. Thomas' work [28], although he did not fully exploit it, but returned rather to a more conventional approach, close to Broyden's method.

Thus, the algorithm is considered as a dynamical system described by a set of state-space equations and unknown inputs. Since, some parts of the dynamics of this system contain unknown parameters, an adaptive identification problem is solved in order to estimate them recursively. The Hessian of the objective function appears to be one of them. In a second period, a regulation problem is solved since the ultimate goal is to force the gradient of the objective function to zero. The resulting regulator will simply use the output of the previous identification

process in order to perform its task.

In Chapter II, a study of different matrix symmetrization methods is presented. Different matrix norms are introduced and their relationship studied. Thereupon, a variational point of view is introduced to find the symmetric solution of the simple algebraic problem,

$$b = X\,a$$

where a and b are known n-vectors and X is an n x n unknown matrix. Finally, the previous solution is generalized to the case where an expansion of the previous solution around any given n x n matrix X^0 is desired. The results of this chapter will, therefore, prove to be useful when studying the previously introduced regulation problem.

Chapter III deals with the construction of an appropriate state-space model for the initial minimization problem. All unknown quantities appearing as input terms of our system are modelled as Gaussian random variables, although the original problem is perfectly deterministic by nature. A filter is constructed at each step in order to estimate the Hessian $G(x_k) = G_k = \nabla^2 f(x_k)$ of the objective function. A recursive procedure is given to propagate directly $H_k = G_k^{-1}$ and, eventually, the initial problem of minimizing f(.) is understood as a stochastic regulation problem.

Chapter IV studies the same dynamical system but, instead of assuming an a priori knowledge of the statistics of the input variables, these quantities will be assumed to be constrained only to some finite ellipsoid-shaped domains. A short review is done about the basic tools needed for understanding the results of set estimation theory and three types of estimation problems are recognized. Finally, a recursive

solution for estimating G_k is presented, its similarity with the minimum mean squares estimate being briefly emphasized.

Chapter V presents the main articulations of the algorithm proposed in this thesis. In particular, some new results are proved about the convergence properties of such an algorithm, and some singularity problems are also analyzed. Finally a short description is given about some necessary tricks, which were actually used when implementing this algorithm.

CHAPTER II

Matrix symmetrization methods under linear constraints.

I-Introduction and statement of the problem.

Consider the following problem:

Problem 0:

Solve the equation $b=Xa$,where b and a are known n-dimensional vectors in R^n and X is some n x n symmetric real matrix,element of $L(R^n)$.

Clearly,this problem is highly underdetermined,especially if n is large,as it contains $\frac{n(n-I)}{2}$ unknown variables-i.e. as many,as there are different coefficients in a symmetric n x n matrix-,for only n equations.

An other way of looking at this problem is to exploit the one to one correspondence existing between elements of $L(R^n)$ and R^{n^2} .Thus, assume that $X=\begin{bmatrix} x_I \\ . \\ x_n \end{bmatrix}$ is an element of $L(R^n)$, then,there exists $\underline{X} \in R^{n^2}$ and $\underline{A} \in L(R^{n^2},R^n)$ such that:

$$b = Xa = \underline{A}\underline{X} \qquad [1]$$

with

$$\underline{A} = \begin{bmatrix} a^T & & 0 \\ & \cdot & \\ 0 & & a^T \end{bmatrix} \qquad [2]$$

and

$$\underline{X} = \begin{bmatrix} a_I^T \\ a_i^T \\ a_n^T \end{bmatrix} \quad \varepsilon \quad R^{n^2} \qquad [3]$$

Now, as X must be symmetric, let S be the subset of R^{n^2} such that:

$$S = \{ \underline{Z} \ \varepsilon R^{n^2} \quad \text{s.t.} \quad Z = Z^T \quad \text{element of } L(R^n) \}$$

then,

Claim I:

S is a linear subspace of R^{n^2} of dimension $\frac{n(n-I)}{2}$.

Proof:

Let Z_I and Z_2 be symmetric matrices. Then, $\underline{Z}_I$ and $\underline{Z}_2$ belong to S and $\alpha_I Z_I + \alpha_2 Z_2$ remains symmetric for any real coefficients α_I and α_2, hence $\alpha_I \underline{Z}_I + \alpha_2 \underline{Z}_2$ S and S is a linear subspace of R^{n^2} . Its dimension is clearly $\frac{n(n-I)}{2}$ as $\frac{n(n-I)}{2}$ coefficients are sufficient to determine uniquely any n x n symmetric matrix. Q.E.D

Consider now in R^{n^2} the set Δ consisting of all possible solutions to the equation b=Xa. Then:

Claim 2:

Let $\Delta = \{ \underline{Z} \ \varepsilon \ R^{n^2}$ such that $b = \underline{A}\ \underline{X}$ for $b \varepsilon \ R^n, \underline{A} \ \varepsilon \ L(R^{n^2}, R^n) \}$

be the set of all solutions of $b=\underline{A}\,\underline{X}$ .Then Δ is a convex subset of R^{n^2}.

Proof:

For $\underline{Z}_I$ and $\underline{Z}_2$ elements of Δ , $b=\underline{A}\,\underline{Z}_I=\underline{A}\,\underline{Z}_2$.Choose then any element $\alpha \in [0,I]$,therefore $\underline{Z}_3=\alpha\,\underline{Z}_I+(I-\alpha)\,\underline{Z}_2$ .Clearly this means also that

$$\underline{A}\,\underline{Z}_3 = \alpha\underline{A}\,\underline{Z}_I + (I-\alpha)\,\underline{A}\,\underline{Z}_2 = b$$

and that $\underline{Z}_3$ belongs to Δ which is itself included in R^{n^2}. Q.E.D

The goal of Problem 0 is to find any element in $\Delta \cap S$.Generally the set Δ will have as dimension n^2 ,whereas S has dimension $\frac{n(n-I)}{2}$: this means that Problem 0 will have an infinite number of solutions, all belonging to $\Delta \cap S$,of dimension smaller or equal to $\frac{n(n-I)}{2}$.

In order to attribute to Problem 0 a more restrictive meaning, a minimal norm condition is introduced.This is achieved in Section 2, where first a quick review is done on some matrix norm candidates. Finally a new formulation of Problem 0 is given and henceforth is refered to as Problem I.

In Section 3 a variational approach is used to find the solution of Problem I.

In Section 4,the recursive symmetrization procedure due to Powell is shortly discussed and compared to the previous result.

In Section 5,Problem I is slightly modified.Instead of looking for the"absolute minimum norm" solution of b=Xa,we shall be interested in discovering the symmetric solution "closest" to any given possibly

non-symmetric matrix. Finally, a geometrical interpretation of these results is given, using the properties of the subspaces Δ and S of R^{n^2}.

2-Matrix norms and reformulation of the initial problem.

Let A be some matrix, element of $L(R^n)$.

$$||A|| = \underset{x \in R^n}{\text{Sup}} \left\{\frac{x^T A x}{x^T x}\right\} \qquad [4]$$

As $||A||$ is a scalar and as $x^T A x = x^T A^T x$, it becomes clear that for any square matrix A, $||A|| = ||A^T||$.

Another possibility is to consider the Euclidien norm in R^{n^2}, sometimes also called the natural Frobenius norm, that is

$$||A||_F = \{Tr[A^T A]\}^{\frac{I}{2}} \qquad [5]$$

which can be shown to be induced by the inner product $< , >_F$ defined by

$$A,B \in L(R^n) \qquad \rightarrow \qquad <A,B>_F = <\underline{A},\underline{B}>_{R^{n^2}} = Tr[A^T B]. \qquad [6]$$

Notice that as,

$$Tr[AB] = Tr[BA] \qquad \leftrightarrow \qquad ||A||_F = ||A^T||_F.$$

Finally, it is easy to show that none of these two norms is sensitive to any change of basis in R^n. Thus, consider

$$A^I = QAQ^T$$

where Q is nonsingular orthogonal n x n matrix,

$$||A|| = \sup_{x\varepsilon R^n} \{\frac{x^T Ax}{x^T x}\} = \sup_{x\varepsilon R^n} \{\frac{x^T Q^T QAQ^T Qx}{x^T Q^T Qx}\} = \sup_{v=Qx} \{\frac{v^T A^I v}{v^T v}\} = ||A^I||$$

and

$$||A||_F^2 = Tr[AA^T] = Tr[Q^T QAQ^T QA^T] = Tr[QAQ^T QA^T Q^T] = Tr[A^I A^{I^T}] = ||A^I||_F^2$$

These properties can also be used to derive the following result:

Claim 1:

For any nonsingular matrix A of $L(R^n)$,

$$||A|| \leq ||A||_F \leq \sqrt{n}\ ||A|| \qquad [7]$$

Proof:

If $\lambda_1 \geq \lambda_2 \geq \lambda_n$ is the sequence of eigenvalues of A, then

$$||A|| = \lambda_1$$

But

$$\lambda_1^2 \leq ||A||_F^2 = \sum_{i=1}^{n} \lambda_i^2 \leq n\,\lambda_1^2 = [\sqrt{n}\ ||A||\ ]^2$$

Q.E.D.

Of course, this Frobenius norm can be extended to the case of any symmetric, positive definite metric induced by $G \geq 0$. Then write ,

$$||A||_{\tilde{G}} = \{Tr[A^T \tilde{G} A]\}^{\frac{1}{2}} \qquad [8]$$

and the previous result is generalized as follows:

Claim 2:

Let $\tilde{G} = RR^T$. As $\tilde{G}$ is non-singular R is also non-singular

and

$$\frac{||A||}{||R^{-1}||} \leq ||R^T A|| = ||A||_{\tilde{G}} \leq ||R||_F ||A||_F \qquad [9]$$

and also

$$\frac{||A||}{||R^{-1}||} \leq ||R^T A|| = ||A||_{\tilde{G}} \leq n||R||.||A|| \qquad [10]$$

Proof:

$$||A||_{\tilde{G}} = \{ Tr[RR^T AA^T] \}^{\frac{1}{2}} \leq \{Tr[RR^T]\}^{\frac{1}{2}} . \{ Tr[AA^T] \}^{\frac{1}{2}} = ||R||_F . ||A||_F$$

and as $\text{Det } \tilde{G} = [\text{Det } R]^2 \neq 0$,R is non-singular and R^{-1} hence exists.

Now,

$$||A||_{\tilde{G}} = \{Tr[A^T \tilde{G} A]\}^{\frac{1}{2}} = \{Tr[A^T RR^T A]\}^{\frac{1}{2}} = ||R^T A||_F$$

The second equation follows immediately from [7].

As $||A|| = ||AR^T R^{-T}|| \leq ||AR^T||.||R^{-T}||$ and therefore the lower bounds are obtained. Q.E.D

It should also be noticed that though their bounds are the same, $||A||_{\tilde{G}} \neq ||A^T||_{\tilde{G}}$ in general.

Because this norm $||\ .||_{\tilde{G}}$ will prove to be useless in the derivation of a minimal norm solution to the equation b=Xa, the following norm has to be introduced:

$$||A||_G = \{Tr[GAGA^T]\}^{\frac{1}{2}} \qquad [11]$$

for any matrix A of $L(R^n)$,

where G is a given symmetric,positive definite matrix,element of $L(R^n)$.

Obviously for this definition,

$$||A||_G = ||A^T||_G \quad .$$

Claim 3:

If G commutes with the matrix A, then

$$||A||_G = ||A^T||_{W=G^2} \qquad [12]$$

Proof:

The proof of this result is trivial, using the definition of the previous norms. Q.E.D

Claim 4:

If $G = SS^T$,

$$||A||_G \leq ||S||_F^2 . ||A||_F \qquad [13\text{-a}]$$

and if furthermore G commutes with A,

$$\frac{||A||_F}{||G^{-1}||} \leq ||A||_G \leq ||S||_F^2 . ||A||_F \qquad [13\text{-b}]$$

Proof:

$$||A||_G = [\langle G, AGA^T \rangle_F]^{\frac{1}{2}} \leq ||G||_F^{\frac{1}{2}} . ||AGA^T||_F^{\frac{1}{2}} \leq ||A||_F . ||G||_F \leq ||S||_F^2 ||G||_F$$

and,

$$||A||_F = ||AGG^{-1}||_F \leq ||AG||_F . ||G^{-1}||_F = ||G^{-1}||_F . \{Tr[AGGA^T]\}^{\frac{1}{2}}$$

and as A and G commute,

$$||A||_F \leq ||G^{-1}||_F . \{Tr[GAGA^T]\}^{\frac{1}{2}} = ||G^{-1}||_F . ||A||_G$$

Q.E.D

For fixed,bounded matrices G,Claim 2 implies that the $\tilde{G}$-norm and the Frobenius norm of a matrix A are equivalent.This is however not the case for G-norms,because for a given bounded matrix G,G-norms and Frobenius norms are only equivalent on the subset of matrices in $L(R^n)$ commuting with G.

Finally,Problem 0 can be restated as follows,

Problem 1:

Given a positive definite,symmetric matrix G in $L(R^n)$,find the minimal G-norm solution X* in $L(R^n)$ of the equation b=Xa, where b and a are given n-vectors,with the constraint that X* must be symmetric..

3-Variational approach:

Let G be some positive definite symmetric matrix which is an element of $L(R^n)$.

a) Problem 1 consists of minimizing $\frac{1}{2}\,\mathrm{Tr}(GXGX^T)$ over the admissible set of values given that the constraints are:

$$\begin{cases} X = X^T \\ b = Xa \text{ with } a \in R^n \quad , \quad b \in R^n \end{cases}$$

The easiest way to solve this problem is by introducing Lagrange multipliers in order to form the following Hamiltonian:

$$H(\lambda,B) = \frac{1}{2}\mathrm{Tr}\,[GXGX^T] + \lambda^T[Xa-b] + \mathrm{Tr}[B(X-X^T)], \qquad [14]$$

where,

$H(.,.) \in R$, $\lambda \in R^n$ and $B \in L(R^n)$.

Next, differentiating $H(\lambda,B)$ with respect to X, and using the fact that,

$$\lambda^T[Xa-b] = Tr[(Xa-b)\lambda^T]$$

$$\frac{\partial}{\partial X}[Tr[XM]] = M^T \qquad \text{and} \qquad \frac{\partial}{\partial X}[Tr[X^TM]] = M \qquad [15]$$

one obtains that,

$$\frac{\partial H}{\partial X} = G\ X\ G + \lambda a^T + B^T - B = 0$$

or also, $$X = -G^{-1}\ [\lambda a^T + B^T - B\ ]\ G^{-1} \qquad [16]$$

But as, $X - X^T = 0$,

$$G^{-1}\ [\lambda a^T - a\lambda^T + 2B^T - 2B\ ]\ G^{-1} = 0$$

$$B^T - B = \frac{1}{2}\ [\ a\lambda^T - \lambda a^T\ ].$$

Substituting this last result into [16], one gets,

$$X = -\frac{G^{-1}}{2}\ [\lambda a^T + a\lambda^T]\ G^{-1} \qquad [17]$$

Furthermore, X must verify the original equation $b=Xa$, hence,

$$b + \frac{G^{-1}}{2}\ [\lambda\ a^T + a\lambda^T]\ G^{-1}a = 0$$

or,

$$2Gb + (\lambda a^T + a\lambda^T)G^{-1}a = 0.$$

Solving partially in λ, the equations become,

$$-[2Gb + a(\lambda^T G^{-1}a)] = \lambda(a^T\ G^{-1}\ a)$$

$$\lambda = -\frac{1}{(a^T G^{-1} a)}\ [\ 2\ G\ b + a\ [\ \lambda^T\ G^{-1}\ a]] \qquad [18]$$

and multiplying on the left by a^{-T},

$$(a^T G^{-1}\lambda) = - \frac{1}{(a^T G^{-1} a)} [2 a^T b + (a^T G^{-1} a)(\lambda^T G^{-1} a)]$$

the result becomes,

$$(\lambda^T G^{-1} a) = - \frac{1}{(a^T G^{-1} a)}$$

Substituting again this expression into [18],λ becomes,

$$\lambda = - \frac{1}{(a^T G^{-1} a)} [2 Gb - \frac{a^T b}{(a^T G^{-1} a)} a]$$

and replacing λ in [17],one obtains,

$$X^* = X^*_I = \frac{1}{(a^T G^{-1} a)} [b a^T G^{-1} + G^{-1} a b^T - \frac{a^T b}{(a^T G^{-1} a)} G^{-1} a a^T G^{-1}] \qquad [19]$$

which is the final result.

Of course,in the case where the Euclidien norm -i.e. only $Tr[XX^T]$ - of the unknown matrix X is considered, $G = I = G^{-1}$; therefore,the result becomes,

$$X^*_I = \frac{1}{a^T a} [b a^T + a b^T - \frac{a^T b}{a^T a} . a a^T] \qquad [20] .$$

b) Sometimes also,a slightly different result is sought,in which case,the problem can be formulated as follows :

-Find the expansion of the symmetric matrix X* solution of the equation $b = X a$,around some known matrix $X^{(0)}$,in order to minimize the G-norm of the difference $X^* - X^{(0)}$.

This implies that $X^* = X^{(0)} + D^*$ with D^* symmetric and also that in the previous computations X can be replaced by D and b by $b - X^{(0)}$.

Therefore, the result becomes in this last case,

$$D^* = \frac{1}{a^T G^{-1} a} \left[(b - X^{(0)}a)a^T G^{-1} + G^{-1}a(b - X^{(0)})^T .. \right.$$

$$\left. .. - \frac{a^T(b - X^{(0)}a)}{a^T G^{-1} a} G^{-1} a a^T G^{-1} \right] \qquad [21\text{-a}]$$

and $X^* = X^{(0)} + D^*$ [21-b]

In fact, this second version of the problem will be discussed at length later on, when studying recursive procedures to perform similar matrix symmetrizations.

4-A recursive procedure to symmetrize the matrix solution of a linear equation.

The procedure below is originally due to Powell (1970) but has been later slightly generalized by Dennis (1972). The purpose for its introduction was to generate approximations for the Hessian matrix of a function; whereas, the previous variational method was introduced by Greenstadt (1970) to approximate the inverse of Hessian matrices.

In this paragraph it is shown, that,in fact, this recursive method is equivalent to the previous one and that it also leads to minimal symmetric solutions X^* of $b = Xa$, with respect to some well defined Euclidien-type norms.

The following procedure was proposed by Powell and Dennis:

Step 0: Let $X^{(0)}$ be any symmetric matrix such that $b \neq X^{(0)}a$.

Step 1: Construct the matrix $X^1 = X^{(0)} + (b - X^{(0)}a)c^T$, where

c is an n-vector verifying $c^T a = 1$.

Step 2: Construct the symmetric matrix $X^{(1)} = \frac{X^1 + X^{1^T}}{2}$.

Step 3: Restart at Step 1 until the procedure converges to X* given by,

$$X^* = X^*_{II} = X^{(0)} + (b - X^{(0)}a)c^T + c(b - X^{(0)}a)^T - (b - X^{(0)}a)^T a.cc^T \quad [22]$$

Remarks:

Let us make some comments before going any further.

a)-The equation b = Xa, where X is unknown, has a whole set of possible solutions. It was Broyden's idea [2] in the case of the "secant equation" to consider the general class of solutions of the form $X = M + (b - Ma)c^T$.

Clearly, a sufficient condition to have b = Xa in this general class is,

$b = Xa \rightarrow (1 - c^T a)Ma + (c^T a)b = b$

$c^T a = 1 \rightarrow$ X solution.

or equivalently,

$c^T a = 1 \leftrightarrow$ there exists d such that $c = \frac{d}{d^T a}$, with $d \in R^n$, $c \in R^n$, $a \in R^n$

b)-All n x n matrices X^n are solutions of b = Xa, but they are not symmetric.

All n x n matrices $X^{(n)}$ are symmetric, but they are not solutions of b = Xa.

The geometrical interpretation given at the end of this paragraph will reveal that as the procedure goes on, the image of the matrix X in R^{n^2} jumps from the hyperplane S to the convex set Δ and so forth, until it reaches their intersection. However, it is remarkable that the point reached through this procedure will also be the "closest" one to the starting point $X^{(0)}$ (in the G-norm sense).

c)-It is readily possible to verify that X^* is effectively the solution of the equation $b = X^*a$, using especially the fact that $c^Ta = 1$:

$$X^*a = X^{(0)}a + (b - X^{(0)}a)c^Ta + c(b - X^{(0)}a)^Ta - c(b - X^{(0)}a)^Ta \cdot c^Ta$$

$$= b + c(b^Ta - a^TX^{(0)}a) - c(b^Ta - a^TX^{(0)}a) = b$$

$$\rightarrow \quad X^* = X^* + (b - X^*a)c^T$$

But X^* is also clearly symmetric $\rightarrow$ $X^* = \lim_{n\to\infty} X^n = \lim_{n\to\infty} X^{(n)}$.

Curiously, no constructive proof of this result exists in the literature.

Finally, one can also notice that this procedure shows the remarkable property, that for each symmetric starting matrix $X^{(0)}$, there exists a symmetric terminal matrix X^*, solution of $b = Xa$. In particular for $X^{(0)} = 0$

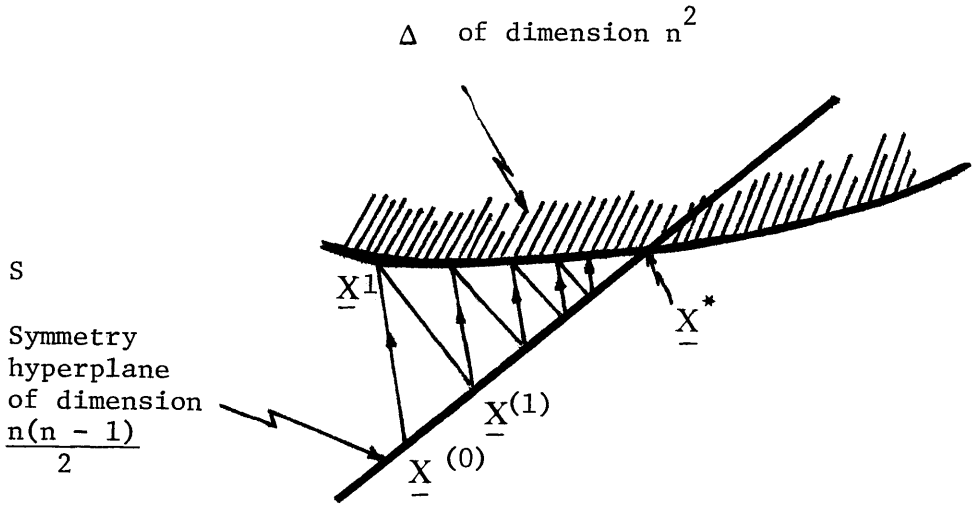
$$\rightarrow \quad X^*_{II} = bc^T + cb^T - (b^Ta)cc^T \ .$$

Geometrical interpretation:

Consider once more the equation $b = \underline{AX}$, where $\underline{X} \in R^{n^2}$ and

$\underline{A} \in L(R^{n^2}, R^n)$. We have already defined in the first section the symmetry hyperplane S of dimension $\frac{n(n-1)}{2}$ in R^{n^2}, along with the convex subset Δ of R^{n^2} containing all solutions $\underline{Z}$ of equation [1].

Powell's symmetrization procedure can be given a nice geometrical interpretation using these sets S and Δ. Consider the point $\underline{X}^{(0)}$ in R^{n^2}, since $X^{(0)}$ is symmetric, $\underline{X}^{(0)}$ belongs to S but not to the convex set Δ, since it generally does not verify equation [1]. From $X^{(0)}$, using the previous recursion , one obtains the point $\underline{X}^1$, which corresponds to X^1, a non-symmetric solution of [1]. $\underline{X}^1$ belongs hence to Δ, but not to the symmetry hyperplane S. The transposed matrix X^{1^T}, corresponding to X^1, has as "image" in R^{n^2} the mirror image of $\underline{X}^1$ in S. $\underline{X}^1$ is then nothing else than the projection of $\underline{X}^1$ on S. The matrix $X^{(1)}$ corresponding to $\underline{X}^{(1)}$ is , therefore, symmetric, but it does no longer verify [1], which proves that it does not belong to Δ. The procedure starts again from $X^{(1)}$, until it converges to $\underline{X}^*$ belonging to S Δ, or, equivalently, in matrix form, until the matrices converge to X^*.



The last point which has to be discussed is whether or not the solution X^* obtained by this method has some norm-minimality property .

Comparison between the minimal G-norm solutions and Powell's solution:

The minimal G-norm solution will be referred to as X_I^* ; whereas, the Powell-type of solution will be referred to as X_{II}^*.
More specifically,

$$X_I^* = \frac{1}{(a^T G^{-1} a)} \left[b a^T G^{-1} + G^{-1} b^T - \frac{(a^T b)}{(a^T G^{-1} a)} \cdot G^{-1} a a^T G^{-1} \right] \qquad [23]$$

and for $X^{(0)} = 0$,

$$X_{II}^* = b c^T + c b^T - (b^T c) \cdot c c^T. \qquad [24]$$

The similarity between these two formulas is especially striking if one chooses,

$$c = \frac{G^{-1} a}{a^T G^{-1} a} \qquad [25]$$

for which the condition $c^T a = 1$ is obviously verified.

As X_I^* is the minimal G-norm solution to Problem 1, it becomes clear that $X_I^* = X_{II}^*$ for the particular choice [25], and it follows that X_{II}^* is also minimal G-norm. This result can be condensed in the following Corollary.

Corollary:

There is a symmetric solution X^* to equation $b = Xa$ minimizing

also $Tr[GXGX^T]$, where G is agiven n x n positive definite symmetric matrix.This solution is given by,

$$X^* = \frac{1}{(a^TGa)}[\ ba^TG^{-1} + G^{-1}ab^T - \frac{a^Tb}{a^TG^{-1}a}\ G^{-1}aa^TG^{-1}\ ]$$

where a , b $\varepsilon\ R^n$ and $G\ \varepsilon\ L(R^n)$, and it is also the limit of the following sequence as n goes to infinity:

$$X^{(0)} = X^0 = 0$$

$$X^{n+1} = X^{(n)} + [\ b - X^{(n)}a\ ]\ \frac{a^TG^{-1}}{a^TG^{-1}a} \qquad \text{with, } X^i\ ,\ X^{(i)} \varepsilon\ L(R^n)$$

$$X^{(n+1)} = \frac{X^n + X^{n^T}}{2} \qquad \text{for } n=0,1,2,.$$

This corollary shows that basically,Greenstadt's variational method and Powell's recursive method- are generating the same kind of updates.

5-Expansion of a symmetric matrix solution of b=X a ,around any given matrix $X^{(0)}$

Let G be some given positive definite symmetric n x n matrix. It is then possible to consider in $L(R^n)$ the following G-norm,

$$||A||_G = \frac{1}{2}\ Tr[GAGA^T] \qquad \text{for any matrix A element of } L(R^n),$$

norm which also induces the G-distance

$$d_G[\ A,B\ ] = ||A - B||_G = \frac{1}{2}\ Tr[\ G(A - B)\ G\ (A - B)^T]$$

defined for any matrices A and B element of $L(R^n)$.

A slight generalization of the initial problem is analyzed in the following section.

Problem:

Given any n x n matrix $X^{(0)}$ - i.e. generally $X^{(0)} - X^{(0)^T} \neq 0$ - , find the G-closest symmetric solution X_G^* of the equation $b = X a$, where b and a are known n-vectors.

Solution:

Consider in $L(R^n)$ the n x n matrix E defined as,

$$X_G^* = X^* = E + X^{(0)} \qquad [26]$$

and let

$$\bar{b} = b - X^{(0)} a \ .$$

One has to minimize $\frac{1}{2} \mathrm{Tr}[\ GEGE^T\ ]$,

given that,

$$\bar{b} = E\ a \qquad \text{and} \qquad E^T + X^{(0)^T} = E + X^{(0)}$$

This leads us to the construction of the following Hamiltonian,

$$H = \frac{1}{2} \mathrm{Tr}[\ GEGE^T\ ] + \lambda^T (E\ a - \bar{b}\ ) + \mathrm{Tr}[\Gamma\ (\ E+X^{(0)}-E^T-X^{(0)^T}\ )]$$

or

$$H = \frac{1}{2} \mathrm{Tr}[\ GEGE^T\ ] + \mathrm{Tr}[(Ea - \bar{b})\lambda^T] + \mathrm{Tr}[\ \Gamma(\ E+X^{(0)} - E^T - X^{(0)^T}\ )] \qquad [27]$$

Next, differentiating with respect to E and using the property that,

$$\frac{\partial}{\partial A} Tr[\, A\, M\,] = M^T \qquad \text{and} \qquad \frac{\partial}{\partial A} Tr[\, A^T M\,] = M$$

$$\frac{\partial H}{\partial E} = GEG + \lambda a^T + \Gamma^T - \Gamma = 0$$

$$GEG = -\,[\, \lambda a^T + \Gamma^T - \Gamma\,]$$

$$E = -\, G^{-1}[\, \lambda a^T + \Gamma^T - \Gamma\,] G^{-1} \qquad [28]$$

But as,

$$E + X^{(0)} - E^T - X^{(0)^T} = 0$$

$$-\, G^{-1}[\, \lambda a^T - a\lambda^T + 2\Gamma^T - 2\,\Gamma] G^{-1} + X^{(0)} - X^{(0)^T} = 0$$

$$\Gamma^T - \Gamma = \frac{1}{2}\,[\, a\lambda^T - \lambda a^T] + \frac{1}{2}\, G\, [\, X^{(0)} - X^{(0)^T}\,]\, G \qquad [29]$$

Substituting this result back into [28],

$$E = -\,\frac{1}{2}[\, X^{(0)} - X^{(0)^T}\,] - G^{-1}[\, \lambda a^T + \frac{a\lambda^T}{2} - \frac{\lambda a^T}{2}\,]\, G^{-1}$$

$$E = -\,\frac{1}{2}[\, X^{(0)} - X^{(0)^T}] - G^{-1}[\frac{\lambda a^T + a\lambda^T}{2}]\, G^{-1} \qquad [30]$$

but still, E must verify

$$E\, a = \bar{b}$$

$$\bar{b} + G^{-1}[\, \frac{\lambda a^T + a\lambda^T}{2}]\, G^{-1} a + \frac{1}{2}\, [\, X^{(0)} - X^{(0)^T}\,]\, a = 0$$

$$2G\bar{b} + [\, \lambda a^T + a\lambda^T\,] G^{-1} a + G[\, X^{(0)} - X^{(0)^T}\,]\, a = 0$$

Solving partially in λ one obtains,

$$-[2G\bar{b} + a(\lambda^T G^{-1} a) + G\, (\, X^{(0)} - X^{(0)^T}\,)\, a\,] = \lambda(\, a^T G^{-1} a)$$

$$\lambda = - \frac{1}{a^T G^{-1} a} [\ 2G\bar{b} + a(\ \lambda^T G^{-1} a) + G(\ X^{(0)} - X^{(0)^T})a\ ] \qquad [31]$$

If one multiplies this expression by $a^T G^{-1}$ on the left, one gets,

$$a^T G^{-1}\lambda = - \frac{1}{a^T G^{-1} a} [\ 2a^T\bar{b} + (\ a^T G^{-1} a\ )(\ \lambda^T G^{-1} a\ ) + a^T(X^{(0)} - X^{(0)^T})a\ ]$$

and as

$$a^T G^{-1}\lambda = \lambda^T G^{-1} a$$

one can solve the previous equation in $a^T G^{-1}\lambda$,

$$a^T G^{-1}\lambda = - \frac{a^T\bar{b}}{a^T G^{-1} a} - \frac{a^T(\ X^{(0)} - X^{(0)^T})a}{2\ (a^T G^{-1} a)} \qquad [32]$$

Substituting this result, back into [31],

$$\lambda = - \frac{1}{a^T G^{-1} a} [\ 2G\bar{b} - \frac{a^T\bar{b}\ a}{a^T G^{-1} a} - \frac{a^T(X^{(0)} - X^{(0)^T})a}{2(a^T G^{-1} a)} + G(X^{(0)} - X^{(0)^T})a\ ]$$

and replacing this value of λ into [30], the expression becomes,

$$\lambda = - \frac{1}{a^T G^{-1} a} [\ 2G\bar{\bar{b}} - \frac{(a^T\bar{b})\ a}{a^T G^{-1} a}\ ]$$

with, $\bar{\bar{b}} = \bar{b} + \frac{X^{(0)} - X^{(0)^T}}{2} a = b - \frac{X^{(0)} - X^{(0)^T}}{2} a$

$$[30] \rightarrow E = - \frac{1}{2} [\ X^{(0)} - X^{(0)^T}\ ] + \frac{1}{a^T G^{-1} a} [\ \bar{\bar{b}} a^T G^{-1} + G^{-1} a\bar{\bar{b}} - \frac{a^T\bar{b}}{a^T G^{-1} a} G^{-1} a a^T G^{-1}\ ]$$

or finally,

$$X^* = \frac{X^{(0)} - X^{(0)^T}}{2} + \frac{1}{a^T G^{-1} a} [(b - \frac{X^{(0)} + X^{(0)^T}}{2} a) a^T G^{-1} + G^{-1} a (b - \frac{X^{(0)} + X^{(0)^T}}{2} a)^T \ ..$$

$$... - a^T (b - \frac{X^{(0)} + X^{(0)^T}}{2} a) . \frac{G^{-1} a a^T G^{-1}}{a^T G^{-1} a}\ ]$$

Theorem:

Given any matrix X in $L(R^n)$, the G closest symmetric matrix X^* solution of $b = X\,a$ is also G-closest to the symmetric matrix $\frac{X + X^T}{2}$ and, hence, is given by:

$$X^* = \frac{X + X^T}{2} + \frac{1}{(a^T G^{-1} a)} \left[(b - \frac{X + X^T}{2} a) a^T G^{-1} + G^{-1} a (b - \frac{X + X^T}{2} a)^T \ldots \right.$$

$$\left. \ldots - a^T (b - \frac{X + X^T}{2} a) . \frac{G^{-1} a a^T G^{-1}}{a^T G^{-1} a} \right]$$

Geometrically this also means that the projection of any matrix on $S \cap \Delta$, the set of all symmetric solutions of $S = Xa$ is also equal to the projection of the symmetrical matrix $\frac{X + X^T}{2}$ on the same set.

CHAPTER III.

State-space model and filtering for minimizing a function.

I-Introduction and statement of the problem.

Consider the function $f:R^n \rightarrow R$ having the property that it is twice continuously differentiable in R^n and let $g(x)=\nabla f(x)$ be its gradient.Assume that the problem consists of finding the minimum of such a function-we are not concerned here with any existence problem and hence,we assume that at least in a certain domain D,such a minimum exists-.This minimum will then be given by the solution of $g(x)=0$, which corresponds to a Newton-type problem.

Assume now that a particular sequence of points $(x_k,k=0,I,2,.)$ in R^n has been found:the previous equation can then be decomposed according to its Taylor expansion:

$$g(x_{k+I})-g(x_k)=G(x_k)(x_{k+I}-x_k)+O(|x_{k+I}-x_k|^2) \qquad [I]$$

for k=0,I,2,.

where $G(x_k)$ is the Hessian matrix of f computed at point x_k,and $O(|x_{k+I}-x_k|^2)$ is a second order term in $(x_{k+I}-x_k)$,term which is constantly equal to zero in the case of quadratic functions.

Now,in this equation [I] x_k and x_{k+I} are assumed to be known and the gradient increment $g(x_{k+I})-g(x_k)$ is supposed also to be exactly computable,but neither $G(x_k)$,nor the correction term are known with precision.The next paragraph shows how to transform equation [I] into a set of stochastic state-space equations and then how to"best" estimate

the value of $G(x_k)$.

2-Presentation of the model.

The state-space equations and the model presented in this section were largely inspired by the work of S.W. Thomas [28] though their interpretation slightly differs from his. A Kalman filtering approach will be used to compute the leastsquares estimate of the Hessian matrix of f(.). In the present case the Kalman filtering method will be utilized as an identification step in the reconstruction of the dynamics of a given system. Thereby, it will differ from Thomas' interpretation since he viewed this method as a peculiarity generating formulas similar to Broyden's, as far as the updating of the Hessian matrix is concerned.

Consider now the following functions:

F_k: [0,1] $\rightarrow$ R^n for all k = 0,1,2,...

such that $F_k(\theta) = F(x_k + \theta s_k) = g(x + \theta s_k)$, where s_k represents the difference $s_k = x_{k+1} - x_k$

and G_k: [0,1] $\rightarrow$ $L(R^n)$

such that $G_k(\theta) = F'(x_k + \theta s_k) = g(x_k + \theta s_k)$ for all k = 0,1,2,..

Now assume that F(.) and the points x_k are such that $F'[x_k + \theta s_k]$ for k = 0,1,2,... is a matrix valued Wiener process. This means in particular that for all k = 0,1,2,...

$G_k[x_k + \theta s_k]$, $\forall\, \theta \in [0,1]$ is such that,

a) $[G_k(\theta),\ \theta\varepsilon[0,1]\ ]$ has stationary and independent increments,

b) for given θ in $[0,1]$, $G_k(\theta)$ is normally distributed,

c) it has zero mean, i.e. $E\,[G_k(\theta)]$ for all $\theta\varepsilon[0,1]$.

Now , using the isomorphism existing between $L(R^n)$ and R^{n^2}, that is between the set of all n x n matrices and R^{n^2}, to compute the covariance of G (see also Parzen (1962)):
for each $G\varepsilon L(R^n)$, one can isomorphically associate a vector $\underline{G}\ \varepsilon R^{n^2}$ such that,

$$\text{if}\quad G = \begin{bmatrix} g_1 \\ \cdots \\ g_2 \\ \cdots \\ \vdots \\ g_n \end{bmatrix} \quad \leftrightarrow \quad \underline{G}^T = [g_1^T,\ g_2^T,\ \ldots\ g_n^T\ ]^T$$

where $g_i\ \varepsilon\ R^n$ is the i-th row of G.

By definition, the covariance of a matrix valued process G is also the covariance of its isomorphically correspondent vector valued process G. Therefore,

$$\text{Cov}[G(\theta_2) - G(\theta_1)] = E\left\{ [\underline{G}(\theta_2) - \underline{G}(\theta_1)]\ [\underline{G}(\theta_2) - \underline{G}(\theta_1)]^T \right\}$$

$$= C_k|\theta_2 - \theta_1| \quad \text{with } C_k \geq 0,\ C_k\varepsilon L(R^{n^2})$$

$$\text{for all} \quad \theta_2,\theta_1\varepsilon[0,1]$$

$$\text{and all } k = 0,1,2,\ldots$$

To simplify the computations , assume that C_k is diagonal and that it takes the form,

$$C_k = ||s_k||\, I_{n2} \qquad [3]$$

which simply means that the rows of $G_k(\theta_2) - G_k(\theta_1)$ are assumed to be uncorrelated with one another; therefore, because of the normality assumption, they are statistically independent.

A state- space model can than be constructed to describe the pair $[F'(x_k),\ F(x_{k+1}) - F(x_k)]$. For this occasion , the previous notation is simplified by using:

$$\begin{aligned} g_k &= F_k(0), \quad g_{k+1} = F_{k+1}(0) = F_k(1) \\ u_k &= g_{k+1} - g_k = F_k(1) - F_k(0) \\ G_k &= G_k(0) = F'(x_k) \end{aligned} \qquad [4]$$

then,

$$\begin{aligned} G_{k+1} &= G_k + V_k \\ u_k &= G_k s_k + w_k \qquad \text{for all } K = 0,1,2,... \end{aligned} \qquad [5]$$

where,

$$V_k = G_{k+1} - G_k = G_k(1) - G_k(0), \quad \text{for all } k = 0,1,2,...,$$

is the matrix valued Wiener process previously described,

and,

$$w_k = g_{k+1} - g_k - G_k s_k$$

$$w_k = F_k(1) - F_k(0) - G_k(0)\, s_k$$

$$w_k = \int_0^1 F_k'(\theta)\ .\ s_k d\theta - G_k(0) s_k$$

$$w_k = \int_0^1 [G_k(\theta) - G_k(0)\,]s_k \, d\theta \qquad \text{for all } k=0,1,2,. \quad [6]$$

The mean of w_k is, therefore, clearly zero since,

$$E(w_k) = \int_0^1 E[G_k(\theta) - G_k(0)\,]s_k \, d\theta = 0 \quad \text{for all } k=0,1,2,.$$

and the joint noise process $\begin{pmatrix} V_k \\ w_k \end{pmatrix}$ has the following covariance:

$$Q_k = E\left\{ \begin{bmatrix} V_k V_k^T & V_k w_k^T \\ w_k V_k^T & w_k w_k^T \end{bmatrix} \right\} = ||s_k|| \begin{bmatrix} I_{n^2} & \dfrac{S_k^T}{2} \\ \dfrac{S_k}{2} & \dfrac{||s_k||^2}{3} I_n \end{bmatrix}$$

equation [7]

where Q_k is an element of $L(R^{n^2+n})$ and where S_k is defined by

$$S_k = \begin{bmatrix} s_k^T & 0 \\ 0 & s_k^T \end{bmatrix} \text{ belonging to } L(R^{n^2}, R^n) \quad \text{for all } k=0,1,2,.$$

with the property that

$$G_k \, s_k = S_k \, \underline{G}_k \qquad [8]$$

For more details,see the Appendix where the computations to get Q_k are explained.

To complete this model,some more assumptions are needed on the statistics of the initial value G_0.In this model we shall assume that G_0 is a zero-mean, Gaussian matrix process taking its values in $L(R^n)$ and having as covariance Π_0 such that :

$$\Pi_0 \geq 0 \quad \text{and} \quad \Pi_0 \text{ is an element of } L(R^{n^2})$$

From equation [5] it appears that the state of the system can be described as an element of $L(R^n)$,or equivalently , using the isomorphism between $L(R^n)$ and R^{n^2} , by the n^2-vector $\underline{G}_k$.

The true process noise is the sequence (V_k k=0,1,2,.),sequence which, because of the non-linearity of [1] induces an observation noise (w_k, k=0,1,2,.) with covariance $\frac{||s_k||^3}{3} I_n$ and with a high correlation with (V_k).

The observed process is the gradient increment process (u_k, k=0,1,2,...) which according to equation [5] is corrupted by the observation noise (w_k). At this point, let us notice that, in fact, the set of equations given at [5] can be rewritten as,

$$\begin{cases} \underline{G}_{k+1} = \underline{G}_k + \underline{V}_k & k=0,1,... \quad \underline{V}_k, \underline{G}_k \quad R^{n^2} \\ u_k = S_k \underline{G}_k + w_k & \quad w_k, u_k \quad R^n \end{cases} \qquad [9]$$

which resembles more the conventional vector-state and vector-observation models encountered in the literature.

Before discussing this model , let us regroup its assumptions in the following manner:

Model:

$$\begin{cases} \text{Dynamics:} & G_{k+1} = G_k + V_k \\ \text{Observations:} & u_k = G_k s_k + w_k \end{cases} \qquad [10]$$

with k=0,1,2,. ,and with V_k, $G_k \in L(R^n)$, w_k, $u_k \in R^n$.

→ G_0 is an n x n matrix valued, Gaussian random variable,with mean,

$$E[G_0] = \hat{G}_0$$

and covariance,

$$\text{Cov}\,[\underline{G}_0, \underline{G}_0] = \Pi_0 = \begin{bmatrix} P_0 & 0 \\ 0 & P_0 \end{bmatrix} \in L(R^n) \qquad [11]$$

where P_0 is a positive semi-definite matrix in $L(R^n)$.

→ $\begin{pmatrix} V_k \\ w_k \end{pmatrix}$, k=0,1,2,. is a Gaussian random process taking its values in $L(R^n) \times R^n$, with mean $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance,

$$Q_k = ||s_k|| \begin{Bmatrix} I_{n^2} & \frac{S_k^T}{2} \\ \frac{S_k}{2} & ||s_k||^2 \frac{I_n}{3} \end{Bmatrix} \in L(R^{n^2+n})$$

In S.W. Thomas' model [see pp. 42-43 of ref. 28], the state

is the couple (g_k, G_k) element of $R^n \times L(R^n)$, and the observation is a function $F_k \varepsilon R^n$ identically equal to the first component of the state for each k = 0,1,2,.. .

The dynamics were described by,

$$\begin{bmatrix} g_{k+1} \\ G_{k+1} \end{bmatrix} = \underline{\Phi}_k \begin{bmatrix} g_k \\ G_k \end{bmatrix} + \begin{bmatrix} w_k \\ \underline{V}_k \end{bmatrix} \qquad [12]$$

for all k = 0,1,2,..., where $(g_k, G_k) \varepsilon R^n \times L(R^n)$

and,

$$\Phi_k = \begin{bmatrix} I & S_k \\ 0 & I_{n^2} \end{bmatrix}$$

and the observations were described by,

$$F_k = M_k \begin{bmatrix} g_k \\ G_k \end{bmatrix}$$

and [13]

$$M_k = [I \vdots 0]$$

for all k = 0,1,2,..., where $F_k \varepsilon R^n$.

At each point x_k, F_k is measured with high precision and consequently g_k as well. Thomas formulates his problem as consisting of estimating $\hat{g}_{k+1/k}$ and $\hat{G}_{k+1/k}$ given all previous F_j (or g_j) for j = o,1,2,...k. This is typically a singular filtering formulation,

where $\hat{g}_{k/k} = g_k$

and strictly speaking, no Kalman Filter can be constructed.

The classical way,however, to transform this mathematically ill-posed problem is to substract the deterministic -or known- parts from the state.

This leads to,

$$\begin{cases} G_{k+1} = G_k + V_k \\ u_k = g_{k+1} - g_k = G_k s_k + w_k \end{cases} \qquad \text{for } k=0,1,2,.$$

Consider now,that after having reached point x_k , u_k is observed but is known to be corrupted by some observation noise w_k of known covariance. From this new observation u_k and all the previous ones - which by simple addition reconstruct the gradient g_k- , the least squares estimate $\hat{G}_{k+1/k}$ of G_{k+1} is sought. This formulation now perfectly fits into the standard Kalman filtering theory,which will be applied in the next section.

In conclusion, consider again the Taylor expansion [1] :

$$g_{k+1} - g_k = u_k = G_k s_k + O[||s_k||^2] \qquad \text{for } k=0,1,2,.$$

Clearly if the function f(.) which has to be minimized , is quadratic , this expansion is reduced to its first term ,

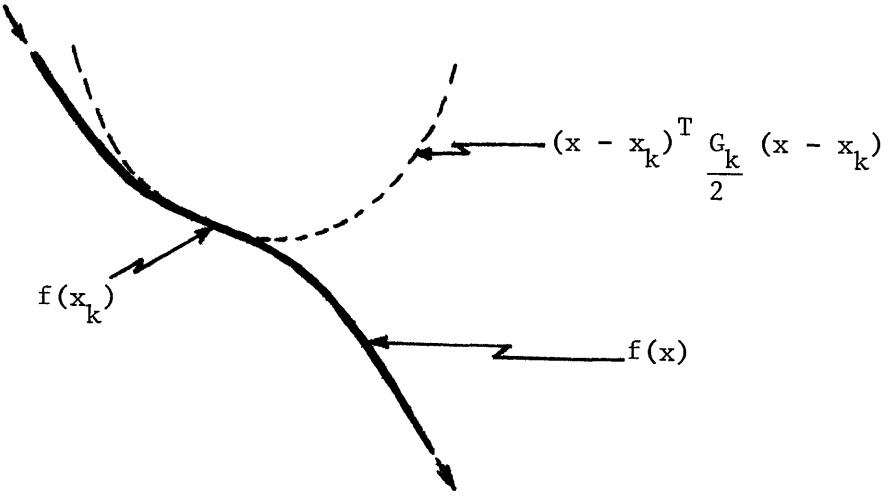
$$u_k = Qs_k \qquad \text{if } f(x) = \frac{1}{2} x^T Q x + b^T x + c \ .$$

In the case of any twice continuously differentiable function f(.) , the model [see equation 10] replaces the correction term $O[||s_k||^2]$ of order at least two in s_k by a Gaussian noise process (w_k) with covariance proportional to $||s_k||^3$ .This last indication

should reinforce our confidence in the consistency of the proposed model.

Other remarks:

In some sense, one can consider that the covariance of the observation noise is a relatively good measure of the distance existing between $f(x_k)$ and its local quadratic hull.



The whole previous derivation was based upon the assumption that the sequence of points $(x_k, k=0,1,2..)$ converging to the minimum of the function was known. This conjecture allowed us to compute for each value of k, the step-length s_k, as well as all the noise covariances. However,this is not the case in general and one has also to construct this sequence of points (x_k) by using for instance Newton's method ,

$$x_{k+1} - x_k = - [G(x_k)]^{-1} g_k \qquad \text{for all } k=0,1,2,.$$

or, $$s_k = - G_k^{-1} g_k \qquad [14]$$

3-The Kalman Filter:

Claim:

Assume that the initial estimation error covariance is given by [11]. The minimum mean squares estimate $\hat{G}_k$ associated with the previous model is then given by:

$$\begin{cases} \hat{G}_{k+1} = \hat{G}_k + \dfrac{(u_k - \hat{G}_k s_k) s_k^T}{s_k^T \left[P_k + \dfrac{||s_k||}{3} I\right] s_k} \left[P_k + \dfrac{||s_k||}{2} I\right] \\ \hat{G}_0 \text{ given} \end{cases} \qquad \text{for } k = 0,1,2,... \qquad [15]$$

and its error covariance by:

$$\begin{cases} P_{k+1} = ||s_k|| I + P_k - \dfrac{\left[P_k + \dfrac{||s_k||}{2} I\right] s_k s_k^T \left[P_k + \dfrac{||s_k||}{2} I\right]}{s_k^T \left[P_k + \dfrac{||s_k||}{3} I\right] s_k} \\ P_0 \text{ given} \end{cases} \qquad \text{for } k = 0,1,2,... \qquad [16]$$

Proof:

Considering the previous model in its vector form (i.e. with state $\underline{G}_k \varepsilon R^{n^2}$ instead of $G_k \varepsilon L(R^n)$):

$$\begin{cases} \underline{G}_{k+1} = \underline{G}_k + \underline{V}_k \\ u_k = S_k \underline{G}_k + w_k \end{cases} \qquad \text{for } k = 0,1,2,... \qquad [17]$$

Note that $\Pi_k = \text{Cov}[G_k, G_k] = \text{Cov}[\underline{G}_k, \underline{G}_k] = E[(\underline{G}_k - \hat{\underline{G}}_k)(\underline{G}_k - \hat{\underline{G}}_k)^T] \varepsilon L(R^{n^2})$ is the error covariance after the k-th step.

Using the notation,

$$Q_{11}(k) = ||s_k||I_{n^2} ,$$

$$Q_{12}(k) = Q_{21}^T(k) = \frac{1}{2}||s_k||S_k^T ,$$

$$Q_{22}(k) = \frac{1}{3}||s_k||^3 I_n ,$$

the resulting Kalman Filter is, for instance, given by Theorem 6.42 in Kwakernaak and Sivan [14]. Applying it to this model,one gets,

$$\hat{\underline{G}}_{k+1} = \hat{\underline{G}}_k + K_g(k)[u_k - S_k\hat{\underline{G}}_k]$$

with

$$K_g(k) = [\Pi_k S_k^T + Q_{12}(k)]\ [Q_{22}(k) + S_k\Pi_k S_k^T]^{-1},$$

and

$$\Pi_{k+1} = Q_{11}(k) + \Pi_k - K_g(k)[Q_{12}^T + S_k\Pi_k] ,$$

or also

$$K_g(k) = [\Pi_k + \frac{1}{2}||s_k||I_{n^2}]\ S_k^T\ [\frac{1}{3}||s_k||^3 I_{n^2} + S_k\Pi_k S_k^T]^{-1}$$

starting with $\Pi_0 = \begin{bmatrix} P_0 & 0 \\ 0 & P_0 \end{bmatrix}$ $\varepsilon L(R^{n^2})$ bloc diagonal, and assuming that Π_k has also the same structure,

$$\Pi_k = \begin{bmatrix} P_k & 0 \\ 0 & P_k \end{bmatrix} \varepsilon L(R^{n^2}), \text{ and that } P_k = P_k^T\ \varepsilon L(R^n)$$

it follows that,

$$K_g(k) = [\ \Pi_k + \frac{1}{2}\ ||s_k||\ I_{n^2}\ ]S_k^T\ [\ \frac{||s_k||^3}{3}\ I_{n^2} + s_k^T\ P_k s_k\ ]^{-1}$$

$$\rightarrow\ K_g(k) = \frac{[\ \Pi_k + \frac{1}{2}||s_k||I_{n^2}\ ]\ S_k^T}{s_k^T\ [\ P_k + \frac{||s_k||}{3}\ I_n\ ]\ s_k}$$

and hence that,

$$\Pi_{k+1} = \Pi_k + ||s_k||I_{n^2} - \frac{[\ \Pi_k + \frac{1}{2}||s_k||I_{n^2}]S_k^T S_k[\Pi_k + \frac{1}{2}||s_k||I_{n^2}]}{s_k^T\ [\ P_k + \frac{||s_k||}{3}\ I_n\ ]\ s_k}$$

$$\rightarrow \Pi_{k+1} = \Pi_k + ||s_k||I_{n^2} - \frac{[\ P_k + \frac{1}{2}||s_k||I_n\ ]s_k s_k^T[P_k + \frac{1}{2}||s_k||I_n\ ]}{s_k^T\ [\ P_k + \frac{||s_k||}{3}\ I_n\ ]\ s_k}\ I_{n^2}$$

It becomes clear that Π_{K+1} remains bloc-symmetric and that,

if $\Pi_{k+1} = \begin{pmatrix} P_{k+1} & 0 \\ 0 & P_{k+1} \end{pmatrix}$, then ,

$$P_{k+1} = P_k + ||s_k||I_n - \frac{[\ P_k + \frac{1}{2}||s_k||I_n]s_k s_k^T[\ P_k + \frac{1}{2}||s_k||I_n\ ]}{s_k^T\ [\ P_k + \frac{||s_k||}{3}\ I_n\ ]\ s_k}$$

hence:

$$K_g(k) = [\ P_k + \frac{1}{2}||s_k||I_n\ ]$$

$$\hat{G}_{k+1} = \hat{G}_k + \frac{[\ \Pi_k + \frac{||s_k||}{2} I_{n^2}\ ]\ s_k^T\ [\ u_k - S_k\ \hat{G}_k]}{s_k^T\ [\ P_k + \frac{||s_k||}{3} I_n\ ]\ s_k}$$

and after rearranging the terms,

$$\hat{G}_{k+1} = \hat{G}_k + [\ u_k - \hat{G}_k s_k]\ \frac{s_k^T\ [\ P_k + \frac{1}{2}||s_k|| I_n\ ]}{s_k^T\ [\ P_k + \frac{||s_k||}{3} I_n\ ]\ s_k} \qquad \text{for } k=0,1,2.$$

Q.E.D

Notice that the Kalman filter formula [15] is a member of a general class of updates of the form,

$$G_{k+1} = G_k + [u_k - G_n s_k] c_k^T \qquad [18]$$

These formulas have the advantage of being particularly simple, however, they also offer the following major inconveniences.

In the case of function minimization problems ,where f, the function to be minimized is assumed to be continuously differentiable, G_k has the meaning of being the Hessian of f(.), and hence should be symmetrical. Consequently, it is somehow disturbing to construct a family of non-symmetrical estimates $\hat{G}_k$.

A second property is also usually required in the gradient algorithm literature. At each step k, the estimate $\hat{G}_{k+1}$ is usually required to verify the so-called "secant equation",

$$u_k = \hat{G}_{k+1} s_k \qquad \text{for all } k = o,1,2,... \qquad [19]$$

Now,forgetting for a moment the previous results obtained by application of the minimum mean squares filtering theory, and considering the following new updates,

$$\hat{G}_{k+1} = \hat{G}_k + [u_k - \hat{G}_k s_k] c_k^T$$

Attempting to also verify the secant equation , it is possible to write,

$$u_k = \hat{G}_{k+1} s_k$$

$$\rightarrow u_k = \hat{G}_k s_k + [u_k - \hat{G}_k s_k] c_k^T s_k$$

$$\rightarrow [u_k - \hat{G}_k s_k] \; [1 - c_k^T s_k] = 0$$

$$\text{or} \quad c_k^T s_k = 1 \qquad [20]$$

which means , also, that there exists some vector $\mu_k \varepsilon R^n$ verifying,

$$c_k = \frac{\mu_k}{\mu_k^T s_k} \qquad [21]$$

Moreover, this condition appears to be necessary and sufficient for any matrix update $\hat{G}_{k+1}$ obtained through the class [18] to verify also the secant equation.

As the Kalman filter [15] belongs to the general class [18], but with a vector c_k equal to,

$$c_k = \frac{[P_k + \frac{||s_k||}{3} I_n]}{s_k^T [P_k + \frac{||s_k||}{3} I_n] s_k} \cdot s_k \quad \rightarrow \quad c_k^T s_k \neq 1,$$

it implies that the secant equation [19] is not verified.

In Chapter V , it will be demonstrated , using the results of Chapter III, how to best approximate in the Euclidien norm sense the least square estimate $\hat{G}_{k+1}$, previously derived, in order to construct symmetric updates which also verify the secant equation [19].

4-Inversion of the Kalman filter.

In the previous section, we saw that the least squares estimate of G_{k+1} ,the Hessian matrix of a function f(.) computed at the point x_{k+1} was given by,

$$\hat{G}_{k+1} = \hat{G}_k + \frac{[\, u_k - \hat{G}_k s_k \,]\, s_k^T \,[\, P_k + \frac{||s_k||}{2} I \,]}{s_k^T \,[\, P_k + \frac{||s_k||}{3} I \,]\, s_k} \qquad [22] \qquad \text{for } k=0,1,..$$

where P_k is generated by [16].

Let us assume for a moment that $\hat{G}_k$ is non-singular and that its inverse $\bar{H}_{k+1} = [\, \hat{G}_{k+1} \,]^{-1}$ exists ,then,

Claim:

If $\bar{H}_k = [\, \hat{G}_k \,]^{-1}$ exists ,then

$$\bar{H}_{k+1} = \bar{H}_k + \alpha_k^{-1} \frac{[\, s_k - \bar{H}_k u_k \,]\, \bar{d}_k^T \bar{H}_k}{1 + \alpha_k^{-1} \bar{d}_k [\, s_k - \bar{H}_k u_k \,]} \qquad [23]$$

where,

$$\alpha_k = \frac{s_k^T \left[P_k + \frac{||s_k||}{3} I \right] s_k}{s_k^T \left[P_k + \frac{||s_k||}{2} I \right] s_k} > 0 \qquad [24]$$

$$\bar{d}_k = \frac{\left[P_k + \frac{||s_k||}{2} I \right] s_k}{s_k^T \left[P_k + \frac{||s_k||}{2} I \right] s_k} \quad , \qquad [25]$$

is the inverse of $\hat{G}_{k+1}$.

Proof:

Our proof is based on an identity due to Sherman and Morrison which can be written as follows:

if B^{-1} exists, then ,

$$\begin{aligned} [\, B - \sigma x y^T \,]^{-1} &= B^{-1} - \tau B^{-1} x y^T B^{-1} \\ \frac{1}{\tau} + \frac{1}{\sigma} &= y^T B^{-1} x \end{aligned} \qquad [26]$$

Some straightforward algebraic computation would be needed in order to verify the previous result , but being irrelevent to the subject , it will not be attempted here.

Now, consider the following correspondences between [22] and [26]:

$$\sigma \leftrightarrow \alpha_k^{-1} \qquad B \leftrightarrow \hat{G}_k$$

$$d_k \leftrightarrow \bar{d}_k \qquad x \leftrightarrow \hat{G}_k s_k - u_k$$

$$-\frac{\alpha_k^{-1}}{1 + \alpha_k^{-1} d_k^T [s_k - \bar{H}_k u_k]} \leftrightarrow \tau = \frac{\sigma}{\sigma y^T B^{-1} x - 1}$$

and hence from [26] , one obtains ,

$$\bar{H}_{k+1} = \bar{H}_k + \frac{[\, s_k - \bar{H}_k u_k \,]\, \bar{d}_k^T\, \bar{H}_k}{\alpha_k + \bar{d}_k^T\, [\, s_k - \bar{H}_k u_k \,]} \qquad [23]$$

Q.E.D

Of course there is no more reason now for deciding that $\bar{H}_{k+1}$ should be symmetric in [23] , nor to have ,

$$s_k = \bar{H}_{k+1}\, u_k$$

the secant equation , verified. Consequently , the updates given by [23] do not belong either to the class of Broyden's formulas described by

$$\bar{H}_{k+1} = \bar{H}_k + [\, s_k - \bar{H}_k u_k \,]\, c_k^T$$

with $c_k^T u_k = 1$.

5-A nonlinear stochastic regulation problem.

Consider once more the problem of minimizing a function $f:R^n \to R$ which is assumed to be twice continuously differentiable and let x^* be its minimum. A necessary condition of minimality for such a function is that $\nabla f(x^*) = g(x^*) = 0$. The problem consists then of building a sequence $[x_k, x \geq 0]$ starting at some point x_0, such that $g(x_0) = g_0$ and converging to x^*. A necessary condition to obtain such a sequence is clearly through the corresponding sequence of gradients $[g_k, k \geq 0]$, that converge also to zero as k goes to

infinity.

The state space equations [10] representing our system are:

$$\begin{cases} G_{k+1} = G_k + V_k \\ u_k = G_k s_k + w_k \end{cases} \qquad [27]$$

$$G_k \in L(R^n)$$
$$u_k \in R^n$$
$$G_0 \sim N[\hat{G}_0, \Pi_0] \qquad \text{for all } k \geq 0$$

where $[V_k]$ and $[w_k]$ were two Gaussian perturbations, and u_k represented the gradient increment at point x_k, that is, $u_k = g_{k+1} - g_k$. An equivalent representation of the same system is then the following:

$$\begin{cases} \begin{pmatrix} \underline{G}_{k+1} \\ g_{k+1} \end{pmatrix} = \begin{pmatrix} I_{n^2} & 0 \\ S_k & I \end{pmatrix} \begin{pmatrix} \underline{G}_k \\ g_k \end{pmatrix} + \begin{pmatrix} \underline{V}_k \\ w_k \end{pmatrix} \\ \\ y_k = (0 \quad I) \begin{pmatrix} \underline{G}_k \\ g_k \end{pmatrix} \qquad \text{for all } k = o,1,2,... \end{cases} \qquad [28]$$

in which no perturbation alters the output y_k.

$\begin{pmatrix} \underline{V} \\ w \end{pmatrix}$ is the input noise with given covariance - see Appendix 1.

$\begin{pmatrix} \underline{G}_0 \\ g_0 \end{pmatrix}$ is also Gaussian with mean $\begin{pmatrix} \hat{\underline{G}}_0 \\ 0 \end{pmatrix}$ and covariance $\begin{pmatrix} \Pi & 0 \\ 0 & 0 \end{pmatrix}$.

y_k takes value g_k at each point x_k and can be interpreted as the output of our system.

Apparently there is no deterministic input to the system, but only noises; this is not quite so since s_k or equivalently S_k (remember that by definition $S_k G_k = G_k s_k$) is a multiplicative input to the system. Consequently, the problem can also be viewed as consisting of finding the function $\Phi(.)$such that $s_k = \Phi_k[G_k, y_k]$ and, which stears the output to zero. In fact we have a regulation problem in the sense that the output y_k must go to zero as k increases, but perhaps a better appelation would be a zero-target problem, as we are interested in only the first time (or point x^*) the system reaches a level zero.

One possibility, as usual, is to restrict ourselves to linear feedback laws for $\Phi_k(.)$ of the form,

$$s_k = \Phi_k g_k \qquad [29]$$

In order to achieve this goal, the classical method consists of linearizing $\Phi_k(\ )$ around the estimates of its arguments, estimates which are usually obtained through a filtering stage. This leads to:

$$s_k = \Phi_k[G_k, g_k\ y_k] \sim \Phi_k[\hat{G}_k, g_k] y_k \qquad [30]$$

$$\text{as } \hat{g}_k = y_k$$

By comparing now this result with Newton's method which uses the fact that,

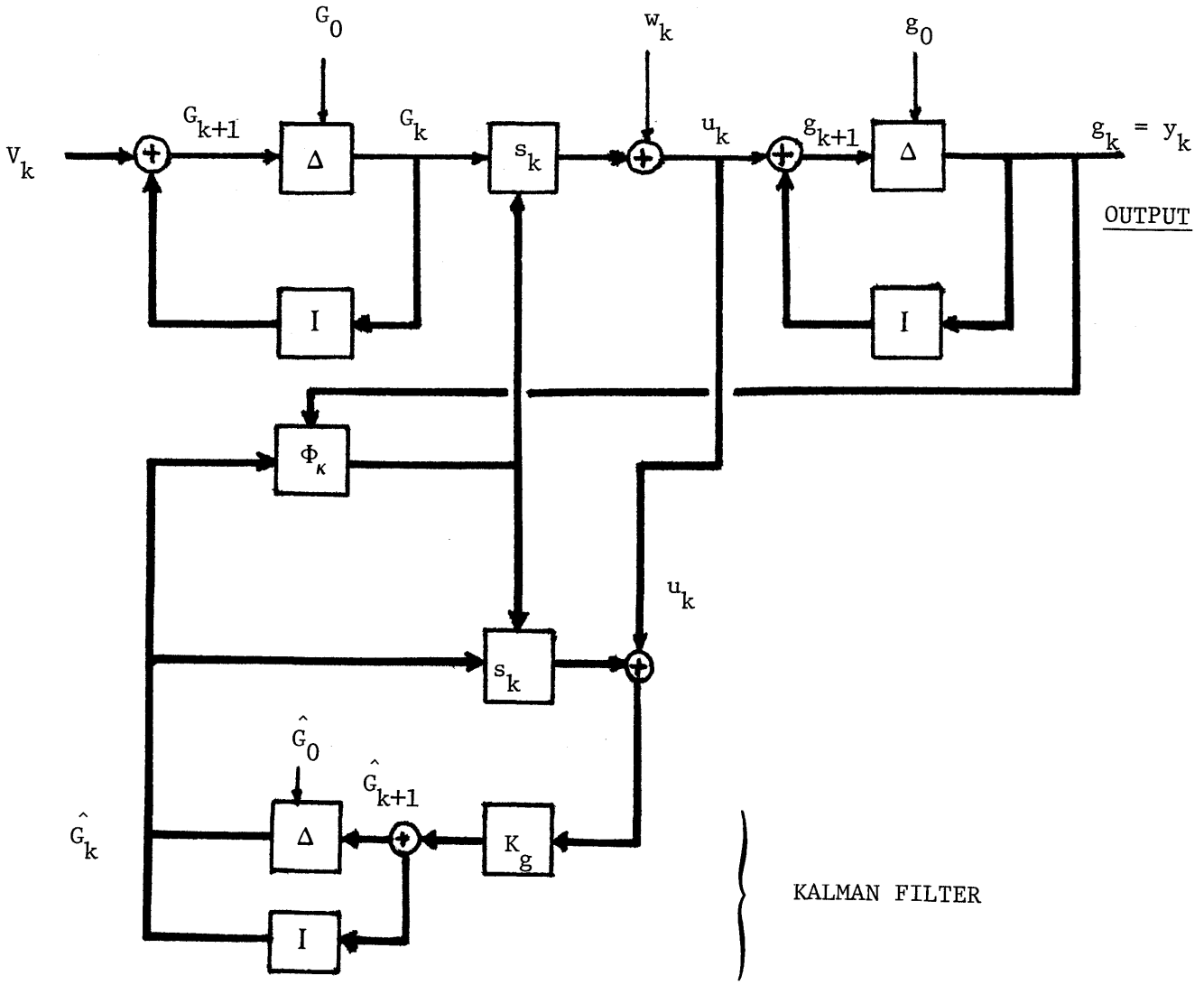
$$x^* - x_k = - [G_k]^{-1} g_k$$

a reasonable guess for $\Phi_k[\hat{G}_k, g_k]$ is the following,

$$\Phi_k[\hat{G}_k, g_k] = -[\hat{G}_k]^{-1} \qquad [31]$$

if $\hat{G}_k$ is non-singular
for each k =0,1,2,...

This whole last dicussion then can be condensed in a bloc diagram:



CHAPTER IV

The set estimation approach applied to function minimization.

1-Introduction.

In the previous chapter we presented a Gaussian model to describe the evolution of the Hessian of a function along each point of a minimization algorithm, and then used the techniques provided by linear filtering theory to construct a recursive sequence of estimates $(\hat{G}_k)$.

Probably a more realistic method would consist of considering that V_k and w_k, respectively the input and observation perturbations of the system considered, have unknown statistics but remain bounded. This will be achieved by using the so-called recursive set estimation procedure originally due to Schweppe [26] and applied later to the control field mainly by Bertsekas [3].

The basic idea of this procedure is to combine knowledge of the systems dynamics with the observations, as well as the bounds, in order to specify a time-varying set in the state-space, which always contains the true state of the system. Hence, the actual estimate consists of a set in state-space rather than a vector (or a matrix as in the case we considered). Unfortunately, this set is usually too hard to characterize and this led Schweppe to introduce the concept of minimal bounding ellipsoids containing the previously obtained set.

The idea of applying the set estimation technique to obtain new gradient algorithms is not new, since it was introduced for the first time by Thomas (see 28); nevertheless, for the sake of completeness we shall devote a whole chapter to it, emphasizing more the model building aspects than the derivation of each result.

In section 2, a short review of some useful concepts, such as the one of support functions for closed convex sets will be introduced and particularly applied to the case of ellipsoidal sets.

In section 3, a discussion will be directed around the correspondence between unknown but bounded models and ordinary linear Gaussian ones. The analogy between correlated Gaussian noises and unknown perturbations bounded by skewed ellipsoids will be particularly emphasized. This last part of the discussion will enable us to choose a model of the skewed ellipsoid type to represent function minimization problems.

Finally, in section 4, the sequences (G_k) and (Π_k) will be derived according to Bertsekas [3] and Thomas' works [28].

2-The basic concepts of set estimation theory.

Consider any subset Ω of R^n. The first notion which we will need is that of a support function.

Definition:

Given a non-empty set Ω in R^n, the support function of Ω, is a mapping from R^n to R defined by:

$$\eta: R^n \rightarrow R \quad \text{s.t.} \quad \eta(y) = \underset{z\varepsilon\Omega}{\text{Sup}} \; \langle y,z\rangle \; [\text{or just } \underset{z\varepsilon\Omega}{\text{Sup}} \; y^T z] \qquad [1]$$

and the following Lemma, due to Rockafellar (see Rockafellar p. 113),

Lemma 1:

Let $\Omega \subseteq R^n$ be non-empty, closed and convex. Then Ω is completely determined by its support function. In particular, Ω may be defined by,

$$\Omega = \left\{ z\varepsilon R^n : \langle y,z\rangle \leq \eta(y) \quad \text{for all } y\varepsilon R^n \right\} \qquad [2]$$

Now, consider the case of a closed, compact ellipsoid Ω with equation,

$$\Omega = \left\{ x : (x - x_c)^T \Gamma^{-1} (x - x_c) \leq 1 \right\} \subset R^n \qquad [3]$$

where x_c is the center and Γ is a positive definite, symmetric matrix describing its excentricity. The support function of Ω is given by the following Lemma.

Lemma 2:

The support function of the ellipsoid Ω [3] is given by,

$$\eta(y) = \langle y,x_c\rangle + \langle y, \Gamma y\rangle^{\frac{1}{2}} \qquad [4]$$

Proof:

From the definition of η one may write,

$$\eta(y) = \underset{x\varepsilon\Omega}{\text{Sup}} \; \langle y,(x - x_c)\rangle + \langle y,x_c\rangle$$

As Γ is a positive definite matrix, there is a self-adjoint $\Gamma^{\frac{1}{2}}$ s.t. $\Gamma = \Gamma^{\frac{1}{2}} \Gamma^{\frac{1}{2}}$.

From [3] and the Cauchy-Schwartz inequality, one now deduces that,

$$\begin{aligned} \langle y, x - x_c \rangle &= \langle \Gamma^{\frac{1}{2}} y, (\Gamma)^{-\frac{1}{2}} (x - x_c) \rangle \\ &\leq \langle y, \Gamma y \rangle^{\frac{1}{2}} \quad \langle (x - x_c), \Gamma^{-1} (x - x_c) \rangle \\ &\leq \langle y, \Gamma y \rangle^{\frac{1}{2}} \end{aligned}$$

for all $y \in R^n$.

Now by [3] $z = \langle y, \Gamma y \rangle^{-\frac{1}{2}} \Gamma y + x_c$ is contained in Ω, and hence verifies $\langle y, z - x_c \rangle = \langle y, \Gamma y \rangle^{\frac{1}{2}}$ Q.E.D. [5]

Lemma 3:

The support function of the vector sum of two closed, convex sets Ω_1 and Ω_2 is the sum of the support functions of each of them.

Proof:

Let $S = \Omega_1 + \Omega_2 = \left\{ z : z = x_1 + x_2 \text{ with } x_1 \in \Omega, \text{ and } x_2 \in \Omega_2 \right\}$

then S has as support ,

$$\eta_s(y) = \sup_{v \in S} v^T y = \sup_{\substack{x_1 \in \Omega_1 \\ x_2 \in \Omega_2}} (x_1^T y + x_2^T y) = \sup_{x_1 \in \Omega_1} x_1^T y + \sup_{x_2 \in \Omega_2} x_2^T y$$

Q.E.D.

Consider now the following linear system:

$$\begin{cases} \dfrac{dx(t)}{dt} = F(t)x(t) + G(t)u(t) & [6] \\ x_0 \text{ given} \\ \\ y(t) = H(t)x(t) + w(t) & [7] \end{cases}$$

for $t \geq 0$, $u(t) \in \Omega_1$, $w(t) \in \Omega_2$, $x_0 \in \Omega_0$,

where u(t) and w(t) are unknown but bounded perturbations belonging respectively at each instant t to the closed, convex sets Ω_1 and Ω_2 and where also the initial state x_0 belongs to some given closed, convex set Ω_0. If, furthermore, $\Phi(t,s)$ corresponds to the transition matrix associated to the matrix F(t), the solution of equation [6] can be written as,

$$\Phi(t,s) \in L(R^n) \qquad \text{for } (t,s) \in [0,T]$$

$$x(t) = \Phi(t,0)x_0 + \int_0^t \Phi(t,s)G(s)u(s)ds, \qquad t \geq 0 \qquad [8]$$

As we implicitly assume that F(t) is such that $\Phi(.,.)$ is a bounded linear operator, the set,

$$\Omega_\Phi(t) = \left\{ x \text{ s.t. } x = \Phi(t,0)z \text{ , } z\in\Omega_0 \right\} , \qquad [9]$$

being the image of a compact set, remains closed and convex for any instant t.

Now the set,

$$\Omega_u(t) = \left\{ x \text{ s.t. } x = \int_0^t \Phi(t,s)G(s)uds, \; u\in\Omega_1 \right\} , \qquad [10]$$

remains also closed and convex for all t because of the linearity of the mapping $u \to t$, and if $\Omega(t)$ represents the set of all possible reachable states, we clearly see that,

$$\Omega(t) = \Omega_u(t) + \Omega_\Phi(t) \qquad [11]$$

Consider now also the set of all possible states which are coherent with the observation y(t) at time t,

$$\Omega_{obs}(t) = \left\{ x : y(t) - H(t)x(t)\in\Omega_2(t) \right\} \qquad [12]$$

By definition the set estimate will be the intersection,

$$\Omega_{\varepsilon s}(t) = \Omega(t) \cap \Omega_{obs}(t) \qquad [13]$$

If Ω_1 and Ω_2 are ellipsoids, it is easy to verify that $\Omega_u(t)$ and $\Omega_\Phi(t)$ are also ellipsoids, but unfortunately, as the use of Lemma 2 proves it, $\Omega(t)$ has no reason for remaining also an ellipsoid (the sum of the support functions of the two ellipsoids does not conserve the structure $y^T x_0 + [y^T \Gamma y]\frac{1}{2}$, except for some very special cases).

In the same way, although $\Omega_2(t)$ is taken to be an ellipsoid, $\Omega_{obs}(t)$ is no longer an ellipsoid, nor the intersection $\Omega_{\varepsilon s}(t)$. However, if one bounds each of these sets by bigger ellipsoids, Schweppe [see 26] proved that a recursive formula could be carried out for the centers $\hat{x}_t$ and the kernels $\Gamma_{(t)}$ of the ellipsoid $\hat{\Omega}_{\varepsilon s}(t)$ containing the "true" set estimate $\Omega_{es}(t)$.

Now, before starting to compare ordinary linear Gaussian models with such unknown but bounded perturbation models, a last comment must be made about the case when no perturbation affects the observation variable y(t). In this case, in fact, the set of all possible states coherent with the observations y(t), that is $\Omega_{obs}(t)$ reduces to a hyperplane of equation,

$$\Omega_{obs}(t) = \left\{ x : y(t) - H(t)x = 0 \right\} \qquad [14]$$

We shall see in the next section that this case is of particular interest as the intersection of the set $\Omega(t)$ with the hyperplane $\Omega_{obs}(t)$ will be much easier to compute. Furthermore, if $\Omega(t)$ is included in an ellipsoid $\hat{\Omega}(t)$, the intersection $\hat{\Omega}(t) \cap \hat{\Omega}_{obs}(t)$ will also be

an ellipsoid (the intersection of an ellipsoid with a hyperplane of lower dimension always being an ellipsoid), but possibly degenerate. [see Appendix 2]

Finally, in the remaining part of this chapter, we shall consider the case where t is a discrete variable taking values 0,1,2,...

3-Three types of estimation problems on unknown but bounded models.

Take the case of a linear discrete-time dynamical system described by,

$$x_{k+1} = A_k x_k + B_k u_k \qquad k = 0,1,2,... \qquad [15]$$

on which noise-corrupted measurements are performed,

$$z_k = C_k x_k + w_k \qquad [16]$$

$x_k \varepsilon R^n$ is the state of the system, $u_k \varepsilon R^r$ is an input disturbance vector and $w_k \varepsilon R^p$ is the measurement noise vector. A_k, B_k, C_k have the appropriate dimensions and N corresponds to the time horizon of this problem.

In this section, the recursive ellipsoidal state set estimates, $\hat{\Omega}_{es}(k)$, shall be constructed, with the following three different types of constraints on the unknown quantities x_0, w_k, u_k.

The first type of constraint is the "energy constraint" type described by,

$$x_0^T \Psi^{-1} x_0 + \sum_{k=1}^{N} (u_{k-1}^T Q_{k-1}^{-1} u_{k-1} + w_k^T R_k^{-1} w_k) \leq 1 \qquad [17]$$

where Ψ, Q_k, R_k are given positive definite symmetric matrices for all k = 0,1,2,... .

Practically, the second type is the more important case, which will be designated as the "separate, instantaneous, constraint type". The uncertain quantities are constrained at each instant of time to lie within the ellipsoids,

$$\begin{cases} x_0^T \Psi^{-1} x_0 \leq 1 \\ u_{k-1}^T Q_{k-1}^T u_{k-1} \leq 1 \\ w_k^T R_k^{-1} w_k \leq 1 \end{cases} \qquad k = 1 \ldots N \qquad [18]$$

Finally, the third type, which will be necessary to study a gradient algorithm, will be referred to as the "global instantaneous constraint" type described by,

$$x_0^T \Psi^{-1} x_0 \leq 1 \qquad (u_k^T, w_k^T) \begin{pmatrix} Q_k & S_k^T \\ S_k & R_k \end{pmatrix}^{-1} \begin{pmatrix} u_k \\ w_k \end{pmatrix} \leq 1 \qquad [19]$$

where $\begin{pmatrix} Q_k & S_k^T \\ S_k & R_k \end{pmatrix}$ is required to be globally positive definite for each instant of time k.

The first two types have already been studied by Bertsekas in his thesis [3, see in particular Chapter IV]. He was able to derive in both cases a recursive procedure to construct at each time k an ellipsoidal estimate for the set of all states, consistent with the measurements z_k. Furthermore, his resulting estimator, though similar to the one proposed by Schweppe [26], has two advantages-

his gain matrices are precomputable , and a time invariant structure exists to the problem, obtained as the limit, when N the time horizon becomes infinitely long.

The third case shown to be related to the correlated Gaussian-input, Gaussian-observation noise model. The estimation problem corresponding to this case will be solved by noticing that such systems are in fact equivalent to perfect observation situations with, however, noisy inputs driving them.

a) The energy constraint case:

The way to solve the problem is to consider at each instant of time k, the linear mappings $L_k : R^{n+N(r+p)} \to R^n$ and $D_k : R^{n+N(r+p)} \to R^{kp}$ such that,

$$\text{for} \quad v^T = (x_0^T, u_0^T, u_1^T, \ldots u_{N-1}^T, w_1^T, w_2^T, \ldots w_N^T) \qquad [20]$$

$$\text{and} \quad \zeta_k^T = (z_1^T, z_2^T, \ldots z_k^T) \qquad [21]$$

$$x_k = L_k v \qquad [22]$$

$$\zeta_k = D_k v \qquad [23]$$

for all k = 0,1,2,...N

The energy constraint [17] implies that the vector v belongs to the ellipsoid,

$$V = \left\{ v \in R^{n+N(r+p)} : v^T M^{-1} v \leq 1 \right\} \qquad M = \begin{bmatrix} {}^{\Psi}Q_0 & & 0 \\ & Q_{n-1} & \\ 0 & & R_1 \\ & & & R_n \end{bmatrix} \in L(R^{n+N(r+p)})$$

Calling $\hat{V}_k$ the set of all possible vectors v consistent with the measurement vector ,

$$\zeta_k : \hat{V}_k = \left\{ v : \zeta_k = D_k v,\ v \varepsilon V \right\} .$$

One can notice that as V is an ellipsoid and $\hat{V}_k$ is the intersection of V with manifold $\left\{ v : \zeta_k = D_k v \right\}$, $\hat{V}_k$ is also an ellipsoid, as well as $\hat{\Omega}_{es}(k) = L_k \hat{V}_k$ obtained through the linear mapping L_k.

The final result can be found in Bertsekas [3, for example, see Proposition 4-2],

$$\hat{\Omega}_{es}(k) = \left\{ x : (x - \hat{x}_k)^T \ \Sigma_k^{-1} (x - \hat{x}_k) \leq 1 - \delta^2(k) \right\}_{k = 0,1,\ldots N} \qquad [24]$$

where Σ_k is given recursively by the Riccati equation,

$$\begin{cases} \Sigma_k^{-1} = C_k^T R^{-1} C_k + \left[A_{k-1} \Sigma_{k-1} A_{k-1}^T + B_{k-1} Q_{k-1} B_{k-1}^T \right]^{-1} \\ \Sigma_0 = \Psi \end{cases} \qquad [25]$$

and

$$\begin{cases} \hat{x}_{k+1} = A_k \hat{x}_k + \Sigma_{k+1} C_{k+1}^T R_{k+1}^{-1} \left[z_{k+1} - C_{k+1} A_k \hat{x}_k \right] \\ \hat{x}_0 = 0 \end{cases} \qquad [26]$$

for

$$\begin{cases} \delta^2(k+1) = \delta^2(k) + \left[z_{k+1} - C_{k+1} A_k \hat{x}_k \right]^T \left[C_{k+1} B_k Q_k B_k^T C_k^T + R_{k+1} \ldots \right. \\ \qquad \left. \ldots + C_{k+1} A_k \Sigma_k A_k^T C_{k+1}^T \right]^{-1} \left[z_{k+1} - C_{k+1} A_k \hat{x}_k \right] \\ \delta^2(0) = 0 \end{cases} \qquad [27]$$

b) The separate instantaneous, constraint problem.

Contrary to the previous case, the problem becomes very difficult when instantaneous constraints defined by inequalities of the form [18] are given for each of the unknown perturbations, and, at least for the moment, no exact solution to this problem has been worked out.

In Schweppe's work [26], the main idea for solving this problem is to bound recursively by ellipsoids the convex sets of all reachable states defined for every instant k in time by the conditions [11-12-13]. In Bertsekas' work [3], however, the same problem is transformed in a first stage into a problem of the energy constraint type, and then, solved by the same methods as in part a). Although the last method leads to some nicer results than the first one, especially in terms of their asymptotic behavior, their exact form will not be discussed here, since their relevance to the present function minimization is quite questionable.

c- The global instantaneously constrained set estimation problem.

Problem:

Consider the dynamical system described by,

$$\begin{cases} x_{k+1} = A_k x_k + B_k u_k \\ z_k = C_k x_k + w_k \end{cases} \qquad k = 0,1,... \qquad [28]$$

with

$$x_k \in R^n, \quad u_k \in R^r, \quad w_k \in R^p, \quad z_k \in R^p$$

and, where A_k, B_k, C_k are known matrices of appropriate dimensions.

The initial state x_0 is unknown but bounded by the ellipsoid,

$$\Omega_0 = \left\{ x \in R^n : (x - \bar{x}_0)^T \Psi_0^{-1} (x - \bar{x}_0) \leq 1 \right\} \qquad [29]$$

where Ψ_0 is a positive definite symmetric matrix.

The perturbations u_k, w_k are jointly bounded for each instant in time k, by the ellipsoids,

$$\Omega_{1k} = \left\{ z \in R^{p+r} : (z_1^T, z_2^T) \begin{pmatrix} Q_k & S_k^T \\ S_k & R_k \end{pmatrix}^{-1} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \leq 1 \right\} \qquad [30]$$

$$k = 0,1,\ldots$$

where $\begin{pmatrix} Q_k & S_k^T \\ S_k & R_k \end{pmatrix} > 0$ for all k.

Find the recursive procedure to build the ellipsoid set estimates of the state x_k for each time k, given the previous observations $z_0, z_1, \cdots z_{k+1}$.

Equations [28] have to be transformed, by defining z_k such that,

$$z_k = y_{k+1} - y_k \in R^p \qquad k = 0,1,\ldots$$

and

$$\theta_k = y_k \in R^p \qquad [31]$$

The equation describing the dynamical evolution of the system is now,

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{bmatrix} A_k & 0 \\ C_k & I \end{bmatrix} \begin{pmatrix} x_k \\ y_k \end{pmatrix} + \begin{bmatrix} B_k & 0 \\ 0 & I \end{bmatrix} \begin{pmatrix} u_k \\ w_k \end{pmatrix} \qquad k=0,1,2,.. \qquad [32]$$

whereas, the observations are described only by,

$$\theta_k = [\ 0 \quad I\ ] \begin{pmatrix} x_k \\ y_k \end{pmatrix} \quad \text{and} \quad z_k = \theta_{k+1} - \theta_k \qquad k=0,1,2,...$$

The state-space has a dimension augmented from n to n+p. Thus also the problem which has been initially formulated, in equations ([28] - [30]) is now transformed into an "instantaneous constraint" type of problem but with perfect observations.

Assuming that at time k, $(x_k^T,\ y_k^T)$ belongs to an ellipsoid having as support function,

$$\eta_k(v_1,v_2) = <(v_1,v_2)\ ;\ \begin{pmatrix} \hat{x}_k \\ \hat{y}_k \end{pmatrix} > \ +\ [\ <(v_1,v_2);\ \Sigma_k \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} >\ ]^{\frac{1}{2}} \qquad [33]$$

The support function of the set containing all possible values taken by $(x_{k+1}^T,\ y_{k+1}^T)^T$, can be computed using Lemma 3,

$$\eta_{k+1}(v_1,v_2) = <(v_1,v_2); \begin{bmatrix} A_k & 0 \\ C_k & I \end{bmatrix} \begin{pmatrix} \hat{x}_k \\ \hat{y}_k \end{pmatrix} > + <(v_1,v_2); \begin{bmatrix} A_k & 0 \\ C_k & I \end{bmatrix} \Sigma_k \begin{bmatrix} A_k^T & C_k^T \\ 0 & I \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} >^{\frac{1}{2}} \ \ldots$$

$$\ldots + <(v_1,v_2); \begin{bmatrix} B_k & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} Q_k & S_k^T \\ S_k & R_k \end{bmatrix} \begin{bmatrix} B_k^T & 0 \\ 0 & I \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} >^{\frac{1}{2}}$$

Using now a majoration technique also used by Thomas [28] (see equation 2.38), the result is that all possible states $(x_{k+1}^T, y_{k+1}^T)^T$ are also contained in an ellipsoid with support function ,

$$\tilde{\eta}_{k+1}(v_1,v_2) = \langle (v_1,v_2);\ (A_k\hat{x}_k,\ C_k\hat{x}_k + \hat{y}_k)\rangle \ \ldots \qquad [34]$$

$$\ldots + \ \langle (v_1,v_2);\ \begin{bmatrix} A_k & 0 \\ C_k & I \end{bmatrix} \Sigma_k \begin{bmatrix} A_k^T & C_k^T \\ 0 & I \end{bmatrix} + \begin{bmatrix} B_k Q_k B_k^T & 0 \\ 0 & R_k \end{bmatrix} (v_1,v_2)\rangle^{\frac{1}{2}}$$

In order now to obtain the set estimate of $(x_{k+1}^T\ ,\ y_{k+1}^T)^T$ which is also the set of all possible values consistent with the observation θ_{k+1} (taken at the same instant of time), one has to compute the intersection of the ellipsoid defined above, with the manifold,

$$\theta_{k+1} = \left\{ \theta \varepsilon R^p \ :\ \theta = y_{k+1} \right\} \qquad [35]$$

This intersection is also an ellipsoid with center defined by,

$$\begin{cases} \hat{x}_{k+1} = [\ \Sigma_{11} - \Sigma_{21}\Sigma_{22}^{-1}\Sigma_{12}]\ \ \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}\ \ [z_k - C_k\hat{x}_k] \\ \hat{x}_0 = \bar{\bar{x}}_0 \end{cases}$$

where,

$$\Sigma_{k+1} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} A_k & 0 \\ C_k & I \end{bmatrix} \Sigma_k \begin{bmatrix} A_k^T & C_k^T \\ 0 & I \end{bmatrix} + \begin{bmatrix} B_k Q_k B_k^T & 0 \\ 0 & R_k \end{bmatrix} > 0$$

for k=0,1,...

[36]

and

$$\Sigma_0 = \begin{bmatrix} \Psi_0 & 0 \\ 0 & 0 \end{bmatrix} \qquad \Psi_0 > 0.$$

The ellipsoid itself can be described by,

$$\left\{ x\varepsilon R^n : < (x-\hat{x}_{k+1}),\ [\Sigma_{11}-\Sigma_{21}^{*}\Sigma_{22}^{-1}\Sigma_{12}]^{-1}\ (x-\hat{x}_{k+1})> \leq 1-\gamma_k \right\}$$

where, [37]

$$\gamma_k = < [z_k - C_k\hat{x}_k]\ ;\ [\Sigma_{22}-\Sigma_{12}\Sigma_{11}^{-1}\Sigma_{21}^{*}]\ (z_k - C_k\hat{x}_k) >$$

These equations can be derived directly following the method indicated here or they can also be derived as a limiting case , when the observation noise becomes zero, of Schweppe's results (see [26] pp. 168).

4- A set estimation problem for minimizing a function.

In this section we review, following Thomas' work [28], how the stochastic model of equations (Chapter III-10,11) can be transformed into an unknown but bounded noise model with global instantaneous constraints. Finally, a recursive solution for the estimates $\hat{G}_k$ is stated.

Consider once more the case of a function f: $R^n \to R$ twice continuously differentiable, such that, $\nabla f(x) = g(x)$ and $\nabla f^2(x) = G(x)$. Let x^* be its (possibly local) minimum and let D be some convex neighborhood of x^*. It is finally assumed that,

$$||G(x_k + \theta_1 s_k) - G(x_k + \theta_2 s_k)|| \leq L|\theta_1 - \theta_2| \cdot ||s_k|| \qquad L>0 \qquad [38]$$

for all $x_k, x_k+s_k \varepsilon\ D$ and all $\theta_1, \theta_2\ \varepsilon[0,1]$.

Let us introduce also the continuous mappings:

$$G_k : [0 . 1] \to L(R^n)$$
$$g_k : [0 . 1] \to R^n$$
$$u_k : [0 . 1] \to R^n$$

such that for each k = 0, 1, 2...

$$G_k(\theta) = G(x_k + \theta s_k)$$
$$g_k(\theta) = g(x_k + \theta s_k), \qquad \text{and} \quad u_k(\theta) = g_{k-1}(\theta) - g_{k-1}(0)$$

The there exists (see Lemma 1.1 in [28]) a Lebesgues integrable function $U_k : [0 . 1] \to L(R^n)$ defined by,

$$G_k(\theta) - G_k(0) = \int_0^\theta U_k(t)dt \qquad [39]$$

and

$$||U_k(\theta)|| \leq L||s_k|| \qquad \text{for all } k=0,1,\ldots \text{ and all } \theta\varepsilon[0,1]$$

On the other hand,

$$g_k(\theta) = g_k(0) + \int_0^\theta G_k(t)s_k dt$$

or
$$u_{k-1}(\theta) = \int_0^\theta G_k(0)\, s_k dt \qquad [40]$$

$$u_{k-1}(\theta) = G_k\theta s_k + \int_0^\theta [G_k(t) - G_k(0)]s_k dt$$

for all $\theta\varepsilon[0 , 1]$

Computing the expressions in [39] and [40] at point θ= 1, one gets that,

$$G_{k+1} = G_k + \int_0^1 U_k(t)dt$$
$$u_k = G_k s_k + \int_0^1 [G_k(t) - G_k(0)]s_k dt \qquad \text{for } k=0,1,... \qquad [41]$$

Let G_0 be the unknown but bounded initial state of the system belonging to the set,

$$\Omega_0 = \left\{ G_0 \varepsilon L(R^n) \mid < (G_0 - \hat{G}_0) \mid \Pi_0^{-1} (G_0 - \hat{G}_0) >_F \leq 1 \right\} \qquad [42]$$

which represents its set estimate with $\Pi_0 > 0$.

Comparing equations [41] to [9] we notice that the input noise V_k and the observation noise w_k have for expression,

$$V_k = \int_0^1 U_k(t)dt$$
$$w_k = \int_0^1 [G_k(t) - G_k(0)]s_k dt \qquad \text{with } V_k \varepsilon L(R^n), w_k \varepsilon R^n \qquad [43]$$

The set of all possible values taken by these two perturbations can be determined through condition [39]. However, using the result of Claim 1 in Chapter II, it becomes also clear that [38] is equivalent to,

$$||G(x_k + \theta_1 s_k) - G(x_k + \theta_2 s_k)||_F \leq L^1 |\theta_1 - \theta_2| \cdot ||s_k|| \qquad [44]$$
$$\text{for } k = 0,1,...$$

where , for example, $L^1 = \frac{L}{\sqrt{n}} > 0$

This implies in particular that,

$$||U_k(\theta)||_F \leq L||s_k|| \qquad \text{for all } k = 0,1,... \text{ and } \theta\varepsilon[0,1] \qquad [45]$$

if we simplify the notation by calling L^1, L.

Now clearly, this last condition means that $U_k(.)$ remains in a bounded set which constraints not only V_k to another bounded set but also w_k. One can also say that [45] is the coupling equation relating the perturbation noises V_k, w_k. Similarly, on the Gaussian model a correlation appeared between V_k, w_k. This means also that "global instantaneous constraints", have to appear on those perturbation terms and, therefore, following the conclusions of the previous section, that a realistic model for the system has to include perfect observations and a state of dimension $n^2 \times n$,

$$\begin{cases} \begin{pmatrix} \underline{G} \\ g \end{pmatrix}_{k+1} = \begin{bmatrix} I_{n^2} & 0 \\ S_k & I \end{bmatrix} \begin{pmatrix} \underline{G} \\ g \end{pmatrix}_k + \int_0^1 \begin{pmatrix} U_k(t) \\ [G_k(t)-G_k(0)]s_k \end{pmatrix} dt & k = 0,1,.. \quad [46] \\ z_k = (0 \quad I) \begin{pmatrix} \underline{G} \\ g \end{pmatrix}_k & [47] \end{cases}$$

with

$$\underline{G}_k \in R^{n^2}, \quad \underline{U}_k \in R^{n^2}, \quad g_k \in R^n, \quad \text{and} \quad S_k \underline{G} = G_k s_k$$

with still, $G_0 \in \Omega_0$ defined by [42]

and

$$||U_k(\theta)||_F \leq L||s_k|| \qquad k = 0,1,... \qquad [48]$$

Let us compute the ellipsoids defining the "global instantaneous constraints". Equation [46] can be written in differential form, using [39] and [40],

$$\frac{d}{d\theta}\begin{bmatrix}\underline{G}_k\\ g_k\end{bmatrix}(\theta) = \left[\begin{array}{cc}0 & 0\\ S_k & 0\end{array}\right]\begin{bmatrix}\underline{G}_k\\ g_k\end{bmatrix}(\theta) + \begin{bmatrix}\underline{U}_k(\theta)\\ 0\end{bmatrix}$$

with $g_k(\theta)\varepsilon R^n$, $\underline{G}_k(\theta)\varepsilon R^{n^2}$

or

$$\frac{d}{d\theta}\begin{bmatrix}G_k\\ g_k\end{bmatrix}(\theta) = \left[\Lambda_k\begin{pmatrix}G_k\\ g_k\end{pmatrix}\right](\theta) + \begin{bmatrix}U_k(\theta)\\ 0\end{bmatrix} \qquad [49]$$

with $G_k(\theta) \varepsilon L(R^n)$, for each k = 0,1,..., and $\theta\varepsilon[0,1]$

Let,

$$\Phi_{\Lambda,k}(\tau,\sigma) = L(R^n) \times R^n \rightarrow L(R^n) \times R^n$$

for all $(\tau,\sigma) \varepsilon [0,1] \times [0,1]$

be the state transition mapping of this system corresponding to Λ.

The solution the differential equation [49] is given by,

$$\begin{pmatrix}G_k(\theta)\\ g_k(\theta)\end{pmatrix} = \Phi_{\Lambda,k}(\theta,0)\begin{pmatrix}G_k\\ g_k\end{pmatrix} + \int_0^\theta \Phi_{\Lambda,k}(\theta,\sigma)\begin{pmatrix}U_k(\sigma)\\ 0\end{pmatrix}d\sigma \qquad [50]$$

for all $\theta\varepsilon[0,1]$,

which for $\theta = 1$ is equivalent to [46].

Let us define the two projections of $\Phi_{\Lambda,k}(.,.)$:

$$\begin{cases}\Phi^0_{\Lambda,k}(.,.) : [0,1] \times [0,1] \rightarrow L\,[\,L[R^n] \times R^n\,;\,L[R^n]]\\ \Phi^1_{\Lambda,k}(.,.) : [0,1] \times [0,1] \rightarrow L\,[\,L[R^n] \times R^n\,;\,R^n]\end{cases}$$

then,

$$\begin{cases} \Phi^{0}_{\Lambda,k}(\theta,\sigma) \cdot \begin{pmatrix} G_k(\sigma) \\ g_k(\sigma) \end{pmatrix} = G_k(\sigma) \\ \\ \Phi^{1}_{\Lambda,k}(\theta,\sigma) \cdot \begin{pmatrix} G_k(\sigma) \\ g_k(\sigma) \end{pmatrix} = g_k(\sigma) + (\tau-\sigma)\, G_k(\sigma) s_k \end{cases} \qquad [51]$$

Let $\mathcal{U}_k$ denote the set of all mappings $U_k(.)$ verifying condition [45]. The support function of the set containing all possible values $(V_k^T,\ w_k^T)$ is obtained by taking,

$$\eta_k(H,h) = \sup_{u \in \mathcal{U}_k} \ll (H,h), \int_0^1 \Phi_{\Lambda,k}(1,\sigma) \begin{pmatrix} U_k(\sigma) \\ 0 \end{pmatrix} d\sigma \gg_T \qquad [52]$$

where $\ll\ ,\ \gg_T$ is the inner product on $L(R^n) \times R^n$ defined by,

$$\ll (A,a)\ ,\ (B,b) \gg_T = \langle A,B \rangle_F + a^T b = \mathrm{Tr}\,[A^T B] + a^T b \qquad [53]$$

for all A,B in $L(R^n)$ and all a,b in R^n.

Defining also by $\Phi^{*}_{\Lambda,k}(\tau,\sigma)$ the adjoint mapping of $\Phi_{\Lambda,k}(\tau,\sigma)$ in terms of this inner product, it is easy to verify that,

$$\Phi^{*}_{\Lambda,k}(\tau,\sigma) \begin{pmatrix} H \\ h \end{pmatrix} = \begin{cases} \Phi^{0*}_{\Lambda,k}(\tau,\sigma) \begin{pmatrix} H \\ h \end{pmatrix} = H + (t-\sigma)\, H s_k \\ \\ \Phi^{1*}_{\Lambda,k}(\tau,\sigma) \begin{pmatrix} H \\ h \end{pmatrix} = h \end{cases}$$

Using this equation [52] becomes ,

$$\eta_k(H,h) = \sup_{U \in \mathcal{U}_k} \int_0^1 \ll (H,h) , \Phi_{\Lambda,k}(1,\sigma) \begin{bmatrix} U(\sigma) \\ 0 \end{bmatrix} \gg_T d\sigma$$

$$= \sup_{U \in \mathcal{U}_k} \int_0^1 \ll \Phi^{*0}_{\Lambda,k}(1,\sigma) \begin{pmatrix} H \\ h \end{pmatrix}, \begin{bmatrix} U(\sigma) \\ 0 \end{bmatrix} \gg_T d\sigma$$

and applying Schwartz' and Holder's inequalities,

$$\eta_k(H,h) = \sup_{U \in \mathcal{U}_k} \int_0^1 < \Phi^{*0}_{\Lambda,k}(1,\sigma) \begin{bmatrix} H \\ h \end{bmatrix}, U(\sigma) >_F d\sigma$$

$$\leq \int_0^1 || \Phi^{*0}_{\Lambda,k}(1,\sigma) \begin{bmatrix} H \\ h \end{bmatrix} ||_F \cdot || U(\sigma) ||_F d\sigma$$

$$\leq \left\{ \int_0^1 || \Phi^{*0}_{\Lambda,k}(1,\sigma) \begin{bmatrix} H \\ h \end{bmatrix} ||_F^2 d\sigma \right\}^{\frac{1}{2}} \times \left\{ \int_0^1 ||U(\sigma)||_F^2 d\sigma \right\}^{\frac{1}{2}}$$

using condition [48] ,

$$\rightarrow \eta_k(H,h) \leq L||s_k||.\left\{ \frac{1}{3} ||s_k||^2 ||h||^2 + h^T H s_k + ||H||_F^2 \right\}^{\frac{1}{2}} \qquad [54]$$

Defining now the mapping,

$Q_k : L(R^n) \times R^n \rightarrow L(R^n) \times R^n$

and its projections,

$Q_{k1}: L(R^n) \times R^n \rightarrow L(R^n)$

$Q_{k2}: L(R^n) \times R^n \rightarrow R^n$

such that,

$$Q_k\begin{bmatrix} H \\ h \end{bmatrix} = \begin{cases} Q_{k1}\begin{bmatrix} H \\ h \end{bmatrix} \\ Q_{k2}\begin{bmatrix} H \\ h \end{bmatrix} \end{cases} \quad \text{for all } H\varepsilon L(R^n),\ h\varepsilon R^n$$

one finally obtains that,

$$\begin{cases} Q_{k1}\begin{pmatrix} H \\ h \end{pmatrix} = \frac{1}{2} Rs_k^T + H \\ Q_{k2}\begin{pmatrix} H \\ h \end{pmatrix} = \frac{1}{3} ||s_k||^2 \cdot ||h||^2 + \frac{1}{2} Hs_k \end{cases} \qquad [55]$$

and inequality [54] becomes,

$$\eta_k(H,h) \leq L||s_k|| \left\{ << \begin{pmatrix} H \\ h \end{pmatrix}, \ Q_k \begin{pmatrix} H \\ h \end{pmatrix} >> _T \right\}^{\frac{1}{2}}$$

It is easy to verify that Q_k is a positive definite operator from $L(R^n) \times R^n$ to itself, for each k = 0,1,...

Now, the choice of ,

$$U(\theta) = L||s_k|| \left\{ << \begin{pmatrix} H \\ h \end{pmatrix}, \ Q_k \begin{pmatrix} H \\ h \end{pmatrix} >> _T \right\}^{\frac{-1}{2}} \Phi^*_{\Lambda,k}(1,\theta) \begin{pmatrix} H \\ h \end{pmatrix} ,$$

verifies condition [45]. Therefore, the previous majorant is effectively reachable and

$$\eta_k(H,h) = L||s_k|| \left\{ << \begin{pmatrix} H \\ h \end{pmatrix}, \ Q_k \begin{pmatrix} H \\ h \end{pmatrix} >> \right\}^{\frac{1}{2}} \qquad [56]$$

To conclude this discussion , these results can be condensed

in the following Proposition:

Proposition 1:

The set estimation version of the stochastic problem defined by equations (III - 10,11) is a set estimation problem with global instantaneous constraints described by,

$$\begin{cases} G_{k+1} = G_k + V_k \\ u_k = G_k s_k + w_k \end{cases} \qquad [57]$$

with $G_k, V_k \in L(R^n)$; $s_k, w_k, u_k \in R^n$ for $k = 0,1,\ldots$

$$G_0 \in \Omega_0 = \left\{ G_0 \in L(R^n) \mid \langle G_0 - \hat{G}_0 | \Pi_0^{-1} (G_0 - \hat{G}_0) \rangle_F \leq 1 \right\} \qquad [58]$$

and

$$\begin{bmatrix} V \\ w \end{bmatrix}_k \in \Omega_k = \left\{ z \in L(R^n) \times R^n : [L||s_k||]^{-1} \langle\langle z, Q_k^{-1}(z) \rangle\rangle_T \leq 1 \right\} \qquad [59]$$

with Q_k defined as in [55] .

Proof:

Since Q_k is positive definite it is also non-singular and hence its inverse does exist.

Equation [56] clearly defines the support function of an ellipsoid in $L(R^n) \times R^n$ - see for instance equation [5] - : this ellipsoid Ω_k is centered at (0,0) and has as kernel Q_k.

Q.E.D

Finally , a solution to this problem can be derived by using equations ([35]-[37]) of the previous section and by noticing the following correspondences ,

$$A_k \quad \leftrightarrow \quad I_{n^2}$$

$$B_k \quad \leftrightarrow \quad I_{n^2}$$

$$C_k \quad \leftrightarrow \quad S_k$$

$$\begin{bmatrix} Q_k & S_k^T \\ S_k & R_k \end{bmatrix} \leftrightarrow \quad \text{the operator } Q_k$$

$$\Psi_0 \quad \leftrightarrow \quad \Pi_0$$

The result one obtains is the following -see also Lemma 1.8 in Thomas [28] -

$$\hat{\Omega}_{k+1} = \left\{ G : \langle (G - \hat{G}_{k+1}) , \Pi_{k+1}^{-1} (G - \hat{G}_{k+1}) \rangle_F \leq 1 - \gamma_k \right\} \qquad [60]$$

where the operator Π_{k+1} verifies

$$\Pi_{k+1} G = G P_{k+1} \qquad \text{for all } k = 0,1,...$$

and P_k is defined by,

$$P_{k+1} = [1 + ||s_k||] \left\{ P_k + L^2 ||s_k|| I - \frac{[P_k + \frac{L^2||s_k||}{2} I] s_k s_k^T [P_k + \frac{L^2||s_k||}{2} I]}{s_k^T [P_k + \frac{L^2||s_k||}{2} I] s_k} \right\} \qquad [61]$$

and

$$\hat{G}_{k+1} = \hat{G}_k + \frac{[\, u_k - \hat{G}_k s_k \,]\, s_k^T \,[\, P_k + \frac{L^2 ||s_k||}{2} I]}{s_k^T \,[\, P_k + \frac{L^2 ||s_k||}{2} I \,] s_k} \qquad [61]$$

$$\gamma_k = \frac{||[\, u_k - \hat{G}_k s_k \,]||^2}{L^2 ||s_k|| [\, 1 + ||s_k||] s_k^T [\, P_k + \frac{L^2 ||s_k||}{3} I] s_k} \qquad [63]$$

In fact, we will usually use L = 1 as value for the Lipschitz constant.

The comparison between equations [61] and [62] with the result of the Kalman filter is striking . In particular,for a Lipschitz constant L equal to 1 , the only difference -see equation [16] of Chapter III - occurs in the generation of the sequence (γ_k) , factor characteristic of the set estimation approach. In particular, this means also that the inverse formula for $\hat{G}_k$ will be the same as (III - 23,24,25) if one does not forget to generate P_k according to (63) instead of (III- 16).

This result is now condensed in the following Proposition.

Proposition 2:

If $\bar{H}_k = [\hat{G}_k]^{-1}$ exists and if $\hat{G}_{k+1}$ is generated according to equations (61) and (62), then

$$\bar{H}_{k+1} = \bar{H}_k + \frac{[s_k - \bar{H}_k u_k]\, \bar{d}_k^T \bar{H}_k}{\alpha_k + \bar{d}_k^T [s_k - \bar{H}_k u_k]} \qquad [64]$$

where

$$\alpha_k = \frac{s_k^T [P_k + \frac{||s_k||}{3} I] s_k}{s_k^T [P_k + \frac{||s_k||}{2} I] s_k} < 1 \qquad [65]$$

$$\bar{d}_k = \frac{[P_k + \frac{||s_k||}{2} I]\, s_k}{s_k^T [P_k + \frac{||s_k||}{2} I] s_k} \qquad [65]$$

is the inverse of $\hat{G}_{k+1}$.

Proof:

The proof is clear using the Sherman-Morisson formula (see equation III-26). Q.E.D.

CHAPTER V

A new quasi- Newton type of algorithm.

1-Introduction.

In his thesis [28] Thomas studies an algorithm based on the set estimation ideas of Chapter IV, but without really using the correct gains or using the actual estimates. Instead, he starts by simplifying them to the Broyden type of update (see III-18) and by taking their symmetric version according to Powell's symmetrization technique (see Chapter II, section 3); finally, in order to be sure that the corresponding sequence is non-singular a parameter θ_k must be computed at each point x_k, as the solution of a quadratic equation. The updating formulas he obtains with this method are:

Broyden:
$$\hat{G}_{k+1} = \hat{G}_k + \frac{(u_k - \hat{G}_k s_k)\, d_k^T}{s_k^T d_k} \qquad k = 0,1,\ldots \qquad [1]$$

modified Powell:
$$\hat{G}_{k+1} = \hat{G}_k + \theta_k \frac{(u_k - \hat{G}_k s_k) d_k^T}{s_k^T d_k} + \theta_k \frac{d_k (u_k - \hat{G}_k s_k)^T}{s_k^T d_k} \quad \ldots \qquad [2]$$

$$\ldots - \theta_k^2 \frac{d_k^T (u_k - \hat{G}_k s_k)}{(s_k^T d_k)^2} \cdot d_k d_k^T$$

where $\hat{G}_0$ is taken to be symmetric and the sequence (P_k) is updated as ,

$$\begin{cases} P_{k+1} = ||s_k||\left(I + P_k - (2 - \theta_k)\theta_k \dfrac{d_k d_k^T}{s_k^T d_k}\right) \\ P_0 = I \end{cases} \qquad [3]$$

$$d_k = \left[P_k + \frac{||s_k||}{2} I \right] s_k \qquad [4]$$

$\hat{G}_k$ is then inverted at each step according to :

$$\bar{H}_{k+1} = \bar{H}_k + \frac{[h_k d_k^T \bar{H}_k + \bar{H}_k d_k h_k^T] d_k^T \bar{H}_k \hat{y}_k - h_k h_k^T (d_k^T \bar{H}_k d_k) - \bar{H}_k d_k d_k^T \bar{H}_k h_k^T \hat{y}_k}{(h_k^T \hat{y}_k)(d_k^T \bar{H}_k d_k) - (d_k^T \bar{H}_k \hat{y}_k)^2} \qquad [5]$$

$$\hat{y}_k = u_k + (1 - \theta_k)\theta_k \frac{[u_k - \hat{G}_k s_k]^T s_k d_k}{s_k^T d_k} - (u_k - \hat{G}_k s_k) \qquad [6]$$

$$h_k = \bar{H}_k \hat{y}_k - s_k \qquad \text{for all } k = 0,1,2,.. \qquad [7]$$

and finally, Powell's " dog-leg " strategy [see 22] is used to compute the next point x_{k+1} such that s_k belongs to the plane spanned by g_k and by $-\bar{H}_k g_k$.

The purpose of this chapter will be to present a new method for updating only $\bar{H}_k$ and P_k , in order to avoid the lengthy equations ([5]-[6]-[7]). Furthermore, our solution will at each step rigorously verify the secant equation $s_k = \bar{H}_{k+1} u_k$.

The main articulations of this new algorithm can be described

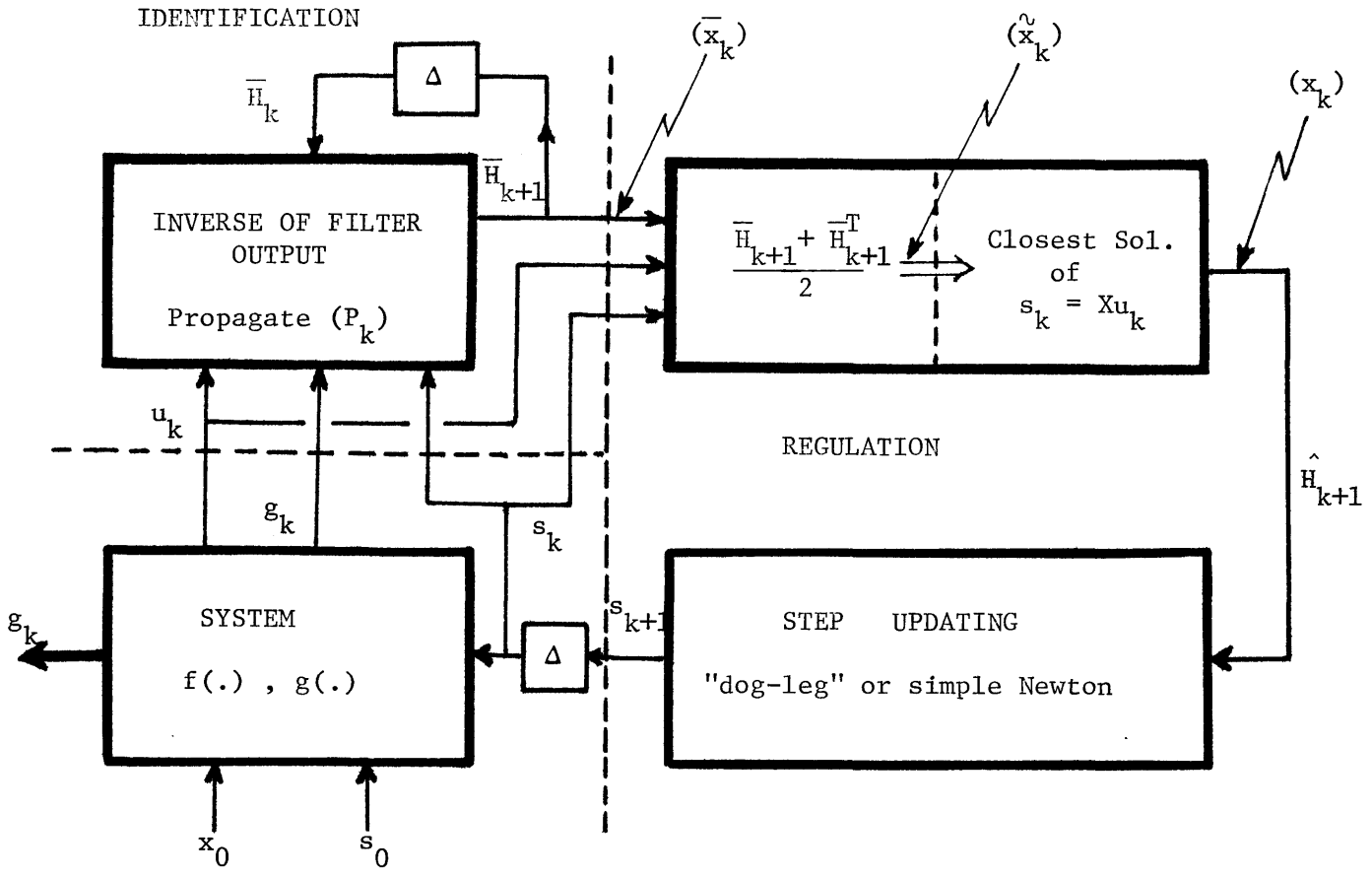
as follows:

a) $\bar{H}_k$ and P_k are updated according to equations (Chapter IV-equations [61] and [64]) and the non-singularity of $\bar{H}_k$ tested for each step K.

b) $\bar{H}_k$ is symmetrized so that the result verifies the secant equation (see Chapter II.). The resulting matrices will be noted $\hat{H}_k$ and will be considered as the actual estimates of the inverse of the Hessian matrix for each point x_k.

c) The last step,finally, is the same as in Thomas' algorithm; that is, the next step is computed also according to a "dog-leg" strategy and no linesearch is necessary for each step k.

This algorithm can be visualized as:



where $(\bar{x}_k)$, $(\tilde{x}_k)$, (x_k) are the sequences of points which would be generated using respectively

$$s_k = -\bar{H}_k g_k \ , \quad s_k = -\left[\frac{\bar{H}_k + \bar{H}_k^T}{2}\right] g_k \ , \text{ or } \quad s_k = -\hat{H}_k g_k$$

to compute the next step.

In section 2 , the convergence properties of the sequence $(\bar{H}_k)$, $\left[\frac{\bar{H}_k + \bar{H}_k^T}{2}\right]$ as well as $(\bar{x}_k)$, $(\tilde{x}_k)$ are analyzed.

In section 3 a short discussion will be held on the singularity problems arising when propagating the sequence $(\hat{G}_k)$.

In section 4, a global description of the algorithm is presented with some partial numerical results interpreted.

2- Convergence properties of the set estimation filter.

For definition of the rates of convergence and a precise treatment of the Q-order convergence, we refer the reader to Ortega and Rheinholt [19-Chapter 9].

For our purpose it is enough to know that a sequence $(x_k) \subset R^n$ converges Q-linearly to x^*, if there is some r in [0,1] and some $k_0 \geq 0$ such that,

$$||x_{k+1} - x^*|| \leq r||x_k - x^*|| \qquad \text{for each } k \geq k_0 \qquad [8]$$

where $||.||$ is an arbitrary vector norm in R^n - usually this will be the Euclidien norm - or the corresponding induced operator norm in $L(R^n)$, the space of real matrices of order n.

Similarly, one says that $(x_k) \subset R^n$ converges Q-superlinearly to

x^* if

$$\lim_{k\to\infty} \frac{||x_{k+1} - x^*||}{||x_k - x^*||} = 0 \qquad \text{for each } k \geq k_0 \qquad [9]$$

The results of this paragraph will treat the updating of the matrices $\bar{H}_k$ according to,

$$\bar{H}_{k+1} = \bar{H}_k + \frac{[s_k - \bar{H}_k u_k]\, \bar{d}_k^T\, \bar{H}_k}{\alpha_k + \bar{d}_K^T [s_k - \bar{H}_k u_k]} \qquad \text{for } k = 0,1,.. \qquad [10]$$

where,

$$\bar{d}_k = \frac{[P_k + \frac{||s_k||}{2} I]\, s_k}{s_k^T [P_k + \frac{||s_k||}{2} I]\, s_k} \qquad [11]$$

and

$$\alpha_k = \frac{s_k^T [P_k + \frac{||s_k||}{3} I]\, s_k}{s_k^T [P_k + \frac{||s_k||}{2} I]\, s_k} < 1 \qquad [12]$$

and the sequence $(P_k,\ k \geq 0\ )$ is generated by,

$$P_{k+1} = (1 + ||s_k||)\ [P_k + ||s_k||\, I - \frac{d_k d_k^T}{\alpha_k s_k^T d_k}\ ] \qquad [13]$$

for $k = 0,1,...$

and P_0 is chosen to be proportional to the identity matrix -

$$P_0 = \sigma^2 I \qquad [14]$$

Of course equation [10] has been shown to be equivalent in the case where $\hat{G}_k$ in non-singular to:

$$\hat{G}_{k+1} = [H_{k+1}]^{-1} = \hat{G}_k + \frac{[u_k - \hat{G}_k s_k]\bar{d}_k^T}{\alpha_k} \qquad k = 0,1,.. \qquad [15]$$

an equation which will be also very useful.

Recall, finally, that to each iteration k, we associate a step $s_k \in R^n$, $s_k \neq 0$ such that ,

$$x_{k+1} - x_k = s_k \qquad [16]$$

and

$$u_k = g_{k+1} - g_k = \nabla f(x_{k+1}) - \nabla f(x_k) \qquad [17]$$

with also,

$$s_k = - \bar{H}_k g_k \qquad [18]$$

/

Consider then the following sequence:

$$\begin{cases} \Pi_{k+1} = (1 + ||s_k||) \left[\Pi_k + ||s_k|| - \dfrac{(2\alpha_k - 1)}{\alpha_k^2} \dfrac{d_k d_k^T}{s_k^T d_k}\right] \\ \Pi_0 = \sigma^2 I \end{cases} \qquad \text{for } k = 0,1,... \qquad [19]$$

Lemma 1:

If (P_k) and (Π_k) are two sequences in $L(R^n)$ generated according to equations [3] and [19] with $\Pi_0 = P_0 = \sigma^2 I$, then for each $k \geq 0$:

$$\Pi_k \geq P_k \geq 0 \qquad [20]$$

Proof:

Consider the difference,

$$P_1 - \Pi_1 = (1 + ||s_0||\ [\ \frac{2\alpha_0 - 1}{\alpha_0^2} - \frac{1}{\alpha_0}\ ]\ \frac{d_0 d_0^T}{s_0^T d_0}$$

$$= 2\ (1 + ||s_0||)\ \frac{d_0 d_0^T}{s_0^T\ d_0}\ \ (\frac{\alpha_0 - 1}{\alpha_0^2}) < 0$$

between $s_0^T d_0 > 0$ and $0 < \alpha_0 < 1 \rightarrow \Pi_1 > P_1 > 0$, the last inequality stems from the fact that $[P_1]^{-1}$, kernel of the set estimation ellipsoid at time k = 0 is positive definite.

Assume now that $0 \leq P_k \leq \Pi_k$ for k=0,1,...m-1 , then,

$$\Pi_m - P_m = (1 + ||s_{m-1}||)\ [\ \Pi_{m-1} - P_{m-1} + 2\ \frac{d_{m-1} d_{m-1}^T}{s_{m-1}^T d_{m-1}}\ (\frac{1 - \alpha_{m-1}}{\alpha_{m-1}^2})$$

$$\geq (\ 1 + \ ||s_{m-1}||)[\ \Pi_{m-1} - P_{m-1}\ ] > 0$$

which proves the recursion. Q.E.D

Notice that equation [13] is nothing else than equation (IV-61) and that, consequently , it defines ellipsoids $\hat{\Omega}_k$ containing all points G_k, or, in other terms, all the possible values of the Hessian of the function to be minimized. These ellipsoids can be in particular inscribed in bigger ellipsoid such that,

$$\hat{\Omega}_k \subset \Omega_k = \left\{ G_k \varepsilon L(R^n) : <(G_k - \hat{G}_k),\ \tilde{\Pi}_k^{-1}\ (G_k - \hat{G}_k)>_F \leq 1 \right\} \qquad [21]$$

see for instance equation [60] in Chapter IV for the definition of $\hat{\Omega}_k$. Recall also that Π_k^{-1} was defined such that , $\tilde{\Pi}_k G_k = G_k P_k$

$$\rightarrow \tilde{\Pi}_k^{-1} [G_k - \hat{G}_k] = [G_k - \hat{G}_k] P_k^{-1} \qquad [22]$$

and that,

$$<A,B>_F = \text{Tr } [AB^T]$$

$$||A||_F = [\text{Tr } [AA^T]]^{\frac{1}{2}}$$

It is now possible to prove the following result:

Corollary 1:

Let $D \subset R^n$ be an open, convex set containing the point x^* such that $g(x^*) = 0$ and let $g:D \rightarrow R^n$ be (Gâteaux) differentiable. Assume that the sequence $(x_k, k \geq 0)$ obtained through [18] is completely contained in D and that g is Lipschitz. Let $\hat{G}_k$ verify [15] and P_k be generated according to [13] with $\hat{G}_0$ given, and with some given symmetric, positive definite matrix $P_0 > 0$; then there exists a constant $\mu > 0$, such that,

$$||G_k - \hat{G}_k||^2 \leq \mu ||P_k|| \qquad [23]$$

for all $k \geq 0$.

Proof:

From the set estimation theory presented in Chapter IV, we already know that if $\hat{G}_k$ and P_k are generated via [15] and [13] then $G_0 \varepsilon \hat{\Omega}_0$ implies that $G_k \varepsilon \hat{\Omega}_k$.

Therefore, it is possible to write for each k that

$$G_o \in \hat{\Omega}_0 \rightarrow G_k \in \hat{\Omega}_k \subset \Omega_k$$

hence,

$$< \Delta G_k, \tilde{\Pi}_k^{-1} \Delta G_k >_F \leq 1 \qquad [24]$$

with the notation, $\Delta G_k = G_k - \hat{G}_k$ [25]

Now, as

$$< \Delta G_k, \tilde{\Pi}_k^{-1} \Delta G_k >_F = < \Delta G_k, \Delta G_k P_k^{-1} >_F$$

$$\rightarrow \text{Tr} [\Delta G_k P_k^{-1} \Delta G_k^T] \leq 1 \qquad [26]$$

But P_k is a positive definite and symmetric matrix for each $k \geq 0$ -see for instance Lemma 1- ; therefore, there exists a non-singular (for example triangular) matrix R_k such that,

$$P_k = R_k^T R_k \quad \text{with } R_k \in L(R^n) \leftrightarrow P_k^{-1} = R_k^{-1} . R_k^{-T}$$

[26] $\rightarrow \text{Tr} [(\Delta G_k R_k^{-1}) (\Delta G_k R_k^{-1})^T] \leq 1$

$$\leftrightarrow \quad ||\Delta G_k R_k^{-1}||_F^2 \leq 1$$

but

$$||\Delta G_k||_F^2 \leq ||\Delta G_k R_k^{-1}||_F^2 \quad ||R_k||_F^2 \leq ||R_k||^2$$

or

$$\text{Tr} [\Delta G_k . \Delta G_k^T] \leq \text{Tr} [P_k]$$

and using the fact that,

$$\text{Tr} [A] \geq ||A|| = \sup_{x \in R^n} \frac{x^T A x}{x^T x} \quad , \text{ if A non-singular,}$$

one obtains that,

$$||\Delta G_k||^2 \leq Tr[P_k] \leq n||\Pi_k|| \qquad \text{for each } k \geq 0$$

from Chapter II- Claim 1.

Q.E.D.

Consider now the sequence $(\bar{x}_k)$ generated according to ,

$$\bar{x}_{k+1} - \bar{x}_k = - \bar{H}_k g_k \qquad [27]$$

the next theorem deals with its convergence properties:

Theorem I:

Let $g:R^n \to R^n$ be (Gâteaux) differentiable in an open convex neighborhood D of x^* for which $g(x^*)= 0$ and $\nabla g(x^*) = G(x^*)$ is non-singular. Assume that for $L \geq 0$, the derivative $\nabla g(x) = G(x)$ verifies,

$$||G(x) - G(y)|| \leq L||x - y|| \qquad \text{for all } x,y \text{ in } D \qquad [28]$$

Therefore,for each $\gamma \geq 0$ and $r \in [0,1]$, there exist positive constants $\delta = \delta(\gamma,r), \varepsilon = \varepsilon(\gamma,r)$, such that for $||x_0 - x^*|| \leq \delta$ and for $\sigma \varepsilon [0,\varepsilon]$ such that $||\hat{G}_0 - G_0|| \leq \gamma\sigma$, the iteration

$$\bar{x}_{k+1} = \bar{x}_k - [\hat{G}_k]^{-1} g_k \qquad [29]$$

with $(\hat{G}_k)$ generated according to [15] and $P_0 = \sigma^2 I$ is well defined for each $k \geq 0$ and the sequence $(\bar{x}_k)$ converges to x^*.

Moreover, $||x_{k+1} - x^*|| \leq r||x_k - x^*||$ for each $k \geq 0$, and the sequences $(||\hat{G}_k||)$ and $(||\bar{H}_k||)$ are uniformly bounded.

Proof:

The proof of this theorem is very similar to the one of Theorem 3.1 in Thomas' thesis [28], but it needs an intermediate result called the Perturbation Lemma (sometimes attributed to Banach) which we first introduce for the sake of completeness.

Perturbation Lemma:

Let $A, C \in L(R^n)$ and assume that A is invertible with $||A^{-1}|| \leq \alpha$. If $||A - C|| \leq \beta$ and $\beta\alpha < 1$ then C is also invertible and

$$||C^{-1}|| \leq \frac{\alpha}{1 - \alpha\beta} \qquad [30]$$

Proof:

see Ortega and Rheinholt pp.45 [19].

Proof of Theorem I (conti.):

The main steps of the proof are the following ones:

choose $\varepsilon > 0$ and $\delta \in [0, \delta]$ for $\mu = \max [\gamma, L, \sqrt{n}]$ such that they satisfy,

$$\eta(1+r) [L\delta + 2\mu\varepsilon] \leq r \qquad [31]$$

$$(1 + \delta + 4\varepsilon^2) \delta \frac{1+r}{1-r} \leq \varepsilon^2 \qquad [32]$$

for $r \in [0,1]$ and $\gamma \geq 0$ given.

Then starting with $||x_0 - x^*|| \leq \delta$ and $||\hat{G}_0 - G_0|| \leq \gamma\sigma \leq 2\mu\varepsilon$ one can easily show that $||x_1 - x^*|| \leq r||x_0 - x^*||$. [33]

Now, if $P_0 = I\sigma^2 \leq I\varepsilon^2$, then as $||P_k||$ has the same type of bounds as $||\Pi_k||$, we get the following,

$$||P_k|| \leq 4\varepsilon^2 \rightarrow ||P_{k+1}|| \leq 4\varepsilon^2 \qquad [35]$$

due to the choice of constants in [31] and [32].

From Corollary 1,

$$||\hat{G}_k - G_k||^2 \leq n||P_k|| \qquad [36]$$

hence, as

$$\mu = \max[\gamma, L, \sqrt{n}] \rightarrow ||\hat{G}_k - G_k||^2 \leq ||P_k|| \mu \leq (2\varepsilon\mu)^2 \qquad [37]$$

and this concludes the induction on

$$||\hat{G}_k - G_k|| \leq 2\mu\varepsilon \qquad \text{for all } k \geq 0 .$$

The last point deals with the sequence itself. Assuming that,

$$||\bar{x}_k - x^*|| \leq \bar{\delta} \qquad \text{for all } k = 0,1,\ldots m-1,$$

from $||\hat{G}_k - G_k|| \leq 2\mu\varepsilon$, and from the fact that $||G_k^{-1}||$ is bounded, by application of the Perturbation Lemma the following is obtained:

$$||\hat{G}_{m-1}^{-1}|| \leq \eta(1+r) \rightarrow ||\bar{x}_m - x^*|| \leq \eta(1+r)[L\delta+2\mu\varepsilon] ||x_{m-1} - x^*||$$

and hence,

$$||\bar{x}_m - x^*|| \leq r||x_{m-1} - x^*|| \qquad \text{Q.E.D.}$$

At this point, only the Q-linear convergence of the sequence $(\bar{x}_k, k \geq 0)$ defined through [29], with $\hat{G}_k$ and P_k generated according to equations [13] and [15] has been proved.

In his thesis Thomas [28] has also proved that a sequence defined by,

$$x_{k+1} = x_k - [\hat{G}_k]^{-1} g_k \qquad [38]$$

and generated according to equation [1] and [19] instead of [15] and [29] was not only Q-linearly, but also Q-superlinearly convergent (see in particular Theorem 3.4 and Corollary 3.5 of [28]).

We shall now prove that the same result is true for updates of the form [29], and hence also, that the use of the true set estimate given by the results of Chapter IV also conserves this former property.

In order to show that our updates lead to a Q-superlinearly convergent method, it will be convenient to apply a result proved by Dennis and More -[6]- . For completeness we present it here:

Theorem II:

Let $g: R^n \to R^n$ be differentiable in the open, convex set D of R^n and assume that $G = \nabla g$ is continuous at some x^* in D and that $G(x^*)$ is non-singular. Let $(\hat{G}_k)$ in $L(R^n)$ be a sequence of non-singular matrices and suppose that for some x_0 in D the sequence $(\bar{x}_k)$ where

$$\bar{x}_{k+1} = \bar{x}_k - [\hat{G}_k]^{-1} g_k \qquad [29]$$

remains in D and converges to x^*. Then $(\bar{x}_k)$ converges Q-superlinearly to x^* and $g(x^*) = 0$ if and only if ,

$$\lim_{k\to\infty} \frac{|| [\hat{G}_k - G(x^*)] [\bar{x}_{k+1} - \bar{x}_k]||}{||\bar{x}_{k+1} - \bar{x}_k ||} = 0 \qquad [39]$$

The following sufficiency condition will lead to the sought after result.

Theorem II:

Let $g:R^n \to R^n$ be a differentiable function on an open convex set D and assume that for some $L \geq 0$ and any $x,y \in D$ the derivative $g'(x) = G(x)$ obeys the condition $||G(x) - G(y)|| \leq L||x-y||$. Assume also that $(\bar{x}_k)$, $(\bar{x}_k + \bar{s}_k)$ are contained in D and that with some $\hat{G}_0 \in L(R^n)$ and positive definite symmetric $P_0 \in L(R^n)$, $(\hat{G}_k)$ is updated according to [15]. Then,

$$\lim_{k\to\infty} \frac{|| [\hat{G}_k - G(x_k)]s_k ||}{||s_k||} = 0 \qquad [40]$$

if $\sum_{k=0}^{\infty} ||s_k||$, is convergent. [41]

Proof:

It is important to notice first the equivalence of [40] and [39] as ,

$$\frac{||[\hat{G}_k - G(x^*)]s_k||}{||s_k||} \leq \frac{||[\hat{G}_k - G(x_k)]s_k||}{||s_k||} + \frac{||[G(x_k) - G(x^*)]s_k||}{||s_k||}$$

$$\frac{||[\hat{G}_k - G(x^*)]s_k||}{||s_k||} \leq \frac{||[\hat{G}_k - G(x_k)]s_k||}{||s_k||} + L||s_k||$$

But $||s_k|| \leq r^k(r+1)||x_0-x^*||$ and consequently, for bounded $||x_0-x^*||$, $\lim_{k\to\infty} ||s_k|| = 0$.

$$\lim_{k\to\infty} \frac{||[G_k - G(x^*)]s_k||}{||s_k||} \leq \lim_{k\to\infty} \frac{||[\hat{G}_k - G(x_k)]s_k||}{||s_k||} \qquad [42]$$

hence, if the right hand side converges to zero , the left hand side will also converge and Q-superlinearity will be obtained through Theorem II.

Now, as we know that (P_k) is updated according to [13], by Corollary I ,

$$||[\hat{G}_k - G(x_k)]^2|| \leq ||\hat{G}_k - G(x_k)||^2 \leq \mu\, ||P_k||$$

for each $k \geq 0$.

From [13] the following inegality is true,

$$Tr[P_{k+1}] \leq [1 + ||s_k||]\,[Tr\,[P_k] + n||s_k|| - \frac{1}{\alpha_k}\frac{||d_k||^2}{d_k^T s_k}] \qquad [43]$$

and, since $d_k^T s_k > 0$ and $\alpha_k > 0$,

$$Tr[P_{k+1}] \leq (1 + ||s_k||)\,[Tr[P_k] + n||s_k||] \qquad [44]$$

Now let

$$\mu_{k+1} = \prod_{j=0}^{k} (1 + ||s_j||) \geq 1 \qquad [45]$$

$$\rightarrow 0 \leq \log\,(\mu_k) = \sum_{j=0}^{k} \log\,(1 + ||s_j||) \leq \sum_{j=0}^{k} ||s_j|| < \infty$$

then,

$$\frac{Tr[P_{k+1}]}{\mu_{k+1}} \leq \frac{Tr[P_k]}{\mu_k} + \frac{n||s_k||}{\mu_k} \leq \frac{Tr[P_k]}{\mu_k} + n||s_k||$$

or if,

$$\Phi_{k+1} = \frac{Tr[P_{k+1}]}{n\,\mu_{k+1}} \rightarrow \Phi_{k+1} \leq \Phi_k + ||s_k||$$

and after summation,

$$\Phi_{k+1} \leq \Phi_0 + \sum_{k=0}^{k} ||s_j|| \qquad [46]$$

As by hypothesis the sum [41] converges, it is possible to conclude that (Φ_k) is bounded as well as $(Tr[P_k])$, because of the boundedness of the μ_k. Hence, (Φ_k) must be convergent-unicity is also clear- as well as $(Tr[P_k])$. Assuming T to be the limit of this last sequence consider equation [13] and choose $\alpha_k \in [0,1]$,

$$\lim_{k\to\infty} \frac{||d_k||^2}{d_k^T s_k} \leq \lim_{k\to\infty} (1 + ||s_k||)\left\{[Tr[P_k] + n||s_k||] - Tr[P_{k+1}]\right\} \qquad [47]$$

and as $Tr[P_k] \rightarrow T$ and $||s_k|| \rightarrow 0$ as $k \rightarrow \infty$

$$\lim_{k\to\infty} \frac{||d_k||^2}{d_k^T s_k} = 0 \qquad [48]$$

Finally, remembering the definition of d_k -see [11]- the Cauchy-Schwartz inequality provides:

$$0 \leq \frac{s_k^T P_k s_k}{||s_k||^2} \leq \frac{d_k^T s_k}{||s_k||^2} \leq \frac{||d_k||^2}{d_k^T s_k}$$

as the right hand side converges to zero when $\sum_{k=0}^{\infty} ||s_k||$ converges, it can be concluded that,

$$\lim_{k\to\infty} \frac{||[\hat{G}_k - G(x_k)]s_k||^2}{||s_k||^2} = 0 \qquad \text{Q.E.D.}$$

Corollary II:

Let $g:R^n \to R^n$ be (Gateaux) differentiable in an open, convex neighborhood D of x^* for which $g(x^*) = 0$ and $g'(x^*) = G(x^*)$ is non-singular. Assume that for $L \geq 0$, the derivative $\nabla g(x) = G(x)$ verifies,

$$||G(x) - G(y)|| \leq L||x-y|| \qquad \text{for each } x,y \text{ in } D,$$

then for each $\gamma \geq 0$ and $r\varepsilon(0,1)$, there exist positive constants $\delta = \delta(\gamma,r)$, $\varepsilon = \varepsilon(\gamma,r)$ such that for $||x_0 - x^*|| \leq \delta$ and for $\sigma\varepsilon(0,\varepsilon)$, $||\hat{G}_0 - G_0|| \leq \gamma\sigma$ the iteration,

$$\bar{x}_{k+1} = \bar{x}_k - [\bar{H}_k]g_k = \bar{x}_k - [\hat{G}_k]^{-1}g_k$$

with $(\hat{G}_k)$ generated according to [19] and P_k to [13] is well defined for each k and the sequence $(\bar{x}_k)$ converges Q-superlinearly to x^*.

Proof:

We have only to verify that $\sum_{k=0}^{\infty} ||s_k|| < \infty$ converges in this case; however,

$$||\bar{x}_{k+1} - x^*|| \leq r||\bar{x}_k - x^*||$$

by Theorem I

$$\rightarrow \quad ||s_k|| = ||x_{k+1} - x_k|| \leq (1+r)||x_k - x^*||$$

$$\leq (1+r)r^k ||x_0 - x^*|| \leq (1+r)r^k \delta .$$

$$\sum_{k=0}^{\infty} ||s_k|| \leq (1+r)\delta \sum_{k=0}^{\infty} r^k$$

and as by construction $r \in [0,1]$,

$$0 \leq \sum_{k=0}^{\infty} ||s_k|| \leq \frac{(1+r)}{(1-r)} \delta$$

The sequence (S_k) with $S_k = \sum_{k=0}^{k} ||s_j||$ is monotonely increasing and has an upper bound; hence, (S_k) converges.

By Theorem III it becomes clear that $(\bar{x}_k)$ is a Q-superlinearly convergent sequence in R^n. Q.E.D.

Up to this point, we succeeded only in proving that a sequence $(\bar{x}_k, k \geq 0)$ of points x_k generated when using directly the output of the filter as actual estimate of the Hessian matrices, was Q-superlinearly convergent. The second half of this section will deal with the sequence $(\tilde{x}_k, k \geq 0)$, that is , the sequence of points generated with as estimate for the Hessian inverse $\frac{\bar{H}_k + \bar{H}_k^T}{2}$ (see also Fig. 2).

For this purpose let us introduce the following notation:

-All variables indexed by 0 will refer to the previous construction, i.e. $(\bar{x}_k^o,\ k \geq 0)$ for example, corresponds to using $\hat{G}_k$ and $\bar{H}_k$ generated by [10-15].

-All variables indexed by 1 will refer to the same type of construction but with $\hat{G}_k^T$ and $\bar{H}_k^T$ instead of $\hat{G}_k$ and $\bar{H}_k[ex(\bar{x}_k^o)]$ otherwise , $\hat{G}_k$ and $\bar{H}_k$ will still be generated according to [10-15].

The following result can be then derived:

Theorem IV:

Let $g:R^n \to R^n$ be (Gateaux) differentiable in an open, convex neighborhood D of x^* for which $g(x^*) = 0$ and for which $g'(x^*) = G(x^*)$ is non-singular. Assume that for $L \geq 0$, the derivative $g'(x)=G(x)$ verifies:

$$||G(x) - G(y)|| \leq L||x-y|| \qquad \text{for any } x,y \text{ in } D$$

then, for each $\gamma \geq 0$ and $r\varepsilon(0,1)$, there exist positive constants $\delta = \delta(\gamma,r)$, $\varepsilon = \varepsilon(\gamma,r)$ such that for $||x_0 - x^*|| \leq \delta$ and for $\sigma\varepsilon(0,\varepsilon)$, $||\hat{G}_0 - G_0|| \leq \gamma\sigma$, the iteration

$$\tilde{x}_{k+1} = \tilde{x}_k - \left[\frac{\bar{H}_k + \bar{H}_k^T}{2}\right] g_k \qquad [49]$$

where $\bar{H}_k = [\hat{G}_k]^{-1}$ and $(\bar{H}_k)$ is generated according to [10] with $P_0 = \sigma^2 I$, being well defined for each $k \geq 0$ and for the sequence $(\tilde{x}_k)$ converging to x^*.

Moreover, $||\tilde{x}_{k+1} - x^*|| \leq r||\tilde{x}_k - x^*||$, for each $k \geq 0$ and the sequences $(||\hat{G}_k||)$ and $\left(\left|\left|\frac{\bar{H}_k + \bar{H}_k^T}{2}\right|\right|\right)$ and uniformly bounded.

Proof:

The first points to remark are that if $\bar{H}_k = [\hat{G}_k]^{-1}$, then $\hat{H}_k^T = [\hat{G}_k^T]^{-1}$ and also that $||A|| = ||A^T||$ for all $A \in L(R^n)$. This means in particular that if $(\hat{G}_k)$ verifies the conditions of Corollary I, so does $(\hat{G}_k^T)$. Furthermore, consider the two sequences generated respectively as,

$$\begin{cases} \bar{x}_{k+1}^o = \bar{x}_k^o - \bar{H}_k g_k^o \\ \bar{x}_o^o = x_o \end{cases} \qquad [50]$$

and

$$\begin{cases} \bar{x}_{k+1}^1 = \bar{x}_k^1 - \bar{H}_k^T g_k^1 \\ \bar{x}_0^1 = x_0 \end{cases} \qquad [51]$$

using then Theorem I with $\delta = \text{Inf}(\delta^0, \delta^1)$ and $\varepsilon = \text{Inf}(\varepsilon^0, \varepsilon^1)$, such that equations [31,32] are still satisfied, the two sequences $(\bar{x}_k^o)$ and $(\bar{x}_k^1)$ will verify,

$$|| \bar{x}_{k+1}^1 - x^* || \leq r || \bar{x}_k^1 - x^* || \qquad [52]$$

$$|| \bar{x}_{k+1}^0 - x^* || \leq r || \bar{x}_k^0 - x^* || \qquad [53]$$

so equivalently,

$$|| \bar{x}_{k+1}^0 - \bar{x}_k^0 || \leq r^k (1 + r)\delta \qquad [54]$$

$$|| \bar{x}_{k+1}^1 - \bar{x}_k^1 || \leq (1 + r) r^k \delta \qquad [55]$$

Now, starting at k=0 and choosing for $\tilde{x}_k = \dfrac{\bar{x}_k^0 + \bar{x}_k^1}{2}$, at each step k, if the recursion defined at [49] is used to generate $(\tilde{x}_k)$, it is

clear that,

$$||\tilde{x}_{k+1} - \tilde{x}_k|| \leq \frac{1}{2}||\bar{x}^o_{k+1} - \bar{x}^o_k|| + \frac{1}{2}||\bar{x}^1_{k+1} - \bar{x}^1_k|| \leq r^k(1+r)\delta$$

$\leftrightarrow$ and as $\tilde{x}_0 = x_0$ is such that $||x_0 - x^*|| \leq \delta$

$$\rightarrow \quad ||\tilde{x}_{k+1} - x^*|| \leq ||\tilde{x}_k - x^*||.r \qquad [56]$$

Now, the boundedness of ($||\hat{G}_k||$) has been already proved by Theorem I as well as by the uniform boundedness of ($\bar{H}_k$). Since ,

$||\frac{\bar{H}_k + \bar{H}_k^T}{2}|| \leq ||\bar{H}_k|| < \infty$,we have the desired result. Q.E.D.

At this point , only the convergence of the sequence $(\bar{x}_k)$ and $(\tilde{x}_k)$ has been proved . Unfortunately, the proof of the convergence of (x_k), when using the updating formulas :

$$s_{k+1} = -\hat{H}_{k+1} g_{k+1} \qquad k = 0,1,..$$

and

$$\hat{H}_{k+1} = \frac{[\bar{H}_{k+1} + \bar{H}_{k+1}^T]}{2} + [s_k - \frac{\bar{H}_{k+1} + \bar{H}_{k+1}}{2} u_k] \frac{u_k^T}{u_k^T u_k} + \frac{u_k}{u_k^T u_k} [s_k - \frac{\bar{H}_{k+1} + \bar{H}_{k+1}^T}{2} u_k]^T ..$$

$$.... - u_k^T [s_k - \frac{\bar{H}_{k+1} + \bar{H}_{k+1}^T}{2} u_k] \frac{u_k u_k^T}{(u_k^T u_k)^2} \qquad k = 0,1,..$$

becomes terribly complex and , therefore, the result that the convergence of (x_k) is effectively preserved can only be assumed. The early numerical results show a relatively slow convergence of such a sequence in the

case of a symmetrization procedure using a natural Frobenius norm -that is $G = I$, following the notations of Chapter II- . Some trials were also done by changing at each step the local metric and by using instead of $G = I$, $G = G_k$, where G_k is the true Hessian matrix of the objective function $f(.)$, at each point x_k. The updating formula for $\hat{H}_{k+1}$ thus becomes ,

$$\hat{H}_{k+1} = \frac{\bar{H}_{k+1} + \bar{H}_{k+1}^T}{2} + [s_k - \frac{\bar{H}_{k+1} + \bar{H}_{k+1}^T}{2} u_k] \frac{s_k^T}{s_k^T u_k} + \frac{s_k}{s_k^T u_k}[s_k - \frac{\bar{H}_{k+1} + \bar{H}_{k+1}^T}{2} u_k]^T \cdot\cdot$$

$$\cdot\cdot - u_k^T[s_k - \frac{\bar{H}_{k+1} + \bar{H}_{k+1}^T}{2} u_k] \frac{s_k s_k^T}{(s_k^T u_k)^2} \qquad \text{for } k = 0,1,..$$

In some cases the speed of convergence could be increased by this means, but in the general case no conclusion could be carried out. The conviction of the author is, however, that a variable metric method, very similar in essence to the Fletcher-Powell method, should give rise to fairly good results.

3- Singularities arising in the computation of the sequence (x_k).

One of the problems arising when propagating the inverse H_k of the minimum mean squares estimate $\bar{G}_k$ is that for some values of the coefficient α_k this inverse becomes quite large:

$$H_{k+1} = H_k + \frac{[s_k - \bar{H}_k u_k]\, \bar{d}_k^T \bar{H}_k}{\alpha_k + \bar{d}_k^T [s_k - \bar{H}_k u_k]} \qquad k = 0,1,2,..$$

$$d_k = \frac{[P_k + \frac{||s_k||}{2} I]s_k}{s_k^T [P_k + \frac{||s_k||}{2} I] s_k}$$

The method used in the algorithm is to compute at each step the denominator of the previous formula, to compare it to some fixed level -for example 0.1-. If this denominator is larger than the prefixed level, the coefficient α_k conserves its previous value. If the denominator on the contrary, becomes smaller or equal to the previous level, α_k is given the new value α_0 solution of,

$$\alpha_0 = 0.1 - \bar{d}_k^T [s_k - \bar{H}_k u_k] .$$

This means also that the updating procedure for the matrices P_k, corresponding to the covariance during the estimation phasis, is restarted in this way.

4- Description of some computational tricks.

All computations should be carried out using double precision arithmetic.

The structure of the algorithm is the following one:

1) Given $x_0, s_0, P_0 = 1$, compute analytically g_0 and u_0. Then determine $\bar{H}_0$ which should verify the corresponding secant equation.

2) Propagate the matrices P_k and $\bar{H}_k$.

3) Test the singularity of $\bar{H}_k$ and modify α_k, if necessary.

4) Compute $\hat{H}_k$ depending on the procedure chosen (simple symmetrization, or closest symmetrized version using the "natural"

Frobenius norm or closest symmetrized using a variable metric method.)

This ends the identification step of the system. The regulation part uses the classical dog-leg method (see Powell [19]), that is:

5) Compute $$T_k = \frac{g_k^T \hat{H}_k g_k}{||g_k||^2}$$

if $T_k < 0$ $\left\{ s_k = - t g_k : t \geq 0 \right\}$

if $T_k \geq 0$ $\left\{ s_k = - t g_k : 0 \leq t \leq T_k \right\} \cup \left\{ s_k = - (1-\lambda) T_k g_k - \lambda \hat{H}_k g_k : 0 \leq \lambda \leq 1 \right\}$

6) Go back to step 2, unless the length of the gradient obtained is smaller than some prechosen level.

Of course, in order not to have instabilities for functions studied in the neighborhood of some local minimum, the step length s_k should be constrained to remain bounded within some fixed length Δ.

For the purpose of the experimentation, three different computations were run starting at the same initial condition. The first one used only $\frac{\bar{H}_k + \bar{H}_k^T}{2}$ as the estimate of the inverse of the Hessian of the function, the second one used $\hat{H}_k$, and finally the third one used a variable metric version of $\hat{H}_k$, with G_k the true (but unknown) Hessian matrix of f(.) used in the Frobenius norm. The algorithm was run on a simple two-dimensional quadratic function. Convergence was observed in each case, but the relative speed of convergence of each method varied depending on the initial point chosen, no serious conclusion could be carried out relatively to their particular advantages or inconveniences.

Lastly,no special problem was noticed when propagating symmetric matrices like P_k or H_k. If this should arise one of the safest ways of avoiding any numerical instability would be to use a square root filter to propagate $P_k^{\frac{1}{2}}$ instead of P_k. Unfortunately, as $\hat{H}_k$ could not be constrained to remain non-negative definite along the propagation, this method does not work. The only possibility is then to propagate directly the vector $\hat{H}_k g_k$ instead of the whole n x n matrix $\hat{H}_k$.

CONCLUSION

Let us conclude this thesis by two short sections. The first one deals with some advantages of the algorithm previously described as well as with some of the performances one could expect from it when implemented on "difficult" test functions. The second section emphasizes the most interesting conceptual aspects of this work, by discussing some of the perspectives it opens for future research.

A-Discussion of the algorithm

Several attractive features are seen in the proposed algorithm. The first one is its simplicity of implementation, especially compared to the algorithm described by Thomas [28]. The second advantage seems to be the fact that no linear search is needed from one step to the other. This means in particular that this algorithm does not converge in n steps for a quadratic function, like the classical conjugate direction algorithms, but rather that it implies an asymptotic convergence to the optimum. Hence, not having at each step a one dimensional minimization to perform, the computational load is also somewhat lightened.

Several numerical tests were performed on a simple, two dimensional quadratic function. Different initial points, as well

as initial steps $s_0 = x_1 - x_0$, were successively chosen. Even in the case of initial guesses starting in the wrong direction, that is starting in directions which increased the objective function, the minimum was obtained in about five to ten steps with a good precision. Furthermore, this precision seems to be rather sensitive on the maximal step-length allowed for each iteration, as this step-length plays an important role in the actual implementation of the dog-leg procedure.

Finally, the same program was run but with different sequences of estimates for the inverse of the Hessian matrix. The first one was the usual symmetrized matrix $\frac{H_k + H_k^T}{2}$, the second one was the closest symmetrized version of the estimate using the natural Frobenius norm as measure of the distance separating two matrices, and ,finally, the third one used the same type of updating formula but with the G_k- Frobenius norm instead , with G_k being the true Hessian matrix of the objective function. In the case of the particular quadratic function which was tested, all three behaved similarly and no criterion, except perhaps the one of simplicity, could be used to decide which of them was the best.

Of course, many other computational tests are needed before one could draw any definite conclusion on the performances of the algorithm contained in this thesis. In particular, more difficult functions should be used and the performances of the algorithm should be closely compared to the behaviour of some other gradient procedures.

B-Some of the new concepts introduced in this thesis and some suggestions for future research.

The main concept introduced in this thesis is the possibility of representing a minimization algorithm which uses the gradients of its objective function, as a system described by some state-space equations. A particular emphasis was made on the identification step needed to determine satisfactorily all parameters of such a model. The associated regulation problem, however, has not been considered in its generality, since we already started by restricting ourselves to Newton-type algorithms,and consequently, in terms of the regulation problem, to linear output feedback policies.

An interesting possibility for future research would be to treat this problem in a more general framework, that is by really associating to the previous regulation problem some explicit cost functional. Of course, this functional should depend at each step on the gradient, possibly last two gradients, of the objective function, and should also contain some kind of penalty depending either on the step-length or on the gain in the objective. A proposition would be,

$$C = \sum_{k=0}^{N} [g_k^T g_k + s_k^T s_k] ,$$

or if Y(x) represents the positive step-function defined as.

$$Y(x) = \begin{cases} 1 & \text{for x positive} \\ 0 & \text{otherwise} \end{cases}$$

$$C = \sum_{k=0}^{N} [g_k^T g_k + Y[f(x_{k+1}) - f(x_k)]. a_k]$$

The problem could therefore be represented as the regulation problem of a system in which an adaptive identification procedure is necessary. In particular, it seems that the use of some of the techniques introduced by Ljung ans his colleagues in the study of self-tuning regulators should be useful.

APPENDIX I .

Computation of the covariance of the noisy process.

Before really computing this covariance, first one can check that it is possible to find a matrix S_k and a matrix $\underline{G}_k$ such that

$$G_k\, s_k = S_k\, \underline{G}_k \qquad \text{with } G_k \in L(R^n),\ s_k \in R^n,\ \underline{G}_k \in L(R^{n^2}),\ S_k \in L(R^{n^2}, R^n)$$

and this for all indices k= 0,1,2,.. .

Proof:

$$G_k\, s_k = \begin{bmatrix} g_k^1 \\ g_k^i \\ g_k^n \end{bmatrix} s_k = \begin{bmatrix} g_k^{11} & g_k^{12} & g_k^{1n} \\ g_k^{i1} & g_k^{i2} & g_k^{in} \\ g_k^{n1} & g_k^{n2} & g_k^{nn} \end{bmatrix} \begin{bmatrix} s_k^1 \\ s_k^i \\ s_k^n \end{bmatrix} =$$

$$= \begin{bmatrix} \sum_{j=1}^{n} g_k^{1j}\, s_k^j \\ \sum_{j=1}^{n} g_k^{ij}\, s_k^j \\ \sum_{j=1}^{n} g_k^{nj}\, s_k^j \end{bmatrix} = \begin{bmatrix} g_k^{1T}\, s_k \\ g_k^{iT}\, s_k \\ g_k^{nT}\, s_k \end{bmatrix} = \begin{bmatrix} s_k^T\, g_k^1 \\ s_k^T\, g_k^i \\ s_k^T\, g_k^n \end{bmatrix} =$$

$$= \begin{bmatrix} s_k^T & & 0 \\ & 0 & \\ 0 & & s_k^T \end{bmatrix} \begin{bmatrix} g_k^1 \\ g_k^i \\ g_k^n \end{bmatrix} = S_k\, \underline{G}_k \qquad \text{Q.E.D}$$

The covariance of the observation noise process (w_k) can be now computed using the previous notation

$$E\left[w_k w_k^T \right] = \int_0^1 \int_0^1 E\left\{ \left[G_k(\theta) - G_k(0) \right] s_k s_k^T \left[G_k(\sigma) - G_k(0) \right]^T \right\} d\theta \, d\sigma$$

$$= \int_0^1 \int_0^1 E\left\{ S_k \left[\underline{G}_k(\theta) - \underline{G}_k(0) \right] \left[\underline{G}_k(\sigma) - \underline{G}_k(0) \right]^T S_k^T \right\} d\theta \, d\sigma$$

$$= \int_0^1 \int_0^1 (\theta - \sigma) \, ||s_k|| \, S_k S_k^T \, d\theta \, d\sigma = 2||s_k|| \, S_k S_k^T \int_0^1 \frac{\theta^2}{2} \, d\theta$$

$$= \frac{||s_k||}{3} S_k S_k^T = \frac{||s_k||^3}{3} I$$

and similarly ,

$$E\left[w_k V_k^T \right] = \int_0^1 E\left\{ \left[G_k(\theta) - G_k(0) \right] s_k V_k^T \right\} d\theta$$

$$= \int_0^1 E\left\{ S_k \left[G_k(\theta) - G_k(0) \right] \left[G_k(1) - G_k(0) \right]^T \right\} d\theta$$

$$= \frac{||s_k||}{2} S_k$$

$$E\left[V_k V_k^T \right] = C_k = ||s_k|| \, I_{n^2} \qquad \text{for all } k = 0,1,..$$

and hence the covariance of the joint noise process $\begin{pmatrix} V_k \\ w_k \end{pmatrix}$ is

described by equation [7] in Chapter III.

APPENDIX II

Intersection of an ellipsoid with a hyperplane

Consider in R^{n+p} the ellipsoid Ω having as equation,

$$\Omega = \left\{ x \in R^{n+p} : (x-x_c)^T \Pi (x-x_c) \leq 1 \right\} \text{ ,with } \Pi > 0 \text{ ,}$$

as well as the lower dimensional hyperplane

$$\Omega_{obs} = \left\{ x \in R^n : y = A x \right\} \text{ , where y is an element of } R^m.$$

The intersection of Ω with Ω_{obs} is the set defined by,

$$\Omega_I = \left\{ x : (x-x_c)^T \Pi (x-x_c) \leq 1 \text{ and } x = A^* y \right\}$$

if A^* represents the Penrose pseudo-inverse of the matrix A.

Now,let x^0, x^1 and x_c^0, x_c^1 be the components of respectively x and x_c on the subspaces R^n and R^p . The matrix Π can also be decomposed as follows,

$$\Pi = \begin{bmatrix} \Pi_{00} & \Pi_{01} \\ \Pi_{10} & \Pi_{11} \end{bmatrix}$$

Clearly, a necessary and sufficient condition for having a non-empty intersection Ω_I is that,

$$(A^*y - x_c^0)^T \Pi_{00} (A^*y - x_c^0) \leq 1 ,$$

furthermore the set Ω_I can be represented in this case as,

$$\Omega_I = \left\{ x^1 \in R^p : [(A^*y - x_c^0),(x^1 - x_c^1)]^T \begin{bmatrix} \Pi_{00} & \Pi_{01} \\ \Pi_{10} & \Pi_{11} \end{bmatrix} \begin{bmatrix} (A^*y - x^0 c) \\ (x^1 - x_c^1) \end{bmatrix} \leq 1 \right\}$$

Call $\bar{x}^0 = A^*y - x_c^0$. It is easy to show after some trivial computations that Ω_I can also be described by,

$$\Omega_I = \left\{ x^1 \in R^p : \quad (x^1 - \bar{x}_c^1)^T \bar{\Pi}_{11} (x^1 - \bar{x}_c^1) \leq 1 \right\}$$

where $\bar{x}_c^1$ is the center of the new ellipsoid,

$$\bar{x}_c^1 = x_c^1 - \Pi_{11} \bar{x}^0 ,$$

and

$$\bar{\Pi}_{11} = [1 - \bar{x}^{0T} (\Pi_{00} - \Pi_{01} \Pi_{11}^{-1} \Pi_{10}) \bar{x}^0]^{-1} \Pi_{11}$$

is the positive definite matrix describing its excentricity.

REFERENCES

[1]. A. Albert, Regression and the Moore-Penrose Pseudoinverse, Academic Press, New York 1970.

[2]. D. Bertsekas, "Control of Uncertain Systems with a Set Membership Description of the Uncertainty", MIT Dpt EE, Thesis 1971, PhD.

[3]. C.G. Broyden, "Quasi-Newton Methods and their Application to Function Minimization", J.O.T.A.,Vol 13 No 6 , 1976.

[4]. W.C. Davidon, "Optimally Conditioned Optimization Algorithms without Line-Search", Math. Program. ,Vol 9 ,1975 , pp. 1-30.

[5]. J.E. Dennis, Jr, "On Some Methods Based on Broyden's Secant Approximation to the Hessian", in Numerical Methods for Non-Linear Optimization, Lootsma Edit. ,Academic Press, New York, 1972.

[6]. J.E. Dennis,Jr, and J.J. Moré, "A Characterization of Superlinear Convergence and its Application to Quasi-Newton Methods", Math. Comput. , Vol 28, 1974, pp. 549-560.

[7]. R. Fletcher, "A New Approach to Variable Metric Algorithms", Computer Journal, Vol 13, 1970, pp. 317-322.

[8]. R. Fletcher, and M.J.D. Powell, "A Rapidly Convergent Descent Method for Minimization", Computer Journal, Vol 6, 1963, pp. 163-168.

[9]. G. Golub, "Numerized Methods for Solving Linear Least Squares Problems", Num. Mathematik, Vol 7, 1965, pp. 206-216.

[10]. J. Greenstadt, "Variations on Variable Metric Methods",Math. Comput. , Vol 24, No 109, Jan 1970.

[11]. A.S. Householder, The Theory of Matrices in Numerical Programming Blaisdell Publishers, New York, 1965.

[12]. H.Y. Huang, "A Unified Approach to Quadratically Convergent Algorithms for Function Minimization", J.O.T.A. , Vol 5, No 6, 1970.

[13]. H.Y. Huang and J.P. Charliss, "Numerical Experiments on Dual Matrix Algorithms for Function Minimization", J.O.T.A. ,Vol 13, No 6, 1974.

[14]. H. Kwakernaak and R. Sivan, Linear Optimal Control Systems John Wiley and Sons, New York, 1972.

[15]. O.I. Laritchev and G.G. Gorvits, "New Approach to Comparison of Search Methods Used in Nonlinear Programming", J.O.T.A. ,Vol 13, No 6, 1974.

[16]. D.G. Luenberger, Introduction to Linear and Nonlinear Programming Addison Wesley, New York, 1973.

[17]. J.S. Meditch, Stochastic Optimal Linear Estimation and Control Mac Graw Hill, New York, 1969.

[18]. N.E. Nahi, Estimation Theory and Applications, John Wiley and Sons, New York, 1969.

[19]. J.M. Ortega and W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.

[20]. E. Parzen, Stochastic Processes, Holden Day, San Francisco, 1962.

[21]. M.J.D. Powell, "Convergence Properties of a Class of Minimization Algorithms", in Nonlinear Programming, Vol 2, Edited by Mangasarian, Meyer, Robinson, Academic Press, New York, 1975.

[22]. M.J.D. Powell, "A Fortran Subroutine for Unconstrained Minimization Requiring First Derivatives of the Objective Function", Report No R6469, AERE, Harwell, 1970.

[23]. M.J.D. Powell, "Recent Advances in Unconstrained Optimization", in Math. Program. , Vol 1, 1971, pp. 26-57.

[24]. C.R. Rao and S.K. Mitra, Generalized Inverse of Matrices and Applications, John Wiley and Sons, New York, 1971.

[25]. H.R. Schwartz, Numerical Analysis of Symmetric Matrices, Prentice Hall, Englewood Cliffs, New Jersey, 1973.

[26]. F.C. Schweppe, "Recursive State Estimation: Unknown but Bounded Errors and Systems Inputs", IEEE, AC Vol 13, No1, Feb 1968.

[27]. G. Strang, Linear Algebra and its Applications, Academic Press, New York, 1976.

[28]. S.W. Thomas, "Sequential Estimation Techniques for Quasi-Newton Algorithms", Cornell University, Dpt. Computer Science, Technical Report TR75-227, 1975.