Constrained Optimization Using Multiplier Methods with Applications to Control Problems

Glad, Torkel

1976

Document Version:
Publisher's PDF, also known as Version of record

Torkel Glad

Constrained Optimization using Multiplier Methods with Applications to Control Problems

REPORT 7603

APRIL 1976

Constrained Optimization using Multiplier Methods with Applications to Control Problems

TORKEL GLAD

LTH

Department of Automatic Control · Lund Institute of Technology

CONSTRAINED OPTIMIZATION USING MULTIPLIER METHODS WITH APPLICATIONS TO CONTROL PROBLEMS

av

Torkel Glad

Civ ing, Mlm

Akademisk avhandling som för avläggande av teknisk doktorsexamen vid tekniska fakulteten vid universitetet i Lund kommer att offentligen försvaras i sal M:B, Maskinhuset, Lunds Tekniska Högskola, onsdagen den 19 maj 1976 kl 10.15.

Torkel Glad

CONSTRAINED OPTIMIZATION USING MULTIPLIER METHODS WITH APPLICATIONS TO CONTROL PROBLEMS

Lund 1976

T A B L E O F C O N T E N T S

1. INTRODUCTION

Optimization techniques have many applications in different fields of engineering. Its use in control theory is the topic of this report. Classical methods of designing a control system are largely based on trial and error. The engineer is given certain specifications and has to use his intuition and practical experience in trying to find a solution. Often the design consists of two steps. First a configuration is chosen, e.g. a PID controller for an industrial process or a lead-lag compensation for a servo motor. Then a number of parameters within the given structure have to be determined, the three gain parameters for the PID controller or the gain and time constants in the lead-lag network. There are rules of thumb for choosing the parameters but sometimes they are not sufficient. The designer might want a tuning of the parameters that is the best one in a certain sense, e.g. the one that minimizes settling time. It may be desired to obtain fixed parameters that give reasonable performance even if the process parameters vary. In some cases no satisfactory tuning is found even after long trials. It is then natural for the designer to suspect that it is actually impossible to satisfy the specifications with the given configuration. However, it is not possible to prove this without a systematic way of choosing the parameters.

One way of getting a systematic approach is to specify a mathematically defined criterion. The goal of the design task is then to find a controller that minimizes the criterion. A survey of the early design methods based on optimization is given in Newton, Gould and Kaiser (1957). A drawback of the methods used there, is that only criteria which can be computed from explicit analytic expressions are considered. This limits the application to linear systems and criteria which are the time integral of the squared error or the expectation of the squared error. With the development of computers and computational methods that has taken place in

the last two decades, it is no longer necessary to use only design criteria that can be manipulated by direct analytical techniques. The modern numerical optimization algorithms can be applied to practically all criteria. This is a great advantage since it permits the designer to choose a criterion that describes exactly the qualities he desires to have in the control system.

A complication that arises in most design problems is that there are several conflicting criteria. The conflict can be resolved in several ways. One possibility is to introduce a new criterion that is a weighted sum of the original criteria. The drawback of this approach is the difficulty of choosing good values of the weights. An alternative is to minimize with respect to one criterion and put upper or lower bounds on the remaining criteria. This results in a constrained optimization problem. Accepting this formulation it is also possible to use optimization theory to explore if the specification can be fulfilled at all with the chosen configuration. If the global optimum does not satisfy the specifications, the design problem is impossible to solve.

In Chapter 2 a survey is given of some different optimization problems originating in control theory. It is shown that these problems usually lead to constrained optimization problems. Some of these are finite dimensional while others are infinite dimensional.

The general constrained finite dimensional problem is also called the nonlinear programming problem. As mentioned above, design problems for control systems with conflicting criteria can be formulated in this form. Since the constrained optimization is important also in other fields of engineering, some methods for solving it are considered in Chapter 3. There exist powerful methods for unconstrained minimization, e.g. the so-called quasi-Newton methods. In this report the

attention will therefore be focused on methods that convert the constrained problem into an unconstrained one. The methods which are studied belong to the combined multiplier and penalty function methods. These methods are based on the use of a function $F(x,u)$ called the augmented Lagrangian. F depends on the independent variables of the constrained problem, x and on a vector of so-called multipliers, u. For a certain value, $\bar{u}$, the function $F(x,\bar{u})$ has a local unconstrained minimum at the point, $\bar{x}$, which solves the original problem. This is the basic property and means that the constrained problem is transformed into an unconstrained one. Since $\bar{u}$ is unknown, it is usually determined iteratively. An alternative is to replace u by a function $\tilde{u}(x)$, having the property that $\tilde{u}(\bar{x}) = \bar{u}$. In 3.3 it is shown that $F(x,\tilde{u}(x))$ has a local minimum at $\bar{x}$, under conditions that are more general than those given in literature. Most of Chapter 3 is devoted to methods where $\bar{u}$ is determined iteratively.

Most convergence results for algorithms based on augmented Lagrangians are derived for methods where $F(x,u)$ is minimized for fixed u before u is updated. In Chapter 3 a method is presented where a quasi-Newton method is used for the minimization with respect to x, and u is updated after each line search in x. It is shown that safeguards can be included in the method to make it globally convergent. It is also shown that the local convergence rate is linear for two of the updating rules for u and superlinear for a third one. The practical usefulness of the algorithms is demonstrated in a comparison with other algorithms using a number of test problems.

In Chapter 4 the infinite dimensional optimal control problem is considered. The goal is to study the extension of the methods of Chapter 3 to infinite dimensional problems. The augmented Lagrangians connected with the equality constraints defined by the differential equation and the terminal

constraints are studied. Conditions that guarantee that this augmented Lagrangian has a local minimum which solves the original optimal control problem are presented. It is shown that a Riccati equation plays a crucial role in these sufficiency conditions.

The connection with the usual second order sufficiency conditions for optimal control problems is given. As in the finite dimensional case it is necessary to use iterative methods for the multipliers. A method proposed in literature is investigated and shown to be linearly convergent. Finally an extension of the augmented Lagrangian approach to handle mixed state control inequality constraints is proposed. It is shown how the method is connected with sufficiency conditions for a local minimum.

In Chapter 5 the use of finite dimensional constrained optimization methods in the design of control systems is discussed. A powerful interactive simulation program called SIMNON was developed by Elmqvist. In Chapter 3 it is described how an optimization routine is added to SIMNON. This gives a versatile tool for control system design. The ability of the optimization routine to handle general nonlinear minimization problems means that many design tasks that are difficult to handle with conventional synthesis techniques can be solved. In particular, nonlinearities in the system can be dealt with. Several examples of synthesis problems solved using optimization techniques are presented.

An on-line application of optimization is presented in Chapter 6. The process consists of an oil burner and the objective is to maximize combustion efficiency. It is shown that the best adjustment can be found automatically by a standard optimization routine of the quasi-Newton type.

Acknowledgements.

I wish to express my sincere gratitude to Professor Karl Johan Åström for his stimulating help and guidance throughout the work.

The people at the Department of Automatic Control have helped in the work by many interesting discussions. In particular I want to thank Krister Mårtensson who initiated the work on multipliers. I also want to thank Gunnar Bengtsson, Olov Einarsson, Hilding Elmqvist and Lennart Ljung who have given valuable comments on the manuscript. Leif Andersson, Rolf Braun and Lars Jensen have been of great help in the experimental part of the work.

I also want to thank l:e forskn. ing. Bertil Reenstierna and Civ. ing. Lennart Sjöstedt at the Department of Machine Design for their cooperation in the experiments with the oil burner.

This work has been supported by the Swedish Institute of Applied Mathematics which is gratefully acknowledged.

Many thanks also to Gudrun Christensen for her excellent typing of the manuscript and to Britt Marie Carlsson who carefully prepared the figures.

Notation.

With the exception of derivatives explained below, all vectors are column vectors. The transpose of a vector or a matrix y is denoted y^T.

The derivative of a vector valued function

$$g(x) = \begin{pmatrix} g_1(x) \\ \\ g_m(x) \end{pmatrix}$$

is denoted

$$g_x = \begin{pmatrix} \partial g_1/\partial x_1 & \cdots & \partial g_1/\partial x_n \\ \\ \partial g_m/\partial x_1 & \cdots & \partial g_m/\partial x_n \end{pmatrix}$$

i. e. derivatives are row vectors. The second derivative of a scalar function f is denoted

$$f_{xx} = \begin{pmatrix} \partial^2 f/\partial x_1^2 & \cdots & \partial^2 f/\partial x_1 \partial x_n \\ \\ \partial^2 f/\partial x_n \partial x_1 & \cdots & \partial^2 f/\partial x_n^2 \end{pmatrix}$$

For vectors, $\| \ \|$ denotes an arbitrary vector norm. For matrices it denotes the corresponding operator norm. $\|a\|_2$ denotes the norm $\sqrt{a^T a}$.

For function spaces the following notation is used.

$C_0^n[t_1,t_2]$ denotes the space of n-vector valued continuous functions on $[t_1,t_2]$. The corresponding norm is

$$\|f\|_0 = \sup_{t_1 \le t \le t_2} \|f(t)\|$$

Similarly

$C_1^n[t_1,t_2]$ denotes the space of n-vector valued continuously differentiable functions on $[t_1,t_2]$. The corresponding norm is

$$\|f\|_1 = \sup_{t_1\leq t\leq t_2} \|f(t)\| + \sup_{t_1\leq t\leq t_2} \|\dot{f}(t)\|$$

where $\dot{f}$ is the derivative of f.

References.

Newton, G.C., Gould, L.A., and Kaiser, J.F. (1957):
Analytical Design of Linear Feedback Controls, Wiley, New York.

2. OPTIMIZATION PROBLEMS IN CONTROL THEORY

There are several different areas of control theory where optimization problems can be formulated. To give a motivation for the optimization theory developed in Chapters 3 and 4, a summary of some important control problems leading to optimization is presented.

A control system, see Fig. 2.1, is characterized by an input u, an output y, a state x and influences from the environment, v. It will be assumed that the system can be described by differential equations

$$\dot{x}(t) = f\big(x(t),u(t),v(t),t\big)$$
$$y(t) = g\big(x(t),u(t),v(t),t\big)$$

or difference equations

$$x(t+\tau) = f\big(x(t),u(t),v(t),t\big)$$
$$y(t) = g\big(x(t),u(t),v(t),t\big)$$

or a combination of these.

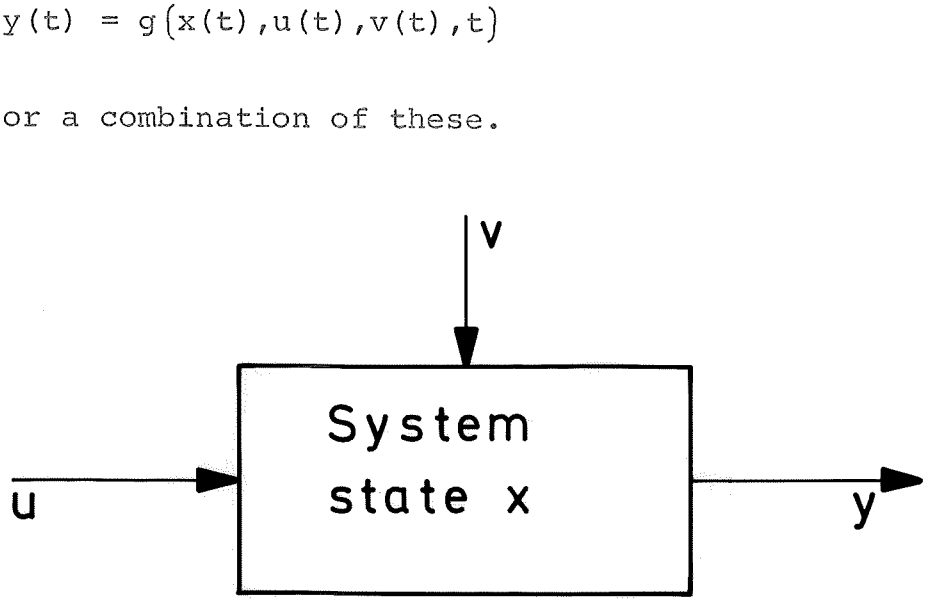


Fig. 2.1.

The problem of designing a controller can in very general terms be stated as follows.

Find a rule for computing u from y such that the system behaves in a desirable way despite the influences from the environment.

There are several ways in which this design problem may lead to an optimization problem.

Controller with fixed structure.

A common situation is that the controller has a fixed structure. The structure may be fixed because a certain type of equipment, e.g. a PID controller, has to be used. The designer can also choose a certain structure on the basis of experience. The chosen structure can also be one in a series to be investigated.

Assume for simplicity that both the controller and the system are described by differential equations. If p denotes the parameters to be adjusted in the controller, the system plus controller is then described by

$$\dot{x}(t) = f(x(t),u(t),t)$$
$$y(t) = g(x(t),u(t),t)$$
$$\dot{z}(t) = k(z(t),y(t),t,p)$$
$$u(t) = m(z(t),y(t),t,p)$$

Introducing a modified state vector

$$\tilde{x} = \begin{pmatrix} x \\ z \end{pmatrix}$$

and the function $\tilde{f}$ satisfying

$$\tilde{f}(\tilde{x}(t),t,p) = \begin{bmatrix} f(x(t),u(t),t) \\ k(z(t),y(t),t,p) \end{bmatrix}$$

the system can be written

$$\dot{\tilde{x}}(t) = \tilde{f}(\tilde{x}(t),t,p)$$

Many criteria that characterize the performance of a control system can be written in the form

$$J(x,p) = \int_0^T L(\tilde{x}(t),t,p)dt$$

If a fixed initial state of $\tilde{x}$ is assumed, then every value of p defines a solution $\tilde{x}_p(t)$. (It is assumed that $\tilde{f}$ satisfies a Lipschitz condition for every p, so that the solution $\tilde{x}$ is unique.) Therefore the function

$$\varphi_0(p) = \int_0^T L(\tilde{x}_p(t),t,p)dt$$

is well defined. The resulting problem is therefore the minimization of a real function of finitely many variables. It is also possible that there are constraints in the problem, e.g.

$$J_i(\tilde{x},p) = \int_0^T L_i(\tilde{x}(t),t,p)dt \leq C_i$$

or

$$J_i(\tilde{x},p) = \max L_i(\tilde{x}(t),t,p) \leq C_i \qquad i = 1, \ldots, m$$

Defining

$$\varphi_i(p) = J_i(\tilde{x}_p,p)$$

the problem then becomes

minimize $\varphi_0(p)$
under the constraints $\varphi_i(p) \leq 0 \qquad i = 1, \ldots, m$

This is an example of a general nonlinear optimization problem, sometimes called the nonlinear programming problem.

The reasoning above can easily be extended to the case where some of the differential equations are replaced by difference equations, e.g. for a system controlled by a digital computer.

Several examples of optimization of systems with fixed structure are given in Section 4.3.

Instead of studying a fixed structure controller one can pose the problem of finding the best control, u on some given time interval. This is an open loop problem. Examples of this problem can be found in the aerospace industry where it is required to compute the optimal trajectory for a space vehicle. An example involving a container crane is given in Section 4.3. There is an important difference between discrete time and continuous time problems.

Discrete time optimal control.

This problem can be formulated:

Minimize

$$J = \sum_{i=0}^{N} L(x(i), u(i), i)$$

for the system

$$x(i+1) = f(x(i),u(i),i)$$

It is possible to give all sorts of different constraints. The usual ones are

$$g_j(x(i),u(i),i) \leq 0 \qquad j = 1, \ldots, m_1$$

$$\psi_j(x(N)) = 0 \qquad j = 1, \ldots, m_2$$

$$x(0) = x_0$$

The problem can either be regarded as a problem in the variables

$$u(0), u(1), \ldots, u(N)$$

or in the variables

$$u(0), \ldots, u(N), x(1), \ldots, x(N)$$

with the difference equation as an equality constraint. In both cases the resulting problem is a finite dimensional constrained problem. This does not mean that a general nonlinear programming algorithm is necessarily the best way of solving this problem because the problem has a special structure that can be taken advantage of in an algorithm.

Continuous time optimal control.

The problem is to find a u to minimize

$$J = \int_0^T L(x(t),u(t),t)dt + F(x(T))$$

for the system

$$\dot{x}(t) = f\big(x(t),u(t),t\big)$$
$$x(0) = a$$

The constraints can be of the form

$$g\big(x(t),u(t),t\big) \leq 0$$
$$\psi\big(x(T)\big) = 0$$

This problem differs from those treated before because here u and x are elements in an infinite dimensional space. The constraints $\dot{x} = f(x,u,t)$ and $g(x,u,t) \leq 0$ are also infinite dimensional in nature.

Although this problem is mathematically much more difficult than the finite dimensional one, there are many parallels in the theory. In Chapter 4 it will be shown that the numerical methods for finite dimensional problems can, to a large extent, be generalized to the optimal control problem.

Optimization problems are not only encountered in the design of control systems. The problem of finding a mathematical model that describes a physical system can also be formulated as an optimization problem.

The identification problem.

Consider the problem of finding the value of a parameter p in a mathematical model of a physical system

$$x(t+1) = f\big(x(t),u(t),t,p\big)$$
$$y(t) = g\big(x(t),u(t),t,p\big)$$

Most identification methods can be formulated as the minimization of suitably chosen criterion, that can usually be expressed as

$$J = L(y(0),y(1),..,y(N),y_m(0),y_m(1),..,y_m(N))$$

where $y_m(t)$ is the output of the real system, when the model and the physical system both have the same input. This problem is a finite dimensional problem.

The conclusion of this brief survey is that, even if optimization problems are derived from different parts of control theory, they can all be formulated as the minimization of an objective function possibly under constraints. An important difference from the mathematical, and to a lesser extent,the algorithmic, point of view is between finite dimensional and infinite dimensional problems. Methods for solving these problems will be considered in the next two chapters.

3. THE FINITE DIMENSIONAL PROBLEM.

In this section methods for solving the constrained finite dimensional problem are investigated. The methods use the idea of converting the constrained problem into an unconstrained one or into a sequence of unconstrained problems. To do this a function called the augmented Lagrangian is introduced in section 3.1. Known results from the literature are also given in that section. These results show that the augmented Lagrangian has a local minimum at the solution to the constrained problem. In section 3.2 it is shown that, with mild restrictions on the objective function, the minimum is actually a global one. Section 3.3 shows that, when the multipliers used in the augmented Lagrangian are treated as functions of x, the augmented Lagrangian has a local minimum under weaker conditions than those given in the literature. In 3.4 iterative methods based on the augmented Lagrangian are discussed. The local convergence properties of three different updating methods are treated in 3.5 and it is shown that two of them converge linearly while the third one has superlinear convergence. Modifications that give global convergence are shown in 3.6. Finally the algorithms are tested on numerical examples and compared with other methods of solving the constrained problem in 3.7.

3.1. Problem Formulation and Basic Results.

After defining the problem, some standard results from optimization theory will be given. First the necessary and sufficient conditions for a constrained optimum are presented. Then the function known as the augmented Lagrangian is defined and three theorems showing its basic properties are quoted. These results form the background of the material presented in sections 3.2 - 3.7.

As shown in section 2, the finite dimensional problem can be sta-

ted as follows:

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && h(x) = 0 \quad h(x) = \left(h_1(x), \ldots, h_p(x)\right)^T \\ &&& g(x) \leq 0 \quad g(x) = \left(g_1(x), \ldots, g_m(x)\right)^T \end{aligned}$$

where x is an n-dimensional real vector and $f(x)$, $h_i(x)$ and $g_i(x)$ are real valued functions. The notation $g(x) \leq 0$ means that $g_i(x) \leq 0$, $i = 1, \ldots, m$. Let $\bar{x}$ be a local minimum to this problem. Define the integer sets $A = \{i \mid g_i(\bar{x}) = 0\}$ and $I = \{i \mid g_i(\bar{x}) < 0\}$, i.e. the sets of indices corresponding to active and inactive inequality constraints respectively.

In what follows it will be assumed that the following conditions hold.

(C1) In a neighbourhood of $\bar{x}$ the functions f, h and g are twice continuously differentiable.

(C2) The vectors $(h_i)_x(\bar{x})$, $(g_j)_x(\bar{x})$, $i = 1, \ldots, p$, $j \in A$ are linearly independent.

The necessary conditions for a local minimum are then given by the following theorem, the well-known Kuhn-Tucker multiplier rule.

Theorem 3.1. Assume that $\bar{x}$ is a local minimum to the constrained optimization problem and that conditions C1 and C2 hold. Then there exist uniquely defined vectors $\bar{\mu} = (\bar{\mu}_1, \ldots, \bar{\mu}_p)^T$ and $\bar{\lambda} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_m)^T$ such that

$$\begin{aligned} &f_x(\bar{x}) + \bar{\mu}^T h_x(\bar{x}) + \bar{\lambda}^T g_x(\bar{x}) = 0 \\ &h(\bar{x}) = 0 \qquad\qquad (3.1) \\ &\bar{\lambda} \geq 0 \quad \bar{\lambda}^T g(\bar{x}) = 0 \quad g(\bar{x}) \leq 0 \end{aligned}$$

Proof. See e.g. Fiacco and McCormick (1968). □

Remark. A point satisfying (3.1) is called a Kuhn-Tucker point. Since equation (3.1) is only a necessary condition, a Kuhn-Tucker point is not always a local minimum. □

To simplify the notation, introduce the vector

$$u^T = (\mu^T, \lambda^T)$$

It is also convenient to use the Lagrangian

$$L(x,u) = L(x,\mu,\lambda) = f(x) + \mu^T h(x) + \lambda^T g(x)$$

Theorem 3.1 then implies that the Lagrangian has a stationary point at $x = \bar{x}$.

In the sufficiency conditions two further assumptions are needed.

(C3) For all vectors $z \neq 0$ such that $(h_i)_x(\bar{x})z = 0$, $i = 1, \ldots, p$, $(g_j)_x(\bar{x})z = 0$, $j \in A$ the inequality $z^T L_{xx}(\bar{x},\bar{u})z > 0$ holds.

(C4) $\bar{\lambda}_i > 0$ for all $i \in A$.

The standard sufficiency conditions used in constrained optimization theory are then given by the following theorem.

Theorem 3.2. Let $\bar{x}$, $\bar{\mu}$ and $\bar{\lambda}$ satisfy equation (3.1) and assume that C1 - C4 hold. Then $\bar{x}$ is an isolated local minimum to the constrained minimization problem.

Proof. See Fiacco and McCormick (1968). □

For convex problems the Lagrangian has the property that

$$L(\bar{x},\bar{u}) = \min_{x} L(x,\bar{u})$$

see Luenberger (1968). In the general case this is not true. As shown by Theorem 3.1, $L_x(\bar{x},\bar{u}) = 0$, but this stationary point is not necessarily a local minimum with respect to x. Therefore Hestenes (1969) and Powell (1969), treating the equality constrained case, i.e. where g(x) is absent, introduced the augmented Lagrangian

$$F(x,\mu,c) = f(x) + \mu^T h(x) + \frac{c}{2} h(x)^T h(x)$$

An interesting property of this function is given by the following theorem, shown by Hestenes (1969).

Theorem 3.3. Suppose that all constraints are equality constraints and that conditions C1, C2 and C3 hold. Then there exists a constant c_0 such that, for $c > c_0$, $F_{xx}(\bar{x},\bar{\mu},c)$ is positive definite and $F(x,\bar{\mu},c)$ has a local isolated minimum at $x = \bar{x}$.

Proof. See Hestenes (1969).

The function F can also be written

$$F(x,\mu,c) = f(x) + \frac{1}{2c}\left(ch(x) + \mu\right)^T\left(ch(x) + \mu\right) - \frac{1}{2c}\mu^T\mu$$

This suggests an extension to inequality constraints which has been used by several authors, see e.g. Rockafellar (1974).

$$F(x,u,c) = f(x) + \frac{1}{2c}\left(ch(x) + \mu\right)^T\left(ch(x) + \mu\right) -$$

$$- \frac{1}{2c}\,\mu^T\mu + \frac{1}{2c}\left(cg(x) + \lambda\right)_+^T\left(cg(x) + \lambda\right)_+ -$$

$$- \frac{1}{2c}\,\lambda^T\lambda \qquad (3.2)$$

where $\left(g(x) + \lambda\right)_+ = \left[\left(g_1(x) + \lambda_1\right)_+, \ldots, \left(g_m(x) + \lambda_m\right)_+\right]^T$ and the notation

$$a_+ = \begin{cases} a & \text{if } a > 0 \\ 0 & \text{if } a \leq 0 \end{cases}$$

is used.

For this choice of F, Bertsekas (1973), (1975) has shown the following result.

Theorem 3.4. Suppose conditions C1 - C4 hold. Then there exists a c_0 such that $F_{xx}(\bar{x},\bar{u},c)$ is positive definite for $c > c_0$. Consequently $F(x,\bar{u},c)$ has an isolated local minimum at $x = \bar{x}$ if $c > c_0$.

Proof. See Bertsekas (1973). □

Since, in general, $\bar{u}$ is not known, it is natural to study what happens when $F(x,u,c)$ is minimized for an arbitrary value of u. Define

$$G_M(u,c) = \inf_{x \in M} F(x,u,c)$$

where M is a neighbourhood of $\bar{x}$. $G_M(u)$ has the following duality property, which has been studied by Glad (1973) in the equa-

lity constrained case and by Rockafellar (1974) and Bertsekas (1973) in the general case.

Theorem 3.5. Assume that conditions C1 - C4 hold and let c be large enough for $F(x,\bar{u},c)$ to have a local minimum at $\bar{x}$. Let M be a neighbourhood of $\bar{x}$ such that

$$\inf_{x\in M} F(x,\bar{u},c) = F(\bar{x},\bar{u},c)$$

Then the function

$$G_M(u,c) = \inf_{x\in M} F(x,u,c)$$

has the following property

$$\max_u G_M(u,c) = G_M(\bar{u},c) = F(\bar{x},\bar{u},c) = f(\bar{x})$$

Proof.

$$G_M(u,c) = \inf_{x\in M} F(x,u,c) \leq F(\bar{x},u,c) \leq$$

$$\leq f(\bar{x}) = F(\bar{x},\bar{u},c) = G_M(\bar{u},c) \qquad \square$$

The first and second derivatives of $G_M(u,c)$ with respect to u can be calculated, see e.g. Fletcher (1974). The result is summarized in the following theorem.

Theorem 3.6. Let conditions C1 - C4 hold and let c be large enough for $F_{xx}(\bar{x},\bar{u},c)$ to be positive definite. Then there exists a continuously differentiable function $\varphi(u,c)$ defined

in a neighbourhood, M_U, of $\bar{u}$ such that

$$F\left(\varphi(u,c),u,c\right) = \inf_{x\in M} F(x,u,c), \qquad u \in M_U$$

$$\varphi(\bar{u},c) = \bar{x}$$

where M is some neighbourhood of $\bar{x}$.

In the same neighbourhood, M_U, the derivatives $G_u(u,c)$ and $G_{uu}(u,c)$ are defined. The first derivative is

$$G_\mu^T(u,c) = h\left(\varphi(u,c)\right)$$

$$G_{\lambda_i}^T(u,c) = \begin{cases} g_i\left(\varphi(u,c)\right) & \text{if } cg_i\left(\varphi(u,c)\right) + \lambda_i > 0 \\ -\frac{1}{c}\lambda_i & \text{if } cg_i\left(\varphi(u,c)\right) + \lambda_i \leq 0 \end{cases} \tag{3.3}$$

Assume for simplicity that the indices are ordered such that

$$cg_i\left(\varphi(u,c)\right) + \lambda_i > 0 \qquad i = 1, \ldots, i_1$$

$$cg_i\left(\varphi(u,c)\right) + \lambda_i \leq 0 \qquad i = i_1+1, \ldots, m$$

and let z be the vector

$$z = \begin{pmatrix} h_x \\ (g_1)_x \\ \vdots \\ (g_{i_1})_x \end{pmatrix}$$

Then the second derivative can be written

$$G_{uu}(u,c) = \begin{pmatrix} -zF_{xx}^{-1}z^T & 0 \\ 0 & -\frac{1}{c} I_{m-i_1} \end{pmatrix} \tag{3.4}$$

where $F_{xx} = F_{xx}(\varphi(u,c),u,c)$.

Proof. The existence and properties of the function φ follow from the implicit function theorem, see Luenberger (1968). The expressions for the derivatives G_u, G_{uu} are found in Fletcher (1974). □

3.2. Properties of the Augmented Lagrangian.

Theorem 3.4 states that $F(x,\bar{u},c)$ has a local minimum at $x = \bar{x}$. Suppose $\bar{x}$ is a global minimum of the constrained problem. It is then natural to ask if $\bar{x}$ is also a global minimum of $F(x,\bar{u},c)$ for some choice of c. Rockafellar (1974) proved $\bar{x}$ to be a global minimum under fairly general conditions, using a duality approach. Here the globality of the minimum at $\bar{x}$ will be investigated, using more direct methods. First it is proved that the area, with respect to which $\bar{x}$ is a local minimum, can be made arbitrarily large.

Lemma 3.1. Assume that conditions C1 - C4 hold and that $f(x) > f(\bar{x})$ for all $x \neq \bar{x}$ satisfying the constraints. Then for every $A > 0$ there exists a constant c_0, depending on A, such that

$$F(x,\bar{u},c) > F(\bar{x},\bar{u},c)$$

for all $x \neq \bar{x}$ such that $\| x - \bar{x} \| \leq A$, provided $c > c_0$.

Proof. Suppose the theorem is false. Then there exist sequences

$\{x^{(k)}\}_1^\infty$ and $\{c^{(k)}\}_1^\infty$ with

$$\lim_{k\to\infty} c^{(k)} = \infty$$

such that

$$F\left(x^{(k)},\bar{u},c^{(k)}\right) \leq F\left(\bar{x},\bar{u},c^{(k)}\right) = f(\bar{x})$$

and

$$x^{(k)} \neq \bar{x},\ \| x^{(k)} - \bar{x} \| \leq A \text{ for all } k$$

Since the set $\| x - \bar{x} \| \leq A$ is compact, there exists a subsequence $x^{(k_j)}$, converging to some point $\hat{x}$ with $\| \hat{x} - \bar{x} \| \leq A$. To simplify the notation, the sequence $x^{(k_j)}$ will be denoted $x^{(j)}$ in the remainder of the proof. Then it follows that

$$f\left(x^{(j)}\right) + \frac{c^{(j)}}{2}\left(h\left(x^{(j)}\right) + \frac{\bar{\mu}}{c^{(j)}}\right)^T\left(h\left(x^{(j)}\right) + \frac{\bar{\mu}}{c^{(j)}}\right) -$$

$$- \frac{\bar{\mu}^T\bar{\mu}}{2c^{(j)}} + \frac{c^{(j)}}{2}\left(g\left(x^{(j)}\right) + \frac{\bar{\lambda}}{c^{(j)}}\right)_+^T\left(g\left(x^{(j)}\right) + \frac{\bar{\lambda}}{c^{(j)}}\right)_+ -$$

$$- \frac{\bar{\lambda}^T\bar{\lambda}}{2c^{(j)}} \leq f(\bar{x}) \qquad (3.5)$$

Taking limits

$$\lim_{j\to\infty} \frac{c^{(j)}}{2}\left|\left(h\left(x^{(j)}\right) + \frac{\bar{\mu}}{c^{(j)}}\right)^T\left(h\left(x^{(j)}\right) + \frac{\bar{\mu}}{c^{(j)}}\right) +\right.$$

$$\left. + \left(g\left(x^{(j)}\right) + \frac{\bar{\lambda}}{c^{(j)}}\right)_+^T\left(g\left(x^{(j)}\right) + \frac{\bar{\lambda}}{c^{(j)}}\right)_+\right| \leq f(\bar{x}) - f(\hat{x})$$

Since $c^{(j)} \to \infty$, this can be true only if

$$h(\hat{x}) = 0$$

$$g(\hat{x}) \leq 0$$

i.e. only if $\hat{x}$ satisfies the constraints. The inequality (3.5) also gives the relation

$$f(x^{(j)}) \leq f(\bar{x}) + \frac{\bar{\mu}^T\bar{\mu}}{2c^{(j)}} + \frac{\bar{\lambda}^T\bar{\lambda}}{2c^{(j)}}$$

which shows that $f(\hat{x}) \leq f(\bar{x})$.

From the assumptions in the theorem it then follows that $\hat{x} = \bar{x}$. This means that $F(x^{(j)},\bar{u},c^{(j)}) \leq F(\bar{x},\bar{u},c^{(j)})$ for a sequence satisfying $x^{(j)} \to \bar{x}$, $x^{(j)} \neq \bar{x}$, which contradicts the fact that $F(x,\bar{u},c)$ has an isolated local minimum at $\bar{x}$ if c is large enough.

□

The fact that the constant A in Lemma 3.1 can be chosen arbitrarily large, does not in general imply that F has a global minimum at $\bar{x}$ as shown by the following example.

Example 3.1.

$$f(x) = x_2^2 - e^{x_1^2} \qquad h(x) = x_1$$

$$\bar{x} = (0,0) \qquad \bar{\mu} = 0$$

$$F(x,0,c) = x_2^2 + \frac{c}{2}x_1^2 - e^{x_1^2}$$

which is not bounded from below. However, with some mild further restrictions on f, h and g, F does have a global minimum at $\bar{x}$, as shown by the following theorem.

Theorem 3.7. Assume that conditions C1 - C4 hold and that $f(x) > f(\bar{x})$ for all $x \neq \bar{x}$ satisfying the constraints. In addition assume that there exist constants A, c_1 and $\varepsilon > 0$ such that $F(x,\bar{u},c_1) \geq f(\bar{x}) + \varepsilon$ for all x satisfying $\| x - \bar{x} \| > A$. Then there exists a constant c_0 such that $F(x,\bar{u},c)$ has a global minimum at $x = \bar{x}$ for $c > c_0$.

Proof. There exists a constant c_1 such that, for $c \geq c_1$, and $\| x - \bar{x} \| > A$

$$F(x,\bar{u},c) \geq f(\bar{x}) + \varepsilon$$

Since from lemma 3.1 there exists a c_0 such that $F(x,\bar{u},c) > F(\bar{x},\bar{u},c)$ for $\| x - \bar{x} \| \leq A$, $x \neq \bar{x}$ and $c > c_0$, it follows that $F(x,\bar{u},c)$ has a global minimum at $\bar{x}$ for $c > \max(c_0,c_1)$. □

The assumptions can be varied somewhat as shown by the following corollaries.

Corollary 1. Assume that conditions C1 - C4 hold and that $f(x) > f(\bar{x})$ for all $x \neq \bar{x}$ satisfying the constraints. Also assume that there exist constants A and $\varepsilon > 0$ such that $f(x) \geq f(\bar{x}) + \varepsilon$ for all x satisfying $\| x - \bar{x} \| > A$. Then there exists a constant c_0 such that $F(x,\bar{u},c)$ has a global minimum at $x = \bar{x}$ for $c > c_0$.

Proof. There exists a constant c_1 such that, for $c \geq c_1$ and $\| x - \bar{x} \| > A$

$$F(x,\bar{u},c) \geq f(x) - \frac{1}{2c}\bar{u}^T\bar{u} \geq f(\bar{x}) + \frac{\varepsilon}{2}$$

Theorem 3.7 is then applicable. □

Corollary 2. Assume that conditions C1 - C4 hold and that $f(x) > f(\bar{x})$ for all $x \neq \bar{x}$ satisfying the constraints. Let $f(x)$ be bounded from below and assume that the constraints satisfy the condition

$$\inf_{\|x-\bar{x}\| > A} \left\{ \frac{c_1}{2} h(x)^T h(x) + \frac{c_1}{2} g(x)_+^T g(x)_+ \right\} = \varepsilon$$

for some constants $c_1 > 0$, $\varepsilon > 0$ and $A > 0$. Then there exists a constant c_0 such that $F(x,\bar{u},c)$ has a global minimum at $x = \bar{x}$ for $c > c_0$.

Proof.

$$F(x,\bar{u},c) \geq \inf_x f(x) + \frac{c}{2c_1}\varepsilon - \frac{1}{2c}\bar{u}^T\bar{u} \geq f(\bar{x}) + 1$$

if c is large enough and $\|x - \bar{x}\| > A$. Theorem 3.7 can then be applied. □

Remark. A simple example of a set of constraints that satisfy the conditions of corollary 2 is

$$g_i(x) = a_i - x_i \qquad i = 1, \ldots, n$$

$$g_{n+i}(x) = x_i - b_i \qquad i = 1, \ldots, n$$

i.e. upper and lower bounds on all the variables. □

The remark shows that theorem 3.7 is applicable to most problems having $f(x)$ bounded from below, because it is almost always possible to specify upper and lower bounds on the variables and then add the extra constraints corresponding to the upper and lower bounds.

Using theorem 3.7, the duality result of theorem 3.5 is immediately generalized. The function $G(u,c)$ can now be defined

$$G(u,c) = \inf_x F(x,u,c)$$

where the infimum is taken with respect to all x.

Theorem 3.8. Let the assumptions of theorem 3.7 be satisfied. Then, for sufficiently large c,

$$G(u,c) = \inf_x F(x,u,c)$$

has the following property

$$\max_u G(u,c) = G(\bar{u},c) = F(\bar{x},\bar{u},c) = f(\bar{x})$$

Proof. From theorem 3.7 it follows that

$$\inf_x F(x,\bar{u},c) = F(\bar{x},\bar{u},c)$$

The rest of the proof is analogous to the proof of theorem 3.5 . □

3.3. Multipliers That Are a Function of x.

In the previous two sections it was shown that $F(x,\bar{u},c)$ has a local and sometimes a global unconstrained minimum at $\bar{x}$. The constrained problem has thus been transformed into an unconstrained one. Since $\bar{u}$ is unknown in most problems, this fact is not immediately useful. In this section one way of overcoming this problem will be discussed.

The method to be considered can be found in Fletcher (1970) and Mårtensson (1972), for the equality constrained case and in Fletcher (1973) for problems with inequality constraints. Here only equality constraints will be considered. The idea is to replace the parameter μ with a function $\tilde{\mu}(x)$. The augmented Lagrangian then becomes

$$F(x,\tilde{\mu}(x),c) = f(x) + \tilde{\mu}(x)^T h(x) + \frac{c}{2} h(x)^T h(x)$$

The function $\tilde{\mu}(x)$ is required to be twice continuously differentiable and have the property $\tilde{\mu}(\bar{x}) = \bar{\mu}$. A natural choice is

$$\tilde{\mu}(x) = -\left(h_x(x) h_x(x)^T\right)^{-1} h_x(x) f_x(x)^T$$

If conditions C1 and C2 are satisfied, this choice of $\tilde{\mu}(x)$ satisfies $\tilde{\mu}(\bar{x}) = \bar{\mu}$. This follows from equation (3.1). Fletcher (1970) and Mårtensson (1972) have proved that if conditions C1 - C3 are satisfied, then $F(x,\tilde{\mu}(x),c)$ has a local minimum at $x = \bar{x}$, provided c is large enough. It turns out, however, that condition C3 is not needed, when $\mu = \tilde{\mu}(x)$ is used.

Theorem 3.9. Assume that $\bar{x}$ is a local minimum to the constrained problem and that conditions C1 and C2 are satisfied. Then there exists a constant c_0 such that $F(x,\tilde{\mu}(x),c)$ has a local minimum at $x = \bar{x}$ for all $c > c_0$.

Proof. Since $h_x(\bar{x})$ has full rank and h_x is continuous, it follows that there exists a $\delta > 0$ such that $h_x(x)$ has full rank in $M = \{x \mid \| x - \bar{x} \| \leq \delta\}$. Then there exists a constant K such that all elements of $h_x^T (h_x h_x^T)^{-1}$ are less than $K/\sqrt{n}$ for all $x \in M$.

Since h is continuous with $h(\bar{x}) = 0$, it follows that there exists a $\delta' > 0$ such that $\| h(x) \| \leq \delta/4K$ if $\| x - \bar{x} \| \leq \delta'$. Let $\delta'' = \min(\delta/2, \delta')$ and define $M' = \{x \mid \| x - \bar{x} \| \leq \delta''\}$.

Let x_0 be a point in M' with $h(x_0) \neq 0$ and study the differential equation

$$\begin{cases} \dot{x}(t) = - h_x(x)^T \left(h_x(x) h_x(x)^T\right)^{-1} h(x) \\ x(0) = x_0 \end{cases}$$

The right hand side of this differential equation is a continuously differentiable function and therefore satisfies a Lipschitz condition, which implies that a solution exists in a neighbourhood of x_0. Study the function

$$P(t) = h\left(x(t)\right)^T h\left(x(t)\right)$$

It satisfies

$$\dot{P}(t) = - 2P(t)$$

which has the solution

$$P(t) = P(0)e^{-2t}$$

Along the curve x(t) it is consequently true that

$$\| h\left(x(t)\right) \|_2 = \| h(x_0) \|_2 e^{-t} \tag{3.6}$$

Since

$$x(t) = x_0 - \int_0^t h_x^T (h_x h_x^T)^{-1} h \, dt$$

it follows that

$$\| x(t) - \bar{x} \| \leq \| x_0 - \bar{x} \| + K \| h(x_0) \| \int_0^t e^{-t} dt \leq \delta'' + \delta/4 \leq 3\delta/4$$

The solution x(t) therefore never leaves the set $\{x \mid \| x - \bar{x} \| \leq \leq 3\delta/4\}$ and can consequently be continued indefinitely. From the inequality

$$\| \dot{x}(t) \| \leq K \| h(x_0) \| e^{-t}$$

it follows that

$$\lim_{t\to\infty} x(t) = \hat{x}$$

exists. Equation (3.6) implies that $h(\hat{x}) = 0$. Then

$$F(x_0, \tilde{\mu}(x_0), c) - F(\hat{x}, \tilde{\mu}(\hat{x}), c) =$$

$$= - \int_0^\infty F_x \left[x(t), \tilde{\mu}(x(t)), c \right] \dot{x}(t) dt =$$

$$= \int_0^\infty [h^T \tilde{\mu}_x h_x^T (h_x h_x^T)^{-1} h + c h^T h] dt$$

The function $\tilde{\mu}_x h_x^T (h_x h_x^T)^{-1}$ is continuous on M and therefore there exists a constant K_1 such that

$$\left| h^T \tilde{\mu}_x h_x^T (h_x h_x^T)^{-1} h \right| \leq K_1 \| h \|^2$$

for all x in M. If c is chosen such that $c > K_1$ then

$$\int_0^\infty [h^T \tilde{\mu}_x h_x^T (h_x h_x^T)^{-1} h + c h^T h] dt \geq$$

$$\geq (c - K_1) \int_0^\infty \| h(x) \|^2 dt = (c - K_1) \| h(x_0) \|^2 > 0$$

This relation implies that

$$F(x_0, \tilde{\mu}(x_0), c) > F(\hat{x}, \tilde{\mu}(\hat{x}), c) \geq F(\bar{x}, \tilde{\mu}(\bar{x}), c)$$

where the last inequality follows from the fact that $\hat{x}$ satisfies $h(\hat{x}) = 0$. Consequently

$$F(x, \tilde{\mu}(x), c) > F(\bar{x}, \tilde{\mu}(\bar{x}), c)$$

for all x with $\| x - \bar{x} \| \leq \delta$" and $h(x) \neq 0$.

Since $F(x, \tilde{\mu}(x), c) \geq F(\bar{x}, \tilde{\mu}(x), c)$ for all x with $h(x) = 0$ the theorem is proved. □

Corollary. If $f(x) > f(\bar{x})$ for all $x \neq \bar{x}$ satisfying $h(x) = 0$, then $F(x, \tilde{\mu}(x), c) > F(\bar{x}, \tilde{\mu}(\bar{x}), c)$ for all $x \neq \bar{x}$ such that $\| x - \bar{x} \|$ is sufficiently small.

The result shows that the method employing $\mu = \tilde{\mu}(x)$ is slightly more general than the one using $\mu = \bar{\mu}$ in the augmented Lagrangian. $F(x, \bar{\mu}, c)$ might not have a minimum for $x = \bar{x}$ in problems not satisfying condition C3. This is illustrated by the following example.

Example 3.2.

$$f(x) = x_2^4 + x_1 x_2 \qquad h(x) = x_1$$

$$\bar{x} = (0,0) \qquad \bar{\mu} = 0$$

$$F(x, 0, c) = x_2^4 + x_1 x_2 + \frac{c}{2} x_1^2$$

Since the second derivative of F at $\bar{x}$ is indefinite, $\bar{x}$ is not a local minimum of F. If $\mu = \tilde{\mu}(x)$ is chosen, then

$h_x(x) = (1,0)$

$\tilde{\mu}(x) = - f_{x_1} = - x_2$

$F\left(x, \tilde{\mu}(x), c\right) = x_2^4 + \frac{c}{2} x_1^2$

which has a local minimum at $x = \bar{x}$.

The global properties of $\tilde{\mu}(x)$ are, however, not as favourable as those of $\mu = \bar{\mu}$. Even for well behaved functions f and h, $F\left(x, \tilde{\mu}(x), c\right)$ might not have a global minimum at $\bar{x}$. This is shown by the following example.

Example 3.3.

$f(x) = x_2^2 + (1+x_2^4)x_1^2$

$h(x) = x_1$

$\bar{x} = (0,0) \qquad\qquad \bar{\mu} = 0$

Then $\tilde{\mu}(x) = - 2x_1(1+x_2^4)$ and

$F\left(x, \tilde{\mu}(x), c\right) = x_2^2 - x_1^2(1+x_2^4) + \frac{c}{2} x_1^2$

which is not bounded from below.

If $\mu = \bar{\mu}$ is used then

$F(x, \bar{\mu}, c) = F(x, 0, c) = x_2^2 + (1+x_2^4)x_1^2 + \frac{c}{2} x_1^2$

which has a global minimum at $x = \bar{x}$.

There are also some practical difficulties in using $\tilde{\mu}(x)$. To compute F, not only f and h, but also f_x and h_x are needed because they are used in the expression for $\tilde{\mu}$.

If the derivative F_x is calculated, second derivatives of f and h are required and so on. The extension to inequality constraints also has some difficulties because the augmented function has a gradient which is discontinuous at some points, see Fletcher (1973). It is therefore interesting to study the iterative methods of updating the multipliers. This is done in the next section.

3.4. Iterative Methods for the Multipliers.

There are many ways of updating the multipliers iteratively. Both x and the multipliers can be changed simultaneously, the multipliers can be updated after some iterations where only x was changed or the multipliers may remain constant until a complete minimization with respect to x is finished. The last alternative, which is probably the most widely used one, can be summarized as follows.

Algorithm 3.1.

(o) Choose $u^{(0)}$ and put $k = 0$.

(i) Minimize $F(x,u^{(k)},c^{(k)})$ with respect to x. Call the result $x^{(k)}$.

(ii) Update $u^{(k)}$ and/or $c^{(k)}$ and denote the result $u^{(k+1)}$, $c^{(k+1)}$. Put $k = k + 1$ and go to step (i).

An example of an updating rule for the multipliers is

$$\mu^{(k+1)} = \mu^{(k)} + c^{(k)} h(x^{(k)})$$

$$\lambda^{(k+1)} = \left(\lambda^{(k)} + c^{(k)} g(x^{(k)})\right)_+$$

This method has been suggested by Powell (1969) and Hestenes (1969) in the equality constrained case and by Rockafellar (1974) and Bertsekas (1973) in the general case. The updating formula can be written

$$\mu_j^{(k+1)} - \mu_j^{(k)} = ch_j \, x^{(k)} \qquad j = 1, \ldots, p$$

$$\lambda_j^{(k+1)} - \lambda_j^{(k)} \begin{cases} = cg_j\left(x^{(k)}\right) & \text{if } cg_j\left(x^{(k)}\right) + \lambda_j^{(k)} \geq 0 \\ \\ = -\lambda_j^{(k)} & \text{otherwise} \end{cases}$$

A comparison with equation (3.3) shows that the changes in the multipliers are in the direction of the gradient. The method can therefore be regarded as a steepest ascent algorithm in the multiplier space.

The following second order formula is then a natural extension.

$$G_{uu}\left(u^{(i+1)} - u^{(i)}\right) = - G_u^T$$

To compute G_{uu} the second derivative of F is needed. A straightforward use of this updating rule is then only possible when the second derivatives of f, h and g can be calculated. However, if a quasi-Newton method is used to minimize F, an estimate of F_{xx} at the minimum is given. If this approximate value of F_{xx} is used to calculate $G_{\mu\mu}$, only function values and gradients are needed. This method has been used by Glad (1973) for equality constraints and by Fletcher (1974) for combined equality and inequality constraints.

The essential feature of algorithm 3.1 is that a complete minimization of F is done before the multipliers are updated. As mentioned before, several methods are based on the idea of updating

u more often. Miele et al (1971) use the formula

$$\mu = - (h_x h_x^T)^{-1} h_x f_x^T$$

to update the multipliers after each iteration when using a conjugate gradient method.

A modification of this method, using the Davidon-Fletcher-Powell method, is investigated by O'Doherty and Pierson (1974). Tripathi and Narendra (1972) use the formula

$$\mu^{(i+1)} = \mu^{(i)} + \alpha c h(x^{(i)})$$

after each line search in a conjugate gradient method. The parameter α is a step length parameter which is given a value between 0 and 1.

The idea of updating the multipliers after each iteration with respect to x seems intuitively to promise faster convergence than the method of updating the multipliers only when a minimum with respect to x has been found. Therefore this class of algorithms will be investigated in the remainder of the chapter. The iterations with respect to x will be assumed to be carried out by a quasi-Newton method, the so called BFGS method, which is known to be efficient when solving unconstrained problems, see Fletcher (1972). The discussion is broken down into three parts. In 3.5 local convergence is studied, in 3.6 global convergence and in 3.7 practical experience.

3.5. Local Convergence

To study the local convergence, it is first shown that only equality constraints need to be considered. Then the class of algorithms to be studied, is presented (algorithm 3.2). Since this algorithm uses the techniques of the so called quasi-Newton methods, results by Dennis and Moré (1974) for the unconstrained quasi-Newton algorithms are generalized to the constrained case in lemmas 3.2 and 3.3. With these tools, two versions of algorithm 3.2 can be shown to be linearly convergent in theorems 3.9 and 3.10 and example 3.4. A third version of algorithm 3.2 is then shown to be superlinearly convergent in theorem 3.12.

Assume that conditions C1 - C4 hold at a point $\bar{x}$, which is a local minimum of the constrained optimization problem. Let c be held constant at a value large enough for $F_{xx}(\bar{x},\bar{u},c)$ to be positive definite. Define

$$M(\delta) = \{(x,u) \mid \| x - \bar{x} \| \leq \delta, \ \| u - \bar{u} \| \leq \delta\}$$

Since local convergence is considered, only points (x,u) that belong to $M(\delta)$ for an arbitrarily small $\delta > 0$, need to be considered. Let δ be chosen small enough for the following propositions to be true.

- $F_{xx}(x,u,c) > 0$ for all (x,u) in $M(\delta)$
- $(cg_i(x) + \lambda_i) < 0$ $i \in I$ and

  $(cg_i(x) + \lambda_i) > 0$ $i \in A$ for all (x,u) in $M(\delta)$

Then the inactive inequality constraints can be disregarded and the active ones treated as equality constraints. Therefore it is sufficient to study only equality constraints in this section.

An algorithm of the quasi-Newton type is used to perform the minimization of $F(x,u,c)$ with respect to x. The constrained minimization algorithm can then be described by the following equations.

$$\mu^{(k+1)} = \Omega\left(\mu^{(k)}, x^{(k)}, H^{(k)}, c\right)$$

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} H^{(k)} F_x^T\left(x^{(k)}, \mu^{(k+1)}, c\right)$$

Here $H^{(k)}$ is an approximation of F_{xx}^{-1}. The step length $\alpha^{(k)}$ is determined by some procedure for approximate minimization of

$$F\left[x^{(k)} - \alpha H^{(k)} F_x^T\left(x^{(k)}, \mu^{(k+1)}, c\right), \mu^{(k+1)}, c\right]$$

with respect to α. Since many linear minimization algorithms first try $\alpha = 1$, it is of interest to study the special case $\alpha^{(k)} = 1$, all k. It will be assumed that $H^{(k)}$ is updated according to the Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula, see Broyden (1972) and Fletcher (1972). The notation $B^{(k)} = \left(H^{(k)}\right)^{-1}$ is used. $B^{(k)}$ is then an approximation of F_{xx}. The full description of the algorithm is

Algorithm 3.2.

$$\mu^{(k+1)} = \Omega\left(\mu^{(k)}, x^{(k)}, H^{(k)}, c\right)$$

$$x^{(k+1)} = x^{(k)} - H^{(k)} F_x^T\left(x^{(k)}, \mu^{(k+1)}, c\right)$$

$$H^{(k+1)} = H^{(k)} + \frac{\left(s^{(k)} - H^{(k)} y^{(k)}\right)\left(s^{(k)}\right)^T + s^{(k)}\left(s^{(k)} - H^{(k)} y^{(k)}\right)^T}{\left(y^{(k)}\right)^T s^{(k)}} -$$

$$- \frac{\left(y^{(k)}\right)^T\left(s^{(k)} - H^{(k)} y^{(k)}\right) s^{(k)}\left(s^{(k)}\right)^T}{\left(\left(y^{(k)}\right)^T s^{(k)}\right)^2}$$

where

$$y^{(k)} = F_x^T\left(x^{(k+1)}, \mu^{(k+1)}, c\right) - F_x^T\left(x^{(k)}, \mu^{(k+1)}, c\right)$$

$$s^{(k)} = x^{(k+1)} - x^{(k)}$$

The properties of this algorithm depend to a great extent on the properties of the updating formula for H. Dennis and Moré (1974) have studied quasi-Newton methods in the unconstrained case and produced the following useful result.

Lemma 3.2. Let M be a nonsingular symmetric matrix such that $\| Mz - M^{-1}w \| \leq b \| M^{-1}w \|$ for some b, $0 \leq b \leq 1/3$ and some vectors z and w, with $w \neq 0$. Then $z^T w > 0$ and $\bar{B}$ can be defined by

$$\bar{B} = B + \frac{(z-Bw)z^T + z(z-Bw)^T}{z^T w} - \frac{w^T(z-Bw)\, z\, z^T}{(z^T w)^2}$$

where B is symmetric. Let $\| \cdot \|_M$ be the matrix norm defined by $\| Q \|_M = \| MQM \|_F$ where $\| \cdot \|_F$ is the Frobenius norm

$$\| P \|_F = \sum_{ij} |p_{ij}|^2$$

Then there are positive constants α, α_1 and α_2 (depending only on M) such that for any symmetric matrix A

$$\| \bar{B} - A \|_M \leq \left\{ \sqrt{1-\alpha\theta^2} + \alpha_1 \| Mz - M^{-1}w \| / \| M^{-1}w \| \right\} \cdot \| B - A \|_M +$$

$$+ \alpha_2 \| z - Aw \| / \| M^{-1}w \|$$

where $0 < \alpha \leq 1$ and

$$\theta = \frac{\| M(B-A)w \|}{\| B - A \|_M \, \| M^{-1}w \|} \qquad \text{for } B \neq A$$

$\theta = 0$ for $B = A$

Proof. See Dennis and Moré (1974). □

The next lemma generalizes this result to the constrained case.

Lemma 3.3. Assume that $F_{xx}(\bar{x},\bar{\mu},c) > 0$ and that there exists a constant K such that

$$\| F_{xx}(x,\mu,c) - F_{xx}(\bar{x},\bar{\mu},c) \| \leq K(\| x - \bar{x} \| + \| \mu - \bar{\mu} \|)$$

for all (x,μ) in some neighbourhood of $(\bar{x},\bar{\mu})$. Then, if the sequence $\{(x_k,\mu_k)\}$ converges to $(\bar{x},\bar{\mu})$, there exists a k_0, such that $H^{(k)}$ satisfies the following inequality for $k \geq k_0$

$$\| H^{(k+1)} - F_{xx}^{-1}(\bar{x},\bar{\mu},c) \|_M \leq$$

$$\leq \left\{ \sqrt{1-\alpha(\theta^{(k)})^2} + \alpha_3 \sigma^{(k)} \right\} \| H^{(k)} - F_{xx}^{-1}(\bar{x},\bar{\mu},c) \|_M + \alpha_4 \sigma^{(k)}$$

where $0 < \alpha \leq 1$, α_3 and α_4 are positive constants and

$$\theta^{(k)} = \frac{\| M(H^{(k)} - F_{xx}^{-1}(\bar{x},\bar{\mu},c)) y^{(k)} \|}{\| H^{(k)} - F_{xx}^{-1}(\bar{x},\bar{\mu},c) \|_M \| M^{-1} y^{(k)} \|}$$

for $H^{(k)} \neq F_{xx}^{-1}(\bar{x},\bar{\mu},c)$

$\theta^{(k)} = 0$

for $H^{(k)} = F_{xx}^{-1}(\bar{x},\bar{\mu},c)$

$$M = F_{xx}(\bar{x},\bar{\mu},c)^{1/2}$$

$$\sigma^{(k)} = \max\left(\| x^{(k)} - \bar{x} \|, \| x^{(k+1)} - \bar{x} \|\right) + \| \mu^{(k+1)} - \bar{\mu} \|$$

Proof. From the mean value theorem, see Ortega and Rheinboldt (1970), it follows that

$$\| F_x^T\left(x^{(k+1)},\mu^{(k+1)},c\right) - F_x^T\left(x^{(k)},\mu^{(k+1)},c\right) - F_{xx}(\bar{x},\bar{\mu},c)\cdot$$

$$\cdot\left(x^{(k+1)} - x^{(k)}\right) \| \leq$$

$$\leq \sup_{0\leq\theta\leq 1} \| F_{xx}\left[x^{(k)} + \theta\left(x^{(k+1)} - x^{(k)}\right),\mu^{(k+1)},c\right] -$$

$$- F_{xx}(\bar{x},\bar{\mu},c) \| \cdot \| x^{(k+1)} - x^{(k)} \| \leq$$

$$\leq K\left[\max\left(\| x^{(k+1)} - \bar{x} \|, \| x^{(k)} - \bar{x} \|\right) + \| \mu^{(k+1)} - \bar{\mu} \|\right] \cdot$$

$$\cdot \| x^{(k+1)} - x^{(k)} \|$$

if $x^{(k)}$ and $x^{(k+1)}$ are sufficiently close to $\bar{x}$ and $\mu^{(k+1)}$ sufficiently close to $\bar{\mu}$. Using the definitions of y, s, M and σ gives

$$\| y^{(k)} - M^2 s^{(k)} \| \leq K\sigma^{(k)} \| s^{(k)} \| \qquad (3.7)$$

Since $\left(x^{(k)},\mu^{(k)}\right)$ is assumed to converge to $(\bar{x},\bar{\mu})$, there is a k_0 such that, for any $\varepsilon > 0$, $\sigma^{(k)} \leq \varepsilon$ for $k \geq k_0$. (3.7) then implies that there exist constants $K_1 > 0$, $K_2 > 0$ and k_1 such that

$$\frac{1}{K_1}\| s^{(k)} \| \leq \| y^{(k)} \| \leq K_2 \| s^{(k)} \| \qquad (3.8)$$

$$\| y^{(k)} - M^2 s^{(k)} \| \leq KK_1\sigma^{(k)} \| y^{(k)} \| \qquad (3.9)$$

for $k \geq k_1$.

Using (3.9) gives

$$\| Ms^{(k)} - M^{-1}y^{(k)} \| = \| M^{-1}(y^{(k)} - M^2 s^{(k)}) \| \leq$$

$$\leq K_3 \sigma^{(k)} \| M^{-1}y^{(k)} \|$$

for some constant $K_3 > 0$. If k_0 is large enough

$$\| Ms^{(k)} - M^{-1}y^{(k)} \| \leq \frac{1}{3} \| M^{-1}y^{(k)} \| \qquad k \geq k_0$$

and Lemma 3.2 is applicable. □

From this lemma the asymptotic properties of $H^{(k)}$ can be deduced. They are given in the following lemma.

Lemma 3.4. Let F_{xx} satisfy the assumptions of lemma 3.2 and assume that

$$\sum_1^\infty \sigma^{(k)} < \infty$$

Then

$$\lim_{k\to\infty} \| F_{xx}^{1/2}(\bar{x},\bar{\mu},c) H^{(k)} F_{xx}^{1/2}(\bar{x},\bar{\mu},c) - I \|_F$$

exists, and

$$\lim_{k\to\infty} \| \left(H^{(k)} - F_{xx}^{-1}(\bar{x},\bar{\mu},c)\right) \hat{y}^{(k)} \| = 0 \qquad (3.10)$$

where $\hat{y}^{(k)} = y^{(k)} / \| y^{(k)} \|$.

Proof. Lemma 3.3 shows that $H^{(k)}$ satisfies an inequality that is completely analogous to the inequality given by Dennis and Moré (1974) in the unconstrained case. The result then follows from their theorem 3.4. □

Corollary. If, in addition to the assumptions of lemma 3.4, $\| B^{(k)} \|$ is uniformly bounded, then

$$\lim_{k\to\infty} \| \left(B^{(k)} - F_{xx}(\bar{x},\bar{\mu},c)\right)\hat{s}^{(k)} \| = 0 \qquad (3.11)$$

where $\hat{s}^{(k)} = s^{(k)}/\| s^{(k)} \|$.

Proof.

$$\| \left(B^{(k)} - F_{xx}(\bar{x},\bar{\mu},c)\right)\hat{s}^{(k)} \| \leq$$

$$\leq \frac{\| y^{(k)} \|}{\| s^{(k)} \|} \| B^{(k)} \| \| \left(H^{(k)} - F_{xx}^{-1}(\bar{x},\bar{\mu},c)\right)\left(\hat{y}^{(k)} + r\right) \|$$

where $\| r \| \leq K\sigma^{(k)}$ for some $K > 0$. □

Equation (3.11) does not imply that $B^{(k)}$ converges to $F_{xx}(\bar{x},\bar{\mu},c)$, because it is quite possible that all $\hat{s}^{(k)}$ lie in a subspace. In fact, Powell has shown examples for the unconstrained case, where $B^{(k)}$ does not converge to the second derivative of the minimum point, see Dennis and Moré (1974). However, since the updating formula for x in algorithm 3.2 can be written

$$B^{(k)}s^{(k)} = -F_x^T\left(x^{(k)},\mu^{(k+1)},c\right)$$

it follows that it is not the matrix $B^{(k)}$ itself, but the product $B^{(k)}s^{(k)}$ that is interesting.

To prove that

$$\sum_{1}^{\infty} \sigma^{(k)}$$

is convergent, which was one of the assumptions made in order to show (3.11), it is necessary to specify the function Ω of algorithm 3.2 more precisely. First study the case where Ω depends only on x

$$\Omega(\mu, x, H, c) = \varphi(x) \tag{3.12}$$

and φ is a continuously differentiable function with $\varphi(\bar{x}) = \bar{\mu}$. The function

$$\varphi(x) = -\left(h_x(x) h_x(x)^T\right)^{-1} h_x(x) f_x(x)^T \tag{3.12a}$$

mentioned in sections 3.3 and 3.4, is an example of a choice of φ that satisfies these conditions. With Ω given by (3.12) it is possible to show linear convergence of algorithm 3.2.

Theorem 3.10. Let c_0 be a number such that $F_{xx}(\bar{x}, \bar{\mu}, c)$ satisfies the assumptions of lemma 3.3 for $c > c_0$ and let r be an arbitrary number with $0 < r < 1$. Assume that algorithm 3.2 with Ω given by (3.12) is used. Then there is a number $c_1(r) > c_0$ such that, for each $c > c_1(r)$ there are constants $\varepsilon_1(r,c) > 0$ and $\varepsilon_2(r,c) > 0$ with the property that, if

$$\| x^{(0)} - \bar{x} \| \leq \varepsilon_1 \quad \text{and} \quad \| H^{(0)} - F_{xx}^{-1}(\bar{x}, \bar{\mu}, c) \| \leq \varepsilon_2$$

then

$$\| x^{(k+1)} - \bar{x} \| \leq r \| x^{(k)} - \bar{x} \|$$

i.e. the convergence is at least linear.

Proof. The formulas of algorithm 3.2 together with (3.12) gives

$$x^{(k+1)} = x^{(k)} - H^{(k)} F_x^T \left(x^{(k)}, \varphi(x^{(k)}), c \right)$$

A Taylor expansion gives

$$x^{(k+1)} = x^{(k)} - H^{(k)} \left(F_{xx}(\bar{x}, \bar{\mu}, c) (x^{(k)} - \bar{x}) + \right.$$

$$\left. + h_x^T(\bar{x}) \varphi_x(\bar{x}) (x^{(k)} - \bar{x}) \right) + H^{(k)} R(c, x^{(k)} - \bar{x})$$

where $\| R(c,z) \| / \| z \| \to 0$ as $z \to 0$.

The formula can be rewritten

$$x^{(k+1)} - \bar{x} = \left[(F_{xx}^{-1}(\bar{x}, \bar{\mu}, c) - H^{(k)}) (F_{xx}(\bar{x}, \bar{\mu}, c) + h_x^T(\bar{x}) \varphi_x(\bar{x})) - \right.$$

$$\left. - F_{xx}^{-1}(\bar{x}, \bar{\mu}, c) h_x^T(\bar{x}) \varphi_x(\bar{x}) \right] (x^{(k)} - \bar{x}) +$$

$$+ H^{(k)} R(c, x^{(k)} - \bar{x}) \qquad (3.13)$$

Study the quantity

$$F_{xx}^{-1}(\bar{x}, \bar{\mu}, c) h_x^T(\bar{x})$$

Let $\tilde{c}$ be a number such that $F_{xx}(\bar{x}, \bar{\mu}, \tilde{c})$ is positive definite. Then, for $c > \tilde{c}$,

$$F_{xx}^{-1}(\bar{x}, \bar{\mu}, c) h_x^T(\bar{x}) =$$

$$= (F_{xx}(\bar{x}, \bar{\mu}, \tilde{c}) + (c - \tilde{c}) h_x^T(\bar{x}) h_x(\bar{x}))^{-1} h_x^T(\bar{x}) =$$

$$= \frac{1}{c - \tilde{c}} F_{xx}^{-1}(\bar{x}, \bar{\mu}, \tilde{c}) h_x^T(\bar{x}) \left(\frac{1}{c - \tilde{c}} I + h_x(\bar{x}) F_{xx}^{-1}(\bar{x}, \bar{\mu}, \tilde{c}) h_x^T(\bar{x}) \right)^{-1}$$

Since $h_x(\bar{x})$ is assumed to have full rank, it follows that

$$F_{xx}^{-1}(\bar{x},\bar{\mu},c)h_x^T(\bar{x}) \to 0 \text{ as } c \to \infty$$

Now choose a value of c such that

$$\| F_{xx}^{-1}(\bar{x},\bar{\mu},c)h_x^T(\bar{x})\varphi_x(\bar{x}) \| \leq \frac{r}{3} \tag{3.14}$$

Next choose ε_H such that

$$\| \left(F_{xx}^{-1}(\bar{x},\bar{\mu},c) - H^{(k)}\right)\left(F_{xx}(\bar{x},\bar{\mu},c) + h_x^T(\bar{x})\varphi_x(\bar{x})\right) \| \leq \frac{r}{3} \tag{3.15}$$

for all $H^{(k)}$ satisfying

$$\| H^{(k)} - F_{xx}^{-1}(\bar{x},\bar{\mu},c) \| \leq \varepsilon_H \tag{3.16}$$

Finally choose ε_x such that

$$\| H^{(k)}R(c,x^{(k)} - \bar{x}) \| \leq \frac{r}{3}\| x^{(k)} - \bar{x} \| \tag{3.17}$$

for all $H^{(k)}$ satisfying (3.16) and all $x^{(k)}$ satisfying

$$\| x^{(k)} - \bar{x} \| \leq \varepsilon_x \tag{3.18}$$

Then, if $H^{(k)}$ satisfies (3.16) and $x^{(k)}$ satisfies (3.18), the relations (3.13), (3.14), (3.15) and (3.17) give

$$\| x^{(k+1)} - \bar{x} \| \leq r\| x^{(k)} - \bar{x} \|$$

Using the notation

$$\| H^{(k)} - F_{xx}^{-1}(\bar{x},\bar{\mu},c) \| = \beta_k$$

$$\| x^{(k)} - \bar{x} \| = \delta_k$$

$$\xi_k = \max(\delta_{k+1}, \delta_k)$$

and applying lemma 3.3 gives the formulas

$$\beta_{k+1} \leq (1+a\xi_k)\beta_k + b\xi_k \qquad a, b \text{ positive constants}$$

$$\delta_{k+1} \leq r\delta_k \quad \text{if} \quad \beta_k \leq \varepsilon_H, \ \delta_k \leq \varepsilon_x$$

Define

$$m = \delta_0 \sum_0^\infty r^j$$

and choose β_0 and δ_0 such that $\beta_0 \leq \varepsilon_H$, $\delta_0 \leq \varepsilon_x$ and $e^{am}(\beta_0+bm) \leq \varepsilon_H$. We will show by induction that $\beta_k \leq \varepsilon_H$ for all k. Suppose that this is true for β_i, $0 \leq i \leq k$. Then $\delta_{i+1} \leq$ $\leq r\delta_i$, $0 \leq i \leq k$, and

$$\beta_{i+1} \leq (1+a\delta_i)\beta_i + b\delta_i \qquad 0 \leq i \leq k$$

Define

$$\gamma_i = \prod_{j=1}^{i-1} (1+a\delta_j)$$

Then

$$\frac{\beta_{k+1}}{\gamma_{k+1}} \leq \frac{\beta_k}{\gamma_k} + b\delta_k \leq \ldots \leq \beta_0 + b \sum_{i=0}^{k} \delta_i$$

Since

$$\gamma_i \leq \exp\Big(a \sum_{j=0}^{i-1} \delta_j\Big)$$

it follows that

$$\beta_{k+1} \leq \exp\Big(a\delta_0 \sum_{j=0}^{k} r^j\Big)\Big(\beta_0 + b\delta_0 \sum_{j=0}^{k} r^j\Big) \leq e^{am}(\beta_0 + bm) \leq \varepsilon_H$$

Consequently $\beta_k \leq \varepsilon_H$ for all k and $\delta_{k+1} \leq r\delta_k$ all k. □

Corollary. If the assumptions of theorem 3.10 are satisfied, then

$$\| \big(B^{(k)} - F_{xx}(\bar{x},\bar{\mu},c)\big)\hat{s}^{(k)} \| \to 0$$

Proof. Follows from the corollary of lemma 3.4 and theorem 3.10. □

A similar result can be established for the choice

$$\Omega(\mu,c,H,c) = \mu + ch(x) \tag{3.19}$$

Theorem 3.11. Let c_0 be a number such that $F_{xx}(\bar{x},\bar{\mu},c)$ satisfies the assumptions of lemma 3.3 for $c > c_0$ and let r be an arbitrary number with $0 < r < 1$. Assume that algorithm 3.2 with Ω given by (3.19) is used. Then there is a constant $c_1(r) > c_0$ such that, for each $c > c_1(r)$ there are constants $\varepsilon_1(r,c) > 0$, $\varepsilon_2(r,c) > 0$ and $\varepsilon_3(r,c) > 0$ with the property that, if

$$\| x^{(0)} - \bar{x} \| \leq \varepsilon_1, \quad \| H^{(0)} - F_{xx}^{-1}(\bar{x},\bar{\mu},c) \| \leq \varepsilon_2 \text{ and}$$

$$\| \mu^{(0)} - \bar{\mu} \| \leq \varepsilon_3 \quad \text{then}$$

$$\| x^{(k+1)} - \bar{x} \| + \| \mu^{(k+2)} - \bar{\mu} \| \leq$$

$$\leq r\left(\| x^{(k)} - \bar{x} \| + \| \mu^{(k+1)} - \bar{\mu} \| \right)$$

Proof. We have

$$x^{(k+1)} = x^{(k)} - H^{(k)} F_x^T \left(x^{(k)}, \mu^{(k+1)}, c \right)$$

$$\mu^{(k+2)} = \mu^{(k+1)} + ch\left(x^{(k+1)} \right)$$

A Taylor expansion gives

$$\begin{pmatrix} x^{(k+1)} - \bar{x} \\ \mu^{(k+2)} - \bar{\mu} \end{pmatrix} = \begin{pmatrix} I - H^{(k)} F_{xx} & - H^{(k)} h_x^T \\ ch_x \left(I - H^{(k)} F_{xx} \right) & I - ch_x H^{(k)} h_x^T \end{pmatrix} \begin{pmatrix} x^{(k)} - \bar{x} \\ \mu^{(k+1)} - \bar{\mu} \end{pmatrix} + R \qquad (3.20)$$

where $F_{xx} = F_{xx}(\bar{x}, \bar{\mu}, c)$, $h_x = h_x(\bar{x})$ and R satisfies

$$\| R \| / \left(\| x^{(k)} - \bar{x} \| + \| \mu^{(k+1)} - \bar{\mu} \| \right) \to 0$$

as $x^{(k)} \to \bar{x}$, $\mu^{(k+1)} \to \bar{\mu}$.

The matrix in (3.20) can be rewritten

$$\begin{pmatrix} \left(F_{xx}^{-1} - H^{(k)} \right) F_{xx} & \left(F_{xx}^{-1} - H^{(k)} \right) h_x - F_{xx}^{-1} h_x^T \\ ch_x \left(F_{xx}^{-1} - H^{(k)} \right) F_{xx} & I - ch_x F_{xx}^{-1} h_x^T - ch_x \left(H^{(k)} - F_{xx}^{-1} \right) h_x^T \end{pmatrix} \qquad (3.21)$$

As shown in the proof of theorem 3.9, $F_{xx}^{-1} h_x^T$ can be made arbitrarily small by choosing c large enough. The matrix $I - ch_x F_{xx}^{-1} h_x^T$ can be written

$$I - ch_x F_{xx}^{-1} h_x^T =$$

$$= \frac{c}{(c-\tilde{c})^2}\left[\frac{1}{c-\tilde{c}} I + h_x(\bar{x}) F_{xx}^{-1}(\bar{x},\bar{\mu},\tilde{c}) h_x^T(\bar{x})\right]^{-1} - \frac{\tilde{c}}{c-\tilde{c}} I$$

where $\tilde{c}$ is a constant, chosen such that $F_{xx}(\bar{x},\bar{\mu},\tilde{c})$ is positive definite. Consequently $I - ch_x F_{xx}^{-1} h_x^T \to 0$ as $c \to \infty$.

Now choose c such that

$$\| I - ch_x F_{xx}^{-1} h_x^T \| \leq \frac{r}{8}$$

$$\| F_{xx}^{-1} h_x^T \| \leq \frac{r}{8}$$

Then choose ε_H such that

$$\| (F_{xx}^{-1} - H^{(k)}) F_{xx} \| \leq \frac{r}{4}$$

$$\| (F_{xx}^{-1} - H^{(k)}) h_x \| \leq \frac{r}{8}$$

$$\| ch_x (F_{xx}^{-1} - H^{(k)}) F_{xx} \| \leq \frac{r}{4}$$

$$\| ch_x (H^{(k)} - F_{xx}^{-1}) h_x^T \| \leq \frac{r}{8}$$

if

$$\| H^{(k)} - F_{xx}^{-1} \| \leq \varepsilon_H \qquad (3.22)$$

Next choose an ε_x such that

$$\| R \| \leq \frac{r}{4}(\| x^{(k)} - \bar{x} \| + \| \mu^{(k+1)} - \bar{\mu} \|)$$

for $x^{(k)}$ and $\mu^{(k+1)}$ satisfying

$$\| x^{(k)} - \bar{x} \| + \| \mu^{(k+1)} - \bar{\mu} \| \leq \varepsilon_x \qquad (3.23)$$

Then, for all $x^{(k)}$, $\mu^{(k+1)}$ and $H^{(k)}$ satisfying (3.22) and (3.23)

$$\| x^{(k+1)} - \bar{x} \| + \| \mu^{(k+2)} - \bar{\mu} \| \leq r\left(\| x^{(k)} - \bar{x} \| + \right.$$

$$\left. + \| \mu^{(k+1)} - \bar{\mu} \| \right)$$

The remainder of the proof is analogous to the proof of theorem 3.10. □

Corollary. If the assumptions of theorem 3.11 are true, then

$$\| \left(B^{(k)} - F_{xx}\right) \hat{s}^{(k)} \| \to 0 \qquad \text{as } k \to \infty$$

Theorems 3.10 and 3.11 might give the impression that it is advantageous to choose c as large as possible to get a small r. This is usually not true in practice because a large value of c results in an ill-conditioned problem. The reason is that the ratio between the largest and smallest eigenvalues of the matrix

$$F_{xx}(\bar{x},\bar{u},c) = \left(L_{xx}(\bar{x},\bar{u}) + c h_x(\bar{x})^T h_x(\bar{x})\right)$$

increases towards infinity as c goes to infinity. This is known from penalty function theory, see Luenberger (1973). It is then difficult to get the small values of

$$\| H^{(k)} - F_{xx}^{-1}(\bar{x},\bar{\mu},c) \| \quad \text{and} \quad \| x^{(k)} - \bar{x} \|$$

that the theorems require.

Theorems 3.10 and 3.11 show that (3.12) and (3.19) in algorithm 3.2 give at least linear convergence locally. Dennis and Moré (1974) showed that, for unconstrained problems, the fact that $\| (B^{(k)} - F_{xx})\hat{s}^{(k)} \| \to 0$ results in superlinear convergence. This is not the case for algorithm 3.2 with Ω defined by (3.12) or (3.19), as shown by the following example.

Example 3.4.

$$f(x) = \tfrac{1}{2}(x_1^2 + x_2^2) \qquad h(x) = x_1 - 1$$

Choose

$$x^{(0)} = (0,0)^T \qquad \mu^{(0)} = c \qquad H^{(0)} = \begin{pmatrix} \frac{1}{1+c} & 0 \\ 0 & 1 \end{pmatrix}$$

Then (3.12a) and (3.19) give the same result and

$$\mu^{(k+1)} = - x_1^{(k)} \qquad H^{(k)} = H^{(0)} \qquad x_2^{(k)} = 0$$

$$x_1^{(k+1)} - 1 = \frac{x_1^{(k)} - 1}{c + 1}$$

i.e. linear convergence.

The formula

$$\Omega(\mu,x,H,c) = (h_x H h_x^T)^{-1} (h - h_x H f_x^T) - ch \tag{3.24}$$

has been suggested by O'Doherty and Pierson (1974), as a modification of the method used by Miele (1971). A different way of deriving this formula is as follows. The first order optimality conditions give the following equations.

$$F_x(x,\mu,c) = 0$$

$$h(x) = 0$$

Solving these equations using a Newton method gives

$$F_{xx}\delta x + h_x^T\delta\mu = -F_x^T$$

$$h_x\delta x = -h$$

Replacing F_{xx}^{-1} by $H^{(k)}$ and eliminating δx gives the formula for Ω shown above. It may also be noted that this updating formula for μ is a natural extension to the case $F_x \neq 0$, of the one given in Glad (1973).

Since the expression for Ω in equation (3.24) depends on c and H, theorem 3.10 is not applicable. Instead we use the representation

$$\begin{pmatrix} B^{(k)} & h_x^T(x^{(k)}) \\ h_x(x^{(k)}) & 0 \end{pmatrix} \begin{pmatrix} x^{(k+1)} - x^{(k)} \\ \mu^{(k+1)} - \mu^{(k)} \end{pmatrix} = - \begin{pmatrix} F_x^T(x^{(k)},\mu^{(k)},c) \\ h(x^{(k)}) \end{pmatrix} \qquad (3.25)$$

to prove the following theorem.

Theorem 3.12. Let $(x^{(k)},\mu^{(k)})$ be generated by equation (3.25). (As shown above, this is equivalent to the use of (3.24) in algorithm 3.2.) Assume that c is chosen large enough for $F_{xx}(\bar{x},\bar{\mu},c)$ to be positive definite and that

$$\| F_{xx}(x,\mu,c) - F_{xx}(\bar{x},\bar{\mu},c) \| \leq K(\| x - \bar{x} \| + \| \mu - \bar{\mu} \|)$$

holds for all (x,μ) in some neighbourhood of $(\bar{x},\bar{\mu})$. Let r be a number, $0 < r < 1$, and let z^T denote the vector $[x^T,\mu^T]$.

Then there exist constants ε_1 and ε_2 such that

$$\| z^{(k+1)} - \bar{z} \| \leq r \| z^{(k)} - \bar{z} \| \quad \text{if}$$

$$\| H^{(0)} - F_{xx}^{-1}(\bar{x}, \bar{\mu}, c) \| \leq \varepsilon_1 \quad \text{and}$$

$$\| z^{(0)} - \bar{z} \| \leq \varepsilon_2$$

Proof. We have

$$\begin{bmatrix} B^{(k)} & h_x^T(x^{(k)}) \\ h_x(x^{(k)}) & 0 \end{bmatrix} \begin{bmatrix} x^{(k+1)} - \bar{x} \\ \mu^{(k+1)} - \bar{\mu} \end{bmatrix} =$$

$$= \begin{bmatrix} B^{(k)} & h_x^T(x^{(k)}) \\ h_x(x^{(k)}) & 0 \end{bmatrix} \begin{bmatrix} x^{(k)} - \bar{x} \\ \mu^{(k)} - \bar{\mu} \end{bmatrix} - \begin{bmatrix} F_x^T(x^{(k)}, \mu^{(k)}, c) \\ h(x^{(k)}) \end{bmatrix} =$$

$$= \begin{bmatrix} B^{(k)} - F_{xx}(\bar{x}, \bar{\mu}, c) & h_x^T(x^{(k)}) - h_x^T(\bar{x}) \\ h_x(x^{(k)}) - h_x(\bar{x}) & 0 \end{bmatrix} \begin{bmatrix} x^{(k)} - \bar{x} \\ \mu^{(k)} - \bar{\mu} \end{bmatrix} +$$

$$+ R(z^{(k)} - \bar{z})$$

where $\| R(z^{(k)} - \bar{z}) \| / \| z^{(k)} - \bar{z} \| \to 0$ as $z^{(k)} \to \bar{z}$. It follows that there exist constants ε_B and ε_z such that

$$\| z^{(k+1)} - \bar{z} \| \leq r \| z^{(k)} - \bar{z} \| \quad \text{if}$$

$$\| H^{(k)} - F_{xx}^{-1}(\bar{x}, \bar{\mu}, c) \| \leq \varepsilon_B \quad \text{and}$$

$$\| z^{(k)} - \bar{z} \| \leq \varepsilon_z$$

The remaining part of the proof is carried out in the same way as in theorem 3.10. □

Corollary. If the assumptions of this theorem are true, then

$$\| \left(B^{(k)} - F_{xx}(\bar{x},\bar{\mu},c)\right)\hat{s}^{(k)} \| \to 0 \quad \text{as } k \to \infty$$

Proof. Combine theorem 3.12 and lemma 3.4. □

To prove superlinear convergence of (3.25), the following lemma from Dennis and Morê (1974) is needed.

Lemma 3.5. Let the equation $\xi(x) = 0$ with solution $x = \bar{x}$ be solved by the iterative method

$$x^{(k+1)} = x^{(k)} - \left(B^{(k)}\right)^{-1} \xi\left(x^{(k)}\right)$$

If $\xi_x(x)$ is continuous and $\xi_x(\bar{x})$ nonsingular, then

$$\lim_{k\to\infty} \frac{\| \left(B^{(k)} - \xi_x(\bar{x})\right)\left(x^{(k+1)} - x^{(k)}\right) \|}{\| x^{(k+1)} - x^{(k)} \|} = 0$$

is equivalent to

$$\lim_{k\to\infty} \frac{\| x^{(k+1)} - \bar{x} \|}{\| x^{(k)} - \bar{x} \|} = 0 \qquad \text{(superlinear convergence)}$$

Proof. See Dennis and Morê (1974). □

The superlinear convergence of the method based on (3.25) can now be established.

Theorem 3.13. Let

$$z^{(k)} = \left((x^{(k)})^T, (\mu^{(k)})^T \right)^T$$

be generated by algorithm 3.2 using equation (3.24) or equivalently by (3.25). Let the assumptions of theorem 3.12 be valid. Then $z^{(k)}$ converges superlinearly to $\bar{z}$, i.e.

$$\lim_{k\to\infty} \frac{\| z^{(k+1)} - \bar{z} \|}{\| z^{(k)} - \bar{z} \|} = 0$$

Proof. Apply lemma 3.5 to equation (3.25). Superlinear convergence is seen to be equivalent to

$$\frac{\left\| \left(\begin{pmatrix} B^{(k)} & h_x^T(x^{(k)}) \\ h_x(x^{(k)}) & 0 \end{pmatrix} - \begin{pmatrix} F_{xx}(\bar{x},\bar{\mu},c) & h_x^T(\bar{x}) \\ h_x(\bar{x}) & 0 \end{pmatrix} \right) \begin{pmatrix} x^{(k+1)} - x^{(k)} \\ \mu^{(k+1)} - \mu^{(k)} \end{pmatrix} \right\|}{\left\| \begin{pmatrix} x^{(k+1)} - x^{(k)} \\ \mu^{(k+1)} - \mu^{(k)} \end{pmatrix} \right\|} \to 0$$

as $k \to \infty$.

This condition is satisfied if

$$\| (B^{(k)} - F_{xx}(\bar{x},\bar{\mu},c)) \hat{s}^{(k)} \| \to 0, \quad k \to \infty$$

An application of the corollary of theorem 3.12 finishes the proof. □

Remark. All the results of this section remain true if the BFGS updating formula is replaced by the Davidon-Fletcher-Powell (DFP) formula. This follows because the DFP formula is the same as the BFGS formula with $H^{(k)}$ replaced by $B^{(k)}$ and $y^{(k)}$ and $s^{(k)}$ changing places, see Dennis and Moré (1974). The detailed calculations for this case can be found in Glad (1975). □

3.6. Global Convergence Properties.

In the last section it was shown that three updating rules for the multipliers used in algorithm 3.2 all gave local convergence. In practice one cannot be sure of having a good enough starting point for this result to apply and it is therefore desirable to be able to show global convergence. To do this, some modifications of the algorithm discussed in the previous section are necessary.

First a test quantity, $a(x,u)$, that shows if progress is made towards a Kuhn-Tucker point, is introduced.

This test quantity is then used in a modified algorithm, which is shown in theorem 3.14 to be globally convergent in the sense that every accumulation point is a Kuhn-Tucker point.

Introduce the vectors

$$w^T(x,u) = \left(h^T(x),\ \min\left(-g_1(x),\lambda_1\right),\ \ldots,\ \min\left(-g_m(x),\lambda_m\right)\right)$$

and

$$t^T(x,u) = \left(L_x(x,u),w^T(x,u)\right)$$

Note that

$$\min\left(-g_i(x), \lambda_i\right) = 0$$

if and only if

$$g_i(x) \leq 0, \quad \lambda_i \geq 0 \quad \text{and} \quad g_i(x)\lambda_i(x) = 0$$

Consequently $t(x,u) = 0$ if and only if (x,u) satisfies the first order conditions for a Kuhn-Tucker point, i.e. equation (3.1) of theorem 3.1.

Define

$$a(x,u) = \| t(x,u) \|$$

$$b(x,u) = \| w(x,u) \|$$

and let η, r, $\varepsilon^{(k)}$ and K be real numbers satisfying

$$0 < \eta < 1, \quad 0 < r < 1, \quad K > 1, \quad \varepsilon^{(k)} > 0, \quad \varepsilon^{(k)} \to 0 \text{ as } k \to \infty$$

Then the modified algorithm can be described as follows.

Algorithm 3.3

0) Choose $x^{(0)}$, $u^{(0)}$ and $c^{(0)}$; put $a_0 = a\left(x^{(0)}, u^{(0)}\right)$; put $i = 0$, $k = 0$; put b = some large number.

i) Make a line search in a direction obtained by a quasi-Newton method for minimizing $F\left(x, u^{(i)}, c^{(k)}\right)$ with respect to x. Call the resulting value $\tilde{x}$. Update the approximation of the second derivative.

ii) If $\| F_x\left(\tilde{x}, u^{(i)}, c^{(k)}\right) \| \leq \varepsilon^{(k)}$, then go to vi), else go to iii).

iii) Update $u^{(i)}$ and $\tilde{x}$ giving $\hat{u}$ and $\hat{x}$. (It is possible that $\hat{x} = \tilde{x}$.) If $\hat{x} \neq \tilde{x}$ update the approximation of the second

derivative, using a quasi-Newton formula.

iv) If $a(\hat{x},\hat{u}) \leq \eta\, a_0$ then put $a_0 = a(\hat{x},\hat{u})$ and go to v), else go to i).

v) Put $x^{(i+1)} = \hat{x}$, $u^{(i+1)} = \hat{u}$, $i = i+1$. Go to i) or to iii). (The choice of i) or iii) does not affect the global convergence properties.)

vi) Put $x^{(i+1)} = \tilde{x}$, $u^{(i+1)} = u^{(i)}$, $i = i+1$.

vii) If $b(x^{(i)},u^{(i)}) \leq rb_0$ then go to viii) else go to ix).

viii) Put $b_0 = b(x^{(i)},u^{(i)})$, $c^{(k+1)} = c^{(k)}$, $k=k+1$. Go to iii).

ix) Put $c^{(k+1)} = Kc^{(k)}$, put $k = k+1$ and go to i).

It is assumed that the updating of u is done in such a way that $\lambda \geq 0$ all the time.

Assume to begin with that there are no inequality constraints. The connection between algorithm 3.3 and algorithm 3.2, which was studied in the last section, is then as follows.

If step iii) uses the updating formula $\hat{u} = \varphi(x)$ or $\hat{u} = u^{(i)} + ch(\tilde{x})$ and the algorithm goes from step v) to step i), then, as long as the test at point iv) is satisfied and the line search at i) gives $\alpha = 1$, the algorithm is identical to algorithm 3.2 using equations (3.12) or (3.19) respectively.

If instead equation (3.24) is used at step iii) to define $\hat{x}$ and $\hat{u}$ and the algorithm goes from step (v) to step (iii), then the algorithm is identical to algorithm 3.2 using equation (3.24) as long as the test in iv) is satisfied.

Inequality constraints are treated in the following way. Define

$$A(x) = \{i \mid cg_i(x) + \lambda_i > 0\}$$
$$I(x) = \{i \mid cg_i(x) + \lambda_i \leq 0\} \qquad (3.26)$$

$\tilde{g}(x)^T = (g_{i_1}(x), g_{i_2}(x), \ldots)$, where $i_1, i_2, \ldots$ are the indices belonging to $A(x)$. The updating formulas presented in the last section are then used for the multipliers corresponding to the constraints $(h(x)^T, \tilde{g}(x)^T)$. If the updating formula gives $\lambda_i < 0$, for some i, then the value $\lambda_i = 0$ is used. The multipliers λ_i, $i \in I(x)$, are all zeroed.

From Algorithm 3.3 it follows that, if the multiplier updating is unsuccessful, the method results in the minimization of $F(x,u^{(i)},c^{(k)})$ with respect to x. Since quasi-Newton methods can be constructed to have the property that $F_x \to 0$, if the iterates remain bounded, see Polak (1970), the test at point ii) will eventually be satisfied. It is shown by Powell (1969), that if the constraints are satisfied at a point where F(x,u,c) has a minimum with respect to x, then this point is a solution to the problem. This means that, if the test quantity b related to the constraints, has decreased since last time an approximate minimum for F was found, then the algorithm is making progress. If this is not the case, then c is increased. Consequently the penaly function method is used as a last resort to ensure convergence, as suggested by Powell (1969). For the penalty function method, convergence can be proved, using a technique given in Polak (1970). In our case the situation is more complex, since $\mu^{(i)}$ and $\lambda^{(i)}$ must be taken care of. However, it is possible to prove the following theorem.

Theorem 3.14. Let (x_a,u_a) be an accumulation point of the infinite sequence $\{(x^{(i)},u^{(i)})\}$. Assume that $(h_j)_x(x_a)$, $j = 1, \ldots, p$, $(g_j)_x(x_a)$, $j \in A(x_a)$, are linearly independent. Then there

exists a u_b such that (x_a, u_a+u_b) is a Kuhn-Tucker point. If $c^{(k)}$ is bounded then $u_b = 0$.

Proof. Let $\{(x^{(i_j)}, u^{(i_j)})\}$ be an infinite subsequence converging to (x_a, u_a). Assume that the test at point iv) is satisfied infinitely many times for $\{(x^{(i_j)}, u^{(i_j)})\}$. Then $a(x_a, u_a) = 0$ and (x_a, u_a) is a Kuhn-Tucker point. If this is not the case then the inequality

$$\| f_x(x^{(i_j)}) + \left(c^{(k_j)} h(x^{(i_j)}) + \mu^{(i_j)}\right)^T h_x(x^{(i_j)}) + \left(c^{(k_j)} g(x^{(i_j)}) + \lambda^{(i_j)}\right)_+^T g_x(x^{(i_j)}) \| \leq \varepsilon^{(k_j)}$$

is satisfied for all $j \geq j_0$, for some j_0.

There are two possibilities. If $c^{(k)}$ is bounded, then the test $b(x^{(i_j)}, u^{(i_j)}) \leq r b_0$ is satisfied infinitely many times and $b(x_a, u_a) = 0$. This means that $h(x_a) = 0$, $g(x_a) \leq 0$ and $\lambda_a^T g(x_a) = 0$. Since $\varepsilon^{(k)} \to 0$, it also follows that

$$f_x(x_a) + \mu_a^T h_x(x_a) + \lambda_a^T g_x(x_a) = 0$$

If on the other hand $c^{(k)} \to \infty$, then from the inequality above it follows, using the linear independence of $(h_j)_x$, $(g_j)_x$, that $h(x_a) = 0$, $g(x_a) \leq 0$. Furthermore it follows that

$$\left(c^{(k_j)} h(x^{(i_j)}) + \mu^{(i_j)}\right) \to \alpha$$

$$\left(c^{(k_j)} g(x^{(i_j)}) + \lambda^{(i_j)}\right)_+ \to \beta \geq 0 \qquad \text{with}$$

$$f_x(x_a) + \alpha^T h_x(x_a) + \beta^T g_x(x_a) = 0$$

$$\beta^T g(x_a) = 0$$

Corollary 1. If, in addition to the assumptions of the theorem, the sequence $\{(x^{(i)}, u^{(i)})\}$ remains bounded, then there is at least one accumulation point, which is a Kuhn-Tucker point.

Corollary 2. If the assumptions of the theorem are satisfied and if there are no Kuhn-Tucker points except $(\bar{x},\bar{u})$, then, if the sequence $\{(x^{(i)},u^{(i)})\}$ remains bounded, it converges to $(\bar{x},\bar{u}+u_b)$ for some u_b.

Remark 1. Since the updating rule for the multipliers has not been specified, the result applies to any updating formula, not just those that were considered in section 3.5.

Remark 2. The only property of the quasi-Newton method, that is used, is that $F_x \to 0$, when the iterates are bounded. Therefore the theorem remains valid if any other unconstrained method having this property is used.

The result shown in the theorem is somewhat weaker than one would wish. It would be nice to be able to show that the accumulation points are not only Kuhn-Tucker points, but local minima. This is, however, hardly realistic, as seen from a comparison with the unconstrained case, see Polak (1970). There it is in general only possible to show that an algorithm has accumulation points that are stationary points.

3.7. Comparison of Algorithms on Test Problems.

To test the practical usefulness of the ideas behind algorithms 3.2 and 3.3 that were presented in the previous sections, three optimization routines have been programmed and tested. The general structure used in these algorithms is that of algorithm 3.3. The differences between the algorithms lie in the updating of the multipliers. This updating is done according to the three different formulas investigated in section 3.5. The algorithms are described below.

Algorithm MINGRA. Step iii) uses the formula

$$\begin{pmatrix} \mu \\ \tilde{\lambda} \end{pmatrix} = - \left[\begin{pmatrix} h_x \\ \tilde{g}_x \end{pmatrix} \begin{pmatrix} h_x \\ \tilde{g}_x \end{pmatrix}^T \right]^{-1} \begin{pmatrix} h_x \\ \tilde{g}_x \end{pmatrix} f_x^T$$

where $\tilde{g}$ was defined in equation (3.26). The algorithm goes to i) after step v). If the test at step iv) is satisfied and the step length $\alpha^{(k)} = 1$ is used in the quasi-Newton method, then the algorithm is identical to algorithm 3.2 of section 3.4, with u given by (3.12a).

Algorithm MINGRB. This algorithm is identical to MINGRA except that the updating formula for u is

$$\mu^{(i+1)} = \mu^{(i)} + ch(x^{(i)})$$

$$\lambda^{(i+1)} = \left[\lambda^{(i)} + cg(x^{(i)}) \right]_+$$

If the test at step iv) is satisfied and $\alpha^{(k)} = 1$ in the line search, this algorithm is identical to algorithm 3.2 using equation (3.19).

Algorithm MINGRC. Step iii) uses equation (3.24) to update the multipliers, with h replaced by $\begin{bmatrix} h \\ \tilde{g} \end{bmatrix}$ and μ replaced by $\begin{bmatrix} \mu \\ \tilde{\lambda} \end{bmatrix}$. From step v) it moves to step iii). If the test under point iv) is satisfied the algorithm is identical to algorithm 3.2 using equation (3.24).

In both MINGRA and MINGRB the line search uses the method of Fletcher (1972). To take into account different scaling of the constraints the parameter c is replaced by different parameters c_i for each constraint. The initial values of c_i are given by

$$c_i = \frac{4 \max\left(1, |f(x) - f(y)| + |f(y) - f(z)| + |f(z) - f(y)|\right)}{h_i^2(x) + h_i^2(y) + h_i^2(z)}$$

where x, y and z are three points near the starting point. The idea behind the choice is to give all h_i the same weight as the objective function f.

The starting values used for H and u are always $H^{(0)} = I$ and $u^{(0)} = 0$.

For comparison two other constrained minimization algorithms are also tested. The first is VFOIA of the Harwell library, see Fletcher (1973), and the second is the GRG algorithm of Abadie (1970).

For the equality constrained problems, the results for the algorithms tested in Glad (1973) will also be given, to get further comparisons. These algorithms are

OPF The ordinary penalty function method. F(x,0,c) is minimized for a sequence $c^{(k)} \to \infty$. The extrapolation method of Fiacco and McCormick (1968) is used.

HEPO The method used in Hestens (1969) and Powell (1969).

FLE This method minimizes the function $F\left(x, \tilde{\mu}(x), c\right)$ given in section 3.3. To avoid computation of second derivatives of f and h, $\tilde{\mu}_x(x)$ is approximated by the matrix Ω, which is updated using the formula

$$\Omega^{(i+1)} := \Omega^{(i)} + \left(\Delta\tilde{\mu} - \Omega^{(i)}\Delta x\right)\Delta x^T / \Delta x^T \Delta x$$

see Fletcher and Lill (1970), where $\Delta\tilde{\mu}$ and Δx are the differences

$$\Delta\tilde{\mu} = \tilde{\mu}\left(x^{(i+1)}\right) - \tilde{\mu}\left(x^{(i)}\right), \quad \Delta x = x^{(i+1)} - x^{(i)}$$

The initial value of Ω is given by a difference approximation.

In all these methods, the unconstrained minimization is done by the quasi-Newton method in Fletcher (1972) .

The equality constrained problems are the following, which were also used in Glad (1973).

POW See Powell (1969).

Minimize $f(x) = x_1x_2x_3x_4x_5$ under the constraints

$$h_1(x) = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 - 10 = 0$$

$$h_2(x) = x_2x_3 - 5x_4x_5 = 0$$

$$h_3(x) = x_1^3 + x_2^3 + 1 = 0$$

starting point $x = (-2,2,2,-1,-1)$
solution $\bar{x} = (-1.7171,1.5957,1.8272,-0.7636,-0.7636)$

PAV See Himmelblau (1972).

Minimize $f(x) = 1000 - x_1^2 - 2x_2^2 - x_3^2 - x_1x_2 - x_1x_3$ under the constraints

$$h_1(x) = x_1^2 + x_2^2 + x_3^2 - 25 = 0$$

$$h_2(x) = 8x_1 + 14x_2 + 7x_3 - 56 = 0$$

starting point $x = (10,10,10)$
solution $\bar{x} = (3.512,0.217,3.552)$

EXP See Himmelblau (1972).

Minimize

$$f(x) = \sum_{i=1}^{10} \left\{ \exp(x_i) \left(c_i + x_i - \ln \sum_{j=1}^{10} \exp(x_j) \right) \right\}$$

under the constraints

$$h_1(x) = \exp(x_1) + 2\exp(x_2) + 2\exp(x_3) + \exp(x_6) +$$

$$+ \exp(x_{10}) - 2 = 0$$

$$h_2(x) = \exp(x_4) + 2\exp(x_5) + \exp(x_6) + \exp(x_7) - 1 = 0$$

$$h_3(x) = \exp(x_3) + \exp(x_7) + \exp(x_8) + 2\exp(x_9) +$$

$$+ \exp(x_{10}) - 1 = 0$$

where

$c_1 = -6.089$ $c_2 = -17.164$ $c_3 = -34.054$

$c_4 = -5.914$ $c_5 = -24.721$ $c_6 = -14.986$

$c_7 = -24.100$ $c_8 = -10.708$ $c_9 = -26.662$

$c_{10} = -22.179$

starting point $x_i = -2.3 \quad i = 1,\ldots,10$

solution $\bar{x} = (-3.2,-1.9,-0,24,-\infty,-0.72,-\infty,-3.6,$
$-4.0,-3.3,-2.3)$

COL1 See Fletcher and Lill (1970) and Colville (1968)

Minimize

$$f(x) = \sum_{j=1}^{5} e_j x_j + \sum_{j=1}^{5} \sum_{i=1}^{5} c_{ij} x_i x_j + \sum_{j=1}^{5} d_j x_j^3$$

under the constraints

$$h_1(x) = -3.5x_1 + 2x_3 + 0.25 = 0$$

$$h_2(x) = -9x_2 - 2x_3 + x_4 - 2.8x_5 + 4 = 0$$

$$h_3(x) = 2x_1 - 4x_3 + 1 = 0$$

$$h_4(x) = x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 - 5 = 0$$

where the constants are given by

e_j	-15	-27	-36	-18	-12
c_{ij}	30	-20	-10	32	-10
	-20	39	- 6	-31	32
	-10	- 6	10	- 6	-10
	32	-31	- 6	39	-20
	-10	32	-10	-20	30
d_j	4	8	10	6	2

starting point $x = (0,0,0,0,1)$
solution $\bar{x} = (0.3000,0.3335,0.4000,0.4283,0.2240)$

TRIG Minimize

$$f(x) = \sum_{i=1}^{n} \left(b_i E_i - f_i(x)\right)^2$$

under the constraints

$$h_i(x) = E_i - f_i(x) = 0 \qquad i = 1,\ldots,m$$

where $b_i = 1$, $i = m+1,\ldots,n$, $b_i \neq 1$, $i = 1,\ldots,m$

$$f_i(x) = \sum_{j=1}^{n} \left(A_{ij} \sin(x_j) + B_{ij} \cos(x_j)\right)$$

$E_i = f_i(\bar{x})$ where $\bar{x}$ is the point chosen to be the minimum. A_{ij}, B_{ij}, E_i, b_i and $\bar{x}$ are given in Glad (1975).

The inequality constrained problems are

ROS See Rosen and Suzuki (1965).

Minimize $f(x) = x_1^2 + x_2^2 + 2x_3^2 + x_4^2 - 5x_1 - 5x_2 - 21x_3 + 7x_4$ under the constraints

$$g_1(x) = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_1 - x_2 + x_3 - x_4 - 8 \leq 0$$

$$g_2(x) = x_1^2 + 2x_2^2 + x_3^2 + 2x_4^2 - x_1 - x_4 - 10 \leq 0$$

$$g_3(x) = 2x_1^2 + x_2^2 + x_3^2 + 2x_1 - x_2 - x_4 - 5 \leq 0$$

starting point $x = (0,0,0,0)$ (ROSa)
$x = (3,3,3,3)$ (ROSb)
solution $\bar{x} = (0,1,2,-1)$

PROG See Colville (1968).

Minimize $f(x) = 5.3578547x_3^2 + 0.8356891x_1x_5 + 37.293239x_1 - 40792.141$ under the constraints

$$0 \leq 85.334407 + 0.0056858x_2x_5 + 0.0006262x_1x_4 - 0.0022053x_3x_5 \leq 92$$

$$90 \leq 80.51249 + 0.0071317x_2x_5 + 0.0029955x_1x_2 + 0.0021813x_3^2 \leq 110$$

$$20 \leq 9.300961 + 0.0047026x_3x_5 + 0.0012547x_1x_3 + 0.0019085x_3x_4 \leq 25$$

$$78 \leq x_1 \leq 102, \qquad 33 \leq x_2 \leq 45, \qquad 27 \leq x_3 \leq 45,$$

$$27 \leq x_4 \leq 45, \qquad 27 \leq x_5 \leq 45$$

starting point $x = (78,33,27,27,27)$
solution $\bar{x} = (78.000,33.000,29.995,45.000,36.776)$

The results are given in the following table. The entries are the number of times that (f, f_x, h, h_x, g, g_x) has been evaluated. For the GRG method a separate table is presented because for this method f, f_x, h, h_x, g and g_x are not evaluated the same number of times.

The first column gives the absolute accuracy in x, that was required.

	acc	OPF	HEPO	FLE	MINGRA	MINGRB	MINGRC	VFOIA
POW	10^{-4}	41	37	42	37	41	18	39
PAV	10^{-3}	64	54	58	70	-	35	-
EXP	10^{-1}	>150	>150	76	95	116	>150	140
COL1	10^{-4}	45	54	18	22	20	15	26
TRIG2	10^{-3}	22	28	13	17	15	11	14
TRIG4	10^{-3}	130	92	53	34	46	21	139
TRIG6	10^{-3}	62	57	35	32	27	25	63
TRIG8	10^{-2}	>150	>150	>150	50	68	148	>150
ROSa	10^{-3}	-	-	-	23	25	26	-
ROSb	10^{-3}	-	-	-	102	106	20	-
PROG	10^{-3}	-	-	-	41	95	67	94

For TRIG2, TRIG4, TRIG6 and TRIG8 the number of variables are 2, 4, 6 and 8 and the number of equality constraints, 1, 2, 3 and 4 respectively. The sign - means that the algorithm was not tested on this problem. The computation was stopped after 150 evaluations and if the desired accuracy was not reached then, the entry is marked ">150".

Results for GRG, number of evaluations to reach the accuracy given above:

	f	f_x	h,g	h_x,g_x
POW	82	33	101	15
PAV	114	26	192	11
EXP	192	73	611	27
COL1	26	17	22	7
TRIG2	24	8	29	6
TRIG4	121	29	230	14
TRIG6	201	59	564	33
TRIG8	235	75	902	29
PROG	67	29	131	11

The comparison between GRG and the other algorithms depends on the computational effort needed to evaluate f, f_x, h, g, h_x and g_x respectively. The other algorithms can be compared directly. The algorithms MINGRA, MINGRB and MINGRC seem to be promising although there are occasional bad results for some problems. Note that, although MINGRC is theoretically superior, having superlinear convergence while MINGRA and MINGRB have only linear, it is not obvious that it is a better algorithm in practice.

3.8. Summary.

This chapter deals with methods of solving the constrained minimization problem. In section 3.1 the Lagrangian $L(x,u)$ is introduced. Its important property is that there is a value, $u = \bar{u}$, of the multipliers such that $L(x,\bar{u})$ has a stationary point at $x = \bar{x}$, where $\bar{x}$ is the solution to the constrained minimization problem. The augmented Lagrangian, $F(x,u,c)$ is then presented. $F(x,\bar{u},c)$ has a local minimum at $x = \bar{x}$ if c is large enough. The results in 3.1 are well known from the recent literature in

optimization and form the basis for the other sections of chapter 3.

Section 3.2 gives additional insight into the behaviour of the augmented Lagrangian by showing that, for a wide class of problems, there is actually a global minimum at $x = \bar{x}$, if the parameter c is chosen large enough.

The fact that $F(x,\bar{u},c)$ has a local or a global minimum at $x = \bar{x}$ is interesting from a computational point of view. It means that the constrained problem has been converted into an unconstrained one and there are very powerful methods for solving unconstrained optimization problems. This fact cannot be used directly, however, because $\bar{u}$ is in general not known. There are two ways of overcoming this difficulty. The first is to replace the parameter u by a function $\tilde{u}(x)$ having the property $\tilde{u}(\bar{x}) = \bar{u}$. The second one is to use an iterative method which updates u in such a way that it converges to $\bar{u}$. Section 3.3 deals with the function $\tilde{u}(x)$ in the equality constrained case. The interesting theoretical result shown there is that $F(x,\tilde{u}(x),c)$ has a local minimum at $\bar{x}$ under weaker conditions than those previously given . Since there are practical drawbacks to this method as pointed out in 3.3, the alternative of updating u iteratively is considered in the following sections.

Section 3.4 presents some different iterative methods suggested in the literature. The algorithm that is chosen for further study uses a quasi-Newton method known as the BFGS method to minimize the augmented Lagrangian with respect to x. After each line search the multipliers are updated. In section 3.5 the local convergence of this algorithm is studied. Three different updating methods for the multipliers are shown to be convergent locally. Two of them converge linearly while the third one shows superlinear convergence. It is also shown that the asymptotic behaviour of the approximation of the second derivative is the same as in the unconstrained case.

In 3.6 modifications are introduced to avoid divergence of the algorithm when the starting point is far from the solution. It is shown that the modifications guarantee global convergence in the sense that every accumulation point satisfies the first order necessary conditions for a constrained minimum.

The algorithm of section 3.6 has been tested on numerical problems, using the three different updating formulas of 3.5. The results are given in section 3.7 where there is also a comparison with other methods. The practical performance of the algorithms investigated in 3.5 and 3.6 seems to be quite promising.

The computationally relevant conclusion for the class of methods treated in 3.5 and 3.6 can then be summarized.

- They have sound theoretical properties:

  a) Fast local convergence.
  b) Global convergence.

- They have solved a number of test problems satisfactorily.

Consequently, these methods appear to represent a useful tool for solving the constrained optimization problem.

3.9. References.

Abadie, J. (1970):
Numerical Experiments with the GRG Method, In Abadie, J., ed.: Integer and Nonlinear Programming, North Holland Publishing Co., Amsterdam.

Bertsekas, D.P. (1973):
On Penalty and Multiplier Functions for Constrained Minimization. EES Department Working Paper, Stanford University, California.

Bertsekas, D.P. (1975):
Combined Primal-Dual and Penalty Methods for Constrained Minimization, SIAM J. Control, Vol. 13, No. 3, pp. 521 - 544.

Broyden, C.G. (1972):
Quasi-Newton Methods, In Murray, W., ed.: Numerical Methods for Unconstrained Optimization, Academic Press, London and New York.

Colville, A.R. (1968):
A Comparative Study of Nonlinear Programming Codes, IBM New York Scientific Center, Report 320-2949.

Dennis, J.E., and Moré, J.J. (1974):
A Characterization of Superlinear Convergence and its Applications to Quasi-Newton Methods, Math. Comp., Vol. 28, No. 126, pp. 549 - 560.

Fiacco, A.V., and Mc Cormick, G.P. (1968):
Nonlinear Programming: Sequential Unconstrained Minimization Techniques, John Wiley and Sons, New York.

Fletcher, R. (1970):
Methods for Nonlinear Programming, In Abadie, ed.: Integer and Nonlinear Programming, North Holland Publishing company.

Fletcher, R., and Lill, S.A. (1970):
A Class of Methods for Nonlinear Programming II. Computational Experience, In Rosen, J.B., Mangasarian, O.L., and Ritter, K., ed.: Nonlinear Programming, Academic Press, London.

Fletcher, R. (1972):
Fortran Subroutines for Minimization by Quasi-Newton Methods, Report AERE-R7125, Atomic Energy Research Establishment, Harwell, England.

Fletcher, R. (1973):
An Exact Penalty Function for Nonlinear Programming with Inqualities, Math. Programming 5, pp. 129-150

Fletcher, R. (1974):
An Ideal Penalty Function for Constrained Optimization, in Mangasarian, O.L., Meyer, R.R. and Robinson, S.M. eds., Nonlinear Programming 2, Academic Press, London.

Glad, T. (1973):
Lagrange Multiplier Methods for Minimization under Equality Constraints, Report 7323, Lund Institute of Technology, Division of Automatic Control.

Glad, T. (1975):
Algorithms for Nonlinear Minimization with Equality and Inequality Constraints based on Lagrange Multipliers, Report 7503, Lund Institute of Technology, Department of Automatic Control.

Hestenes, M.R. (1969):
Multiplier and Gradient Methods, J. Opt. Theory Appl., Vol. 4, No. 5, pp. 303 - 320.

Himmelblau, D.M. (1972):
Applied Nonlinear Programming, McGraw Hill, New York.

Luenberger, D.G. (1968):
Optimization by Vector Space Methods, John Wiley and Sons, New York.

Luenberger, D.G. (1973):
Introduction to Linear and Nonlinear Programming, Addison-Wesley.

Miele, A., Cragg, E.E., Iyer, R.R., and Levy, A.V. (1971):
Use of the Augmented Penalty Function in Mathematical Programming Problems, J. Opt. Theory Appl., Vol. 8, No. 2, pp. 115 - - 130, pp. 131 - 153.

Mårtensson, K. (1972):
New Approaches to the Numerical Solution of Optimal Control Problems, Report 7206, Lund Institute of Technology, Division of Automatic Control.

O'Doherty, R.J., and Pierson, B.L. (1974):
A Numerical Study of Multiplier Methods for Constrained Parameter Optimization, Int. J. Systems Sci., Vol. 5, No. 2, pp. 187 - 200.

Ortega, J.M., and Rheinboldt, W.C. (1970):
Iterative Solution of Nonlinear Equations in Several Variables, Academic Press.

Polak, E. (1971):
Computational Methods in Optimization, Academic Press, New York and London.

Powell, M.J.D. (1969):
A Method for Nonlinear Constraints in Minimization Problems, In Fletcher, R., ed.: Optimization, Academic Press, London.

Rockafellar, R.T. (1974):
Augmented Lagrange Multiplier Functions and Duality in Nonconvex Programming, SIAM J. Control, Vol. 12, No. 2, pp. 268 - 285.

Rosen, J.B., and Suzuki, S. (1965):
Construction of Nonlinear Programming Test Problems, Comm. ACM, Vol. 8, p. 113.

Tripathi, S.S., and Narendra, K.S. (1972):
Constrained Optimization Problems Using Multiplier Methods, J. Opt. Theory Appl., Vol. 9, No. 1, pp. 59 - 70.

4. OPTIMAL CONTROL PROBLEMS

As was shown in Section 2, the optimal control problem is infinite dimensional. Three types of constraints are considered, the differential equation, terminal constraints and mixed state-control inequality constraints. The goal is to treat these constraints with methods similar to those used in the finite dimensional case. The differential equation constraint has been treated by Hestenes (1947), (1969), Rupp (1972a) and di Pillo et. al. (1974) using an augmented functional which is a direct generalization of the finite dimensional augmented Lagrangian. A similar technique has also been used by Rupp (1972b) for isoperimetric constraints and by Nahra (1971), Mårtensson (1972) and O'Doherty and Pierson (1974) for terminal constraints. In 4.2 and 4.3 the properties of the augmented functional are examined, using a Riccati equation approach. Iterative methods based on the augmented functional are discussed in Section 4.4. Finally an extension to mixed state control inequality constraints is considered in 4.5.

4.1. Problem Formulation.

The optimal control problem to be studied in 4.2 - 4.4 can be formulated:

Minimize

$$I(x,u) = \int_0^T L\big(x(t),u(t),t\big)dt + F\big(x(T)\big) \tag{4.1}$$

subject to

$$\dot{x}(t) = f\big(x(t),u(t),t\big) \qquad 0 \le t \le T$$
$$x(0) = a$$
$$\psi\big(x(T)\big) = 0$$

Here x and u are functions belonging to $C_1^n[0,T]$ and $C_0^m[0,T]$ respectively. The functions f and ψ are vector valued with n and r components respectively.

The following assumptions are made

- L and f are three times continuously differentiable with respect to x and u.
- L and f together with their first and second derivatives with respect to x and u, are continuous with respect to t.
- F and ψ are three times continuously differentiable.
- The minimization problem has a solution denoted $(\bar{x}(t),\bar{u}(t))$.

Define the Hamiltonian

$$H(x(t),u(t),p(t),t) =$$

$$= L(x(t),u(t),t) + p^T(t)f(x(t),u(t),t) \qquad (4.2)$$

where p is a continuous function of time.

The standard necessary conditions for the optimal control problem are given by the following theorem.

Theorem 4.1. Let $\bar{x},\bar{u}$ be the solution to the problem defined by (4.1) and assume that the following regularity conditions are satisfied.

(i) The matrix $\psi_x(\bar{x}(T))$ has rank r.

(ii) Given any vector z, it is possible to find a continuous function v such that

$$\dot{h}(t) = f_x(\bar{x}(t),\bar{u}(t),t)h(t) + f_u(\bar{x}(t),\bar{u}(t),t)v(t)$$

$$h(0) = 0$$

has a solution satisfying $h(T) = z$, i.e. the linearized system is controllable.

Then there is an n-dimensional vector valued function $\bar{p}(t)$ and an r-dimensional vector $\bar{b}$ such that for all $t \in [0,T]$.

$$
\begin{aligned}
-\dot{\bar{p}}(t) &= H_x\left(\bar{x}(t),\bar{u}(t),\bar{p}(t),t\right) \\
\bar{p}(T) &= F_x^T\left(\bar{x}(T)\right) + \psi_x^T\left(\bar{x}(T)\right)\bar{b} \\
H_u\left(\bar{x}(t),\bar{u}(t),\bar{p}(t),t\right) &= 0
\end{aligned}
\tag{4.3}
$$

Proof. See Luenberger (1968). □

4.2. The Optimal Control Problem with a Differential Equation Constraint

In this case the problem can be written:

Minimize

$$I(x,u) = \int_0^T L\left(x(t),u(t),t\right)dt + F\left(x(T)\right)$$

subject to

$$
\begin{aligned}
\dot{x}(t) &= f\left(x(t),u(t),t\right) \\
x(0) &= a
\end{aligned}
\tag{4.4}
$$

The idea in Hestenes (1947) and Hestenes (1969) is to form the augmented function

$$
\begin{aligned}
J(x,u,p,c) = \int_0^T \Big\{ & L\left(x(t),u(t),t\right) + p^T(t)\left(f\left(x(t),u(t),t\right) - \dot{x}(t)\right) + \\
& + \frac{c}{2}\left(f\left(x(t),u(t),t\right) - \dot{x}(t)\right)^T \\
& \cdot \left(f\left(x(t),u(t),t\right) - \dot{x}(t)\right)\Big\}dt + F\left(x(T)\right)
\end{aligned}
\tag{4.5}
$$

Here c is a positive real number and p is a continuous function. The functions x and u are now allowed to take arbitrary values, not necessarily satisfying the differential equation $\dot{x} = f(x,u,t)$. The condition $x(0) = a$, however, is still applied. Note that J has a form which is analogous to the augmented Lagrangian of Chapter 3.

It has been proved in Hestenes (1947) that $J(x,u,\bar{p},c)$ has a local minimum at $(\bar{x},\bar{u})$ if c is large enough. Here a different proof, based on the Riccati equation, will be given. It has the advantage of showing what the lower bound of c is. The connection with the sufficiency conditions in Bryson and Ho (1969) is also given.

To show that $J(x,u,\bar{p},c)$ has a local minimum at $(\bar{x},\bar{u})$ an expansion is used.

$$x(t) = \bar{x}(t) + h(t)$$
$$u(t) = \bar{u}(t) + k(t)$$

Since the function x belongs to $C_1^n[0,T]$, with $x(0) = a$, and u belongs to $C_0^m[0,T]$, only variations h and k satisfying $h(0) = 0$, h belonging to $C_1^n[0,T]$ and k belonging to $C_0^m[0,T]$ are of interest. The functions h and k satisfying these conditions will be called admissible. The following norms are used

$$\| h \|_1 = \sup_{0\leq t\leq T} \| h(t) \| + \sup_{0\leq t\leq T} \| \dot{h}(t) \|$$

$$\| k \|_0 = \sup_{0\leq t\leq T} \| k(t) \|$$

where $\| \cdot \|$ denotes an arbitrary vector norm. In what follows, x and u will often be written for x(t) and u(t), to simplify the notation.

The expansion can now be written

$$J(\bar{x}+h,\bar{u}+k,\bar{p},c) = \int_0^T \Big\{ H(\bar{x}+h,\bar{u}+k,\bar{p},t) - p^T(\dot{\bar{x}}+\dot{h}) +$$

$$+ \frac{c}{2}\big(f(\bar{x}+h,\bar{u}+k,t) - \dot{\bar{x}} - \dot{h}\big)^T\big(f(\bar{x}+h,\bar{u}+k,t) -$$

$$- \dot{\bar{x}} - \dot{h}\big)\Big\}dt + F\big(\bar{x}(T) + h(T)\big)$$

where the Hamiltonian $H = L + p^T f$ is used. Expanding H, f and F in a Taylor series give

$$J(\bar{x}+h,\bar{u}+k,\bar{p},c) = J(\bar{x},\bar{u},\bar{p},c) + \int_0^T \{H_x h + H_u k - \bar{p}^T\dot{h}\}dt + F_x h(T) +$$

$$+ \frac{1}{2}\int_0^T \{h^T H_{xx} h + 2h^T H_{xu} k + k^T H_{uu} k\}dt +$$

$$+ \frac{1}{2}\int_0^T c\{h^T f_x^T f_x h + k^T f_u^T f_u k + \dot{h}^T\dot{h} +$$

$$+ 2h^T f_x^T f_u k - 2h^T f_x^T\dot{h} - 2k^T f_u^T\dot{h}\}dt +$$

$$+ \frac{1}{2} h^T(T) F_{xx} h(T) + R(h,k)$$

where H_x, H_u, H_{xx} etc. are evaluated along $(\bar{x},\bar{u})$ and

$$|R(h,k)| \leq \varepsilon(h,k)\int_0^T (h^T h + \dot{h}^T\dot{h} + k^T k)dt$$

and $\varepsilon(h,k) \to 0$ as $(h,k) \to 0$ in the norms given above.

Since $\bar{p}$ satisfies the conditions of Theorem 4.1, it follows from a partial integration that the linear terms disappear. Then

$$J(\bar{x}+h,\bar{u}+k,\bar{p},c) - J(\bar{x},\bar{u},\bar{p},c) =$$

$$= \frac{1}{2}\int_0^T \Big\{ h^T(H_{xx}+cf_x^Tf_x)h + 2h^T(H_{xu}+cf_x^Tf_u)k +$$

$$+ k^T(H_{uu}+cf_u^Tf_u)k + c\dot{h}^T\dot{h} - 2ch^Tf_x^T\dot{h} -$$

$$- 2ck^Tf_u^T\dot{h}\Big\}dt + h^T(T)F_{xx}h(T) + R(h,k) =$$

$$= \delta^2 J(h,k) + R(h,k)$$

To prove that $\delta^2 J$ is positive, it is transformed into a perfect square. First observe that for any continuously differentiable matrix function S, it is true that

$$\int_0^T \{h^T\dot{S}h + 2h^TS\dot{h}\}dt - h^T(T)S(T)h(T) = 0$$

for all continuously differentiable h satisfying $h(0) = 0$. The addition of a term of this form to get a perfect square is used in the calculus of variations, see Gelfand and Fomin (1963). Adding this quantity to $\delta^2 J$ gives

$$\delta^2 J = \frac{1}{2}\int_0^T \Big\{ h^T(H_{xx}+cf_x^Tf_x+\dot{S})h + 2h^T(H_{xu}+cf_x^Tf_u)k +$$

$$+ k^T(H_{uu}+cf_u^Tf_u)k + c\dot{h}^T\dot{h} + 2h^T(S-cf_x^T)\dot{h} - 2ck^Tf_u^T\dot{h}\Big\}dt +$$

$$+ \frac{1}{2}h^T(T)[F_{xx} - S(T)]h(T) =$$

$$= \frac{1}{2}\int_0^T \begin{Vmatrix} k + H_{uu}^{-1}(H_{ux}+f_u^TS)h \\ \dot{h} + [f_uH_{uu}^{-1}(H_{ux}+f_u^TS) + \frac{1}{c}S - f_x]h \end{Vmatrix}^T \cdot$$

$$\cdot \begin{pmatrix} H_{uu} + cf_u^T f_u & -cf_u^T \\ -cf_u & cI \end{pmatrix} \cdot$$

$$\cdot \begin{pmatrix} k + H_{uu}^{-1}(H_{ux}+f_u^T S)h \\ \dot{h} + [f_u H_{uu}^{-1}(H_{ux}+f_u^T S) + \frac{1}{c} S - f_x]h \end{pmatrix} dt +$$

$$+ \frac{1}{2} \int_0^T h^T [\dot{S} + H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} + (f_x - f_u H_{uu}^{-1}H_{ux})^T S +$$

$$+ S(f_x - f_u H_{uu}^{-1}H_{ux}) - S f_u H_{uu}^{-1} f_u^T S - \frac{1}{c} S^2]dt +$$

$$+ \frac{1}{2} h^T(T)[F_{xx} - S(T)]h(T) \qquad (4.6)$$

Here it is assumed that H_{uu} is nonsingular. Now the following theorem is an immediate consequence.

Theorem 4.2. Let $(\bar{x},\bar{u})$ be a solution to (4.1) and let $\bar{p}$ satisfy equation (4.3) of Theorem 4.1. Also assume that

$$c > 0$$
$$H_{uu}(\bar{x}(t),\bar{u}(t),\bar{p}(t),t) > 0 \qquad t \in [0,T]$$

and that the Riccati equation

$$-\dot{S} = H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} + (f_x - f_u H_{uu}^{-1}H_{ux})^T S +$$

$$+ S(f_x - f_u H_{uu}^{-1}H_{ux}) - S(f_u H_{uu}^{-1} f_u^T + \frac{1}{c} I)S \qquad (4.7)$$

$$S(T) = F_{xx}$$

where H_{xx} etc. are evaluated along $\bar{x}$, $\bar{u}$, has a solution over

the whole interval $[0,T]$. Then $\delta^2 J(h,k) > 0$ for all admissible h and k that are not both identically zero.

Proof. First it is shown that the matrix

$$\begin{pmatrix} H_{uu} + cf_u^T f_u & -cf_u^T \\ -cf_u & cI \end{pmatrix}$$

is positive definite for all $c > 0$. Form

$$\begin{bmatrix} z^T & w^T \end{bmatrix} \begin{pmatrix} H_{uu} + cf_u^T f_u & -cf_u^T \\ -cf_u & cI \end{pmatrix} \begin{pmatrix} z \\ w \end{pmatrix} =$$

$$= z^T H_{uu} z + c(f_u z - w)^T (f_u z - w) \geq 0$$

Equality is attained only for $z = 0$, $w = 0$.

Since the second and third terms of $\delta^2 J$ in (4.6) disappear, it follows that $\delta^2 J \geq 0$. If $\delta^2 J = 0$ then

$$k + H_{uu}^{-1}(H_{ux} + f_u^T S)h = 0 \qquad 0 \leq t \leq T$$

$$\dot{h} + \left(f_u H_{uu}^{-1}(H_{ux} + f_u^T S) + \frac{1}{c} S - f_x\right)h = 0 \qquad 0 \leq t \leq T$$

Since $h(0) = 0$, it follows from the uniqueness theorem for linear differential equations that $h(t)$ is identically zero. Then k is also identically zero. □

To show that J has a minimum at $(\bar{x},\bar{u})$, it is not enough to know that $\delta^2 J$ is positive. It must also dominate the higher order terms. The result of Theorem 4.2 can, however, be strengthened.

Theorem 4.3. With the same assumptions as in Theorem 4.2 there exists a constant $\eta > 0$ such that

$$\delta^2 J(h,k) \geq \eta \int_0^T (h^T h + \dot{h}^T \dot{h} + k^T k)\, dt$$

Proof. Study

$$A(h,k,\eta) = \delta^2 J(h,k) - \eta \int_0^T (h^T h + \dot{h}^T \dot{h} + k^T k)\, dt$$

where $\eta > 0$. The value of $A(h,k,\eta)$ is the same as the value of $\delta^2 J(h,k)$ with $h^T(H_{xx}+cf_x^T f_x)h$ replaced by $h^T(H_{xx}+cf_x^T f_x-\eta I)h$, $c\dot{h}^T\dot{h}$ replaced by $(c-\eta)\dot{h}^T\dot{h}$ and $k^T(H_{uu}+cf_u^T f_u)k$ replaced by $k^T(H_{uu}+cf_u^T f_u-\eta I)k$. It then follows from Lemma A.5 in the appendix that, if η is chosen sufficiently small, then the Riccati equation corresponding to $A(h,k,\eta)$ exists over the interval $[0,T]$. Since the matrix

$$\begin{pmatrix} H_{uu} + cf_u^T f_u - \eta I & -cf_u^T \\ -cf_u & (c-\eta)I \end{pmatrix}$$

is still positive definite for sufficiently small η, it follows that $A(h,k,\eta) \geq 0$ and the theorem is proved. □

This leads directly to the following result, showing that J has a local minimum at $(\bar{x},\bar{u})$.

Theorem 4.4. If the assumptions of Theorem 4.2 are satisfied, then $J(\bar{x}+h,\bar{u}+k,\bar{p},c) > J(\bar{x},\bar{u},\bar{p},c)$ for all admissible h and k, not both identically zero, and with $\| h \|_1$ and $\| k \|_0$ sufficiently small.

Proof. From Theorem 4.3

$$J(\bar{x}+h,\bar{u}+k,\bar{p},c) - J(\bar{x},\bar{u},\bar{p},c) = \delta^2 J(h,k) + R(h,k) \geq$$

$$\geq \left(\eta - |\varepsilon(h,k)|\right)\int_0^T (h^T h+\dot{h}^T\dot{h}+k^T k)\,dt > 0$$

if $\|h\|_1$ and $\|k\|_0$ are sufficiently small and h and k are not both identically zero. □

It is interesting to note that the magnitude of c that is required only depends on the Riccati equation (4.7) (provided $c > 0$). The interesting question is of course: is there any c for which (4.7) has a solution over $[0,T]$? First note the following result.

Theorem 4.5. If (4.7) has a solution on $[0,T]$ for $c = c_1$, then it has a solution for any $c \geq c_1$.

Proof. Let $c_2 > c_1$ and define

$$P_1 = f_u H_{uu}^{-1} f_u^T + \frac{1}{c_1} I$$

$$P_2 = f_u H_{uu}^{-1} f_u^T + \frac{1}{c_2} I$$

Then $P_1 - P_2 \geq 0$. It now follows from Lemma A.2 in the appendix that $S_2(t) \geq S_1(t)$, where S_1 and S_2 are the solutions corresponding to c_1 and c_2 respectively. Since from Lemma A.4, the only way the solution S can fail to exist on an interval $[t_1,T]$, is by going off to minus infinity, it follows that S_2 exists on any interval where S_1 exists. □

Corollary. Either there are no values of c for which (4.7) has

a solution on $[0,T]$ or else there is a number c_0 such that S exists on $[0,T]$ for $c > c_0$ and goes to minus infinity for some $t_1 \in [0,T]$ when $c < c_0$.

Proof. Take $c_0 = \inf\{$all $c > 0$ such that S exists on the whole interval$\}$. □

Theorem 4.6. Let the Riccati equation

$$-\dot{S} = H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} + (f_x - f_uH_{uu}^{-1}H_{ux})^TS +$$

$$+ S(f_x - f_uH_{uu}^{-1}H_{ux}) - Sf_uH_{uu}^{-1}f_u^TS$$

$$S(T) = F_{xx} \tag{4.8}$$

have a solution defined in the whole interval $[0,T]$. Then there exists a $c_0 \geq 0$ such that (4.7) also has a solution over $[0,T]$ for all $c > c_0$.

Proof. Since the difference between the matrices $f_uH_{uu}^{-1}f_u^T$ and $(f_uH_{uu}^{-1}f_u^T + \frac{1}{c}\,I)$ can be made arbitrarily small by choosing c large, the result follows from Lemma A.5 in the appendix. □

An immediate consequence is

Theorem 4.7. Let $(\bar{x},\bar{u})$ be the solution to (4.1) and let $\bar{p}$ satisfy equations (4.3) of Theorem 4.1. Also assume that

$$H_{uu}(\bar{x}(t),\bar{u}(t),\bar{p}(t),t) > 0 \qquad t \in [0,T]$$

and that the Riccati equation (4.8) has a solution over $[0,T]$. Then there exists a constant $c_0 \geq 0$ such that $J(x,u,\bar{p},c)$ has a

local minimum at $(\bar{x},\bar{u})$ for all $c > c_0$.

Proof. Follows directly from Theorems 4.6 and 4.4. □

The assumptions made in this theorem are the standard second order sufficiency conditions of problem (4.1), see e.g. Bryson and Ho (1969). If J has a minimum with respect to arbitrary (x,u), then it has also a minimum with respect to the special choice of (x,u) which satisfies the differential equation $\dot{x} = f(x,u,t)$. Since J = I for these (x,u), Theorem 4.7 actually forms an alternative proof of the sufficiency conditions.

So far, it has been shown that, when $H_{uu} > 0$, the existence of a solution to (4.7) over $[0,T]$ is a sufficient condition for J to have a local minimum at $(\bar{x},\bar{u})$. The condition is almost necessary in the sense explained in the following theorem.

Theorem 4.8. Let $\bar{p}$ satisfy (4.3) and assume that $J(x,u,\bar{p},c)$ has a local minimum at $(\bar{x},\bar{u})$ for some $c > 0$. Assume that $H_{uu}(\bar{x}(t), \bar{u}(t),\bar{p}(t),t) > 0$, $t \in [0,T]$. Then the Riccati equations (4.7) and (4.8) have a solution over $[\varepsilon,T]$ for all $\varepsilon > 0$.

Proof. For J to have a local minimum it is necessary that $\delta^2 J(h,k) \geq 0$ for all admissible h and k. Since the solution of (4.7) exists on $[t_1,T]$ for some $t_1 < T$ (local existence theorem for differential equations, see Coddington and Levinson (1955)), it follows that

$$\delta^2 J(h,k) = \frac{1}{2}\int_0^{t_1} \{h^T(H_{xx}+cf_x^T f_x)h + 2h^T(H_{xu}+cf_x^T f_u)k +$$

$$+ k^T(H_{uu}+cf_u^T f_u)k + c\dot{h}^T\dot{h} -$$

$$- 2ch^T f_x^T\dot{h} - 2ck^T f_u^T\dot{h}\}dt +$$

$$+ \frac{1}{2} \int_{t_1}^{T} \begin{pmatrix} k + H_{uu}^{-1}(h_{ux}+f_u^T S)h \\ \dot{h} + \left[f_u H_{uu}^{-1}(H_{ux}+f_u^T S) + \frac{1}{c} S - f_x\right]h \end{pmatrix}^T \cdot$$

$$\cdot \begin{pmatrix} H_{uu} + cf_u^T f_u & -cf_u^T \\ -cf_u & cI \end{pmatrix} \cdot$$

$$\cdot \begin{pmatrix} k + H_{uu}^{-1}(H_{ux}+f_u^T S)h \\ \dot{h} + \left[f_u H_{uu}^{-1}(H_{ux}+f_u^T S) + \frac{1}{c} S - f_x\right]h \end{pmatrix} dt +$$

$$+ \frac{1}{2} h^T(t_1) S(t_1) h(t_1)$$

Now choose

$$k(t) = 0 \qquad t \in [0,t_1]$$

$$h(t) = \frac{t}{t_1} a \qquad t \in [0,t_1]$$

where a is an arbitrary constant vector, and let h and k be the solutions of

$$k = - H_{uu}^{-1}(H_{ux}+f_u^T S)h$$

$$\dot{h} = - \left[f_u H_{uu}^{-1}(H_{ux}+f_u^T S) + \frac{1}{c} S - f_x\right]h$$

$$h(t_1) = a$$

in $[t_1,T]$. For this choice of h and k

$$\delta^2 J(h,k) = \frac{1}{2} a^T \int_0^{t_1} \left\{\frac{t^2}{t_1^2}(H_{xx}+cf_x^T f_x) + \frac{c}{t_1^2} I - c\frac{t}{t_1^2}(f_x+f_x^T)\right\} dt\, a +$$

$$+ \frac{1}{2} a^T S(t_1) a$$

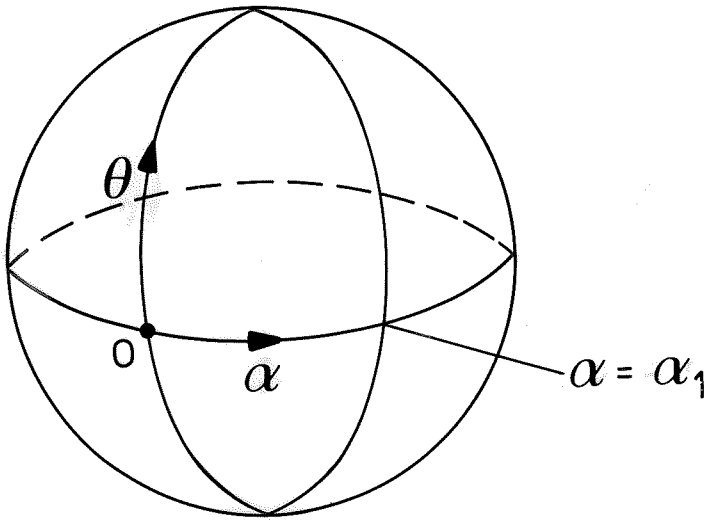
Since the h and k used here can be approximated arbitrarily well with continuous k and continuously differentiable h, it follows that $\delta^2 J(h,k) \geq 0$ also for this choice of h and k. Then

$$a^T S(t_1) a \geq - a^T \int_0^{t_1} \left\{ \frac{t^2}{t_1^2} (H_{xx} + c f_x^T f_x) + \frac{c}{t_1^2} I - \frac{ct}{t_1^2} (f_x + f_x^T) \right\} dt \; a \qquad (4.9)$$

for any vector a. Now suppose that S goes to minus infinity for $t = t_2$, $0 < t_2 < T$. Then (4.9) must be violated for some $t_1 \in [t_2, T]$. Consequently the solution to (4.7) exists on $[\varepsilon, T]$ for any ε. From Theorem 4.5 this is true also for the solution to (4.8). □

Corollary. If the solution to (4.7) goes to minus infinity for some $t \in (0,T)$, then $J(x,u,\bar{p},c)$ does not have a local minimum at $(\bar{x},\bar{u})$.

Example 4.1. Shortest distance between a point and a great circle on a unit sphere.



Let the given point be at the origin 0 of a latitude-longitude coordinate system and let the great circle be the meridian $\alpha = \alpha_1$.

Then $ds^2 = (d\theta)^2 + (\cos\theta\, d\alpha)^2$ and the problem is to minimize

$$I = \int_0^{\alpha_1} \sqrt{u^2 + \cos^2\theta}\, d\alpha$$

where

$$\dot{\theta} = u \qquad \theta(0) = 0$$

The Hamiltonian is given by

$$H = \sqrt{u^2 + \cos^2\theta} + pu$$

The first order necessary conditions are

$$\frac{u}{\sqrt{u^2 + \cos^2\theta}} + p = 0$$

$$\dot{p} = \frac{\cos\theta \sin\theta}{\sqrt{u^2 + \cos^2\theta}} \qquad p(T) = 0$$

They are satisfied by $\bar{u} = 0$, $\bar{\theta} = 0$, $\bar{p} = 0$. The second derivatives of H evaluated along $\bar{u}$, $\bar{\theta}$, $\bar{p}$ are

$$H_{uu} = 1 \qquad H_{u\theta} = 0 \qquad H_{\theta\theta} = -1$$

The Riccati equation (4.8) then becomes

$$-\frac{dS}{d\alpha} = -1 - S^2 \qquad S(\alpha_1) = 0$$

with solution

$$S(\alpha) = -\tan(\alpha_1 - \alpha)$$

The second order sufficiency conditions are satisfied if $0 < \alpha_1 < \pi/2$. The Riccati equation (4.7) becomes

$$-\frac{dS}{d\alpha} = -1 - (1 + \frac{1}{c})S^2 \qquad S(\alpha_1) = 0$$

with solution

$$S = -\frac{\tan \sqrt{1 + \frac{1}{c}}\,(\alpha_1 - \alpha)}{\sqrt{1 + \frac{1}{c}}}$$

The lower bound of c is then

$$c_0 = \frac{\alpha_1^2}{\frac{\pi^2}{4} - \alpha_1^2} \quad \text{for } 0 < \alpha_1 < \frac{\pi}{2}$$

4.3. Extension to Terminal Constraints.

With terminal constraints the problem can be written:

Minimize

$$I(x,u) = \int_0^T L\big(x(t),u(t),t\big)dt + F\big(x(T)\big)$$

subject to

$$\dot{x}(t) = f\big(x(t),u(t),t\big)$$
$$x(0) = a$$
$$\psi\big(x(T)\big) = 0$$

Terminal constraints have been treated by Nahra (1971), Mårtensson (1972) and O'Doherty and Pierson (1974). They replaced $F\big(x(T)\big)$ by

$$F(x(T)) + b^T\psi(x(T)) + \frac{c_2}{2}\psi^T(x(T))\psi(x(T))$$

and iterated on the multipliers b. The combination of this idea with the methods of the preceding section will now be studied.

Define

$$J(x,u,p,b,c_1,c_2) = \int_0^T \Big\{L(x,u,t) + p^T(f(x,u,t) - \dot{x}) +$$

$$+ \frac{c_1}{2}(f(x,u,t) - \dot{x})^T(f(x,u,t) - \dot{x})\Big\}dt +$$

$$+ F(x(T)) + b^T\psi(x(T)) + \frac{c_2}{2}\psi^T(x(T))\psi(x(T))$$

The following theorems, analogous to the ones of Section 4.2 can be proved.

Theorem 4.9. Let $\bar{p}$ and $\bar{b}$ satisfy equation (4.3). Assume that $c_1 > 0$, $c_2 \geq 0$, $H_{uu}(\bar{x}(t),\bar{u}(t),\bar{p}(t),t) > 0$, $t \in [0,T]$, and that the Riccati equation

$$-\dot{S} = H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} + (f_x - f_uH_{uu}^{-1}H_{ux})^TS + S(f_x - f_uH_{uu}^{-1}H_{ux}) -$$

$$- S(f_uH_{uu}^{-1}f_u^T + \frac{1}{c_1}I)S$$

$$S(T) = F_{xx} + c_2\psi_x^T\psi_x + \Sigma\bar{b}_i(\psi_i)_{xx} \qquad (4.10)$$

has a solution over $[0,T]$. Then $J(x,u,\bar{p},\bar{b},c_1,c_2)$ has a local minimum at $(\bar{x},\bar{u})$.

Proof. Follows from Theorems 4.2 - 4.4 with F replaced by $F + \bar{b}^T\psi + \frac{c_2}{2}\psi^T\psi$. □

Theorem 4.10. Assume that $J(x,u,\bar{p},\bar{b},c_1,c_2)$ has a local minimum at $(\bar{x},\bar{u})$ and that $H_{uu}(\bar{x},\bar{u},\bar{p},t) > 0$, $t \in [0,T]$. Then the Riccati equation (4.10) has a solution over $[\varepsilon,T]$ for arbitrary $\varepsilon > 0$.

Proof. Analogous to the proof of Theorem 4.8. □

It is interesting to study some special cases. First, let ψ determine $x(T)$ completely.

Theorem 4.11. Let $\psi(x(T))$ have dimension n. Assume that the regularity conditions of Theorem 4.1 hold and that $\bar{p}$ and $\bar{b}$ satisfy equation (4.3). Assume that

(i) $H_{uu}(\bar{x}(t),\bar{u}(t),\bar{p}(t),t) > 0 \qquad t \in [0,T]$

(ii) There exists a symmetric matrix S_0 such that the Riccati equation

$$-\dot{S} = H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} + (f_x - f_u H_{uu}^{-1}H_{ux})^T S +$$

$$+ S(f_x - f_u H_{uu}^{-1}H_{ux}) - S f_u H_{uu}^{-1} f_u^T S$$

$$S(T) = S_0 \tag{4.11}$$

has a solution in $[0,T]$.

Then there exist constants $c_1 > 0$ and $c_2 \geq 0$ such that $J(x,u,\bar{p},\bar{b},c_1,c_2)$ has a local minimum at $(\bar{x},\bar{u})$.

Proof. There exists a value of c_2 such that $F_{xx} + c_2\psi_x^T\psi_x + \Sigma\bar{b}_i(\psi_i)_{xx} \geq S_0$. The difference between $(f_u H_{uu}^{-1} f_u^T + \frac{1}{c_1} I)$ and $f_u H_{uu}^{-1} f_u^T$ can be made arbitrarily small by choosing c_1 large

enough. The result then follows from Lemmas A.1 and A.5 in the appendix. □

The simplest type of terminal constraint is $x_i(T) = d_i$ for some indices i. For easier notation assume that the variables are ordered such that

$$\begin{aligned} &x_i(T) = d_i \qquad i = 1, \ldots, r \\ &x_i(T) \text{ is free} \quad i = r+1, \ldots, n \end{aligned} \tag{4.12}$$

Theorem 4.12. Let the terminal constraint be given by (4.12) and assume that $\bar{b}$ and $\bar{p}$ are defined by (4.3). Assume that

(i) $H_{uu}(\bar{x}(t),\bar{u}(t),\bar{p}(t),t) > 0 \qquad t \in [0,T]$

(ii) There exists an r×r-matrix A such that the Riccati equation (4.11) with

$$S_0 = \left(\begin{array}{c|c} A & 0 \\ \hline 0 & F_{x_i x_j} \end{array}\right)$$

has a solution on $[0,T]$.

Then there exist constants c_1 and c_2 such that $J(x,u,\bar{p},\bar{b},c_1,c_2)$ has a local minimum at $(\bar{x},\bar{u})$.

Proof. Analogous to Theorem 4.11. □

Example 4.2. Shortest distance between two points on a sphere.

The difference compared with Example 4.1 is the boundary condition $\theta(\alpha_1) = 0$. The Riccati equation (4.11) becomes

$$-\frac{dS}{d\alpha} = -1 - S^2 \qquad S(\alpha_1) \text{ arbitrary}$$

with solution $S = -\tan(\alpha_0-\alpha)$, where α_0 can be chosen arbitrarily. To prolong the existence of S as much as possible α_0 should be taken close to $\alpha_1 - \pi/2$, which corresponds to large values of $S(\alpha_1)$. The sufficiency conditions are then satisfied on the interval $0 \leq \alpha \leq \pi - \varepsilon$ for any $\varepsilon > 0$.

The Riccati equation (4.10) becomes

$$-\frac{dS}{d\alpha} = -1 - (1+\frac{1}{c_1})S^2$$

$$S(\alpha_1) = c_2$$

with solution

$$S = -\frac{\tan\sqrt{1+\frac{1}{c_1}}(\alpha_0-\alpha)}{\sqrt{1+\frac{1}{c_1}}}$$

where

$$\alpha_0 = \alpha_1 - \frac{\arctan\sqrt{1+\frac{1}{c_1}}\,c_2}{\sqrt{1+\frac{1}{c_1}}}$$

The values c_1 and c_2 for which S exists on $[0,\alpha_1]$ are given by

$$c_2 + \frac{\tan\left(\frac{\pi}{2} - \sqrt{1+\frac{1}{c_1}}\,\alpha_1\right)}{\sqrt{1+\frac{1}{c_1}}} \geq 0 \qquad \text{see Fig. 4.1.}$$

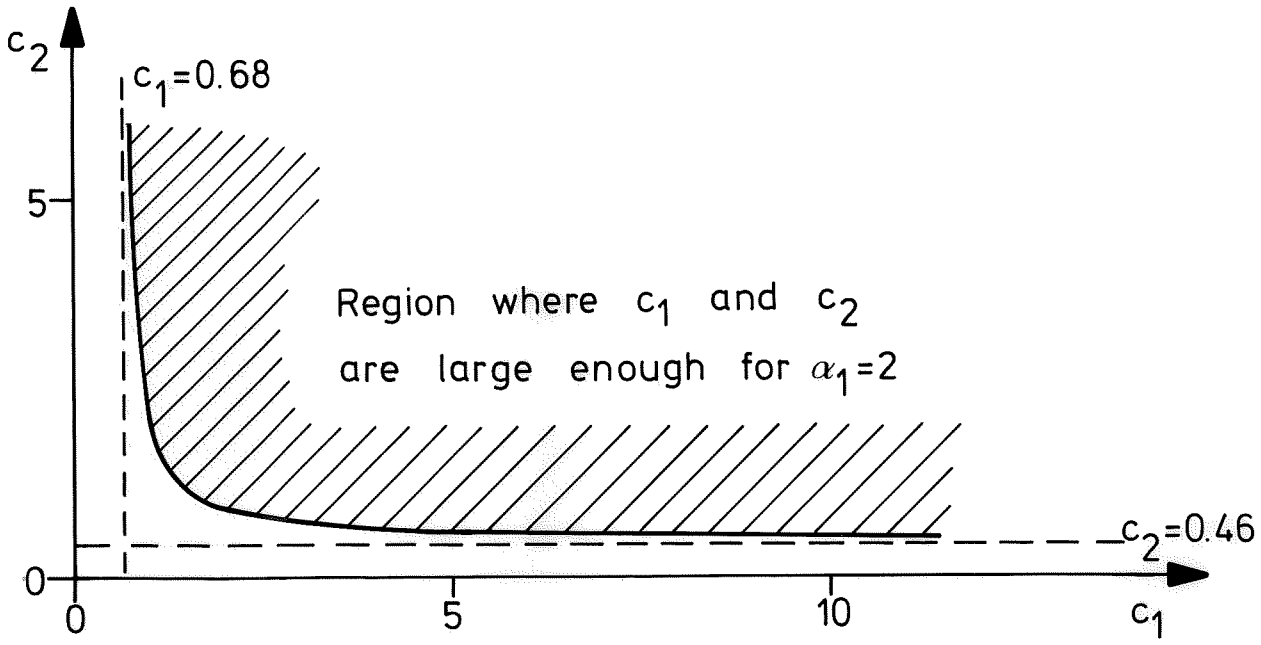
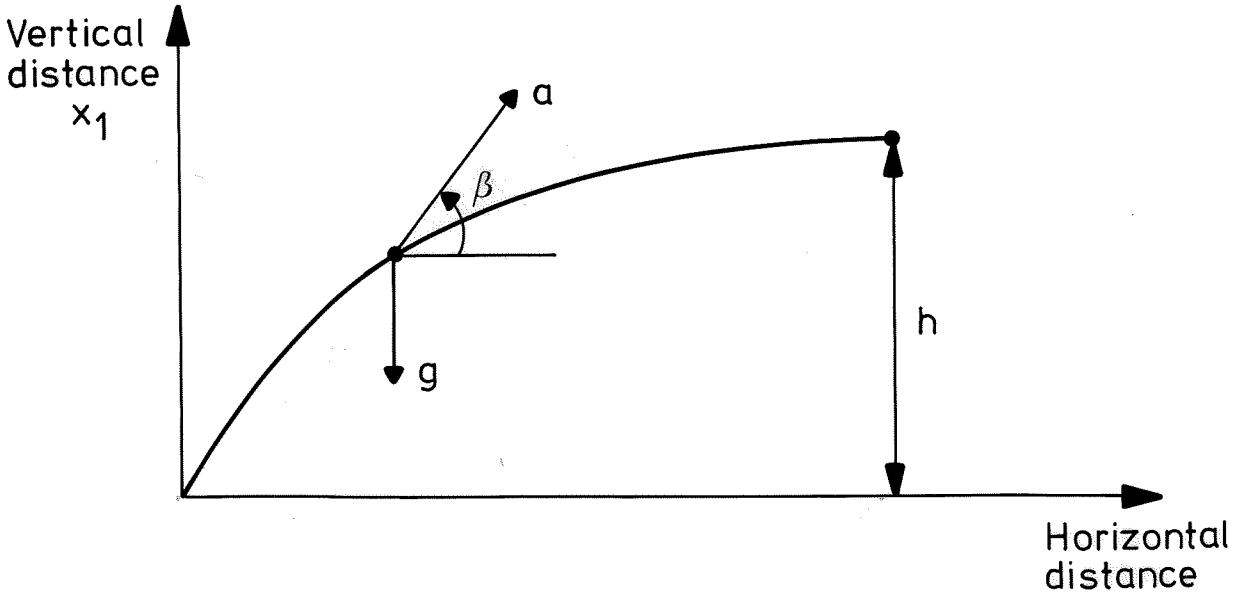


Fig. 4.1.

Example 4.3. See Bryson and Ho (1969).



Study the motion of a rocket in a constant gravitational field. The thrust has a constant magnitude a, but the thrust angle β

is a control variable. If x_2 denotes the vertical component of the velocity, the equations of motion are

$$\dot{x}_1 = x_2$$

$$\dot{x}_2 = a \sin \beta - g$$

$$x_1(0) = x_2(0) = 0$$

$$x_1(T) = h \qquad x_2(T) = 0$$

The objective is to maximize the horisontal velocity component for $t = T$. This gives the loss

$$J = - a \int_0^T \cos \beta \, dt$$

The Hamiltonian is

$$H = - a \cos \beta + p_1 x_2 + p_2(a \sin \beta - g)$$

The first order necessary conditions are

$$\dot{p}_1 = 0$$

$$\dot{p}_2 = - p_1$$

$$\sin \beta + p_2 \cos \beta = 0$$

This gives a control strategy of the form

$$\tan \beta = At + B$$

where A and B are determined by the boundary conditions. Along the optimal trajectory we have

$$H_{xx} = 0 \qquad H_{x\beta} = 0$$

$$H_{\beta\beta} = a \cos \beta - a p_2 \sin \beta = \frac{a}{\cos \beta} > 0$$

$$f_x = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \qquad f_u = \begin{pmatrix} 0 \\ a \cos \beta \end{pmatrix}$$

The Riccati equation (4.10) becomes

$$-\dot{S} = f_x^T S + S f_x - S(f_\beta H_{\beta\beta}^{-1} f_\beta^T + \frac{1}{c_1} I)S$$

$$S(T) = c_2 I$$

For $c_2 = 0$ the solution is $S(t) = 0$ all t. This means that any $c_1 > 0$ and $c_2 \geq 0$ will be sufficient for J to have a local minimum at the solution to the problem.

4.4. Iterative Methods.

The results of the previous two sections are only useful if p and b have the correct values $\bar{p}$ and $\bar{b}$. Therefore iterative methods of updating p and b in such a way that they converge to $\bar{p}$ and $\bar{b}$ must be studied. A natural way of updating p was suggested by Hestenes (1969) and used by di Pillo et.al. (1974). The updating rule is

$$p(t)^{(i+1)} = p(t)^{(i)} + c_1\left[f(x(t)^{(i)}, u(t)^{(i)}, t) - \dot{x}(t)^{(i)}\right]$$

where $x^{(i)}$ and $u^{(i)}$ are the values that minimize the functional $J(x,u,p^{(i)},b^{(i)},c_1,c_2)$. The multiplier b is updated using a similar rule by Nahra (1971) and O'Doherty and Pierson (1974).

$$b^{(i+1)} = b^{(i)} + c_2 \psi(x^{(i)}(T))$$

Note the similarity between these updating rules and the updating rules used in Chapter 3, algorithm 3.1. The convergence properties of these updating methods will now be studied. First consider the minimization of J for fixed p and b.

$$J(x,u,p,b,c_1,c_2) = \int_0^T \Big\{ L(x,u,t) + p^T\big(f(x,u,t) - \dot{x}\big) +$$

$$+ \frac{c_1}{2}\big(f(x,u,t) - \dot{x}\big)^T\big(f(x,u,t) - \dot{x}\big)\Big\} dt +$$

$$+ F\big(x(T)\big) + b^T\psi\big(x(T)\big) + \frac{c_2}{2}\,\psi\big(x(T)\big)^T\psi\big(x(T)\big) \qquad (4.13)$$

This problem is of a standard form studied in the calculus of variations. The minimum therefore satisfies the Euler equations, see Gelfand and Fomin (1963).

$$-\frac{d}{dt}\big(p + c_1(f-\dot{x})\big) = L_x^T + f_x^T p + c_1 f_x^T(f-\dot{x})$$

$$\big[p+c_1(f-\dot{x})\big]_{t=T} = F_x^T + \psi_x^T b + c_2\psi_x^T\psi$$

$$L_u^T + f_u^T p + c_1 f_u^T(f-\dot{x}) = 0 \qquad (4.14)$$

Introducing

$$p + c_1(f-\dot{x}) = \xi \qquad\qquad b + c_2\psi = \zeta$$

$H(x,u,p,t) = L(x,u,t) + p^T f(x,u,t)$ and

$\varphi(x,b) = F(x) + b^T\psi(x)$

these equations can be written

$$\dot{x} = f(x,u,t) + \frac{1}{c_1}(p-\xi)$$

$$-\dot{\xi} = H_x^T(x,u,\xi,t)$$

$$H_u(x,u,\xi,t) = 0$$

$$x(0) = a \qquad \psi\big(x(T)\big) = \frac{1}{c_2}(\zeta - b)$$

$$\xi(T) = \varphi_x^T\big(x(T), \zeta\big)$$

Let h, k, η, θ, q and d denote the deviations from the optimum i.e.

$$h = x - \bar{x} \qquad k = u - \bar{u} \qquad \eta = \xi - \bar{p} \qquad \theta = \zeta - \bar{b} \qquad q = p - \bar{p}$$

$$d = b - \bar{b}$$

Then the equations are

$$\dot{h} = f(\bar{x}+h, \bar{u}+k, t) - f(\bar{x}, \bar{u}, t) + \frac{1}{c_1}(q - \eta)$$

$$-\dot{\eta} = H_x^T(\bar{x}+h, \bar{u}+k, \bar{p}+\eta, t) - H_x^T(\bar{x}, \bar{u}, \bar{p}, t)$$

$$H_u(\bar{x}+h, \bar{u}+k, \bar{p}+\eta, t) = 0$$

$$h(0) = 0 \qquad \psi\big(\bar{x}(T) + h(T)\big) = \frac{1}{c_2}(\theta - d)$$

$$\eta(T) = \varphi_x^T\big(\bar{x}(T) + h(T), \bar{b} + \theta\big) - \varphi_x^T\big(\bar{x}(T), \bar{b}\big) \tag{4.15}$$

The linearized version of these equations is

$$\dot{h} = f_x h + f_u k - \frac{1}{c_1}\eta + \frac{1}{c_1} q$$

$$-\dot{\eta} = H_{xx} h + H_{xu} k + f_x^T \eta$$

$$H_{uu} k + H_{ux} h + f_u^T \eta = 0 \tag{4.16}$$

$$h(0) = 0 \qquad \psi_x h(T) = \frac{1}{c_2}(\theta - d)$$

$$\eta(T) = \varphi_{xx} h(T) + \psi_x^T \theta$$

where H_{xx}, H_{xu} etc. are evaluated along $(\bar{x},\bar{u},\bar{p})$.

If $H_{uu} > 0$, k can be expressed as

$$k = - H_{uu}^{-1}H_{ux}h - H_{uu}^{-1}f_u^T\eta$$

This gives the following two point boundary value problem.

$$\dot{h} = (f_x - f_u H_{uu}^{-1}H_{ux})h - (f_u H_{uu}^{-1}f_u^T + \frac{1}{c_1} I)\eta + \frac{1}{c_1} q$$

$$-\dot{\eta} = (H_{xx} - H_{xu}H_{uu}^{-1}H_{ux})h - (H_{xu}H_{uu}^{-1}f_u^T - f_x^T)\eta$$

$$h(0) = 0 \qquad \psi_x h(T) = \frac{1}{c_2}(\theta - d)$$

$$\eta(T) = \varphi_{xx}h(T) + \psi_x^T\theta \tag{4.17}$$

Let

$$\Phi(t,s) = \begin{pmatrix} \Phi_{11}(t,s) & \Phi_{12}(t,s) \\ \Phi_{21}(t,s) & \Phi_{22}(t,s) \end{pmatrix}$$

be the fundamental matrix of this system of linear differential equations and let S be the solution of the associated Riccati equation

$$-\dot{S} = H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} + (f_x - f_u H_{uu}^{-1}H_{ux})^T S +$$

$$+ S(f_x - f_u H_{uu}^{-1}H_{ux}) - S(f_u H_{uu}^{-1}f_u^T + \frac{1}{c_1} I)S$$

$$S(T) = F_{xx} + \Sigma\bar{b}_i(\psi_i)_{xx} + c_2\psi_x^T\psi_x \tag{4.18}$$

Note that this Riccati equation is identical to (4.10). Assume that there exist c_1^0 and c_2^0 such that (4.18) has a solution on

$[0,T]$ for $c_1 \geq c_1^0$, $c_2 \geq c_2^0$. In what follows, only values of c_1 and c_2 satisfying $c_1 \geq c_1^0$, $c_2 \geq c_2^0$ will be studied.

The two point boundary value problem (4.15) can be represented as an integral equation, using the technique of Falb and de Jong (1969). A short description is given in the appendix. It is convenient to regard θ as a function on $[0,T]$ satisfying the differential equation $\dot{\theta} = 0$. The boundary conditions of the linearized problem (4.17) can then be written

$$\begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} h(0) \\ \eta(0) \\ \theta(0) \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ \varphi_{xx} & -I & \psi_x^T \\ \psi_x & 0 & -\frac{1}{c_2} I \end{pmatrix} \begin{pmatrix} h(T) \\ \eta(T) \\ \theta(T) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -\frac{1}{c_2} d \end{pmatrix}$$

The linearized problem is boundary compatible if the following matrix is nonsingular

$$A = \begin{pmatrix} I & 0 & 0 \\ \varphi_{xx}\Phi_{11} - \Phi_{21} & \varphi_{xx}\Phi_{12} - \Phi_{22} & \psi_x^T \\ \psi_x\Phi_{11} & \psi_x\Phi_{12} & -\frac{1}{c_2} I \end{pmatrix} \tag{4.19}$$

where $\Phi_{ij} = \Phi_{ij}(T,0)$.

A is nonsingular if

$$\left(\Phi_{22}(T,0) - (\varphi_{xx}+c_2\psi_x^T\psi_x)\Phi_{12}(T,0)\right)$$

is nonsingular. Since this matrix is related to the solution of the Riccati equation (4.18) by

$$S(t) = \left(\Phi_{22}(T,t) - (\varphi_{xx}+c_2\psi_x^T\psi_x)\Phi_{12}(T,t)\right)^{-1} \cdot$$

$$\cdot \left((\varphi_{xx}+c_2\psi_x^T\psi_x)\Phi_{11}(T,t) - \Phi_{21}(T,t)\right)$$

the nonsinularity follows from the assumption that S(t) exists on [0,T].

Note that the equation

$$H_u(\bar{x}+h,\bar{u}+k,\bar{p}+\eta,t) = 0$$

defines k uniquely in terms of h and η if h and η are sufficiently small. This follows from the implicit function theorem, see Luenberger (1968), since $H_{uu}(\bar{x},\bar{u},\bar{p},t) > 0$. The solution of (4.15) can now be written

$$\begin{pmatrix} h \\ \eta \\ \theta \end{pmatrix} = K(t) \begin{pmatrix} 0 \\ \varphi_x^T(\bar{x}(T) + h(T), \bar{b} + \theta) - \varphi_x^T(\bar{x}(T),\bar{b}) - \varphi_{xx}h(T) - \psi_x^T\theta \\ \psi(\bar{x}(T) + h(T)) - \psi_x h(T) - \frac{1}{c_2} d \end{pmatrix} +$$

$$+ \int_0^T G(t,s) \begin{pmatrix} f(\bar{x}+h,\bar{u}+k,s) - f(\bar{x},\bar{u},s) - f_x h - f_u k + \frac{1}{c_1} q \\ H_x^T(\bar{x},\bar{u},\bar{p},s) - H_x^T(\bar{x}+h,\bar{u}+k,\bar{p}+\eta,s) + H_{xx}h + H_{xu}k + f_x^T\eta \\ 0 \end{pmatrix} ds \tag{4.20}$$

with k given by $H_u^T(\bar{x}+h,\bar{u}+k,\bar{p}+\eta,t) = 0$. K(t) and G(t,s) are the Green's matrices associated with the linear two point boundary value problem (see Lemma A.6 in the appendix).

Theorem 4.13. There exist constants $\varepsilon > 0$ and $\delta > 0$ such that for all continuous functions q and all d with $\| q \|_0 \leq \delta$ and $\| d \| \leq \delta$, there exists a unique solution, (h,η), to the integral equation (4.20) satisfying $\| h \|_0 + \| \eta \|_0 + \| \theta \|_0 \leq \varepsilon$.

Proof. The integral equation can be written as an operator equation

$$h = T_1(h,\eta,\theta) + A_1 \cdot q$$
$$\eta = T_2(h,\eta,\theta)$$
$$\theta = T_3(h,\eta,\theta) + A_2 \cdot d$$

where T_i are maps from $C_0^{3n}[0,T]$ to $C_0^n[0,T]$ and A_1 and A_2 are linear maps. Let α be a real number, $0 \leq \alpha < 1$. Then from equation (4.20) it follows that there exists an $\varepsilon > 0$ and a $\delta > 0$ such that

$$\| T_i(h_1,\eta_1,\theta_1) - T_i(h_2,\eta_2,\theta_2) \|_0 \leq$$
$$\leq \alpha(\| h_1-h_2 \|_0 + \| \eta_1-\eta_2 \|_0 + \| \theta_1-\theta_2 \|_0) \qquad i = 1, 2, 3$$

for all h_1, h_2, η_1, η_2, θ_1 and θ_2 satisfying

$$\| h_i \|_0 + \| \eta_i \|_0 + \| \theta_i \|_0 \leq \varepsilon \qquad i = 1, 2$$

It also follows that

$$\| T_1(0,0,0) + A_1 \cdot q \| \leq \| A_1 \| \cdot \| q \|$$
$$\| T_3(0,0,0) + A_2 \cdot d \| \leq \| A_2 \| \cdot \| d \|$$

Define $\eta = \max(\| A_1 \| \cdot \| q \|, \| A_2 \| \cdot \| d \|)$. Choose δ such that

$$\frac{1}{1-\alpha}\,\eta \leq \varepsilon$$

for q and d satisfying $\| q \|_0 \leq \delta$, $\| d \| \leq \delta$. The conditions of the contraction mapping Lemma A.7 in the appendix are then satisfied and the theorem is proved. □

To study the solution h, k and η for small values of q and b it is desirable to have an approximate representation.

Theorem 4.14. Let (h,η) be the solution of the nonlinear problem (4.15). Then

$$\begin{pmatrix} h \\ \eta \\ \theta \end{pmatrix} = K(t) \begin{pmatrix} 0 \\ 0 \\ -\frac{1}{c_2} d \end{pmatrix} + \int_0^T \frac{1}{c_1} G(t,S) \begin{pmatrix} q \\ 0 \\ 0 \end{pmatrix} dS + r(q,d)$$

where $\| r(q,d) \|_0 / (\| q \|_0 + \| d \|) \to 0$ as $(q,d) \to 0$.

Proof. From (4.20) it follows that

$$\| T_i(h,k,\eta) \| \leq \frac{1}{3} K_1 (\| h \|_0 + \| \eta \|_0 + \| \theta \|_0)^2$$

and consequently

$$\| h \|_0 + \| \eta \|_0 + \| \theta \|_0 \leq K_1 (\| h \|_0 + \| \eta \|_0 + \| \theta \|_0)^2 +$$

$$+ K_2 (\| q \|_0 + \| d \|)$$

for some constants K_1 and K_2. Let ε be the constant defined in Theorem 4.13 and let $\varepsilon_1 = \min(1/2K_1, \varepsilon)$. Then, for sufficiently small $\| q \|$ and $\| d \|$, $(1/1-\alpha)\eta \leq \varepsilon_1$, where α and η are defined as in Theorem 4.13. Consequently $\| h \|_0 + \| \eta \|_0 + \| \theta \|_0 \leq$ $\leq 1/2K_1$ for sufficiently small $\| q \|_0$ and $\| d \|$. This gives

$$\| h \|_0 + \| \eta \|_0 + \| \theta \|_0 \leq 2K_2 (\| q \|_0 + \| d \|)$$

Using this in the expression for $\| T_i(h,\eta,\theta) \|$ gives the desired bound on $r(q,d)$. □

Corollary. Let $\tilde{h}$, $\tilde{\eta}$ and $\tilde{\theta}$ denote the solution to the linearized boundary value problem (4.17). Then the solutions of the nonlinear and the linearized problem are related by

$$\begin{pmatrix} h \\ \eta \\ \theta \end{pmatrix} = \begin{pmatrix} \tilde{h} \\ \tilde{\eta} \\ \tilde{\theta} \end{pmatrix} + r(q,d)$$

where $\| r(q,d) \|_0 / (\| q \|_0 + \| d \|) \to 0$ as $(q,d) \to 0$.

Proof. The solution to (4.17) is given by

$$\begin{pmatrix} \tilde{h} \\ \tilde{\eta} \\ \tilde{\theta} \end{pmatrix} = K(t) \begin{pmatrix} 0 \\ 0 \\ -\frac{1}{c_2} d \end{pmatrix} + \int_0^T G(t,s) \begin{pmatrix} \frac{1}{c_1} q \\ 0 \\ 0 \end{pmatrix} ds \qquad (4.21)$$

□

With this result it is possible to investigate the convergence rate of the iterate method of updating the multipliers. As mentioned at the beginning of this section, the algorithm will be assumed to be the following:

Algorithm 4.1.

(i) Choose starting values $p^{(0)}$, $b^{(0)}$, put $i = 0$.

(ii) Minimize $J(x,u,p^{(i)},b^{(i)},c_1,c_2)$; let the result be $x^{(i)}$, $u^{(i)}$.

(iii) Update the multipliers

$$p^{(i+1)}(t) = p^{(i)}(t) + c_1 \left(f(x^{(i)},u^{(i)},t) - \dot{x}^{(i)}(t) \right)$$

$$b^{(i+1)} = b^{(i)} + c_2 \psi(x^{(i)}(T))$$

put $i = i + 1$ and go to (ii).

It is assumed that c_1 and c_2 are held constant, $c_1 \geq c_1^0$, $c_2 \geq c_2^0$.

Theorem 4.15. Let $p^{(i)}$ and $b^{(i)}$ be generated by Algorithm 4.1. Assume that

(i) $\bar{p}$ and $\bar{b}$ satisfy equation (4.3).

(ii) the linearized system

$$\dot{x} = f_x(\bar{x},\bar{u},t)h + f_u(\bar{x},\bar{u},t)k$$

is controllable.

(iii) $H_{uu}(\bar{x},\bar{u},\bar{p},t) > 0 \qquad t \in [0,T]$

(iv) The Riccati equation (4.18) has a solution on $[0,T]$ for $c_1 = c_1^0$, $c_2 = c_2^0$.

Then there are constants $c_1 \geq c_1^0$, $c_2 \geq c_2^0$ such that, if $p^{(0)}$ and $b^{(0)}$ are sufficiently close to $\bar{p}$ and $\bar{b}$ then

$$\| p^{(i+1)} - \bar{p} \|_0 + \| b^{(i+1)} - \bar{b} \| \leq K(\| p^{(i)} - \bar{p} \|_0 + \| b^{(i)} - \bar{b} \|)$$

where K is an arbitrary number in (0,1).

Proof. Using the notation

$$p^{(i)} - \bar{p} = q^{(i)} \qquad x^{(i)} - \bar{x} = h^{(i)}$$

$$u^{(i)} - \bar{u} = k^{(i)} \qquad b^{(i)} - \bar{b} = d^{(i)}$$

the updating formula can be written

$$q^{(i+1)} = q^{(i)} + c_1(f_x h^{(i)} + f_u k^{(i)} - \dot{h}^{(i)}) +$$

$$+ c_1 R_1(h^{(i)},k^{(i)})$$

$$d^{(i+1)} = d^{(i)} + c_2\psi_x h^{(i)}(T) + c_2 R_2\left(h^{(i)}(T)\right)$$

where

$$\| R_1(h,k) \|_0 / (\| h \|_0 + \| k \|_0) \to 0, \qquad (h,k) \to 0$$

$$\| R_2(z) \| / \| z \| \to 0, \qquad z \to 0$$

If $\tilde{h}$, $\tilde{\eta}$, $\tilde{\theta}$ and $\tilde{k}$ denote the solution to the linear two point boundary value problem, then it follows from Theorem 4.14 that

$$q^{(i+1)} = q^{(i)} + c_1\left(f_x\tilde{h}^{(i)} + f_u\tilde{k}^{(i)} - \dot{\tilde{h}}^{(i)}\right) + R_3\left(q^{(i)}, d^{(i)}\right)$$

$$d^{(i+1)} = d^{(i)} + c_2\psi_x\tilde{h}^{(i)}(T) + R_4\left(q^{(i)}, d^{(i)}\right) \qquad (4.22)$$

where

$$\| R_i(q,d) \|_0 / (\| q \|_0 + \| d \|) \to 0, \qquad (q,d) \to 0$$

From (4.16) it follows that

$$f_x\tilde{h}^{(i)} + f_u\tilde{k}^{(i)} - \dot{h}^{(i)} = \frac{1}{c_1}\left(\eta^{(i)} - q^{(i)}\right)$$

$$\psi_x\tilde{h}(T) = \frac{1}{c_2}\left(\theta^{(i)} - d^{(i)}\right)$$

Using these expressions in (4.22) results in

$$q^{(i+1)} = \eta^{(i)} + R_3\left(q^{(i)}, d^{(i)}\right)$$

$$d^{(i+1)} = \theta^{(i)} + R_4\left(q^{(i)}, d^{(i)}\right)$$

From conditions (i) - (iv) it follows that the linear problem (4.17) has a solution for $c_1 = \infty$, $c_2 = \infty$. Then $K(t)$ and $G(t,S)$ go to finite limits as $c_1 \to \infty$, $c_2 \to \infty$. From (4.21) it then fol-

lows that there are values $c_1 \geq c_1^0$ and $c_2 \geq c_2^0$ such that

$$\| \eta^{(i)} \|_0 + \| \theta^{(i)} \|_0 \leq \frac{K}{2}(\| q^{(i)} \|_0 + \| d^{(i)} \|)$$

For these values of c_1 and c_2 then choose δ such that

$$\| R_i(q^{(i)}, d^{(i)}) \|_0 \leq \frac{K}{4} (\| q^{(i)} \|_0 + \| d^{(i)} \|)$$

for $\| q^{(i)} \|_0 + \| d^{(i)} \| \leq \delta$. Then

$$\| q^{(i+1)} \|_0 + \| d^{(i+1)} \| \leq K(\| q^{(i)} \|_0 + \| d^{(i)} \|)$$

for $\| q^{(0)} \|_0 + \| d^{(0)} \| \leq \delta$. □

Theorem 4.15 shows that Algorithm 4.1 can be used to solve the optimal control problem. However, this algorithm is based on the minimization of J for fixed values of the multipliers. This is not a trivial problem, even if it is simpler than the original optimization problem because the differential equation and terminal constraints are eliminated. Di Pillo et al. (1974) have studied this problem and shown that a conjugate gradient method can be used. The optimization problem can then be solved using only quadrature and without the solution of any differential equations.

4.5. Extension to Mixed State-Control Inequality Constraints.

The results will now be extended to inequality constraints on the state and control variables. The problem is then:

minimize

$$I(x,u) = \int_0^T L\big(x(t),u(t),t\big)dt + F\big(x(T)\big)$$

under the constraints

$$\begin{aligned} &\dot{x}(t) = f\big(x(t),u(t),t\big) && t \in [0,T] \\ &x(0) = a \\ &g\big(x(t),u(t),t\big) \leq 0 && t \in [0,T] \end{aligned} \qquad (4.23)$$

where g is a q-dimensional vector function, three times continuously differentiable with respect to its arguments. For simplicity it is assumed that there are no terminal constraints. The extension to terminal constraints is only a straightforward combination of the results of this section and those of 4.3.

The necessary conditions of the problem defined by (4.23) are given by the following theorem.

Theorem 4.16. Let $(\bar{x},\bar{u})$ be the solution to (4.23) and assume that $g_u(\bar{x},\bar{u},t)$ has full rank in $[0,T]$. Define

$$H(x,u,p,\lambda,t) = L(x,u,t) + p^T f(x,u,t) + \lambda^T g(x,u,t)$$

Then there exist functions $\bar{p}(t) \in C_1^n[0,T]$, $\bar{\lambda}(t) \in C_0^n[0,T]$ such that

$$
\begin{aligned}
&\dot{\bar{p}} = - H_x(\bar{x},\bar{u},\bar{p},\bar{\lambda},t) && t \in [0,T] \\
&\bar{p}(T) = F_x^T(\bar{x}(T)) \\
&H_u(\bar{x},\bar{u},\bar{p},\bar{\lambda},t) = 0 && t \in [0,T] \\
&\lambda^T(t) g(\bar{x},\bar{u},t) = 0 && t \in [0,T] \\
&g(\bar{x},\bar{u},t) \leq 0 && t \in [0,T] \\
&\lambda(t) \geq 0 && t \in [0,T]
\end{aligned}
\tag{4.24}
$$

Proof. See Luenberger (1968). □

The logical extension of the finite dimensional method of treating inequality constraints would be to use the augmented loss function

$$
\begin{aligned}
J(x,u,p,\lambda,c,\sigma) = \int_0^T \Big\{ & L\big(x(t),u(t),t\big) + p^T(t)\Big(f\big(x(t),u(t),t\big) - \dot{x}\Big) + \\
& + \frac{c}{2}\Big(f\big(x(t),u(t),t\big) - \dot{x}\Big)^T\Big(f\big(x(t),u(t),t\big) - \dot{x}\Big) + \\
& + \frac{1}{2\sigma}\Big[\Big(\sigma g\big(x(t),u(t),t\big) + \lambda(t)\Big)_+^T \cdot \\
& \cdot \Big(\sigma g\big(x(t),u(t),t\big) + \lambda(t)\Big)_+ - \lambda^T(t)\lambda(t)\Big]\Big\} dt + \\
& + F\big(x(T)\big)
\end{aligned}
\tag{4.25}
$$

It is difficult to analyze this expression directly because the last term under the integral sign is not twice continuously differentiable. Therefore J will be derived in the following way, using an idea of Rockafellar (1973) for the finite dimensional case. Transform the inequality $g\big(x(t),u(t),t\big) \leq 0$ to an equality

$$\begin{bmatrix} g_1(x(t),u(t),t) + v_1^2(t) \\ \vdots \\ g_q(x(t),u(t),t) + v_q^2(t) \end{bmatrix} = 0$$

using the auxiliary variables $v_1(t)$, ..., $v_q(t)$. Then study

$$J_1(x,u,v,\bar{p},\bar{\lambda},c,\sigma) = \int_0^T \Big\{ L(x,u,t) + \bar{p}^T\big(f(x,u,t) - \dot{x}\big) +$$

$$+ \frac{c}{2}\big(f(x,u,t) - \dot{x}\big)^T\big(f(x,u,t) - \dot{x}\big) +$$

$$+ \bar{\lambda}^T\big(g(x,u,t) + \tilde{V}\big) + \frac{\sigma}{2}\big(g(x,u,t) + \tilde{V}\big)^T \cdot$$

$$\cdot \big(g(x,u,t) + \tilde{V}\big)\Big\}dt + F(x(T))$$

where

$$\tilde{V} = \begin{pmatrix} v_1^2 \\ \vdots \\ v_q^2 \end{pmatrix}$$

For simplicity the time dependence of x, u and v has been suppressed in the terms in the integral.

Now expand J_1 to second order around the optimum $(\bar{x},\bar{u},\bar{v})$. Then

$$J_1(\bar{x}+h,\bar{u}+k,\bar{v}+w,\bar{p},\bar{\lambda},c,\sigma) - J_1(\bar{x},\bar{u},\bar{v},\bar{p},\bar{\lambda},c,\sigma) =$$

$$= \int_0^T \{H_x h + H_u k - \bar{p}^T\dot{h}\}dt + F_x h(T) +$$

$$+ \frac{1}{2}\int_0^T \Big\{ h^T H_{xx} h + 2h^T H_{xu} k + k^T H_{uu} k +$$

$$+ w^T \Lambda w + c(f_x h + f_u k - \dot{h})^T (f_x h + f_u k - \dot{h}) +$$

$$+ \sigma \left(g(\bar{x}+h, \bar{u}+k, t) + \tilde{V}(\bar{v}+w) \right)^T \left(g(\bar{x}+h, \bar{u}+k, t) + \tilde{V}(\bar{v}+w) \right) \Big\} dt +$$

$$+ \frac{1}{2} h^T(T) F_{xx} h(T) + R(h,k) \qquad (4.26)$$

where

$$\Lambda = \begin{pmatrix} \bar{\lambda}_1 & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & \bar{\lambda}_q \end{pmatrix}$$

The term $R(h,k)$ satisfies

$$|R(h,k)| / \int_0^T (h^T h + \dot{h}^T \dot{h} + k^T k)\, dt \to 0, \quad (h,k) \to 0$$

Assume the following regularity conditions

- For each g_i there is a finite set of disjoint intervals

  $$I_i = \{[t_i^1, t_i^2], [t_i^3, t_i^4], \ldots\}$$

  such that $g_i(\bar{x}(t), \bar{u}(t), t) = 0$, $t \in I_i$, and $g_i(\bar{x}(t), \bar{u}(t), t) < 0$, $t \notin I_i$.

- The multipliers $\lambda_i(t)$ satisfy

  $$\lambda_i(t) > 0, \quad t \in \{(t_i^1, t_i^2), (t_i^3, t_i^4), \ldots\}$$

The main result is then

Theorem 4.17. Let $Q(t,\varepsilon)$ be a diagonal matrix whose elements satisfy

$$Q_{ii}(t,\varepsilon) = \begin{cases} 1 \text{ if } t \in \left\{[t_i^1+\varepsilon, t_i^2-\varepsilon],\ [t_i^3+\varepsilon, t_i^4-\varepsilon], \ldots\right\} \\ \\ 0 \text{ otherwise} \end{cases}$$

and define

$$\tilde{H}_{xx} = H_{xx} + \sigma g_x^T Q g_x$$
$$\tilde{H}_{xu} = H_{xu} + \sigma g_x^T Q g_u$$
$$\tilde{H}_{uu} = H_{uu} + \sigma g_u^T Q g_u$$

Assume that $c > 0$, $\sigma > 0$, that

$$H_{uu}\left(\bar{x}(t), \bar{u}(t), \bar{p}(t), \bar{\lambda}(t), t\right) > 0 \quad t \in [0,T]$$

and that the Riccati equation

$$-\dot{S} = \tilde{H}_{xx} - \tilde{H}_{xu}\tilde{H}_{uu}^{-1}\tilde{H}_{ux} + S(f_x - f_u \tilde{H}_{uu}^{-1}\tilde{H}_{ux}) +$$

$$+ (f_x - f_u\tilde{H}_{uu}^{-1}\tilde{H}_{ux})^T S - S(f_u\tilde{H}_{uu}^{-1}f_u^T + \frac{1}{c} I)S$$

$$S(T) = F_{xx} \tag{4.27}$$

has a solution in $[0,T]$ for some value $\varepsilon > 0$.

Then $J_1(x,u,v,\bar{p},\bar{\lambda},c,\sigma)$ has a local minimum at $\bar{x},\bar{u},\bar{v}$.

Proof. To simplify the notation, the case where there is only one constraint, inactive in $[0,t)$, active in $[t_1,T]$, will be considered. The general case is a straightforward extension. Equation (4.26) can then be written

$$\Delta J_1 = J_1(\bar{x}+h, \bar{u}+k, \bar{v}+w, \bar{p}, \bar{\lambda}, c, \sigma) - J_1(\bar{x}, \bar{u}, \bar{v}, \bar{p}, \bar{\lambda}, c, \sigma) =$$

$$= \frac{1}{2}\int_0^T \begin{pmatrix} k + A_1 h \\ \dot{h} + A_2 h \end{pmatrix}^T \begin{pmatrix} \tilde{H}_{uu} + cf_u^T f_u & -cf_u^T \\ -cf_u & cI \end{pmatrix} \begin{pmatrix} k + A_1 h \\ \dot{h} + A_2 h \end{pmatrix} dt +$$

$$+ \frac{1}{2}\int_0^{t_1+\varepsilon} \sigma\left(g(\bar{x}+h,\bar{u}+k,t) + (\bar{v}+w)^2\right)^2 dt + \frac{1}{2}\int_{t_1}^{t_1+\varepsilon} 2\Lambda w^2 d +$$

$$+ \frac{1}{2}\int_{t_1+\varepsilon}^{T} 2\Lambda w^2 dt + R_1(h,k) + R_2(w)$$

where

$$A_1 = \tilde{H}_{uu}^{-1}(\tilde{H}_{ux}+f_u^T S)$$

$$A_2 = f_u\tilde{H}_{uu}^{-1}(\tilde{H}_{ux}+f_u^T S) - f_x + \frac{1}{c} S$$

S is the solution to the Riccati equation (4.27). R_1 and R_2 satisfy

$$|R_1(h,k)|/\int_0^T (h^T h+\dot{h}^T\dot{h}+k^T k)\,dt \to 0 \quad (h,k) \to 0$$

$$|R_2(w)|/\int_{t_1+\varepsilon}^{T} w^T w dt \to 0 \qquad w \to 0$$

Using the same reasoning as in Theorems 4.2 and 4.3 it now follows that

$$\Delta J_1 \geq \left(\eta - \rho_1(h,k)\right)\int_0^T (\dot{h}^T\dot{h}+h^T k^T + k^T k)\,dt + \left(\Lambda - \rho_2(w)\right)\int_{t_1+\varepsilon}^{T} w^2 dt +$$

$$+ \frac{1}{2}\int_0^{t_1+\varepsilon} \left(g(\bar{x}+h,\bar{u}+k,t) + (\bar{v}+w)\right)^2 dt$$

where $\eta > 0$ and $\rho_1(h,k) \to 0$, $(h,k) \to 0$ and $\rho_2(w) \to 0$ as $w \to 0$. Then $\Delta J_1 > 0$ for all (h,k,w) that are sufficiently small and not all identically zero. □

Corollary. If (4.27) has a solution on $[0,T]$ with $Q = Q(t,0)$ and the other assumptions are valid, then J_1 has a local minimum at $(\bar{x},\bar{u},\bar{v})$.

Proof. If (4.27) has a solution on $[0,T]$ for $Q = Q(t,0)$, then from Lemma A.5 in the appendix it has a solution for $Q = Q(t,\varepsilon)$ for some $\varepsilon > 0$. □

The minimization with respect to v can be done explicitly. The part of J_1 that depends on v_i is

$$(J_1)_i = \int_0^T \Big[\bar{\lambda}_i(t) \Big(g_i\big(x(t),u(t),t\big) + v_i^2(t) \Big) + \\ + \frac{\sigma}{2} \Big(g_i\big(x(t),u(t),t\big) + v_i^2(t) \Big)^2 \Big] dt$$

The minimum is reached by minimizing the integrand at each instant of time. The value of $v_i(t)$ is therefore the one that minimizes

$$\bar{\lambda}_i(t) v_i^2(t) + \sigma g_i\big(x(t),u(t),t\big) v_i^2(t) + \frac{\sigma}{2} v_i^4(t)$$

The result is

$$v_i^2(t) = \frac{1}{\sigma} \Big(- \bar{\lambda}_i - \sigma g_i\big(x(t),u(t),t\big) \Big)_+$$

Using this value of v_i^2 in $(J_1)_i$ gives

$$(J_1)_i = \int_0^T \frac{1}{2\sigma} \Big[\big(\sigma g_i(x,u,t) + \bar{\lambda}_i\big)_+^2 - \bar{\lambda}_i^2 \Big] dt$$

It then follows that

$$J(x,u,p,\lambda,c,\sigma) = \min_v J_1(x,u,v,p,\lambda,c,\sigma)$$

Theorem 4.18. Under the same assumptions as in Theorem 4.17 it follows that $J(x,u,\bar{p},\bar{\lambda},c,\sigma)$ has a local minimum at $(\bar{x},\bar{u})$.

Proof. Since $J_1(x,u,v,\bar{p},\bar{\lambda},c,\sigma) \geq J_1(\bar{x},\bar{u},\bar{v},\bar{p},\bar{\lambda},c,\sigma)$ it follows that $J(x,u,\bar{p},\bar{\lambda},c,\sigma) = \min_v J_1(x,u,v,\bar{p},\bar{\lambda},c,\sigma) \geq J_1(\bar{x},\bar{u},\bar{v},\bar{p},\bar{\lambda},c,\sigma) = J(\bar{x},\bar{u},\bar{p},\bar{\lambda},c,\sigma)$ for all (x,u) in a neighbourhood of $(\bar{x},\bar{u})$.

Theorem 4.19. Let the assumptions of Theorem 4.16 be valid. A sufficient condition for $(\bar{x},\bar{u})$ to be a solution to (4.23) is that $\bar{p}$ and $\bar{\lambda}$ satisfy (4.24), that

$$H_{uu}(\bar{x}(t),\bar{u}(t),\bar{p},\bar{\lambda},t) > 0 \qquad 0 \leq t \leq T$$

and that the Riccati equation

$$\begin{aligned} -\dot{S} = {} & H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} + \\ & + (H_{xu}H_{uu}^{-1}\tilde{g}_u^T - \tilde{g}_x^T)(\tilde{g}_u H_{uu}^{-1}\tilde{g}_u^T)^{-1}(\tilde{g}_u H_{uu}^{-1}H_{ux} - \tilde{g}_x) + \\ & + S(f_x - f_u H_{uu}^{-1}H_{ux} + f_u H_{uu}^{-1}\tilde{g}_u^T(\tilde{g}_u H_{uu}^{-1}\tilde{g}_u^T)^{-1}(\tilde{g}_u H_{uu}^{-1}H_{ux} - \tilde{g}_x)) + \\ & + (f_x - f_u H_{uu}^{-1}H_{ux} + f_u H_{uu}^{-1}\tilde{g}_u^T(\tilde{g}_u H_{uu}^{-1}\tilde{g}_u^T)^{-1}(\tilde{g}_u H_{uu}^{-1}H_{ux} - \tilde{g}_x))^T S - \\ & - S(f_u H_{uu}^{-1}f_u^T - f_u H_{uu}^{-1}\tilde{g}_u^T(\tilde{g}_u H_{uu}^{-1}\tilde{g}_u^T)^{-1}\tilde{g}_u H_{uu}^{-1}f_u^T)S \end{aligned}$$

$$S(T) = F_{xx} \tag{4.28}$$

has a solution over the whole interval $0 \leq t \leq T$. Here $\tilde{g}_x = Q(t,0)\tilde{g}_x$, $\tilde{g}_u = Q(t,0)g_u$, i.e.

$$\left(\tilde{g}_x(\bar{x}(t),\bar{u}(t),t)\right)_i = \begin{cases} \left(g_x(\bar{x}(t),\bar{u}(t),t)\right)_i & \text{for } t \text{ where } g_i \text{ is active} \\ 0 & \text{for } t \text{ where } g_i \text{ is inactive} \end{cases}$$

and analogously for $\tilde{g}_u$.

Moreover, if these sufficiency conditions are satisfied, then there exist c and σ such that $J(x,u,\bar{p},\bar{\lambda},c,\sigma)$ has a local minimum at $(\bar{x},\bar{u})$.

Proof. The matrices $\tilde{H}_{xx}$, $\tilde{H}_{xu}$ and $\tilde{H}_{uu}$ in equation (4.27) can be rewritten as follows

$$\tilde{H}_{uu} = H_{uu} + \sigma \tilde{g}_u^T \tilde{g}_u$$

$$\tilde{H}_{uu}^{-1} = H_{uu}^{-1} - H_{uu}^{-1}\tilde{g}_u^T\left(\frac{1}{\sigma} I + \tilde{g}_u H_{uu}^{-1}\tilde{g}_u^T\right)^{-1}\tilde{g}_u H_{uu}^{-1}$$

$$\tilde{H}_{uu}^{-1}\tilde{H}_{ux} = H_{uu}^{-1}H_{ux} - H_{uu}^{-1}\tilde{g}_u^T\left(\frac{1}{\sigma} I + \tilde{g}_u H_{uu}^{-1}\tilde{g}_u^T\right)^{-1}\left(\tilde{g}_u H_{uu}^{-1}H_{ux} - \tilde{g}_x\right)$$

Then the Riccati equation (4.27) can be written

$$\begin{aligned} -\dot{S} = {} & H_{xx} - H_{xu}H_{uu}^{-1}H_{ux} + \left(H_{xu}H_{uu}^{-1}\tilde{g}_u^T - \tilde{g}_x^T\right)\left(\frac{1}{\sigma} I + \tilde{g}_u H_{uu}^{-1}\tilde{g}_u^T\right)^{-1}\left(\tilde{g}_u H_{uu}^{-1}H_{ux} - \tilde{g}_x\right) \\ & + \left(f_x - f_u H_{uu}^{-1}H_{ux} + f_u H_{uu}^{-1}\tilde{g}_u^T\left(\frac{1}{\sigma} I + \tilde{g}_u H_{uu}^{-1}\tilde{g}_u^T\right)^{-1} \cdot \right. \\ & \left. \cdot \left(\tilde{g}_u H_{uu}^{-1}H_{ux} - \tilde{g}_x\right)\right)^T S + \\ & + S\left(f_x - f_u H_{uu}^{-1}H_{ux} + f_u H_{uu}^{-1}\tilde{g}_u^T\left(\frac{1}{\sigma} I + \tilde{g}_u H_{uu}^{-1}\tilde{g}_u^T\right)^{-1}\left(\tilde{g}_u H_{uu}^{-1}H_{ux} - \tilde{g}_x\right)\right) - \\ & - S\left(f_u H_{uu}^{-1}f_u^T - f_u H_{uu}^{-1}\tilde{g}_u^T\left(\frac{1}{\sigma} I + \tilde{g}_u H_{uu}^{-1}\tilde{g}_u^T\right)^{-1}\tilde{g}_u H_{uu}^{-1}f_u^T + \frac{1}{c} I\right)S \end{aligned}$$

$$S(T) = F_{xx}$$

Since the difference between the coefficients of the two Riccati equations can be made arbitrarily small by choosing c and σ large enough, it follows that if (4.28) has a solution over the whole interval, so has (4.27). □

Example 4.4. Study the linear-quadratic approximation of the problem in Example 4.1

$$\dot{x} = u$$

$$x(0) = 0$$

$$J = \frac{1}{2}\int_0^T (u^2 - x^2)\,dt$$

As was shown, $x = 0$, $u = 0$, is a solution for $T < \pi/2$. For $T > \pi/2$, there is no lower bound on J. Introduce the constraint

$$|u| \leq 1$$

or

$$g_1(u) = u - 1 \leq 0$$
$$g_2(u) = -u - 1 \leq 0$$

Due to the symmetry of the problem, there are two symmetrical solutions $(\bar{u},\bar{x})$ and $(\tilde{u},\tilde{x})$ with $\bar{u} = -\tilde{u}$, $\bar{x} = -\tilde{x}$. Study the solution with positive u. Then only g_1 is of interest

$$H = \frac{1}{2}u^2 - \frac{1}{2}x^2 + pu + \lambda(u-1)$$

The necessary conditions are

$$\dot{x} = 0 \qquad x(0) = 0$$
$$\dot{p} = x \qquad p(T) = 0$$

$$u + p + \lambda = 0$$

For $T > \pi/2$, they are satisfied by

$$u = \begin{cases} 1 & t \in [0,t_1] \\ -B \sin(t-T) & t \in [t_1,T] \end{cases}$$

$$x = \begin{cases} t & t \in [0,t_1] \\ B \cos(t-T) & t \in [t_1,T] \end{cases}$$

$$\lambda = \begin{cases} \frac{1}{2}(t_1^2-t^2) & t \in [0,t_1] \\ 0 & t \in [t_1,T] \end{cases}$$

where $B = -1/(\sin(t_1-T))$. The point t_1 is given by $1 + t_1 \cdot \tan(t_1-T) = 0$, which has a unique solution in $(T-\pi/2,T)$. The Riccati equation (4.28) is in this case

$$-\dot{S} = -1 - S^2 \qquad t \in [t_1,T]$$

$$-\dot{S} = -1 \qquad t \in [0,t_1]$$

$$S(T) = 0$$

This gives

$$S = -\tan(T-t) \qquad t \in [t_1,T]$$

$$S = t - t_1 - \tan(T-t_1) \qquad t \in [0,t_1]$$

and the second order sufficiency conditions are satisfied for any $T > 0$.

4.6. Summary

In the optimal control problem three types of constraints have been considered in this chapter. These are the differential equation, the terminal constraints and mixed state control inequality constraints. The constraints have been treated by the formation of a functional J, which is a direct generalization of the augmented Lagrangian in the finite dimensional case. J depends on a parameter c, which is the weight assigned to the part of J that is a quadratic form in the constraints. This parameter must be chosen large enough for J to have a local minimum at the solution to the original problem. It is shown that the magnitude of c that is required, is determined by a Riccati equation, which is closely connected to the ordinary sufficiency conditions. The theorems in this chapter only prove that J has a local minimum. For the finite dimensional case it was possible to prove global optimality under fairly general conditions. We remark that it is not straightforward to extend these results to the optimal control problem. The reason is that lemma 3.1 uses the fact that a closed bounded set is compact. This is not true in an infinite dimensional space.

The iterative method investigated in Section 4.4 is completely analogous to the Hestenes-Powell algorithm discussed in 3.4. It was shown in Chapter 3 that methods that update the multipliers after each iteration in x, are efficient. A natural continuation of the work presented in this chapter would therefore be to investigate the generalization of these methods to the optimal control problem.

4.7. Appendix.

Properties of the Riccati Equation.

Here some interesting results about the Riccati equation are collected. Most of them can be found in books about linear quadratic control theory, e.g. Brockett (1970) or Anderson and Moore (1971). Another useful reference is Mårtensson (1972).

We will write the Riccati equation in the form

$$-\dot{S}(t) = A^T(t)S(t) + S(t)A(t) + Q(t) - S(t)P(t)S(t)$$

$$S(T) = Q_0$$

where A, Q and P are matrices whose elements are continuous functions of t and Q_0, Q and P are symmetric. It follows from standard theorems for differential equations that $S(t)$ exists at least on a sufficiently small interval $t_0 \leq t \leq T$. Moreover, the only way in which S can fail to exist is by having some element which becomes unbounded. In what follows, $M \geq N$, where M and N are symmetric matrices, means that $M - N$ is nonnegative definite and $M > N$ means that $M - N$ is positive definite.

It is useful to rewrite the Riccati equation as an integral equation. Introduce the fundamental matrix $\phi(t,T)$ satisfying

$$\frac{d}{dt}\phi(t,T) = \left(A(t) - \frac{1}{2}P(t)S(t)\right)\phi(t,T)$$

$$\phi(T,T) = I$$

Then we have

$$S(t) = \int_t^T \phi^T(s,t)Q(s)\phi(s,t)dS + \phi^T(T,t)Q_0\phi(T,t)$$

Lemma A.1. For the Riccati equation

$$-\dot{S} = A^T S + SA + Q - SPS$$

let S_1 and S_2 be the solutions corresponding to $S(T) = Q_0^1$ and $S(T) = Q_0^2$ respectively. Then if $Q_0^2 \geq Q_0^1$ it follows that $S_2(t) \geq S_1(t)$ for all $t \in [t_0,T]$ where $[t_0,T]$ is an interval on which both solutions exist.

Proof. We have

$$-\frac{d}{dt}(S_2-S_1) = (A-PS_1)^T(S_2-S_1) + (S_2-S_1)(A-PS_1) -$$

$$- (S_2-S_1)P(S_2-S_1)$$

$$S_2(T) - S_1(T) = Q_0^2 - Q_0^1$$

Regarding this as a Riccati equation in $S_2 - S_1$ we get, using the integral equation representation above

$$S_2 - S_1 = \phi^T(T,t)(Q_0^2-Q_0^1)\phi(T,t)$$

where $\phi(t,T)$ now is the fundamental matrix corresponding to

$$A - PS_1 - \frac{1}{2}P(S_2-S_1)$$

Lemma A.2. Let S_1 and S_2 be the solutions of the Riccati equations

$$-\dot{S} = A^T S + SA + Q - SP_1S \qquad S(T) = Q_0$$

$$-\dot{S} = A^T S + SA + Q - SP_2S \qquad S(T) = Q_0$$

respectively. If $P_1 \geq P_2$ then $S_2(t) \geq S_1(t)$ for all $t \in [t_0,T]$, where $[t_0,T]$ is any interval on which both solutions exist.

Proof. We have

$$-\frac{d}{dt}(S_2-S_1) = (A-P_2S_1)^T(S_2-S_1) + (S_2-S_1)(A-P_2S_1) -$$

$$- (S_2-S_1)P_2(S_2-S_1) + S_1(P_1-P_2)S_1$$

$$S_2(T) - S_1(T) = 0$$

Using the integral equation form this can be written

$$S_2(t) - S_1(t) = \int_t^T \phi^T(s,t)S_1(P_1-P_2)S_1\phi(s,t)dS$$

Lemma A.3. Let S_1 and S_2 be the solutions of the Riccati equations

$$-\dot{S} = A^TS + SA + Q_1 - SPS \qquad S(T) = Q_0$$

and

$$-\dot{S} = A^TS + SA + Q_2 - SPS \qquad S(T) = Q_0$$

respectively. Then if $Q_2 \geq Q_1$ it follows that $S_2(t) \geq S_1(t)$ $t \in [t_0,T]$, where $[t_0,T]$ is any interval on which both solutions exist.

Proof. We have

$$S_2(t) - S_1(t) = \int_t^T \phi^T(s,t)(Q_2-Q_1)\phi(s,t)dS$$

where ϕ is the fundamental matrix corresponding to

$$A - PS_1 - \frac{1}{2}P(S_2-S_1)$$

We can now deduce the following result.

Lemma A.4. If $P > 0$ then there exists a continuous matrix $R(t)$ such that $S(t) \leq R(t)$ on any interval $[t_0,T]$ where S exists.

Proof. From Lemma A.2 it follows that $S(t) \leq R(t)$ where R is the solution to the linear differential equation.

$$-\dot{R} = A^TR + RA + Q \qquad R(T) = Q_0$$

From this lemma it follows that, to prove existence of $S(t)$ on some interval, all that is needed is a lower bound on S on that interval.

Lemma A.5. Let S be the solution of the Riccati equation

$$-\dot{S} = A^TS + SA + Q - SPS \qquad S(T) = Q_0$$

and assume that S exists on the interval $[t_0,T]$. Let $\tilde{S}$ be the solution to the Riccati equation where $\tilde{A}$, $\tilde{Q}$ and $\tilde{P}$ have replaced A, Q and P. Then there exists an $\varepsilon > 0$ such that $\tilde{S}$ also exists on $[t_0,T]$ if $\|\tilde{A} - A\| \leq \varepsilon$, $\|\tilde{Q} - Q\| \leq \varepsilon$ and $\|\tilde{P} - P\| \leq \varepsilon$.

Proof. Since the right hand side of the Riccati equation is a continuous function of S, A, Q and P the result follows from general results for nonlinear differential equations, see Coddington and Levinson (1955).

Two point boundary value problems.

A linear two point boundary value problem can be written

$$\dot{y} = V(t)y + f(t) \qquad My(0) + Ny(1) = c$$

where V, M, N are p p matrices and f,c and y are p vectors.

Definition. (Falb and de Jong (1969))
The set $\{V,M,N\}$ is called boundary compatible if (i) $V(t)$ is measurable with $||V(t)|| < m(t)$ for an integrable $m(t)$, and (ii) $\det(M + N\Phi(1,0)) \neq 0$ where $\Phi(t,s)$ is the fundamental matrix of $\dot{y} = V(t)y$.

$\{V,M,N\}$ is a boundary compatible set if and only if the linear two point boundary value problem has a solution for all f and c.

Lemma A.6. Let D be an open set in R^p and let I be an open set in R containing [0,1]. Suppose that (i) $F(y,t)$ is a map of D x I into D which is measurable in t for each fixed y and continuous in y for each fixed t; (ii) there is an integrable function $m(t)$ such that $||F(y,t)|| < m(t)$ on D x I; (iii) $g(y)$ och $h(y)$ are maps of D into D; and (iv) $\{V(t),M,N\}$ is a boundary compatible set. Then the boundary value problem

$$\dot{y} = F(y,t), \qquad g(y(0)) + h(y(1)) = c$$

has the equivalent representation

$$y(t) = H(t)\{c-g(y(0))-h(y(1))+My(0)+Ny(1)\} +$$

$$+ \int_0^1 G(t,s)\{F(y(s),s)-V(s)y(s)\}\,ds$$

where the Green's functions H(t) and G(t,s) are given by

$$H(t) = \Phi(t,0)\ (M+N\Phi(1,0))^{-1}$$

and

$$G(t,s) = \begin{cases} \Phi(t,0)(M+N\Phi(1,0))^{-1}M\Phi(0,s) & 0<s<t \\ -\Phi(t,0)(M+N\Phi(1,0))^{-1}N\Phi(1,s) & t<s<1 \end{cases}$$

where $\Phi(t,s)$ is the fundamental matrix of the linear system $y = V(t)y$.

Proof See Falb and de Jong (1969).

Contraction mapping theorem.

Lemma A.7. Let Y be a Banach space and let $S(y_0,r)$ be the closed sphere in Y with center y_0 and radius r. Let T map Y into Y and suppose that (i) T is defined on $D(y_0,r)$ and (ii) there are real numbers η and α with $\eta \geq 0$ and $0 \leq \alpha < 1$ such that

$$||T(y_0)-y_0|| \leq \eta$$

$$\sup_{u,v \in S} ||T(u)-T(v)||/||u-v|| \leq \alpha < 1$$

and $\eta/(1-\alpha) \leq r$

Then there is a unique fixed point of T in S.

Proof. See Falb and de Jong (1969).

4.8. References.

Anderson, B.D.O., Moore, J.B. (1971):
Linear Optimal Control, Prentice Hall, New Jersey.

Brockett, R.W. (1970):
Finite Dimensional Linear Systems, John Wiley and Sons, New York.

Bryson, A.E., Ho, Y.C. (1969):
Applied Optimal Control, Blaisdell Publishing Company.

Coddington, E.A., Levinson, N. (1955):
Theory of Ordinary Differential Equations, Mc Graw-Hill, New York.

Di Pillo, G., Grippo, L., Lampariello, F. (1974):
The Multiplier Method for Optimal Control Problems; Conference on Optimization Problems in Engineering and Economics, Naples, Italy, Dec. 16-20, 1974.

Falb, P.L., and De Jong, J.L. (1969):
Some Successive Approximation Methods in Control and Oscillation Theory, Academic Press, New York.

Gelfand, I.M., and Fomin , S.V. (1963):
Calculus of Variations, Prentice Hall, London.

Hestenes, M.R. (1947):
An Indirect Sufficiency Proof for the Problem of Bolza in Non-parametric Form, Transactions of the American Mathematical Society, Vol. 62, No. 3, pp. 509 - 535.

Hestenes, M.R. (1969):
Multiplier and Gradient Methods, Journal of Optimization Theory and Applications, Vol. 4, pp. 303 - 320.

Luenberger, D. (1968):
Optimization by Vector Space Methods.
John Wiley and Sons, New York

Mårtensson, K. (1972):
New Approaches to the Numerical Solution of Optimal Control Problems; Report 7206, Lund Institute of Technology, Division of Automatic Control.

Nahra, J.E. (1971):
Balance Function for the Optimal Control Problem, J. Optimization Theory Appl., Vol. 8, No. 1, pp. 35 - 48.

O'Doherty, R.J., and Pierson, B.L. (1974):
A Numerical Study of Augmented Penalty Function Algorithms for Terminally Constrained Optimal Control Problems, J. Optimization Theory Appl., Vol. 14, No. 4, pp. 393 - 403.

Rockafellar, R. T. (1973):
A Dual Approach to Solving Nonlinear Programming Problems by Unconstrained Optimization, Math. Programming, Vol. 5, pp. 354 - - 373.

Rupp, R.D. (1972a):
A Method for Solving a Quadratic Optimal Control Problem, J. Optimization Theory Appl., Vol. 9, No.4 , pp. 238 - 250.

Rupp, R.D. (1972b):
Approximation of the Classical Isoperimetric Problem, J. Optimization Theory Appl., Vol. 9, No.4 , pp. 251 - 264.

5. INTERACTIVE OPTIMIZATION OF DYNAMIC SYSTEMS WITH RESPECT TO PARAMETERS.

It was shown in Chapter 2 that the optimization of a control system with fixed structure gives rise to a finite dimensional optimization problem with constraints. In this chapter the problem will be considered more in detail.

A number of design methods proposed in the literature are discussed in Section 5.1 and it is shown that they all lead to a constrained optimization problem where the criterion and constraints have a particular structure.

A problem when using optimization methods is that it is often not clear which mathematical criterion is the appropriate one. Usually the designer has a vague intuitive idea of what he means by a "good" system, but it is difficult to translate it into a criterion and constraints that can be used by an optimization method. In practice, the optimization is therefore an iterative procedure. The designer has to try several different criteria and study the corresponding optimal systems before he is satisfied. It is also often necessary to investigate several different controller structures. To make the communication between designer and optimization program efficient, it is therefore very useful to have an interactive program. The implementation of the optimization as a part of an interactive simulation program is considered in Section 5.2.

Finally, the usefulness of the design approach considered in this chapter is demonstrated in a number of examples in Section 5.3.

5.1. Discussion of Criteria.

A large number of design criteria have been suggested in the literature. In this section some of the more common ones will be examined and it will be shown that they all lead to the following problem formulation:

$$\begin{aligned} &\text{minimize} \quad \varphi_0(p) \\ &\text{subject to } \varphi_i(p) \leq 0 \qquad i = 1, \ldots, m \end{aligned} \tag{5.1}$$

where p is a vector containing the controller parameters to be determined. The functions φ_i are given by

$$\varphi_i(p) = \int_0^T L\big(x(t),p,t\big)dt + M\big(x(T),p\big) \tag{5.2}$$

or

$$\varphi_i(p) = \max_{t_1 \leq t \leq t_2} L\big(x(t),p,t\big) \tag{5.3}$$

where x is the solution of

$$\begin{aligned} \dot{x}(t) &= f\big(x(t),p,t\big) \\ x(0) &= a(p) \end{aligned} \tag{5.4}$$

In some cases the original formulation may not involve any criterion to be minimized, only a set of inequalities to be satisfied

$$\varphi_i(p) \leq 0 \qquad i = 1, \ldots, m \tag{5.5}$$

Some or all of the differential equations in (5.4) may also be replaced by difference equations and the integral in (5.2) may be replaced by a sum. To simplify notation, the following discussion will be confined to the continuous time formulation of (5.2) - (5.4). The extension to discrete time is usually trivial.

Most criteria given in the literature are related to a specific type of input and initial condition. The input might be either a disturbance or the command signal for a servo-mechanism. For the specified input the criteria are then functionals defined on the state space trajectory of the system. If the input is stochastic in nature, the criteria are usually expectations of the functionals. Criteria related to deterministic inputs are first considered.

The quadratic criterion

$$J = \int_0^\infty e^2(t)dt \tag{5.6}$$

where e(t) is the difference between the actual output and the desired output is widely used for single output systems with deterministic input signals,see Newton, Gould and Kaiser (1957). This criterion can be approximated arbitrarily well by a criterion of the type in (5.2), if T is chosen large enough. The advantage of this criterion is that, for a linear system, J can be calculated as an explicit function of the coefficients describing the system. For low order systems the optimization problem can then be solved analytically. The criterion is often combined with a constraint of the type

$$\int_0^\infty z^2(t)dt \leq C$$

where z is the signal at some point in the system where it is essential to limit the magnitude. For example z might be the input of an amplifier that saturates for high signal levels.

Graham and Lathrop (1953) investigated different criteria for the design of servo mechanisms required to give an output reproducing the input. They found that the criterion (5.6), used for linear systems with a step as input, often give an optimal system that is poorly damped. The optimum is usually very flat;

parameter changes that are large enough to give a significant change in the response of the system, alter the value of the criterion very little. Therefore Graham and Lathrop also studied the criteria

$$J_1 = \int_0^\infty |e(t)|dt$$

$$J_2 = \int_0^\infty t|e(t)|dt$$

$$J_3 = \int_0^\infty t\, e^2(t)dt$$

J_2 was especially recommended because systems that are optimal with respect to J_2 usually also are "good" from many other points of view.

Martens and Larsen (1975) have suggested the following criterion for a step input.

$$J = \int_0^\infty \theta(t-\tau)e^2(t)dt \qquad (5.7)$$

where

$$\theta(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$$

This criterion does not penalize the error until after the time τ. Martens and Larsen show that with an appropriate choice of τ, the criterion gives a response that is very well damped.

For a linear system on state space form

$$\dot{x} = Ax + Bu$$
$$y = Cx$$

it is natural to use the quadratic criterion

$$J = \int_0^\infty (x^T Q_1 x + u^T Q_2 u)\,dt$$

which can be considered as a generalization of the criterion

$$\int_0^\infty e^2(t)\,dt$$

discussed above.

If all the state variables are available for the controller, it is well known, see e.g. Anderson and Moore (1971), that the optimal control is given by

$$u = Kx$$

where

$$K = - Q_2^{-1} B^T S$$

and S is given by the solution to the stationary Riccati equation

$$A^T S + SA + Q_1 - SBQ_2^{-1}B^T S = 0$$

It is interesting to note that the control law is optimal for all choices of initial conditions x(0).

Since the solution in this case has a special structure that can be used in the computations, there is no point in using a general nonlinear optimization routine.

However, the fact that a feedback from all state variables is

required can be a disadvantage in some cases. Therefore it is also interesting to study output feedback

$$u = Ky$$

The problem then becomes:

Minimize

$$J(K,x_0) = \int_0^\infty x^T(t)(Q_1 + C^T K^T Q_2 KC)x(t)dt \tag{5.8}$$

where x satisfies

$$\dot{x}(t) = (A + BKC)x(t)$$

In general the optimal K depends on what value of x(0) is chosen. One way of dealing with this problem is to choose a specific value of x(0), e.g. a value corresponding to the most common disturbance. A different approach is to use a weighted value of the initial states. The criterion can be written

$$J(K,x(0)) = x(0)^T S(K)x(0)$$

where

$$S(K) = \int_0^\infty e^{(A+BKC)^T t}(Q_1 + C^T K^T Q_2 KC)e^{(A+BKC)t}dt$$

If the initial states are assumed to be distributed with equal weight over the unit sphere, the criterion becomes

$$J_1 = \text{tr } S(K)$$

see Levine and Athans (1970).

If different values of x(0) are given different weights, the criterion becomes

$$J_2 = \text{tr } S(K)R \tag{5.9}$$

where R is a weighting matrix. This form of the criterion can also be derived from a stochastic point of view, see Mårtensson (1970). Let x(0) be a stochastic variable. Then, if E denotes expectation,

$$E\, x(0)^T S(K) x(0) = \text{tr } S(K)R$$

where

$$R = E\, x(0)x(0)^T$$

S(K) can be computed in several different ways. It is the solution of the algebraic equation

$$(A+BKC)^T S(K) + S(K)(A+BKC) + Q_1 + C^T K^T Q_2 KC = 0 \tag{5.10}$$

It can also be computed from a matrix differential equation

$$\dot{X} = (A+BKC)X$$

$$X(0) = I$$

$$S(K) = \int_0^\infty X^T(t)(Q_1 + C^T K^T Q_2 KC)X(t)dt$$

which is of the form (5.2). A discussion and comparison of different ways of calculating S can be found in Hagander (1972).

An important class of systems consists of servo-mechanism that are required to produce an output that is a reproduction of the input. The system behaviour is then often specified for step and

ramp command signals. The asymptotic behaviour is defined by the error coefficients e_0, e_1, ... If e(t) is the error, i.e. the difference between the command signal and the output, and the command can be expressed as a polynomial in t then

$$e(t) \to e_0 u + e_1 \dot{u} + e_2 \ddot{u} + \ldots \qquad \text{as } t \to \infty$$

The transient behaviour can be fairly well described by the following data for the step response, see Fig. 5.1. The output is denoted y(t). It is assumed for simplicity that the amplitude of the step is 1.

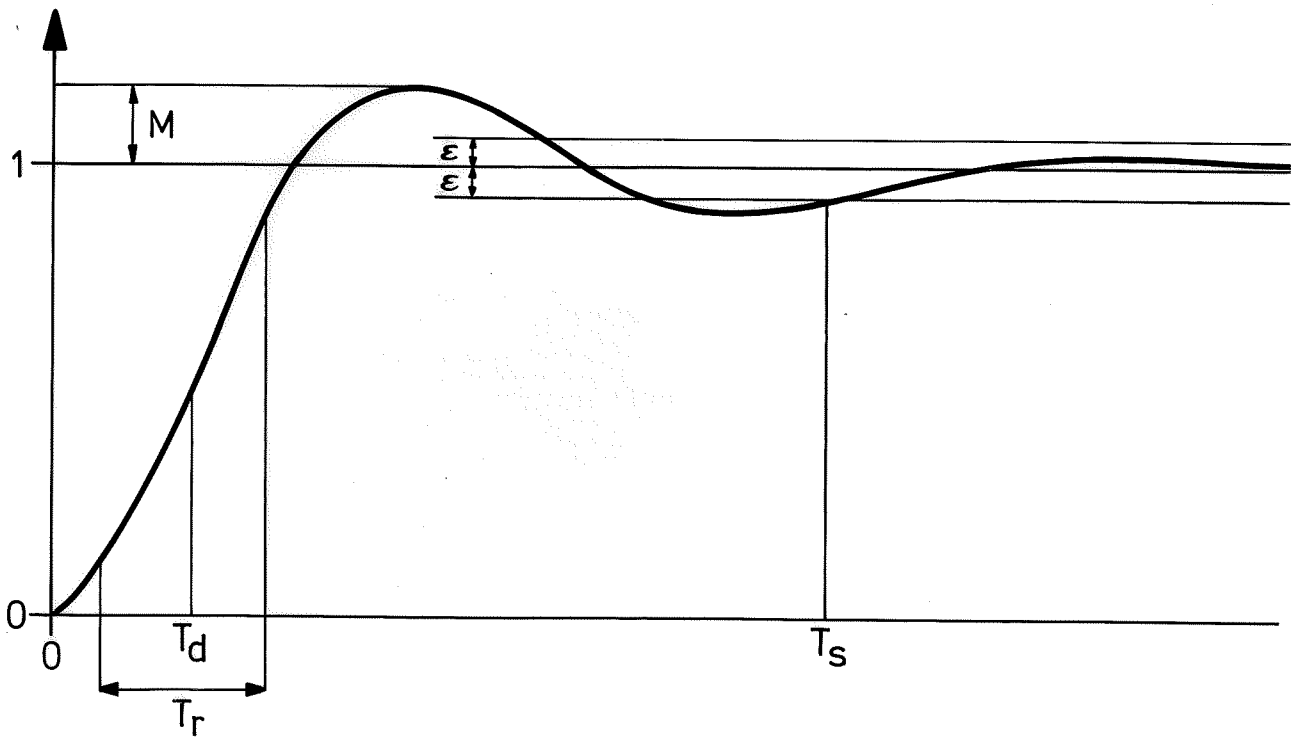


Fig. 5.1.

o The settling time T_s is the smallest value of time for which

$$|y(t) - y_{ref}| \leq \varepsilon\, y_{ref} \qquad \text{all } t \geq T_s$$

The value of ε is usually 0.01 or 0.05.

- The delay time T_d is the time it takes for the step response to reach half its final value.

- The rise time T_r is the time it takes for the step response to go from 0.1 to 0.9 of its final value.

- The overshoot M is $(\max y(t) - y_{ref})/y_{ref}$.

Usually these specifications for the step response are combined with restrictions on the error coefficients.

We will now study how these quantities can be expressed in the form of (5.2) or (5.3). Since the error coefficients can be expressed explicitly in p, they are a special case of (5.2). If T is sufficiently large the overshoot can be written

$$M_0 = \max_{0 \le t \le T} y(t) - 1$$

which is of the same type as (5.3). The settling time, T_s, is more difficult to handle. T_s itself is often a discontinuous function of p and is therefore not well suited for numerical calculations. A specification of the type

$$T_s \le t_1$$

can, however, be rewritten

$$\max_{t \ge t_1} y(t) \le 1 + \varepsilon$$

$$\min_{t \ge t_1} y(t) \ge 1 - \varepsilon$$

and these specifications are of the type (5.3).

The delay time T_d and the rise time T_r can be expressed in the following way

$$T_d = \int_0^T \theta(0.5-y(t))dt$$

$$T_r = \int_0^T \{\theta(y(t) - 0.1) - \theta(y(t) - 0.9)\}dt$$

where θ has the property

$$\theta(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$$

and T is sufficiently large. The use of these formula assumes that y(t) remains above the values 0.5 and 0.9 respectively once it has passed them. For very oscillative systems the expressions will therefore give an incorrect value. Since the specifications are likely to define a system whose oscillations are fairly well damped, this will, however, usually not cause too much trouble.

A different type of design criteria are based on disturbance models using stochastic processes. The criterion is then usually to minimize the average deviation from the desired output. If the system is linear we can write

$$E(s) = G(s)V(s)$$

where E and V are the Laplace transforms of the error and the disturbance respectively. If the performance of the system is defined as the mean of the squared error

$$J = Ee^2$$

and if the disturbance is a weakly stationary stochastic process, then

$$J = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} G(s)G(-s)\phi_V(s)ds$$

where ϕ_V is the spectral density of V. If ϕ_V is a rational function, J can be expressed explicitly in the coefficients of G and ϕ_V, see Åström (1970). This means that the criterion is a special case of (5.2). An alternative way of evaluating the criterion is to compute

$$J = \int_0^\infty e^2(t)dt$$

for the deterministic input

$$v(t) = h(t)$$

where h is impulse response of the system $G_V(s)$, satisfying

$$G_V(s)G_V(-s) = \phi_V(s)$$

In this way the stochastic has been converted to a deterministic problem of the type considered before.

An application of parametric optimization of a system subject to stochastic disturbances is given by Vandierendonk (1972), who describes the design of an autopilot for airplanes. The system is of the form

$$\begin{aligned} \dot{x} &= Fx + G_1u + G_2\eta \\ r &= Hx + Du \\ y &= Mx \end{aligned} \qquad (5.11)$$

where x is the state variables, u the input, y the measurements and η white noise. The vector r contains physical quantities for which values close to zero are desired. The control law has

the form

$$u = Ky$$

and the elements in K have to be determined. The criterion to be minimized is

$$J = Er^T Qr$$

when the system is in steady state. The criterion can be rewritten

$$J = tr(H+DKM)^T Q(H+DKM)P \tag{5.12}$$

where $P = Exx^T$. P can be calculated from

$$(F+G_1KM)P + P(F+G_1KM)^T + G_2G_2^T = 0 \tag{5.13}$$

The criterion is of the same form as the deterministic criterion (5.9), (5.10).

As shown above, minimization of the average value of a quadratic criterion for a linear system leads to a minimization problem equivalent to a deterministic one.

In a more general case, where the system is described by nonlinear differential equations, the computation of the criterion involves the solution of a partial differential equation, the Fokker-Planck equation. Since the computational work involved when doing this is usually excessive, one might try a simpler approach. One way is to optimize the regulator parameter for a particular realization of the stochastic process describing the disturbance and replace expectations with time averages. The problem is then converted to a problem of the deterministic type. If the realization is long enough one can expect that the solution obtained is close to the solution of the original problem.

So far it has been shown that, for a wide variety of proposed design criteria, the functions φ_i are of the type described by equations (5.2) or (5.3). As remarked in (5.5) the problem might not be an optimization problem, but a problem of the form

Find a p such that

$$\varphi_i(p) \leq 0 \qquad i = 1, \ldots, m \tag{5.14}$$

A natural way of solving this problem, however, is to convert it into an optimization problem. One way of doing this is to choose one of the functions φ_i as a criterion. This gives the problem

$$\begin{aligned} &\text{minimize} \quad \varphi_k(p) \\ &\text{subject to } \varphi_i(p) \leq 0 \qquad i = 1, \ldots, m, \; i \neq k \end{aligned} \tag{5.15}$$

Let the solution to this optimization problem be denoted $\bar{p}$. If $\varphi_k(\bar{p}) \leq 0$ the original problem is solved. If on the other hand $\varphi_k(\bar{p}) > 0$, then the original problem is impossible to solve. In principle the optimization therefore makes it possible to determine when a set of specifications are impossible to satisfy. One has to be careful in practice, however, because most numerical methods give only local minima, and it is difficult to ascertain if a result is actually the global minimum.

A different approach is to form the loss function

$$F(p) = \sum_{i=1}^{m} \varphi_i(p)_+^2 \tag{5.16}$$

where

$$\varphi_i(p)_+ = \begin{cases} \varphi_i(p) & \text{if } \varphi_i(p) \geq 0 \\ 0 & \text{if } \varphi_i(p) < 0 \end{cases}$$

This loss function has been suggested by Zakian (1973). F(p) has the property that all points p satisfying (5.5) are global minima to F(p) with F(p) = 0. The problem is therefore solved by the minimization of F. Since the minimum of F is in general not unique, the solution which is obtained will depend on the starting point of the numerical optimization algorithm. A third method is to solve the problem

$$\text{minimize } \psi(p)$$

$$\text{subject to } \varphi_i(p) \leq 0 \qquad i = 1, \ldots, m \qquad (5.17)$$

where ψ is an arbitrary function. If ψ = constant is chosen and the augmented Lagrangian of Chapter 3 is used with u = 0 as starting value for the multipliers, this method will be equivalent to the use of (5.16).

5.2. Implementation.

From the standard form of the design problem described in the previous section, (5.1) - (5.5), it follows that a program that can solve this problem must contain the following numerical algorithms.

- An algorithm to solve differential equations.
- An algorithm for constrained optimization.

As mentioned in the introduction to this chapter, it is also a great advantage if the program can be used interactively. A possibility to plot the optimal response is also desirable.

The desired facilities have been obtained by the inclusion of an optimization routine in the interactive simulation program SIMNON. This program is described in Elmqvist (1975). Some of its important properties are:

- Systems consisting of differential and difference equations can be handled at the same time.

- The system can be described by several separate subsystems that are connected together.

- A system can be described either in FORTRAN or in a special simulation language.

- It is easy to change both parameter values and the structure of the system.

- The response of the system can be plotted on a display.

Algorithms for constrained minimization are described in Chapter 3, where it is shown that methods based on the so-called augmented Lagrangian are quite promising. The methods discussed in Chapter 3 use derivatives of the objective function and constraints. In this application, however, it is more convenient to use algorithms that only need function values. It is possible to compute derivatives of most of the criteria discussed in 5.1, but it involves more work for the user, who has to supply more information.

For unconstrained problem it is known, see Fletcher (1972), that quasi-Newton methods, where the analytic gradient is replaced by a difference approximation, are quite efficient. It is therefore natural to use one of the algorithms of Section 5, and use difference approximations where gradients occur. The natural choice is then MINGRB since, in that algorithm, gradients are not used in the updating of the multipliers. An algorithm called OPTA, based on the algorithm MINGRB and with gradients of the augmented Lagrangian computed by difference approximations has therefore been included in SIMNON.

The procedure when designing a control system using the optimization facility in SIMNON is then as follows. The user writes

down a description of the system and the criteria either in FORTRAN or in the simulation language of SIMNON. The user then gives initial values to the regulator parameters that are to be optimized. When the optimization is started, the user can watch the response of the system plotted on a display, and observe what progress is made by the optimization routine. The fact that the optimization is done interactively has the following advantages.

- Choice of step length for difference approximation of derivatives. This choice can have great effect on the performance of the optimization routine and is often difficult to make. When using an interactive program it is easy to do a few trial runs to see that the chosen step length gives reasonable changes in the performance index.

- Choice of starting point. It is easy to make some preliminary experiments to find a reasonable starting point.

- Choice of time interval. In the integral in equation (5.2), T should usually be large enough to cover all the transient behaviour of the system. This is easily checked because the system response is plotted.

- Stopping criterion. It is difficult to find good stopping criteria for optimization problems where only function values are available. The interaction helps in two ways. The user can restart the algorithm if he is suspicious of the result or he can stop the algorithm before the exact optimum is found if he is already satisfied with the performance.

- Choice of controller structure. It is easy to try out the effects of different controller structures.

- Choice of criteria. It is easy to see how changes in the criteria affect the performance.

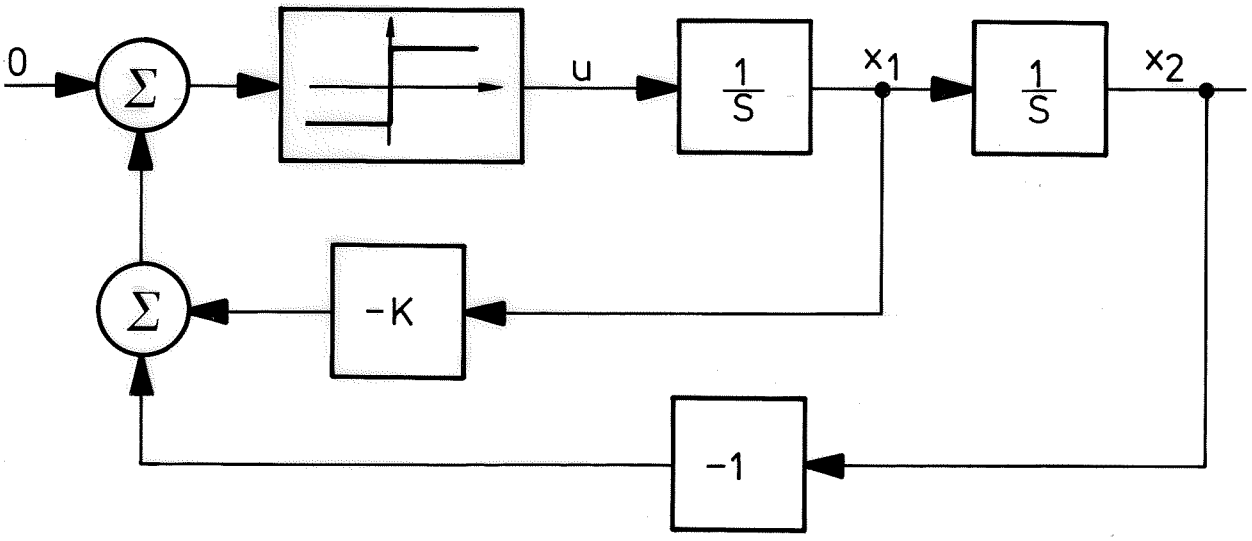
- Choice of weights for the constraints. The parameter c in equation (3.2) must be chosen large enough, as discussed in Chapter 3. With the interactive program, too small a value of c can quickly be detected and a more suitable value inserted.

- "Helping the algorithm". The user can immediately observe if the algorithm is making slow progress. He can then use his knowledge of the problem in trying to figure out the reason, e.g. bad scaling of variables, bad choice of step length for the difference approximation, bad starting point. The required changes can then be made immediately.

5.3. Examples.

The first example is a simple one, containing only one parameter to be optimized.

Example 5.1. Double integrator controlled by relay.



This problem is studied by Fuller (1967). The problem is to determine K, i.e. the slope of the switching line of the relay. The system behaviour is measured by the integral of the squared

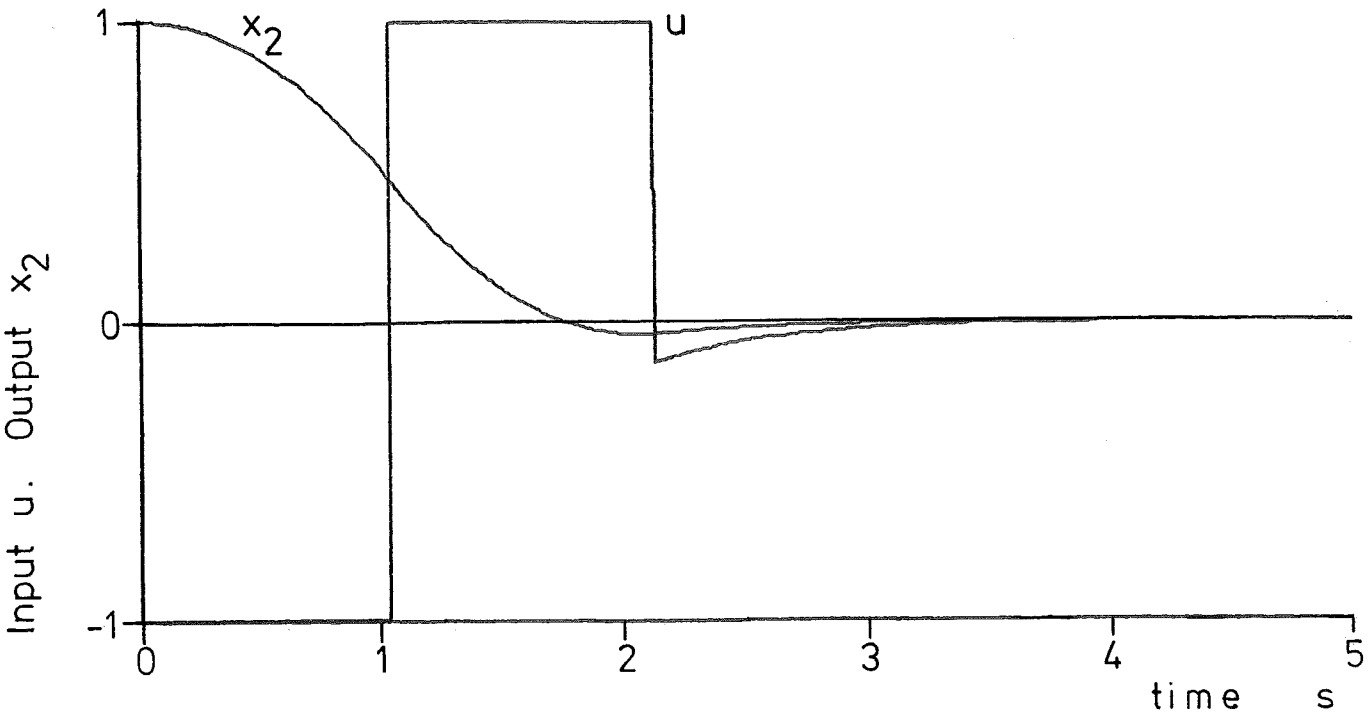


Fig. 5.2. Response of optimal relay control system for $x(0)=(0,1)^T$

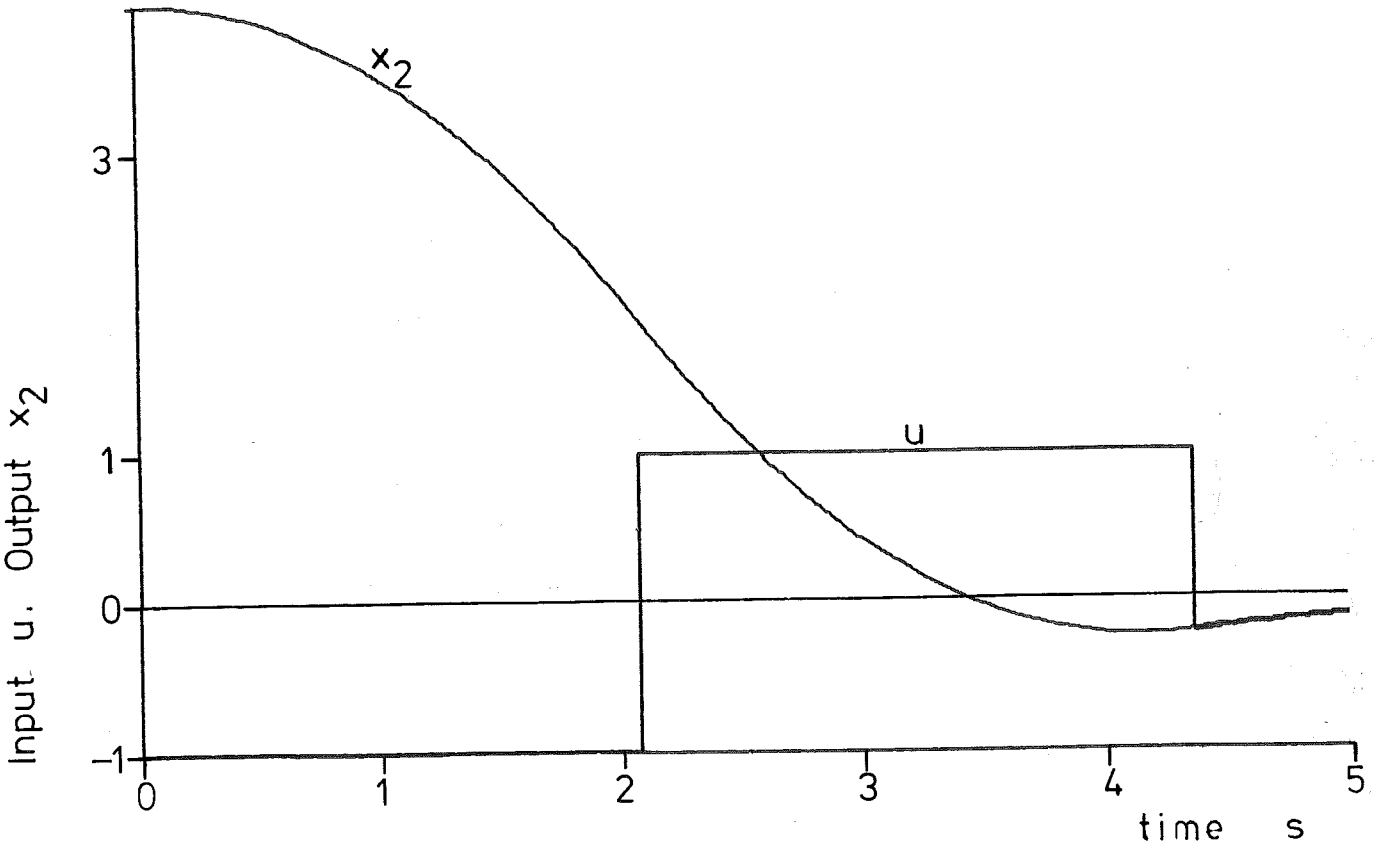


Fig. 5.3. Response of optimal relay control system for $x(0)=(0,4)^T$

position error

$$J = \int_0^\infty x_2^2(t)\,dt$$

for initial conditions

$x_1(0) = 0$

$x_2(0) = a$

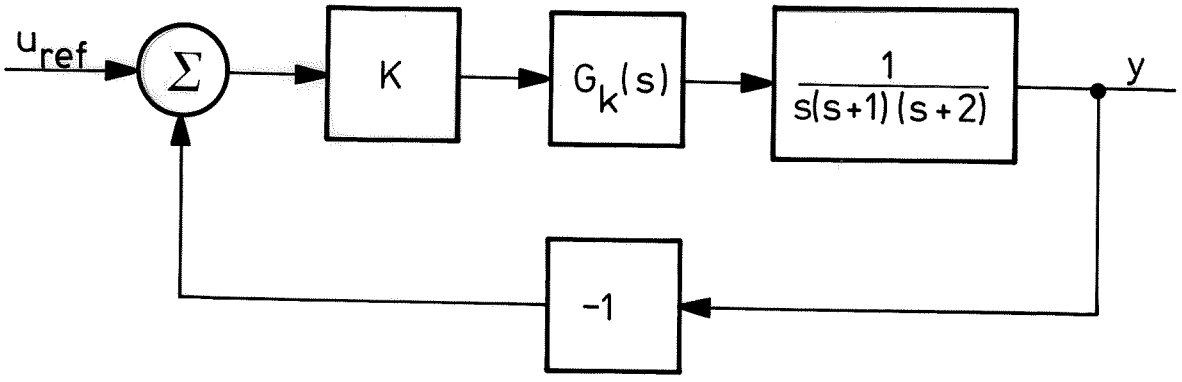
The following results were obtained

		theoretical values
a = 1	K = 0.4652	K = 0.4610
	J = 0.7645	J = 0.7643
a = 4	K = 0.9224	K = 0.9219
	J = 24.46	J = 24.46

The theoretical values are computed from $K = 0.46096|a|^{1/2}$ and $J = 0.7643|a|^{5/2}$, see Fuller (1967). The response of the optimal system is given in Figures 5.2 and 5.3. The SIMNON commands needed to solve this problem are presented in the appendix.

Example 5.2. Lead compensation.

This is an example of a classical design problem for feed back systems. The following system is given.



With $G_k(s) = 1$ and $K = 1$ the system has the step response shown in Fig. 5.4. The performance can be described by the following data:

Settling time (5%)	$T_s = 8.6s$
Overshoot	$M = 15\%$
$e_0 = 0$	
$e_1 = 2s$	

where e_0 and e_1 are the error coefficients.

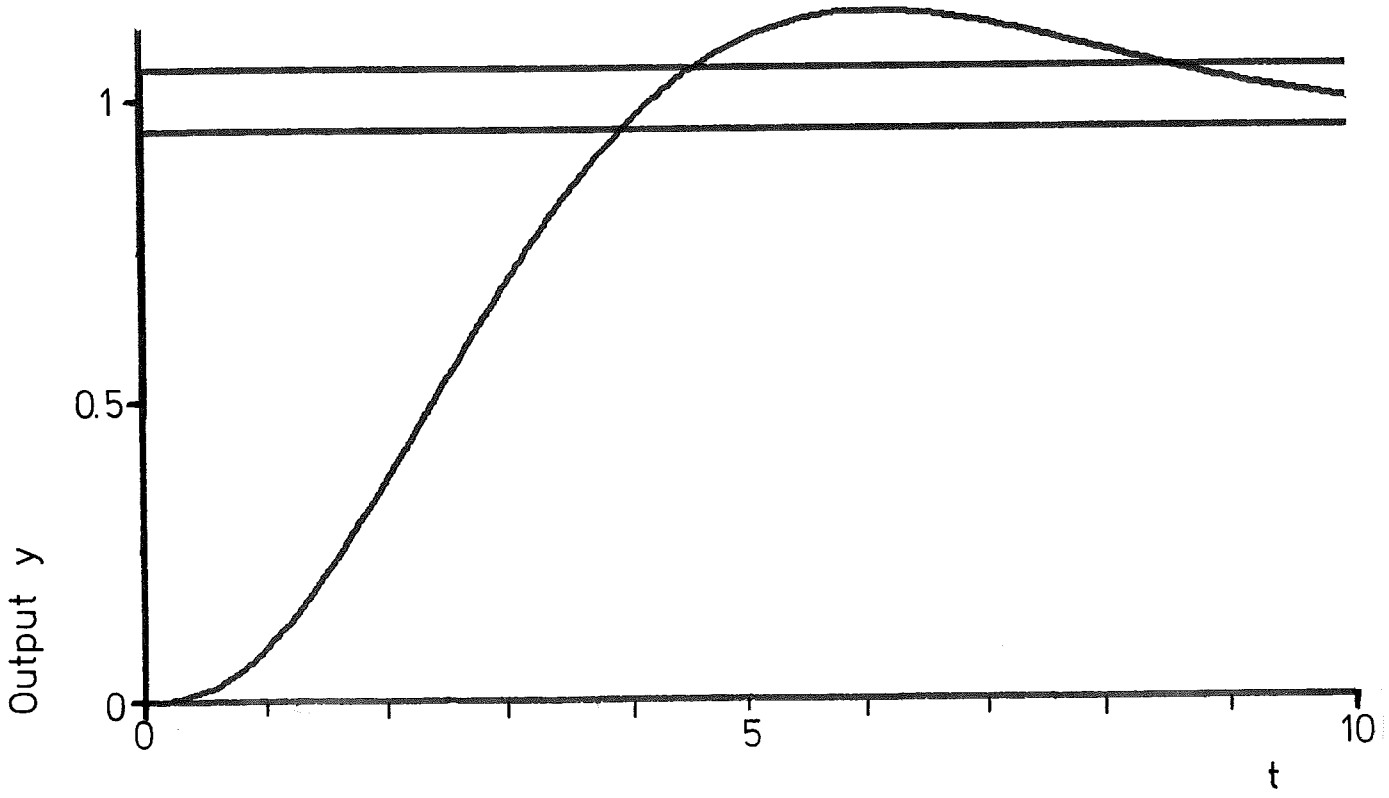


Fig. 5.4. Step response for $G_k = 1$ and $K = 1$.

It is desired to make the system about three times as fast without sacrificing other properties. The desired performance is then

$T_s \leq 2.9s$

$M_0 \leq 15\%$

$e_0 = 0$

$e_1 \leq 2s$

The classical way of achieving this is to use a lead network.

$$G_k(s) = \frac{s + b}{s + bN}$$

There are then three coefficients to determine: K, N and b. The error coefficient e_0 is automatically zero because of the integration in the open system. The coefficient e_1 can be calculated analytically

$$e_1 = \frac{2N}{K}$$

The condition $T_s \leq 2.9$ can be rewritten as follows:

$$\max_{2.9 \leq t} y(t) \leq 1.05$$

$$\min_{2.9 \leq t} y(t) \geq 0.95$$

The overshoot condition can be written

$$\max_{0 \leq t} y(t) \leq 1.15$$

One possible way of attacking the design problem is to formulate the following optimization problem.

$$\text{Minimize } \max_{0 \leq t} y(t)$$

under the constraints

$$\max_{2.9 \leq t} \; y(t) - 1.05 \leq 0$$

$$0.95 - \max_{2.9 \leq t} \; y(t) \leq 0$$

$$N - K \leq 0$$

The resulting parameters are

$$K = 4.83 \qquad N = 4.37 \qquad b = 0.54$$

The optimal step response is shown in Fig. 5.5.

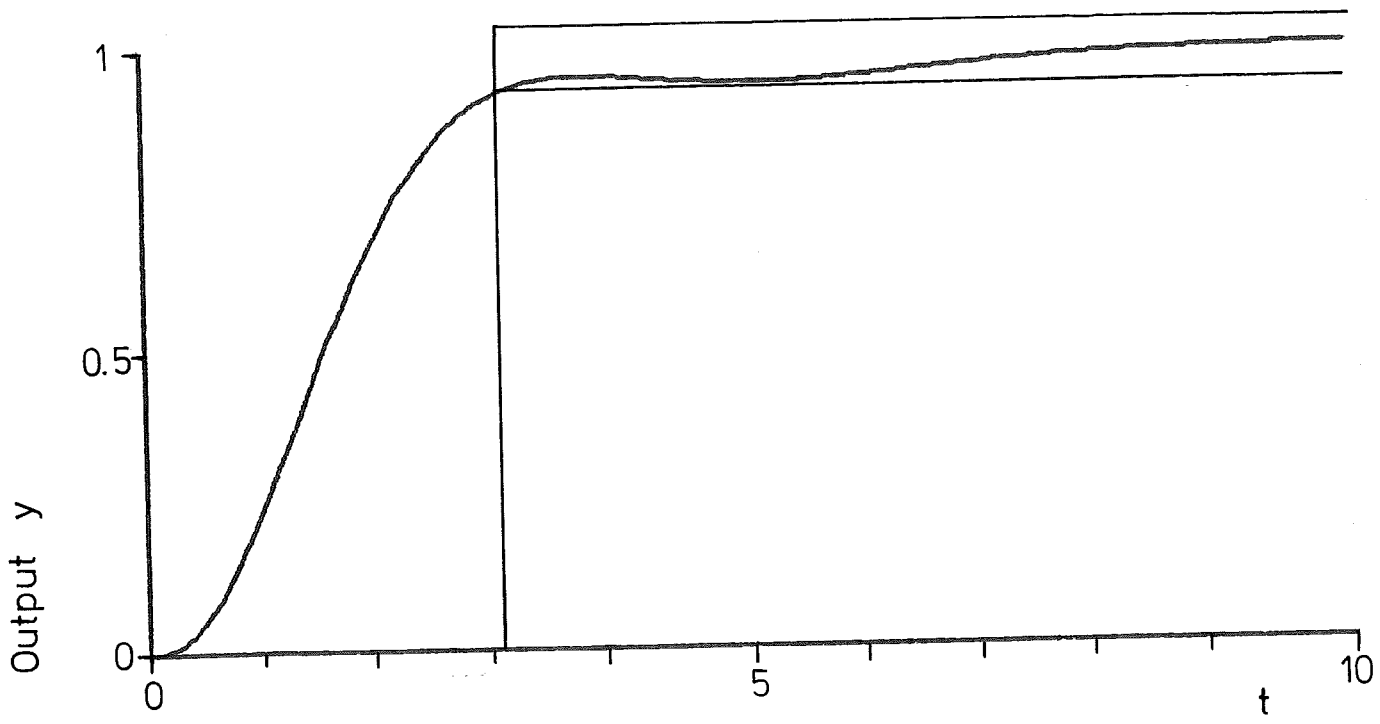


Fig. 5.5. Step response for K = 4.83, N = 4.37 and b = 0.54.

Another way to convert the design problem to an optimization problem is to use the criterion, see (5.16)

$$J = \left(\max\, y(t) - 1.17\right)_+^2 + \left(\max_{2.9\leq t}\, y(t) - 1.05\right)_+^2 +$$

$$+ \left(0.95 - \max_{2.9\leq t}\, y(t)\right)_+^2 + (N-K)_+^2$$

This criterion does not give a unique value to K, N and b because there is a whole area in the parameter space where J is identically zero. The choice of K, N and b therefore depends on the starting point for the minimization. For the starting point

$$K = 4.0 \qquad N = 4.0 \qquad b = 0.5$$

the result is

$$K = 4.19 \qquad N = 3.72 \qquad b = 0.54$$

with the response shown in Fig. 5.6.

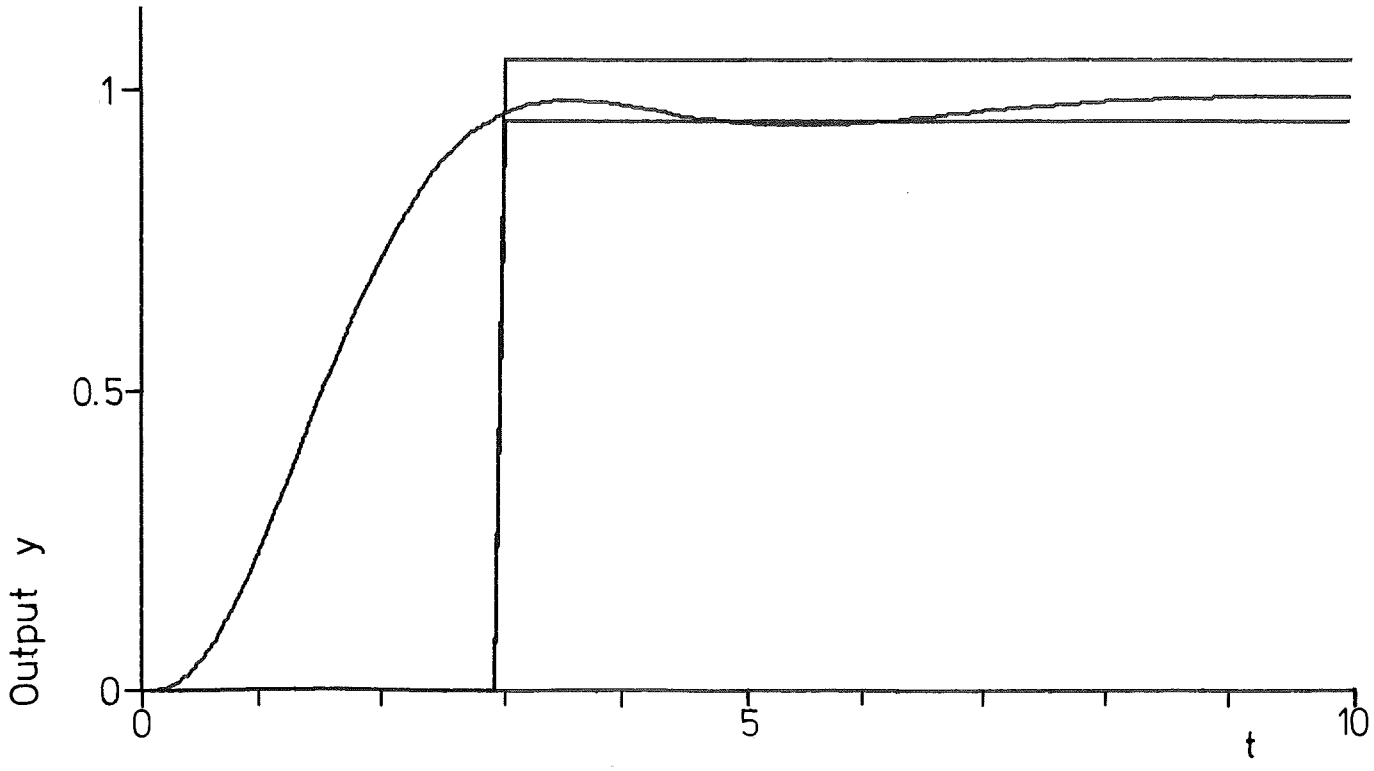


Fig. 5.6. Step response for K = 4.19, N = 3.72 and b = 0.54.

If the starting point

K = 1.0 N = 1.0 b = 1.0

is chosen instead, the result is

K = 3.08 N = 2.56 b = 0.64

with the response of Fig. 5.7.

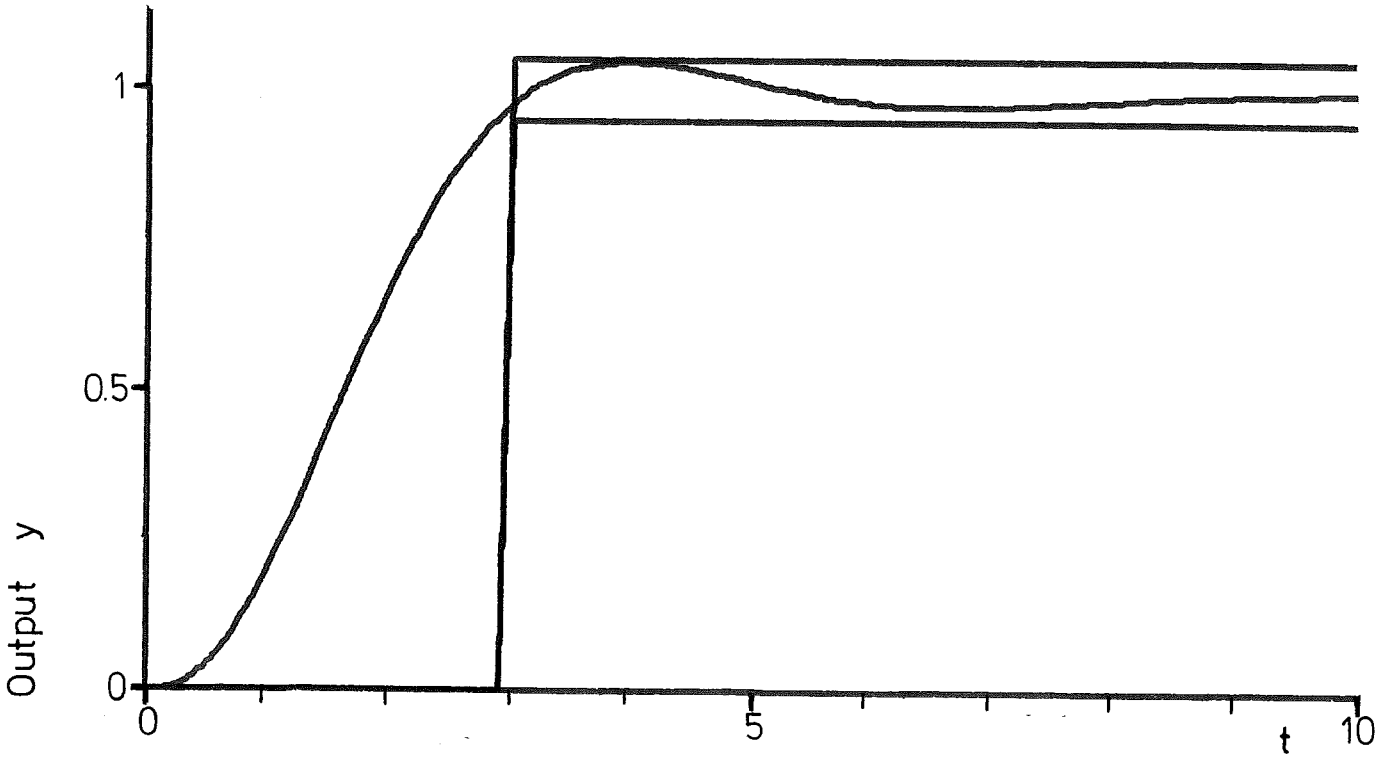
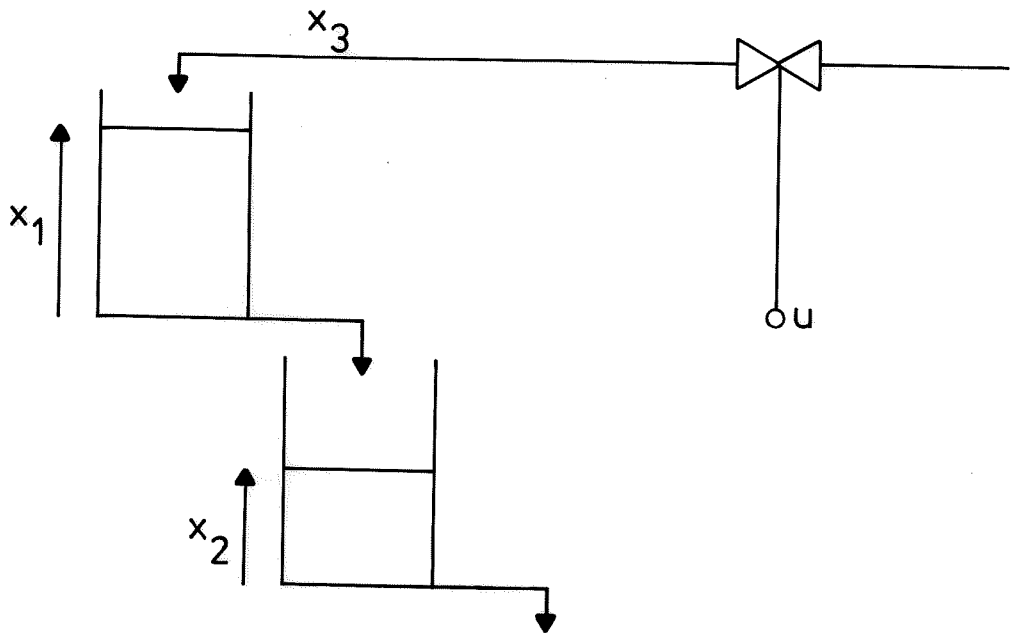


Fig. 5.7. Step response for K = 3.08, n = 2.56 and b = 0.64.

Which one of the step responses shown in figs. 5.5-5.7 that is to be preferred depends on the particular application. If no overshoot at all is tolerated, the responses in fig. 5.5 and fig. 5.6 can be used. For manual control, e.g. in an aircraft, the response in fig. 5.7 would probably be preferred because it gives the operator a more distinct feeling of the system response.

Example 5.3. Tuning of PI- and PID-controller.

The system consists of two identical tanks and a valve controlling the input flow as shown below.



It is assumed that the transfer function from the input of the valve, u, to the input flow, x_3, is 1/(1+Ts). The equations describing the system are

$$\dot{x}_1 = -\frac{a}{A}\sqrt{2gx_1} + \frac{x_3}{A}$$

$$\dot{x}_2 = \frac{a}{A}\sqrt{2gx_1} - \frac{a}{A}\sqrt{2gx_2}$$

$$\dot{x}_3 = -\frac{1}{T}x_3 + \frac{1}{T}u$$

where a is the effective area of the tank outlet and A the cross section of the tank. The control is done by a PI-controller using the measurement of the level in the lower tank.

$$u(t) = K\left(e(t) + \frac{1}{T_I}\int^{t} e(\tau)d\tau\right)$$

where $e = r - x_2$ and r is the desired value. The constants K and T_I are chosen to give a good response when r is changed suddenly at a time when the system is in a steady state. As measure of a good response a criterion of the type proposed in Martens and Larsen (1975) is used.

$$J = \int_{t_0+t_1}^{\infty} |e(t)|dt$$

where t_0 is the time when r is changed and t_1 is chosen to indicate how quickly the system is required to response. With the following values of the constants

$A = 19.6\ cm^2$ $\qquad$ $r = 30\ cm$

$a = 0.43\ cm^2$ $\qquad$ $x_1(0) = 20\ cm$

$T = 3s$ $\qquad$ $x_2(0) = 20\ cm$

$t_1 = 30s$ $\qquad$ $x_3(0) = 85.18$

the result of the optimization is

$K = 1.88$ $\qquad$ $T_I = 18.7$

with the response shown in Fig. 5.8.

For a PID-controller

$$u(t) = K\left(e(t) + \frac{1}{T_I}\int^{t} e(\tau)d\tau + T_D\dot{e}(t)\right)$$

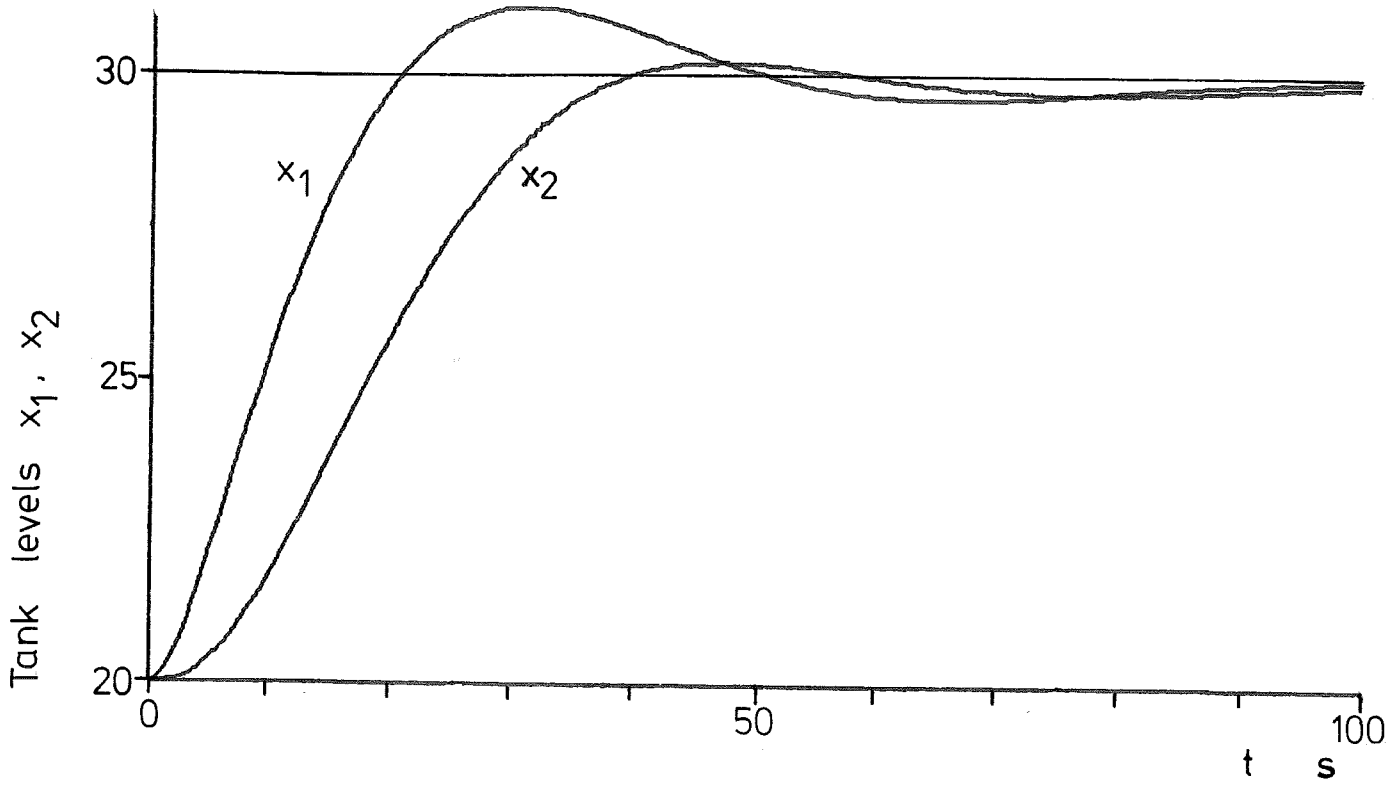


Fig. 5.8. Optimal response of tank system with PI-controller.

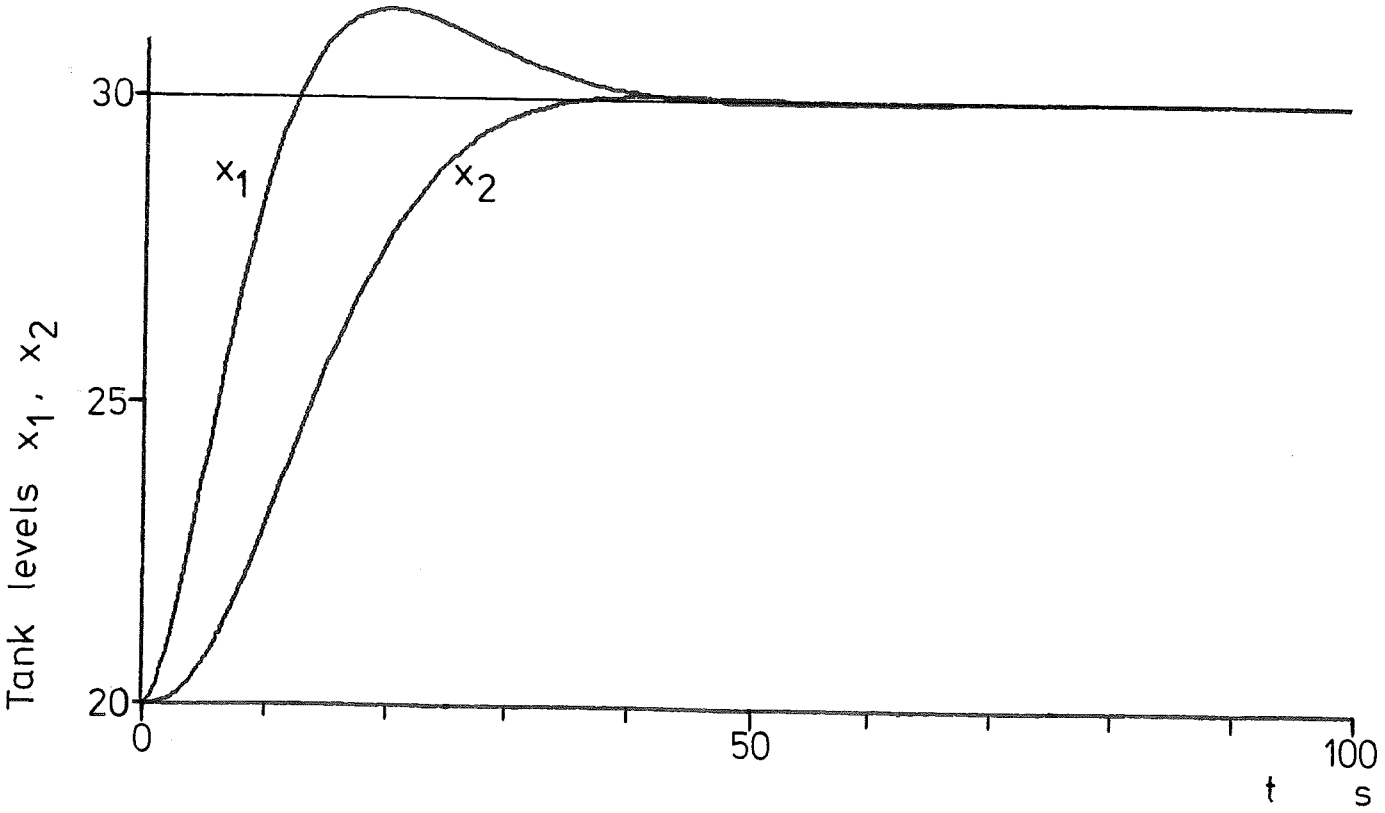


Fig. 5.9. Optimal response of tank system with PID-controller.

the optimal choice of K, T_I and T_D, using the same criterion as for the PI-controller, is

$$K = 3.96 \qquad T_I = 3.05 \qquad T_D = 5.37$$

The optimal response is shown in Fig. 5.9.

Example 5.4. Min-max criterion.

In Example 5.1 it was shown that the optimal slope of the switching line was dependant on the initial value $x_2(0)$. If K-values computed for a large $x_2(0)$ are used for small $x_2(0)$-values, the system becomes slow, see fig. 5.10.

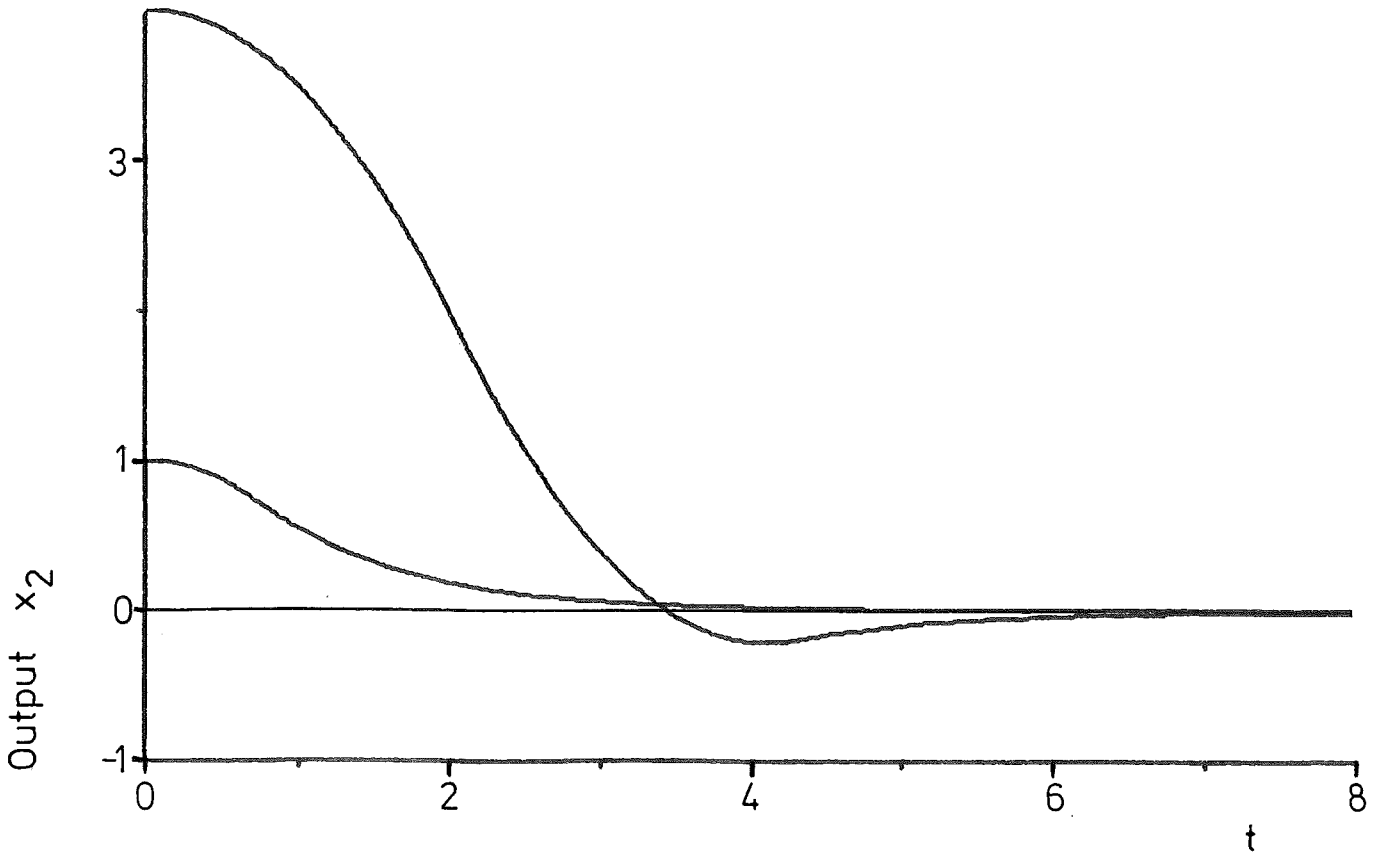


Fig. 5.10. Response of relay control system for K = 0.92.

If, on the other hand, a K-value computed for small $x_2(0)$-values is used when $x_2(0)$ is large, the system becomes very oscillative, see fig. 5.11.

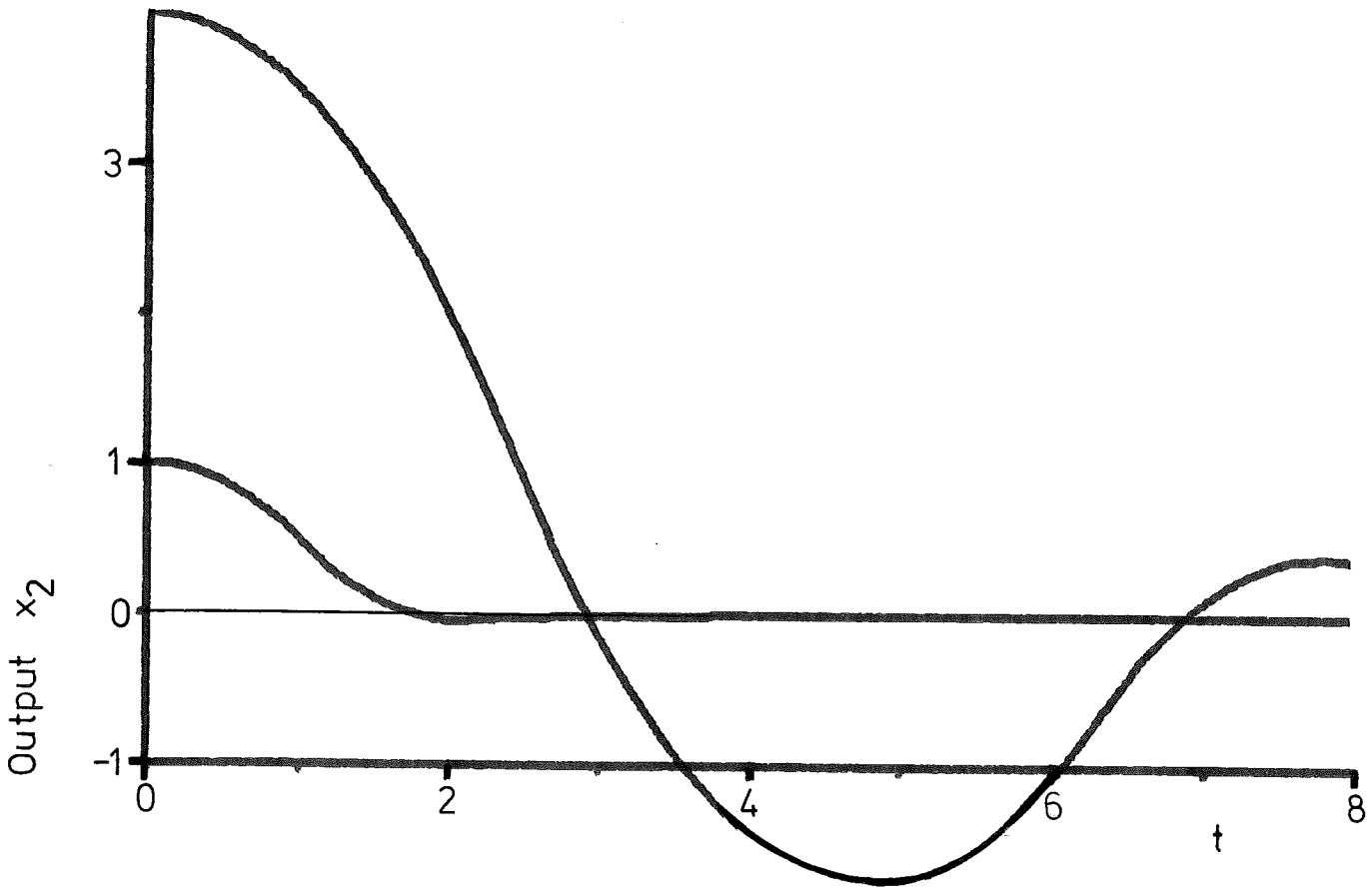


Fig. 5.11. Response of relay control system for K = 0.46.

One way of getting a compromise value for K is to use the following criterion

$$J(K) = \max\left(J_1(K), J_2(K)\right)$$

where

$$J_1(K) = \int_0^\infty x_2^2(t)\,dt \qquad \text{for } x_2(0) = a_1$$

$$J_2(K) = \beta \int_0^\infty x_2^2(t)\,dt \qquad \text{for } x_2(0) = a_2$$

and β is a weighting factor, which is necessary, since without it the loss corresponding to the large $x_2(0)$ would always dominate. To minimize J directly can lead to problems for the minimization routine, since J usually has a discontinuous derivative at the minimum. Therefore the problem is reformulated, using an auxilliary variable z.

Minimize z

under constraints $J_1(K) \leq z$

$J_2(K) \leq z$

The problem has been solved for $a_1 = 1$, $a_2 = 4$ and $\beta = 32$. (This is the natural weight, considering the theoretical results in Ex. 1.)

The result is

$K = 0.71$ $\qquad\qquad$ $J_1 = J_2 = 0.80$

The multipliers corresponding to the two constraints are

$\lambda_1 = 0.62$ $\qquad\qquad$ $\lambda_2 = 0.40$

The response is shown in Fig. 5.12.

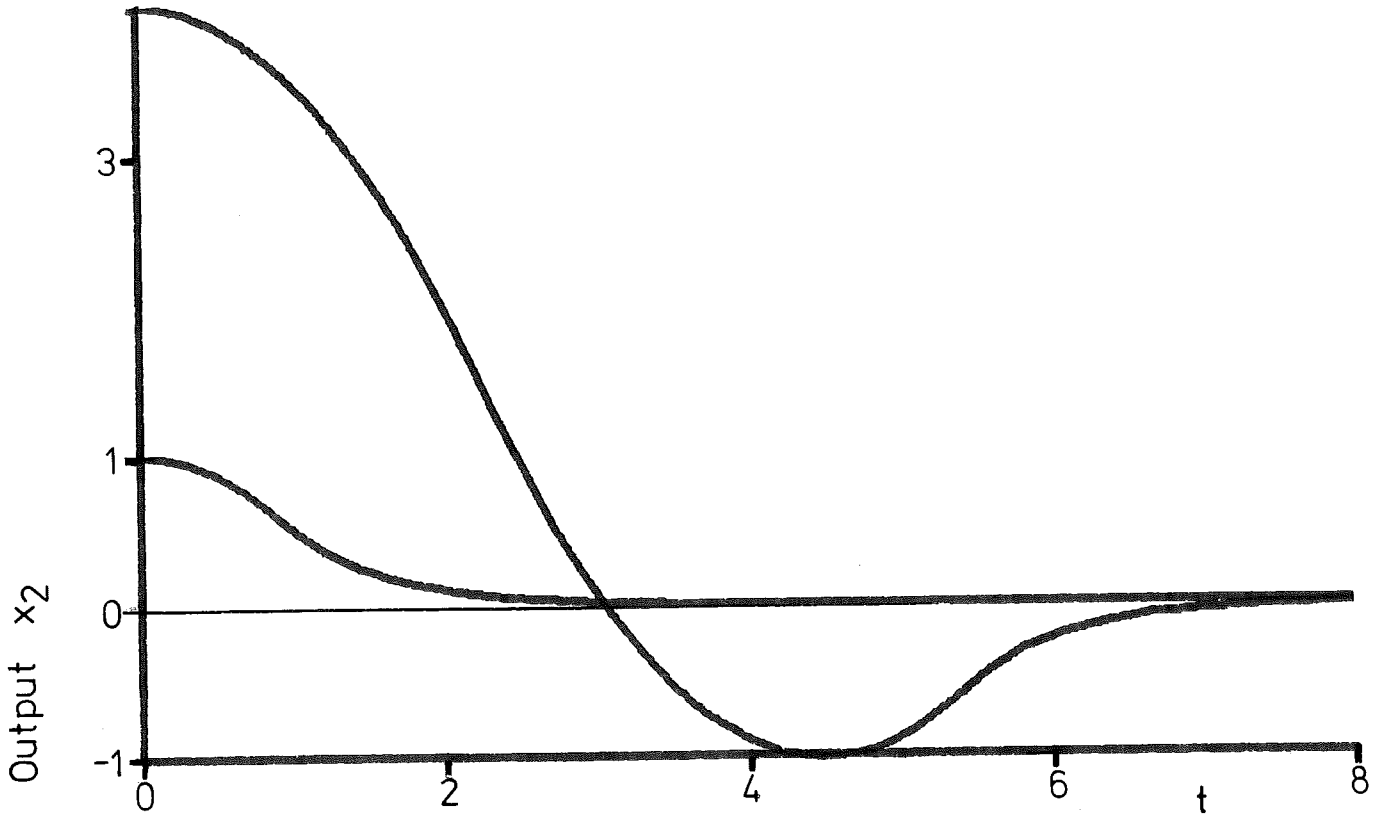


Fig. 5.12. Optimal response of relay control system with min-max criterion.

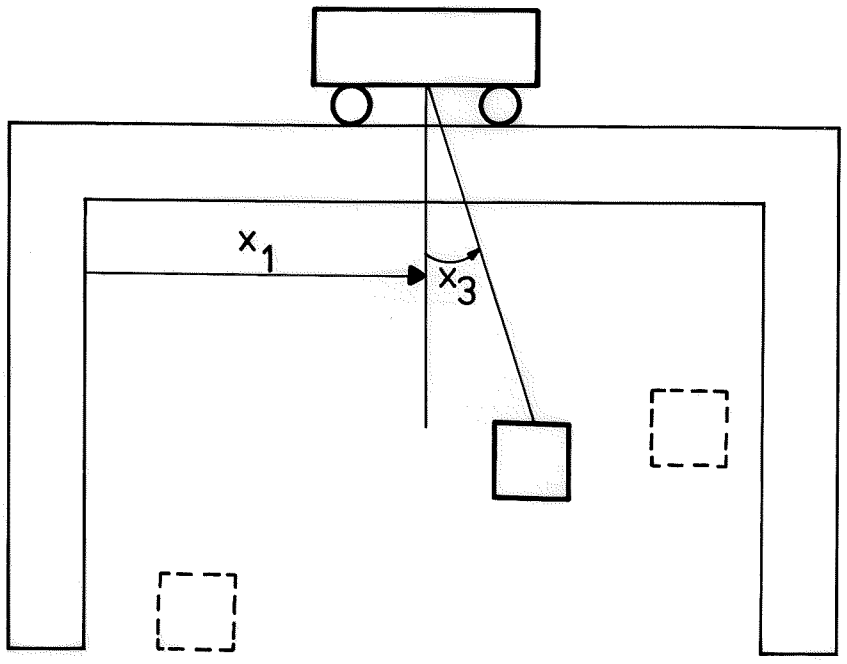
For the weighting factor $\beta = 16$ the result is

$$K = 0.89 \qquad J_1 = 0.85 \qquad J_2 = 1.53$$
$$\lambda_1 = 0 \qquad \lambda_2 = 1.00$$

In this case J_2 dominates and the solution is the same as in Example 5.1.

Example 5.5. Control of a container crane.

The crane which is more fully described in Mårtensson (1972), consists of a trolley moving on a gauntry. The container to be lifted is attached to the end of a cable. The configuration is shown below.



We consider the problem of moving the container between two given positions with the least control effort. The system is described by the following equations.

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = u_1$$
$$\dot{x}_3 = x_4$$

$$\dot{x}_4 = - g \sin x_3/x_5 - 2x_6x_4/x_5 - u_1 \cos x_3/x_5$$

$$\dot{x}_5 = x_6$$

$$\dot{x}_6 = u_2$$

where

x_1 = horizontal position of trolley

x_2 = horizontal velocity of trolley

x_3 = angle between cable and the vertical

x_4 = derivative of x_3

x_5 = length of cable

x_6 = derivative of x_5

u_1 = acceleration of trolley

u_2 = acceleration of cable length

The optimization problem is:

Find u_1 and u_2 such that the system is transferred from

$$x(0)^T = (0,0,0,0,12,0)$$

to

$$x(20)^T = (20,0,0,0,6,0)$$

and

$$J = \int_0^{20} (u_1^2+u_2^2)dt$$

is minimized. This problem is infinite dimensional (u_1 and u_2 are functions on $[0,20]$) and can therefore not be attacked directly by the optimization routine used. However, we will transform the problem into a finite dimensional one. The method cho-

sen is probably not the most efficient one for solving the problem, but it illustrates the versatility of the combined optimization and simulation program.

From Lemma 4.1 it follows that the necessary conditions are

$$\dot{\lambda}_1 = 0$$
$$\dot{\lambda}_2 = -\lambda_1$$
$$\dot{\lambda}_3 = \lambda_4(g \cos x_3 - u_1 \sin x_3)/x_5$$
$$\dot{\lambda}_4 = -\lambda_3 + 2\lambda_4 x_6/x_5$$
$$\dot{\lambda}_5 = -\lambda_4(g \sin x_3 + 2x_4x_6 + u_1 \cos x_3)/x_5^2$$
$$\dot{\lambda}_6 = 2\lambda_4 x_4/x_5 - \lambda_5$$
$$u_1 = -0.5\lambda_2 + 0.5\lambda_4 \cos x_3/x_5$$
$$u_2 = -0.5\lambda_6$$

together with the differential equations for $x_1, \ldots, x_6$. The problem is then a two point boundary value problem. This problem can be solved by choosing $\lambda_1(0), \ldots, \lambda_6(0)$ as variables to be optimized and

$$J = (x_1(20) - 20)^2 + x_2(20)^2 + x_3(20)^2 + x_4(20)^2 +$$
$$+ (x_5(20) - 6)^2 + x_6(20)^2$$

as criterion to be optimized. The result is shown in Figs. 5.13 and 5.14.

This method of solving the two point boundary value problem, so-called "shooting", sometimes leads to difficulties. The reason is that to each eigenvalue μ of the linearization of the x-equations, there corresponds an eigenvalue $-\mu$ of the λ-equations. If the physical system is very well damped the λ-equations are then highly unstable and numerical difficulties occur. In the equations describing the crane, all eigenvalues lie on the imaginary axis and no problems arise.

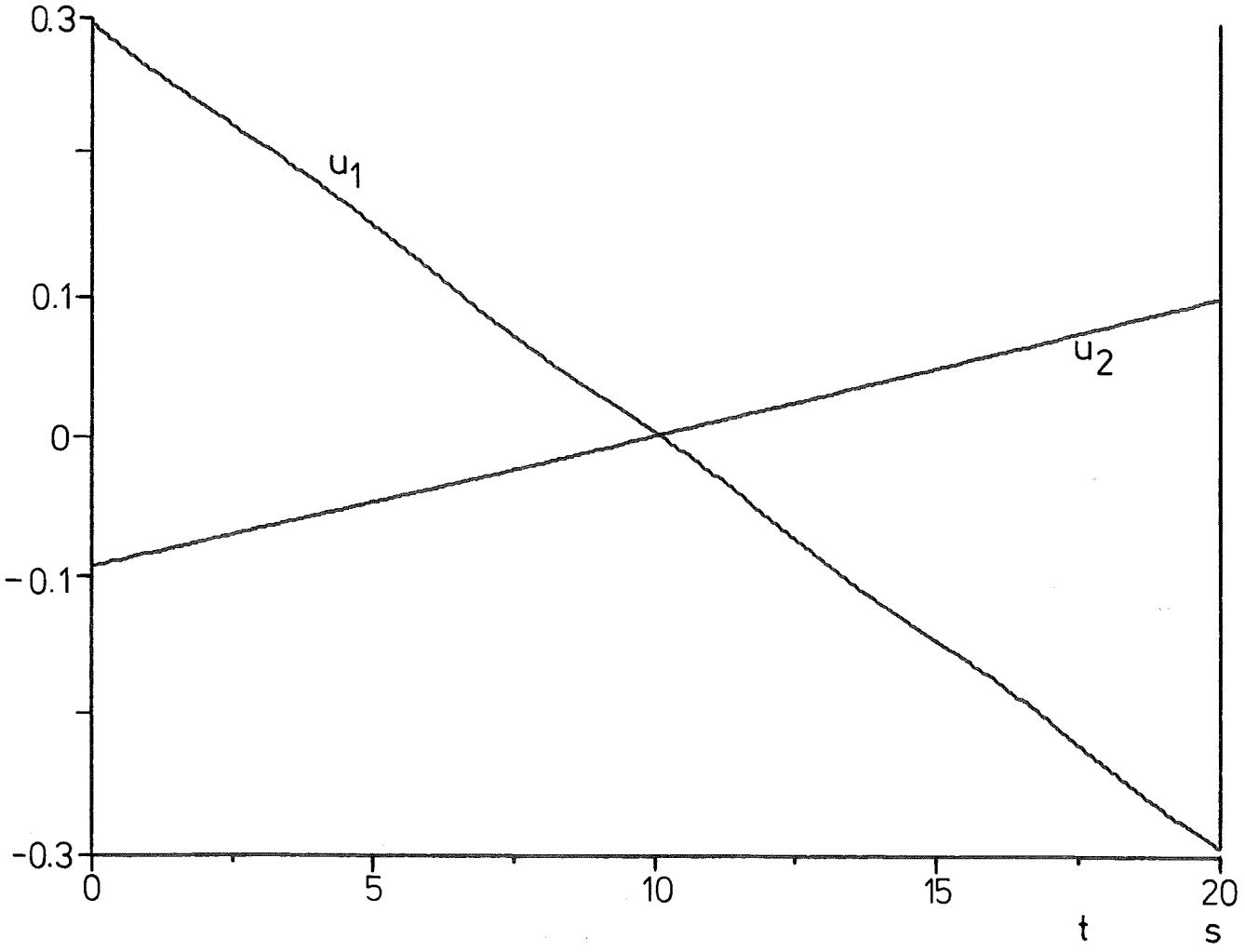


Fig. 5.13. Optimal control policy for the container crane.

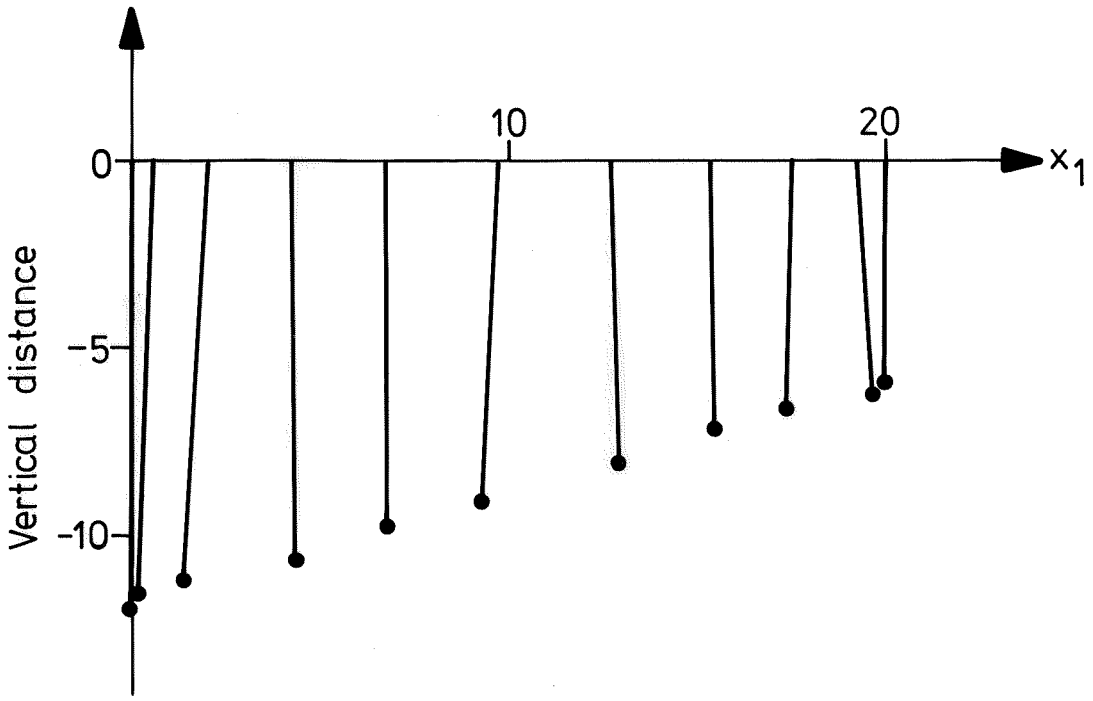


Fig. 5.14. Optimal trajectory for the container. The positions of the container and cable are shown with 2 s intervals.

Also notice that in the problem formulation given here there are no state constraints. In practice, see Mårtensson (1972), such constraints may be necessary. Some other method, e.g. differential dynamic programming , must then be used.

An Application.

The optimization facility of SIMNON has also been used by Holmberg and Svensson (1975) to compute a regulator for a boiling water reactor. The model of the physical process has 14 state variables, 2 inputs and 5 outputs. The 10 coefficients of an output feedback control law were computed by the optimization routine.

5.4 Appendix.

Here the SIMNON systems used in Examples 5.1 - 5.5 are presented. For Example 5.1 the command sequence is also given. A description of the SIMNON language used to define systems and of the SIMNON commands can be found in Elmqvist (1975). The use of optimization in SIMNON is described in Glad (1974).

Example 5.1

The system description is

```
CONTINUOUS SYSTEM RELAY
STATE X1 X2 ILOS
DER DX1 DX2 DILOS
INPUT K
DYNAMICS
CX=K*X1+X2
U=IF CX>0 THEN -1. ELSE 1.
UCH=-ALFA*X1-X1/K-ALFA*X2/K
DX1=IF ABS(CX)<EPS AND ABS(X1)<K THEN UCH ELSE U
DX2=X1
DILOS=X2*X2

ALFA:1
EPS:0.01
END
```

- -

```
CONNECTING SYSTEM CONN
TIME TIM
LOSS[OPTA]=ILOS[RELAY]
K[RELAY]=P1[OPTA]
T=TIM-TBEG[OPTA]
END
```

The variable UCH is used to describe the chattering motion, which takes place on the switching line when $|x_1| < K$. UCH gives the mean value of the actual control which switches infinitely fast between +1 and -1. The background theory is described in Anderson and Moore (1971).

The following command sequence is used. The text within quotation marks is only used for comments and does not belong to the commands.

```
>LET NPAR.OPTA=1  "NUMBER OF PARAMETERS TO OPTIMIZE"
>,NCONS.OPTA=0    "NUMBER OF CONSTRAINTS"
>SYST OPTA RELAY CONN  "DEFINE THE SYSTEM"
>PAR TINC:5  "LENGTH OF TIME FOR EVALUATION OF THE LOSS"
>INIT X2:1   "INITIAL VALUE FOR X2"
>,P11:1      "INITIAL VALUE FOR THE PARAMETER TO BE OPTIMIZED"
>PLOT X2 DX1 (T)  "VARIABLES TO BE PLOTTED ON DISPLAY"
>AXES H 0 5 V -1 1  "SCALING OF AXES ON DISPLAY"
>SIMU 0 100 1  "START THE SIMULATION AND OPTIMIZATION"
```

Example 5.2

```
CONTINUOUS SYSTEM S012
STATE X1 X2 X3 ZU ZL MX T
DER DX1 DX2 DX3 DZU DZL DMX DT
INPUT U K
DYNAMICS
DX1=-3*X1-2*X2+K*U
DX2=X1
DX3=X2
DZU=IF T<TS THEN X2 ELSE IF X2<0 OR X3<ZU THEN 0 ELSE X2
DZL=IF T<TS THEN X2 ELSE IF X2>0 OR X3>ZL THEN 0 ELSE X2
DMX=IF X2<0 OR X3<MX THEN 0 ELSE X2
DT=1.
TS:2.9
END
```

```
CONTINUOUS SYSTEM LEAD
STATE X
DER DX
INPUT U B N
OUTPUT Y
OUTPUT
Y=U-X
DYNAMICS
DX=-B*N*X+B*(N-1)*U
END
```

```
CONNECTING SYSTEM EX524
UREF=1.
U[LEAD]=UREF-X3[S012]
U[S012]=Y[LEAD]
A=MAX(0.,MX[S012]-1.17)
B=MAX(0.,ZU[S012]-1.05)
CC=MAX(0.,0.95-ZL[S012])
D=MAX(0.,PD2[OPTA]-PD1[OPTA])
LOSS[OPTA]=A*A+B*B+CC*CC+D*D
K[S012]=P1[OPTA]
N[LEAD]=P2[OPTA]
B[LEAD]=P3[OPTA]
UU=K[S012]*U[S012]
UB=IF T[S012]>TS THEN 1.05 ELSE 0
LB=IF T[S012]>TS THEN 0.95 ELSE 0
MXX=1.17
TS:2.9
END
```

Example 5.3

```
CONTINUOUS SYSTEM TANKS
STATE X1 X2 V Z J T
DER DX1 DX2 DV DZ DJ DT
INPUT K TI TD
DYNAMICS
S1=MAX(0.,2*G*X1)
SQ1=A*SQRT(S1)/AA
S2=MAX(0.,2*G*X2)
SQ2=A*SQRT(S2)/AA
ERR=REF-X2
UNOM=K*(ERR+Z/TI+TD*(SQ2-SQ1))+U0
U0:85.18
U=IF UNOM>UMAX THEN UMAX ELSE IF UNOM<0 THEN 0 ELSE UNOM
DV=-V/TV+U/TV
DX1=-SQ1+V/AA
DX2=SQ1-SQ2
DZ=ERR
DT=1.
A:0.43
AA:19.6
G:981
TV:3
REF:20
UMAX:120.
DJ=IF T>TST THEN ABS(ERR) ELSE 0.
TST:10
END
```

```
CONNECTING SYSTEM COTAN
LOSS[OPTA]=J[TANKS]
K[TANKS]=P1[OPTA]
TI[TANKS]=P2[OPTA]
TD[TANKS]=0.
END
```

Example 5.4

Two copies of the system RELAY described in example 5.1 are used. They are called RELAY and RELCO and are started with different values of x_2. The connecting system is

```
CONNECTING SYSTEM CONN
TIME TIM
LOSSCOPTA]=P02COPTA]
CON1COPTA]=ILOSERELAY]-P02COPTA]
A=ILOSERELCO]/FACT
FACT:32
CON2COPTA]=A-PD2COPTA]
KERELAY]=P1COPTA]
KERELCO]=P1COPTA]
T=TIM-TSEGCOPTA]
END
```

Example 5.5

```
CONTINUOUS SYSTEM CRANE
STATE X1 X2 X3 X4 X5 X6 L1 L2 L3 L4 L5 L6
DER DX1 DX2 DX3 DX4 DX5 DX6 DL1 DL2 DL3 DL4 DL5 DL6
DYNAMICS
S3=SIN(X3)
C3=COS(X3)
U1N=-0.5*L2+0.5*L4*C3/X5
U2N=-0.5*L6
U1=IF U1N>S1 THEN S1 ELSE IF U1N<I1 THEN I1 ELSE U1N
U2=IF U2N>S2 THEN S2 ELSE IF U2N<I2 THEN I2 ELSE U2N
S1:0.46
I1:-0.46
S2:0.61
I2:-0.61
DX1=X2
DX2=U1
DX3=X4
DX4=-G*S3/X5-2*X6*X4/X5-U1*C3/X5
DX5=X6
DX6=U2
DL1=0
DL2=-L1
DL3=-L4*(-G*C3+U1*S3)/X5
DL4=-L3+2*L4*X6/X5
DL5=-L4*(G*S3+2*X4*X6+U1*C3)/(X5*X5)
DL6=-L5+2*L4*X4/X5
G:9.81
XX=X1+X5*S3
YY=-X5*C3
END
```

```
DISCRETE SYSTEM CR
INPUT A1 A2 A3 A4 A5 A6
TIME TCR
TSAMP TSCR
DYNAMICS
L1[CRANE]=A1
L2[CRANE]=A2
L3[CRANE]=A3
L4[CRANE]=A4
L5[CRANE]=A5
L6[CRANE]=A6
TSCR=TCR+DT
DT:20
END
```

```
CONNECTING SYSTEM OPCRA
TIME TIM
B1=X1[CRANE]-D
D:20
B2=X2[CRANE]
B3=X3[CRANE]
B4=X4[CRANE]
B5=X5[CRANE]-H1
H1:6
B6=X6[CRANE]
LOSS[OPTA]=B1*B1+B2*B2+KK*(B3*B3+B4*B4)+B5*B5+B6*B6
KK:10
A1[CR]=P1[OPTA]
A2[CR]=P2[OPTA]
A3[CR]=P3[OPTA]
A4[CR]=P4[OPTA]
A5[CR]=P5[OPTA]
A6[CR]=P6[OPTA]
T=TIM-TBEG[OPTA]
END
```

5.5. References.

Åström, K.J. (1970):
Introduction to Stochastic Control Theory, Academic Press, New York.

Anderson, B.D.O., and Moore, J.B. (1971):
Linear Optimal Control, Prentice Hall, New Jersey.

Elmqvist, H. (1975):
SIMNON - An Interactive Simulation Program for Nonlinear Systems - User's Manual, Report 7502, Lund Institute of Technology, Division of Automatic Control.

Fletcher, R. (1972):
Fortran Subroutines for Minimization by Quasi-Newton Methods, Report AERE-R7125, United Kingdom Atomic Energy Research Establishment, Harwell.

Fuller, A.T. (1967):
Linear Control of Non-Linear Systems, Int. J. Control, Vol. 5, No. 3, pp. 197 - 243.

Glad, T. (1974):
A Program for Interactive Solution of Parametric Optimization Problems in Dynamic Systems, Report 7424, Lund Institute of Technology, Division of Automatic Control.

Graham, D., and Lathrop, R.C. (1953):
The Synthesis of "Optimum" Transient Response: Criteria and Standard Forms, Trans. Amer. Inst. Electrical Engineers, Vol. 72, pp. 273 - 288. Reprinted in Oldenburger, R. ed.: Optimal and Self-Optimizing Control, MIT Press, 1966.

Hagander, P. (1972):
Numerical Solution of $A^TS + SA + Q = 0$, Information Sciences 4, pp. 35 - 50.

Holmberg, N. and Svensson, J. A. (1975):
Återkopplingsreducering i reglersystem för kokarreaktor.(Reduced Feedback of a Boiling Water Reactor), Report RE-168, Division of Automatic Control, Lund Institute of Technology.(in Swedish)

Levine, W.S., and Athans, M. (1970):
On the Determination of the Optimal Constant Output Feedback Gains for Linear Multivariable Systems, IEEE Trans. Automatic Control, AC-15, pp. 44 - 49.

Martens, H.R., and Larsen, G.R. (1975):
A New Performance Index for Improved Computer Aided Design of Control Systems, Journal of Dynamic Systems, Measurement and Control, Trans. ASME, Vol. 97, Ser. G, No. 1, pp 69-74

Mårtensson, K. (1970):
Suboptimal Linear Regulators for Linear Systems with Known Initial-State Statistics, Report 7004, Lund Institute of Technology, Division of Automatic Control.

Mårtensson, K. (1972):
New Approaches to the Numerical Solution of Optimal Control Problems, Report 7206, Lund Institute of Technology, Division of Automatic Control.

Newton, G.C., Gould, L.A., and Kaiser, J.F. (1957):
Analytical Design of Linear Feedback Controls, Wiley, New York.

Vandierendonck, A.J. (1972):
Design Method for Fully Augmented Systems for Variable Flight Conditions, Report AFFDL-TR-71-152, Air Force Flight Dynamics Laboratory, Wright-Patterson Air Force Base, Ohio.

Zakian, V. (1973):
Design of Dynamical and Control Systems by the Method of Inequalities, Proc. IEE, Vol. 120, No. 11, pp. 1421 - 1427.

6. ON LINE OPTIMIZATION OF AN OIL BURNER

An example of a practical optimization problem is the adjustment of a burner to obtain the most efficient combustion. To do this manually can be time consuming and it is therefore of interest to try out automatic methods for reaching the optimum.

This chapter describes the use of a numerical optimization algorithm on line to perform this task. Section 6.1 describes the oil burner and the arrangements that were made to allow computer control. In 6.2 the modifications of the algorithm required by the on-line use are discussed. The results of practical experiments are presented in 6.3.

6.1. Description of the Process.

The physical process is an oil burner installation of the type used for heating of small houses. The burner itself is of a new type developed at the department of Machine Design at the Lund Institute of Technology, see Reenstierna (1975). The essential difference compared to an ordinary burner is that the combustion takes place in the alcohol-aldehyde phase and gives a blue flame. This makes it possible to get complete combustion with little excess air. The physical background is described in Fritsch (1961).

The new burner was to be tested with various nozzles and turbulators. In each case the burner has to be adjusted for optimum combustion. Since it is fairly laborious to do the adjustment manually, it is of interest to study the possibility of adjusting the burner automatically using a computer.

A schematic description of the process is given in Fig. 6.1. The oil flow is controlled by an oil pump and the air flow by the velocity of a fan and the variable opening at the inlet to

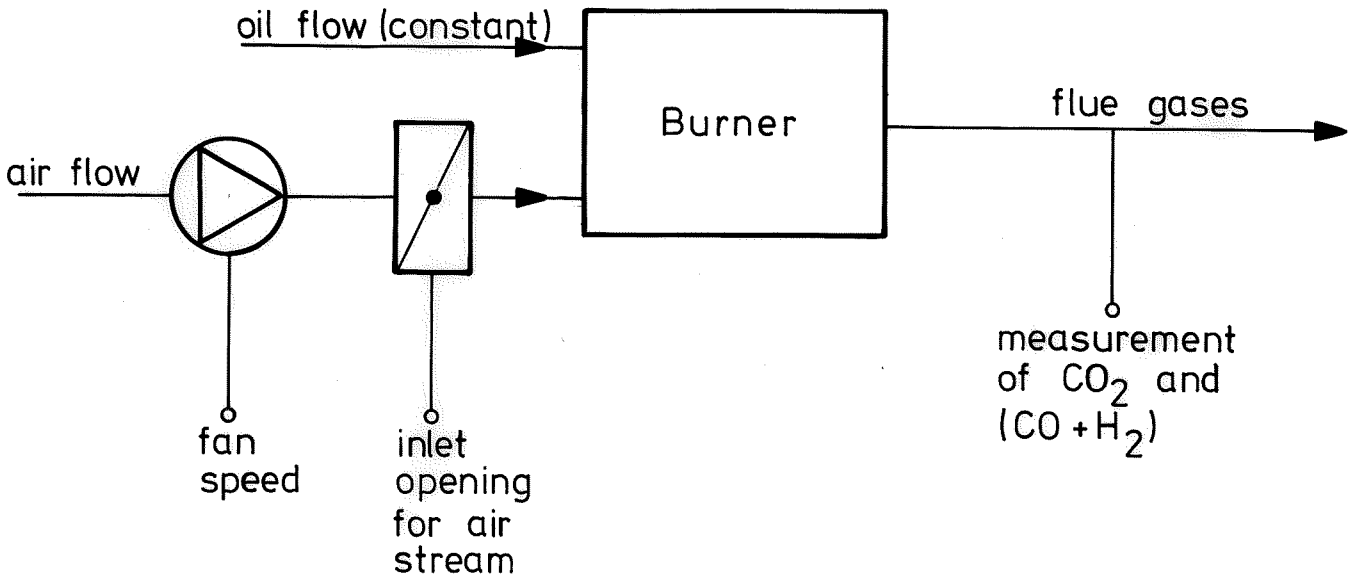


Fig. 6.1

the burner. The draught through the chimney can also be controlled. The contents of carbon dioxide (CO_2) and carbon monoxide and unburnt hydrogen compounds ($CO + H_2$) in the flue gases can be measured. The problem that is studied is to find those values of fan speed and inlet opening for the air stream, which give the best combustion, for fixed values of oil pressure and draught.

The combustion should be complete, i.e. give a negligible value of the ($CO + H_2$) measurement. On the other hand there should be as little excess air as possible to minimize the heat loss through the chimney. When complete combustion takes place, the incoming oxygen either remains as oxygen or reacts to form carbon dioxide. The minimization of the oxygen content of the flue gases is therefore equivalent to the maximization of the CO_2 content.

The connections between the computer and the process were done in the following way. Because of the distance between the computer and the process, a Hewlett-Packard Coupler/Controller was

placed at the process. The data transmission between the Coupler/Controller and the computer, a PDP-15, was then done digitally. The output from the PDP-15 was given by the Coupler to a digital-analog converter where it was converted to voltages in the range -10V to 10V. Measurements were made by a digital voltmeter operated by the Coupler and sent to the computer. Fig. 6.2 shows the arrangement. A description of the Coupler/ /Controller and the programs used can be found in Jensen (1973), (1974).

The fan was driven by a thyristor controlled DC-motor and the thyristor control unit could be connected directly to the DA converter.

The air intake was controlled by an electric motor via a chain, see Fig. 6.3. The position was measured by a potentiometer connected to the shaft of the motor. The air inlet opening could then be measured as a voltage. To control this opening, the DA output gave a reference voltage to a relay servo which drove the motor until the measured and desired voltages agreed.

The measurement apparatus, see Fig. 6.4, takes a sample of flue gas every two minutes. The CO_2 content is observed by measuring the decrease in volume of the gas sample after it has passed through a liquid absorbing the CO_2. Every second measurement the gas sample is passed through an oven, where unburnt components are oxidized to carbon dioxide and water, before the gas passes through the absorbing liquid. This means that every second measurement gives the CO_2-content and the measurements in between give $(CO_2 + CO + H_2)$-content. The $(CO + H_2)$-content is thus arrived at by taking the difference between two successive measurements.

The output of the measuring apparatus is in the form of a pointer writing on a moving chart. To get an output that could be read by a voltmeter, a mirror was fastened to the axis of the pointer. A light beam was reflected by this mirror on to a light sensitive potentiometer, see Fig. 6.5. An overall picture of the process is shown in Fig. 6.6.

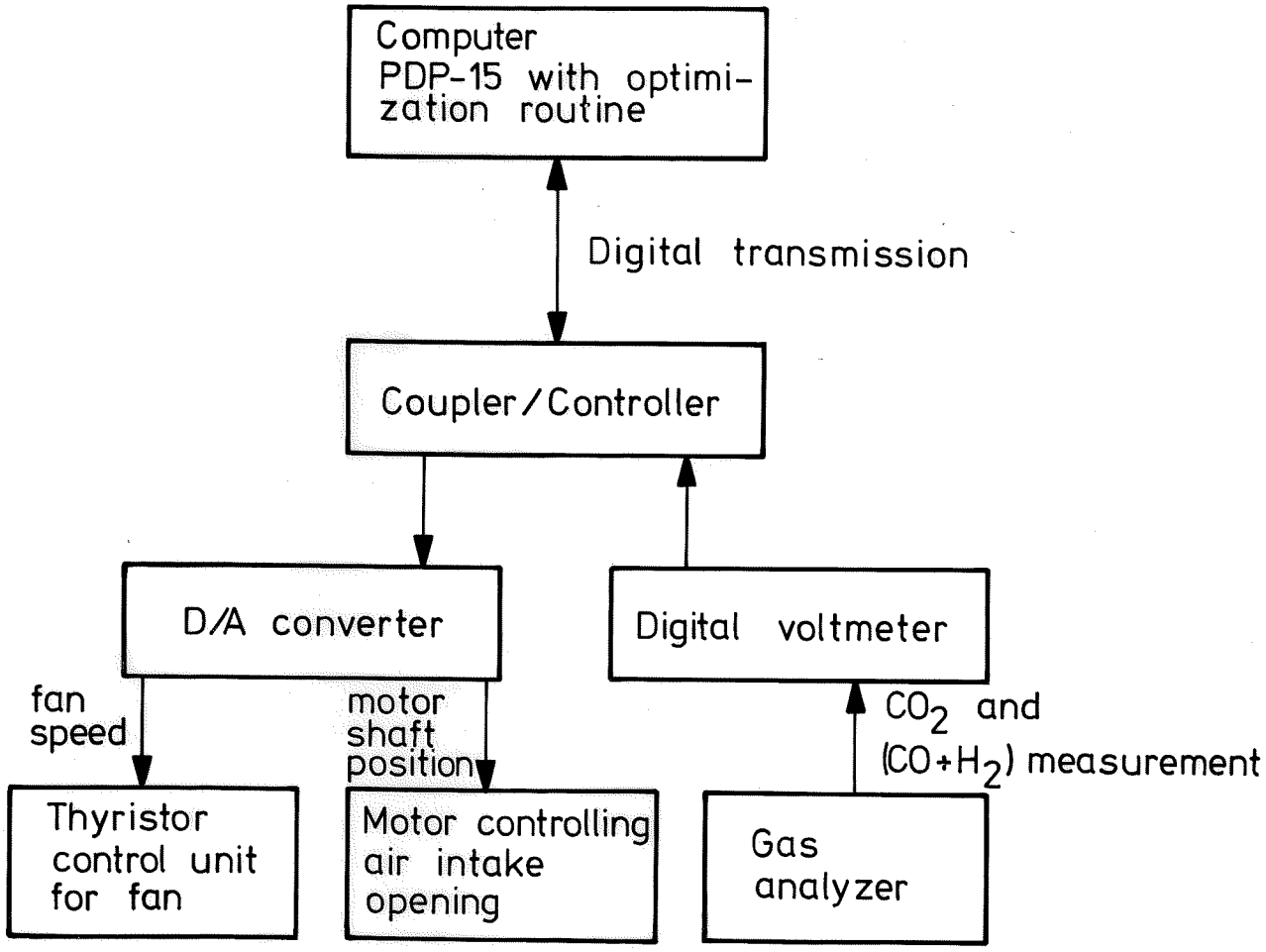


Fig. 6.2. On line optimization of oil burner.

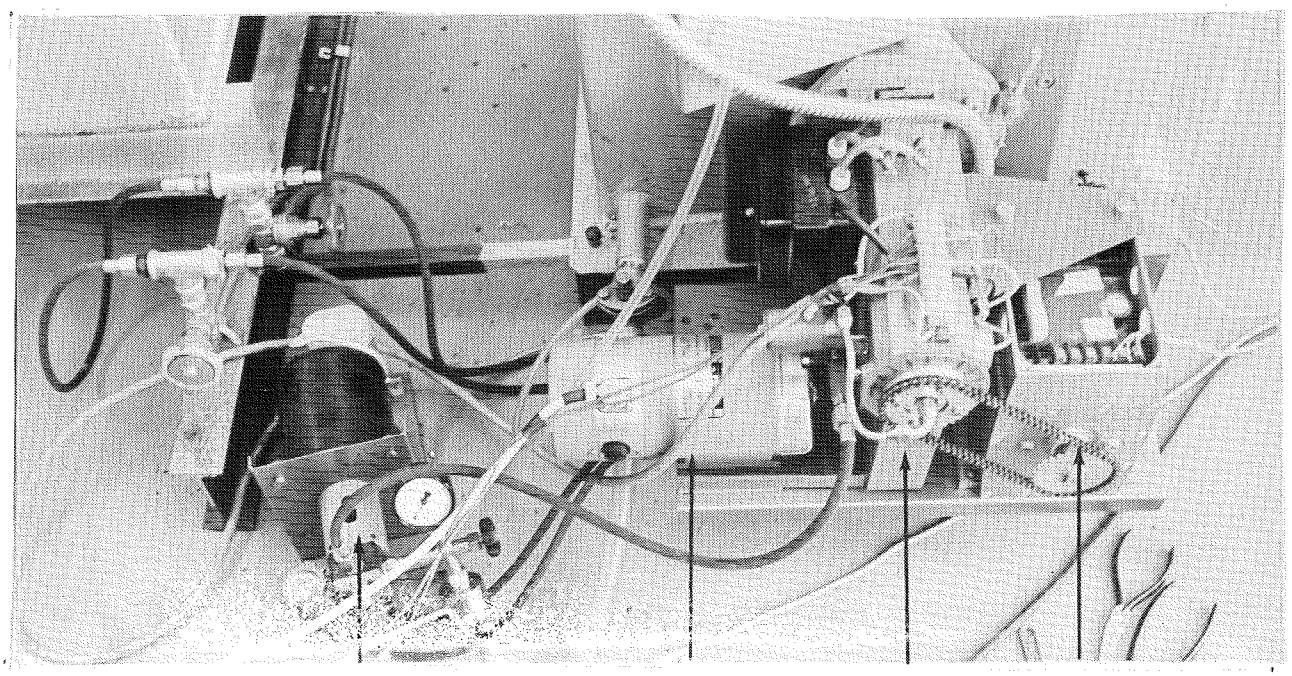


Fig. 6.3. Burner with oil pump, DC-motor for the fan and positioning motor for the variable air intake.

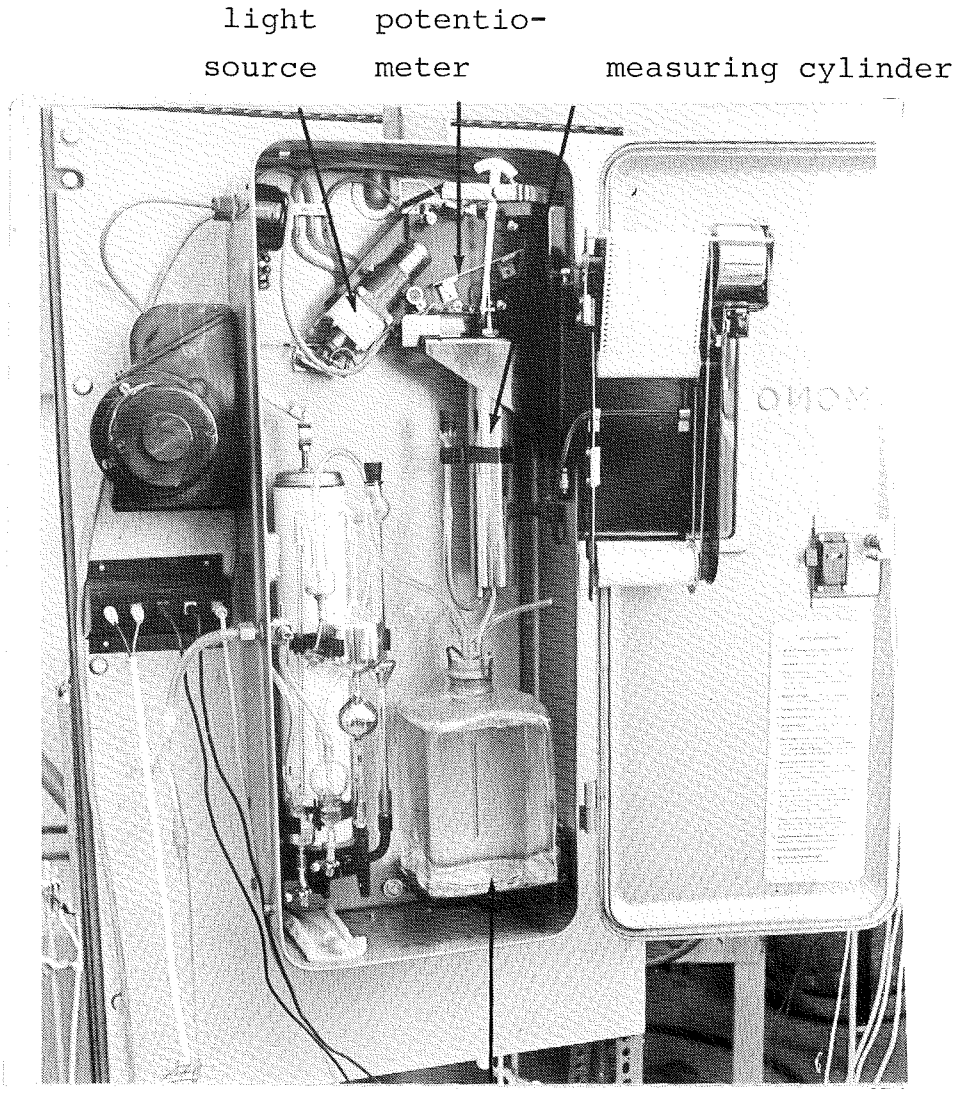


Fig. 6.4. Gas analyzer. A sample of gas is taken into the cylinder to the left. After passing through the big bottle containing a liquid that absorbs CO_2, the gas volume is measured in the cylinder to the right. The volume measurement is transferred to the pointer.

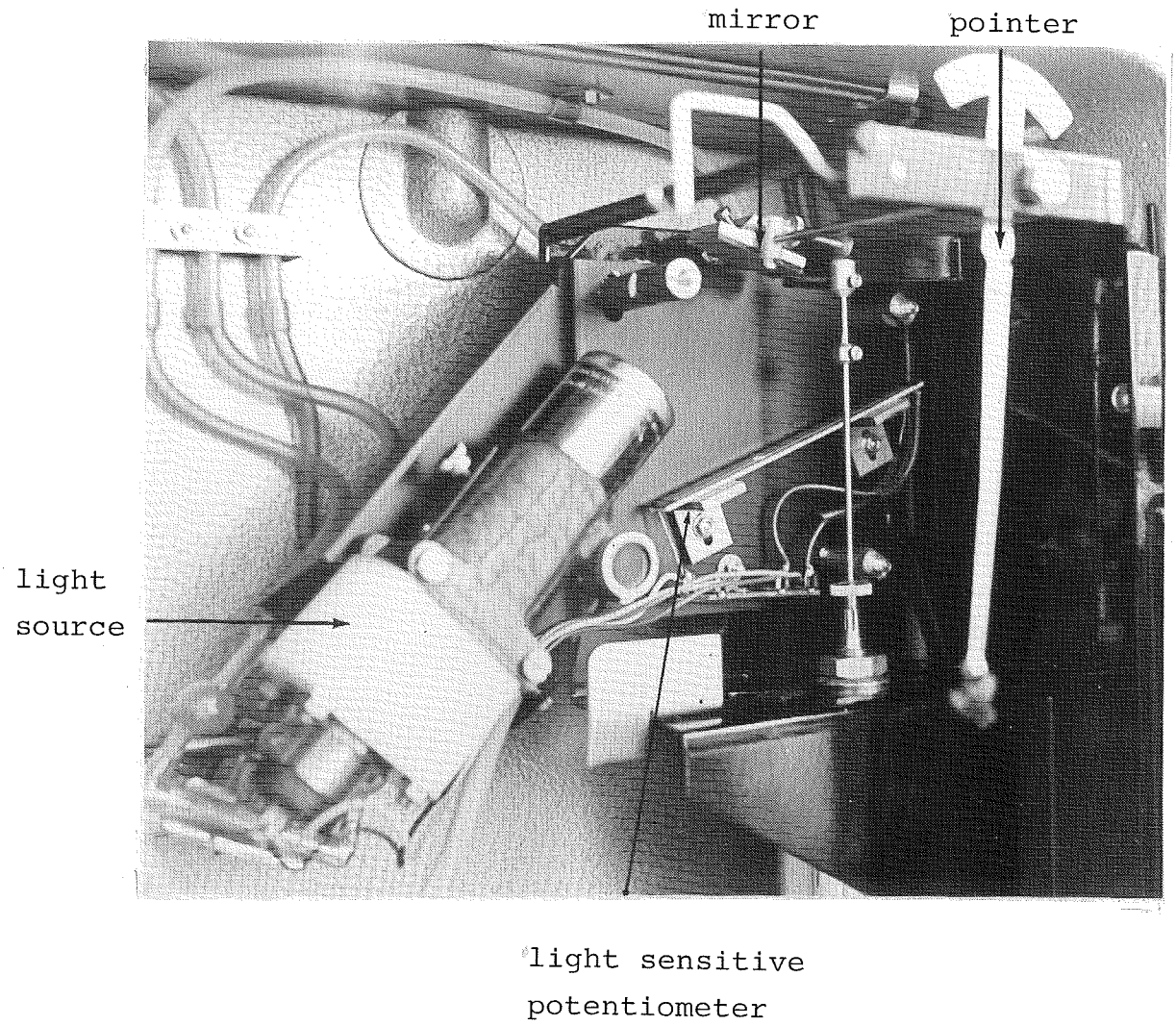


Fig. 6.5. Readout arrangement for gas analyzer. A mirror is attached to the same axis as the pointer. A light beam is reflected by the mirror on to the light sensitive potentiometer in the centre.

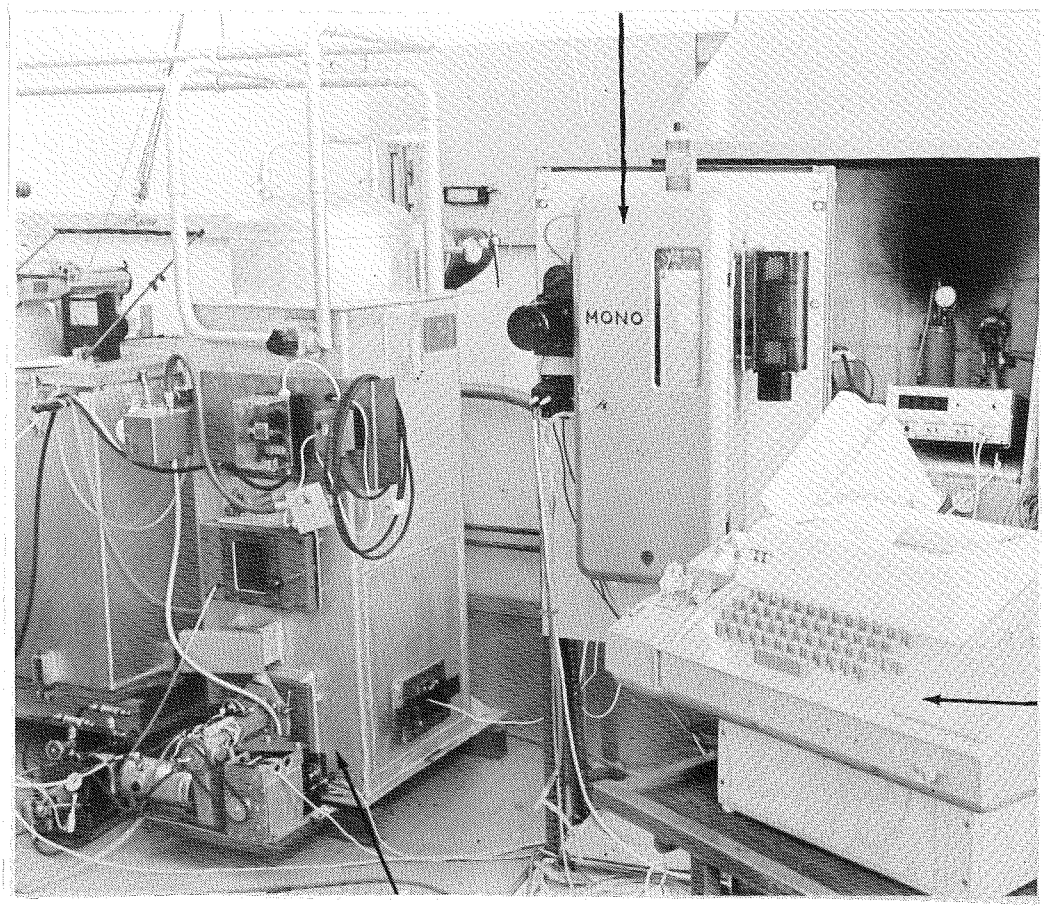


Fig. 6.6. The oil burner installation with analyzer. The teletype is used for communication with the Coupler/Controller.

6.2. Problems that arise in On Line Optimization.

In principle the optimization algorithm works in the same way when it minimizes a mathematically defined function and when it minimizes a criterion given by a physical process on line. In the first case the function value corresponding to a certain parameter value is requested from a subroutine, in the second case the parameter values are given as inputs to the process, and the loss function received as output. The main difference is, of course, that the physical process gives a value of the loss function which is influenced by measurement errors. This may impair the working of the algorithm quite seriously. In this application, an algorithm based on a quasi-Newton method, see Fletcher (1972), is used. The gradients are calculated from difference approximations. It is possible that an unfortunate combination of measurement errors will make the computed search direction an uphill direction, stopping (at least momentarily) further progress.

Simplified versions of minimization algorithms based on difference approximations can be analyzed from a stochastic point of view. The stochastic approximation algorithm

$$x^{(n+1)} = x^{(n)} - \gamma^{(n)} Df\left(x^{(n)}\right)$$

where $Df\left(x^{(n)}\right)$ is the difference approximation of f_x at $x^{(n)}$

$$\left(Df\left(x^{(n)}\right)\right)_i = \frac{1}{2h^{(n)}}\left[f\left(x_1^{(n)}, \ldots, x_i^{(n)} + h^{(n)}, \ldots\right) - f\left(x_1^{(n)}, \ldots, x_i^{(n)} - h^{(n)}, \ldots\right)\right]$$

has been analyzed by Kiefer and Wolfowitz (1952). Kushner (1972) has studied some generalizations of this scheme. It is possible to prove that the algorithm converges with probability one to a point where $f_x = 0$, if

$$\Sigma\gamma^{(n)} = \infty$$

$$\Sigma\gamma^{(n)} h^{(n)} < \infty$$

$$\Sigma \frac{\gamma^{(n)2}}{h^{(n)2}} < \infty$$

and with weak restrictions on the noise and the loss function. In Kushner (1972), the modification of the search direction to

$$- H^{(n)} Df\left(x^{(n)}\right)$$

is also considered. If the matrices $H^{(n)}$ satisfy

$$0 < \varepsilon_1 I \leq H^{(n)} \leq \varepsilon_2 I \text{ for some } \varepsilon_i$$

convergence with probability one can still be proved.

Since the line search in the Quasi-Newton method is more complicated than the one considered above, the theoretical results do not apply directly. Still, they show that it is reasonable to try a method based on difference approximations of the gradient.

A problem that has to be considered is that the algorithm might take too large steps. This can lead to an adjustment of the air flow that extinguishes the flame and stops the optimization. To reduce this risk, the line search algorithm was altered so that an upper limit on the change in each variable was inserted.

Another problem arising from the fact that the values of the loss function are disturbed by noise is the following. The algorithm stores the value, f^*, of the best point obtained. During the line search the new values of the loss function are compared with f^* and a new point is only accepted if a value lower than f^* is obtained. This is a normal procedure in most optimization algorithms irrespective of the method used. However, when the

loss function is obtained from a physical process, difficulties might occur in the following way. The stochastic disturbances acting at the moment of the measuring of f^* might give an abnormally low value of f^*. Since it will then be very difficult to find any value which is lower, the progress of the algorithm will be stopped. Another cause of difficulties is drift in the process. If the value of f^* used for comparison was obtained long before the present measurement, it might not be relevant for the comparison. To get round these difficulties, it is necessary to reset the value of f^* by making a new measurement of it. In the algorithm used this is done every time there is a failure to obtain an improved point in a line search. This resetting of the value of f^* can also be initiated by manual interaction during the minimization, if some abnormal behaviour of the process is observed.

As pointed out in the discussion in 6.1, the criterion can be stated as the maximization of CO_2-content while keeping $(CO + H_2)$-content low. A natural criterion is then

$$J = - y_{CO_2} + k \cdot y_{CO}$$

where

y_{CO_2} = measurement of CO_2-concentration

y_{CO} = measurement of $(CO + H_2)$-concentration

and J is to be minimized. The constant k gives a suitable weighting of the objectives of maximizing CO_2-content and obtaining complete combustion. A disadvantage with this criterion is that a large weight is given to y_{CO}, which is a more uncertain measurement than y_{CO_2}, because it is a difference between two measurements of the same order of magnitude. A possible modification is to use

$$J = - y_{CO_2} + [\max(0, y_{CO} - d)]^n$$

Here d is threshold level. If y_{CO} is below this level, the $(CO + H_2)$-content is considered to be zero. This is natural, since even if the combustion were ideal, one would obtain a value of y_{CO} different from zero caused by the normal fluctuation of the measurements. By choosing a value of n greater than 1, it is possible to penalize large values of y_{CO} more than small ones.

6.3. Experimental Results.

The results of three experiments are given below. Since the measurement apparatus was fairly slow - each function evaluation required 8 minutes - the experiment length has been limited to about 25 function evaluations. Therefore the asymptotic properties of the optimization algorithm remain to a large extent unknown. What the experiments show is the ability of the algorithm to give a fairly rapid decrease of the loss during the first iterations.

The numerical values of fan speed, p_1, and air inlet opening, p_2, that are given are the output voltages from the D/A converter. For the fan, 0 corresponds to zero speed and 10 to maximum speed. For the air opening, -10 corresponds to the fully open and +10 to the fully closed position. For the loss function the lower bound corresponding to ideal combustion is about 2.9, corresponding to a CO_2-content of 15.2%. An increase in the loss function of 1 unit corresponds to a drop in the CO_2-content to about 11.5%, assuming there is no $(CO + H_2)$-content. These figures are only approximative because the calibration of the measurements varies a little from one experiment to another. The exact values of CO_2 and $(CO + H_2)$-content at the optimum are given for each experiment.

Different nozzles and turbulators were used during the different experiments. Since the physical properties are altered appreciably by the change of nozzle and/or turbulator, the shapes of the loss functions are quite different for the different experiments.

Experiment 1. This experiment was performed April 24, 1975. The lowest loss function value was registered after 17 evaluations for $p_1 = 4.9$ and $p_2 = 4.1$. The value of the loss is 3.14 and corresponds to 14.6% CO_2 and 0.1% ($CO + H_2$). The loss is computed from $- y_{CO_2} + 7.5y_{CO}$. See fig. 6.7.

Experiment 2. This experiment was performed on May 15, 1975. The best value of the loss is the one reached on the 15th function evaluation. It is attained for $p_1 = 6.3$, $p_2 = 3.6$ and the loss is 3.32. The CO_2 content at this point was 14.7% and the ($CO + H_2$)-content less than 0.1%. The loss is computed from $- y_{CO_2} + 7.5y_{CO}$. See fig. 6.8.

Experiment 3. This experiment was performed on June 18, 1975. The lowest loss occured after 9 and 19 evaluations. The loss was in both cases 2.68 corresponding to 14.4% CO_2 and less than 0.1% ($CO + H_2$). It was attained for $p_1 = 5.5$, $p_2 = 4.1$, and $p_1 = 5.4$, $p_2 = 3.8$ respectively. The loss is computed from $- y_{CO_2} + 40 \cdot [\max(y_{CO}-0.02,0)]^2$. See fig. 6.9.

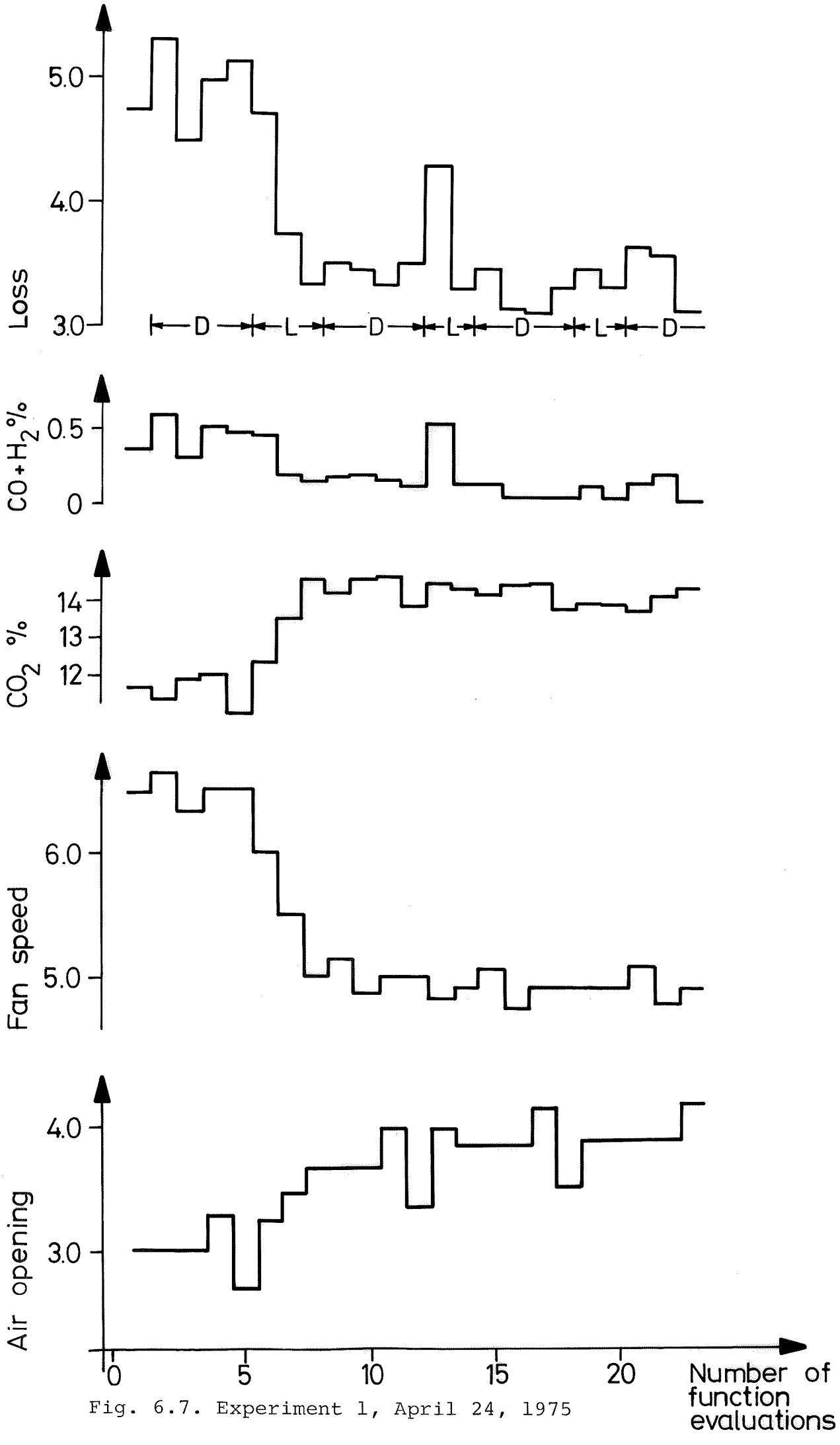


Fig. 6.7. Experiment 1, April 24, 1975

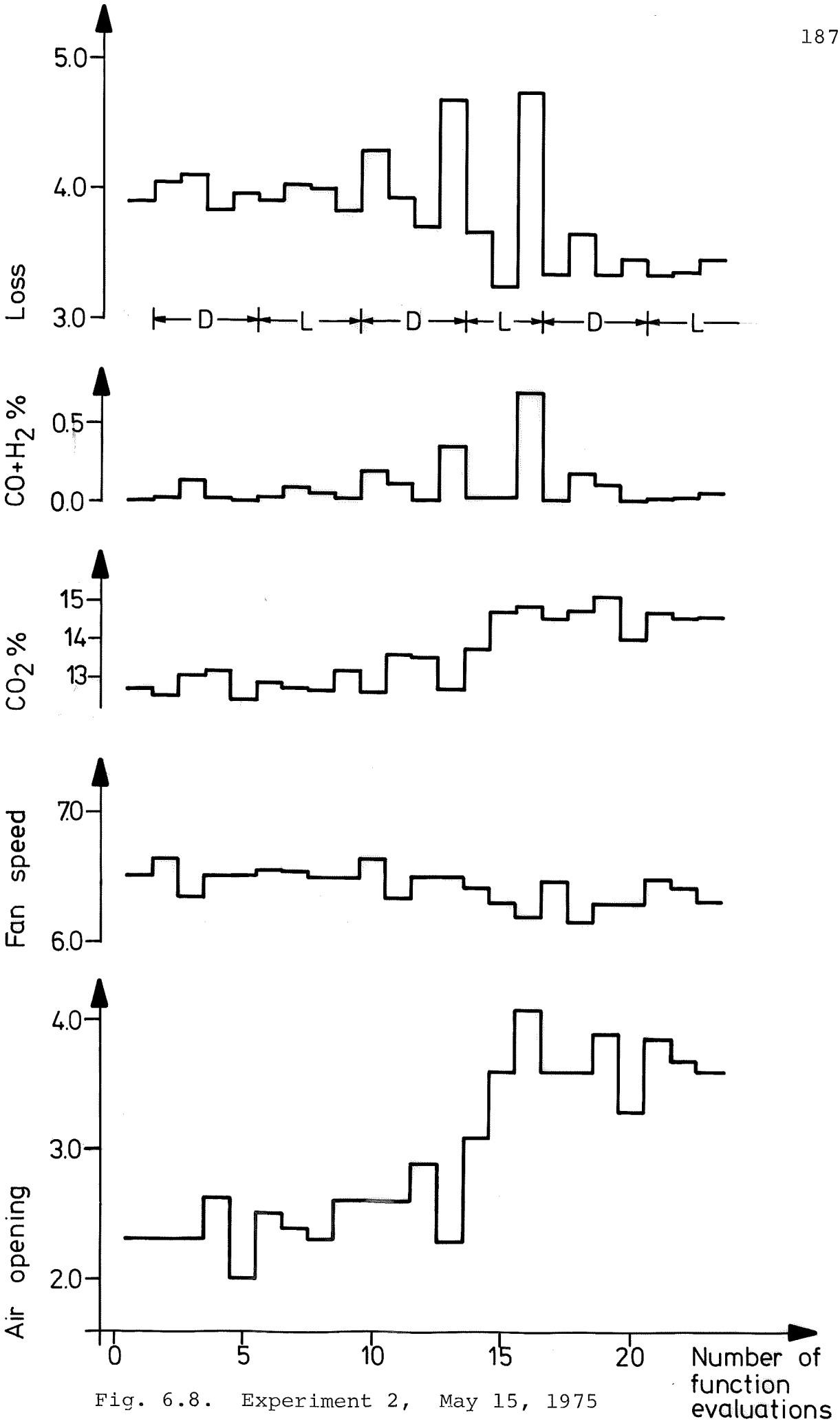


Fig. 6.8. Experiment 2, May 15, 1975

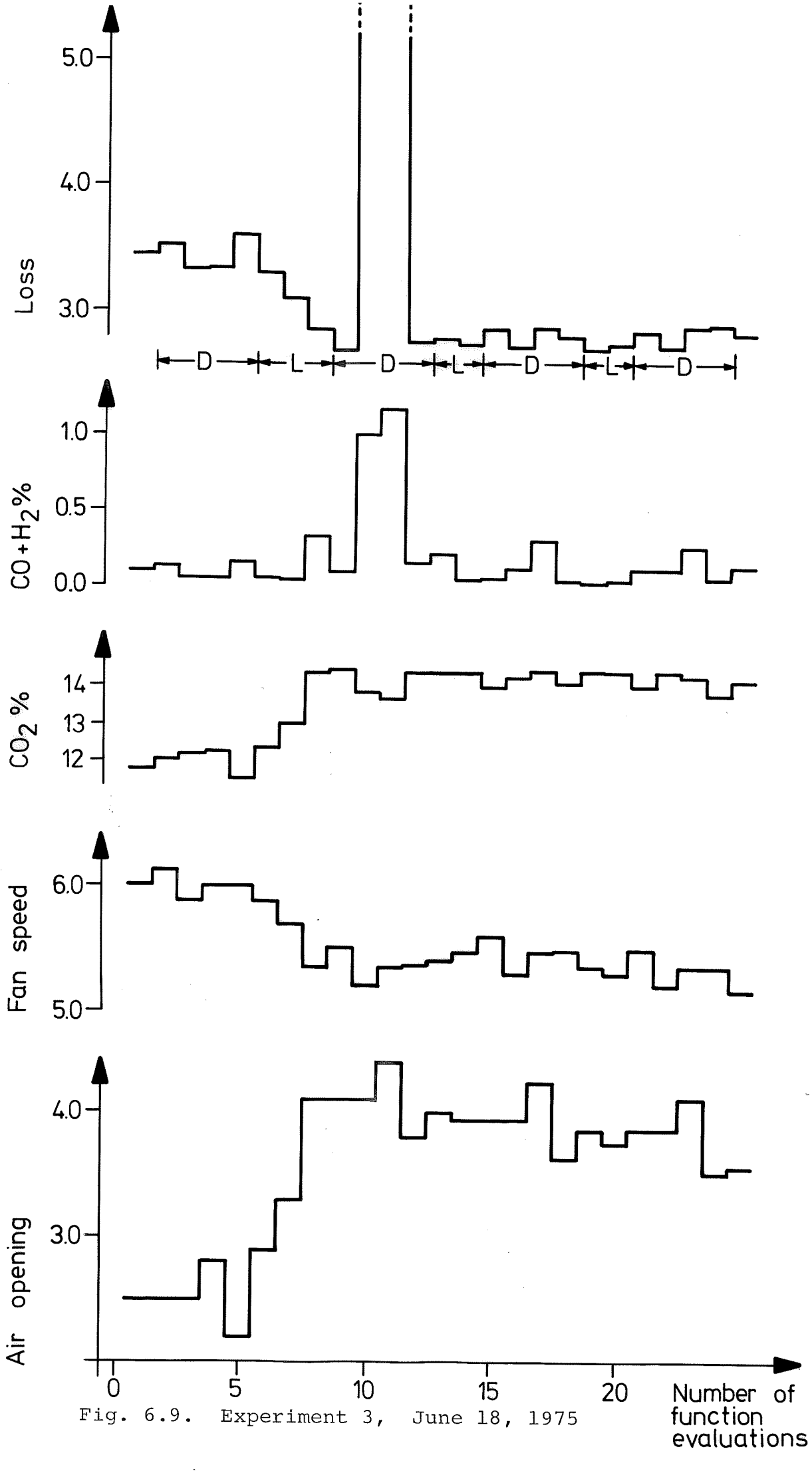


Fig. 6.9. Experiment 3, June 18, 1975

At the end of the experiments the differences between loss function values are so small that stochastic errors dominate. As discussed in Section 6.2, a suitable step length reduction rule should make it possible for the algorithm to converge despite these errors. The limit on experiment length has not made it possible to test this in practice.

However, the results of the experiments show that it is quite possible to use an optimization algorithm in on line applications. In this case the optimization is intended only for the experimental testing of the oil burner. In normal use the burner would be used with a fixed setting. However, for burners operating in a big plant, it might be possible to consider permanent use of an on line optimizer. The burner could then be held at an optimal operating point under all conditions. This may be an interesting comtinuation of the present work.

The different phases of the optimization are indicated below the plot of the loss function. "D" means calculation of the difference approximation of the gradient and "L" means line search. The difference approximation requires four function evaluations because central differences are used. Notice that most of the improvement is achieved in a single line search. In experiment 2 there is a delay because the first search direction is a bad one and the improvement comes in the second line search.

6.4. References.

Fritsch, W.H. (1961):
Chemisch-physikalische Grundlagen der blauen Flamme, Die Ölfeuerung, 1961, pp. 662-682.

Fletcher, R. (1972):
Fortran Subroutines for Minimization by Quasi-Newton Methods, Report AERE-R7121, United Kingdom Atomic Energy Research Establishment, Harwell.

Jensen, L.H. (1973):
Ett Coupler/Controllersystem. Report 7339, Lund Institute of Technology, Division of Automatic Control (in Swedish).

Jensen, L.H. (1974):
Computer Programs for Fullscale Experiments, Report 7424, Lund Institute of Technology, Division of Automatic Control.

Kiefer, J., and Wolfowitz, J. (1952):
Stochastic Estimation of the Maximum of a Regression Function, Ann. Math. Stat. 23, pp. 462-466.

Kushner, J.K. (1972):
Stochastic Approximation Type Algorithms for the Optimization of Constrained and Multinode Stochastic Problems, CDS Technical Report 72-1, Brown University.

Reenstierna, B. (1973):
Projekt Blåbrännare,(Project Blue Burner), Report, Division of Machine Design, Lund Institute of Technology.