Monograph:

Owens, D.H. and Jones, R.P. (1976) Iterative Solution of Constrained Differential/Algebraic Systems. Research Report. ACSE Report 52 . Department of Control Engineering, University of Sheffield, Mappin Street, Sheffield

ITERATIVE SOLUTION OF CONSTRAINED DIFFERENTIAL/ALGEBRAIC SYSTEMS

by

D. H. OWENS, B.Sc.,A.R.C.S.,Ph.D.,A.F.I.M.A.

and

R. P. JONES, B.Sc.,M.Sc.Tech.,Grad.I.M.A.

Department of Control Engineering,
University of Sheffield,
Mappin Street,
Sheffield S1 3JD

Research Report No. 52

December 1976

1. INTRODUCTION

The dynamics of a large class of engineering systems can be approximately described by coupled algebraic and differential equations of the form (e.g. see appendix A.1)

$$\dot{x}(t) = f(x(t),u(t),t), \quad x(o) = x_o \tag{1}$$

$$g(x(t),u(t),t) = 0 \tag{2}$$

where $x \in R^n$, $u \in R^\ell$, $g : R^n x R^\ell x[0,T] \to R^m$ and $f : R^n x R^\ell x[0,T] \to R^n$, e.g. the dynamics of a thermal nuclear reactor (Owens, 1973). A discrete representation of such systems takes the form

$$x'(k+1) = f'(x'(k),u'(k),k) \;, \quad k = 0,1,\ldots,N-1 \tag{1'}$$

$$g'(x'(k),u'(k),k) = 0 \;, \quad k = 0,1,\ldots,N \tag{2'}$$

where, now, $g':R^n x R^\ell \to R^m$, $f':R^n x R^\ell \to R^n$, $x'(k) \in R^n$, $k = 0,1,\ldots,N$, and $u(k) \in R^\ell$, $k = 0,1,\ldots,N$. The algebraic equations (2) (and equivalently (2')) are, typically, formed by neglecting fast stable time constants and may also incorporate other algebraic constraints. An important feature of such systems is that the equations may have no solution, a unique solution or an infinite number of solutions each of which is not necessarily isolated. The problem discussed in this paper is, given state and control constraints of the form

$$x(t) \in \Omega_x(t) \;, \quad u(t) \in \Omega_u(t) \;, \quad t \in [0,T] \;, \tag{3}$$

in the continuous case, or

$$x'(k) \in \Omega'_x(k) \;, \quad u'(k) \in \Omega'_u(k) \;, \quad k = 0,1,\ldots,N \;, \tag{3'}$$

for the discrete case, find a solution pair (x,u) of (1)-(3) (or (1')-(3') if one exists. It is assumed that any solution satisfying the equations and constraints given above is acceptable as a solution to the engineering problem. In general, it is not possible to achieve an analytic solution and so iterative techniques are required to generate a sequence (x_j, u_j) tending to a limit (x,u), where (x,u) lies in the solution set defined by (1)-(3) (or (1')-(3')).

One method of solution is to embed the above problem in an optimal control setting, i.e. to solve the system of equations (1)-(3) (respectively (1')-(3')) whilst minimising the cost criterion

$$J(u) = \int_o^T L(x,u,t)dt \quad (\text{respectively, } J'(u'_o, u'_1, \ldots, u'_{N-1}) = P(x'_o, \ldots, x'_N, u'_o, \ldots u'_{N-1}))$$

An approach along these lines has been suggested (Owens, 1973) where the minimum of $J(u) = \int_o^T \langle g(x,u,t), Q(t)g(x,u,t)\rangle \, dt$, $Q(t) > 0$ for all $t \in [0,T]$, is sought subject to the constraints (1) and (3). This approach is somewhat artificial and numerical problems with the minimisation algorithm can arise. It is important to realise that the problem is not in itself an optimal control problem, although optimisation techniques may help in its solution.

This paper presents a method for the systematic iterative solution of a linearised form of the above system (1)-(3)

$$\dot{x}(t) = Ax(t) + Bu(t) \quad , \quad x(o) = x_o \tag{4}$$

$$Eu(t) + Fx(t) = 0 \tag{5}$$

$$u(t) \in \Omega_u(t) \quad , \quad x(t) \in \Omega_x(t) \quad , \quad 0 \leq t \leq T \tag{6}$$

(and the equivalent linearised form of the discrete system (1')-(3')), which has guaranteed convergence in a well-defined computational sense, and requires only standard Riccati and minimisation routines for implementation. The formulation is quite general and can also be applied to linear, constrained algebraic problems, continuous problems with integral constraints and, in fact, any problem where the solution set is the intersection of two closed convex sets in a suitable real Hilbert space.

2. PROBLEM FORMULATION

Let H be a real Hilbert space with $K_1 \subset H$, $K_2 \subset H$ two closed convex sets representing the system constraints and consider the general problem of finding a point $y \in K_1 \cap K_2$. A sequence $(y_j) \subset H$ is sought having one of the following properties:-

(i) $y_j \to y^*$, in the sense of the norm, for some $y^* \in K_1 \cap K_2$

(ii) for each real number $\varepsilon > 0$, there exists an integer N such that whenever $j \geq N$, $\max \{\inf_{Z \in K_1} \|y_j - Z\|, \inf_{Z \in K_2} \|y_j - Z\|\} < \varepsilon$.

Property (i) represents the case of strong convergence where it is possible to construct a sequence (y_j) whose strong limit lies in $K_1 \cap K_2$. This paper considers the case where convergence, in the strong sense, to an element of $K_1 \cap K_2$ cannot necessarily be guaranteed, although it is possible to construct a sequence which will generate a point arbitrarily close to both K_1 and K_2. This case is represented by property (ii) and is an acceptable form of convergence in the engineering problem in the sense that the system constraints are satisfied to an arbitrary accuracy.

Specific examples of the general problem defined above are:-

(1) The linear system defined by (4)-(6), with Ω_u, Ω_x closed convex sets, and $H = L_2^n[0,T] \times L_2^\ell[0,T]$, with associated norm

$$\|(x,u)\| = \{ \int_0^T (x^TQ(t)x + u^TR(t)u)\,dt\}^{\frac{1}{2}}, \quad Q(t)>0, \ R(t)>0 \ \forall\, t \in [0,T].$$

Here $K_1 = \{(x(t),u(t)) \in H : x(t) \in \Omega_x(t),\ u(t) \in \Omega_u(t),\ \text{a.e.},\ t \in [0,T]\}$, which is a closed convex set, and

$$K_2 = \{(x(t),u(t)) \quad H : x(t) = e^{At}x(o) + \int_0^t e^{A(t-s)}Bu(s)\,ds \text{ and}$$

$$Eu(t) + Fx(t) \equiv 0\},$$ which is a closed linear variety in H.

(2) The problem of choosing u(t) such that x(t) satisfies

$$\dot{x}(t) = ax(t) + bu(t) \quad , \quad x(o) = x_o, \; x(T) = x_f$$

subject to the constraint $|u(t)| \leqslant 1 \; \forall \; t \in [0,T]$. Noting that $x(T) - \exp(aT)x_o = \int_o^T \exp(-a\tau)bu(\tau)d\tau$ and defining $H = L_2[0,T]$, then

$$K_1 = \{u \in H : \int_o^T \exp(-a\tau)bu(\tau)d\tau = x_f - \exp(aT)x_o\} \text{ and}$$

$$K_2 = \{u \in H : |u(t)| < 1 \; \forall \; t \in [0,T]\}$$

(3) The problem of obtaining a solution of the linear algebraic equation Ax = d, where $d \in \mathbb{R}^\ell$, $x \in \mathbb{R}^n$, $\ell<n$, satisfying the constraints $|x_j - \hat{x}_j| \leqslant r_j$, $1 \leqslant j \leqslant q$, for some $q \leqslant n$. Here, $H = \mathbb{R}^n$,

$$K_1 = \{x \in H : Ax = d\} \text{ and } K_2 = \{x \in H : |x_j - \hat{x}_j| \leqslant r_j, \; 1 \leqslant j \leqslant q\}.$$

3. ITERATIVE SOLUTION VIA SEQUENTIAL PROJECTION

This Hilbert space formulation of the problem enables the simple geometric ideas of orthogonal projection (see, for example, Luenberger, 1969) to be utilised in the development of an algorithm for its solution. In this section an iterative scheme, based upon sequential application of the Projection Theorem, is developed.

The general problem outlined in section 2 is first considered, the results being presented in the following theorem.

Theorem 1 Let $K_1 \subset H$, $K_2 \subset H$, be two closed convex sets in a real Hilbert space H with $K_1 \cap K_2$ nonempty. Define

$$K_j = \begin{cases} K_1 \quad , \quad j \text{ odd} \\ K_2 \qquad j \text{ even} \end{cases}$$

Then, given the initial guess $k_o \in H$, the sequence (k_j), $j = 0,1,2,...,$ given by

$$\|k_j - k_{j-1}\| = \inf_{k \in K_j} \|k-k_j\| , \qquad j \geqslant 1 \tag{7}$$

with $k_j \in K_j$, $j \geqslant 1$, is uniquely defined for each $k_o \in H$ and satisfies

$$\|k_{j+1} - k_j\| < \|k_j - k_{j-1}\| , \qquad j \geqslant 2 \tag{8}$$

Furthermore, for any $x \in K_1 \cap K_2$,

$$\sum_{j=1}^{\infty} \|k_{j+1} - k_j\|^2 \leqslant \|x - k_1\|^2 \tag{9}$$

and, hence, for each $\varepsilon>0$, there exists an integer N such that for $j \geqslant N$

$$\inf_{k \in K_{j+1}} \|k - k_j\| < \varepsilon \tag{10}$$

<u>Proof</u>

Since K_j is a closed convex set in a Hilbert space, then, given $k_j \in K_j$, the Projection Theorem guarantees the existence of a well-defined and unique $k_{j+1} \in K_{j+1}$ such that $\|k_{j+1}-k_j\| \leqslant \|x-k_j\|$ for all $x \in K_{j+1}$ proving uniqueness. Moreover, for any $x \in K_{j+1}$, $<x-k_{j+1}, k_j-k_{j+1}> \leqslant 0$ and, in particular, $<k_j-k_{j+1}, k_{j+1}-k_{j-1}> \geqslant 0$ and, hence,

$$\|k_j-k_{j-1}\|^2 = \|k_j-k_{j+1}\|^2 + \|k_{j+1}-k_{j-1}\|^2 + 2<k_j-k_{j+1}, k_{j+1}-k_{j-1}>$$

$$> \|k_j-k_{j+1}\|^2 ,$$

which verifies (8).

If $x \in K_1 \cap K_2$, then $<x-k_{j+1}, k_{j+1}-k_j> \geqslant 0$ for all j, and so

$$\|x-k_j\|^2 = \|x-k_{j+1}\|^2 + \|k_{j+1}-k_j\|^2 + 2<x-k_{j+1}, k_{j+1}-k_j>$$

$$\geqslant \|x-k_{j+1}\|^2 + \|k_{j+1}-k_j\|^2$$

An induction argument then gives

$$||x-k_1||^2 \geqslant ||x-k_j||^2 + \sum_{\ell=1}^{j-1} ||k_{\ell+1}-k_\ell||^2$$

for all j, so that (in the limit)

$$||x-k_1||^2 \geqslant \sum_{\ell=1}^{\infty} ||k_{\ell+1}-k_\ell||^2$$

as required. (10) now follows from this result.

Q.E.D.

Theorem 1 presents an iterative scheme satisfying the convergence criterion, property (ii) of section 2, and, using equation (8), each iteration is a better approximation to the solution of the problem. This scheme, based on the basic geometrical concept of orthogonal projection, is outlined in Figure 1.

Figure 2 describes the case where the tangent hyperplanes to K_1 and K_2 at the points k_i and k_{i+1} are nearly parallel and suggests that in this case convergence will be slow. The question of whether the scheme can be modified to incorporate some form of extrapolation parameter to speed up convergence is investigated in the next theorem. Attention is restricted to the case where K_2 is a closed linear variety and a modified scheme is outlined in Figure 3. Note that the result reduce to theorem 1 if $\lambda_i = 1$, $i \geqslant 1$.

Theorem 2 Let $K_1 \subset H$ be a closed convex set and $K_2 = a+M$, $K_2 \subset H$, a closed linear variety in a real Hilbert space H such that $K_1 \cap K_2$ is nonempty. ($M \subset H$ is a closed subspace and $a \in H$). Then, given $r_1 \in K_2$, a sequence $\{r_1,k_1,s_1,r_2,k_2,s_2,\ldots\}$ given by

$$||k_i-r_i|| = \inf_{y\in K_1} ||y-r_i|| \quad , \qquad k_i \in K_1 \ , \tag{11}$$

$$\|s_i-k_i\| = \inf_{y\in K_2} \|y-k_i\| \quad , \qquad s_i \in K_2 \ , \tag{12}$$

and

$$r_{i+1} = r_i + \lambda_i(s_i-r_i) \ , \tag{13}$$

with

$$1 \leqslant \lambda_i \leqslant \frac{\|k_i-r_i\|^2}{\|s_i-r_i\|^2} \ , \tag{14}$$

is well-defined for each $r_1 \in K_2$. Furthermore,

$$\|r_1-x\|^2 \geqslant \sum_{j=1}^{\infty} \|k_j-r_j\|^2 \tag{15}$$

and, hence, for each $\varepsilon>0$ there is an integer N such that for $j \geqslant N$

$$\inf_{y\in K_1} \|y-r_j\| < \varepsilon \tag{16}$$

Proof

Given $r_i \in K_2$, then since K_1 is a closed convex set and K_2 a closed linear variety in a Hilbert space, the Projection Theorem guarantees the existence and uniqueness of a k_i and s_i satisfying (11) and (12), respectively. Furthermore, $\langle r_i-k_i, x-k_i\rangle \leqslant 0$ for all $x \in K_1$ and $k_i-s_i \perp M$. It is therefore only necessary to show that λ_i, as given in (14), is well defined, i.e. that $\|s_i-r_i\|^2 > 0$ and $\|k_i-r_i\|^2 \geqslant \|s_i-r_i\|^2$. It is assumed that $\|k_i-r_i\|^2 > 0$ since otherwise $k_i = r_i$ and the algorithm has converged. For this case, suppose that $\|s_i-r_i\|^2 = 0$, i.e. $s_i = r_i$. Then, for all $x \in K_1$, $\langle x-k_i, k_i-s_i\rangle \geqslant 0$ and, for all $x \in K_2$, $\langle x-s_i, s_i-k_i\rangle = 0$ or, equivalently, $\langle x-k_i, k_i-s_i\rangle = -\|k_i-s_i\|^2 < 0$ ie $K_1 \cap K_2$ is empty contrary to assumption. Noting that

$$\begin{aligned}\|k_i-r_i\|^2 &= \|k_i-s_i\|^2 + \|s_i-r_i\|^2 + 2\langle k_i-s_i, s_i-r_i\rangle \\ &= \|k_i-s_i\|^2 + \|s_i-r_i\|^2 \ ,\end{aligned}$$

since $s_i - r_i \in M$ and $k_i - s_i \perp M$, and so, if the algorithm has not converged, $\|k_i - r_i\|^2 > \|s_i - r_i\|^2$.

Now let $x \in K_1 \cap K_2$ and consider

$$\langle r_{i+1} - r_i, r_i - x\rangle = \lambda_i \langle s_i - r_i, r_i - x\rangle = \lambda_i \langle s_i - k_i + k_i - r_i, r_i - x\rangle$$

$$= \lambda_i \langle k_i - r_i, r_i - x\rangle,$$

since $r_i - x \in M$ and $s_i - k_i \perp M$. Then

$$\langle r_{i+1} - r_i, r_i - x\rangle = \lambda_i \langle k_i - r_i, r_i - k_i + k_i - x\rangle$$

$$= -\lambda_i \|k_i - r_i\|^2 + \lambda_i \langle k_i - r_i, k_i - x\rangle$$

$$\leqslant -\lambda_i \|k_i - r_i\|^2,$$

by the definition of k_i. Also, for λ_i satisfying (14),

$$\lambda_i \|k_i - r_i\|^2 = \lambda_i \frac{\|k_i - r_i\|^2}{\|s_i - r_i\|^2} \cdot \|s_i - r_i\|^2 \geqslant \lambda_i^2 \|s_i - r_i\|^2 = \|r_{i+1} - r_i\|^2.$$

Hence, for $x \in K_1 \cap K_2$ and for any i,

$$\|r_{i+1} - x\|^2 = \|r_i - x\|^2 + \|r_{i+1} - r_i\|^2 + 2\langle r_{i+1} - r_i, r_i - x\rangle$$

$$\leqslant \|r_i - x\|^2 + \|r_{i+1} - r_i\|^2 - 2\lambda_i \|k_i - r_i\|^2$$

and, rearranging,

$$\|r_i - x\|^2 \geqslant \|r_{i+1} - x\|^2 + (\lambda_i \|k_i - r_i\|^2 - \|r_{i+1} - r_i\|^2) + \lambda_i \|k_i - r_i\|^2$$

$$\geqslant \|r_{i+1} - x\|^2 + \lambda_i \|k_i - r_i\|^2,$$

and, since $\lambda_i \geqslant 1$,

$$\|r_i - x\|^2 \geqslant \|r_{i+1} - x\|^2 + \|k_i - r_i\|^2.$$

An induction argument now gives,

$$\|r_1-x\|^2 \geq \|r_{i+1}-x\|^2 + \sum_{j=1}^{i} \|k_j-r_j\|^2 \text{ , for all } i \text{, and so}$$

$$\|r_1-x\|^2 \geq \sum_{j=1}^{\infty} \|k_j-r_j\|^2 \text{ , as required.}$$ (16) now follows immediately.

Q.E.D.

The iterative schemes presented in Theorems 1 and 2 will not, in a general Hilbert space, converge to a solution in a finite number of iterations. It can be shown, however, that, for the case where K_1 is a closed hyperplane and K_2 a closed linear variety convergence can be obtained in one iteration. In this case it is, in fact, possible to obtain a minimum norm solution. These results are formalised in the following theorem.

Theorem 3 Let $K_1 = \{x \in H: \langle\alpha, x-\alpha\rangle = 0, \alpha \in H, \|\alpha\| > 1\}$ be a closed hyperplane in a Hilbert space H and define K_2 as in Theorem 2. Given $r_1 \in K_2$, then for k_1 and s_1 as defined in (11) and (12) of Theorem 2, respectively,

$$r_2 = r_1 + \frac{\|k_1-r_1\|^2}{\|s_1-r_1\|^2}(s_1-r_1) \in K_1 \cap K_2 \ .$$

Furthermore, if $\|r_1\| \leq \|y\|$ for all $y \in K_2$, then $\|r_2\| \leq \|x\|$ for all $x \in K_1 \cap K_2$.

Proof

By translation, and without loss of generality, it is assumed that $r_1 = 0$ so that, from the definition of k_1, $\langle k_1, x-k_1\rangle = 0$ for all $x \in K_1$ and, hence, $\langle k_1, x-k_1\rangle = 0$ is an alternative definition of K_1. Since, by construction, $\langle s_1, k_1-s_1\rangle = 0$, it follows that $\langle s_1, k_1\rangle = \|s_1\|^2$ and so

$$\langle k_1, r_2 - k_1\rangle = \langle k_1, \frac{\|k_1\|^2}{\|s_1\|^2} s_1 - k_1\rangle = \frac{\|k_1\|^2}{\|s_1\|^2}\langle k_1, s_1\rangle - \|k_1\|^2 = 0$$

which implies that $r_2 \in K_1$. By definition, $r_2 \in K_2$ and hence $r_2 \in K_1 \cap K_2$.

If $y \in K_1 \cap K_2$, then

$$\langle y - r_2, r_2\rangle = \frac{\|k_1\|^2}{\|s_1\|^2}\langle y - \frac{\|k_1\|^2}{\|s_1\|^2} s_1, s_1\rangle = \frac{\|k_1\|^2}{\|s_1\|^2}\{\langle y, s_1\rangle - \|k_1\|^2\}$$

$$= \frac{\|k_1\|^2}{\|s_1\|^2}\{\langle y, s_1 - k_1\rangle + \langle y, k_1\rangle - \|k_1\|^2\}$$

and, since $s_1 - k_1 \perp K_2$,

$$\langle y - r_2, r_2\rangle = \frac{\|k_1\|^2}{\|s_1\|^2}\{\langle y, k_1\rangle - \|k_1\|^2\} = \frac{\|k_1\|^2}{\|s_1\|^2}\langle y - k_1, k_1\rangle = 0 ,$$

by definition of k_1. It now follows that

$$\|y\|^2 = \|y - r_2\|^2 + \|r_2\|^2 + 2\langle y - r_2, r_2\rangle = \|y - r_2\|^2 + \|r_2\|^2 \geqslant \|r_2\|^2$$

as required.

In the general case with $r_1 \neq 0$,

$$\|y - r_1\|^2 = \|y - r_2\|^2 + \|r_2 - r_1\|^2 ,$$

and if $\|r_1\| \leqslant \|y\|$ for all $y \in K_2$ it follows that $r_1 \perp M$. Therefore, since $y - r_1 \in M$,

$$\|y\|^2 = \|y - r_1\|^2 + \|r_1\|^2 = \|y - r_2\|^2 + \|r_2 - r_1\|^2 + \|r_1\|^2$$

$$\geqslant \|r_2 - r_1\|^2 + \|r_1\|^2 = \|r_2\|^2 ,$$

as $r_2 - r_1 \in M$.

Comments

1. In general, the existence of a strong limit point to the sequence generated in Theorems 2 and 3 has not been established. This is of interest theoretically but is of little consequence from a practical point of view since it has been demonstrated that the convergence property (ii) of section 2 is satisfied. However, for the case where H is a finite dimensional space, a proof along the following lines can be obtained.

For any $K_1 \cap K_2$ and r_i defined by Theorem 2, $\|r_1-x\|^2 \geq \|r_i-x\|^2$ for all i. Furthermore $\|r_i\| \leq \|r_i-x\| + \|x\| \leq \|r_1-x\| + \|x\|$, for all i, and so the sequence (r_i) is bounded. Then, as H is a finite dimensional space, (r_i) is relatively compact and has at least one cluster value $r \in K_1 \cap K_2$. If r and $\hat{r}$ are distinct cluster values of the sequence (r_i) then there are subsequences (r_{i_k}) and (r_{i_ℓ}) of (r_i) such that $(r_{i_k}) \to r$ and $(r_{i_\ell}) \to \hat{r}$. Defining $\varepsilon = \frac{1}{2}\|r-\hat{r}\|$, there exists an integer N such that for $k,\ell>N$, $r_{i_k} \in B(r,\varepsilon)$ and $r_{i_\ell} \in B(\hat{r},\varepsilon)$, where $B(x,\varepsilon)$ is the open ball centred on x with radius ε. Taking $i_k > i_\ell$, for some $k,\ell>N$, it follows that $\|r_{i_k}-\hat{r}\| > \|r_{i_\ell}-\hat{r}\|$, since $r_{i_k} \in B(r,\varepsilon)$, $r_{i_\ell} \in B(\hat{r},\varepsilon)$ and $\|r-\hat{r}\| = 2\varepsilon$, contradicting the result that $\|r_{i+1}-x\| < \|r_i-x\|$ for all $x \in K_1 \cap K_2$ (see proof of Theorem 2). Then (r_i) must have a unique cluster value r and hence $(r_i) \to r \in K_1 \cap K_2$.

2. If K_1 and K_2 are disjoint, the algorithm defined by Theorem 2 with $\lambda_i \gg 1$ may exhibit wild oscillation, as illustrated in Figure 4. With $\lambda_i = 1$, however, the algorithm is well behaved and, intuitively, converges to points $r_1 \in K_1$, $r_2 \in K_2$ defining the minimum distance between the two sets. A proof of this observation for the case where H is finite dimensional now follows.

Taking $\lambda_i = 1$ for all i, Theorem 2 gives $\|r_{i+1}-k_{i+1}\| \leq \|r_{i+1}-k_i\| \leq \|r_i-k_i\| \leq \|r_1-k_1\|$ which implies that $\beta_i = \|r_i-k_i\|$ has limit β and, since H is finite dimensional, the sequence (r_i-k_i) has cluster values $r \in K_2$, $k \in K_1$

with $\beta = \|k-r\| > 0$, for each k_i, r_i, $\langle k'-k_i, k_i-r_i\rangle > 0$ for all $k' \in K_1$, giving $\|k-r\| = \inf_{y \in K_1} \|y-r\|$ and, by similar reasoning, $\|k-r\| = \inf_{y \in K_2} \|k-y\|$.

Then, for all $k' \in K_1$, $r' \in K_2$,

$$\|k'-r'\|^2 = \|k'-k+k-r+r-r'\|^2 = \|k'-k\|^2 + \|k-r\|^2 + \|r-r'\|^2 + 2\{\langle k'-k, k-r\rangle$$

$$+\langle k'-k, r-r'\rangle + \langle k-r, r-r'\rangle\} \geqslant \|k-r\|^2 + \|k'-k\|^2 + \|r-r'\|^2 + 2\langle k'-k, r-r'\rangle$$

$$= \|k-r\|^2 + \|k'-k+r-r'\|^2 \geqslant \|k-r\|^2 \quad \text{as required.}$$

3. The introduction of an extrapolation parameter $\lambda_i >> 1$ can cause numerical errors introduced into the calculation at each iteration to be magnified at successive iterations. For, if errors ε_i^r, ε_i^s are introduced at the i-th iteration in the calculation of r_i and s_i, respectively, then (14) gives

$$\varepsilon_{i+1}^r = (1-\lambda_i)\varepsilon_i^r + \lambda_i \varepsilon_i^s$$

and, for $\lambda_i > 1$ and under worst-case conditions (i.e. $\varepsilon_i^r = -\varepsilon_i^s$), $\|\varepsilon_{i+1}^r\| = (2\lambda_i - 1)\|\varepsilon_i^s\| >> \|\varepsilon_i^s\|$ if $\lambda_i >> 1$. In practice this problem can be removed by setting $\lambda_i = 1$ every few iterations to reset the magnitude of the computational errors.

4. In general, the sequence (r_i, k_i) does not converge to a minimum norm solution as a simple three dimensional example will testify. For the case where K_1 is a closed hyperplane, a minimum norm solution is obtained, however, (Theorem 3) and the algorithm converges in one iteration.

4. <u>EXAMPLES</u>

1. Consider the system $\dot{x}(t) = u(t)$, $x(o) = 0$, $x(1) = 1$, where $u(t)$ is constrained to satisfy $\int_o^1 tu(t)dt = 1$. Defining $H = L_2[0,1]$ with

$||f||^2 = \frac{1}{2}\int_0^1 f^2(t)dt$, $f \in H$, $K_1 = \{u \in H : \int_0^1 tu(t)dt = 1\}$ and $K_2 = \{u \in H : \int_0^1 u(t)dt = 1\}$, then K_1 is a closed hyperplane and K_2 a closed linear variety in H. Application of Theorem 3 and Pontryagin's Minimum Principle therefore gives:

$r_1(t) = 1$ for all t $|0,1|$, $k_1(t) = 1 + \frac{3}{2}t$ and $s_1(t) = \frac{1}{4} + \frac{3}{2}t$. Then $\lambda_1 = (\frac{3}{4})/(\frac{3}{16})$ and so $r_2(t) = 1 + 4\{\frac{1}{4} + \frac{3}{2}t - 1\} = -2+6t$. In this case, $r_2 \in K_1 \cap K_2$ and is of minimum norm, i.e. $u = r_2$ is also a solution of the associated minimum energy problem.

2. Consider, now, the discrete form of the problem outlined in equations (4)-(6) of section 2. Given a system described by equations of the form

$$x(k+1) = \Phi x(k) + \Delta u(k) \;, \quad x(o) = x_o \;, \; k = 0,1,\ldots,N-1 \;, \tag{17}$$

$$Eu(k) + Fx(k) = 0 \;, \qquad k = 0,1,\ldots,N \tag{18}$$

where $x(k) \in \mathcal{R}^n$, $u(k) \in \mathcal{R}^\ell$, $k = 0,1,\ldots,N$, and Φ_{nxn}, $\Delta_{nx\ell}$, $E_{mx\ell}$, F_{mxn} $(m<n)$ are real matrices, a solution (x,u) of (17) and (18) is sought satisfying the constants $x(k) \in \Omega_x$, $u(k) \in \Omega_u$, $k = 0,1,\ldots,N$, where Ω_x and Ω_u are two closed convex sets. In the example discussed here, x is unconstrained (i.e. $\Omega_x = \mathcal{R}^n$) and Ω_u is defined by

$$\Omega_u = \{u \in \mathcal{R}^\ell : u_i^{min} \leq u_i \leq u_i^{max} \;, \quad i = 1,2,\ldots,\ell\} \tag{19}$$

However, the ideas are trivially extended to include convex state constraints.

The problem formulation is analogous to that outlined in example (1) of section 2 and is not repeated here. The spaces are, of course, now finite-dimensional and a suitable norm defined as

$$||(x,u)||^2 = \frac{1}{2} \sum_{k=o}^{N} \{x^T(k)Qx(k) + u^T(k)Ru(k)\} , \quad Q>0, R>0 .$$

The results of Theorem 2 are employed:

For each i, given $k_i = (x^{ref},u^{ref}) \in K_1$, $s_i = (x,u) \in K_2$ is calculated as $\min_{y \in K_2} ||y-k_i||$, i.e.

$$\min_{(x,u)} \frac{1}{2} \sum_{k=o}^{N} \{[x(k)-x^{ref}(k)]^T Q[x(k)-x^{ref}(k)]+[u(k)-u^{ref}(k)]^T R[u(k)-u^{ref}(k)]\}$$

where (x,u) satisfies equations (17) and (18). This is a linear quadratic optimal control problem and has solution

$$u(N-k) = -K(k)x(N-k) + g(k) \quad , \qquad k = 0,1,\ldots,N , \qquad (20)$$

with x(k) given by equation (17). Expressions for the 'Riccati matrix' K(k) and 'tracking vector' g(k) are given in appendix A2. An initial estimate $r_1 = (x,u)$ can be obtained setting $x^{ref} = 0$, $u^{ref} = 0$.

Given $r_i = (x^{ref},u^{ref}) \in K_2$, k_i is calculated as $\min_{y \in K_1} ||y-r_i||$, i.e.

$$\min_{(x,u)} \frac{1}{2} \sum_{k=o}^{N} \{[x(k)-x^{ref}(k)]^T Q[x(k)-x^{ref}(k)]+[u(k)-u^{ref}(k)]^T R[u(k)-u^{ref}(k)]\}$$

with u satisfying the constraint (19), which, if R is diagonal, has solution

$$u_i = \begin{cases} u_i^{max} , & \text{for } u_i^{ref} \geqslant u_i^{max} \\ u_i^{ref} , & \text{for } u_i^{min} \leqslant u_i^{ref} \leqslant u_i^{max} \\ u_i^{min} , & \text{for } u_i^{ref} \leqslant u_i^{min} \end{cases} , \quad i = 1,\ldots,\ell, \text{ and } x = x^{ref}$$

Hence computation of k_i simply involves 'clipping off' the components of r_i where they violate the constraint set Ω_u.

r_{i+1} is now generated by $r_{i+1} = r_i+\lambda_i(s_i-r_i)$, for suitable λ_i, $1 \leqslant \lambda_i \leqslant ||k_i-r_i||^2 / ||s_i-r_i||^2$ and the iterative process repeated until numerical convergence is obtained.

A solution to equations (17) and (18) satisfying constraints of the form (19) can therefore be generated by iterative application of equations (17) and (20) and 'clipping off' the resulting control trajectories where they violate the constraints. The 'Riccati matrix' K is independent of (x^{ref},u^{ref}) and need only be calculated once whereas the 'tracking vector' g has to be updated at each iteration. Two numerical examples of the application of this algorithm are given below.

(a) 4-th order integrator plant. It is desired to find a solution (x,u) of the algebraic/differential system

$$\dot{x}_1 = x_2 , \qquad x_1(o) = -0.5 ,$$
$$\dot{x}_2 = x_3+u_1 , \qquad x_2(o) = 0.5 ,$$
$$\dot{x}_3 = x_4+u_2 , \qquad x_3(o) = -0.5 ,$$
$$\dot{x}_4 = u_3 , \qquad x_4(o) = 0.5 ,$$

$$x_1 + x_2 + u_1 + u_2 + u_3 = 0 ,$$
$$x_3 + x_4 + u_1 - u_2 + u_3 = 0 ,$$

defined on the time interval $[0,1]$, subject to the constraint $u_3 \geqslant 0$. The time interval is divided into N steps of length $h = 1/N$ and the problem put into discrete form with $\Phi = I + hA + \frac{h^3A^2}{2!} + \ldots$, $\Delta = hI + \frac{h^2A}{2!} + \frac{h^3A^2}{3!} + \ldots$

20 time steps were employed and the weighting matrices Q,R in the norm were taken to be $Q = I_4$, $R = I_3$. The extrapolation factor λ_i was set at $\lambda_i = \|k_i-r_i\|^2/\|s_i-r_i\|^2$, throughout, and convergence to an accurate solution in 8 iterations is shown in Table 1 in terms of variation in λ_i and distance between K_1 and K_2 with iteration. The initial and final trajectories of u_3 are given in Figure 5, Figure 6 describing the corresponding plant outputs x_1.

(b) Nuclear Reactor Control Problem. Large thermal nuclear power reactors can exhibit unstable or underdamped oscillations in the power distribution, with periods of 30-40 hours, due to the effects of the fission product poison xenon-135 (Owens, 1973). The dynamics of such systems can be approximated by equations of the form

$$\dot{x}(t) = Ax(t) + Bu(t) \quad , \qquad x(o) = x_o \ ,$$
$$Eu(t) + Fx(t) = 0 \quad ,$$

where $x(t) \in \mathbb{R}^{2m}$, $u(t) \in \mathbb{R}^{m+n}$, $t \in [0,T]$, and $A_{2m\times 2m}$, $B_{2m\times(m+n)}$, $E_{(m+1)\times(m+n)}$, $F_{(m+1)\times 2m}$ are real matrices. $x(t)$ represents the internal states inherent in the system, namely, xenon and its precursor iodine, and the power distribution and control rod reactivity are lumped together in $u(t)$. A typical one-dimensional model has system matrices

$$A = \begin{pmatrix}
-0.29\times10^{-4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.29\times10^{-4} & -0.119\times10^{-3} & 0 & 0 & 0 & 0.195\times10^{-4} & 0 & 0 \\
0 & 0 & -0.29\times10^{-4} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.29\times10^{-4} & -0.991\times10^{-4} & 0 & 0 & 0 & 0.223\times10^{-4} \\
0 & 0 & 0 & 0 & -0.29\times10^{-4} & 0 & 0 & 0 \\
0 & 0.195\times10^{-4} & 0 & 0 & 0.29\times10^{-4} & -0.963\times10^{-4} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -0.29\times10^{-4} & 0 \\
0 & 0 & 0 & 0.223\times10^{-4} & 0 & 0 & 0.29\times10^{-4} & -0.954\times10^{-4}
\end{pmatrix}$$

$$B = \begin{pmatrix}
0.112\times10^{-3} & 0 & 0 & 0 & 0 & 0 & 0 \\
-0.902\times10^{-4} & 0 & 0.503\times10^{-5} & 0 & 0 & 0 & 0 \\
0 & 0.112\times10^{-3} & 0 & 0 & 0 & 0 & 0 \\
0 & -0.852\times10^{-4} & 0 & 0.762\times10^{-5} & 0 & 0 & 0 \\
0 & 0 & 0.112\times10^{-3} & 0 & 0 & 0 & 0 \\
0.503\times10^{-5} & 0 & -0.83\times10^{-4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.112\times10^{-3} & 0 & 0 & 0 \\
0 & 0.762\times10^{-5} & 0 & -0.817\times10^{-4} & 0 & 0 & 0
\end{pmatrix}$$

$$E =$$

$$\begin{pmatrix} -0.297\text{x}10^{-4} & 0 & 0.594\text{x}10^{-5} & 0 & 0.461\text{x}10^{-1} & 0.128\text{x}10^{-1} & 0.128\text{x}10^{-1} \\ 0 & -0.443\text{x}10^{-4} & 0 & 0.679\text{x}10^{-5} & 0 & 0.156\text{x}10^{-1} & -0.156\text{x}10^{-1} \\ 0.594\text{x}10^{-5} & 0 & -0.777\text{x}10^{-4} & 0 & -0.802\text{x}10^{-2} & 0.62\text{x}10^{-2} & 0.62\text{x}10^{-2} \\ 0 & 0.679\text{x}10^{-5} & 0 & -0.125\text{x}10^{-3} & 0 & -0.81\text{x}10^{-2} & 0.81\text{x}10^{-2} \\ 1.0 & 0 & 0.333 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$F =$$

$$\begin{pmatrix} 0 & -0.102\text{x}10^{-3} & 0 & 0 & 0 & 0.204\text{x}10^{-4} & 0 & 0 \\ 0 & 0 & 0 & -0.815\text{x}10^{-4} & 0 & 0 & 0 & 0.233\text{x}10^{-4} \\ 0 & 0.204\text{x}10^{-4} & 0 & 0 & 0 & -0.786\text{x}10^{-4} & 0 & 0 \\ 0 & 0 & 0 & 0.233\text{x}10^{-4} & 0 & 0 & 0 & -0.776\text{x}10^{-4} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

In this case $u_2(t)$ represents the dominant first spatial mode of oscillation, $u_5(t)$ a bulk control action and $u_6(t)$, $u_7(t)$ two trimming controllers. In the absence of trimming controls u_6, u_7, the mode $u_2(t)$ is unstable, and with initial condition

$$x(o) = (10^{14}, -10^{14}, -10^{14}, 10^{14}, 0, 0, 0, 0)^T \tag{21}$$

exhibits the transient shown in Figure 7, with peak magnitude $\max_t |u_2(t)| = 5.39\text{x}10^{14}$ over a time interval of 40 hours. The practical problem considered here is the choice of trimming control action $u_6(t), u_7(t)$ to ensure that the power mode $u_2(t)$, resulting from initial conditions (21), is adequately damped and satisfies the constraint

$$|u_2(t)| \leqslant 0.4 \text{ x } 10^{14}, \qquad 0 \leqslant t \leqslant 40$$

which would obviously be a considerable improvement on the open loop behaviour.

The problem was solved in discrete time as in example (a). In order to offset the difference in magnitude in the power and control components of u(t), the weighting matrices Q,R in the norm were taken to be $Q = I_8$, $R = \mathrm{diag}(1,1,1,1,10^5,10^8,10^8)$. 20 time steps were employed and the extrapolation factor was initially set at $\lambda_i = \|k_i - r_i\|^2 / \|s_i - r_i\|^2$ for each iteration i. In this case, an error growth, as predicted in section 2, was observed and the algorithm broke down. A choice of $\lambda_k = 1$ whenever $\prod_{j=\ell}^{k} \lambda_j > 10^3$, where ℓ was the previous iteration at which λ_i was set to unity, had the effect of introducing an acceptable upper bound on the growth in errors and the algorithm now converged rapidly. To ensure the highest accuracy in the solution of the equations, λ_i was also set to unity on the final (converged) iteration.

Table 2 shows the rate of convergence and variation in λ_i with iteration. The initial and final iterates for the power mode $u_2(t)$ are given in Figure 8 and Figure 9 describes the final trimming control trajectories $u_6(t), u_7(t)$. For comparison purposes, the convergence rate for the case where λ_i was set to unity throughout, is indicated in Table 3. It is noted that, for the case where extrapolation was employed, the algorithm converged to an acceptable solution in 4 iterations, whereas, with no extrapolation, convergence has not been achieved after 50 iterations.

5. CONCLUSIONS

An iterative scheme for the solution of constrained algebraic/differential systems, based upon sequential application of optimisation techniques, has been presented. The algorithm has been derived in a Hilbert space setting and the formulation is quite general. Attention has been restricted to linear systems and for this case a Riccati-type solution is obtained. In this context, it is important to realise that although optimisation procedures have been used in the solution of this problem, in general, the resulting solution is not optimal. The use of an extrapolation factor has been incorporated in the algorithm and it has been demonstrated, with the aid of a numerical example, that this can have a highly significant improvement on the convergence rate. Since this extrapolation parameter is always greater than unity, numerical errors can propogate but the scheme is easily adapted to contain such an error growth. Two illustrative control problems of moderate state dimension have been investigated and accurate solutions to both problems were obtained in a small number of iterations. Finally, it is noted that the norms used for the solution of the problem are unspecified and hence, intuitively, can be used to improve the conditioning of the algorithm.

REFERENCES

OWENS, D.H., 1973, University of London, Ph.D. Thesis.

LUENBERGER, D.G., 1969, Optimisation by Vector Space Methods, Wiley.

APPENDICES

A1. Consider a control problem governed by equations of the form

$$\dot{x}_1 = Q(x_1, x_2, t) \tag{22}$$

$$\dot{x}_2 = \psi(x_1, x_2, u_c, t) \tag{23}$$

where it is required to control the state x_2. If equation (23) is stable and has a fast acting time constant it can be reduced to the algebraic equation

$$\psi(x_1, x_2, u_c, t) = 0 \tag{24}$$

Then, if x_2 and u_c are lumped together as a pseudo 'control vector' u, i.e. $u = (x_2, u_c)^T$, and taking $x = x_1$, equations (22) and (24) can be rewritten as

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

$$g(x,u,t) = 0$$

A2. The 'Riccati matrix' K and 'tracking vector' g of equation (20) are given by the following recurrence relations

$$R(k) = \Delta^T Q(k-1)\Delta + R \quad ,$$

$$S(k) = \Delta^T Q(k-1)\Phi \quad ,$$

$$h(k) = Ru^{ref}(N-k) - \Delta^T p(k-1) \quad ,$$

$$K(k) = R^{-1}(k)S(k) + R^{-1}(k)E^T\left[ER^{-1}(k)E^T\right]^{-1}\left[F - ER^{-1}(k)S(k)\right] \quad ,$$

$$g(k) = R^{-1}(k)h(k) - R^{-1}(k)E^T\left[ER^{-1}(k)E^T\right]^{-1}ER^{-1}(k)h(k) \quad ,$$

$$k = 1, \ldots, N \quad ,$$

$$K(o) = R^{-1}E^T|ER^{-1}E^T|^{-1}F \quad ,$$

$$g(o) = -R^{-1}E^T\left[ER^{-1}E^T\right]^{-1}Eu^{ref}(N) + u^{ref}(N) \quad ,$$

where

$$Q(k) = Q + \left[\Phi - \Delta K(k)\right]^T Q(k-1)\left[\Phi - \Delta K(k)\right] + K^T(k)RK(k) \quad ,$$

$$p^T(k) = -x^{ref^T}(N-k)Q + g^T(k)\Delta^T Q(k-1)\left[\Phi - \Delta K(k)\right] + p^T(k-1)\left[\Phi - \Delta K(k)\right] - \left|g(k) - u^{ref}(N-k)\right|^T RK(k) \ ,$$

$$k = 1,\ldots,N-1 \ ,$$

and

$$Q(o) = Q + K^T(o)RK(o) \ ,$$

$$p^T(o) = -\left[g(o) - u^{ref}(N)\right]^T RK(o) - x^{ref^T}(N)Q \ .$$

ITERATION (i)	λ_i	$\|\|k_i-s_i\|\|$
1	3.68	0.351
2	3.34	0.102
3	2.62	0.263×10^{-1}
4	2.67	0.534×10^{-2}
5	2.46	0.972×10^{-3}
6	2.50	0.160×10^{-3}
7	2.40	0.282×10^{-4}
8	2.30	0.219×10^{-5}

Table 1

ITERATION (i)	λ_i	$\|\|k_i-s_i\|\|$
1	34.6	50.4
2	59.6	10.08
3	1.0	0.35×10^{-4}
4	1.0	0.107×10^{-7}

Table 2

ITERATION (i)	λ_i	$\|\|k_i-s_i\|\|$
1	1	50.4
10	1	42.8
50	1	16.9

Table 3

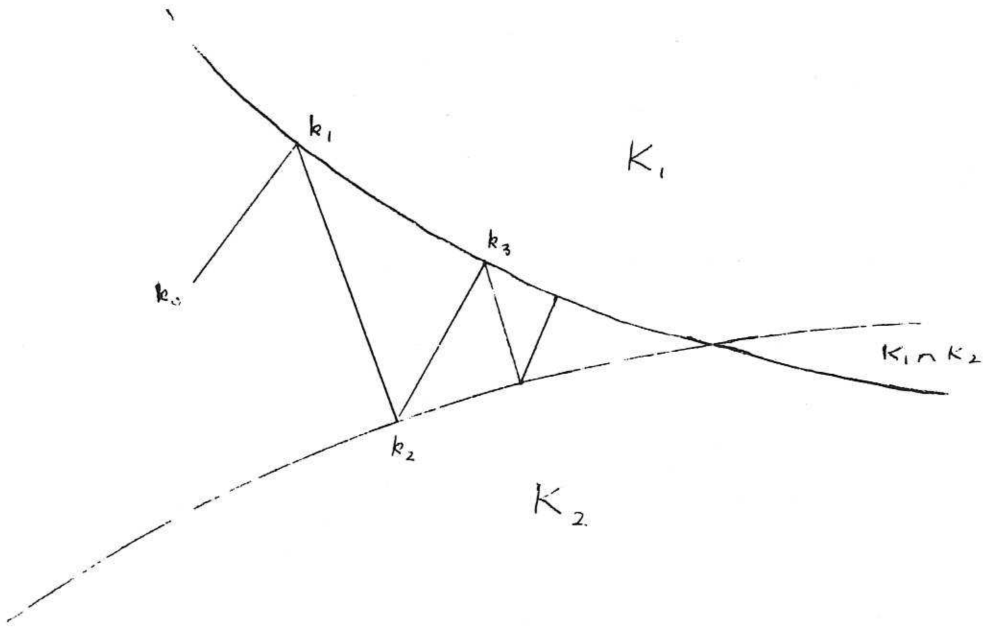

Figure 1.

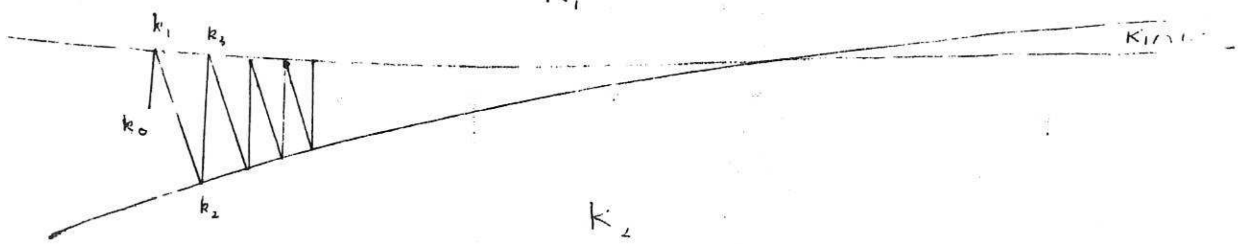

Figure 2

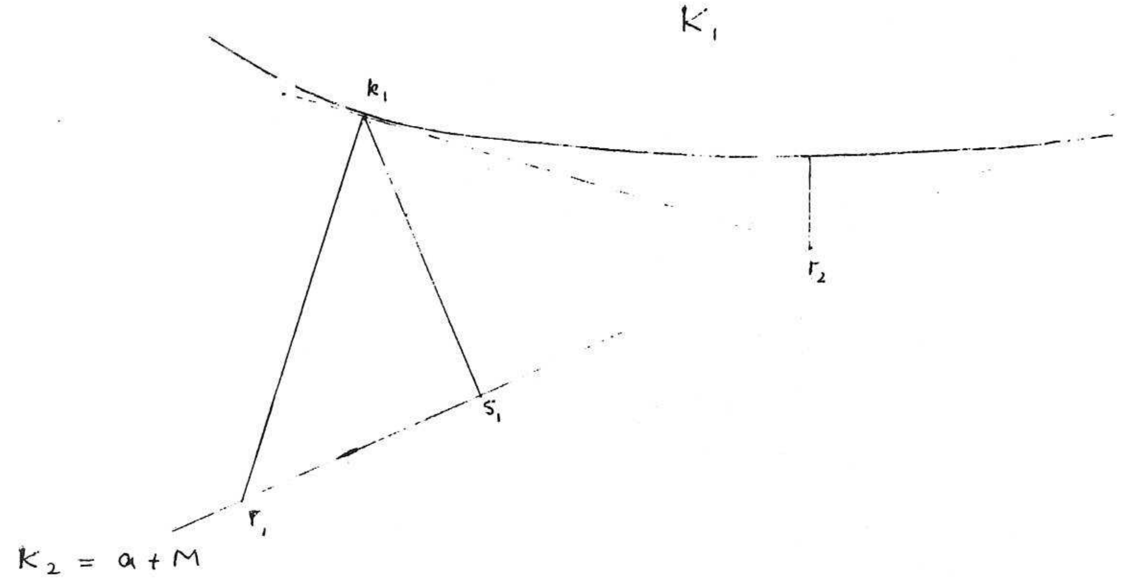

Figure 3

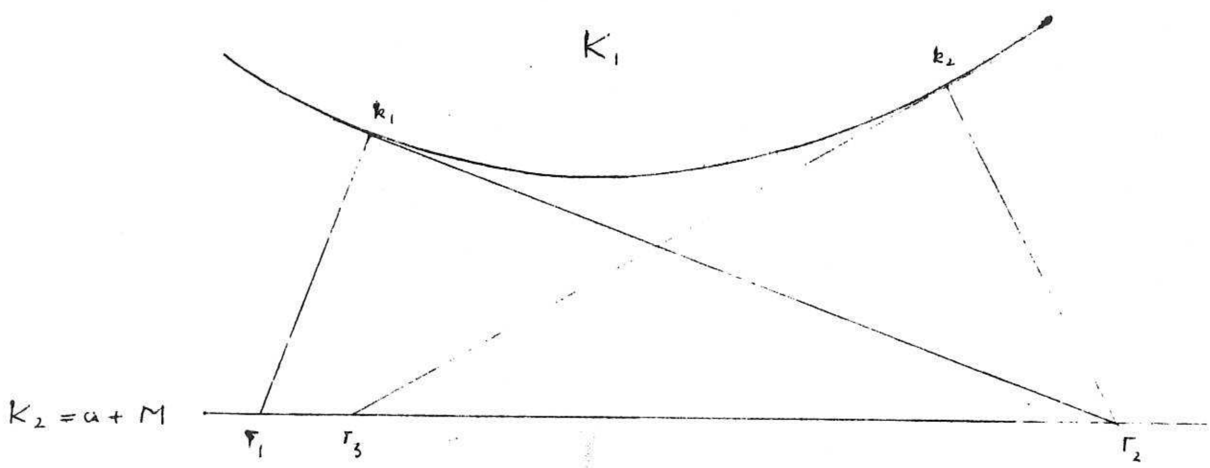

Figure 4.

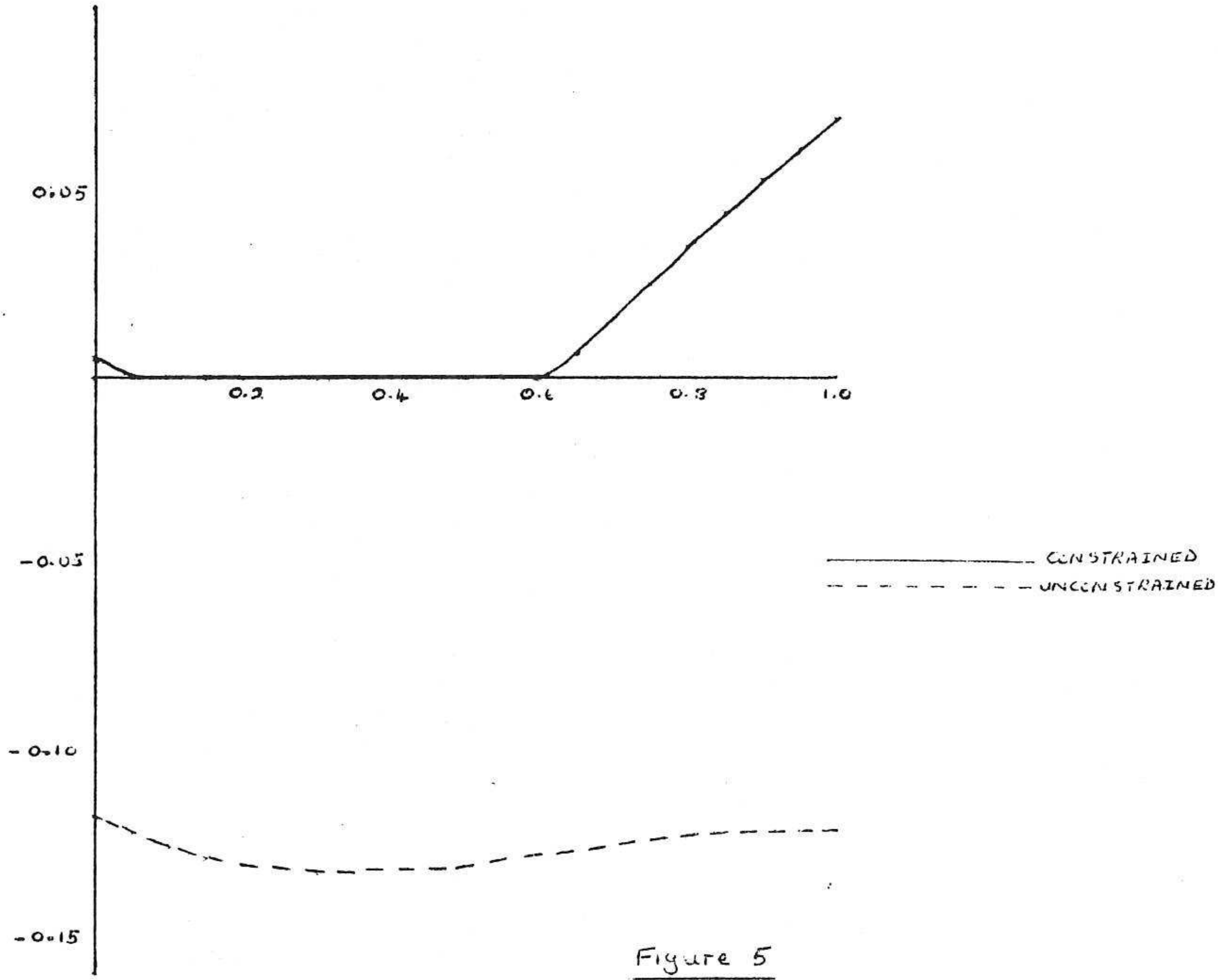

Figure 5

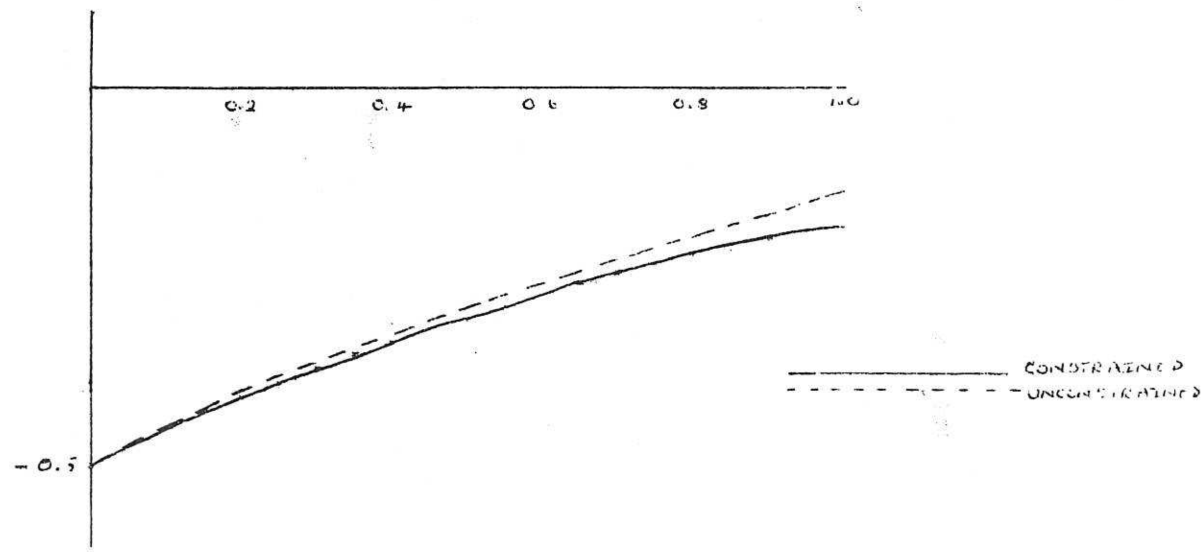

Figure 6

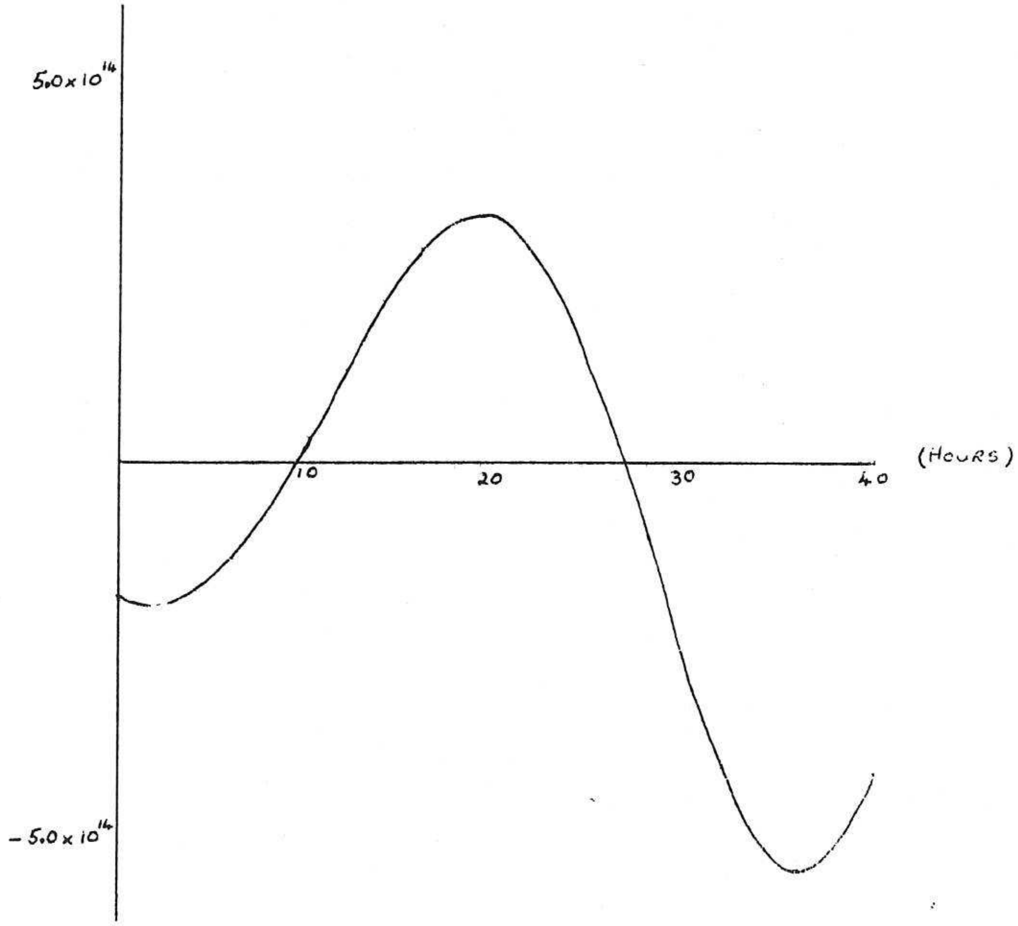

Figure 7

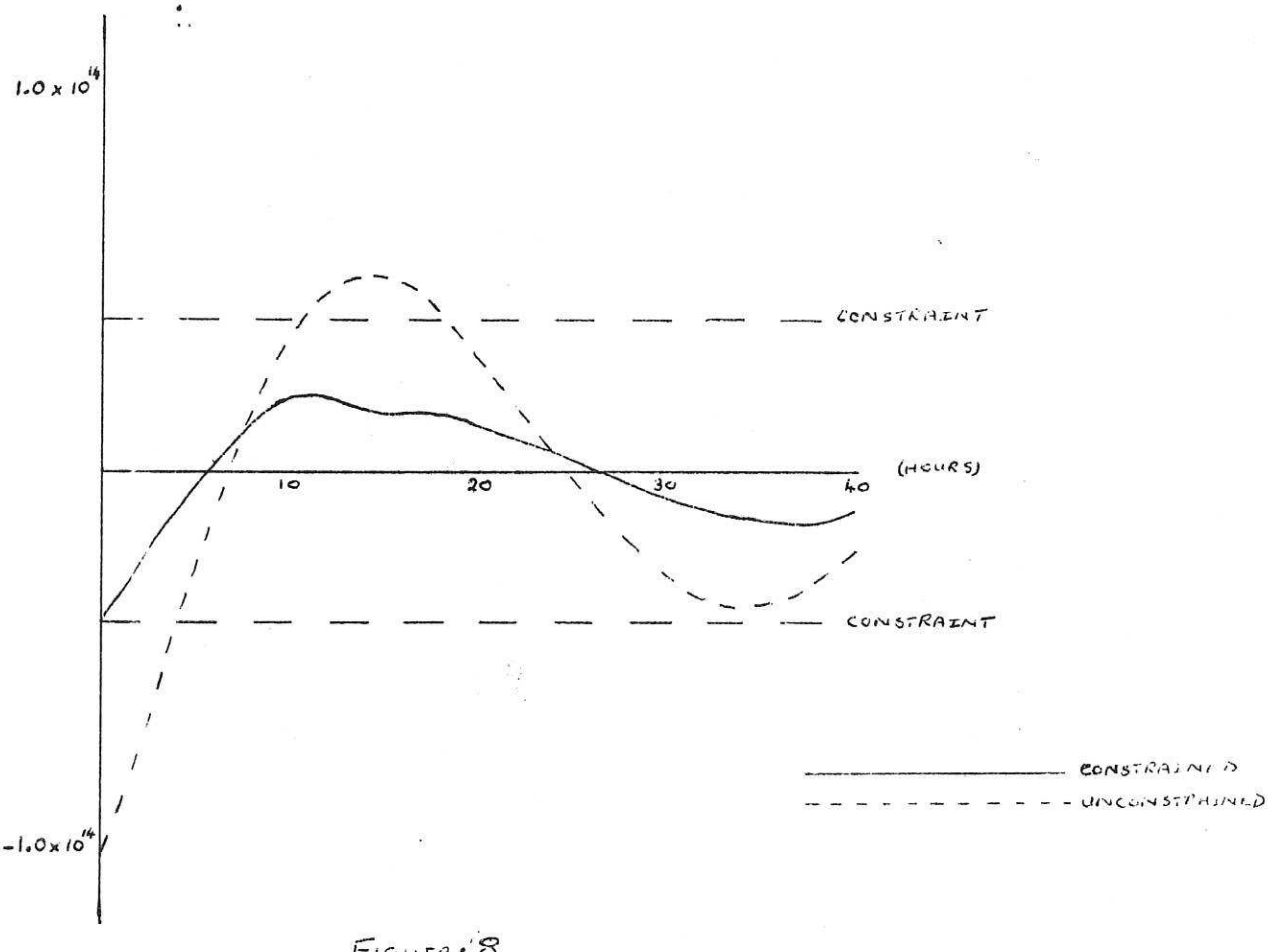

Figure 8

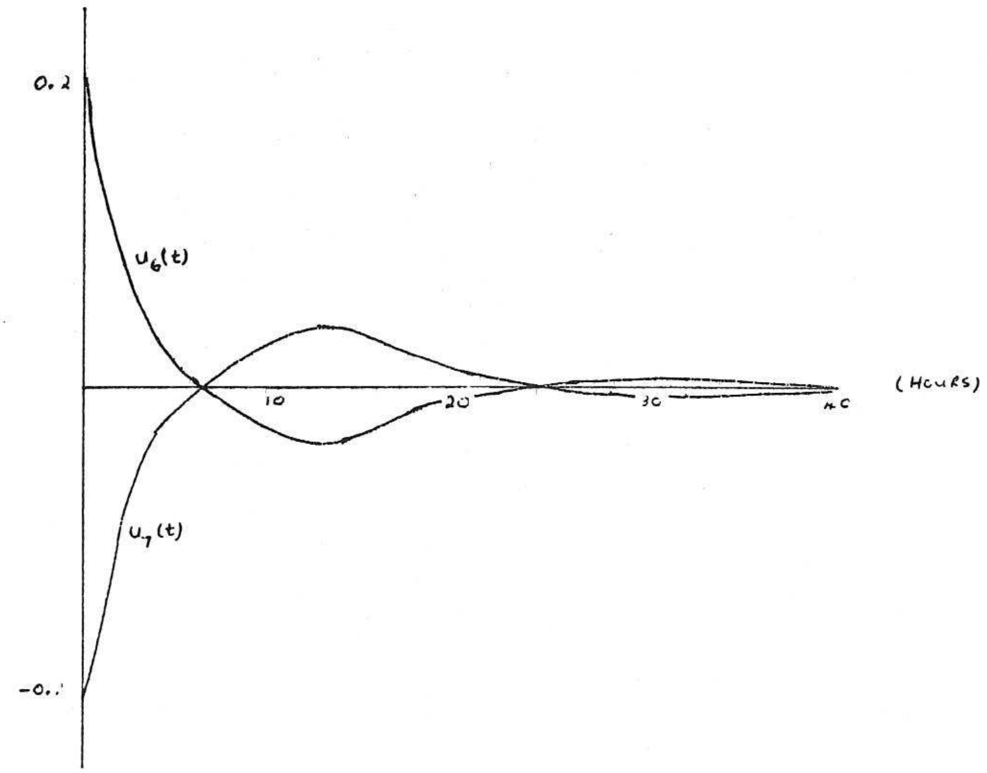

Figure 9