On the Removal of Ill Conditioning Effects in the Computation of Optimal Controls*

by

E. Polak

Department of Electrical Engineering and Computer Sciences
and the Electronics Research Laboratory
University of California
Berkeley, California

N68-34520
(ACCESSION NUMBER)
22
(PAGES)
CR-96765
(NASA CR OR TMX OR AD NUMBER)
(THRU)
1
(CODE)
10
(CATEGORY)
FACILITY FORM 602

*The research reported herein was supported by the National Aeronautics and Space Administration under Grant NsG-354, Suppl. 4.

INTRODUCTION

It has become clearer and clearer in recent years, as can be seen from such work as [1,2], that problems in the calculus of variations, optimal control and nonlinear programming can be treated in a completely unified manner, at least as far as optimality conditions are concerned. It has also become clear that discrete optimal control problems can be recast as standard nonlinear programming forms which are solvable by standard algorithms.

The purpose of this paper is to show that the seeming ease with which nonlinear programming algorithms can be applied to discrete optimal control problems is deceptive, and that severe ill conditioning may occur due to the exponential nature of solutions of a difference equation. However, this paper also shows that it is, nevertheless, possible to make effective use of nonlinear programming results, such as the convergence theory discussed in [3], for the construction of very efficient, large step, optimal control algorithms. Incidentally, it should be pointed out that the algorithm presented in this paper is not the only one which combines both optimal control and nonlinear programming ideas. For a comparison, see the algorithm by Barr and Gilbert [4], which also uses a geometric transcription but has entirely different rules for selecting the next point.

In conclusion, the author wishes to express the hope that the tentative steps presented in this paper for merging optimal control and

nonlinear programming ideas will contribute to a new generation of very efficient optimal control algorithms.

I. A CLASSICAL APPROACH

Statement of the Minimum Engergy Problem: We are given a dynamical system described by the difference equation

(1) $$x_{i+1} = Ax_i + bu_{i+1}, \quad i = 0, 1, \ldots, N-1$$

where $x_i \in R^n$ is the state of the system at time i, $i = 0, 1, 2, \ldots, N$, $u_{i+1} \in R^1$ is the system input at time i, $i = 0, 1, \ldots, N-1$, and A is a $n \times n$ matrix, and $b \in R^n$.

We are required to find a control sequence $\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_N)$ which minimizes the cost

(2) $$\frac{1}{2} \sum_{i=1}^{N} u_i^2,$$

subject to the constraints that $|u_i| \leq 1$, for $i = 1, 2, \ldots, N$, and that the corresponding trajectory $\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_N$, determined by (1) must satisfy $\hat{x}_0 = c_0$ (a given vector in R^n) and $\hat{x}_N \in \Omega$, (a given set in R^n).

Case 1: The set Ω consists of the point $c_N \in R^n$ only, i.e. $\Omega = \{c_N\}$.

This case of the minimum energy problem can be solved by standard quadratic programming algorithms, such as, say, that of Wolfe [5]. To transcribe the minimum energy problem into a standard quadratic programming problem, we proceed as follows.

Solving (1) for x_N, we obtain (with $x_0 = c_0$)

(3) $$x_N = A^N c_0 + \sum_{i=0}^{N-1} A^{N-i-1} b u_{i+1} .$$

Let $r_i = A^{N-i} b$ for $i = 1, 2, \ldots, N$ and let $d = A^N c_0$, then (3) becomes

(4) $$x_N = d + \sum_{i=1}^{N} r_i u_i .$$

Letting R be a $n \times N$ matrix whose i-th column is r_i and setting $c = c_N - d$, the minimum energy problem becomes

(5) $$\text{minimize} \quad \frac{1}{2} \sum_{i=1}^{N} u_i^2 \text{ subject to } Ru = c, \; -1 \le u_i \le +1 \text{ for } i = 1, \ldots, N,$$

which is a standard bounded variable, quadratic programming problem with a unique solution. To apply the Wolfe method [5], or any other method which utilizes the Simplex algorithm [5], we perform one more transformation. Thus, for $i = 1, 2, \ldots, N$, let $2w_i = u_i + 1$, $2v_i = -u_i + 1$. Then (5) becomes,

(6) $$\text{minimize } \frac{1}{2} \sum_{i=1}^{N} (w_i - v_i)^2 \text{ subject to } R(w-v) = c, \; w_i + v_i = 1 \text{ and}$$

$$w_i \ge 0, \; v_i \ge 0, \text{ for } i = 1, 2, \ldots, N.$$

Applying the Kuhn-Tucker necessary and sufficient conditions [6] to (6), we find that $\hat{u}_i = \hat{w}_i - \hat{v}_i$ for $i = 1, 2, \ldots, N$, is optimal if and only if for some vectors ψ^+, ψ^- in R^n and vectors ξ_1, ξ_2 in R^N,

$$
(7) \qquad \left(\begin{array}{c|c|c|c|c}
I & -I & R^T & I & 0 \\ \hline
-I & I & -R^T & 0 & I \\ \hline
R & -R & 0 & 0 & 0 \\ \hline
I & I & 0 & 0 & 0
\end{array}\right)
\begin{pmatrix} \hat{w} \\ \hat{v} \\ \psi^+ \\ \psi^- \\ \xi_1 \\ \xi_2 \end{pmatrix}
=
\begin{pmatrix} 0 \\ 0 \\ 0 \\ c \\ \underline{1} \end{pmatrix}
$$

with $\xi_1 \geq 0$, $\xi_2 \geq 0$ and $\langle \xi_1, w \rangle = \langle \xi_2, v \rangle = 0$, and $\underline{1} = (1, 1, \ldots, 1)$. Wolfe's (or, for that matter, many related algorithms, such as Lemke's [7]) algorithm [5] solves (7) by a modification of the Simplex algorithm [5], and which in turn requires the inversion of $(3N+n) \times (3N+n)$ submatrices of the matrix in (7). Since the top row of R^T (and the first columns of R) will be very close to zero when N is large, it is clear that such submatrices will often be very difficult to invert, resulting in severe ill conditioning. The severity of this ill conditioning, of course, depends on N, since for N large, most of the rows of R^T will appear to be zero to a digital computer.

Thus, standard quadratic programming algorithms become ill-conditioned when used for solving certain optimal control problems.

Case 2: The set Ω is a unit ball with center at the origin, i.e. $\Omega = \{x \mid ||x|| \leq 1\}$. Because Ω is a ball, we can solve this case by modifying a gradient method due to J. Plant [8] which he used for continuous time problems. Thus, applying necessary conditions of optimality (which in this case are also sufficient) (see [1]) directly to

the minimum energy problem, we find that the control sequence $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_N$ (with corresponding trajectory $c_0, \hat{x}_1, \ldots, \hat{x}_N$) is optimal if and only if there exist co-state vectors $p_0, p_1, p_2, \ldots, p_N$ in R^n such that

(8) $$p_i = A^T p_{i+1} \text{ for } i = 0, 1, \ldots, N$$

(9) $$p_N = \left\{ \begin{array}{ll} \beta \hat{x}_N & \text{if } ||\hat{x}_N|| = 1 \text{ for some } \beta < 0 \\ 0 & \text{if } ||\hat{x}_N|| < 1 \end{array} \right\}$$

and, for $i = 1, 2, \ldots, N$,

(10) $$\left\{ \begin{array}{l} -\hat{u}_i + \langle p_i, b \rangle = 0 \text{ if } |\hat{u}_i| < 1 \\ -\hat{u}_i + \langle p_i, b \rangle \geq 0 \text{ if } \hat{u}_i = 1 \\ -\hat{u}_i + \langle p_i, b \rangle \leq 0 \text{ if } \hat{u}_i = -1 \end{array} \right.$$

For A nonsingular (which will be the case if (1) is a sampled-data continuous system), we have that $p_i = (A^{N-i-1})^T p_N$, and from (10) $\hat{u}_i = \text{sat}(\langle p_i, b \rangle)$ (where $\text{sat}(x) = x$ for $|x| \leq 1$, and $\text{sat}(x) = \text{sgn } x$ for $|x| > 1$). Setting $\hat{x}_N \triangleq v$ and making use of (9), we therefore obtain

(11) $$\hat{u}_i = \text{sat} \langle \beta v, r_i \rangle \text{ for } i = 1, 2, \ldots, N,$$

where $r_i = A^{N-i} b$, as before.

(12) <u>Note:</u> It is rather easy to show that $||\hat{x}_N|| = 1$ (i.e. that it is on the boundary of the ball), that $\hat{x}_N$ must satisfy $\langle \hat{x}_N, A^N c_0 \rangle \geq 0$, and that the optimal control sequence $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_N$, and, consequently, also the

corresponding optimal trajectory, $c_0, \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N$, are unique.

To apply the gradient method, we first express the initial state x_0 in terms of a terminal state $v \in S$, where $S = \{v \in E^n \mid ||v|| = 1\}$, by means of (11) and (3), (4). Thus,

$$x_0 = A^{-N} v - \sum_{i=1}^{N} A^{-N} r_i \text{ sat } \langle \beta v, r_i \rangle \triangleq f(\beta, v) \tag{13}$$

If $\beta < 0$ and $v \in S$ are chosen properly, then $x_0 = c_0$, and (11) gives the desired solution. Now, the funtion $f(\beta, v)$ can be shown to be one-to-one for (β, v) in the set $\left\{(\beta, v) \mid \langle v, A^{-N} c_0 \rangle \geq 0,\ ||v|| = 1 \text{ and } |\langle \beta v, r_i \rangle| \leq 1, \text{ for at least one } i \in \{1, 2, \ldots, N\}\right\}$. Hence it can be inverted to find the $\hat{\beta}, \hat{v}$ such that $c_0 = f(\hat{\beta}, \hat{v})$ and hence, by (11) the optimal control sequence.

Now, since v is a point on a unit sphere, $v = h(\theta)$, where θ is its (n-1) spherical co-ordinates. Let $I(\beta, v) \subset \{1, 2, \ldots, N\}$ be an index set such that $i \in I(\beta, v)$ if and only if $\{|\langle \beta v, r_i \rangle| \leq 1\}$. Suppose now we have guessed a $\beta < 0$ and a v satisfying $||v|| = 1$, $\langle v, A^{-N} c_0 \rangle \geq 0$, and $I(\beta, v)$ is not empty. Let $x_0 = f(\beta, v)$. Then

$$x_0 = f(\beta, v) = A^{-N} v - \sum_{i \in \bar{I}(\beta, v)} A^{-N} r_i \rangle . \operatorname{sgn} \langle \beta v, r_i \rangle - \sum_{i \in I(\beta, v)} \beta A^{-N} r_i \rangle \langle r_i, v \rangle , \tag{14}$$

where $\bar{I}$ is the compliment of I, and we have started using the Dirac bracket notation: $\rangle$ indicating a column vector and $\langle$ a row vector. Now,

for small perturbations in β and v, $\bar{I}(\beta, v)$ will not change, and hence, to first order terms, (14) expands as follows:

(15)
$$f(\beta + \Delta\beta,\ v + \Delta v) - f(\beta, v) = (x_0 + \Delta x_0) - x_0$$

$$= A^{-N}\Delta v - \Delta\beta \sum_{i \in I(\beta, v)} A^{-N} r_i \rangle \langle r_i, v \rangle$$

$$- \beta \sum_{i \in I(\beta, v)} A^{-N} r_i \rangle \langle r_i, \Delta v \rangle .$$

Now, since $v = h(\theta)$ and h is obviously differentiable, $\Delta v = \frac{\partial h(\theta)}{\partial \theta} \Delta\theta$, to first order terms. Hence

(16)
$$\Delta x_0 = \left[\left(A^{-N} - \beta \sum_{i \in I(\beta, v)} A^{-N} r_i \rangle \langle r_i \right) \frac{\partial h(\theta)}{\partial \theta} \;\middle|\; - \sum_{i \in I(\beta, v)} A^{-N} r_i \rangle \langle r_i, v \rangle \right] \begin{bmatrix} \Delta\theta \\ \Delta\beta \end{bmatrix}$$

Since, $f(\beta, v)$ is 1 - 1, the matrix in (16) will be nonsingular almost always. Hence, if we choose $\Delta x_0 = \lambda (c_0 - x_0)$ with $\lambda > 0$ sufficiently small for (16) to remain valid, we can compute $(\Delta\theta, \Delta\beta)$ by inverting the matrix in (16) and move from x_0 to $x_0 + \Delta x_0$ (near enough) which is closer to c_0 than x_0 was.

Again, by inspection of (16) and the way $I(\beta, v)$ was defined, we see that matrix inversion will incapacitate this method when N is large, since the matrix in (16) will have an extremely small determinant. Incidentally, apart from this bad ill-conditioning the above approach also suffers from the fact that it cannot be demonstrated to converge.

Thus, once again we find that the straightforward, textbook type approach to a simple problem like our minimum energy problem, may not get one very far on the way to finding a solution.

II. A PARAMETRIC APPROACH

We shall now describe a new algorithm which suffers from none of the defects we have encountered in the previous section. The price one pays for this in Case 1 is that (5) can no longer be solved in a finite number of steps (there seems to be no finite algorithm for Case 2, so we only have gains here). The trick, of course, is to avoid matrix inversions by means of parametrization.

Assumption: We shall assume that $\Omega = \{x \mid ||x-c|| \leq \rho\}$, i.e. that it is a closed ball with center c and radius ρ. We do not exclude the possibility that $\rho = 0$, i.e., that Ω is just a point.

We must first recast the minimum energy problem in geometric terms. Thus, the constraint $|u_i| \leq 1$, for $i = 1, 2, \ldots, N$, defines a hypercube K in R^N (i.e. $K = \{u \mid |u_i| \leq 1 \text{ for } i = 1, 2, \ldots, N\}$). Now, for $\alpha \in [0, \sqrt{N}]$ ($\sqrt{N}$ is the distance from a vertex of K to the origin of R^N), let $\Sigma(\alpha) \subset R^N$ be a closed ball of radius α with center at the origin, i.e. $\Sigma(\alpha) = \{u \mid ||u|| \leq \alpha\}$. Then it is easily seen that our problem is equivalent to finding the smallest α in $[0, \sqrt{N}]$, say $\hat{\alpha}$, for which

$$r(\Sigma(\alpha)) \cap r(K) \cap \Omega \neq \phi , \tag{17}$$

where ϕ is the empty set and $r: R^N \to R^n$ is defined by (4), i.e.

(18) $$r(u) = d + \sum_{i=1}^{N} r_i u_i ,$$

and the corresponding $\hat{u}$ which has the property that

(19) $$r(\Sigma(\hat{\alpha})) \cap r(K) \cap \Omega = \{r(\hat{u})\}.$$

Since the optimal control sequence $\hat{u}$ is unique when Ω is either a point or a closed unit ball, (19) must be true, i.e. the intersection $r(\Sigma(\hat{\alpha})) \cap r(K) \cap \Omega$ must consist of the unique terminal state $\hat{x}_N = r(\hat{u})$. Since the sets Ω and $r(\Sigma(\hat{\alpha})) \cap r(K)$ are both convex, there is a plane $\hat{P}$ passing through the point $\hat{x}_N = r(\hat{u})$ which separates Ω from the set $r(\Sigma(\alpha)) \cap r(K)$. If Ω has points which are in the interior of $r(K)$ then the point $r(\hat{u})$ is also the only point of tangency for $\hat{P}$ and $r(\Sigma(\alpha)) \cap r(K)$. Otherwise the plane $\hat{P}$ will have many points in common with $r(\Sigma(\alpha)) \cap r(K)$.

The gist of the algorithm we are about to describe is as follows: (i) Insert a hyperplane P between the point $r(\Sigma(0)) \cap r(K) = \{r(0)\} = \{d\}$ and the set Ω, with P being a tangent plane to Ω at a point v; (ii) Increase α until $r(\Sigma(\alpha)) \cap r(K)$ touches P at a point w_i; (iii) If v = w then we are done, since we must have found the smallest α satisfying (19). If $v \neq w$, then we can rotate the hyperplane P in such a way that it stays in contact with Ω but breaks away from $r(\Sigma(\alpha)) \cap r(K)$. We can then increase α and check again if v = w.

Thus, we should stop if either v = w, in which case if v is the optimal terminal state $\hat{x}_N$, or else, if we have established that $v = \hat{x}_N$ by independent means. It will readily be seen that our algorithm

handles the case $\Omega \cap r(K) = \{\hat{x}_N\}$ automatically so that we only need to check the optimality of the initial guess by independent means and then proceed as if $v = w$ is the case for the optimal solution.

Obviously, the rotation of the plane P cannot be done in any old way if one wishes to insure convergence. Hence we shall need the following machinery to make it work.

(20) Definition: Let $P(v, s)$ denote the hyperplane in R^n which passes through the point $v \in R^n$, with unit normal s, i.e.

$$P(v, s) = \{ x \mid \langle x - v, s \rangle = 0 \}$$

(21) Definition: Let $S = \{s \mid ||S|| = 1\}$ be a unit sphere in R^n and let $v: S \to \partial\Omega$ (the boundary of Ω) be the contact function defined by the relation

$$\langle x - v(s), s \rangle \leq 0 \text{ for all } x \in \Omega \text{, i.e.,}$$

since $\Omega = \{x \mid ||x - c|| \leq \rho\}$ is a closed ball of radius ρ and center c, $v(s) = c + \rho s$. Thus, if $\rho = 0$, $v(s) = c$ for all s.

(22) Definition: Let $V \subset \partial\Omega$ be the set of all points in $\partial\Omega$ which can be separated by a hyperplane from the point $d = r(0)$, i.e., for each $v^* \in V$ there is a $s \in S$, such that $v(s) = v^*$ and

$$\langle d - v(s), s \rangle \geq 0.$$

(Note, $V = \{c\}$ when $\Omega = \{c\}$).

Furthermore, let T be the set of all points s in S such that $v(s) \in V$. (Note, T is a closed set).

(23) Definition: For $\alpha \in [0, \sqrt{N}]$, let $\mathcal{C}(\alpha)$ denote the set

$$\mathcal{C}(\alpha) = r(\Sigma(\alpha)) \cap r(K),$$

and let $c: T \to [0, \sqrt{N}]$ be a surrogate cost function defined by

$$c(s) = \min_{\alpha \in [0, \sqrt{N}]} \{\alpha \mid P(v(s), s) \cap \mathcal{C}(\alpha) \neq \phi\},$$

i.e. $c(s)$ is the smallest α for which the intersection of $\mathcal{C}(\alpha)$ with $P(v(s), s)$, the tangent hyperplane to Ω at $v(s)$ with normal s, is not empty. Observing now, that $P(v(s), s) \cap \mathcal{C}(c(s))$ is the set of terminal states for the minimum energy problem with Ω set equal to $P(v(s), s)$, and that the solution to this new problem is also unique, we conclude that

(24) $$P(v(s), s) \cap \mathcal{C}(c(s)) = \{w(s)\},$$

i.e. that it must consist of one point only.

(25) Definition: We define the map $w: T \to R^n$ by (24).

Now, it is not difficult to see that small changes in s produce correspondingly small changes in $v(s)$, $w(s)$ and $c(s)$, i.e. that all these functions are continuous.

(26) Definition: For any $s \in T$, let $\sigma(s)$ be the arc in T defined by

$$\sigma(s) = \left\{ s' \in T \mid s' = \frac{s + \lambda(w(s) - v(s))}{\|s + \lambda(w(s) - u(s))\|}, \ 0 \leq \lambda \leq 1 \right\}$$

Again we see that small variations in s cause only small variations in $\sigma(s)$, (i.e. that it is a continuous map from T into the set of all subsets

of T with respect to the Hausdorff metric).

We now have all the parts we need to define our algorithm.

(27) Definition: Let $a: T \to T$ be the algorithm defined by

$$c(a(s)) = \max_{s' \in a(s)} c(s'),$$

i.e. it is the point s' on $\sigma(s)$ which maximizes $c(s')$. Simple geometric considerations lead one to believe that the algorithm $a(\cdot)$ is continuous, except, perhaps, at the optimal point $\hat{s}$ (i.e. $r(\hat{u}) = v(\hat{s})$ and $P(v(\hat{s}), \hat{s})$ separates Ω from $r(\Sigma(\hat{\alpha})) \cap r(K)$. That this is indeed so is proven in [9].

(28) Theorem: Let $s_0, s_1, s_2, \ldots,$ be a sequence in T generated according to the law

$$s_{i+1} = a(s_i), \quad i = 0, 1, 2, \ldots$$

Then s_i converges to a point $s^* \in T$ such that $v(s^*) = w(s^*)$.

Proof: (We shall only prove this theorem for the case when $a(\cdot)$ is continuous on its entire domain of definition). First, we observe that for every $s \in T$ such that $v(s) \neq w(s)$, we must have $c(a(s)) > c(s)$. Next, we observe that for any s, s' in T, with $s' \neq s$, $w(s') \neq w(s)$, i.e. $w(\cdot)$ is one-to-one on T.

Since $c(s_{i+1}) > c(s_i)$, $i = 0, 1, 2, \ldots,$ is a monotonic increasing sequence which is bounded from above, we must have,

(29) $$c(s_i) \to c^* < \infty \quad i \in \{0, 1, 2, \ldots\}.$$

Since T is closed and bounded, the sequence $\{s_i\}$ must contain a convergent subsequence, $\{s_i\}$, $i \in K$, an index set contained in $\{0, 1, 2, \ldots\}$, with limit point $s^* \in T$, say. Hence, since $c(\cdot)$ and $a(\cdot)$ are continuous

(30) $$c(a(s_i) \to c(a(s^*)) = c(s^*) = c^* \text{ for } i \in K$$

But this implies that

(31) $$v(s^*) = w(s^*)$$

Now, since $w(\cdot)$ is one-to-one on T, and since the intersection $\mathcal{C}(c(s^*)) \cap \Omega$ consists of exactly one point w^* (by uniqueness of solution), it follows that the limit point s^* is the only point to which a subsequence of $\{s_i\}$ can converge and hence $\{s_i\}$ itself converges to this point.

(32) <u>Corollary:</u> Let $\lambda \in (0, 1)$ and let $a_\lambda : T \to T$ be defined by

(33) $$a_\lambda(s) = \frac{s + \lambda(a(s) - s)}{||s + \lambda(a(s) - s)||} .$$

where $a(\cdot)$ is defined by (27). Then any sequence $\{s_i\}$ in T generated according to the law

(33) $$s_{i+1} = a_\lambda(s_i)$$

converges to apoint $s^* \in T$ such that $v(s^*) = w(s^*)$.

Proof: We simply note that $a_\lambda(\cdot)$ is continuous and that $c(a_\lambda(s)) > c(s)$ for all $s \in T$ such that $v(s) \neq w(s)$. Hence the proof is exactly as for theorem (28).

(34) Remark: We conclude from the above corollary (since $\lambda > 0$ may be taken to be quite small without affecting convergence) that even a very approximate evaluation of $a(s)$ should be compatible with convergence.

We shall now give the algorithm for carrying out the computation of the optimal control sequence $\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_N)$ for the minimum energy problem.

III. THE ALGORITHM

P1. Initial Guess Procedure:

Set $s_0 = \dfrac{d-c}{||d-c||}$, where c is the center of $\Omega = \{x \mid ||x-c|| \leq \rho\}$ and $d = A^{-N} c_0$, as before. Clearly, s_0 is in T.

P2. Computation of $v(s_0), w(s_0), c(s_0), \sigma(s_0)$

Step 1: By (21),

(35) $$v(s_0) = c + \rho s_0$$

Step 2: Note that $w(s_0)$ is the point on $P(v(s_0), s_0)$, which is the terminal state for the minimum energy problem when Ω is set equal to $P(v(s_0), s_0)$. Hence, from necessary and sufficient conditions, we obtain (as in (10), (11)) that

(36) $$w(s_0) = d + \sum_{i=1}^{N} \text{sat} < \beta_0 s_0, r_i > r_i$$

for some $\beta_0 < 0$. To compute β, substitute (36) into the expression for $P(v(s_0), s_0)$, and set β_0 to satisfy (see (20)),

$$\langle d - c - \rho s_0 + \sum_{i=1}^{N} \text{sat} \langle \beta s_0, r_i \rangle \; r_i, s_0 \rangle = 0 \tag{37}$$

This computation is quite easy since (37) is a piecewise linear expression.

Step 3: By inspection of (36) and (23),

$$c(s_0) = \left(\sum_{i=1}^{N} (\text{sat} \langle \beta_0 s_0, r_i \rangle)^2 \right)^{1/2} \tag{38}$$

Step 4: By definition of $\sigma(s_0)$ (26),

$$\sigma(s_0) = \{ s' \in T \mid s' = \frac{s + \lambda(w(s_0) - c - \rho s_0)}{\| s + \lambda(w(s_0) - c - \rho s_0) \|} \; , \; 0 \le \lambda \le 1 \} \tag{39}$$

P3. Computation of s_1

Step 1: Compute $M + 1$ points $y_1, y_2, \ldots, y_M$, in $\sigma(s_0)$, by setting $\lambda = 0, \frac{1}{M}, \frac{1}{2M}, \ldots$ in (39), (assuming of course, that for $\lambda = 1$, the point is in $\sigma(s_0)$).

Step 2: Compute $c(y_i)$ for $i = 1, 2, \ldots, M$ by setting $s_0 = y_i$ in P2. Find a $j \in \{1, 2, \ldots, M\}$ such that

$$c(y_j) \ge c(y_i) \text{ for all } i \in \{1, 2, \ldots, M\}. \tag{40}$$

Step 3: Set

$$s_1 = y_j \tag{41}$$

Note that it is usually best to start with $M = 2$ and to increase M only if (40) cannot be satisfied.

P4. Verification of Feasibility.

Suppose that for some $s_0 \in T$, we find that the set $\mathcal{C}(\sqrt{N}) = r(K)$ does not intersect the plane $P(v(s_0), s_0)$, then it is clear that there is no admissible control sequence which will take the system from $x_0 = c_0$ to Ω in N steps, i.e. the problem has no solution.

Step 4: Compute a $\beta^* < 0$ such that

(42) $$|\langle \beta^* s_0, r_i \rangle| \geq 1 \text{ for } i = 1, 2, \ldots, N$$

If

(43) $$\langle d-c-\rho s_0 + \sum_{i=1}^{N} \operatorname{sgn} \langle \beta^* s_0, r_i \rangle r_i, s_0 \rangle > 0$$

then there is no solution to the minimum energy problem.

P5. Verification of Optimality of s_0

Step 1: Minimize with respect to β the function

(44) $$g(\beta, s_0) \triangleq ||d-c-\rho s_0 + \sum_{i=1}^{N} \operatorname{sat} \langle \beta s_0, r_i \rangle r_i||^2$$

If $\min_{\beta} g(\beta, s_0) = 0$ then s_0 is optimal.

P6. Computation of the Optimal Control Sequence.

Suppose s_0 satisfies either $v(s_0) = w(s_0)$ or $\min_{\beta} g(\beta, s_0) = 0$. Then s_0 is optimal and

(45) $$\hat{u}_i = \operatorname{sat} \langle \beta_0 s_0, r_i \rangle \text{ for } i = 1, 2, \ldots, N$$

where β_0 is determined either by (36) or by $g(\beta_0, s_0) = 0$.

The manner in which these six procedures are combined into an algorithm is best illustrated by the following flow chart. Note that when $d \in \Omega$, $\hat{u} = (0, 0, \ldots, 0)$ and hence the problem becomes trivial. Also note that the use of P5 can be omitted since the chances of guessing right the very first time are very slim indeed. However, if s_0 were the optimal solution, the algorithm would not recognize it if P5 were omitted. It would yield an $s_1 \neq s_0$ and would then proceed to construct a sequence $s_2, s_3, \ldots$ which would converge to s_0. The use of a truncation error $\epsilon > 0$ is introduced to stop computations after a finite number of steps when $||v(s_0) - w(s_0)|| \leq \epsilon$.

IV. CONCLUSION

The algorithm presented in section III is the least obvious and least direct one of the three methods presented. However, its computational behaviour is considerably better than that of the other two. As a reference point of comparison, the author would like to indicate that on a number of specific runs on a problem with N = 50 and n = 10 (a tenth order system with eigenvalues of A ranging from 0.9 to 0.5), the parametric method took three iterations and 3 - 5 sec. of IBM 7094 time to obtain a solution, while the gradient method of section I required 40 - 60 sec. to compute.

The reason for the good behavior of the parametric method is obvious: it requires no matrix inversions, and the substitute step, the

solution of (37), which it introduces is easy to carry out.

With nonlinear programming becoming more and more relevant to control problems, it is hoped that the present work will facilitate the task of modifying and adapting standard algorithms to the specific structure of optimal control problems.

REFERENCES

1. M. Canon, C. Cullum, and E. Polak, "Constrained Minimization Problems in Finite Dimensional Spaces, J. SIAM Control, Vol. 4, pp. 528-547, 1966.

2. H. Halkin and L. W. Neustadt, "General Necessary Conditions for Optimization Problems," USCEE Report 173, 1966.

3. W. I. Zangwill, "Applications of the Convergence Conditions," Working Paper No. 231, University of California, August, 1967.

4. R. O. Barr and E. G. Gilbert, "Some Iterative Procedures for Computing Optimal Controls," Third Congress of the International Federation of Automatic Control, London 20-25, Paper No. 24.D, June, 1966.

5. P. Wolfe, "The Simplex Method for Quadratic Programming," Econometrica, Vol. 27, pp. 382-398, 1959.

6. H. W. Kuhn and A. W. Tucker, "Nonlinear Programming, Proc. of the Second Berkeley Symposium on Mathematic Statistics and Probability," University of California Press, Berkeley, California, pp. 481-492, 1951.

7. C. E. Lemke, "A Method for Solution of Quadratic Programs," Management Science, Vol. 8, pp. 442-453, 1962.

8. J. B. Plant, "An Iterative Procedure for the Computation of Optimal Controls," Ph.D. Dissertation, Department of Electrical Engineering, MIT, June, 1965.

9. E. Polak and M. Deparis, "An Algorithm for Minimum Energy Control," University of California, Electronics Research Laboratory, Berkeley, California, ERL Memorandum M225, November 1, 1967.

FLOW CHART

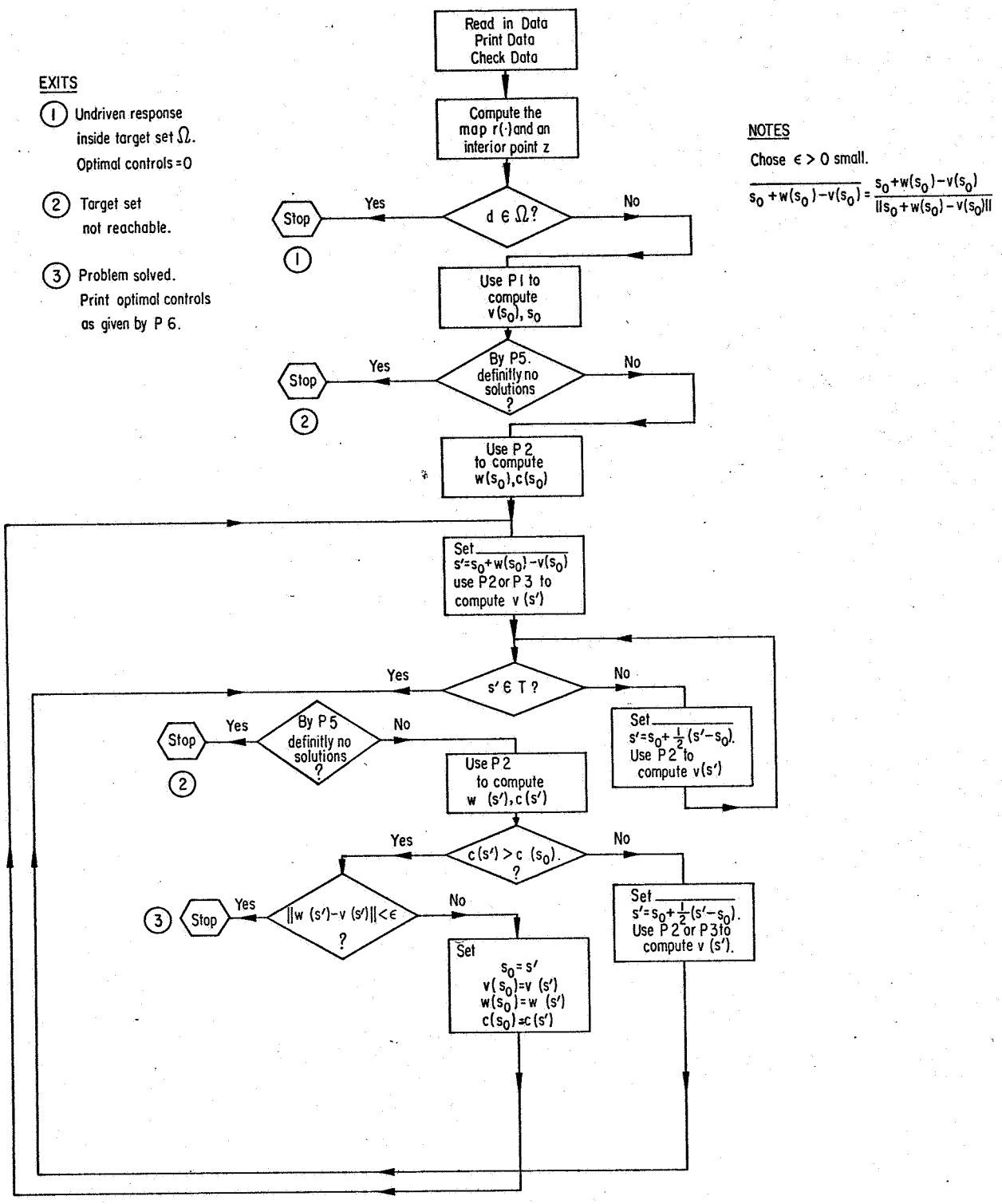

Fig. 1

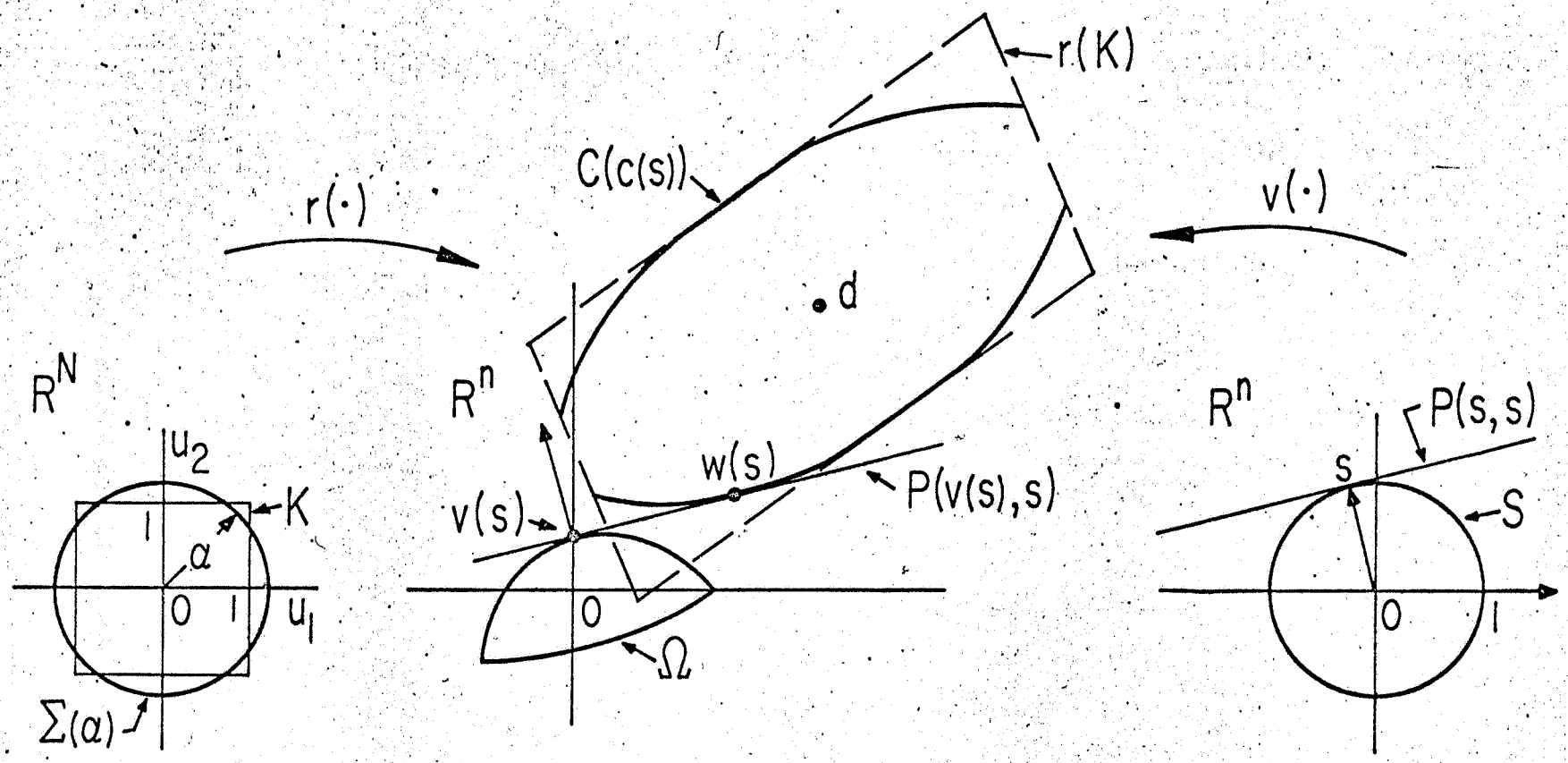

Fig. 2. The geometry of the problem.