J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

DYNAMIC PROGRAMMING, SYSTEM IDENTIFICATION, AND SUBOPTIMIZATION*

RICHARD BELLMAN†

1. Introduction. The problem we start with appears to be quite specialized. Given a function $u(t)$ defined over the interval [0, a], we wish to find a polygonal approximation which is a best fit in a mean-square sense. (See Fig. 1.) The analytic problem for N is that of minimizing the function

$$R_N = \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} (u(t) - a_i - b_i t)^2 \, dt, \tag{1.1}$$

over the quantities a_i, b_i, and t_i. Here $t_0 = 0$, $t_N = a$.

This can be treated in a number of direct fashions, using search and gradient techniques. We wish, however, to employ dynamic programming, which appears to be superior even in this case, and then gradually to enlarge the scope of the problem until it covers a question in the identification of systems and a version of the general problem of considering suboptimal policies in control processes. Results related to what follows have been presented in [1], [2], [3].

2. Adaptive curve fitting. The foregoing problem can be considered to fall within the new area of sequential computation. In place of choosing the t_i in advance, we allow the structure of the function $u(t)$ to determine their positions. Similar techniques can be applied in connection with the numerical integration of ordinary and partial differential equations. Write

$$\min_{a_i, b_i, t_i} R_N = f_N(a), \tag{2.1}$$

defined for $N = 0, 1, 2, \cdots$, and $a \geqq 0$. Introduce the function of two variables,

$$\Delta(s_1, s_2) = \min_{a,b} \int_{s_1}^{s_2} (u(t) - a - bt)^2 \, dt, \tag{2.2}$$

for $0 \leqq s_1 \leqq s_2 < \infty$. That this happens in this case to be explicitly calculable is of no particular significance at the moment. In general, this function will be obtained via numerical methods.

* Received by the editors May 19, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† The RAND Corporation, 1700 Main Street, Santa Monica, California. This research was sponsored by the United States Air Force under Project RAND—Contract No. AF 49(638)-700 monitored by the Directorate of Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF.

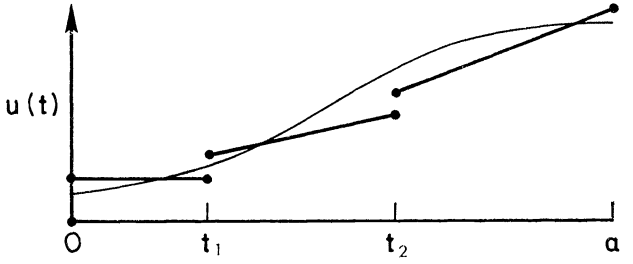

FIG. 1

Then

$$f_0(a) = \Delta(0, a), \tag{2.3}$$

and the principle of optimality yields the recurrence relation

$$f_N(a) = \min_{0 \leqq t_N \leqq a} [\Delta(t_N, a) + f_{N-1}(t_N)], \tag{2.4}$$

for $N \geqq 1$.

This leads to a quite simple and efficient computational algorithm.

3. Discussion. Perhaps the first point to note in connection with what has been given above is that the computational feasibility of the algorithm inherent in (2.4) is not strongly dependent upon the mean-square norm in (2.2). We could just as easily use

$$\Delta(s_1, s_2) = \min_{a,b} \max_{s_1 \leqq t \leqq s_2} |u(t) - a - bt|, \tag{3.1}$$

or allow approximation by polynomials of higher degree. This brings us into contact with the theory of spline approximations, but we shall not pursue that here; see [4] for an extensive set of references.

As soon as we start pursuing the idea of approximating to $u(t)$ over the interval $[s_1, s_2]$ by a function of simple analytic form, we enter the domain of differential approximation [5]. We recognize that a polynomial of degree M satisfies the differential equation

$$\frac{d^{(M+1)}v}{dt^{M+1}} = 0, \tag{3.2}$$

that the exponential polynomial $\sum_{k=1}^{M} a_k e^{\lambda_k t}$ satisfies the differential equation

$$\frac{d^{(M)}v}{dt^{(M)}} + b_1 \frac{d^{(M-1)}v}{dt^{(M-1)}} + \cdots + b_M v = 0, \tag{3.3}$$

and that $\sum_{k=1}^{M} a_k \cos(\lambda_k + \phi_k)$ satisfies a similar equation of degree $2M$.

It follows that a substantial extension of straight-line approximation is the following. Determine the parameters a_i and initial conditions c_i so that

$$\| u - v \| \tag{3.4}$$

is minimized, where u is given and v is determined by the ordinary differential equation

$$\frac{d^{(M)}v}{dt^{(M)}} = \left(t, v, \cdots, \frac{d^{(M-1)}v}{dt^{(M-1)}}, \quad a_1, \cdots, a_M\right), \tag{3.5}$$

$v^{(i)}(0) = c_i$, $i = 0, 1, \cdots, M - 1$. Here, we can use a mean-square norm, or some other convenient norm.

Problems of this nature can be attacked by means of quasilinearization and other techniques [5].

4. Identification of systems. The foregoing remarks and techniques allow us to approach an interesting problem in the identification of systems. Suppose that we know that a function $u(t)$ is generated in the following manner. In the interval $t_i \leqq t \leqq t_{i+1}$, $t_0 \leqq t_1 \leqq \cdots \leqq t_{n+1}$, $t_0 = 0$, $t_{N+1} = a_0$, it satisfies the equation

$$\frac{d^M v}{dt^{M+1}} = g\left(t, v, \cdots, \frac{d^{(M-1)}v}{dt^{(M-1)}}, a_i\right), \tag{4.1}$$

$$v^{(j)}(t_i) = c_{ij}, \qquad j = 0, 1, \cdots, M - 1.$$

Given the values of $u(t)$ in $[0, a]$, we wish to determine the vector parameters a_i, the parameters c_{ij}, and the switching points t_i, and, occasionally, N itself. This is a particular type of pattern recognition problem.

We begin by introducing the function

$$\Delta(s_1, s_2) = \min_{a, c_j} \int_{s_1}^{s_2} (u - v)^2 \, dt, \tag{4.2}$$

where $v(t)$ satisfies (4.1), $0 \leqq s_1 \leqq s_2 \leqq a$. Our assumption is that we can compute this function of two variables. This will, in general, however, be a nontrivial task. If then we introduce the function

$$f_N(a) = \min_{\{a_i, c_{ij}\}} \int_0^a (u - v)^2 \, dt, \tag{4.3}$$

$a \geqq 0$, allowing N switch points, or transition points, we obtain exactly the same recurrence relation as in (2.4). If $u(t)$ is actually determined by (4.1), we will have $f_N(a_0) = 0$ for the correct choice of t_N.

5. Suboptimization. For analytic, economic, and engineering convenience, it is often useful to consider the approximation of optimal control policies by simple, feasible control policies.

Thus, for example, in the minimization of

$$J(u) = \int_0^T g(u, u')\, dt, \quad u(0) = c, \tag{5.1}$$

we may wish to consider as admissible functions only those for which

$$u'(t) = b_i, \quad s_i \leqq t \leqq s_{i+1}, \tag{5.2}$$

with $s_0 = 0$, $s_{N+1} = T$, where the b_i and s_i are to be chosen.

Let us define

$$f_N(T, c) = \min J(u), \tag{5.3}$$

where the minimum is now over the class of suboptimal policies defined above. Then, as before, the principle of optimality yields the relation

$$f_N(T, c) = \min_{b_0, s_1} \left[\int_0^{s_1} g(u(b_0, t), b_0)\, dt + f_{N-1}(T - s_1, u(b_0, s_1)) \right], \tag{5.4}$$

for $N \geqq 1$, with

$$f_0(T, c) = \min_{b_0} \int_0^T g(u(b_0, t), b_0)\, dt.$$

Here $u(b_0, t)$ denotes the function over the relevant t-interval determined by the nature of the suboptimal policy and the initial state c. In this case, $u(b_0, t) = c + b_0 t$.

6. Reduction of dimensionality. One of the purposes of using suboptimal policies is to bypass some of the analytic and computational difficulties of the original optimization problem. This is particularly the case when we have a control process involving either a high-dimensional state vector, or an infinite-dimensional vector.

In this situation, we can often replace the actual state vector at time t by a record of the control policies used, and thus obtain a more manageable computational algorithm. Furthermore, we can use new types of approximation methods. For a detailed discussion of this technique, see [6].

REFERENCES

[1] R. Bellman, *On the approximation of curves by line segments using dynamic programming*, Comm. ACM, 4 (1961), p. 284.

[2] R. Bellman, B. Gluss, and R. Roth, *On the identification of systems and the unscrambling of data: some problems suggested by neurophysiology*, Proc. Nat. Acad. Sci. U.S.A., 52 (1964), pp. 1239–1249.

[3] ———, *Segmental differential approximation and the "black box" problem*, The RAND Corporation, RM-4269-PR, 1964; to appear in J. Math. Anal. Appl.

[4] J. H. Ahlberg, E. N. Nilson, and J. L. Walsh, *Best approximation and con-*

vergence properties of higher-order spline approximations, J. Math. Mech., 14 (1965), pp. 231–244.

[5] R. Bellman and R. Kalaba, *Quasilinearization and Nonlinear Boundary-value Problems*, American Elsevier, New York, 1965.

[6] R. Bellman, *Dynamic programming, generalized states, and switching systems*, The RAND Corporation, RM-4474-PR, 1965; to appear in J. Math. Anal. Appl.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

PROGRAMMING AND CONTROL PROBLEMS ARISING FROM OPTIMAL ROUTING IN TELEPHONE NETWORKS*

V. E. BENEŠ†

1. Introduction. A telephone connecting network invariably provides many paths on which a particular call could be completed. Thus there naturally arise problems of optimal routing, that is, of making choices of routes so as to achieve a maximum of some measure of system performance, such as the loss (probability of blocking).

It is the aim of this work to formulate, study, and (in part) solve a general class of optimal routing problems for telephone networks. The formulation of these problems is undertaken insofar as possible within the classical dynamical theory of telephone traffic initiated by A. K. Erlang, that is, in terms of Markov processes based on the assumptions of (i) negative exponential distributions for mutually independent holding-times, and (ii) randomly originating traffic. To these assumptions is added a description of how attempted calls are accepted and assigned routes.

The problem of choosing "good" routes for information flow in a communications network is vastly complicated by the difficult questions surrounding the collection, updating, and relevance of information (about the state of the system) on the basis of which routing decisions are to be made. Thus, one of the items to be chosen in designing a routing scheme is the information on which the routing is to be based. Indeed there is a whole spectrum of possible choices for this information, from no information at all (except what is unwittingly discovered in making call attempts) to the opposite extreme of full knowledge of the state of the connecting network. Clearly, a practical compromise between total ignorance and a very expensive, complex scheme based on many data must usually be made.

Our considerations in this work will be limited to the case of perfect information, in which the microscopic state of the connecting network is assumed known and available for making routing decisions. This case is of course very far from realistic: few existing or envisaged systems utilize even a small fraction of this possible information for routing. Indeed, much of it is likely to be of very little relevance. Nevertheless, it is important to

* Received by the editors August 5, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

The present text, intended as part of the Proceedings of the Conference, is only a prolonged abstract, formulating the problem and giving sample results. The full version of the paper will be submitted for publication to the Bell System Technical Journal.

† Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey.

know what would be good routing if we could implement it and could afford it, so the full information case to be considered here forms at worst a limiting situation for which some theory is available, and a natural starting point for investigation.

In this discussion of the involved problem of routing calls, one of the difficulties that arises deserves special mention. At first sight, the problem of routing with full information seems to boil down to the question, "Which of the paths available for call c in state x should be used?" This form of the problem overlooks the possibility that perhaps the best thing to do is not to put c in at all when the state is x! In other words, it assumes that, naturally, c will be put up in state x if it is attempted in x and is not blocked. Previous uses of the stochastic model for telephone traffic which we employ have always made this assumption [1], [2].

Conceivably, then, it is better to reject a call c that is not blocked in a state x. Thus the problem of routing should be phrased, "Should a call c, free and not blocked in state x, be completed; and if so, by which route?"

It turns out that answering the first part of the question, as to which calls should be completed in which states, is often the hardest part of the problem. Examples can be given in which it is fairly easy to solve the route selection part of the problem, but for which the question of whether a call should go in or not is not settled, and seems far from being settled. That this question has substantial practical import is apparent from the simulation studies carried out by Weber (3), which clearly show how prohibition of circuitous routes (and thus rejection of certain unblocked calls) can improve system performance.

We conclude this introduction with a brief summary of the entire paper. As is customary in telephone traffic theory, a Markov process is used to describe the operation of the connecting network under study. The Kolmogorov equations for this process then constitute a set of linear differential equations describing the controlled system; in these the control functions expressing the routing method being used appear among the coefficients. It is natural to restrict attention to asymptotic behavior; this leads to a problem of maximizing a bilinear (or linear fractional) form subject to linear constraints; this problem is equivalent to a linear programming problem. An alternative approach first shows that minimizing the probability of loss, and maximizing the fraction of events that are successful call attempts, are equivalent. This fact permits a classical dynamic programming approach. The remainder of the paper attempts to use this approach to establish relations between the combinatorial properties of the network and the policy (or policies) optimal for given criteria of performance. In particular it is shown that for connecting networks having certain "hereditary" properties, optimal policies for minimizing loss correspond

closely to the heuristic advice, "Prefer those states in which as few calls are blocked as possible."

It must be stressed that the problem is far from fully understood, and that much remains to be done.

2. States, events, and rules. The elements of the mathematical model to be used for our study of routing separate naturally into combinatorial ones and probabilistic. The former arise from the structure of the connecting network and from the ways in which calls can be put up in it; the latter represent assumptions about the random traffic the network is to carry. The combinatorial and structural aspects are discussed in this section; terminology and notation for them are introduced. The probabilistic aspects are considered in the next section.

A *connecting network* ν is a quadruple $\nu = (G, I, \Omega, S)$, where G is a graph depicting network *structure*, I is the set of nodes of G which are *inlets*, Ω is the set of nodes of G that are *outlets*, and S is the set of permitted *states*. Variables $x, y, z, \cdots$ at the end of the alphabet denote states, while u and v (respectively) denote a typical inlet and a typical outlet. A state x can be thought of as a set of disjoint chains on G, each chain joining I to Ω. Not every such set of chains represents a state: sets with wastefully circuitous chains may be excluded from S. It is possible that $I = \Omega$, that $I \cap \Omega = \varnothing$ = null set, or that some intermediate condition is obtained, depending on the "community of interest" aspects of the network ν.

The set S of states is *partially ordered* by *inclusion* $\leqq$, where $x \leqq y$ means that state x can be obtained from state y by removing zero or more calls. If x and y satisfy the same *assignment* of inlets to outlets, i.e., are such that all and only those inlets $u \in I$ are connected in x to outlets $v \in \Omega$ which are connected to the same v in y (though possibly by different *routes*), then we say that x and y are equivalent, written $x \sim y$.

The set S of states determines another set E of *events*, either *hangups* (terminations of calls), *successes* (successful call attempts), or *blocks* (blocked call attempts). The occurrence of an event in a state may lead to a new state obtained by adding or removing a call in progress, or it may, if it is a blocked call or one that is rejected, lead to no change of state. Not every event can occur in every state: naturally, only those calls can hang up in a state which are in progress in that state, and only those inlet-outlet pairs can ask for a connection between them in a state that are idle in that state. The notation e is used for a (general) event, h for a hangup, and c for an attempted call. If e can occur in x we write $e \in x$. A call $c \in x$ is *blocked* in a state x if there is no $y \in S$ which covers x in the sense of the partial ordering $\leqq$ and in which c is in progress.

We denote by A_x the set of states that are immediately above x in the partial ordering $\leqq$, and by B_x the set of those that are immediately below.

Thus

$$A_x = \{\text{states accessible from } x \text{ by adding a call}\},$$

$$B_x = \{\text{states accessible from } x \text{ by a hangup}\}.$$

For an event $e \in x$, the set A_{ex} is to consist of those states to which the network might pass upon the occurrence of e in x. Thus, if e is a blocked call, $A_{ex} = \{x\}$; also

$$\bigcup_{h \in x} A_{hx} = B_x,$$

$$\bigcup_{\substack{c \in x \\ c \text{ not blocked in } x}} A_{cx} = A_x.$$

The *number of calls in progress* in state x is denoted by $| x |$. The number of call attempts $c \in x$ which are not blocked in x is denoted by $s(x)$, for "*successes in x.*" The functions $|\cdot|$ and $s(\cdot)$ defined on S play important roles in the stochastic process to be used for studying routing.

It will be assumed throughout this work that attempted calls to busy terminals are rejected and have no effect on the state of the network; similarly, blocked attempts to call an idle terminal are refused, with no change of state. Attempts to place a call are completed instantly with some choice of route, or are rejected, in accordance with a *routing matrix.*

A routing matrix $R = (r_{xy})$, $x, y \in S$, has the following properties: for each $x \in S$, let Π_x be the partition of A_x induced by the equivalence relation $\sim$ of "having the same calls up," or satisfying the same assignment of inlets to outlets; then for each $Y \in \Pi_x$, r_{xy} for $y \in Y$ is a *possibly improper* probability distribution over Y (that is, it may not sum to unity over Y)

$$r_{xx} = s(x) - \sum_{y \in A_x} r_{xy},$$

and $r_{xy} = 0$ in all other cases.

The interpretation of the routing matrix R is to be this: any $Y \in \Pi_x$ represents all the ways in which a particular call c not blocked in x (between an inlet idle in x and an outlet idle in x) *could* be completed when the network is in state x; for $y \in Y$, r_{xy} is the chance that if this call c is attempted in x, it will be completed by being routed through the network so as to take the system to state y. That is, we assume that if c is attempted in x, then with probability

$$1 - \sum_{y \in A_{cx}} r_{xy} \tag{1}$$

it is rejected (even though it is not blocked), and with probability r_{xy} it is completed by being assigned the route which would change the state

x to y, for $y \in A_{cx}$. The possibly improper distribution of probability $\{r_{xy}, y \in Y\}$ indicates how the calling rate λ due to c is to be spread over the possible ways of putting up the call c, while the improper part (1) is just the chance that it is rejected outright.

This description of routing matrices is a generalization of that used in [1] and [2] in that it permits, in the nonvanishing of (1), the rejection of unblocked calls, forbidden in the cited references.

Thus a routing matrix R is any function on S^2 with $r_{xy} \geqq 0$, $r_{xy} = 0$ unless $y \in A_x$ or $y = x$, and such that

$$r_{xx} = s(x) - \sum_{y \in A_x} r_{xy}$$

and

$$\sum_{y \in A_{cx}} r_{xy} \leqq 1,$$

for all $c \in x$ not blocked in x. A routing matrix corresponds to a *fixed rule* if $r_{xy} = 0$ or 1 for $x \neq y$; otherwise it corresponds to a *randomized rule.* The convex set of all possible routing matrices is denoted by C.

The routing rules and doctrines that might be considered here are of course more numerous by far than those we have introduced above. In particular, time-dependent rules and history-dependent rules are natural generalizations. However, since we will be considering only time-invariant traffic and ergodic Markov processes as representations of operating networks, such generalizations add very little of any significance.

An important point, however, is that the routing methods considered here are based on a complete knowledge of the state of the system, i.e., we postulate that we are in the case of "perfect information." This postulate is grossly unrealistic for present day electromechanical telephone systems; for an electronic system with a very large and very cheap memory, it becomes realistic: the state of the network can actually be stored and the routing rule in use represented by a giant translator. Such a procedure overcomes the obvious impracticality of determining the state by examination of the actual network, and is actually used in the Bell System's No. 1 ESS (Electronic Switching System) (see [4] and the references therein).

The routing matrices R used in [1], [2] have the property that if a call is not blocked in a state, then it is completed in *some* way; *only* blocked attempts or attempts to busy terminals are rejected. Thus none of these rules for routing resembles the methods that are at present likely to be used in practice. However, since C contains rules that reject certain calls in certain states, even though these calls are not blocked, it turns out that a large class of routing rules which do mirror what might happen in practice is included in C.

Some of the simplest routing rules are not based on any knowledge

about the current state of the network. Given a call c that has been attempted, they provide a list of routes to be tried in order; the first route found available is used for the call. The list may include all possible routes for c, or only some of them. It is easy to construct a routing matrix to represent such a rule. Let $r_1, r_2, \cdots, r_n$ be the routes to be tried for a call c. For each state x in which c can occur, let $r_{xy} = 1$ if use of the first r_i that is available in x takes the system from x to y, and let $r_{xy} = 0$ for all other $y \in A_{cx}$. If no route for c that is available in x is among $r_1, \cdots, r_n$, then c is rejected in x even though it may not be blocked, simply because the "sieve" for finding routes is too coarse.

It was assumed in the previous paragraph that no information about the state was used. If it is known, e.g., in which element A of a partition Π of S the state currently is, a similar rule can be represented by a class of lists (of routes to be tried in order), one for each $A \in \Pi$. The same kind of construction then yields the appropriate R. Here the A such that $x_t \in A$ is acting as the "information state."

Thus many R from C which reject certain calls in certain states describe a rule which closely resembles what is done in practice, e.g., in the translator of the Bell System No. 4A Crossbar Switching System.

3. Probabilistic assumptions and stochastic processes. A Markov stochastic process x_t taking values on S is used as a mathematical description of an operating connecting network subject to random traffic. It is assumed that this operation is in accordance with one of the routing matrices R of §2. The rest of the process x_t is based on two simple probabilistic assumptions:

(1) holding-times of calls are mutually independent variates, each with the negative exponential distribution of unit mean;

(2) if u is an inlet idle in state x, and $v \neq u$ is any outlet, there is a (conditional) probability,

$$\lambda h + o(h), \quad \lambda > 0,$$

that u attempts a call to v in $(t, t + h)$ if $x_t = x$, as $h \to 0$.

The choice of unit mean for the holding-times merely means that the mean holding-time is being used as the unit of time, so that only the traffic parameter λ needs to be specified.

It is convenient to collect these assumptions and the chosen routing matrix R into one transition rate matrix $Q = (q_{xy})$ characteristic of x_t: this matrix is given by

$$q_{xy} = \begin{cases} 1 & \text{if } y \in B_x, \\ \lambda r_{xy} & \text{if } y \in A_x, \\ -|x| - \lambda[s(x) - r_{xx}] & \text{if } y = x, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In terms of the transition rate matrix Q it is possible to define an ergodic stationary Markov stochastic process $\{x_t, t \text{ real}\}$ taking values on S. The matrix $P(t)$ of transition probabilities,

$$p_{xy}(t) = \Pr\{x_t = y \mid x_0 = x\},$$

satisfies the equations of Kolmogorov,

$$\frac{d}{dt}P(t) = QP(t) = P(t)Q, \quad P(0) = I,$$

and is given formally by the formula

$$P(t) = \exp tQ.$$

Since the zero state (the state with no calls in progress) is accessible from any state in a finite number of steps with positive probability, the process has only one ergodic class, and there exists a unique nonnegative row vector

$$p = \{p_x, x \in S\}$$

such that as $t \to \infty$,

$$P(t) \to \begin{pmatrix} p \\ \vdots \\ p \end{pmatrix},$$

and p satisfies the "statistical equilibrium" or stationarity condition $p'Q = 0$, which can be written out in full in the simple form

$$[\,|\,x\,|\, + \lambda s(x) - \lambda r_{xx}]p_x = \sum_{y \in A_x} p_y + \lambda \sum_{y \in B_x} p_y r_{yx}, \qquad x \in S.$$

It is possible that confusion might arise in the mind of the reader as to whether we are talking about central office connecting networks or large trunk networks such as the toll system. For in telephone traffic theory these two areas of application are often described by different models: a "finite source" model like the present one, in which the conditions of the inlets and outlets form a significant part of the state of the system, is commonly used for the former; an "infinite source" model, with groups of customers' lines reduced to Poisson sources of traffic, is frequently used for the latter. The reason for this difference is that it has simply turned out to be sufficient, in the toll case, to restrict attention to the trunking network as the object of principal interest, and to use the simpler Poisson description of sources.

In principle, of course, the model to be used here serves to describe either area listed above, although in the toll case it naturally demands use of a

very large number of states. Thus in the sequel we make no attempt to distinguish the toll case from the central office case. This viewpoint is justified by the fact that the results to be obtained are robust under passage from finite to infinite source models, or they can be reformulated and reproved in the infinite source context.

4. Formulation of the routing problem. The most common figure of merit used by telephone traffic engineers for evaluating connecting networks is the probability of blocking, the fraction of call attempts that are blocked. It is natural, therefore, to use this quantity as the objective function in our optimization problem of routing. It has been shown [2] for the process x_t to be studied here that if no unblocked call is rejected, then the probability of blocking (in the mnemonic form Pr{bl}) is given in terms of the stationary state probability vector p by the formula

$$\Pr\{\text{bl}\} = \frac{\sum_{x \in S} p_x \beta_x}{\sum_{x \in S} p_x \alpha_x} = \frac{p'\beta}{p'\alpha},$$

where β_x is the number of idle inlet-outlet pairs that are blocked in state x, and α_x is the number of idle inlet-outlet pairs in state x.

By the same methods it follows that for a process x_t defined in terms of an $R \in C$ the fraction of attempted calls which are not completed (are "lost"), be it because they are blocked or simply rejected, is given by

$$\frac{p'(\beta + r)}{p'\alpha},$$

where $r = \{r_{xx}, x \in S\}$ is the diagonal of the routing matrix R.

We can now replace the informal problem of minimizing, by suitable routing, the fraction of call attempts that are lost by a precise problem of mathematical programming, as follows: choose $R \in C$ so as to achieve

$$\min \frac{p'(\beta + r)}{p'\alpha}, \tag{3}$$

subject to $p'Q = 0$, $p'1 = 1$, and $p \geqq 0$. (The "1" in "$p'1$" is the vector with all components 1.) Of the constraints, the first is the equilibrium condition on p, the second states that the components of p sum to one, and the third says that p is nonnegative. It is understood, of course, that Q is to be related to R by (2) or, what is the same, by

$$Q = H + \lambda R + \operatorname{diag}(\,|\,x\,| + \lambda s(x) - 2\lambda r_{xx}) = Q(R),$$

where $H = (h_{xy})$ is the "hangup matrix" such that $h_{xy} = 1$ or 0 according as $y \in B_x$ or not.

5. Optimality of fixed rules. If a routing matrix has any entries other than integers, its use introduces a certain amount of additional randomness into the operation of the network, over and above that due to the random traffic, and may be said to represent a "mixed" strategy. It is a natural intuition that since minimizing the probability of loss is a game played against nature, rather than against an intelligent adversary, there can be no real gain from this additional randomization, i.e., that a fixed rule can be found that is as good as any "mixed strategy." To this effect we prove

THEOREM 1.

$$\min \frac{p'(\beta + r)}{p'\alpha},$$

subject to $R \in C$, $p'Q = 0$, $p'1 = 1$, $p \geqq 0$, $Q = Q(R)$, *is achieved by a fixed rule.*

6. Reduction to linear programming problems. It is possible to describe *linear* programming problems which are equivalent to our *nonlinear* problem of optimal routing [5]. One such reduction will be given as an example. Consider the "adjoined" *linear* problem of finding $R \in C$ and $t \geqq 0$ such that

$$q'(\beta + r) = \min,$$

subject to $q'Q = 0$, $q \geqq 0$, $q'1 - t = 0$, $q'\alpha = a$, and $Q = Q(R)$ as before, where a is a given positive number. (To see that this problem is indeed linear, we change variables to $u_{xy} = q_x r_{xy}$.) We can then prove:

THEOREM 2. *For any* $a > 0$, R *and* t *are a solution of the adjoined linear problem if and only if* R *is a solution of the routing problem* (3).

7. Trying to get closer to the optimal routing rules. It is particularly important to try to verbalize, and eventually to mechanize, routing strategies that are optimal, near optimal, or by some yardstick just "good." In this endeavor, the fact that the original routing problem (3) can be formulated as a linear programming problem, while interesting theoretically and perhaps reassuring, is nevertheless of limited usefulness. For this reason we have attempted to take advantage of some of the special properties of the problem that are due to its telephonic origins, and to describe at least part of the optimal policy in terms of combinatorial properties of the connecting network.

In order to illustrate one possible approach we limit our discussion to a very simple example; once the idea is understood, the principle involved can be abstracted, and a general theorem proved.

It can be shown that minimizing the probability of blocking is equivalent to maximizing the fraction of events that are successful attempts, where an event is either a hangup, a blocked attempt, or a successful one. This

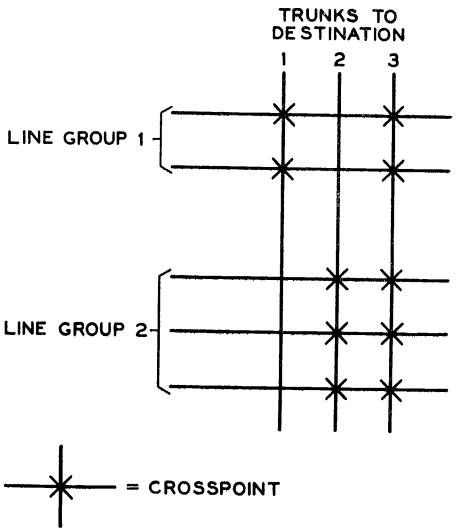

FIG. 1. *Overflow system*

maximal fraction is the limit, as n becomes large, of

$$\frac{1}{n} E_x(n),$$

where $E_x(n)$ is the expected number of successful calls in n events, if the network starts in state x and an optimal policy is followed. We shall base our approach on the vectors $E(n)$.

For our example we choose the overflow system or grading depicted in Fig. 1. There are two groups of lines, one of two lines, the other of three lines. Each has access to one primary trunk to which the other does not have access, and they share a single common overflow trunk. The possible states of this system form the partially ordered system shown in Fig. 2. Alternative ways of putting up particular calls are marked with "ch", for "choice."

After inspecting the system and its state diagram, intuition tells us that, as a first guess, calls should use the primary trunks whenever they can, so as to leave the overflow open as much as possible. Let us, on this basis, formulate some preferences for certain routes.

Clearly, in state 0 a call from group 1 should go on trunk 1, so in state 0 we prefer state (1-1) to (1-3); similarly we prefer (2-2) to (2-3). The same principle should apply if certain calls are already in progress. Thus in state (2-2) we prefer (1-1) (2-2) over (1-3) (2-2), and in state (1-1) we prefer (1-1) (2-2) to (1-1) (2-3).

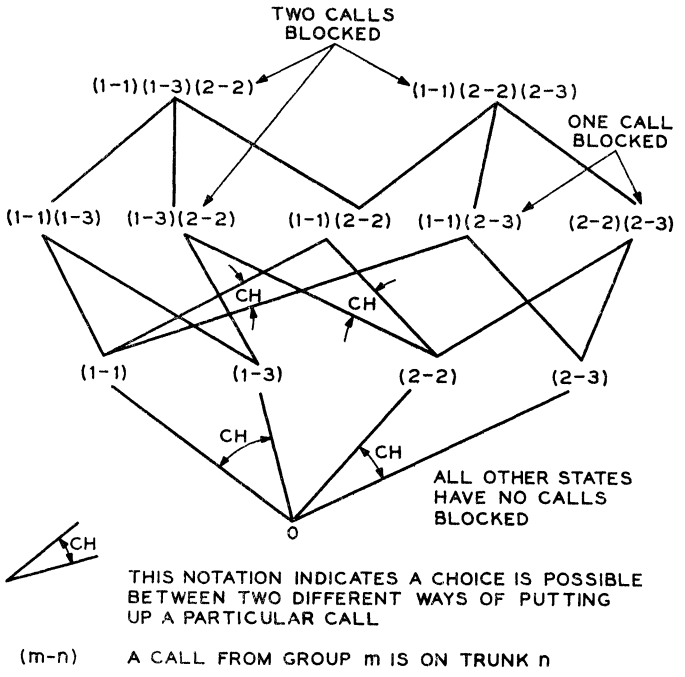

FIG. 2. *State diagram for overflow system*

If taken seriously and followed, the preferences listed above define a policy for putting in calls. We shall show that this policy differs from the optimal policy only in that the latter may reject some calls, while the former accepts all unblocked calls. To do this write xPy if state x is preferred to state y. Thus the relation P is defined by the conditions

$$(1\text{-}1)\ P\ (1\text{-}3),$$

$$(2\text{-}2)\ P\ (2\text{-}3),$$

$$(1\text{-}1)\ (2\text{-}2)\ P\ (1\text{-}3)\ (2\text{-}2),$$

$$(1\text{-}1)\ (2\text{-}2)\ P\ (1\text{-}1)\ (2\text{-}3).$$

We let $E_x(n)$ be the expected number of successful call attempts in n events, if the system starts in state x and an optimal policy is used. It must be explained here that by "use of an optimal policy" over n steps we mean simply that we use a policy which will maximize the average number of successful attempts *among those n events*; the policies that achieve this may, for all we know at this point, be different for different n.

A slight departure from the probabilistic model of §3 is necessary here: we assume that an idle line generates calls to the trunk destination at a

rate $\lambda > 0$, instead of assuming that an idle inlet-outlet pair generates calls at λ. Also, we let α_x be the number of idle lines in x, rather than that of idle inlet-outlet pairs, and $s(x)$ that of idle lines that are not blocked.

THEOREM 3. *If xPy then*

$$E_x(n) \geqq E_y(n), \quad n = 1, 2, 3, \cdots .$$

Proof.

$$E_x(1) = \frac{\lambda s(x)}{|x| + \lambda\alpha_x},$$

and xPy implies $s(x) \geqq s(y)$, so the theorem is true for $n = 1$. Assume that the theorem holds for some $n \geqq 1$. There are four cases, corresponding to the four conditions defining P. We shall give the argument for the case where

$$x = (1\text{-}1)\ (2\text{-}2), \qquad y = (1\text{-}3)\ (2\text{-}2),$$

and (as we know) xPy; the others are similar.

Now apparently,

$$E_{(1\text{-}1)(2\text{-}2)}(n+1) = \frac{1}{2+3\lambda}\{E_{(2\text{-}2)}(n) + E_{(1\text{-}1)}(n)\}$$
$$+ \frac{\lambda}{2+3\lambda}\max\{E_{(1\text{-}1)(2\text{-}2)}(n),\ 1 + E_{(1\text{-}1)(1\text{-}3)(2\text{-}2)}(n)\}$$
$$+ \frac{2\lambda}{2+3\lambda}\max\{E_{(1\text{-}1)(2\text{-}2)}(n),\ 1 + E_{(1\text{-}1)(2\text{-}2)(2\text{-}3)}(n)\},$$

and

$$E_{(1\text{-}3)(2\text{-}2)}(n+1) = \frac{1}{2+3\lambda}\{E_{(2\text{-}2)}(n) + E_{(1\text{-}3)}(n)\}$$
$$+ \frac{\lambda}{2+3\lambda}\max\{E_{(1\text{-}3)(2\text{-}2)}(n),\ 1 + E_{(1\text{-}1)(1\text{-}3)(2\text{-}2)}(n)\}$$
$$+ \frac{2\lambda}{2+3\lambda}E_{(1\text{-}3)(2\text{-}2)}(n).$$

By the induction hypothesis,

$$E_{(1\text{-}1)}(n) \geqq E_{(1\text{-}3)}(n),$$
$$E_{(1\text{-}1)(2\text{-}2)}(n) \geqq E_{(1\text{-}3)(2\text{-}2)}(n);$$

hence,

$$E_x(n+1) \geqq E_y(n+1),$$

for the given x and y.

The point is that each event that can occur leads to a "worse" state in y than it does in x. Thus the hangup of the group 1 call leads both to the state (2-2), a standoff; hangup of the group 2 call takes x into (1-1) and y into (1-3), and (1-1)P(1-3); one of the possible new calls leads both x and y to the state (1-1) (1-3) (2-2), another standoff; the other two possible new calls are blocked in y but not in x, so that by the induction hypothesis, rejecting one of them and staying in x is at least as good as having one of these blocked calls make an attempt in y.

We conclude from Theorem 3 that in the optimal policy the calls which are not rejected are put on the primary trunks if these are available, and on the overflow only if the primary trunk appropriate to the call is already busy. This is entirely in agreement with our original intuition.

REFERENCES

[1] V. E. Beneš, *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, New York, 1965.

[2] ———, *Markov processes representing traffic in connecting networks*, Bell System Tech. J., 42(1963), pp. 2795–2838.

[3] J. H. Weber, *Some traffic characteristics of communication networks with automatic alternate routing*, Ibid., 41(1962), pp. 1201–1247.

[4] R. W. Ketchledge, *The No. 1 Electronic Switching System*, IEEE Trans. Comm. Technology, COM-13(1965), pp. 38–41.

[5] A. Charnes and W. W. Cooper, *Programming with linear fractional functionals*, Naval Res. Logist. Quart., 9(1962), pp. 181–185.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

FURTHER INVESTIGATION INTO THE GEOMETRY OF OPTIMAL PROCESSES*

A. BLAQUIÈRE†

This paper is a continuation of the work developed particularly in [1], [2], namely, it is a further investigation into the geometry of optimally controlled systems. In order to limit the length of our account, we will refer the reader to these papers, chiefly to §§1–8 of [2] which contain the basis of our theory. Here we shall suppose that the reader is acquainted with the primary concepts. We shall start from [2, §9] and investigate from a different viewpoint some local properties of limiting surfaces Σ (whose definition is given in the above references). We shall limit our discussion to the case of interior points of Σ.

Indeed we shall use the same notations, except for the fact that $\bar{P}$ and $\bar{\Gamma}$ will be replaced by $\hat{P}$ and $\hat{\Gamma}$ respectively. Moreover $\mathcal{P}(\mathbf{x})$, or $\mathcal{P}(\mathbf{x}(t))$, will denote the head of vector $\mathbf{x}$ in E^{n+1}. Also we shall comply with the following classical notations. If Ω is any set in a topological space, then Ω° is the largest open set contained in Ω and $\bar{\Omega}$ is the topological closure of Ω.

Most of our derivations will be based on the concept of *separability* without recourse to the concept of convexity. More precisely we will meet with two kinds of properties, some of which will rely on the *separability*, others on the *nonseparability*, of tangent local cones which we shall next define. By the way, as pointed out in [1], [2], the maximum principle appears as a consequence of this geometrical analysis. However, in this paper, we will not lay stress upon its derivation.

1. Local properties of Σ surfaces. First of all we will define the *tangent cone* $\mathcal{S}(\mathbf{x})$ and the regions $A/\mathcal{S}(\mathbf{x})$, $B/\mathcal{S}(\mathbf{x})$, at any interior point $\mathcal{P}(\mathbf{x})$ of Σ.

1.1. Regions $A/\mathcal{S}(\mathbf{x})$ and $B/\mathcal{S}(\mathbf{x})$. Consider the region $\widetilde{A/\Sigma}$, namely,

$$\widetilde{A/\Sigma} \overset{\Delta}{=} (A/\Sigma) \cup \Sigma, \quad (\widetilde{A/\Sigma}) \cap (B/\Sigma) = \varnothing,$$

and let $\mathbf{n}$ be any bound vector at point $\mathcal{P}(\mathbf{x})$.

As a first basic assumption let us assume that there exists $\sigma > 0$ such

* Received by the editors May 26, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Institut d'Électronique, Faculté des Sciences de Paris, Paris, France. This research was supported in part by the National Science Foundation under Grant GP-803 and has been reported in [1] and [8].

that, for every ϵ such that $0 < \epsilon < \sigma$, $\mathcal{P}(\mathbf{x} + \epsilon\,\mathbf{n})$ belongs to *one* of the two regions $\widetilde{A/\Sigma}$, B/Σ.

Now let us introduce the following definitions.

DEFINITION 1. $\mathbf{n} \in A/\mathcal{S}(\mathbf{x})$ if $\exists\alpha > 0\ \forall\epsilon\colon 0 < \epsilon < \alpha$, $\mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}) \in \widetilde{A/\Sigma}$.

DEFINITION 2. $\mathbf{n} \in B/\mathcal{S}(\mathbf{x})$ if $\exists\beta > 0\ \forall\epsilon\colon 0 < \epsilon < \beta$, $\mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}) \in B/\Sigma$.

$A/\mathcal{S}(\mathbf{x})$ and $B/\mathcal{S}(\mathbf{x})$ are sets of vectors emanating from $\mathcal{P}(\mathbf{x})$. Namely, $A/\mathcal{S}(\mathbf{x})$ and $B/\mathcal{S}(\mathbf{x})$ are local cones:

$$A/\mathcal{S}(\mathbf{x}) \overset{\Delta}{=} \{\mathbf{n}\colon \mathbf{n} \text{ obeys Definition 1}\},$$

$$B/\mathcal{S}(\mathbf{x}) \overset{\Delta}{=} \{\mathbf{n}\colon \mathbf{n} \text{ obeys Definition 2}\}.$$

1.2. Interior vectors of $A/\mathcal{S}(\mathbf{x})$ and $B/\mathcal{S}(\mathbf{x})$. First of all we shall define interior vectors of $A/\mathcal{S}(\mathbf{x})$ and $B/\mathcal{S}(\mathbf{x})$.

Given $\mathbf{n}$: $|\,\mathbf{n}\,| = 1$, $\mathbf{n} \in A/\mathcal{S}(\mathbf{x})$, for instance, let us consider a conic neighborhood Δ of $\mathbf{n}$ in E^{n+1}. Δ is composed of vectors $\mathbf{n}^i$, $|\,\mathbf{n}^i\,| = 1$, emanating from $\mathcal{P}(\mathbf{x})$. We shall say that $\mathbf{n}$ is an interior vector of $A/\mathcal{S}(\mathbf{x})$ if there exists a conic neighborhood $\Delta = \Delta_A$ of $\mathbf{n}$, in E^{n+1}, such that

$$\forall\mathbf{n}^i \in \Delta_A\ \exists\alpha > 0\ \forall\epsilon\colon 0 < \epsilon < \alpha, \quad \mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}^i) \in \widetilde{A/\Sigma}.$$

Likewise we shall say that $\mathbf{n}$ is an interior vector of $B/\mathcal{S}(\mathbf{x})$ if there exists a neighborhood $\Delta = \Delta_B$ of $\mathbf{n}$, in E^{n+1}, such that

$$\forall\mathbf{n}^i \in \Delta_B\ \exists\beta > 0\ \forall\epsilon\colon 0 < \epsilon < \beta, \quad \mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}^i) \in B/\Sigma.$$

Then we define

$$(A/\mathcal{S}(\mathbf{x}))^\circ \overset{\Delta}{=} \{k\mathbf{n}\colon |\,\mathbf{n}\,| = 1,\ \mathbf{n} \text{ is interior vector of } A/\mathcal{S}(\mathbf{x}),\ k > 0\},$$

$$(B/\mathcal{S}(\mathbf{x}))^\circ \overset{\Delta}{=} \{k\mathbf{n}\colon |\,\mathbf{n}\,| = 1,\ \mathbf{n} \text{ is interior vector of } B/\mathcal{S}(\mathbf{x}),\ k > 0\}.$$

As a second basic assumption, we shall assume that, if $\mathbf{n}$ is an interior vector of $A/\mathcal{S}(\mathbf{x})$ (or $B/\mathcal{S}(\mathbf{x})$), $|\,\mathbf{n}\,| = 1$, then there exists a Δ_A (or Δ_B) such that

$$\inf_{\mathbf{n}^i \in \Delta_A} \alpha = \alpha_m, \quad \alpha_m > 0$$

(or

$$\inf_{\mathbf{n}^i \in \Delta_B} \beta = \beta_m, \quad \beta_m > 0).$$

Now let us prove the following property.

PROPERTY 1. *If* $\mathbf{n} \in (A/\mathcal{S}(\mathbf{x}))^\circ$, *then*

$$\exists\alpha > 0\ \forall\epsilon\colon 0 < \epsilon < \alpha, \quad \mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}) \in A/\Sigma.$$

Proof. Obviously, we can assume without loss of generality that $|\,\mathbf{n}\,| = 1$. Consider a neighborhood Δ_A of $\mathbf{n}$ in $A/\mathcal{S}(\mathbf{x})$, and the ball S_m defined by

$$S_m = \{\mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}^i)\colon |\,\mathbf{n}^i\,| = 1, \quad 0 < \epsilon \leqq \alpha_m\}.$$

Then the set of points

$$\Omega_m = \{\mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}^i) : |\mathbf{n}^i| = 1, \quad 0 < \epsilon \leqq \alpha_m, \quad \mathbf{n}^i \in \Delta_A\}$$

belongs to $\widetilde{A/\Sigma}$, and for every ϵ such that $0 < \epsilon < \alpha_m$, $\mathcal{P}(\mathbf{x} + \epsilon\mathbf{n})$ is an interior point of Ω_m, so it is an interior point of $\widetilde{A/\Sigma}$, which proves Property 1.

1.3. Cone $\mathcal{S}(\mathbf{x})$ tangent to a limiting surface. As a matter of fact, regions $A/\mathcal{S}(\mathbf{x})$ and $B/\mathcal{S}(\mathbf{x})$ may be neither open nor closed regions, depending on the local properties of Σ. One can easily prove that they have the following properties.

PROPERTY 2.

$$(A/\mathcal{S}(\mathbf{x})) \cup (B/\mathcal{S}(\mathbf{x})) = E^{n+1}.$$

Proof. This is due to the fact that, whatever $\mathbf{n} \in E^{n+1}$, it belongs either to $A/\mathcal{S}(\mathbf{x})$ or to $B/\mathcal{S}(\mathbf{x})$, which are the only two possible cases.

PROPERTY 3. *Let L_- and L_+ be open rays emanating from $\mathcal{P}(\mathbf{x})$, parallel to the x_0-axis, pointing into the negative x_0-direction and into the positive x_0-direction, respectively. Then*

$$L_+ \subset A/\mathcal{S}(\mathbf{x}), \quad L_- \subset B/\mathcal{S}(\mathbf{x}).$$

Hence $A/\mathcal{S}(\mathbf{x}) \neq \varnothing$, $B/\mathcal{S}(\mathbf{x}) \neq \varnothing$.

Proof. This property is a consequence of the definition of regions A/Σ, B/Σ, according to which

$$\forall \mathbf{n} \in L_+ \; \forall \epsilon > 0, \quad \mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}) \in A/\Sigma,$$

and

$$\forall \mathbf{n} \in L_- \; \forall \epsilon > 0, \quad \mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}) \in B/\Sigma.$$

Then Property 3 is a consequence of Definitions 1 and 2.

PROPERTY 4.

$$(A/\mathcal{S}(\mathbf{x})) \cap (B/\mathcal{S}(\mathbf{x})) = \varnothing.$$

Proof. In order to prove Property 4, assume that

$$(A/\mathcal{S}(\mathbf{x})) \cap (B/\mathcal{S}(\mathbf{x})) \neq \varnothing,$$

and let

$$\mathbf{n} \in A/\mathcal{S}(\mathbf{x}), \quad \mathbf{n} \in B/\mathcal{S}(\mathbf{x}).$$

Then

$$\exists \alpha > 0 \; \forall \epsilon : 0 < \epsilon < \alpha, \quad \mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}) \in \widetilde{A/\Sigma},$$

$$\exists \beta > 0 \ \forall \epsilon\colon 0 < \epsilon < \beta, \quad \mathcal{P}(\mathbf{x} + \epsilon \mathbf{n}) \in B/\Sigma.$$

Assume for instance that $\alpha < \beta$, then

$$\forall \epsilon\colon 0 < \epsilon < \alpha, \quad \mathcal{P}(\mathbf{x} + \epsilon \mathbf{n}) \in \widetilde{A/\Sigma} \text{ and } \mathcal{P}(\mathbf{x} + \epsilon \mathbf{n}) \in B/\Sigma,$$

which would imply $(\widetilde{A/\Sigma}) \cap (B/\Sigma) \neq \varnothing$. But since this is not true, Property 4 is established.

From Properties 2, 3, 4, one can conclude that the couple of regions $\{A/\mathcal{S}(\mathbf{x}), B/\mathcal{S}(\mathbf{x})\}$ is a *partition* of E^{n+1}. Hence

$$A/\mathcal{S}(\mathbf{x}) = \complement B/\mathcal{S}(\mathbf{x}).$$

Definition 3. The *local cone* $\mathcal{S}(\mathbf{x})$ *associated to* Σ *at point* $\mathcal{P}(\mathbf{x})$ is the common boundary of $A/\mathcal{S}(\mathbf{x})$ and $B/\mathcal{S}(\mathbf{x})$, namely,

$$\mathcal{S}(\mathbf{x}) \overset{\Delta}{=} (\overline{A/\mathcal{S}(\mathbf{x})}) \cap (\overline{B/\mathcal{S}(\mathbf{x})}).$$

1.4. Further properties of regions $A/\mathcal{S}(\mathbf{x})$ and $B/\mathcal{S}(\mathbf{x})$. By means of Definitions 1, 2, 3, and Property 1 we can deduce:

Lemma 1. *Given a vector* $\mathbf{n}(\epsilon)$ *from a point* $\mathcal{P}(\mathbf{x})$ *of* Σ, *which is a continuous function of parameter* ϵ, $\epsilon > 0$; *if*

$$\mathbf{n}(\epsilon) \to 1 \quad as \quad \epsilon \to 0$$

and

$$\mathcal{P}(\mathbf{x} + \epsilon \mathbf{n}(\epsilon)) \in \widetilde{A/\Sigma} \quad \forall \epsilon < \gamma,$$

where γ *is a positive number, then*

$$1 \in \overline{A/\mathcal{S}(\mathbf{x})}.$$

Lemma 2. *Under the same assumptions, if*

$$\mathcal{P}(\mathbf{x} + \epsilon \mathbf{n}(\epsilon)) \in \widetilde{B/\Sigma} \quad \forall \epsilon < \gamma,$$

where $\widetilde{B/\Sigma} \overset{\Delta}{=} (B/\Sigma) \cup \Sigma$, *then*

$$1 \in \overline{B/\mathcal{S}(\mathbf{x})}.$$

Proofs. In order to prove Lemma 1, assume that our assertion is incorrect, namely,

$$1 \notin \overline{A/\mathcal{S}(\mathbf{x})},$$

then

$$1 \in (B/\mathcal{S}(\mathbf{x}))^{\circ}.$$

Then since $\mathbf{n}(\epsilon) \to 1$ there exists a positive number $\delta < \gamma$ such that

$$\epsilon < \delta \Rightarrow \mathbf{n}(\epsilon) \in (B/\mathcal{S}(\mathbf{x}))^\circ \Rightarrow \mathbf{n}(\epsilon) \in B/\mathcal{S}(\mathbf{x}).$$

Then in view of Definition 2 and of our second basic assumption, there exists a positive number σ, $\sigma < \delta < \gamma$, such that

$$\epsilon < \sigma \Rightarrow \mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}(\epsilon)) \in B/\Sigma,$$

which is in contradiction with our assumption.

To prove Lemma 2, assume that

$$\mathbf{1} \notin \overline{B/\mathcal{S}(\mathbf{x})}, \quad \text{say } \mathbf{1} \in (A/\mathcal{S}(\mathbf{x}))^\circ.$$

Then there exists a positive number $\delta < \gamma$, such that

$$\epsilon < \delta \Rightarrow \mathbf{n}(\epsilon) \in (A/\mathcal{S}(\mathbf{x}))^\circ.$$

Then in view of Property 1,

$$\exists \sigma > 0\colon\ \sigma < \delta < \gamma\ \forall \epsilon\colon\ 0 < \epsilon < \sigma, \quad \mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}) \in A/\Sigma,$$

which again contradicts our assumption.

The following corollary can be readily deduced from the above lemmas.

COROLLARY 1. *If* $\mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}(\epsilon)) \in \Sigma\ \forall \epsilon < \gamma$, *then* $\mathbf{1} \in \mathcal{S}(\mathbf{x})$.

Indeed

$$\Sigma = (\widetilde{A/\Sigma}) \cap (\widetilde{B/\Sigma}),$$

$$\mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}(\epsilon)) \in \widetilde{A/\Sigma} \Rightarrow \mathbf{1} \in \overline{A/\mathcal{S}(\mathbf{x})},$$

$$\mathcal{P}(\mathbf{x} + \epsilon\mathbf{n}(\epsilon)) \in \widetilde{B/\Sigma} \Rightarrow \mathbf{1} \in \overline{B/\mathcal{S}(\mathbf{x})}.$$

Finally, $\mathbf{1} \in (\overline{A/\mathcal{S}(\mathbf{x})}) \cap (\overline{B/\mathcal{S}(\mathbf{x})})$ says that $\mathbf{1} \in \mathcal{S}(\mathbf{x})$.

1.5. Cone of f vectors. Next we shall consider systems governed by the set of differential equations [2, §5]

$$\dot{x}_j = f_j(x_1, \cdots, x_n, u_1, \cdots, u_m), \qquad j = 1, \cdots, n, \tag{1}$$

where $u_1, \cdots, u_m$ are control variables.

Given functions of time $u_1(t), \cdots, u_m(t)$, $t_0 \leqq t \leqq t_1$, these equations define a set of rules which govern the behavior of the system during time interval $[t_0, t_1]$.

We shall assume that vector $\mathbf{u}$, whose components are $u_1, \cdots, u_m$, belongs to a prescribed subset U of E^m.

Moreover, following the assumptions of [1], [2], we shall consider the functional

$$\int_{t_0}^{t_1} f_0(\mathfrak{x}(t), \mathbf{u}(t))\, dt,$$

where $\mathbf{u}(t)$ transfers the system, in E^n, from initial state P^0 either to pre-

scribed terminal state P^1 or to some other terminal state $\hat{P}^1$, in unspecified time $t_1 - t_0$.

Then one can easily show that the equation of Γ, or $\hat{\Gamma}$, is

$$x_0(t) + \int_t^{t_1} f_0(\mathfrak{x}(s), \mathbf{u}(s))\, ds = C,$$

whence

$$\dot{x}_0 = f_0(\mathfrak{x}, \mathbf{u}). \tag{2}$$

If we combine (1) and (2) into a single vector equation, we g

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}), \tag{3}$$

where

$$\mathbf{f}(\mathbf{x}, \mathbf{u}) = \begin{bmatrix} f_0(\mathfrak{x}, \mathbf{u}) \\ \vdots \\ f_n(\mathfrak{x}, \mathbf{u}) \end{bmatrix}.$$

Furthermore we shall assume that the functions

$$f_j(x_1, \cdots, x_n, u_1, \cdots, u_m), \qquad j = 0, 1, \cdots, n,$$

and

$$\frac{\partial f_j(x_1, \cdots, x_n, u_1, \cdots, u_m)}{\partial x_\alpha}, \qquad \alpha = 1, 2, \cdots, n$$

are defined and continuous on $E^n \times U$.

Consequently, given any constant vector $\mathbf{u}^b \in U$, (3) defines a constant *vector field*, namely a field of velocity vectors in E^{n+1}, which has the following properties:

(i) The lines of force of the field are integral curves of (3), namely, they are continuous solutions in E^{n+1}.

(ii) Through every point of E^{n+1}, there passes one and only one such trajectory, whose tangent is uniquely defined at that point.

Now consider such an integral curve‡ L whose running point passes through $\mathcal{P}(\mathbf{x}) \in \Sigma$ at any time t, which can be arbitrarily chosen, and let

$$\mathbf{n}'(\Delta t) \triangleq \frac{\Delta \mathbf{x}'}{\Delta t}, \quad \mathbf{n}''(\Delta t) \triangleq \frac{\Delta \mathbf{x}''}{\Delta t},$$

where

$$\Delta \mathbf{x}' = \Delta \mathbf{x}(t') \triangleq \mathbf{x}(t + \Delta t) - \mathbf{x}(t), \quad \Delta \mathbf{x}'' = \Delta \mathbf{x}(t'') \triangleq \mathbf{x}(t - \Delta t) - \mathbf{x}(t),$$
$$t' = t + \Delta t, \quad t'' = t - \Delta t, \quad \Delta t > 0,$$

‡ Contrary to Γ or $\hat{\Gamma}$, L has neither initial point nor endpoint, since the vector field stretches throughout the whole space.

$$\mathcal{P}(\mathbf{x} + \Delta\mathbf{x}') \in L, \quad \mathcal{P}(\mathbf{x} + \Delta\mathbf{x}'') \in L,$$

$\mathbf{n}'(\Delta t)$ and $\mathbf{n}''(\Delta t)$ are continuous functions of $\Delta t > 0$, and

$$\mathbf{n}'(\Delta t) \to \mathbf{f}(\mathbf{x}, \mathbf{u}^b),$$

$$\mathbf{n}''(\Delta t) \to -\mathbf{f}(\mathbf{x}, \mathbf{u}^b), \qquad \text{as } \Delta t \to 0;$$

and, as a consequence of the global properties of Σ-surfaces (see [2, Theorem I])

$$\mathcal{P}(\mathbf{x} + \Delta\mathbf{x}') = \mathcal{P}(\mathbf{x} + \Delta t\, \mathbf{n}'(\Delta t)) \in \widetilde{A/\Sigma},$$

$$\mathcal{P}(\mathbf{x} + \Delta\mathbf{x}'') = \mathcal{P}(\mathbf{x} + \Delta t\, \mathbf{n}''(\Delta t)) \in \widetilde{B/\Sigma}, \qquad \forall \Delta t.$$

Then from Lemmas 1 and 2 follows:

LEMMA 3. *At any point* $\mathcal{P}(\mathbf{x})$ *of a* Σ*-surface,*

$$\mathbf{f}(\mathbf{x}, \mathbf{u}^b) \in \overline{A/\mathcal{S}(\mathbf{x})},$$

$$-\mathbf{f}(\mathbf{x}, \mathbf{u}^b) \in \overline{B/\mathcal{S}(\mathbf{x})}, \qquad \forall \mathbf{u}^b \in U.$$

1.6. Cone $\mathcal{C}_n(\mathbf{x})$. First of all let us introduce the following definitions:

DEFINITION 4. *An n-dimensional separating hyperplane* $\mathfrak{I}(\mathbf{x})$ of the closed cone $\overline{A/\mathcal{S}(\mathbf{x})}$, or $\overline{B/\mathcal{S}(\mathbf{x})}$, at point $\mathcal{P}(\mathbf{x}) \in \Sigma$, is an n-dimensional hyperplane through $\mathcal{P}(\mathbf{x})$ such that all the vectors $\mathbf{n}$ of $\overline{A/\mathcal{S}(\mathbf{x})}$, or $\overline{B/\mathcal{S}(\mathbf{x})}$, lie in one of the closed halfspaces determined by $\mathfrak{I}(\mathbf{x})$.

Let us call $\bar{R}_A$ or $\bar{R}_B$, depending on the case which is considered, the corresponding closed halfspace, and R_A or R_B, respectively, the corresponding open halfspace.

DEFINITION 5. If there exists an n-dimensional separating hyperplane of $\overline{A/\mathcal{S}(\mathbf{x})}$ or of $\overline{B/\mathcal{S}(\mathbf{x})}$, the cone $\overline{A/\mathcal{S}(\mathbf{x})}$ or $\overline{B/\mathcal{S}(\mathbf{x})}$ is *separable.*

When $\overline{A/\mathcal{S}(\mathbf{x})}$ or $\overline{B/\mathcal{S}(\mathbf{x})}$ is separable, let us consider any separating hyperplane $\mathfrak{I}(\mathbf{x})$, and vector $\mathbf{n}(\mathbf{x})$, $|\mathbf{n}(\mathbf{x})| = 1$, normal to $\mathfrak{I}(\mathbf{x})$ at point $\mathcal{P}(\mathbf{x})$.

(i) If $\overline{A/\mathcal{S}(\mathbf{x})}$ is separable, we choose $\mathbf{n}(\mathbf{x})$ such that $\mathbf{n}(\mathbf{x}) \in \complement \bar{R}_A$.

(ii) If $\overline{B/\mathcal{S}(\mathbf{x})}$ is separable, we choose $\mathbf{n}(\mathbf{x})$ such that $\mathbf{n}(\mathbf{x}) \in R_B$.

Note that $\overline{A/\mathcal{S}(\mathbf{x})}$ and $\overline{B/\mathcal{S}(\mathbf{x})}$ are both separable at regular interior points of Σ. Then obviously assumptions (i) and (ii) are equivalent.

DEFINITION 6. $\mathcal{C}_n(\mathbf{x})$ is the set of all vectors $\mathbf{n}(\mathbf{x})$:

$$\mathcal{C}_n(\mathbf{x}) \overset{\Delta}{=} \{\mathbf{n}(\mathbf{x})\}.$$

Remark. It may easily be seen that, if $\overline{A/\mathcal{S}(\mathbf{x})}$ is separable, then $\overline{B/\mathcal{S}(\mathbf{x})}$ is nonseparable and conversely, except if $\mathcal{P}(\mathbf{x})$ is a regular interior point of Σ.

On the other hand it may be that, at some points of Σ, neither $\overline{A/\mathcal{S}(\mathbf{x})}$ nor $\overline{B/\mathcal{S}(\mathbf{x})}$ is separable.

2. Linear transformations. To each optimal trajectory Γ^*, generated by control $\mathbf{u}^*(t)$, $t_0 \leqq t \leqq t_1$, we have associated a linear transformation, defined by the variational equation

$$\dot{\mathbf{n}} = \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right)\Big|_{\mathbf{x}=\mathbf{x}^*(t)} \mathbf{n}. \tag{4}$$

Equation (4) defines a nonsingular linear operator $A(t', t)$ such that

$$\mathbf{n}(t) = A(t', t)\mathbf{n}', \qquad t' \leqq t \leqq t_1,$$

where $\mathbf{n}(t)$ is defined at point $\mathcal{P}(\mathbf{x}^*(t))$, and $\mathbf{n}' \overset{\Delta}{=} \mathbf{n}(t')$ at point $\mathcal{P}(\mathbf{x}^*(t'))$.

Some properties of this linear operator have been given in [1], [2]. Here let us only recall the most important ones which will be useful for the following arguments:

(i) The equation adjoint to variational equation (4) is

$$\dot{\lambda} = -\left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right)^T\Big|_{\mathbf{x}=\mathbf{x}^*(t)} \lambda. \tag{5}$$

For given initial condition $\lambda(t_0) = \lambda^0$, the solution of (5) is unique and continuous on $[t_0, t_1]$.

(ii) $\lambda(t)\cdot\mathbf{n}(t) = \text{const.}$, $t_0 \leqq t \leqq t_1$.

Moreover one can prove the following lemmas. Let

$$\mathbf{n}' \overset{\Delta}{=} \mathbf{n}(t'), \quad \mathbf{n}'' \overset{\Delta}{=} \mathbf{n}(t''), \quad \mathbf{n}'' = A(t', t'')\mathbf{n}', \quad t'' \geqq t'.$$

LEMMA 4. *If* $\mathbf{n}' \in \overline{A/\mathcal{S}(\mathbf{x}^*(t'))}$, *then* $\mathbf{n}'' \in \overline{A/\mathcal{S}(\mathbf{x}^*(t''))}$.

LEMMA 5. *If* $\mathbf{n}'' \in \overline{B/\mathcal{S}(\mathbf{x}^*(t''))}$, *then* $\mathbf{n}' \in \overline{B/\mathcal{S}(\mathbf{x}^*(t'))}$.

3. Theorems of separability; attractive and repulsive subsets of Σ. Now we shall prove the following theorems. Let $t'' > t'$.

THEOREM 1. *If* $\overline{B/\mathcal{S}(\mathbf{x}^*(t'))}$ *is separable, then* $\overline{B/\mathcal{S}(\mathbf{x}^*(t''))}$ *is separable.*

THEOREM 2. *If* $\overline{A/\mathcal{S}(\mathbf{x}^*(t''))}$ *is separable, then* $\overline{A/\mathcal{S}(\mathbf{x}^*(t'))}$ *is separable.*

Proofs. In order to prove Theorem 1, let $\mathfrak{I}(\mathbf{x}^*(t'))$ be an n-dimensional separating hyperplane of $\overline{B/\mathcal{S}(\mathbf{x}^*(t'))}$ at point $\mathcal{P}(\mathbf{x}^*(t'))$, which determines the closed halfspace $\overline{R_B'}$ and the open halfspace $\mathbf{C}\overline{R_B'}$ (following the notations of §1.6), namely all the vectors of $\overline{B/\mathcal{S}(\mathbf{x}^*(t'))}$ belong to $\overline{R_B'}$.

Let $\Pi(\mathbf{x}^*(t''))$, $\overline{\rho_B''}$, and $\mathbf{C}\overline{\rho_B''}$ be the transforms of $\mathfrak{I}(\mathbf{x}^*(t'))$, $\overline{R_B'}$, and $\mathbf{C}\overline{R_B'}$, respectively, at point $\mathcal{P}(\mathbf{x}^*(t''))$, by the linear transformation $A(t', t'')$.

Because the transformation $A(t', t'')$ is linear and nonsingular, the common boundary of $\overline{\rho_B''}$ and $\mathbf{C}\overline{\rho_B''}$ is $\Pi(\mathbf{x}^*(t''))$.

Consider any vector $\mathbf{n}''$, $\mathbf{n}'' \in \overline{B/\mathcal{S}(\mathbf{x}^*(t''))}$ at point $\mathcal{P}(\mathbf{x}^*(t''))$, and sup-

pose that it belongs to $\mathbf{C}\overline{\rho_B''}$. Then it is the transform of vector $\mathbf{n}'$ at $\mathcal{P}(\mathbf{x}^*(t'))$ such that

$$\mathbf{n}' \in \mathbf{C}\,\overline{R_B'}.$$

Moreover, because of Lemma 5,

$$\mathbf{n}' \in \overline{B/\mathcal{S}(\mathbf{x}^*(t'))}.$$

Since this is impossible we conclude that

$$\mathbf{n}'' \in \overline{\rho_B''} \quad \forall \mathbf{n}'' \in \overline{B/\mathcal{S}(\mathbf{x}^*(t''))}.$$

Accordingly,
(i) $\overline{B/\mathcal{S}(\mathbf{x}^*(t''))}$ is separable,
(ii) $\Pi(\mathbf{x}^*(t'')) = \mathfrak{H}(\mathbf{x}^*(t''))$ is a separating hyperplane of $\overline{B/\mathcal{S}(\mathbf{x}^*(t''))}$, and
(iii) $\overline{\rho_B''} = \overline{R_B''}$.

Theorem 2 can be readily proved by similar arguments, using Lemma 4.

From Theorems 1 and 2 follow:

COROLLARY 2. *An optimal trajectory* Γ^* *cannot join points* $\mathcal{P}(\mathbf{x}') = \mathcal{P}(\mathbf{x}^*(t'))$ *and* $\mathcal{P}(\mathbf{x}'') = \mathcal{P}(\mathbf{x}^*(t''))$, $t'' > t'$, *if* $\overline{B/\mathcal{S}(\mathbf{x}^*(t'))}$ *is separable and* $\overline{B/\mathcal{S}(\mathbf{x}^*(t''))}$ *is not separable.*

COROLLARY 3. *An optimal trajectory* Γ^* *cannot join points* $\mathcal{P}(\mathbf{x}') = \mathcal{P}(\mathbf{x}^*(t'))$ *and* $\mathcal{P}(\mathbf{x}'') = \mathcal{P}(\mathbf{x}^*(t''))$, $t'' > t'$, *if* $\overline{A/\mathcal{S}(\mathbf{x}^*(t'))}$ *is not separable and* $\overline{A/\mathcal{S}(\mathbf{x}^*(t''))}$ *is separable.*

Now let us define *attractive* and *repulsive* subsets of Σ.

DEFINITION 7. An *attractive subset of* Σ is a nonregular interior subset $M_a \subset \Sigma$, at all of whose points $\overline{B/\mathcal{S}(\mathbf{x})}$ is separable, and indeed $\overline{A/\mathcal{S}(\mathbf{x})}$ is not separable.

DEFINITION 8. A *repulsive subset of* Σ is a nonregular interior subset $M_r \subset \Sigma$, at all of whose points $\overline{A/\mathcal{S}(\mathbf{x})}$ is separable, and indeed $\overline{B/\mathcal{S}(\mathbf{x})}$ is not separable.

The names *attractive* and *repulsive* are explained by the following corollaries which are straightforward consequences of Corollaries 2 and 3.

COROLLARY 4. *If a point of an optimal trajectory* Γ^* *belongs to an attractive subset* M_a *of* Σ, *then* Γ^* *cannot go from* M_a *to a regular interior point of* Σ, *or to a point on a repulsive subset.*

However apparently, an optimal trajectory which starts at a regular interior point of Σ or at a point on a repulsive subset can reach an attractive subset.

COROLLARY 5. *An optimal trajectory* Γ^* *which starts at a regular interior point of* Σ, *or at a point on an attractive subset, cannot reach a repulsive subset.*

However apparently, an optimal trajectory which starts on, or belongs for a nonzero interval to, a repulsive subset can leave it.

4. Regular subsets, antiregular subsets of Σ. As pointed out earlier, at regular interior points of Σ both $\overline{A/\mathcal{S}(\mathbf{x})}$ and $\overline{B/\mathcal{S}(\mathbf{x})}$ are separable. Conversely if, at a point $\mathcal{P}(\mathbf{x})$ of Σ, $\overline{A/\mathcal{S}(\mathbf{x})}$ and $\overline{B/\mathcal{S}(\mathbf{x})}$ are both separable, $\mathcal{P}(\mathbf{x})$ is a regular point of Σ. Moreover one can prove the following.

COROLLARY 6. *If* $\mathcal{P}(\mathbf{x}^*(t'))$ *and* $\mathcal{P}(\mathbf{x}^*(t''))$, $t'' > t'$, *are regular points on an optimal trajectory* Γ^* *whose points are interior points of* Σ, *then* $\mathcal{P}(\mathbf{x}^*(t))$, $t' \leqq t \leqq t''$, *is a regular point of* Σ.

Indeed since $\overline{A/\mathcal{S}(\mathbf{x}^*(t''))}$ is separable, $\overline{A/\mathcal{S}(\mathbf{x}^*(t))}$ is separable, and since $\overline{B/\mathcal{S}(\mathbf{x}^*(t'))}$ is separable, $\overline{B/\mathcal{S}(\mathbf{x}^*(t))}$ is separable; accordingly, $\mathcal{P}(\mathbf{x}^*(t))$ is a regular point.

We shall also consider the case of interior subsets of Σ at all of whose points neither $\overline{A/\mathcal{S}(\mathbf{x})}$ nor $\overline{B/\mathcal{S}(\mathbf{x})}$ is separable. We shall call such subsets *antiregular* subsets. Properties of antiregular subsets are codified by the following corollary.

COROLLARY 7. *If a point of an optimal trajectory* Γ^* *belongs to an antiregular subset of* Σ, *then* Γ^* *cannot go from this subset to a regular interior point of* Σ *or to a point on a repulsive subset. An optimal trajectory* Γ^* *which starts at a regular interior point of* Σ, *or at a point on an attractive subset, cannot reach an antiregular subset.*

Again this Corollary is a straightforward consequence of Theorems 1 and 2.

5. Symmetrical subsets of $\mathcal{S}(\mathbf{x})$.

DEFINITION 9. *A symmetrical subset,* $I(\mathbf{x})$, of $\mathcal{S}(\mathbf{x})$ is defined by:

$$I(\mathbf{x}) \overset{\Delta}{=} \{\mathbf{n} : \mathbf{n} \in \mathcal{S}(\mathbf{x}) \text{ and } -\mathbf{n} \in \mathcal{S}(\mathbf{x})\}.$$

Consider points $\mathcal{P}(\mathbf{x}^*(t'))$ and $\mathcal{P}(\mathbf{x}^*(t''))$, $t'' > t'$, along optimal trajectory Γ^*, and assume that $\overline{A/\mathcal{S}(\mathbf{x}^*(t''))}$ is separable; then, according to Theorem 2, $\overline{A/\mathcal{S}(\mathbf{x}^*(t'))}$ is separable.

Let $\mathfrak{I}(\mathbf{x}^*(t'))$ and $\mathfrak{I}(\mathbf{x}^*(t''))$ be separating hyperplanes§ at $\mathcal{P}(\mathbf{x}^*(t'))$ and $\mathcal{P}(\mathbf{x}^*(t''))$, and $\overline{R_A}'$, $\overline{R_A}''$, the closed halfspaces which they respectively determine, following the definitions of §1.6, namely,

$$\overline{A/\mathcal{S}(\mathbf{x}^*(t'))} \subset \overline{R_A}',$$

$$\overline{A/\mathcal{S}(\mathbf{x}^*(t''))} \subset \overline{R_A}''.$$

On the other hand consider any vector $\mathbf{n}' = \mathbf{n}(t')$ at $\mathcal{P}(\mathbf{x}^*(t'))$ such that

$$\mathbf{n}' \in I(\mathbf{x}^*(t')).$$

This implies

$$\mathbf{n}' \in \mathcal{S}(\mathbf{x}^*(t')).$$

§ $\mathfrak{I}(\mathbf{x}^*(t''))$ is the transform of $\mathfrak{I}(\mathbf{x}^*(t'))$ due to $A(t', t'')$.

and, according to Lemma 4,

$$\mathbf{n}'' \in \overline{A/\mathcal{S}(\mathbf{x}^*(t''))}, \quad \mathbf{n}'' = A(t', t'')\mathbf{n}'.$$

Now if

$$\mathbf{n}'' \in (A/\mathcal{S}(\mathbf{x}^*(t')))^\circ,$$

then

$$\mathbf{n}'' \in R_A'',$$

which implies

$$-\mathbf{n}'' \in \mathbf{C}\,\overline{R_A''}.$$

But this is impossible since

$$-\mathbf{n}' \in \mathcal{S}(\mathbf{x}^*(t')) \Rightarrow -\mathbf{n}' \in \overline{R_A'} \Rightarrow -\mathbf{n}'' \in \overline{R_A''}.$$

Consequently

$$\mathbf{n}'' \in \mathcal{S}(\mathbf{x}^*(t'')).$$

Moreover, since $-\mathbf{n}' \in I(\mathbf{x}^*(t'))$, the above result also applies to the vector $-\mathbf{n}'$, say,

$$-\mathbf{n}'' \in \mathcal{S}(\mathbf{x}^*(t'')).$$

Accordingly,

$$\mathbf{n}'' \in I(\mathbf{x}^*(t'')).$$

We reach the following conclusion.

LEMMA 6. *Along a piece of optimal trajectory at all of whose points* $\overline{A/\mathcal{S}(\mathbf{x}^*(t))}$ *is separable,*

$$\mathbf{n}' \in I(\mathbf{x}^*(t')) \Rightarrow \mathbf{n}'' \in I(\mathbf{x}^*(t'')),$$

where

$$\mathbf{n}'' = A(t', t'')\mathbf{n}', \qquad t'' \geqq t'.$$

By similar arguments one can easily prove the following.

LEMMA 7. *Along a piece of optimal trajectory at all of whose points* $\overline{B/\mathcal{S}(\mathbf{x}^*(t))}$ *is separable,*

$$\mathbf{n}'' \in I(\mathbf{x}^*(t'')) \Rightarrow \mathbf{n}' \in I(\mathbf{x}^*(t')), \qquad t'' \geqq t'.$$

As a matter of fact these conclusions trivially apply to the case of zero vectors.

From Lemmas 6 and 7 one can easily deduce the following theorems.

THEOREM 3. *Along a piece of optimal trajectory at all of whose points*

$\overline{A/\mathcal{S}(\mathbf{x}^*(t))}$ *is separable, the dimension of* $I(\mathbf{x}^*(t))$ *is a nondecreasing function of* t.

THEOREM 4. *Along a piece of optimal trajectory at all of whose points* $\overline{B/\mathcal{S}(\mathbf{x}^*(t))}$ *is separable, the dimension of* $I(\mathbf{x}^*(t))$ *is a nonincreasing function of* t.

THEOREM 5. *If* $I(\mathbf{x}^*(t'))$ *and* $I(\mathbf{x}^*(t''))$ *are hyperplanes at points* $\mathcal{P}(\mathbf{x}^*(t'))$ *and* $\mathcal{P}(\mathbf{x}^*(t''))$, $t'' \geqq t'$, *of optimal trajectory* Γ^*, *where* $\overline{A/\mathcal{S}(\mathbf{x}^*(t'))}$ *and* $\overline{A/\mathcal{S}(\mathbf{x}^*(t''))}$, *or* $\overline{B/\mathcal{S}(\mathbf{x}^*(t'))}$ *and* $\overline{B/\mathcal{S}(\mathbf{x}^*(t''))}$, *are separable, and if* $I(\mathbf{x}^*(t'))$ *is of the same dimension as* $I(\mathbf{x}^*(t''))$, *then* $I(\mathbf{x}^*(t''))$ *is the transform of* $I(x^*(t'))$ *due to the linear transformation* $A(t', t'')$.

6. Degenerated case.

DEFINITION 10. A subset of Σ at all of whose points
(i) either $(B/\mathcal{S}(\mathbf{x}))^\circ = \varnothing$,
(ii) or $(A/\mathcal{S}(\mathbf{x}))^\circ = \varnothing$,
is called a *degenerated subset.*

In case (i) the degenerated subset will be called a *B-degenerated* subset, and in case (ii) it will be called an *A-degenerated* subset.

In this section we shall investigate some properties of an optimal trajectory a portion (or a point) of which belongs to a degenerated subset.

First of all let us consider the case of a B-degenerated subset, namely

$$(B/\mathcal{S}(\mathbf{x}^*(t)))^\circ = \varnothing.$$

Assume

$$(B/\mathcal{S}(\mathbf{x}^*(t')))^\circ \neq \varnothing, \qquad t' > t,$$

and let

$$\mathbf{n}' \in (B/\mathcal{S}(\mathbf{x}^*(t')))^\circ, \text{ where } \mathbf{n}' = \mathbf{n}(t').$$

Then $\mathbf{n}'$ is the transform of vector $\mathbf{n} = \mathbf{n}(t)$ due to the linear transformation $A(t, t')$. Moreover, as a straightforward consequence of Lemmas 4 and 5,

$$\mathbf{n} \in (B/\mathcal{S}(\mathbf{x}^*(t)))^\circ,$$

which contradicts the assumption. Accordingly $(B/\mathcal{S}(\mathbf{x}^*(t')))^\circ = \varnothing$. Hence:

THEOREM 6. *If point* $\mathcal{P}(\mathbf{x}^*(t))$ *of optimal trajectory* Γ^* *belongs to a B-degenerated subset and* $\mathcal{P}(\mathbf{x}^*(t'))$ *is an interior point of* Σ, *then* $\mathcal{P}(\mathbf{x}^*(t'))$ *belongs to a B-degenerated subset, whatever* $t' \geqq t$.

Now consider vector $\mathbf{n}'$ such that

$$\mathbf{n}' \in \mathcal{S}(\mathbf{x}^*(t')).$$

According to Lemma 5, it is the transform of vector $\mathbf{n} = \mathbf{n}(t)$, $t \leqq t'$, such that

$$\mathbf{n} \in \overline{B/\mathcal{S}(\mathbf{x}^*(t))};$$

and since we assume

$$(B/\mathcal{S}(\mathbf{x}^*(t)))^\circ = \varnothing,$$

it follows that

$$\mathbf{n} \in \mathcal{S}(\mathbf{x}^*(t)).$$

Hence:

THEOREM 7. *Along a piece of optimal trajectory* Γ^* *which belongs to a B-degenerated subset,*

$$\mathbf{n}' \in \mathcal{S}(\mathbf{x}^*(t')) \Rightarrow \mathbf{n} \in \mathcal{S}(\mathbf{x}^*(t)),$$

where $\mathbf{n}' = \mathbf{n}(t')$, $\mathbf{n} = \mathbf{n}(t)$, $\mathbf{n}' = A(t, t')\mathbf{n}$, $t' \geqq t$.

Moreover, along a piece of optimal trajectory which belongs to a B-degenerated subset, Lemma 3 has the following corollary.

COROLLARY 8.

$$-\mathbf{f}(\mathbf{x}^*, \mathbf{u}^b) \in \mathcal{S} \qquad \forall \mathbf{u}^b \in U.$$

By similar arguments one can prove the following.

THEOREM 8. *If point* $\mathcal{P}(\mathbf{x}^*(t))$ *of optimal trajectory* Γ^* *belongs to an A-degenerated subset and* $\mathcal{P}(\mathbf{x}^*(t''))$ *is an interior point of* Σ, *then* $\mathcal{P}(\mathbf{x}^*(t''))$ *belongs to an A-degenerated subset, whatever* $t'' \leqq t$.

THEOREM 9. *Along a piece of optimal trajectory* Γ^* *which belongs to an A-degenerated subset,*

$$\mathbf{n}'' \in \mathcal{S}(\mathbf{x}^*(t'')) \Rightarrow \mathbf{n} \in \mathcal{S}(\mathbf{x}^*(t)),$$

where $\mathbf{n}'' = \mathbf{n}(t'')$, $\mathbf{n} = \mathbf{n}(t)$, $\mathbf{n} = A(t'', t)\mathbf{n}''$, $t'' \leqq t$.

COROLLARY 9.

$$\mathbf{f}(\mathbf{x}^*, \mathbf{u}^b) \in \mathcal{S} \qquad \forall \mathbf{u}^b \in U.$$

At last from Property 3 it follows that if $B/\mathcal{S}(\mathbf{x}) = \varnothing$, then $L_- \subset \mathcal{S}$; and if $A/\mathcal{S}(\mathbf{x}) = \varnothing$, then $L_+ \subset \mathcal{S}$.

7. The maximum principle. As an example, consider an optimal trajectory Γ^*, no point of which belongs to a repulsive subset of Σ. Let $\mathcal{P}^0 \in E^{n+1}$ and $\mathcal{P}^1 \in E^{n+1}$ be its starting point and its endpoint (namely, the target) at times t_0 and t_1, respectively. $\mathcal{P}^0$ may be regular or nonregular; in any case, $\overline{B/\mathcal{S}(\mathbf{x}^*(t_0))}$ is separable.

Let $\mathcal{I}(\mathbf{x}^*(t_0))$ be a separating hyperplane, and $\overline{R_B{}^0}$ the closed halfspace which contains $\overline{B/\mathcal{S}(\mathbf{x}^*(t_0))}$ (following the notation of §1.6).

From Theorem 1 we know that $\overline{B/\mathcal{S}(\mathbf{x}^*(t))}$ is separable for every t such that $t_0 \leqq t \leqq t_1$, and from remarks (ii), (iii) of §3 we know that

the transform $\mathfrak{I}(\mathbf{x}^*(t))$ of $\mathfrak{I}(\mathbf{x}^*(t_0))$ due to the linear transformation $A(t_0, t)$ is a separating hyperplane of $\overline{B/\mathcal{S}(\mathbf{x}^*(t))}$, and that

$$\overline{B/\mathcal{S}(\mathbf{x}^*(t))} \subset \overline{R_B},$$

where $\overline{R_B}$ is the transform of $\overline{R_B{}^0}$.

Let $\mathbf{n}^0$, $|\mathbf{n}^0| = 1$, be the vector normal to $\mathfrak{I}(\mathbf{x}^*(t_0))$ at point $\mathcal{P}^0$ such that

$$\mathbf{n}^0 \in \overline{R_B{}^0},$$

and let us choose $\mathfrak{A}(t_0) = \lambda^0\mathbf{n}^0$, $\lambda^0 > 0$.

From remark (ii) of §2, it follows that $\mathfrak{A}(t)$ is normal to $\mathfrak{I}(\mathbf{x}^*(t))$ and

$$\mathfrak{A}(t) \in \overline{R_B}.$$

At last, from Lemma 3,

$$-\mathbf{f}(\mathbf{x}^*(t), \mathbf{u}) \in \overline{B/\mathcal{S}(\mathbf{x}^*(t))} \qquad \forall \mathbf{u} \in U,$$

say,

$$-\mathbf{f}(\mathbf{x}^*(t), \mathbf{u}) \in \overline{R_B} \qquad \forall \mathbf{u} \in U.$$

Hence,

$$\mathfrak{A}(t)\cdot\mathbf{f}(\mathbf{x}^*(t), \mathbf{u}) \leqq 0 \qquad \forall \mathbf{u} \in U. \tag{6}$$

A more complete discussion would lead to the further conditions that

$$\mathfrak{A}(t)\cdot\mathbf{f}(\mathbf{x}^*(t), \mathbf{u}^*(t)) = 0 \qquad \forall t \in [t_0, t_1], \tag{7}$$

and

$$\lambda_0(t) = \text{const.} \leqq 0. \tag{8}$$

Conditions (6)–(8) embody the maximum principle of Pontryagin for the case of optimal trajectories along regular or nonregular attractive subsets of Σ.

The same conclusions hold for pieces of optimal trajectories along nonregular repulsive subsets of Σ. They can be easily obtained by similar arguments.

8. Trivial maximum principle. From the analysis of the degenerated case it follows that if

(i) the starting point $\mathcal{P}(\mathbf{x}^\circ)$, $\mathbf{x}^\circ = \mathbf{x}^*(t_0)$, of optimal trajectory Γ^* is a B-degenerated point,

(ii) $\mathcal{S}(\mathbf{x}^\circ) \subseteq T(\mathbf{x}^\circ)$, where $T(\mathbf{x}^\circ)$ is an n-dimensional hyperplane through $\mathbf{x}^\circ$, and

(iii) all the points of Γ^* are interior points of Σ,

then there exists a nonzero continuous vector function $\mathfrak{A}(t)$ which is a solu-

tion of adjoint equation (5), such that

$$\lambda_0(t) = 0,$$

$$\boldsymbol{\lambda}(t)\cdot\mathbf{f}(\mathbf{x}^*(t), \mathbf{u}) = 0 \qquad \forall \mathbf{u} \in U,$$

for all t on $[t_0, t_1]$ along optimal trajectories (or pieces of optimal trajectories) which belong to degenerated subsets of Σ.

REFERENCES

[1] A. Blaquière and G. Leitmann, *On the geometry of optimal processes—Part* I, Rpt. AM 64-10, I.E.R., University of California, National Science Foundation Grant GP-803, 1964.

[2] G. Leitmann, *Some geometrical aspects of optimal processes*, this Journal, 3 (1965), pp. 53–65.

[3] H. Halkin, *The principle of optimal evolution*, Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963.

[4] E. Roxin, *A geometric interpretation of Pontryagin's maximum principle*, Ibid.

[5] A. Blaquière, *Sur la théorie de la commande optimale*, Course at the Faculty of Sciences, Institute Henri Poincaré, University of Paris, 1963.

[6] L. S. Pontryagin et al., *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[7] R. E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[8] A. Blaquière and G. Leitmann, *On the geometry of optimal processes—Part II*, Rpt. AM 65-11, I.E.R., University of California, Berkeley, 1965.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

A NEW ALGORITHM FOR A CLASS OF QUADRATIC PROGRAMMING PROBLEMS WITH APPLICATION TO CONTROL*

M. D. CANON† AND JAMES H. EATON‡

Abstract. The control problem considered is that of determining an input which will take a linear sampled system from a specified initial state to a desired terminal state in minimum time, subject to amplitude constraints on the input. The problem is reduced to solving a sequence of simple quadratic programming problems; a new algorithm is presented for solving this class of problems. Preliminary computational results for a fourth-order system are favorable.

1. Introduction. Numerous techniques for obtaining solutions to the time optimal control problem for linear discrete systems are now available in the literature [1], [2], [3]. The practicality of many of these techniques is severely limited by the computational time required which precludes their employment in a feedback mode. The primary justification for further consideration of this problem lies in reducing the computation time required.

In this paper the control problem is reduced to a specific quadratic programming problem (QPP) and a new algorithm is then developed for solving this class of problems. For a fixed number of sampling periods, a quadratic function of the controls is minimized, subject to the constraint that the control sequence takes the sampled system from a given initial state to the desired target state. It is shown that if a solution to this QPP exists, the solution can be written in closed form. As a result, a canonical representation is obtained for terminal states which can be reached in a fixed number of sampling periods. The time optimal control problem is then solved by finding the smallest number of sampling periods for which the target state can be expressed in the canonical form.

There are physical reasons for using a quadratic cost function of the controls in solving the minimum time problem. For a large class of systems this cost function is a measure of the amount of energy supplied by the controller. Since the time optimal control problem does not, in general, have a unique solution, it is desirable to find that time optimal control

* Received by the editors July 13, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Department of Electrical Engineering, University of California, Berkeley, California. This work was partially supported by the National Aeronautics and Space Administration under Grant No. NSG 354-(S-2).

‡ IBM Research Laboratory, International Business Machines Corporation, San Jose, California.

which, among all time optimal controls, also minimizes the energy supplied to the system. In addition, when considering feedback implementation, it may be desirable to obtain a solution for which the unsaturated controls tend to appear at the end of the control sequence. This can be done by a proper choice of the quadratic cost function.

2. Description of the system. We shall consider an nth order, time invariant, linear discrete system χ whose state $\mathbf{x}_N \in E^n$ at time N is given by

$$\mathbf{x}_N(u_1, u_2, \cdots, u_N) = \sum_{i=1}^{N} \mathbf{r}_i u_i, \tag{1a}$$

where $\mathbf{r}_i \in E^n$, $i = 1, 2, \cdots$, are known constant vectors and the scalar control variables u_i are constrained in magnitude by $-1 \leqq u_i \leqq 1$, $i = 1, 2, \cdots$.[1] For a given input sequence $u_1, \cdots, u_N$, it is assumed that the energy supplied to the system χ is given by

$$J(\mathbf{u}) = \frac{1}{2}\sum_{i=1}^{N} u_i^2. \tag{1b}$$

Note that no difficulties are introduced by considering $J(\mathbf{u}) = \sum_{i=1}^{N} \lambda_i u_i^2$, $\lambda_i > 0$; however we choose (1b) for notational simplification.

DEFINITION. A control sequence $(u_1, \cdots, u_N) = \mathbf{u}$ of length N is said to belong to the *constraint set* Ω_N if $|u_i| \leqq 1$, $i = 1, \cdots, N$. If $\mathbf{u} \in \Omega_N$, then $\mathbf{u}$ is called an *admissible control.*

Let $\mathbf{v}_N \in E^n$, $N = 1, 2, \cdots$, represent a moving target $\mathcal{V}$ at time N. The system χ *intercepts* the target $\mathcal{V}$ at time N, if there exists an admissible control sequence $u_1, \cdots, u_N$ for χ, such that $\mathbf{x}_N(\mathbf{u}) = \mathbf{v}_N$. Roughly speaking, the problem is to find an admissible control for χ which minimizes the intercept time, N. If there is more than one admissible control sequence of minimum length N satisfying the relation $\mathbf{x}_N(\mathbf{u}) = \mathbf{v}_N$, then choose that control sequence which minimizes the energy delivered to χ, i.e., minimizes $J(\mathbf{u})$.

DEFINITION. Let $\mathcal{R}_N$ be the set of states of χ which can be reached at time N with an admissible control sequence, i.e.,

$$\mathcal{R}_N = \left\{ \mathbf{x}_N : \quad \mathbf{x}_N = \sum_{i=1}^{N} \mathbf{r}_i u_i, \mathbf{u} \in \Omega_N \right\}. \tag{2}$$

Observe that $\mathcal{R}_N$ is a compact convex set since it is the image of the compact convex set Ω_N under a continuous linear map. Clearly, the target can be intercepted in time N if and only if $\mathbf{v}_N \in \mathcal{R}_N$. We shall assume that $\mathbf{v}_N \in \mathcal{R}_N$ for some finite N.

[1] The formulation is easily extended to the multiple input case.

3. Statement of the problem. For the system χ and the target $\mathcal{U}$, find the smallest integer N with $\mathbf{v}_N \in \mathcal{R}_N$ and an input sequence $\mathbf{u}^0 \in \Omega_N$ which minimizes $J(\mathbf{u})$, subject to the constraints $\mathbf{v}_N = \mathbf{x}_N(u)$ and $\mathbf{u} \in \Omega_N$.

4. A canonical representation of points in $\mathcal{R}_N$. The principal result in this section is stated in the following theorem.

THEOREM 1. *Each point $\mathbf{x}_N \in \mathcal{R}_N$ can be represented in the form*[2]

$$\mathbf{x}_N = \sum_{i=1}^{N} \mathbf{r}_i \operatorname{sat} \langle \mathbf{r}_i , \mathbf{c} \rangle, \tag{3}$$

for some vector $\mathbf{c} \in E^n$. Furthermore, the control sequence $\operatorname{sat} \langle \mathbf{r}_i , \mathbf{c} \rangle$, $i = 1, 2, \cdots, N$, *satisfying* (3) *is the solution to the quadratic programming problem* (QPP): *Find N real variables $u_1^0, u_2^0, \cdots, u_N^0$, which minimize*

$$J(\mathbf{u}) = \frac{1}{2} \sum_{i=1}^{N} u_i^2 \tag{4}$$

subject to the $n + N$ constraints

$$\sum_{i=1}^{N} \mathbf{r}_i u_i = \mathbf{x}_N , \tag{5}$$

$$\mathbf{u} \in \Omega_N . \tag{6}$$

Remark. $J(\mathbf{u})$ is a continuous, strictly convex function and the constraints (5) and (6) form a nonvoid (since $\mathbf{x}_N \in \mathcal{R}_N$) compact support for J. Consequently, a solution to the QPP exists and is unique.

Proof of the theorem. Define the function $\mathcal{H}(\mathbf{u}, \mathbf{c})$ by

$$\begin{aligned} \mathcal{H}(\mathbf{u}, \mathbf{c}) &= -\langle \mathbf{x}_N , \mathbf{c} \rangle + J(\mathbf{u}) \\ &= \sum_{i=1}^{N} [-\langle \mathbf{r}_i , \mathbf{c} \rangle u_i + \tfrac{1}{2} u_i^2]. \end{aligned} \tag{7}$$

It has been shown by Canon [4] that a necessary and sufficient condition for $\mathbf{u}^0$ to be a solution of the QPP is that for some $\mathbf{c} \in E^n$,

$$\mathcal{H}(\mathbf{u}^0, \mathbf{c}) = \min_{\mathbf{u} \in \Omega_N} \mathcal{H}(\mathbf{u}, \mathbf{c}). \tag{8}$$

Using (7) and (8) it follows directly that $u_i^0 = \operatorname{sat} \langle \mathbf{r}_i , \mathbf{c} \rangle$, $i = 1, 2, \cdots, N$, is the solution to the QPP. Since a solution to the QPP exists for every $\mathbf{x}_N \in \mathcal{R}_N$ the proof of the theorem is complete.

To summarize, the moving target problem can now be stated as follows: Find the smallest integer N and a vector $\mathbf{c} \in E^n$ such that

$$\mathbf{v}_N = \sum_{i=1}^{N} \mathbf{r}_i \operatorname{sat} \langle \mathbf{r}_i , \mathbf{c} \rangle \triangleq \mathbf{f}_N(\mathbf{c}). \tag{9}$$

[2] $\operatorname{sat}(y) = y$ if $| y | \leqq 1$, $\operatorname{sat}(y) = y/| y |$ if $| y | > 1$.

For each integer N, $\mathbf{f}_N$ is a vector function which maps E^n onto $\mathcal{R}_N$. However, the map is not one-to-one; hence, the inverse function is not defined. We will now show that it is possible to restrict the domain of $\mathbf{f}_N$ to a subset of E^n in such a manner as to make $\mathbf{f}_N$ a bijective bicontinuous function. Since our ultimate goal is to find an algorithm for determining if $\mathbf{v}_N \in \mathcal{R}_N$, i.e., if there exists a $\mathbf{c}$ such that $\mathbf{v}_N = \mathbf{f}_N(\mathbf{c})$, the continuity of the inverse function is of major importance.

5. The vector function $\mathbf{f}_N$. For each $\mathbf{c} \in E^n$, let $I_N(\mathbf{c}) \subset \{1, 2, \cdots, N\}$ be an index set such that if $i \in I_N(\mathbf{c})$, then $|\langle \mathbf{r}_i, \mathbf{c}\rangle| > 1$, and let $\bar{I}_N(\mathbf{c})$ be the complement of this set relative to $\{1, 2, \cdots, N\}$. Using this notation, $\mathbf{f}_N(\mathbf{c})$ can be written as

$$\mathbf{f}_N(\mathbf{c}) = \sum_{i \in I_N(\mathbf{c})} \mathbf{r}_i \operatorname{sat} \langle \mathbf{r}_i, \mathbf{c}\rangle + \sum_{i \in \bar{I}_N(\mathbf{c})} \mathbf{r}_i \langle \mathbf{r}_i, \mathbf{c}\rangle. \tag{10}$$

Whenever $\mathbf{f}_N$ and I_N occur together, we shall drop the subscript N on I_N in order to simplify notation.

DEFINITION. For each integer N, let $\mathcal{C}_N$ be the set of points $\mathbf{c} \in E^n$ for which the vectors $\{\mathbf{r}_i : i \in \bar{I}_N(\mathbf{c})\}$ span E^n.

Assumption 1.[3] The vectors $\mathbf{r}_i$, $i = 1, 2, \cdots, n$, span E^n.

In the remaining lemmas of this section we will prove that the map $\mathbf{f}_N : \mathcal{C}_N \to \mathcal{R}_N$, $N \geqq n$, is bicontinuous, one-to-one, and onto. In the statement and proof of the lemmas it will be assumed that $N \geqq n$.

LEMMA 1. *The map $\mathbf{f}_N : \mathcal{C}_N \to \mathcal{R}_N$ is continuous and onto.*

Proof. The continuity of $\mathbf{f}_N$ is obvious. Let $\mathbf{x}_N \in \mathcal{R}_N$ be arbitrary. Using Theorem 1, there exists some vector $\mathbf{c}^0 \in E^n$ such that

$$\mathbf{x}_N = \sum_{i \in I(\mathbf{c}^0)} \mathbf{r}_i \operatorname{sat} \langle \mathbf{r}_i, \mathbf{c}^0\rangle + \sum_{i \in \bar{I}(\mathbf{c}^0)} \mathbf{r}_i \langle \mathbf{r}_i, \mathbf{c}^0\rangle. \tag{11}$$

Suppose the vectors $\{\mathbf{r}_i : i \in \bar{I}(\mathbf{c}^0)\}$ do not span E^n, i.e., $\mathbf{c}^0 \notin \mathcal{C}_N$. Then there is a unit vector $\mathbf{c}' \in E^n$ which is orthogonal to $\mathbf{r}_i$, $i \in \bar{I}(\mathbf{c}^0)$. Since the $\mathbf{r}_i$, $i = 1, 2, \cdots, n$, span E^n there is at least one vector, say $\mathbf{r}_k$, $k \in I(\mathbf{c}^0)$, which is not orthogonal to $\mathbf{c}'$. Hence, there is a scalar σ such that $|\langle \mathbf{r}_k, \mathbf{c}^0 + \sigma\mathbf{c}'\rangle| = 1$ and therefore $k \in \bar{I}(\mathbf{c}^0 + \sigma\mathbf{c}')$. If the $\{\mathbf{r}_i : i \in \bar{I}(\mathbf{c}^0 + \sigma\mathbf{c}')\}$ do not span E^n the process can be repeated, thus we can construct a $\mathbf{c} \in \mathcal{C}_N$ such that $\mathbf{f}_N(\mathbf{c}) = \mathbf{f}_N(\mathbf{c}^0)$.

LEMMA 2. *The map $\mathbf{f}_n : \mathcal{C}_N \to \mathcal{R}_N$ is one-to-one.*

Proof (By contradiction). Suppose there are two distinct points $\mathbf{c}^0$ and $\mathbf{c}'$ in $\mathcal{C}_N$ with

$$\sum_{i=1}^{N} \mathbf{r}_i \operatorname{sat} \langle \mathbf{r}_i, \mathbf{c}^0\rangle = \sum_{i=1}^{N} \mathbf{r}_i \operatorname{sat} \langle \mathbf{r}_i, \mathbf{c}'\rangle. \tag{12}$$

[3] Any sample data system which is controllable will satisfy this assumption.

Since the solution of the QPP is unique, the condition expressed in (12) implies

$$\text{sat}\,\langle \mathbf{r}_i , \mathbf{c}^0\rangle = \text{sat}\,\langle \mathbf{r}_i , \mathbf{c}'\rangle, \qquad i = 1, 2, \cdots, N, \tag{13}$$

and in particular this means $\langle \mathbf{r}_i , \mathbf{c}^0\rangle = \text{sat}\,\langle \mathbf{r}_i , \mathbf{c}'\rangle$ for all $i \in \bar{I}(\mathbf{c}^0)$, thus $|\langle \mathbf{r}_i , \mathbf{c}^0\rangle| \leqq |\langle \mathbf{r}_i , \mathbf{c}'\rangle|$ for all $i \in \bar{I}(\mathbf{c}^0)$. By hypothesis, the vectors $\{\mathbf{r}_i : i \in \bar{I}(\mathbf{c}^0)\}$ span E^n and $\mathbf{c}^0 \neq \mathbf{c}'$. Consequently there is at least one index $k \in \bar{I}(\mathbf{c}^0)$ with $|\langle \mathbf{r}_k , \mathbf{c}^0\rangle| < |\langle \mathbf{r}_k , \mathbf{c}'\rangle|$; therefore $\| \mathbf{c}^0 \| < \| \mathbf{c}' \|$. But, using the same argument, replacing $\bar{I}(\mathbf{c}^0)$ by $\bar{I}(\mathbf{c}')$ it can be shown that $\| \mathbf{c}' \| < \| \mathbf{c}^0 \|$. Therefore, we must conclude that for each $\mathbf{x}_N \in \mathcal{R}_N$ there is one and only one $\mathbf{c} \in \mathcal{C}_N$ with $\mathbf{f}_N(\mathbf{c}) = \mathbf{x}_N$.

Thus far it has been established that the map $\mathbf{f}_N : \mathcal{C}_N \to \mathcal{R}_N$ is continuous, one-to-one, and onto. To prove that $\mathbf{f}_N^{-1}$ is continuous we need one further lemma.

LEMMA 3. *The set $\mathcal{C}_N$ is compact.*

Proof. We show first that $\mathcal{C}_N$ is bounded. For each $\mathbf{c} \in \mathcal{C}_N$, $N \geqq n$, the $n \times n$ matrix[4]

$$\sum_{i \in \bar{I}(\mathbf{c})} \mathbf{r}_i\!>\!<\!\mathbf{r}_i \tag{14}$$

is symmetric and *positive definite*. Let $\lambda_{\min}(\mathbf{c})$ be the smallest eigenvalue of the matrix in (14). Then we have the inequality

$$0 < \| \mathbf{c} \|^2 \lambda_{\min}(\mathbf{c}) \leqq \langle \mathbf{c}, \Big(\sum_{i \in \bar{I}(\mathbf{c})} \mathbf{r}_i\!>\!<\!\mathbf{r}_i\Big)\mathbf{c}\rangle \tag{15}$$

for every $\mathbf{c}$ in $\mathcal{C}_N$ with $\| \mathbf{c} \| \neq 0$. There can be at most a finite number of distinct scalars $\lambda_{\min}(\mathbf{c})$, $\mathbf{c} \in \mathcal{C}_N$. Hence let

$$\lambda^* = \min \{\lambda_{\min}(\mathbf{c}) : \mathbf{c} \in \mathcal{C}_N\}.$$

Using (15) and recalling that if $i \in \bar{I}(\mathbf{c})$ then $|\langle \mathbf{r}_i , \mathbf{c}\rangle| \leqq 1$, we have

$$\lambda^* \| \mathbf{c} \|^2 \leqq \lambda_{\min}(\mathbf{c}) \| \mathbf{c} \|^2 \leqq \langle \mathbf{c}, \Big(\sum_{i \in \bar{I}(\mathbf{c})} \mathbf{r}_i\!>\!<\!\mathbf{r}_i\Big)\mathbf{c}\rangle$$

$$= \sum_{i \in \bar{I}(\mathbf{c})} \langle \mathbf{r}_i , \mathbf{c}\rangle^2 \leqq N. \tag{16}$$

Therefore, $\| \mathbf{c} \|^2 \leqq N/\lambda^*$ for each $\mathbf{c} \in \mathcal{C}_N$.

To prove that $\mathcal{C}_N$ is closed it is sufficient to show that for any arbitrary $\mathbf{c}^0 \notin \mathcal{C}_N$, there is an open set $V_{\mathbf{c}^0}$, containing $\mathbf{c}^0$, which is disjoint from $\mathcal{C}_N$. If $\mathbf{c}^0 \notin \mathcal{C}_N$ the index set $I(\mathbf{c}^0)$ is nonvoid. Choose $\epsilon > 0$ such that $|\langle \mathbf{r}_i , \mathbf{c}^0\rangle| > 1 + \epsilon$ for each $i \in I(\mathbf{c}^0)$ and let $\pi = \max \{\| \mathbf{r}_i \| : i \in I(\mathbf{c}^0)\}$. The set $V = \{\mathbf{c} : \| \mathbf{c} \| < \epsilon/\pi\}$ is an open neighborhood of the origin. Hence, $V_{\mathbf{c}^0} = \{\mathbf{c}^0 + \mathbf{c} : \mathbf{c} \in V\}$ is an open neighborhood of $\mathbf{c}^0$. If $\mathbf{c}_0 + \mathbf{c}$ is an

[4] Here $<\!\mathbf{r}_i$ denotes a row vector and $\mathbf{r}_i\!>$ a column vector, thus $\mathbf{r}_i\!>\!<\!\mathbf{r}_i$ is an $n \times n$ square matrix.

arbitrary element of $V_{\mathbf{c}^0}$ then for each $i \in I(\mathbf{c}^0)$ we have

$$|\langle \mathbf{r}_i , \mathbf{c}^0 + \mathbf{c}\rangle| \geqq |\langle \mathbf{r}_i , \mathbf{c}^0\rangle| - |\langle \mathbf{r}_i , \mathbf{c}\rangle|$$

$$\geqq |\langle \mathbf{r}_i , \mathbf{c}^0\rangle| - \| \mathbf{r}_i \| \cdot \| \mathbf{c} \| > (1 + \epsilon) - \epsilon = 1.$$

Therefore, if $i \in I(\mathbf{c}^0)$, then $i \in I(\mathbf{c}^0 + \mathbf{c})$; hence $I(\mathbf{c}^0) \subset I(\mathbf{c}^0 + \mathbf{c})$ and consequently $\bar{I}(\mathbf{c}^0) \supset \bar{I}(\mathbf{c}^0 + \mathbf{c})$. This means, since by hypothesis the vectors $\{\mathbf{r}_i : i \in \bar{I}(\mathbf{c}^0)\}$ do not span E^n, that the vectors $\{\mathbf{r}_i : i \in \bar{I}(\mathbf{c}^0 + \mathbf{c})\}$ do not span E^n and $\mathbf{c}^0 + \mathbf{c} \notin \mathcal{C}_N$. Thus we have the desired result, that $V_{\mathbf{c}^0} \cap \mathcal{C}_N$ is empty.

The following theorem states the main result of this section.

THEOREM 2. *For each finite integer N, $N \geqq n$, the function $\mathbf{f}_N$ with domain $\mathcal{C}_N$ and range $\mathcal{R}_N$ is a homeomorphism.*

Proof. From Lemmas 1, 2, and 3, $\mathbf{f}_N$ is a continuous bijection from the compact set $\mathcal{C}_N$ onto $\mathcal{R}_N$. It remains for us to show that $\mathbf{f}_N^{-1}$ is continuous. To do this, it is sufficient to show that the image of any closed subset of $\mathcal{C}_N$ under $\mathbf{f}_N$ is closed. Let A be a closed subset in $\mathcal{C}_N$. Then, since $\mathcal{C}_N$ is compact, A is compact and hence $\mathbf{f}_N(A)$ is compact since $\mathbf{f}_N$ is continuous; therefore, $\mathbf{f}_N(A)$ is closed.

6. The algorithm. It will be assumed in this section that N is fixed, that $N \geqq n$, and that $\mathbf{v}_N$ is given. We shall develop an algorithm which, if $\mathbf{v}_N \in \mathcal{R}_N$, can be used to solve the equation $\mathbf{v}_N = \mathbf{f}_N(c)$ for $\mathbf{c} \in \mathcal{C}_N$, and we shall prove that the algorithm converges in a finite number of steps. If $\mathbf{v}_N \notin \mathcal{R}_N$, then it will be shown that the algorithm terminates in a finite number of steps. A brief description of the algorithm follows: We choose some $\mathbf{c}^0 \in \mathcal{C}_N$ as an initial estimate of $\mathbf{c}$ and let $\mathbf{v}^0$ be the corresponding point in $\mathcal{R}_N$, i.e.,

$$\mathbf{v}^0 = \sum_{i \in I(\mathbf{c}^0)} \mathbf{r}_i \text{ sat } \langle \mathbf{r}_i , \mathbf{c}^0\rangle + \sum_{i \in \bar{I}(\mathbf{c}^0)} \mathbf{r}_i \langle \mathbf{r}_i , \mathbf{c}^0\rangle. \tag{17}$$

Let the error $\mathbf{v}_N - \mathbf{v}^0$ be denoted by $\mathbf{e}^0$. We next construct a vector $\mathbf{c}'$ and a scalar σ' such that for each $0 < \sigma \leqq \sigma'$, $(\mathbf{c}^0 + \sigma\mathbf{c}') \in \mathcal{C}_N$ and $\mathbf{v}' = \mathbf{v}^0 + \sigma\mathbf{e}^0 = \mathbf{f}_N(\mathbf{c}^0 + \sigma\mathbf{c}')$. Thus, as σ increases the point $\mathbf{v}'$ moves directly along the error vector $\mathbf{e}^0$. Note that if $\sigma' \geqq 1$ we may set $\sigma = 1$ and a solution has been obtained. If $0 < \sigma' < 1$, then the new error is $(1 - \sigma')\mathbf{e}^0$; and the procedure is then repeated.

We now give the algorithm in full. First we will show how to construct $\sigma'\mathbf{c}'$ by determining the relationship between changes in $\mathbf{v}^0$ and $\mathbf{c}^0$ in a small neighborhood of $\mathbf{v}^0$. If we assume that $\mathbf{v}_N \in \mathcal{R}_N$ then, since $\mathbf{v}^0 \in \mathcal{R}_N$ and $\mathcal{R}_N$ is convex, $\mathbf{v}^0 + \sigma'\mathbf{e}^0 \in \mathcal{R}_N$ for all $0 \leqq \sigma' \leqq 1$. Using Theorem 2, there is some vector $[\mathbf{c}^0 + \mathbf{c}(\sigma')] \in \mathcal{C}_N$ for which

$$\mathbf{v}^0 + \sigma'\mathbf{e}^0 = \sum_{i\in I[\mathbf{c}^0+\mathbf{c}(\sigma')]} \mathbf{r}_i \text{ sat } \langle \mathbf{r}_i, \mathbf{c}^0 + \mathbf{c}(\sigma')\rangle + \sum_{i\in \bar{I}[\mathbf{c}^0+\mathbf{c}(\sigma')]} \mathbf{r}_i\langle \mathbf{r}_i, \mathbf{c}^0 + \mathbf{c}(\sigma')\rangle. \tag{18}$$

Clearly, as σ' tends to zero, $\mathbf{v}^0 + \sigma'\mathbf{e}^0$ tends to $\mathbf{v}^0$; therefore, since $\mathbf{f}_N$ is a homeomorphism, $\mathbf{c}^0 + \mathbf{c}(\sigma')$ tends to $\mathbf{c}^0$. Referring to (17) and (18), σ' can be chosen sufficiently small (but positive) such that (i) each of the saturated terms in (17) is also saturated in (18), with sat $\langle \mathbf{r}_i, \mathbf{c}^0\rangle$ = sat $\langle \mathbf{r}_i, \mathbf{c}^0 + \mathbf{c}(\sigma')\rangle$, and (ii) if $|\langle \mathbf{r}_i, \mathbf{c}^0\rangle| < 1$, then $|\langle \mathbf{r}_i, \mathbf{c}^0 + \mathbf{c}(\sigma')\rangle| \leqq 1$. The only terms which have not been accounted for are those on the boundary, i.e., the terms in (17) for which $|\langle \mathbf{r}_k, \mathbf{c}^0\rangle| = 1$. There are two possibilities for the terms with index k in (18), either $|\langle \mathbf{r}_k, \mathbf{c}^0 + \mathbf{c}(\sigma')\rangle| \leqq 1$, or $|\langle \mathbf{r}_k, \mathbf{c}^0 + \mathbf{c}(\sigma')\rangle| > 1$. In the second case, we may choose σ' such that sat $\langle \mathbf{r}_k, \mathbf{c}^0 + \mathbf{c}(\sigma')\rangle = \langle \mathbf{r}_k, \mathbf{c}^0\rangle$. Roughly speaking then, we choose σ' such that $\bar{I}[\mathbf{c}^0 + \mathbf{c}(\sigma')] \subset \bar{I}(\mathbf{c}^0)$, where the "difference" between the index sets is accounted for by the terms on the boundary in (17) which, now, in (18) are saturated. Subtracting (17) from (18) we obtain

$$\sigma'\mathbf{e}^0 = \Big(\sum_{i\in \bar{I}[\mathbf{c}^0+\mathbf{c}(\sigma')]} \mathbf{r}_i\mathbf{>\!<}\mathbf{r}_i\Big)\mathbf{c}(\sigma'). \tag{19}$$

Since the vectors $\{\mathbf{r}_i : i \in \bar{I}(\mathbf{c}^0 + \mathbf{c}(\sigma'))\}$ span E^n, the $n \times n$ matrix in (19) is nonsingular. Thus,

$$\mathbf{c}(\sigma') = \sigma'\Big(\sum_{i\in \bar{I}[\mathbf{c}^0+\mathbf{c}(\sigma')]} \mathbf{r}_i\mathbf{>\!<}\mathbf{r}_i\Big)^{-1}\mathbf{e}^0. \tag{20}$$

Note that $\mathbf{c}(\sigma')$ depends linearly on σ', and for this reason it is easy to see that the constraints which were previously imposed for choosing σ' are satisfied for any scalar $\sigma \in (0, \sigma']$. Consequently, $\mathbf{v}^0 + \sigma\mathbf{e}^0 = \mathbf{f}_N(\mathbf{c}^0 + \mathbf{c}(\sigma))$ for all $0 < \sigma \leqq \sigma'$. To be consistent with the notation used at the beginning of this section, let $\mathbf{c}' = \mathbf{c}(\sigma)/\sigma$. We then obtain

$$\mathbf{v}^0 + \sigma\mathbf{e}^0 = \mathbf{f}_N(\mathbf{c}^0 + \sigma\mathbf{c}'), \qquad 0 < \sigma \leqq \sigma', \tag{21}$$

where $\mathbf{c}^0 + \sigma\mathbf{c}' \in \mathcal{C}_N$ with

$$\mathbf{c}' = \Big(\sum_{i\in \bar{I}(\mathbf{c}^0+\sigma\mathbf{c}')} \mathbf{r}_i\mathbf{>\!<}\mathbf{r}_i\Big)^{-1}\mathbf{e}^0. \tag{22}$$

Henceforth, we shall denote $\bar{I}(\mathbf{c}^0 + \sigma\mathbf{c}')$ by $\bar{S}$ and refer to the set $\{\mathbf{r}_i : i \in \bar{S}\}$, or more simply to $\bar{S}$, as the basis; S denotes the complement of $\bar{S}$ relative to $\{1, 2, \cdots, N\}$. To reiterate, at each step of the algorithm we must determine the basis $\bar{S}$, solve for $\mathbf{c}'$ in (22) and select σ'. These quantities are determined in accordance with the constraints previously imposed. The constraints are restated below in more compact form.

P1. *The index set* $\bar{S}$: A necessary and sufficient condition for index k to belong to $\bar{S}$ is that $\mathbf{r}_k$ satisfy either

(a) $|\langle \mathbf{r}_k , \mathbf{c}^0 \rangle| < 1$
or
(b) $|\langle \mathbf{r}_k , \mathbf{c}^0 \rangle| = 1$ and $\operatorname{sgn} \langle \mathbf{r}_k , \mathbf{c}^0 \rangle = -\operatorname{sgn} \langle \mathbf{r}_k , \mathbf{c}' \rangle$, where $\mathbf{c}'$ is obtained from (22).

P2. *The scalar* σ': σ' is the largest value of σ, $0 < \sigma \leqq 1$, such that
(a) $|\langle \mathbf{r}_k , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle| \leqq 1$ for $k \in \tilde{S}$,
(b) $|\langle \mathbf{r}_k , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle| \geqq 1$ for $k \in S$.
In other words, once the basis has been chosen, none of the controls are allowed to cross the boundary.

Remark. Computationally, there is no problem determining which of the terms on the boundary at the end of step m are to remain in the basis at step $m + 1$. For, as the algorithm progresses, one is able to deduce from the previous step the proper choice. For example, if at the start of step m, $|u_k| < 1$ and at the end of the step $|u_k| = 1$, the vector $\mathbf{r}_k$ is removed from the basis at step $m + 1$, if this is the only control on the boundary. If more than one control is on the boundary at the end of step $m + 1$ (which computationally is unlikely), then one must guess whether the vector $\mathbf{r}_k$ remains in the basis. After computing $\mathbf{c}'$, one then checks the second condition in P1. Several examples are worked out in the Appendix.

THEOREM 3. *If* $\mathbf{v}_N \in \mathcal{R}_N$, *with* $\mathbf{v}_N = \mathbf{f}_N(\mathbf{c})$, *then the algorithm can be used to determine* $\mathbf{c}$ *in a finite number of steps.*

Proof. Let $\mathbf{c}^0 \in \mathcal{C}_N$ be the initial estimate of $\mathbf{c}$, and let $\mathbf{e}^0 = \mathbf{v}_N - \mathbf{v}^0$ denote the error. Since $\mathbf{v}_N \in \mathcal{R}_N$ we can use P1 and P2 to determine $\sigma' \mathbf{c}'$ at each step in the algorithm. Thus if $\mathbf{c}^m \in \mathcal{C}_N$ is the estimate of $\mathbf{c}$ at step m, the error e^m at step m is

$$\mathbf{e}^m = \prod_{i=1}^{m} (1 - \sigma^i)\mathbf{e}^0 = \mathbf{f}_N(\mathbf{c}) - \mathbf{f}_N(\mathbf{c}^m), \tag{23}$$

where σ^i, $0 < \sigma^i \leqq 1$, $i = 1, 2, \cdots, m$, is that scalar determined by P2 at step i. As m tends to infinity, $\prod_{i=1}^{m} (1 - \sigma^i)$ converges to some scalar, say σ^*. Consequently, the algorithm converges to some $\mathbf{c}^*$. Without loss of generality assume that $\mathbf{c} = \mathbf{c}^*$ and $\sigma^* = 0$. We now show that the algorithm converges to $\mathbf{c}$ in a finite number of steps. Since both $\mathbf{f}_N(\mathbf{c})$ and $\mathbf{f}_N(\mathbf{c}^0)$ belong to $\mathcal{R}_N$, we can use P1 and P2 to find a scalar β', $0 < \beta' \leqq 1$, and a vector $\mathbf{c}'$ such that for each $\beta \in (0, \beta']$ it is true that $\mathbf{c} + \beta \mathbf{c}' \in \mathcal{C}_N$ and

$$\mathbf{v}_N - \beta \mathbf{e}^0 = \mathbf{f}_N(\mathbf{c} + \beta \mathbf{c}'). \tag{24}$$

Choose an integer m such that $0 < \prod_{i=1}^{m} (1 - \sigma^i) \leqq \beta'$; then setting $\beta = \prod_{i=1}^{m} (1 - \sigma^i)$ and using (23) and (24) we have

$$\begin{aligned} \mathbf{f}_N(\mathbf{c} + \beta \mathbf{c}') &= \mathbf{v}_N - \mathbf{e}^m = \mathbf{v}_N - (\mathbf{v}_N - \mathbf{f}_N(\mathbf{c}^m)) \\ &= \mathbf{f}_N(\mathbf{c}^m). \end{aligned} \tag{25}$$

This implies $\mathbf{c}^m = \mathbf{c} + \beta \mathbf{c}'$. Now, since (24) holds for all $0 < \beta \leqq \beta'$, the

error at step $m+1$ can be reduced only if $-\beta\mathbf{c}'$ is that vector obtained at step $m+1$ using P1. Furthermore, the maximum reduction in the error at step $m+1$ is obtained in accordance with P2 by setting $\sigma^{m+1}=1$. Consequently, at step $m+1$ we have $\mathbf{c}^{m+1}=\mathbf{c}^m-\beta\mathbf{c}'=\mathbf{c}$, hence $\mathbf{f}_N(\mathbf{c}^{m+1})=\mathbf{v}_N$.

COROLLARY. *If* $\mathbf{v}_N \notin \mathcal{R}_N$, *then the algorithm terminates in a finite number of steps.*

Proof. Let ξ, $0<\xi<1$, be the largest scalar for which $\xi\mathbf{v}_N \in \mathcal{R}_N$. By Theorem 3, the algorithm converges to $\xi\mathbf{v}_N$ in a finite number of steps, say m. At step $m+1$ it will be impossible to construct $\sigma'\mathbf{c}'$ in the prescribed manner; the algorithm then terminates.

The following lemma will be useful in determining when the computation should be terminated.

LEMMA 4. *Let* $\mathbf{c}$ *be a vector in* $\mathcal{C}_N$ *with*

$$\mathbf{f}_N(\mathbf{c})=\sum_{i\in M}\mathbf{r}_i \operatorname{sat}\langle\mathbf{r}_i,\mathbf{c}\rangle+\sum_{i\in\bar{M}}\mathbf{r}_i\langle\mathbf{r}_i,\mathbf{c}\rangle,$$

where $M\subset\{1,2,\cdots,n\}$ *is an index set defined as follows:* $i\in M$ *if* $|\langle\mathbf{r}_i,\mathbf{c}\rangle|\geqq 1$; $\bar{M}$ *denotes the complement of this set. Then, if the vectors* $\{\mathbf{r}_i : i\in\bar{M}\}$ *do not span* E^n, $\mathbf{f}_N(\mathbf{c})$ *belongs to the boundary of* $\mathcal{R}_N$.

Proof. For each scalar $\alpha>1$, $\alpha\mathbf{c}$ does not belong to $\mathcal{C}_N$. By hypothesis $\mathbf{c}\in\mathcal{C}_N$, consequently, $\mathbf{c}$ belongs to the boundary of $\mathcal{C}_N$. Thus, $\mathbf{f}_N(\mathbf{c})$ belongs to the boundary of $\mathcal{R}_N$ since $\mathbf{f}_N$ is a homeomorphism.

One possible technique which can be used to solve the time optimal control problem is the following. Using the algorithm developed in this section, suppose that for a fixed number of sampling periods N it has been determined that $\mathbf{v}_N\notin\mathcal{R}_N$. Let $\mathbf{f}_N(\mathbf{c}^m)$ be the point at which the algorithm terminated. We then increase N by one and seek a solution to the equation $\mathbf{v}_{N+1}=\mathbf{f}_{N+1}(\mathbf{c})$. As an initial estimate of $\mathbf{c}$ we can use $\mathbf{c}^m$, since if $\mathbf{c}^m\in\mathcal{C}_N$, then $\mathbf{c}^m\in\mathcal{C}_{N+1}$. Now depending on the magnitude of $\langle\mathbf{r}_{N+1},\mathbf{c}^m\rangle$, the vector $\mathbf{r}_{N+1}$ is placed in or out of the basis. The new error is $\mathbf{v}_{N+1}-\mathbf{f}_N(\mathbf{c}^m)-\operatorname{sat}\langle\mathbf{r}_{N+1},\mathbf{c}^m\rangle$ and the algorithm is initiated at this point. Example 2 in the Appendix illustrates this technique.

7. Conclusions. This paper presents a simple algorithm for solving a class of quadratic programming problems as well as minimum time problems which are reducible to a problem in quadratic programming. To test the computational efficiency of the algorithm, the authors have used it in conjunction with an IBM 7090 computer to solve the following problem: For a simple data system of the fourth order (i.e., the vectors $\mathbf{r}_i\in E^4$), time-optimal controls were computed for targets which could be intercepted in at least 20 sampling periods. The maximum computation time was 0.3 seconds. Results indicate that computation time will be roughly

proportional to the order of the system. In addition, the computations were carried out in single precision and no difficulties were encountered performing the required matrix inversions. Note that when the dimension of the system becomes large it is computationally efficient to use formula for inverting a matrix plus a dyad, where the inverse of the matrix is known.

Appendix. We present here two numerical examples. In Example 1, the number of sampling periods N is fixed and the algorithm is used to solve the equation $\mathbf{v}_N = \mathbf{f}_N(\mathbf{c})$ for $\mathbf{c} \in \mathcal{C}_N$. Example 2 illustrates how the algorithm can be used to solve a minimum time control problem.

Example 1. Let $N = 3$, $n = 2$, and take $\mathbf{r}_1 = \text{col } (0, 1)$, $\mathbf{r}_2 = \text{col } (1, 1)$ and $\mathbf{r}_3 = \text{col } (0, 2)$. Let the target be $\mathbf{v}_3 = \text{col } (1, 4)$. We seek a solution to the equation $\mathbf{v}_3 = \mathbf{f}_3(\mathbf{c})$ and, as an initial estimate $\mathbf{c}^0$ of $\mathbf{c}$, we choose $\mathbf{c}^0 = \text{col } (0, 0)$. The initial error is then $\mathbf{e}^0 = \text{col } (1, 4)$.

(i) The basis $\bar{S} = \{1, 2, 3\}$, and from (22) we have

$$\mathbf{c}' = \Big(\sum_{i \in \bar{S}} \mathbf{r}_i \times \mathbf{r}_i\Big)^{-1} \mathbf{e}^0 = \tfrac{1}{5} \text{col } (2, 3).$$

The scalar σ' is now chosen to be the largest value of σ, $0 < \sigma \leqq 1$, such that

$$|\langle \mathbf{r}_1 , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle| = \sigma |\, \tfrac{3}{5} \,| \leqq 1,$$

$$|\langle \mathbf{r}_2 , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle| = \sigma |\, 1 \,| \leqq 1,$$

$$|\langle \mathbf{r}_3 , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle| = \sigma |\, \tfrac{6}{5} \,| \leqq 1.$$

This gives $\sigma' = \frac{5}{6}$. The new estimate of $\mathbf{c}$ and the new error are respectively, $(\mathbf{c}^0 + \sigma' \mathbf{c}') = \frac{1}{6} \text{col } (2, 3)$ and $(1 - \sigma')\mathbf{e}^0 = \frac{1}{6} \text{col } (1, 4)$.

(ii) To keep the same notation as used in P1 and P2, we denote, ambiguously, the new estimate of $\mathbf{c}$ and the new error as:

$$\mathbf{c}^0 = \tfrac{1}{6} \text{col } (2, 3), \quad \mathbf{e}^0 = \tfrac{1}{6} \text{col } (1, 4).$$

The vector $\mathbf{r}_3$ must be removed from the basis used in (i) in order to reduce the error. Thus, the new basis is $\bar{S} = \{1, 2\}$, and

$$\mathbf{c}' = \Big(\sum_{i \in \bar{S}} \mathbf{r}_i \times \mathbf{r}_i\Big)^{-1} \mathbf{e}^0 = \tfrac{1}{6} \text{col } (-2, 3).$$

Again σ' is chosen to be the largest value of σ, $0 < \sigma \leqq 1$, for which

$$|\, \langle \mathbf{r}_1 , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle \,| = \left| \frac{1}{2} + \frac{\sigma}{2} \right| \leqq 1,$$

$$|\, \langle \mathbf{r}_2 , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle \,| = \left| \frac{5}{6} + \frac{\sigma}{6} \right| \leqq 1,$$

$$|\, \langle \mathbf{r}_3 , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle \,| = |\, 1 + \sigma \,| \geqq 1.$$

We find that $\sigma' = 1$, hence a solution has been obtained with $\mathbf{c} = \mathbf{c}^0 + \sigma'\mathbf{c}' = \text{col } (0, 1)$ and

$$u_1 = \text{sat } \langle \mathbf{r}_1 , \mathbf{c} \rangle = 1,$$

$$u_2 = \text{sat } \langle \mathbf{r}_2 , \mathbf{c} \rangle = 1,$$

$$u_3 = \text{sat } \langle \mathbf{r}_3 , \mathbf{c} \rangle = 1.$$

Example 2. Let $\mathbf{r}_1 = \text{col } (0, 1)$, $\mathbf{r}_2 = \text{col } (1, 1)$, $\mathbf{r}_3 = \text{col } (0, 2)$, $\mathbf{r}_4 = \text{col } (2, 0)$, and let the target state $\mathbf{v}_N$ at time N be given by $\mathbf{v}_1 = \mathbf{v}_2 = \mathbf{v}_3 = \mathbf{v}_4 = \text{col } (3, 3)$. The problem is to find the smallest integer $N \in \{1, 2, 3, 4\}$ and a vector $\mathbf{c} \in \mathcal{C}_N$ such that $\mathbf{v}_N = \mathbf{f}_N(c)$.

I. Starting with $N = 2$ we try to solve the equation $\mathbf{v}_2 = \mathbf{f}_2(\mathbf{c})$, and as an initial estimate $\mathbf{c}^0$ of $\mathbf{c}$ we take $\mathbf{c}^0 = \text{col } (0, 0)$; the error is $\mathbf{e}^0 = \text{col } (3, 3)$. Proceeding as in Example 1:

(i) $\bar{S} = \{1, 2\}$; $\quad \mathbf{c}' = (\sum_{i \in \bar{S}} \mathbf{r}_i \times \mathbf{r}_i)^{-1}\mathbf{e}^0 = \text{col } (3, 0)$; $\quad \sigma' = \frac{1}{3}$.

The new estimate of $\mathbf{c}$ is $\sigma'\mathbf{c}^0 = \text{col } (1, 0)$ and the new error is $(1 - \sigma')\mathbf{e}^0 = \text{col } (2, 2)$.

(ii) Using Lemma 4, it follows that we are on the boundary of $\mathcal{R}_2$, consequently, the target is not reachable in $N = 2$ sampling periods.

II. Increasing N by one, we seek a solution to the equation $\mathbf{v}_3 = \mathbf{f}_3(\mathbf{c})$. As out initial estimate of $\mathbf{c}$ we take $\mathbf{c}^0 = \text{col } (1, 0)$; the new error is then

$$\mathbf{e}^0 = \text{col } (2, 2) - \mathbf{r}_3 \text{ sat } \langle \mathbf{r}_3 , \mathbf{c}^0 \rangle = \text{col } (2, 2) - 0 \cdot \text{col } (0, 2)$$

$$= \text{col } (2, 2).$$

(i) Noting that $\langle \mathbf{r}_1 , \mathbf{c}^0 \rangle = 0$, $\langle \mathbf{r}_2 , \mathbf{c}^0 \rangle = 1$, and $\langle \mathbf{r}_3 , \mathbf{c}^0 \rangle = 0$, the vectors $\mathbf{r}_1$ and $\mathbf{r}_3$ must remain in the basis. Let us assume that $\mathbf{r}_2$ also remains in the basis, checking this assumption after $\mathbf{c}'$ is computed (see, e.g., condition (b) in P1). Taking $\bar{S} = \{1, 2, 3\}$ we find that $\mathbf{c}' = \frac{1}{5} \text{ col } (10, 0)$ and hence $\langle \mathbf{r}_2 , \mathbf{c}^0 \rangle = \text{sgn } \langle \mathbf{r}_2 , \mathbf{c}' \rangle$. Thus $\mathbf{r}_2$ must be removed from the basis if the error is to be reduced. However, the remaining vectors $\mathbf{r}_1$, $\mathbf{r}_3$ do not span E^2 and, therefore, we are on the boundary of $\mathcal{R}_3$. Consequently, the target is not reachable in $N = 3$ sampling periods.

III. Again increasing N by one we try to solve for $\mathbf{c}$ in the equation $\mathbf{v}_4 = \mathbf{f}_4(\mathbf{c})$, taking $\mathbf{c}^0 = \text{col } (1, 0)$ as the initial estimate of $\mathbf{c}$. This results in an error given by

$$\mathbf{e}^0 = \text{col } (2, 2) - \mathbf{r}_4 \text{ sat } \langle \mathbf{r}_4 , \mathbf{c}^0 \rangle = \text{col } (2, 2) - \text{col } (2, 0)$$

$$= \text{col } (0, 2).$$

(i) Since $\langle \mathbf{r}_1 , \mathbf{c}^0 \rangle = 0$, $\langle \mathbf{r}_2 , \mathbf{c}^0 \rangle = 1$, $\langle \mathbf{r}_3 , \mathbf{c}^0 \rangle = 0$, $\langle \mathbf{r}_4 , \mathbf{c}^0 \rangle = 2$, we assume

the basis to be $\bar{S} = \{1, 2, 3\}$. From (22), $\mathbf{c}' = \text{col}\ (-\frac{2}{5}, \frac{2}{5})$ and $\text{sgn}\ \langle \mathbf{r}_2 , \mathbf{c}^0 \rangle = -\text{sgn}\ \langle \mathbf{r}_2 , \mathbf{c}' \rangle$; therefore $\mathbf{r}_2$ remains in the basis. Choosing σ' to be the largest scalar σ, $0 < \sigma \leqq 1$, such that

$$| \langle \mathbf{r}_1 , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle | = | \tfrac{2}{5}\sigma | \leqq 1,$$

$$| \langle \mathbf{r}_2 , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle | = | 1 + 0\sigma | \leqq 1,$$

$$| \langle \mathbf{r}_3 , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle | = | \tfrac{4}{5}\sigma | \leqq 1,$$

$$| \langle \mathbf{r}_4 , \mathbf{c}^0 + \sigma \mathbf{c}' \rangle | = | 2 - \tfrac{4}{5}\sigma | \geqq 1,$$

we find that $\sigma' = 1$, and therefore, a solution has been obtained with $\mathbf{c} = \mathbf{c}^0 + \mathbf{c}' = \text{col}\ (\frac{3}{5}, \frac{2}{5})$. The time optimal control is given by

$$u_1 = \text{sat}\ \langle \mathbf{r}_1 , \mathbf{c} \rangle = \tfrac{2}{5}, \quad u_2 = \text{sat}\ \langle \mathbf{r}_2 , \mathbf{c} \rangle = 1,$$

$$u_3 = \text{sat}\ \langle \mathbf{r}_3 , \mathbf{c} \rangle = \tfrac{4}{5}, \quad u_4 = \text{sat}\ \langle \mathbf{r}_4 , \mathbf{c} \rangle = 1.$$

REFERENCES

[1a] C. A. Desoer and J. Wing, *A minimal time discrete system*, IRE Trans. Automatic Control, AC-6 (1961), pp. 111–125.

[1b] ———, *The minimal time regulator problem for linear sampled-data systems: General theory*, J. Franklin Inst., 272 (1961), pp. 208–228.

[2] L. A. Zadeh and B. H. Whalen, *On optimal control and linear programming*, IRE Trans. Automatic Control, AC-7 (1962), pp. 45–46.

[3] J. H. Eaton, *An on line solution to sampled-data time optimal control*, J. Electronics Control, 15 (1963), pp. 333–341.

[4] M. D. Canon, *A new algorithm for bounded variable quadratic programming problems*, Ph.D. Dissertation, University of California, Berkeley, 1966.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

GENERAL THEORY OF OPTIMAL PROCESSES*

SHELDON S. L. CHANG†

In the present paper a general version of the maximum principle is formulated and proved. Pontryagin's maximum principle [1] and various extensions [2]–[9] thereof become special cases which can be readily derived from the general version. Of special interest are the following generalizations: (1) discrete systems, (2) systems with multiple merit criteria, (3) restriction of the control function u to a special class of functions, and (4) systems with bounded state variables.

Operative addition and convexity. Let ϵ denote an infinitesimal quantity and $\hat{u}(t)$ a given function defined on $T = \{t: t_1 \leqq t \leqq t_2\}$. A function $u(t)$ is said to vary infinitesimally from $\hat{u}(t)$ if

$$\int_{t_1}^{t_2} \| u(t) - \hat{u}(t) \| \, dt < \epsilon A, \tag{1}$$

where A is a positive constant. Obviously if $u(t)$ is different from $\hat{u}(t)$ for a finite amount, it can be only for an infinitesimal interval of time. In the present paper, one needs only to consider infinitesimal variations of the following form:

$$\begin{gathered} \delta u(t) \equiv u(t) - \hat{u}(t) = a_i \text{ on } T_i' = \{t: t_i' < t < t_i' + \epsilon\Delta_i\}, \\ \delta u(t) = \epsilon\, \xi(t) \text{ on } T - T', \qquad T' = \bigcup T_i', \end{gathered} \tag{2}$$

where i may range from 1 to any finite number, $t_i' \in T$, and a_i, Δ_i and $\xi(t)$ are finite numbers and function, respectively.

Operative addition is denoted by $\oplus$ and is defined as follows: let $\delta u_1(t)$ and $\delta u_2(t)$ denote two infinitesimal variations from $\hat{u}(t)$. If the two sets of t_i' have no element in common, the two sets T' are disjoint for sufficiently small ϵ. Then

$$\delta u_1(t) \oplus \delta u_2(t) = \delta u_1(t) + \delta u_2(t). \tag{3}$$

If finite variations occur at the same instant, the variations are rearranged in sequence in $\delta u_1(t) \oplus \delta u_2(t)$. For instance, given

$$\delta u_1(t) = a_k', \quad t_k' \leqq t < t_k' + \epsilon\Delta_k',$$

* Received by the editors June 29, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Department of Electrical Sciences, State University of New York at Stony Brook, New York. This work was sponsored by the Office of Scientific Research, Air Research and Developmen Command, Washington, D. C., under Grant No. AF-AFOSR-542-64.

$$\delta u_2(t) = a_k'', \quad t_k' \leqq t < t_k' + \epsilon\Delta_k'',$$

then

$$\delta u_1(t) \oplus \delta u_2(t) = a_k', \quad t_k' \leqq t < t_k' + \epsilon\Delta_k',$$
$$\delta u_1(t) \oplus \delta u_2(t) = a_k'', \quad t_k' + \epsilon\Delta_k' \leqq t < t_k' + \epsilon(\Delta_k' + \Delta_k''),$$

A variation from a to b can be considered as a variation from b to a for a negative interval.

A set of functions C is *operatively convex* if it has the following property: given any u and infinitesimal variations δu_1 and δu_2 such that all three functions u, $u + \delta u_1$, and $u + \delta u_2$ belong to C, then

$$u + [h\,\delta u_1 \oplus (1 - h)\delta u_2]$$

belongs to C for all values of h in the interval $0 < h < 1$.

The control problem. The controlled system is described by the following set of differential equations:

$$\frac{dx_i}{dt} \equiv \dot{x}_i = f_i(x, u, t), \qquad i = 1, 2, \cdots, n, \tag{4}$$

where x and u are the state vector and the control vector, respectively, and f is a vector function having continuous and bounded first derivatives in x and being continuous in u. From known existence theorems, given $x(t_1)$ and $u(t)$ on $T = \{t: t_1 \leqq t \leqq t_2\}$, $x(t)$ is completely determined.

The control function $u(t)$ is required to satisfy three conditions:
(a) $u(t)$ belongs to an operatively convex set of functions C on T,
(b) the $x(t)$ resulting from $u(t)$ stays within an allowed region X, $x(t) \in X$ for all $t \in T$,
(c) the path terminates at a point $x(t_2)$, where t_2 may be fixed or arbitrary, $t_2 \leqq T$.

A set $u(t)$ satisfying (a) and (b) is called an *admissible control.* When all three conditions are satisfied, $u(t)$ is called an *allowed control.* The merit of an allowed control is judged by a set of variables y_i, where

$$y_i(t) = \int_{t_1}^{t} g_i(x, u, t)\, dt, \qquad i = 1, 2, \cdots, N. \tag{5}$$

An allowed control A is said to be *inferior to* B if

$$y_i(t_2)|_A \geqq y_i(t_2)|_B, \tag{6}$$

and the inequality sign holds for at least one value of i. An allowed control is said to be noninferior if it is not inferior to any other allowed control in the sense defined above.

The noninferior controls are generalizations of optimal controls for a system with multivalued criteria.

First variations of state and merit variables. Due to the infinitesimal variation in $u(t)$ (see (2)), $x(t)$ and $y(t)$ are different from $\hat{x}(t)$ and $\hat{y}(t)$:

$$\begin{aligned} x(t) - \hat{x}(t) &= \epsilon \Delta x(t) + O(\epsilon), \\ y(t) - \hat{y}(t) &= \epsilon \Delta y(t) + O(\epsilon), \end{aligned} \tag{7}$$

where $\Delta x(t)$ and $\Delta y(t)$ are finite and are called the first variations of $x(t)$ and $y(t)$, respectively.

Let z denote the $(n + N)$-dimensional vector

$$\begin{pmatrix} x \\ y \end{pmatrix}$$

and $h(x, u, t)$ denote the $(n + N)$-dimensional vector function

$$\begin{pmatrix} f \\ g \end{pmatrix}.$$

Equations (4) and (5) can be combined as

$$\dot{z} = h(x, u, t). \tag{8}$$

The first variation in z for $t \notin T'$ is readily shown to be

$$\begin{aligned} \Delta z \equiv \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} &= \sum_{\text{all } k \text{ with } t_k' < t} A(t, t_k')[h(\hat{x}, u, t_k') - h(\hat{x}, \hat{u}, t_k')]\Delta_k \\ &+ \int_{t_1}^{t} A(t, t') \frac{\partial h(\hat{x}, \hat{u}, t)}{\partial \hat{u}} \xi(t')\, dt', \end{aligned} \tag{9}$$

where $A(t, t')$ is an $(n + N)$-dimensional square matrix satisfying

$$\frac{\partial A(t, t')}{\partial t'} = \frac{\partial h(\hat{x}, \hat{u}, t)}{\partial \hat{z}} A(t, t'), \tag{10}$$

$$\frac{\partial A(t, t')}{\partial t'} = - A(t, t') \frac{\partial h(\hat{x}, \hat{u}, t')}{\partial \hat{z}}, \tag{11}$$

$$A(t, t) = 1,$$

and $\partial h/\partial z$ is an $(n + N)$-dimensional square matrix with

$$\left(\frac{\partial h}{\partial z}\right)_{i,j} = \frac{\partial h_i}{\partial z_j}.$$

Since the vector function h is independent of y, the last N columns of the matrix $\partial h/\partial z$ are identically zero.

The following theorem is obvious from (9).

THEOREM 1. *Let* $(\delta u)_1$, $(\delta u)_2$ *and* $(\delta u)_3$ *represent infinitesimal variations about* $\hat{u}(t)$ *related by*

$$(\delta u)_1 \oplus (\delta u)_2 = (\delta u)_3 .$$

Let $(\Delta z)_1$, $(\Delta z)_2$ *and* $(\Delta z)_3$ *denote the first variations in* z *resulting from* $(\delta u)_1$, $(\delta u)_2$ *and* $(\delta u)_3$, *respectively. Then*

$$(\Delta z)_1(t) + (\Delta z)_2(t) = (\Delta z)_3(t).$$

COROLLARY. *The set of admissible first variations about any terminal point* $z(t_2)$ *is convex.*

General theorems on optimal control.

THEOREM 2. *Given fixed points* $x(t_1)$ *and* $x(t_2)$, *and letting* X *be the* x*-space (unbounded), a necessary condition for a control and path pair* $\hat{u}(t)$, $\hat{x}(t)$ *to be a noninferior control is that there exists a set of vector functions* $\hat{\psi}(t)$ *and* $H(\hat{\psi}, \hat{x}, \hat{u}, t)$ *satisfying*

$$H(\hat{\psi}, \hat{x}, \hat{u}, t) \equiv \sum_{i=1}^{n} \hat{\psi}_i(t) f_i(\hat{x}, \hat{u}, t) - \sum_{k=1}^{N} c_k g_k(\hat{x}, \hat{u}, t), \tag{12}$$

$$\partial_u \int_{t_1}^{t_2} H(\hat{\psi}, \hat{x}, \hat{u}, t)\, dt \leqq 0, \tag{13}$$

and

$$\frac{d\hat{\psi}_i(t)}{dt} = -\frac{\partial H}{\partial \hat{x}_i}, \qquad i = 1, 2, \cdots, n, \tag{14}$$

where

$$C_k \geqq 0, \tag{15}$$

and the equality sign in (15) *cannot hold for all values of* k; ∂_u *is the first variation of the subsequent integral due to an infinitesimal variation of* $u(t)$, *with* $\hat{\psi}$, $\hat{x}$ *and* t *considered as fixed.*

Note that there is no restriction on the infinitesimal variation δu except that $\hat{u} + \delta u$ belongs to C. In this and later theorems ∂_u is interpreted in the same way. From the definition of ∂_u, (9) can be written as

$$\Delta z(t) = \partial_u \int_{t_1}^{t} A(t, t')h(\hat{x}, \hat{u}, t')\, dt'.$$

Before proving Theorem 2 some geometrical notions in z-space will be established. A point p in z-space is called *accessible* if there exists an admissible $u(t)$ which brings the system from $z(t_1)$ to p at some time t_2. The set of all accessible points at fixed t_2 and at $t_2 \leqq T$ are denoted by $\Omega(t_2)$ and Ω, respectively.

The set of allowed terminal points is an N-dimensional plane, P, at $x = x(t_2)$. The intersection of Ω with P is denoted by I.

LEMMA 1. *The terminal point $\hat{z}(t_2)$ of a noninferior control is a boundary point of I and Ω.*

LEMMA 2. *Let Ω_ϵ denote the set $\hat{z}(t_2) + \epsilon\Delta z$ for all admissible first variations Δz. Then $\hat{z}(t_2)$ is a boundary point of Ω_ϵ.*

Lemma 2 follows intuitively from Lemma 1. It has also been proven rigorously by previous authors [1, pp. 86–106].

LEMMA 3. *There exists a vector l such that*

$$\sum_{i=1}^{n+N} l_i(\Delta z)_i \leqq 0 \tag{16}$$

for all allowed first variations Δz,

$$l_i \leqq 0, \quad i = n+1, n+2, \cdots, n+N, \tag{17}$$

and the equality sign in (17) cannot hold for all values of i.

Proof. Let P_s denote a section in P which satisfies

$$z_i - \hat{z}_i(t_2) \leqq 0, \quad i = n+1, n+2, \cdots, n+N.$$

Since $\hat{z}(t_2)$ is the terminal point of a noninferior control, Ω_ϵ and P_s do not have interior points in common. Furthermore Ω_ϵ is convex because of the Corollary of Theorem 1. There exists a support plane S which separates Ω_ϵ and P_s. Let l be the normal to S. Points on the Ω_ϵ side of S are represented by (16) and all the points on P_s satisfy

$$\sum_{j=n+1}^{n+N} l_j[z_j - \hat{z}_j(t_2)] \geqq 0. \tag{18}$$

By choosing

$$z_j - \hat{z}_j < 0 \quad \text{for } j = i,$$

$$z_j - \hat{z}_j = 0 \quad \text{for all } j \neq i,$$

(18) leads to (17). Because S can coincide at most with one boundary plane of S, the final assertion of the lemma is obtained.

Proof of Theorem 2. Let l' represent the row vector $(l_1, l_2, \cdots, l_{n+N})$. Inequality (16) can be written as

$$l'\Delta z(t_2) \leqq 0. \tag{19}$$

Since X is the entire x-space, the class of admissible controls is identical with C. From (9),

$$\Delta z(t_2) = \partial_u \int_{t_1}^{t_2} A(t_2, t)h(\hat{x}, \hat{u}, t)\, dt. \tag{20}$$

Substituting (20) into (19) gives

$$\partial_u \int_{t_1}^{t_2} l'A(t_2, t)h(\hat{x}, \hat{u}, t)\, dt \leqq 0. \tag{21}$$

Let $\hat{\psi}'(t)$ denote the row vector $l'A(t_2, t)$, and let $H(\psi, x, u, t)$ be defined as

$$H(\psi, x, u, t) \equiv \psi'(t)\, h(x, u, t). \tag{22}$$

Inequality (21) is identical with (13). Multiplying (11) on the left by l' gives

$$\frac{d\hat{\psi}'(t)}{dt} = -\hat{\psi}' \frac{\partial h}{\partial z}. \tag{23}$$

The first n components of (23) give (14). The last N components of (23) give

$$\psi_i(t) = \text{const.} = l_i, \qquad i = n + 1, \cdots, n + N.$$

Let $C_k \equiv -l_{n+k}$. Equation (22) is then identical with (12).

THEOREM 3. *Given fixed points $x(t_1)$ and $x(t_2)$, and letting X be an n-dimensional smooth region in x-space, a necessary condition for a control and path pair $\hat{u}(t)$, $\hat{x}(t)$ to be a noninferior control is that there exist $\hat{\psi}(t)$, $\zeta(\hat{x}, t)$ and $H(\hat{\psi}, \hat{x}, \hat{u}, t)$ satisfying* (12), (13), (15), *and the following:*

$$\frac{d\hat{\psi}_i}{dt} = -\frac{\partial H}{\partial \hat{x}_i} \zeta(\hat{x}, t)\eta_i(\hat{x}), \tag{24}$$

where

$$\zeta(\hat{x}, t) \begin{cases} = 0 \text{ if } \hat{x} \text{ is an interior point of } X, \\ \geqq 0 \text{ if } \hat{x} \text{ is on the boundary of } X, \end{cases} \tag{25}$$

and η is the normal to X at $\hat{x}$ pointing away from X.

Proof. The proof of Theorem 3 is identical with that of Theorem 2 up to (19). Inequality (19) holds only for infinitesimal admissible variations of u. Condition (2) defining admissible variations can be written as

$$\sum_{i=1}^{n} \eta_i(\hat{x})\Delta x_i(t) \leqq 0 \quad \text{on} \quad \Gamma', \tag{26}$$

where Γ' is the part of the path $\hat{x}(t)$ lying on the boundary of X. Thus (19) is replaced by itself together with the side condition (26). Let η' represent the $(n + N)$-component row vector with the η_i as its first n components and 0 for the remaining N components. Inequality (26) is identical with

$$\eta'(\hat{x})\Delta z(t) \leqq 0 \quad \text{on} \quad \Gamma'. \tag{27}$$

From a well-known result in variation calculus, (19) together with the side condition (27) is equivalent to the existence of a $\zeta(t)$ such that

$$l'\Delta z(t_2) - \int_\Gamma \zeta(t)\eta'(\hat{x})\Delta z(t)\,dt \leqq 0. \tag{28}$$

The integral is taken over periods of time in which x lies on the boundary of X.

LEMMA 4. *A necessary condition for $\hat{u}(t)$ and $\hat{x}(t)$ to be a noninferior control and path pair is that there exists a $\zeta(t)$ such that (28) is satisfied by all first variations $\Delta(z)$ resulting from $\delta u(t)$ with $\hat{u} + \delta u$ belonging to C.*

Let $\zeta(\hat{x}, t) = \zeta(t)$ when $\hat{x}(t)$ is a boundary point, and equal zero when $\hat{x}(t)$ is an interior point. From (9),

$$\begin{aligned}\int_{\Gamma'} \zeta(t)\eta'(\hat{x})\Delta z(t)\,dt &= \int_{t_1}^{t_2} \zeta(\hat{x}, t)\eta'(\hat{x})\partial_u \int_{t_1}^{t} A(t, t')h(\hat{x}, \hat{u}, t')dt'\,dt \\ &= \partial_u \int_{t_1}^{t_2}\int_{t_1}^{t} \zeta(\hat{x}, t)\eta'(\hat{x})A(t, t')h(\hat{x}, \hat{u}, t')\,dt'\,dt.\end{aligned}$$

Changing the order of integration but integrating over the same area gives

$$\int_{t_1}^{t_2}\int_{t_1}^{t} \cdots\, dt'\,dt = \int_{t_1}^{t_2}\int_{t}^{t_2} \cdots\, dt\,dt'.$$

Interchanging the notations t and t', one finally obtains

$$\begin{aligned}&\int_{\Gamma'} \zeta(t)\eta'(\hat{x})\Delta z(t)\,dt \\ &\qquad = \partial_u \int_{t_1}^{t_2}\int_{t}^{t_2} \zeta(\hat{x}, t')\eta'(\hat{x}(t'))A(t', t)h(\hat{x}, \hat{u}, t)\,dt'\,dt.\end{aligned} \tag{29}$$

Substituting (20) and (29) into (28) gives

$$\partial_u \int_{t_1}^{t_2} \left[l'A(t_2, t) - \int_t^{t_2} \zeta(\hat{x}, t')\eta'(\hat{x})A(t', t)\,dt'\right] h(\hat{x}, \hat{u}, t)\,dt \leqq 0. \tag{30}$$

Let $\hat{\psi}'(t)$ be defined as

$$\hat{\psi}'(t) \equiv l'A(t_2, t) - \int_t^{t_2} \zeta(\hat{x}, t')\eta'(\hat{x})A(t', t)\,dt', \tag{31}$$

and let $H(\psi, x, u, t)$ be defined the same way as in (22); then (30) is identical with (13). From (31),

$$\begin{aligned}\frac{d\psi'(t)}{dt} &= l'\frac{\partial A(t_2, t)}{\partial t} + \zeta(\hat{x}, t)\eta'(\hat{x}) - \int_t^{t_2} \zeta(\hat{x}, t')\eta'(\hat{x})\frac{\partial A(t', t)}{\partial t}\,dt' \\ &= -\psi'\frac{\partial h(\hat{x}, \hat{u}, t)}{d\hat{z}} + \zeta(\hat{x}, t)\eta'(\hat{x}).\end{aligned} \tag{32}$$

The first n components of (32) give (24). The last N components give

$$\psi_{n+k}(t) = -C_k, \quad k = 1, 2, \cdots, N, \tag{33}$$

and (12).

The condition (25) is proved in a previous paper for a less general problem (10). It also follows from the intuitive notion that $y_i(t_2)$ can be reduced if the path $x(t)$ is allowed an excursion beyond X.

The following theorem is the well-known transversality condition, and will be stated without a proof.

THEOREM 4. *Let $u(t)$, $t_1 \leqq t \leqq t_2$, be an admissible control which transfers the phase point from some position $x(t_1) \in S_1$ to the position $x(t_2) \in S_2$, where S_1 and S_2 are smooth regions of points satisfying the following equations and inequalities:*

$$\begin{aligned} S_1: \quad & F_i(x) = 0, \quad i = 1, 2, \cdots, k \leqq n, \\ & G_i(x) \leqq 0, \quad i = 1, 2, \cdots, m, \\ S_2: \quad & F_i'(x) = 0, \quad i = 1, 2, \cdots, k' \leqq n, \\ & G_i'(x) \leqq 0, \quad i = 1, 2, \cdots, m'. \end{aligned}$$

In order that $u(t)$ and $z(t)$ yield the solution of the noninferior problem with variable endpoints, it is necessary that there exists a nonzero continuous vector function $\psi(t)$ which satisfies the conditions of Theorem 3 and, in addition, the transversality condition at both endpoints of the trajectory $z(t)$,

$$\psi_i(t_1) = \sum_{j=1}^{k} a_j \frac{\partial F_j}{\partial x_i} + \sum_{j=1}^{m} b_j \frac{\partial G_j}{\partial x_i}, \tag{34}$$

$$\psi_i(t_2) = \sum_{j=1}^{k'} a_j' \frac{\partial F_j'}{\partial x_i} - \sum_{j=1}^{m'} b_j' \frac{\partial G_j'}{\partial x_i}, \tag{35}$$

where a_j and a_j' are arbitrary constants, and b_j and b_j' are nonnegative constants such that

$$\begin{aligned} & b_j = 0 \quad \text{if} \quad G_j(x) < 0 \text{ or if } x(t_1) \text{ is an interior point of } S_1, \\ & b_j \geqq 0 \quad \text{if} \quad G_j(x) = 0 \text{ and the equation actually defines the boundary of } S_1 \text{ at } x(t_1), \end{aligned} \tag{36}$$

and similar conditions hold for b_j'. In (34), (35) and (36) the values of the functions and partial derivatives are evaluated at the corresponding endpoints.

THEOREM 5. *Consider a control problem satisfying the following conditions:*

$$f(x, u, t) = A(t)x + B(t)u + f(t), \tag{37}$$

$$g(x, u, t) = p(x, t) + q(u, t), \tag{38}$$

where $A(t)$ and $B(t)$ are matrices, $f(t)$, $p(x, t)$, and $q(u, t)$ are vector functions, $p(x, t)$ is convex in x, and $q(u, t)$ is convex in u, X, S_1, S_2, and the class C are convex.

If for an allowed control and path pair, $\hat{u}(t)$ and $\hat{x}(t)$, a set of functions, $H(\hat{\psi}, \hat{x}, \hat{u}, t)$, $\hat{\psi}(t)$, $\zeta(\hat{x}, t)$, and $C_i > 0$, $i = 1, 2, \cdots, N$, can be found such that (12), (13), (24), (25), and the transversality condition are satisfied, then $\hat{u}(t)$ is a noninferior control among all admissible controls which transfer the phase point from a point on S_1 at t_1 to a point on S_2 at t_2.

Proof. Let C' denote the row vector $(C_1, C_2, \cdots, C_N)$. From (12), (37), and (38),

$$H(\psi, x, u, t) = \psi' A(t)x + \psi' B(t)u + \psi' f(t) - C'p(x, t) - C'q(u, t). \tag{39}$$

From (24),

$$\frac{d}{dt}\hat{\psi}' = -\hat{\psi}' A(t) + \frac{\partial C'p(\hat{x}, t)}{\partial \hat{x}} + \zeta(\hat{x}, t)\eta'(\hat{x}). \tag{40}$$

Consider any other allowed control and path pair $u(t)$, $x(t)$ which satisfy the same terminal conditions. Evaluate the following total derivative:

$$\begin{aligned}\frac{d}{dt}[\hat{\psi}'(\hat{x} - x)] &= \hat{\psi}' B(t)(\hat{u} - u) \\ &\quad + \frac{\partial C'p(\hat{x}, t)}{\partial \hat{x}}(\hat{x} - x) + \zeta(\hat{x}, t)\eta'(\hat{x})(\hat{x} - x).\end{aligned} \tag{41}$$

Subtracting $C'p(\hat{x}, t) + C'q(\hat{u}, t) - C'p(x, t) - C'q(u, t)$ from both sides of (41) and integrating from t_1 to t_2 give

$$\begin{aligned}&\hat{\psi}'(\hat{x} - x)\Big|_{t_1}^{t_2} - C'[\hat{y}(t_2) - y(t_2)] \\ &\quad = \int_{t_1}^{t_2}\left\{\hat{\psi}' B(t)(\hat{u} - u) - \frac{\partial C'q(\hat{u}, t)}{\partial \hat{u}}(\hat{u} - u)\right] dt \\ &\quad\quad + \int_{t_1}^{t_2}\left[\frac{\partial C'q(\hat{u}, t)}{\partial \hat{u}}(\hat{u} - u) - C'q(\hat{u}, t) + C'q(u, t)\right] dt \\ &\quad\quad + \int_{t_1}^{t_2}\left[\frac{\partial C'p(\hat{x}, t)}{\partial \hat{x}}(\hat{x} - x) - C'p(\hat{x}, t) + C'p(x, t)\right] dt \\ &\quad\quad + \int_{t_1}^{t_2}\zeta(\hat{x}, t)\eta'(\hat{x})(\hat{x} - x)\, dt.\end{aligned} \tag{42}$$

On the right-hand side of (42), the first integral is nonnegative because of (13) and the convexity of the class C. The second and third integrals are nonnegative because of the convexity of the functions $p(x, t)$ and $q(u, t)$, and $C_i > 0$, $i = 1, 2, \cdots, N$. The last integral is nonnegative because of (25) and the convexity of X. On the left-hand side of (42),

$$\hat{\psi}'(\hat{x} - x) \Big|_{t_1}^{t_2}$$

is nonpositive because of the transversality condition and convexity of S_1 and S_2. Therefore

$$C'[\hat{y}(t_2) - y(t_2)] \leqq 0.$$

Examples of application of the theorems to discrete systems and other special cases are given in a companion paper.

REFERENCES

[1] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze and E. F. Mischchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[2] S. S. L. Chang, *Computer optimization of nonlinear control systems by means of digitized maximum principle*, IRE International Convention Record, New York, 1961, part 4, pp. 48–55.

[3] ———, *Synthesis of Optimum Control Systems*, McGraw-Hill, New York, 1961, Chap. 12.

[4] S. Katz, *A discrete version of Pontryagin's maximum principle*, J. Electronics Control, 13 (1962), pp. 179–184.

[5] L. T. Fan and C. S. Wang, *The Discrete Maximum Principle*, John Wiley, New York, 1964.

[6] R. V. Gamkrelidze, *Optimal control processes with restricted phase coordinates*, Izv. Akad. Nauk SSSR Ser. Mat., 24 (1960), pp. 315–356.

[7] S. S. L. Chang, *Minimal time control with multiple saturation limits*, IEEE Trans. Automatic Control, AC-8 (1963), pp. 35–42.

[8] ———, *Optimal control in bounded phase space*, Automatica, 1(1963), pp. 55–67.

[9] ———, *A modified maximum principle for optimum control of systems with bounded phase coordinates*, Second IFAC Congress, Basel, Switzerland, 1963, pp. 405/1–4.

[10] ———, *An extension of Ascoli's theorem and its application to the theory of optimal control*, Trans. Amer. Math. Soc., (1965), to appear.

[11] ———, *General theory of optimal processes with applications*, Proceedings of IFAC Tokyo Symposium, 1965, to appear.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

LINEAR CONTROL PROCESSES AND MATHEMATICAL PROGRAMMING*

GEORGE B. DANTZIG†

Linear control process defined [8], [14]. We shall consider an "object" defined by its $n + 1$ coordinates $X = (x_0, x_1, \cdots, x_n)$, whose "motion" described as a function of a parameter, "time" (t), can be written as a linear system of differential equations

$$\frac{dX}{dt} = A^t X + B^t u, \tag{1}$$

where A^t, B^t are known matrices that may depend on t and

$$u = (u_1, u_2, \cdots, u_p)$$

is a control vector that must be chosen from a convex set, $u \in U(t)$ for every $0 \leqq t \leqq T$. The time period $0 \leqq t \leqq T$ is fixed and known in advance. The coordinate $x_0 = x_0(t)$ represents the "cost" of moving the object from its initial position to $x_0(t)$. For this purpose it may be assumed that $x_0(0) = 0$. Defining

$$\bar{X} = (0, x_1, x_2, \cdots, x_n), \tag{2}$$

the object is required to start somewhere in a convex domain $\bar{X}(0) \in S_0$ and to terminate at $t = T$ somewhere on another convex domain $\bar{X}(T) \in S_T$.

Problem. Find $u \in U(t)$ and boundary values $\bar{X}(0) \in S_0$, $\bar{X}(T) \in S_T$, such that $x_0(T)$ is minimized.

Assuming $u \in U(t)$ is known, the system of differential equations can be integrated to yield an expression for $X(T)$ in terms of $X(0)$ and $u \in U(t)$. This is true in general but will be illustrated for the case when A^t and B^t do not depend on t; in this case

$$X(T) = e^{TA}X(0) + \int_0^T e^{(T-t)A}Bu(t)\,dt, \tag{3}$$

where $u(t) \in U(t)$ is a convex set and where we assume the integral exists whatever be the choice of the $u(t) \in U(t)$ for $0 \leqq t \leqq T$.

Generalized linear program [2]. Our general objective is to illustrate

* Received by the editors January 12, 1965, and in revised form March 25, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Operations Research Center, University of California, Berkeley, California. This research has been partially supported by the National Science Foundation under Grant GP-2633 with the University of California.

how *mathematical programming* and, in particular, how the *decomposition principle* in the form of the generalized linear program can be applied to this class of problems. An elegant constructive theory emerges, [10], [11], [12], [13].

A *generalized linear program* differs from a standard linear program in that the vector of coefficients, say P, associated with any variable μ need not be constant but can be selected from a convex set C. For example:

Problem. Find $\max \lambda$, $\mu \geqq 0$ such that

$$U_0\lambda + P\mu = Q_0, \quad \mu = 1, \tag{4}$$

where U_0, Q_0 are specified vectors and $P \in C$ convex.

It is assumed that the elements of C are only known implicitly (for example, as some solution to a linear program) but that particular choices of P can be easily obtained which minimize any given linear form in the components of P.

The method of solution assumes we have initially[1] on hand m particular choices $P_i \in C$ with the property that

$$\begin{aligned} U_0\lambda + P_1\mu_1 + P_2\mu_2 + \cdots + P_m\mu_m &= Q_0, \\ \mu_1 + \mu_2 + \cdots + \mu_m &= 1, \end{aligned} \tag{5}$$

has a unique "feasible" solution; that is to say, $\lambda = \lambda^0$, $\mu_i = \mu_i{}^0 \geqq 0$ and the matrix

$$B^0 = \begin{bmatrix} U_0 & P_1 & \cdots & P_m \\ 0 & 1 & \cdots & 1 \end{bmatrix} \tag{6}$$

is nonsingular (i.e., the columns of B^0 form a basis). Because $P_i \in C$, the vector $P^0 = \sum P_i\mu_i{}^0$ constitutes a solution $P = P^0$ for (4) except that $\lambda = \lambda^0$ may not yield the maximal λ.

To test whether or not P^0 is an optimal solution one determines a row vector $\bar{\pi} = \bar{\pi}^0$ such that

$$\bar{\pi}^0 B^0 = (1, 0, \cdots, 0), \tag{7}$$

and then a value δ and a vector $P_{m+1} \in C$ such that

$$\delta = \bar{\pi}^0 \bar{P}_{m+1} = \min_{P \in C} \bar{\pi}^0 \bar{P}, \tag{8}$$

where we denote

$$\bar{P} = \begin{bmatrix} P \\ 1 \end{bmatrix}. \tag{9}$$

If it turns out that $\delta = 0$, then $P = P^0$ is an optimal solution.

[1] This is not a restrictive assumption since there is an analogous method for obtaining such a starting solution, see [2].

If P^0 is not optimal, system (5) is augmented by P_{m+1}. After one or several iterations k the augmented system takes the form of a linear program:

Problem. Find $\max \lambda$, $\mu_i \geqq 0$,

$$U_0\lambda + \sum_1^{m+k} P_i\mu_i = Q_0, \quad \sum_1^{m+k} \mu_i = 1. \tag{10}$$

Letting B^k denote the basis associated with an optimal basic feasible solution $\mu_i = \mu_i{}^k$ to (10), π^k is defined analogous to (7) and δ^{k+1} and P_{m+k+1} analogous to (8). If it turns out that $\delta = 0$, the solution

$$P^k = \sum_1^{m+k} P_i\mu_i{}^k \tag{11}$$

is optimal. If not the system is augmented by P_{m+k+1} and the iterative process is repeated.

It is known under certain general conditions, such as C bounded and the initial solution nondegenerate (i.e., $\mu_i{}^0 > 0$), that $\bar{\pi}^k \to \bar{\pi}^*$ and $P^k \to P^*$ on some subsequence k and that $P = P^*$ is optimal. The two fundamental properties of $\bar{\pi}^*$ are

$$\bar{\pi}^* \neq 0 \quad \text{and} \quad \bar{\pi}^*\bar{P} \geqq \bar{\pi}^*\bar{P}^* = 0 \quad \text{for all} \quad P \in C. \tag{12}$$

The entire process can be considered as constructive providing it is not difficult to compute the various P_{m+k+1} from (8) with $\bar{\pi} = \bar{\pi}^{m+k}$. For example, if C is a parallelepiped or more generally a convex polyhedral set, then $\min \bar{\pi}\bar{P}$ constitutes the minimization of a linear form with known coefficients $\bar{\pi} = \bar{\pi}^{m+k}$ subject to linear inequality constraints in the unknown components of $\bar{P}$, i.e., a linear program. In this case the iterative process terminates in a finite number of steps and P_{m+k} constitute extreme solutions from it. In all cases an estimate is available on how close the kth solution is to an optimal value of λ.

Application of the generalized program to the linear control process. Let us denote

$$P = \int_0^T e^{(T-t)A} Bu(t)\, dt, \tag{13}$$

and note that P is an element of a convex set C_u generated by choosing all possible $u(t) \in U(t)$. We specify that $U_0 = (1, 0, \cdots, 0)$, and denote by $\lambda = -X_0(T)$, where $X_0(T)$ is the coordinate of $X(T)$ to be minimized. Then

$$X(T) = -U_0\lambda + \bar{X}(T). \tag{14}$$

We further define Q_0 by

$$\bar{X}(T) = e^{TA}X(0) + Q_0. \tag{15}$$

Substitution of these into (3) formally converts[2] the integrated form of the control problem into a generalized linear program (4).

Each cycle of the iterative process yields a known row vector, which we partition

$$\bar{\pi}^{k+1} = [\pi, \theta], \tag{16}$$

where π represents its first $n + 1$ components corresponding to P and θ its last component. Since π is known, our choice for P_{m+k+1} is

$$\pi P_{m+k+1} = \min\left\{\int_0^T \pi e^{(T-t)A}Bu(t)\,dt\right\} = \int_0^T \{\min_{u\in U(t)} e^{(T-t)A}Bu(t)\}\,dt, \tag{17}$$

where clearly the minimum is obtained when, in (17), the integrand for each t is selected to be minimum.

Note that

$$\phi_{t,\pi} = \pi e^{(T-t)A}B \tag{18}$$

is a row vector that can be computed for each t. For example, $\phi_{t,\pi}$ can be represented by a finite sum of vectors whose weights depend on t and the eigenvalues of A. The new extremal solution P_{m+k+1} is obtained by choosing the control which minimizes the linear form in u for each t; i.e., find

$$\min\,(\phi_{t,\pi}u), \quad u \in U(t). \tag{19}$$

For example, if $U(t)$ is a polyhedral set then (19) is a linear program. If $U(t)$ is the same for all t, then only the objective form, $\phi_{t,\pi}u$, varies for different t; except for the objective form the linear programs are the same for all t.

If optimal π^* is used, then the optimal control u (except for a set of measure zero) satisfies

$$\min\,[\phi^*(t)u], \quad u \in U(t), \tag{20}$$

where $\phi^*(t) = \pi^* e^{(T-t)A}B$. Pontryagin refers to this as *the maximal principle*. It is, as we have just shown, also a consequence of the decomposition principle of linear programming.

Conclusion. In our approach the general control obtained for each cycle is a linear combination of exactly $n + 1$ special controls obtained by mini-

[2] Actually Q_0 is not given but is an element of a convex set. To simplify the discussion which follows we assume Q_0 is a fixed vector.

mizing for each t, the linear expression (19) in u for $n + 1$ choices of π. These special controls may be referred to as extreme controls. The latter each in themselves do not maintain feasibility, that is to say, guarantee that the object will move from $\bar{X}(0)$ to $\bar{X}(T)$. Each new linear combination of these special controls will, however, generate a new feasible control with a lower value[3] for the total cost $X_0(T)$. Under the conditions stated this iterative process is known to converge.

REFERENCES

[1] R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, 1961.

[2] G. B. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton, 1963.

[3] A. F. Filippov, *On certain questions in the theory of optimal control*, English transl., this Journal, 1(1962), pp. 76–84.

[4] H. Halkin, *On the necessary conditions for optimal control of nonlinear systems*, J. Analyse Math., 12 (1964), pp. 1–82.

[5] J. P. LaSalle, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. V, Princeton University Press, Princeton, 1958.

[6] J. P. LaSalle and S. Lefschetz, *Stability by Liapunov's Direct Method*, Academic Press, New York, 1961.

[7] G. Leitmann, ed., *Optimization Techniques*, Academic Press, New York, 1962.

[8] L. W. Neustadt, *Discrete time optimal control systems*, Nonlinear Differential Equations and Nonlinear Mechanics, J. P. LaSalle and S. Lefschetz, eds., Academic Press, New York, 1963.

[9] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimum Processes*, Interscience, New York, 1962.

[10] B. H. Whalen, *Linear programming for optimal control*, Ph.D. dissertation, University of California, Berkeley, 1962.

[11] ———, *On linear programming and optimal control*, Correspondence to IRE Trans. on Automatic Control, AC-7 (1962), p. 46.

[12] R. M. Van Slyke, *Mathematical programming*, Ph.D. dissertation, University of California, Berkeley, 1965.

[13] L. A. Zadeh, *A note on linear programming and optimal control*, Correspondence to IRE Trans. on Automatic Control, AC-7 (1962), p. 46.

[14] L. A. Zadeh and C. A. Desoer, *Linear System Theory, The State-Space Approach*, McGraw-Hill, New York, 1963.

[3] If basic solution is nondegenerate.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

AN ITERATIVE PROCEDURE FOR COMPUTING THE MINIMUM OF A QUADRATIC FORM ON A CONVEX SET*

ELMER G. GILBERT†

1. Introduction. This paper presents an iterative procedure for computing the minimum of a quadratic form on a compact convex set C. The sole characterization required of C is the availability of a method for solving linear programs on C. This characterization differs from the usual set of functional inequalities given in quadratic programming problems [6], and is particularly appropriate to the solution of problems in optimal control. In fact, some of the results presented here arose from an attempt to provide a convergence proof for the extension by Fancher [5] of a procedure due to Ho [8]. Section 8 and [1] give several direct applications of the iterative procedure to problems in optimal control. By using the algorithm of this paper as a means of projecting points into convex sets it is possible to develop additional algorithms for solving other problems in programming and control [1], [7].

It should be noted that the iterative procedure of this paper is very similar to that given in the latter part of the paper by Frank and Wolfe [6]. However the emphasis and setting of the two papers are quite different, and the overlap is small.

The paper is organized as follows: in §2 notation, definitions, and a basic problem (BP) are considered; in §§3, 4, and 5 the algorithm for BP is described, error bounds are derived, and convergence is proved and investigated in detail; in §6 the algorithm is related to a gradient method for solving BP; in §7 the previous results are extended to a general quadratic programming problem GP; and in §8 the connection with problems in optimal control is made.

2. Preliminaries, the basic problem. The following notation is employed: $z = (z^1, \cdots, z^n)$, a vector in Euclidean n-space E^n; $y \cdot z = \sum_{i=1}^n y^i z^i$; $|z| = (z \cdot z)^{1/2}$; $N(x; \omega) = \{y \mid |y - x| < \omega\}$, $\omega > 0$, the open sphere at x with radius ω; $\bar{N}(x; \omega) = \{y \mid |y - x| \leqq \omega\}$, the corresponding closed sphere; $L(x; y) = \{z \mid z = x + \omega(y - x), -\infty < \omega < \infty\}$, $x \neq y$, the line passing through x and y; $Q(x; y) = \{z \mid z \cdot y = x \cdot y\}$, $y \neq 0$,

* Received by the editors June 30, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Information and Control Engineering, University of Michigan, Ann Arbor, Michigan. This research was supported by the United States Air Force under Grant No. AF-AFOSR-814-65.

the hyperplane (dimension $n-1$) through x with normal y; ∂X, the boundary of the set X.

Now consider some notation and results applicable to a set $K \subset E^n$, which is compact and convex. Let $\eta(y) = \max_{z \in K} z \cdot y$ denote the *support function* of K. Since K is compact, $\eta(y)$ is defined for all y. Furthermore, it can be shown that $\eta(y)$ is a convex function on E^n, a result which implies that $\eta(y)$ is continuous on E^n [3]. Let $P(y)$, $y \neq 0$, be the hyperplane $\{x \mid x \cdot y = \eta(y)\}$. Since $z \cdot y \leqq \eta(y)$ for all $z \in K$ and $P(y) \cap K$ is not empty, $P(y)$ is the (unique) *support hyperplane* of K with outward normal y. For each $y \neq 0$ the set $S(y) = P(y) \cap K$ is called the *contact set* of K. It follows that $S(y)$ is not empty, $S(y) \subset \partial K$, $S(\omega y) = S(y)$ for $\omega > 0$. If for every $y \neq 0$, $S(y)$ contains only a single point, then K is *strictly convex*.

Definition. A function, $s(y)$, defined on E^n is a *contact function* of K if $s(y) \in S(y)$, $y \neq 0$, and $s(0) \in K$.

From the preceding it may be concluded that $s(\cdot)$ is bounded; $s(y) = s(\omega y)$, $\omega > 0$; and $\eta(y) = s(y) \cdot y$. Furthermore, on the set $\{y \mid\mid y \mid > 0\}$ each of the following is true if and only if K is strictly convex: $s(\cdot)$ is uniquely determined, $s(\cdot)$ is continuous. The continuity result is proved in [10].

If for every y there is a method for determining a point $x(y) \in K$ such that $x(y) \cdot y = \max_{z \in K} z \cdot y = \eta(y)$, then this method may be used to evaluate a contact function of K. Such an evaluation, which corresponds to the solution of a linear programming problem on K (see §1), is essential to the computing procedure which follows. Consider now the basic problem:

BP. *Given K, a compact convex set in E^n, find a point $z^* \in K$ such that* $\mid z^* \mid = \min_{z \in K} \mid z \mid$.

Since K is compact and $\mid z \mid$ is a continuous function of z, a solution z^* exists. The following additional results hold:

Solution properties. (i) z^* *is unique,* (ii) $\mid z^* \mid = 0$ *if and only if* $0 \in K$, (iii) *for* $\mid z^* \mid > 0$, $z^* \in \partial K$, (iv) *for* $\mid z^* \mid > 0$, $z = z^*$ *if and only if* $z \in P(-z) \cap K = S(-z)$.

Properties (ii) and (iii) are obvious. Property (i) is proved by contradiction. Suppose z_1^* and z_2^* are distinct solutions. Then by convexity $\tilde{z} = \frac{1}{2} z_1^* + \frac{1}{2} z_2^* \in K$, which means $\mid \tilde{z} \mid \geqq \mid z_1^* \mid = \mid z_2^* \mid$. But this implies

$$\mid \tfrac{1}{2} z_1^* + \tfrac{1}{2} z_2^* \mid^2 \geqq \tfrac{1}{2} \mid z_1^* \mid^2 + \tfrac{1}{2} \mid z_2^* \mid^2,$$

which can be written $\mid z_1^* - z_2^* \mid^2 \leqq 0$, an inequality which is only true for $z_1^* = z_2^*$. Consider (iv). The condition $z \in P(-z) \cap K$ implies $z \in P(-z) = Q(z; z)$. But $Q(z; z)$ is the support hyperplane for the closed sphere $\bar{N}(0; \mid z \mid)$ whose outward normal is z and whose contact point is z. Therefore $Q(z; z)$ is a (separating) support hyperplane for K

and $\bar{N}(0; |z|)$. Thus $K \cap N(0; |z|)$ is empty. Since $z \in K \cap \bar{N}(0; |z|)$, this implies $z = z^*$. The steps of this argument may be reversed to obtain the converse result.

3. The iterative procedure for the basic problem. In this section the iterative procedure for computing the solution to BP is described.

As a first step, let $s(\cdot)$ be a specific contact function of K and consider

$$\beta(z) = \begin{cases} |z - s(-z)|^{-2} z \cdot (z - s(-z)) & \text{if } z - s(-z) \neq 0, \\ 0 & \text{if } z - s(-z) = 0, \end{cases} \tag{3.1}$$

and

$$\gamma(z) = \begin{cases} |z|^{-2} z \cdot s(-z) & \text{if } |z| > 0,\ z \cdot s(-z) > 0, \\ 0 \quad \text{if } z = 0 \text{ or } |z| > 0,\ z \cdot s(-z) \leqq 0. \end{cases} \tag{3.2}$$

Thus $\beta(\cdot)$ and $\gamma(\cdot)$ are functions which are defined on K. Their geometric significance is as follows: $x = z + \beta(z)\ (s(-z) - z)$ is the point on the line $L(z; s(-z))$ with minimum Euclidean length; $\gamma(z)z$ is either the point $L(0; z) \cap P(-z)$ or the origin, depending on whether or not $L(0; z) \cap P(-z)$ is on the line segment connecting z and the origin. The functions $\beta(\cdot)$ and $\gamma(\cdot)$ have the following properties.

THEOREM 1. *Let K be the set described in BP and restrict z to K. Then*

(i) $\beta(z) \geqq 0$,

(ii) $\beta(z) = 0$ *if and only if* $z = z^*$,

(iii) $0 \leqq \gamma(z) \leqq 1$,

(iv) *if* $0 \in K$, $\gamma(z) \equiv 0$,

(v) *if* $0 \notin K$, $\gamma(z) = 1$ *if and only if* $z = z^*$,

(vi) $\gamma(z)$ *is continuous.*

Proof. In this paragraph z always denotes a point in K. In §4 (inequality (4.5)) it is shown that $0 \leqq z \cdot (z - s(-z))$. Hence, (i) and (iii) follow from (3.1) and (3.2). For the time being assume $|z^*| > 0$. The conditions $\beta(z) = 0$ and $\gamma(z) = 1$ both imply $z \cdot (-z) = s(-z) \cdot (-z) = \eta(-z)$ which requires $z \in P(-z)$. Since $z \in K$, solution property (iv) yields $z = z^*$. Reversing these arguments completes the proof of (ii) for $|z^*| > 0$ and of (v). Now take $|z^*| = 0$. Inequality (4.4) then implies $s(-z) \cdot z \leqq 0$ which by (3.2) yields (iv). If $\beta(z) = 0$, then it must follow from (3.1) that $s(-z) \cdot z = |z|^2$. Because of $s(-z) \cdot z \leqq 0$ this implies $z = 0 = z^*$. Since $z = z^* = 0$ also yields $\beta(z) = 0$, the proof of (ii) is complete. For $|z| \geqq |z^*| > 0$, the continuity of $\gamma(z)$ follows from (3.2) and the conti-

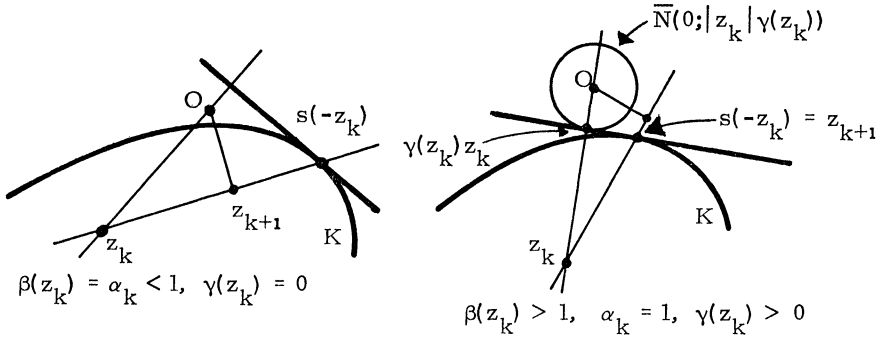

FIG. 1. *Geometric interpretation of the iterative procedure (O origin)*

nuity of the support function $\eta(y) = s(y) \cdot y$. For $| z^* | = 0$, it is trivially true from (iv).

It is of interest to note that $\beta(\cdot)$ may be discontinuous on K, even though $s(\cdot)$ is continuous on K. See Example 3, §5.

The iterative procedure defines a sequence of vectors $\{z_k\}$ by

$$z_{k+1} = z_k + \alpha_k(s(-z_k) - z_k), \qquad k = 0, 1, 2, \cdots, \tag{3.3}$$

where z_0 is an arbitrary point in K and the scalars α_k are selected arbitrarily from the closed interval $I(z_k)$,

$$I(z) = [\min\{\delta\beta(z), 1\}, \min\{(2 - \delta)\beta(z), 1\}], \qquad 0 < \delta = \text{fixed number} \leqq 1. \tag{3.4}$$

Fig. 1 gives the geometric interpretation of the iterative procedure for the case where $\delta = 1$ and $\alpha_k \in I(z_k)$ reduces to $\alpha_k = \operatorname{sat} \beta(z_k)$ ($\operatorname{sat} \omega = \omega$, $0 \leqq \omega \leqq 1$; $\operatorname{sat} \omega = 1$, $\omega > 1$). If $\beta(z_k) > 0$ an improvement is obtained on the kth step, i.e., $| z_{k+1} | < | z_k |$; if $\beta(z_k) = 0$, $z_k = z^*$ and the iterative process is finite, i.e., the solution has been obtained in k steps. From Fig. 1 it is also clear that $| z_k | \gamma(z_k) \leqq | z^* | \leqq | z_k |$. Thus on each step upper and lower bounds on $| z^* |$ may be computed. Notice that in applying the iterative procedure it is not necessary to know beforehand whether or not $0 \in K$. A more precise and complete statement of results is contained in the following theorem.

THEOREM 2. *Let $s(\cdot)$ be an arbitrary contact function of the set K specified in BP. Take $z_0 \in K$ and, by means of* (3.3) *with $\alpha_k \in I(z_k)$, generate the sequence $\{z_k\}$. Then for $k \geqq 0$ and $k \to \infty$:*

(i) $z_k \in K$,

(ii) *the sequence $\{| z_k |\}$ is decreasing and $| z_k | \to | z^* |$,*

(iii) $z_k \to z^*$,

(iv) $|z_k|\gamma(z_k) \leqq |z^*|$ *and* $|z_k|\gamma(z_k) \to |z^*|$,

(v) $|z_k - z^*| \leqq \sqrt{1-\gamma(z_k)}\,|z_k|$ *and* $\sqrt{1-\gamma(z_k)}\,|z_k| \to 0$,

(vi) $|s(-z_k) - z^*| \leqq |s(-z_k) - \gamma(z_k)z_k|$.

Since the bounds given in parts (iv), (v), and (vi) are computable as the iterative process proceeds, they may be used to generate stopping criteria for the termination of the iterative process. Example problems show $\{|z_k|\gamma(z_k)\}$ is not necessarily increasing. Thus $|z_k| - \max_{i\leqq k}|z_i|\gamma(z_i)$ is more satisfactory as an upper bound for $|z_k| - |z^*|$ than $|z_k| - |z_k|\gamma(z_k)$. Since examples also show that $\{|z_k - z^*|\}$ and $\{|s(-z_k) - z^*|\}$ are not necessarily decreasing, it is not possible to improve similarly the bounds given in (v) and (vi).

Suppose $|z^*| > 0$ and $s(\cdot)$ is continuous in a neighborhood of $-z^*$ (the latter is certainly implied if K is strictly convex). Then it follows from the continuity of $\gamma(\cdot)$ and (iii) that the upper bound in (vi) converges to zero. Thus $\{s(-z_k)\}$ may be used as an approximating sequence, an approach which may be advantageous in some situations (see §8). In addition it is clear from (iv) that

$$|s(-z_k)| - |z^*| \leqq |s(-z_k)| - \max_{i\leqq k}|z_i|\gamma(z_i),$$

where the right side converges to zero. Therefore meaningful stopping criteria are available.

4. Proof of Theorem 2. First, some basic inequalities are stated. From $z^* \in P(-z^*)$, $0 \notin K$, and $s(-y) \in P(-y)$, $y \neq 0$, it follows by the definition of $P(\cdot)$ that

$$z^*\cdot z^* \leqq z\cdot z^*, \quad z \in K, 0 \notin K; \tag{4.1}$$

$$s(-y)\cdot y \leqq z\cdot y, \quad z \in K, y \in E^n. \tag{4.2}$$

These inequalities lead to

$$|z^*|^2 \leqq s(-y)\cdot z^*, \quad 0 \notin K, y \in E^n; \tag{4.3}$$

$$s(-y)\cdot y \leqq z^*\cdot y, \quad y \in E^n; \tag{4.4}$$

$$s(-z)\cdot z \leqq |z|^2, \quad z \in K; \tag{4.5}$$

$$|y - z^*|^2 + z^*\cdot(y - z^*) \leqq y\cdot(y - s(-y)), \quad y \in E^n; \tag{4.6}$$

$$|z - z^*|^2 \leqq z\cdot(z - s(-z)), \quad z \in K, 0 \notin K. \tag{4.7}$$

Inequalities (4.3), (4.4), and (4.5) are deduced from (4.1) and (4.2) by obvious substitutions. From the identity

$$|y - z^*|^2 + z^* \cdot (y - z^*) + y \cdot (z^* - s(-y)) = y \cdot (y - s(-y)),$$

(4.6) follows by (4.4). Inequality (4.7) follows from (4.6) by use of (4.1).

Part (i) of the theorem depends on $\alpha_k \in I(z_k)$ which insures $0 \leqq \alpha_k \leqq 1$. Thus from (3.3), $s(-z_k) \in K$, and the convexity of K, $z_k \in K$ implies $z_{k+1} \in K$.

Consider now the inequalities in (iv), (v), and (vi). From (4.4) and the Schwarz inequality, $s(-y) \cdot y \leqq |y| \cdot |z^*|$. Thus (iv) follows from (3.2). The proof of the inequalities in (v) and (vi) makes use of $z = z_k \in K$. For $s(-z) \cdot z > 0$, $z \cdot (z - s(-z)) = |z|^2 (1 - \gamma(z))$ and from (4.7) the inequality in (v) is true. Now consider $s(-z) \cdot z \leqq 0$, which corresponds to $\gamma(z) = 0$. For $z^* = 0$, $\gamma(z) = 0$ (Theorem 1) and (v) holds as an equality; for $z^* \neq 0$, the inequality in (v) follows from (4.1) which insures

$$|z - z^*|^2 = |z|^2 - 2z \cdot z^* + 2|z^*|^2 - |z^*|^2 \leqq |z|^2.$$

If $z^* = 0$ the inequality in (vi) is trivially true. Consider now $z^* \neq 0$. If $s(-z) \cdot z \leqq 0$, (vi) reduces to $-2s(-z) \cdot z^* + |z^*|^2 \leqq 0$ which is true by (4.3). The following identity is easily verified:

$$|s(-z) - z^*|^2 = |s(-z) - \gamma(z)z|^2 + |z|^{-2}(s(-z) \cdot z)^2 + |z^*|^2 - 2s(-z) \cdot z^*.$$

Assuming $s(-z) \cdot z > 0$ and using $s(-z) \cdot z \leqq |z| \cdot |z^*|$ yields $|z|^{-2}(s(-z) \cdot z)^2 \leqq |z^*|^2$. Thus

$$|s(-z) - z^*|^2 \leqq |s(-z) - \gamma(z)z|^2 + 2(|z^*|^2 - s(-z) \cdot z^*)$$

and by (4.3) the inequality in (vi) follows.

In order to complete the proof of the theorem, the function

$$\Gamma(z) = |z|^2 - |z^*|^2 = |z - z^*|^2 + 2z^* \cdot (z - z^*) \tag{4.8}$$

is introduced. For $0 \notin K$ inequality (4.1) gives

$$0 \leqq |z - z^*|^2 \leqq \Gamma(z), \quad z \in K, \tag{4.9}$$

a result which is obviously true for $0 \in K$. In the following paragraphs it will be shown that $\{\Gamma(z_k)\}$ is decreasing and $\Gamma(z_k) \to 0$. By (4.8) and (4.9) this proves (ii) and (iii). The remaining results in (iv) and (v) follow from the known value of $\gamma(z^*)$, the continuity of $\gamma(\cdot)$, and (iii).

For simplicity let

$$\Delta(z; \alpha) = \Gamma(z) - \Gamma(z + \alpha(s(-z) - z)), \tag{4.10}$$

and assume tacitly in what follows that $z \in K$. Then from (4.8),

$$\Delta(z; \alpha) = 2\alpha(|z|^2 - s(-z) \cdot z) - \alpha^2 |z - s(-z)|^2. \tag{4.11}$$

Because the coefficient of α^2 is not positive, $\min_{\alpha \in I(z)} \Delta(z; \alpha)$ is attained at one of the end points of $I(z)$. It is readily shown that

$$\Delta(z; \delta\beta(z)) = \Delta(z; (2 - \delta)\beta(z)).$$

Thus from the definition of $I(z)$,

$$\min_{\alpha \in I(z)} \Delta(z; \alpha) = \begin{cases} \Delta(z; \delta\beta(z)) & \text{if } \beta(z) \leqq \delta^{-1}, \\ \Delta(z; 1) & \text{if } \beta(z) \geqq \delta^{-1}. \end{cases} \tag{4.12}$$

Equation (4.12) is now used to obtain a lower bound on $\Delta(z; \alpha)$, $\alpha \in I(z)$. From (4.11) and (3.1) it follows that

$$\Delta(z; \delta\beta(z)) = |z - s(-z)|^{-2}[z \cdot (z - s(-z))]^2(2\delta - \delta^2). \tag{4.13}$$

Let

$$\mu = \max_{z_1, z_2 \in K} |z_1 - z_2| \tag{4.14}$$

denote the diameter of K and recall that $0 < \delta \leqq 1$. Then

$$\Delta(z; \delta\beta(z)) \geqq \mu^{-2}\delta[z \cdot (z - s(-z))]^2. \tag{4.15}$$

From (4.8) and (4.6),

$$\Gamma(z) \leqq 2|z - z^*|^2 + 2z^* \cdot (z - z^*) \leqq 2z \cdot (z - s(-z)) \tag{4.16}$$

(for $z^* = 0$ this may be sharpened to $\Gamma(z) \leqq z \cdot (z - s(-z))$). Thus

$$\Delta(z; \delta\beta(z)) \geqq \tfrac{1}{4}\mu^{-2}\delta\Gamma^2(z). \tag{4.17}$$

For $\beta(z) \geqq 1$, $z \cdot (z - s(-z)) \geqq |z - s(-z)|^2$ and consequently

$$\Delta(z; 1) = 2z \cdot (z - s(-z)) - |z - s(-z)|^2 \geqq z \cdot (z - s(-z)).$$

Therefore (4.16) yields

$$\Delta(z; 1) \geqq \tfrac{1}{2}\Gamma(z), \quad \beta(z) \geqq 1. \tag{4.18}$$

Finally, utilizing (4.17) and (4.18) in (4.12) yields

$$\Delta(z; \alpha)|_{\alpha \in I(z)} \geqq \min\{\tfrac{1}{4}\mu^{-2}\delta\Gamma^2(z), \tfrac{1}{2}\Gamma(z)\}. \tag{4.19}$$

Letting $z = z_k$ in (4.19), using (3.3), and returning to (4.10), it is seen that

$$\Gamma(z_k) - \Gamma(z_{k+1}) \geqq \min\{\tfrac{1}{4}\mu^{-2}\delta\Gamma^2(z_k), \tfrac{1}{2}\Gamma(z_k)\} \geqq 0. \tag{4.20}$$

Therefore the sequence $\{\Gamma(z_k)\}$ is decreasing and, since it is bounded from below by zero, has a limit point. Thus passing to the limit on the left side of (4.20) gives zero and therefore from the right side $\Gamma(z_k) \to 0$.

5. Nature of convergence. This section gives further results on the

convergence of the iterative procedure. Theorem 3 establishes upper bounds on the elements of the sequences $\{|\, z_k \,|\}$ and $\{|\, z_k - z^* \,|\}$. Several example problems are analyzed to demonstrate still more fully the nature of convergence. Finally, a few numerical results are given. Emphasis is on the case $0 \notin K$, since it appears that it is most important in applications.

THEOREM 3. *Let*

$$\theta_k = \theta_0(1 + \tfrac{1}{4}\mu^{-2}\delta\theta_0 k)^{-1}, \quad \theta_0 = |\, z_0 \,|^2 - |\, z^* \,|^2, \tag{5.1}$$

and assume that $|\, z_0 \,|^2 \leqq |\, z^* \,|^2 + 2\mu^2\delta^{-1}$. *Then if* $\{z_k\}$ *is generated by the iterative procedure, the following inequalities hold for* $k \geqq 0$:

$$|\, z_k \,| \leqq \sqrt{\theta_k + |\, z^* \,|^2}, \tag{5.2}$$

$$|\, z_k - z^* \,| \leqq \sqrt{\theta_k}\,. \tag{5.3}$$

The assumption on $|\, z_0 \,|$ is often met in practice. For example, it is easily shown that it must be satisfied if $|\, z^* \,| \leqq \tfrac{1}{2}(2\delta^{-1} - 1)\mu$. In any case, z_0 may be interpreted as a suitable intermediate point in the iterative process, and inequalities (5.2) and (5.3) may be used to estimate the subsequent rate of convergence.

For $|\, z^* \,| > 0$ and $k \geqq 1$ inequalities (5.2) and (5.3) imply

$$|\, z_k \,| - |\, z^* \,| < 2\mu^2\, |\, z^* \,|^{-1}\delta^{-1}k^{-1}, \tag{5.4}$$

$$|\, z_k - z^* \,| < 2\mu\delta^{-1/2}k^{-1/2}, \tag{5.5}$$

results which conform closely to (5.2) and (5.3) for k sufficiently large. In Examples 1 and 2, which appear later in this section, it is demonstrated that within a constant multiplicative factor it is impossible to obtain bounds on $|\, z_k \,| - |\, z^* \,|$ and $|\, z_k - z^* \,|$ which approach zero more rapidly than those given in (5.4) and (5.5).

Proof of Theorem 3. Since $|\, z_0 \,|^2 \leqq |\, z^* \,|^2 + 2\mu^2\delta^{-1}$, it follows from the previous section that $\Gamma(z_k) \leqq \Gamma(z_0) \leqq 2\mu^2\delta^{-1}$, $k \geqq 0$. From (4.20) this implies

$$\Gamma(z_{k+1}) \leqq \Gamma(z_k) - \tfrac{1}{4}\mu^{-2}\delta\Gamma^2(z_k), \quad k \geqq 0.$$

Since

$$1 - \tfrac{1}{4}\mu^{-2}\delta\Gamma \leqq (1 + \tfrac{1}{4}\mu^{-2}\delta\Gamma)^{-1}$$

for all $\Gamma \geqq 0$, it is possible to write

$$\Gamma(z_{k+1}) \leqq \Gamma(z_k)(1 + \tfrac{1}{4}\mu^{-2}\delta\Gamma(z_k))^{-1}, \quad k \geqq 0. \tag{5.6}$$

But substitution shows that θ_k is the solution of

$$\theta_{k+1} = \theta_k(1 + \tfrac{1}{4}\mu^{-2}\delta\theta_k)^{-1}, \tag{5.7}$$

with $\theta_0 = |z_0|^2 - |z^*|^2 = \Gamma(z_0)$. Thus comparison of (5.6) and (5.7) yields $\Gamma(z_k) \leqq \theta_k$, $k \geqq 0$. Finally, (5.2) and (5.3) follow from (4.8) and (4.9).

The complexity of the difference equation (3.3) makes it difficult to obtain more specific analytic results than those obtained in Theorem 3. Thus the remainder of this section is limited to the presentation and discussion of three, somewhat specialized, example problems and a few numerical results.

Example 1. Take $\delta = 1$ and let K be the convex hull of three points in 2-space, $(1, \nu)$, $(-1, \nu)$, $(0, 1 + \nu)$, where $\nu > 0$. Clearly $z^* = (0, \nu)$ and $|z^*| = \nu$. Simple inspection shows that the iterative process is finite ($z_1 = z^*$) if and only if z_0 is on the line segment connecting $(1, \nu)$ and $(-1, \nu)$. Moreover when the process is not finite, z_k, $k \geqq 1$, is determined by the scalar $\psi_k = |z_k^1|(z_k^2)^{-1}$. Thus the second order nonlinear difference equation (3.3) may be replaced by a first order difference equation in ψ_k. It is not difficult to show that

$$\psi_{k+1} = \psi_k(1 - \nu\psi_k)(1 + \nu\psi_k + 2\psi_k^2)^{-1}, \quad k \geqq 1. \tag{5.8}$$

For $\psi_k \ll 1$ this equation is approximated by $\tilde{\psi}_{k+1} = \tilde{\psi}_k(1 + 2\nu\tilde{\psi}_k)^{-1}$, an equation of the same form as (5.7). These observations and some tedious, but straightforward, computations lead to (the notation $o(\omega)$ means $\lim_{\omega \to 0} \omega^{-1} o(\omega) = 0$)

$$|z_k| - |z^*| = (2\nu k)^{-1} + o(k^{-1}), \tag{5.9}$$

$$|z_k - z^*| = (2\nu k)^{-1}\sqrt{1 + \nu^2} + o(k^{-1}). \tag{5.10}$$

Equation (5.9) demonstrates that it is impossible to obtain an upper bound on $|z_k| - |z^*|$ which approaches zero more rapidly than (const.)k^{-1}. For large k the upper bound in (5.4) is conservative by a factor of sixteen. This factor can be traced to two sources each of which contributes a factor of four: in (4.15), μ is an unsatisfactory estimate of $|z_k - s(-z_k)|$, in the derivation of (4.6) the term $y \cdot (z^* - s(y))$ has been omitted. For this example the upper bound in (5.5) is a poor estimate because it is order $k^{-1/2}$ rather than order k^{-1}.

It is also possible to show that

$$|z^*| - \gamma(z_k)|z_k| = (2\nu k)^{-1} + o(k^{-1}), \tag{5.11}$$

$$\sqrt{1 - \gamma(z_k)}\,|z_k| = k^{-1/2} + o(k^{-1/2}). \tag{5.12}$$

By comparing (5.11) with (5.9) and (5.12) with (5.10) it is seen that in Theorem 2, part (iv) provides a reasonably good stopping criterion while (v) does not.

Example 2. Take $\delta = 1$ and let K be the convex hull of three points in

3-space, $(1, 0, \nu)$, $(-1, 0, \nu)$, $(0, 1, \nu)$, where $\nu > 0$. Thus $z^* = (0, 0, \nu)$ and $|z^*| = \nu$. The iterative process is much the same as in Example 1, the points $z_k \in K$, $k \geqq 1$, being determined by the scalar $\psi_k = |z_k^1|(z_k^2)^{-1}$. The first order difference equation for ψ_k is (5.8) with $\nu = 0$. By using the fact that $\tilde{\psi}_k = \tilde{\psi}_0(1 + 4\tilde{\psi}_0^2 k)^{-1/2}$ is the solution of $\tilde{\psi}_{k+1} = \tilde{\psi}_k(1 + 4\tilde{\psi}_k^2)^{-1/2}$ and that $(1 + 4\tilde{\psi}_k^2)^{1/2} \cong 1 + 2\psi_k^2$ for $\psi_k \ll 1$, the following results can be derived:

$$|z_k| - |z^*| = (8\nu k)^{-1} + o(k^{-1}), \tag{5.13}$$

$$|z_k - z^*| = (2k^{1/2})^{-1} + o(k^{-1/2}), \tag{5.14}$$

$$|z^*| - \gamma(z_k)|z_k| = 3(8\nu k)^{-1} + o(k^{-1}), \tag{5.15}$$

$$\sqrt{1 - \gamma(z_k)}\,|z_k| = (2k)^{-1/2} + o(k^{-1/2}). \tag{5.16}$$

Equation (5.14) shows that the asymptotic behavior of $|z_k - z^*|$ matches the bound given in (5.5), except for a multiplicative factor of eight. The bound given in (5.4) is conservative by a multiplicative factor of 64. Comparison of (5.15) with (5.13) and (5.16) with (5.14) shows that (iv) and (v) of Theorem 2 both provide reasonable stopping criteria.

Example 3. Take $\delta = 1$ and in n-space let

$$K = \{z \mid z^1 \geqq \nu + \frac{1}{2}\sum_{i=2}^{n} (z^i)^2\lambda_i^{-1}, z^1 \leqq 2\nu\}, \qquad \nu, \lambda_2, \cdots, \lambda_n > 0. \tag{5.17}$$

In the neighborhood of $z^* = (\nu, 0, 0, \cdots, 0)$, ∂K is the elliptic hyperparaboloid

$$z^1 = \nu + \frac{1}{2}\sum_{i=2}^{n} (z^i)^2\lambda_i^{-1},$$

where $\lambda_2, \cdots, \lambda_n$ are the principal radii of curvature at the vertex z^*. For many convex sets K, ∂K in the neighborhood of z^* may be closely approximated by such an elliptic hyperparaboloid. Thus this example is of more general interest than the previous examples.

For $y^1 < 0$ and

$$\frac{1}{2}\sum_{i=2}^{n} (y^1)^{-2}\lambda_i(y^i)^2 < \nu,$$

it is easy to show that

$$\begin{aligned} s^1(y) &= \nu + \frac{1}{2}\sum_{i=2}^{n} (y^1)^{-2}\lambda_i(y^i)^2, \\ s^i(y) &= -(y^1)^{-1}\lambda_i y^i, \qquad i = 2, \cdots, n. \end{aligned} \tag{5.18}$$

Let $\bar{\lambda} = \max_{i=2,\cdots,n}\{\lambda_i\}$ and assume the conditions

$$\zeta = \nu\sqrt{1 + 2\nu\bar{\lambda}^{-1}} > |y|, \quad -\nu \geqq y^1 \tag{5.19}$$

are satisfied, which in turn imply

$$\frac{1}{2}\sum_{i=2}^{n} (y^1)^{-2}\lambda_i (y^i)^2 < \nu.$$

Thus (5.19) defines a set on which (5.18) is valid. Using this fact, $z^1 \geqq \nu$ for $z \in K$, and (3.1) gives

$$\beta(z) = \Bigg((z^1 - \nu)^2 + \nu(z^1 - \nu) + \sum_{i=2}^{n}\left(1 + \frac{1}{2}(z^1)^{-1}\lambda_i\right)(z^i)^2\Bigg)\Big/$$

$$\Bigg((z^1 - \nu)^2 + \sum_{i=2}^{n}[1 + (z^1)^{-1}\lambda_i + (z^1)^{-2}\nu\lambda_i + (z^1)^{-2}\lambda_i^2] \cdot (z^i)^2 + \left[\frac{1}{2}\sum_{i=2}^{n}(z^1)^{-2}\lambda_i (z^i)^2\right]^2\Bigg), \quad z \neq z^*, z \in K, |z| < \zeta. \tag{5.20}$$

Because $z \in K$, $|z| < \zeta$ imply $z^1 \geqq \nu$ and

$$\frac{1}{2}\sum_{i=2}^{n}(z^1)^{-2}\lambda_i (z^i)^2 < \nu,$$

it follows that

$$\begin{aligned}\beta(z) &\geqq \frac{(z^1 - \nu)^2 + \sum_{i=2}(z^i)^2}{(z^1 - \nu)^2 + (1 + \frac{5}{2}\bar{\lambda}\nu^{-1} + \bar{\lambda}^{-2}\nu^{-2})\sum_{i=2}^{n}(z^i)^2} \\ &\geqq \frac{1}{1 + \frac{5}{2}\bar{\lambda}\nu^{-1} + \bar{\lambda}^{-2}\nu^{-2}} = \underline{\beta}, \qquad z \neq z^*, z \in K, |z| < \zeta.\end{aligned} \tag{5.21}$$

Because $\beta(z^*) = 0$ this inequality implies that $\beta(z)$ is discontinuous on K at z^*.

By starting with (4.13) and repeating the derivation of §4 with

$$[z \cdot (z - s(-z))]\,|z - s(-z)|^{-2} = \beta(z) \geqq \underline{\beta},$$

it can be shown that

$$\Gamma(z_{k+1}) \leqq \Gamma(z_k)(1 - \tfrac{1}{2}\underline{\beta}\delta), \quad z_k \neq z^*. \tag{5.22}$$

For $z_k = z^*$, $\Gamma(z_{k+1}) = 0$ and (5.22) is trivially true. Thus

$$\Gamma(z_k) \leqq \Gamma(z_0)(1 - \tfrac{1}{2}\underline{\beta}\delta)^k, \quad k \geqq 0, \quad z_0 \in K, \quad |z_0| < \zeta.$$

Using (4.8) and (4.9) this leads to

$$|z_k| - |z^*| \leqq \tfrac{1}{2}\nu^{-1}\theta_0(1 - \tfrac{1}{2}\underline{\beta}\delta)^k, \tag{5.23}$$

$$|z_k - z^*| \leqq \sqrt{\theta_0}(1 - \tfrac{1}{2}\underline{\beta}\delta)^{k/2}, \tag{5.24}$$

TABLE 1

Number of iterations to satisfy error criteria

First column: first k for which $|z_k| - |z^*| \leqq \epsilon$

Second column: first k for which $|z^*| - \gamma(z_k)|z_k| \leqq \epsilon$

Case / λ_2 / λ_3 / ϵ	1		2		3		4		5		6	
	1		10		100		1000		100		1000	
	1		10		100		1000		10		10	
10^{-3}	3	2	28	20	59	18	216	83	27	14	229	82
10^{-4}	5	4	31	27	59	35	250	88	52	26	267	125
10^{-5}	5	4	38	30	74	58	290	162	73	51	298	167
10^{-6}	5	4	41	37	111	58	340	215	81	51	359	218

where θ_0 is given as before in (5.1). Since $\beta > 0$ inequalities (5.23) and (5.24) guarantee that the convergence of $\{|z_k|\}$ and $\{|z_k - z^*|\}$ is geometric. However, the guaranteed rate of convergence is not rapid if $\beta \ll 1$, i.e., $\nu^2 \ll \bar{\lambda}^2$.

Table 1 presents some numerical results for Example 3, $\nu = 1$, $n = 3$, and $z_0 = (6, 2, 2)$. Similar results are obtained for different z_0. The extent of K has been increased beyond $z^1 = 2\nu$ so that (5.18) is valid even though (5.19) is violated. Note that convergence is slow when $\nu \ll \bar{\lambda}$. Although the bounds derived in the preceding paragraph follow the same pattern it may be concluded from Table 1 that they are not sharp estimates of actual convergence rate. Better estimates than (5.23) and (5.24) have been obtained but their derivation is too lengthy to present here. It is interesting to note that Cases 3 and 4 exhibit rates of convergence which are respectively similar to Cases 5 and 6. Thus $\nu/\bar{\lambda}$ seems to be the key parameter while λ_3/λ_2 has little effect. This is not true when the gradient method of the next section is used ($\lambda_3 \gg \lambda_2$ corresponds to a "ridge" of $f(y)$).

Fig. 2 shows the details of Case 5. The irregularity of the sequences shown is typical. Various methods for accelerating convergence (based on different rules for selecting $\alpha_k \in I(z_k)$, the results of the next section, etc.) are being investigated and will be reported in a later paper.

6. Relation to a gradient method. The iterative procedure described in the preceding sections is related to a gradient method, which is similar in approach to certain gradient based methods which have been proposed for the solution of a variety of problems in optimal control [2], [4], [9], [10], [11]. The purpose of this section is to illustrate both the differences

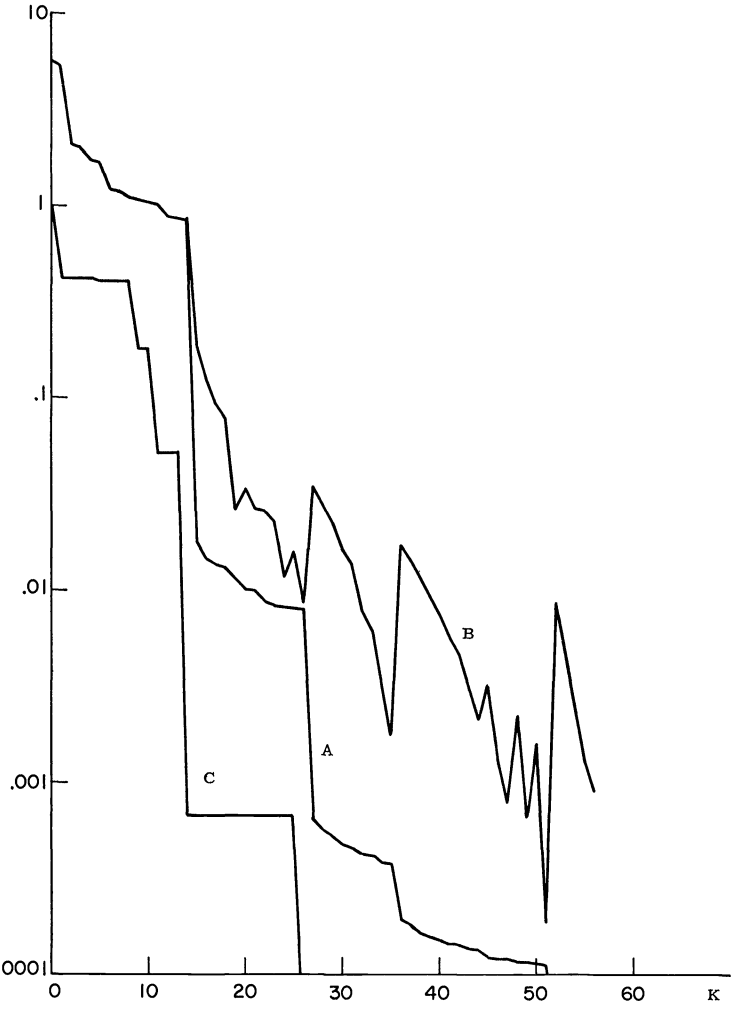

FIG. 2. *Numerical results for Case 5 of §6:* (A) $|z_k| - |z^*|$, (B) $|z_k - z^*|$, (C) $|z^*| - \max_{i \leqq k} |z_i| \gamma(z_i)$. *For* $k \leqq 14$, $|z_k - z^*| \cong |z_k| - |z^*|$.

and strong connections between the two approaches. For brevity the developments which follow are presented somewhat superficially and without proof.

THEOREM 4. *Assume* $0 \notin K$ *and let* $J = \{y \mid y \cdot s(-y) > 0\}$. *Then for* $y \in J$ *the scalar function*

$$f(y) = |y|^{-1}(y \cdot s(-y)) = \gamma(y)|y| \tag{6.1}$$

is defined and has the following properties:

(i) $0 < f(y) \leqq |z^*|$,

(ii) $f(y) = |z^*|$ *if and only if* $y = \rho z^*$, $\rho > 0$.

Further assume that K *is strictly convex. Then*:

(iii) $s(-\rho z^*) = z^*$, $\rho > 0$,

(iv) *the gradient of* $f(y)$ *exists and is given by*

$$\nabla f(y) = |y|^{-1} s(-y) - |y|^{-3}(y \cdot s(-y))y,$$

(v) $\nabla f(y) = 0$ *if and only if* $y = \rho z^*$, $\rho > 0$.

Theorem 4 forms the basis for the gradient method. A sequence of vectors $\{y_k\}$ is generated by

$$y_{k+1} = y_k + \sigma_k \nabla f(y_k), \quad y_0 \in J. \tag{6.2}$$

If K is strictly convex, $0 \notin K$, and the positive numbers σ_k are appropriately chosen, it can be shown that $y_k \in J$, $k \geqq 0$, $\{f(y_k)\}$ is increasing, and $y_k \to \rho z^*$, $\rho > 0$, for $k \to \infty$. Strict convexity of K also assures that $s(y)$ is continuous on J. This, $s(\omega y) = s(y)$ for $\omega > 0$, and solution property (iv) (§2) guarantee that $\{s(-y_k)\}$ is an approximating sequence for z^*, i.e., $s(-y_k) \to z^*$. Disadvantages of the gradient method, relative to the procedure of §3, are: K must be strictly convex, methods for choosing the values of σ_k may be cumbersome and time consuming, the selection of a y_0 in J may be difficult. On the other hand it is conceivable that the gradient method may yield more rapid convergence, particularly when variations of (6.2) are employed.

Consider now a modified version of the gradient method. Since from (iv) of Theorem 4, $\nabla f(\rho^{-1}y) = \rho \nabla f(y)$, $\rho > 0$, the difference equation

$$z_{k+1} = \rho_{k+1}\rho_k^{-1}(z_k + \sigma_k \rho_k^{\,2} \nabla f(z_k)), \quad z_0 = y_0 \in J, \tag{6.3}$$

with $\rho_0 = 1$ and $\rho_k > 0$, $k > 0$, yields a sequence $\{z_k\}$ such that $z_k = \rho_k y_k$, $k \geqq 0$. Thus $s(-y_k) = s(-z_k)$, $k \geqq 0$. By letting

$$\rho_{k+1} = \rho_k[1 + \sigma_k \rho_k^{\,2} |z_k|^{-1}(1 - \gamma(z_k))]$$

and

$$\alpha_k = \sigma_k \rho_k \rho_{k+1} |z_k|^{-1}, \; k \geqq 0, \tag{6.4}$$

it is easy to show that (6.3) becomes (3.3). Thus if $z_0 \in K \cap J$, (3.3) realizes the modified version of the gradient method, where the selection rule for α_k is (6.4) rather than (3.4). If the α_k as obtained from (6.4) happen to be in $I(z_k)$, $k \geqq 0$, then all the results of Theorem 2 follow; in particular $\{z_k\}$, whose elements are in K, is also an approximating sequence.

In any case the iterative procedure described in §3 takes "steps" in the same direction as those indicated by the modified gradient method. The "step size" prescribed by (3.4) may be much larger than that prescribed by (6.4). Thus with (3.4) the sequence $\{f(z_k)\}$ is not necessarily increasing.

7. Extension to more general quadratic forms. The iterative procedure for the Basic Problem can be extended without great difficulty to the general problem:

GP. *Given C, a compact convex set in E^m, and the quadratic form*

$$q(x) = |x|_G^2 + g\cdot x, \tag{7.1}$$

where $|x|_G^2 = x\cdot Gx$, G is a symmetric nonnegative definite $m \times m$ matrix, and g is an m-vector in the range of G, find a point $x^* \in C$ such that

$$q(x^*) = q^* = \min_{x\in C} q(x).$$

Clearly a solution x^* exists. In order to obtain its essential properties and derive the iterative procedure it is convenient to write $q(x)$ as

$$q(x) = |Hx - a|^2 + q_0, \tag{7.2}$$

where H is an $n \times m$ matrix, $n = \operatorname{rank} G$, $G = H'H$ (the $'$ denotes matrix transpose), $a = \frac{1}{2}(HH')^{-1}Hg$ or equivalently $g = 2H'a$, and

$$q_0 = -|a|^2 = \min_{x\in E^m} q(x).$$

The existence of such a representation is a consequence of the hypotheses in the statement of GP. Introducing the set $K = \{z \mid z = Hx + a, x \in C\}$ it is clear that

$$q^* = \min_{z\in K} |z|^2 + q_0 = |z^*|^2 + q_0, \tag{7.3}$$

where z^* is defined as before. Furthermore since $z^* \in K$ is unique it follows that $F = \{x \mid Hx + a = z^*, x \in C\}$ is the set of all solutions of GP. Since F may sometimes contain more than a single point, x^* is not necessarily unique.

The iterative procedure for GP is developed from the results of §3 by noting that for every point $x \in C$ there is, by means of

$$z = Hx + a, \tag{7.4}$$

a corresponding point $z \in K$. Thus, for example,

$$\max_{z\in K} y\cdot z = \max_{x\in C} y\cdot(Hx + a) = s_C(H'y)\cdot H'y + y\cdot a = y\cdot(Hs_C(H'y) + a),$$

where $s_C(\cdot)$ is a contact function of C. Therefore a contact function of K is

$$s(y) = Hs_C(H'y) + a. \tag{7.5}$$

Using this result and $H'(Hx + a) = Gx + \frac{1}{2}g$, it is further seen that the equation

$$x_{k+1} = x_k + \alpha_k(s_C(-Gx_k - \tfrac{1}{2}g) - x_k) \tag{7.6}$$

when transformed by (7.4) yields the same sequence as (3.3). Hence if $\alpha_k \in I(Hx_k + a)$ and $x_0 \in C$, (7.6) yields a sequence $\{x_k\}$ with elements in C such that $q(x_k)$ converges downward to q^*. This and other results are summarized in the following.

THEOREM 5. *Let $s_C(\cdot)$ be a contact function of the set C specified in GP. Define*

$$v_k = Gx_k + \tfrac{1}{2}g; \tag{7.7}$$

$$\beta_k = \begin{cases} |x_k - s_C(-v_k)|_G^{-2}[v_k \cdot (x_k - s_C(-v_k))] \text{ if } G(x_k - s_C(-v_k)) \neq 0, \\ 0 \text{ if } G(x_k - s_C(-v_k)) = 0; \end{cases} \tag{7.8}$$

$$\gamma_k = \begin{cases} \max\{(1 - (q(x_k) - q_0)^{-1}[v_k \cdot (x_k - s_C(-v_k))]), 0\} \\ \qquad\qquad \text{if } q(x_k) \neq q_0, \\ 0 \text{ if } q(x_k) = q_0; \end{cases} \tag{7.9}$$

$$I_k = [\min\{\delta\beta_k, 1\}, \min\{(2 - \delta)\beta_k, 1\}], \quad 0 < \delta \leqq 1; \tag{7.10}$$

$$x_{k+1} = x_k + \alpha_k(s_C(-v_k) - x_k), \quad x_0 \in C, \quad \alpha_k \in I_k. \tag{7.11}$$

By means of (7.11) *generate* $\{x_k\}$. *Then* $\beta_k \geqq 0$, $k \geqq 0$; *and* $\beta_k = 0$ *implies* $x_k \in F$. *Furthermore, for* $k \geqq 0$ *and* $k \to \infty$:

(i) $x_k \in C$,

(ii) $\{q(x_k)\}$ *is decreasing and* $q(x_k) \to q^*$,

(iii) *there is a convergent subsequence of* $\{x_k\}$ *and every convergent subsequence of* $\{x_k\}$ *has its limit point in* F,

(iv) $\gamma_k^2 q(x_k) + (1 - \gamma_k^2)q_0 \leqq q^*$ *and* $\gamma_k^2 q(x_k) + (1 - \gamma_k^2)q_0 \to q^*$,

(v) $|x_k - \tilde{x}|_G^2 \leqq (1 - \gamma_k)(q(x_k) - q_0)$ *for all* $\tilde{x} \in F$ *and* $(1 - \gamma_k)(q(x_k) - q_0) \to 0$,

(vi) $|s_C(-v_k) - \tilde{x}|_G^2 \leqq |s_C(-v_k) - \gamma_k x_k|_G^2 + (1 - \gamma_k)g \cdot (s_C(-v_k) - \gamma_k x_k) - (1 - \gamma_k)^2 q_0$ *for all* $\tilde{x} \in F$.

Proof. Part (iii) follows from (i), (v), the definition of F, and the compactness of C. The remaining parts follow from Theorems 1 and 2 by straightforward substitutions.

If C is strictly convex and $q^* > q_0$, the upper bound in (vi) converges to zero and $\{s_C(-v_k)\}$ serves as an approximating sequence (see remarks

after Theorem 2). Also the results of §§5, 6 may be extended in an obvious way. For most applications the hypothesis that g is in the range of G holds. When it does not hold, by a different line of attack it is still possible to derive a theorem similar to Theorem 5.

The approach taken by Frank and Wolfe [6] to the concave programming problem can be extended to give a direct proof of (i), (ii), and (iii) of Theorem 5. A lower bound for $q(x_k)$ is also obtainable but it is not as sharp as (iv).

8. Application to problem in optimal control. Consider the dynamical system

$$\dot{x} = A(t)x + f(u(t); t), \quad x(0), \tag{8.1}$$

where x is the m-dimensional state vector, $\dot{x}$ is its time derivative, $x(0)$ is the initial state; $u(t)$ is an r-dimensional vector control function, admissible if measurable on the control interval $[0, T]$, $0 < T < \infty$, with range in a compact set U; $A(t)$ is an $m \times m$ matrix function continuous on $[0, T]$; $f(\cdot\,; \cdot)$ is an m-dimensional vector function defined and continuous on $U \times [0, T]$. For every admissible control $u(t)$ there is an absolutely continuous solution function, $x_{u(t)}(t) = x_u(t)$, which satisfies (8.1) almost everywhere in $[0, T]$. It is desired to find an admissible control $u^*(t)$ such that $q(x_{u^*}(T)) = q^* \leqq q(x_u(T))$ for all admissible controls $u(t)$, where $q(\cdot)$ is prescribed in GP, §7. This optimal control problem has a number of practical applications [1].

Under the conditions just stated, Neustadt [12] has shown that the set

$$C = \{x \mid x = x_u(T), u(t) \text{ admissible}\}$$

is compact and convex. Thus if a method for evaluating a contact function of C exists, the iterative procedure of §7 can be used to obtain approximations for $x_{u^*}(T) = x^*$ and q^*.

To obtain a contact function of C, $s_C(\cdot)$, note that

$$w \cdot x_u(T) = \psi(0; w) \cdot x(0) + \int_0^T \psi(\sigma; w) f(u(\sigma); \sigma)\, d\sigma, \tag{8.2}$$

where $\psi(t; w)$, defined on $[0, T] \times E^n$, is the solution of the adjoint differential equation

$$\dot{\psi} = -A'(t)\psi, \quad \psi(T) = w. \tag{8.3}$$

Equation (8.2) follows from (8.1) by integrating $\dfrac{d}{dt}(\psi(t; w) \cdot x_u(t))$. Suppose there exists an admissible control $u(t; w)$ such that almost everywhere in $[0, T]$,

$$\psi(t;w)\cdot f(u(t;w),t) = \max_{\tilde{u}\in U} \psi(t;w)\cdot f(\tilde{u},t). \tag{8.4}$$

Then from (8.2) it is clear that $w\cdot x_{u(t;\,w)}(T) \geqq w\cdot x_{u(t)}(T)$ for every admissible control $u(t)$. Thus from the definition of C a contact function of C is

$$s_C(w) = x_{u(t;\,w)}(T). \tag{8.5}$$

This result agrees with the well-known fact that boundary points of the reachable set C must "satisfy" the Pontryagin maximum principle. For all but the most elementary systems (8.1), $s_C(\cdot)$ is the only reasonable means for numerically characterizing the set C.

In most practical problems it is not difficult to obtain a function $u(t;w)$ which satisfies (8.4). Consider, for example, the case where $f(u;t) = B(t)u$, $B(t)$ is an $m \times r$ matrix function continuous on $[0, T]$, and U is the unit hypercube $\{u \mid |u^i| \leqq 1,\ i = 1, \cdots, r\}$. Notice that (8.4) may not uniquely define $u(t;w)$ almost everywhere in $[0, T]$; suppose for instance that in the example of the preceding sentence $B'(t)\psi(t;w)$ has at least one component which is identically zero on $[0, T]$. This is of no concern, since different choices for $u(t;w)$ will at most lead only to different contact functions of C. Previous computational procedures [2], [4], [9], [10], [11] have required assumptions which correspond to a unique determination of $u(t;w)$ by (8.4). Such "unique maximum" assumptions imply strict convexity of C.

Computer evaluation of $s_C(\cdot)$ entails three steps: evaluation of $\psi(t;w)$ by solving (8.3) backwards from $t = T$ to $t = 0$, determination of $u(t;w)$ from $\psi(t;w)$ by (8.4), solution of (8.1) with $u(t) = u(t;w)$ from $t = 0$ to $t = T$. Thus when the iterative procedure is applied to the optimal control problem each iteration involves the sequential solution of two differential equations. This situation is handled efficiently by a hybrid computer which includes both digital and analog elements.

The details of applying the iterative procedure should be clear. There is no difficulty in choosing $x_0 \in C$, it is only necessary to set $x_0 = x_{u^0}(T)$ where $u^0(t)$ is an arbitrary admissible control. In the sense of Theorem 5, $\{x_k\}$ and $\{q(x_k)\}$ (and if C is strictly convex and $q^* > q_0$, $\{s_C(-v_k)\}$ and $\{q(s_C(-v_k))\}$) are approximating sequences and error bounds may be computed.

The issue of finding admissible control functions corresponding to x_k or $s_C(-v_k)$ remains. The control corresponding to $s_C(-v_k)$ is $u(t;-v_k)$, i.e.,

$$s_C(-v_k) = x_{u(t;\,-v_k)}(T).$$

Finding an admissible control which produces the terminal state x_k is

more difficult. From (7.11) it follows that

$$x_k = \sum_{i=1}^{k-1} \lambda_i s_C(-v_i) + \lambda_0 x_0 ,$$

where $\lambda_i \geqq 0$, $0 \leqq i < k$, and $\sum_{i=0}^{k-1} \lambda_i = 1$. Suppose $u_k(t)$ is an admissible control such that almost everywhere in $[0, T]$,

$$f(u_k(t), t) = \sum_{i=1}^{k-1} \lambda_i f(u(t; -v_i); t) + \lambda_0 f(u^0(t); t).$$

Then from the form of (8.1) it may be deduced that $x_{u_k(t)}(T) = x_k$. If for all $t \in [0, T]$ the sets $f(U; t)$ are convex such a choice is possible. If this is not the case an additional approximation process, the construction of a chattering control, is necessary [1]. For $f(u, t) = B(t)u$ and U convex it follows that

$$u_k(t) = \sum_{i=1}^{k-1} \lambda_i u(t; -v_i) + \lambda_0 u_0(t)$$

or equivalently

$$u_{i+1}(t) = u_i(t) + \alpha_i[u(t, -v_i) - u_i(t)], \qquad i = 0, \cdots, k-1.$$

For additional details on application of the iterative procedure to a variety of problems in optimal control, see [1].

9. Acknowledgments. The author wishes to thank L. W. Neustadt and Robert O. Barr for helpful comments during the development of the material reported above. The computational results in §5 are due to Robert O. Barr.

REFERENCES

[1] R. O. Barr and E. G. Gilbert, *Some iterative procedures for computing optimal controls*, to appear.

[2] J. H. Eaton, *An iterative solution to time-optimal control*, J. Math. Anal. Appl., 5 (1962), pp. 329–344.

[3] H. G. Eggleston, *Convexity*, Cambridge University Press, Cambridge, 1958.

[4] E. J. Fadden and E. G. Gilbert, *Computational aspects of the time optimal problem*, Computing Methods in Optimization Problems, Academic Press, New York, 1964, pp. 167–192.

[5] P. S. Fancher, *Iterative computation procedures for an optimum control problem*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 346–348.

[6] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[7] A. A. Goldstein, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.

[8] Y. C. Ho, *A successive approximation technique for optimal control systems subject*

to input saturation, Trans. ASME Ser. D. J. Basic Engrg., 84 (1962), pp. 33–40.

[9] L. W. NEUSTADT, *Synthesizing time-optimal control systems*, J. Math. Anal. Appl., 1 (1960), pp. 484–492.

[10] ———, *On synthesizing optimal controls*, Proceedings of Second Congress of IFAC, Butterworth, London, 1964.

[11] ———, *Minimum effort control systems*, this Journal, 1 (1962), pp. 16–31.

[12] ———, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

MINIMIZING FUNCTIONALS ON NORMED-LINEAR SPACES*

A. A. GOLDSTEIN†

Summary. This paper extends results of the author [1], [2] and of Vainberg [3] concerning steepest descent and related topics. An example is given taken from a simple rendezvous problem in control theory. The problem is one of minimizing a norm on an affine subspace of a Banach space and is solved here in the "primal". A solution in the "dual" is given by Neustadt [4].

1. Generation of minimizing sequences. Let E be a normed linear space, x_0 an arbitrary point of E, and f a functional defined on E. Let S denote the level set $\{x \in E: f(x) \leqq f(x_0)\}$ defined at an arbitrary fixed $x_0 \in E$. We denote by $f'(x)$ the Fréchet or F-derivative of f at x. We call f *uniformly F-differentiable* on S if f is F-differentiable on S and if $\delta(\epsilon)$ in the definition of the F-derivative is constant on S. The F-derivative of f at x will be denoted by $f'(x)$. If $g \in E^*$ the value of g at x will be denoted by $[g, x]$, and if $h \in E^{**}$ the value of h at $g \in E^*$, by $[h, g]$. Recall that if E and F are normed linear spaces, A is a bounded linear operator from E to F, in short $A \in B(E, F)$, and if A is onto, then A^{-1} exists and belongs to $B(F, E)$ if and only if for some $m > 0$ and all $x \in E$, $\| Ax \| \geqq m \| x \|$; and furthermore, that $m \| x \| \leqq \| Ax \| \leqq M \| x \|$ for all x in E implies that $M^{-1} \| y \| \leqq \| A^{-1}y \| \leqq m^{-1} \| y \|$ for all y in F.

We observe that if E is a reflexive Banach space, $A \in B(E, E^*)$, and $[Ax, x] \geqq m \| x \|^2$ for all $x \in E$, then A is onto, and thus has an inverse. The proof is via the Hahn-Banach theorem. For, on the contrary supposition, take $f^0 \notin M = \text{range } A$. Choose g in E^{**} such that $g(f^0) = \text{dist } (f^0, M) > 0$, $\| g \| = 1$, and $g(f) = 0$ for all f in M. Take $\bar{g}$ in E so that $[g, f] = [f, \bar{g}]$ for all f in E^*. Then $0 = [g, Ax] = [Ax, \bar{g}]$ for all x in E. Thus $[A\bar{g}, \bar{g}] = 0$ while $\| g \| = \| \bar{g} \| = 1$.

Let ϕ denote a bounded map from S to E satisfying the two conditions $[f'(x), \phi(x)] \geqq 0$, and given $\epsilon > 0$ there exists $\delta > 0$ such that $[f'(x), \phi(x)] < \delta$ implies $\| f'(x) \| < \epsilon$. Some examples of such mappings are the following:

(1) Let $A \in B(E^*, E)$ such that $[y, Ay] \geqq \sigma \| y \|^2$ for all $y \in E^*$ and some $\sigma > 0$. Let $\phi(x) = Af'(x)$ and choose $\delta = \epsilon^2\sigma$. Then $\| f'(x) \| < \epsilon$.

* Received by the editors July 2, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 16, 1965.

† Department of Mathematics, University of Washington, Seattle, Washington. This work was supported by the Boeing Scientific Research Laboratories and by the United States Air Force under Grant AF-AFOSR-937-65.

As a likely candidate for the operator A, suppose f is twice F-differentiable on E. Assume that for some $\mu > 0$ and some x in S the operator $f''(x)$ in $B(E, E^*)$ is onto and "bounded below", that is, the bilinear functional satisfies $[f''(x)z, z] \geqq \mu \| z \|^2$ for all z in E. Then $\| f''(x)z \| \geqq \mu \| z \|$ showing that $f''(x)$ has an inverse $[f''(x)]^{-1} = A \in B(E^*, E)$. Since A has a bounded inverse, there exists a number $\sigma > 0$ such that $\| Ay \| \geqq \sigma \| y \|$ for all $y \in E^*$. Set $z = Ay$. Then $[f''(x)z, z] = [y, Ay] \geqq \mu\sigma^2 \| y \|^2$ showing the candidacy of A.

(2) Suppose E is a reflexive Banach space. By the weak compactness of the unit sphere in E it follows that for some z_0, $\| z_0 \| = 1$, $[f'(x), z_0] = \| f'(x) \|$. Set $\phi(x) = z_0 \| f'(x) \|$. Because $[f'(x), \phi(x)] = \| f'(x) \|^2, \phi(x)$ is the analogue of the gradient in Hilbert space. When E is an L_p space the point z_0 is obtained by considerations of equality in Hölder's inequality.

(3) Since $\| f'(x) \| = \sup \{[f'(x), z]: \| z \| = 1\}$, if $0 < \alpha < 1$ a point z_0 exists such that $[f'(x), z_0] \geqq \alpha \| f'(x) \|$. If for fixed α and all $x \in S$ we can find such z_0, we may take $\phi(x) = z_0 \| f'(x) \|$.

In what follows let $\Delta(x, \rho) = f(x) - f(x - \rho\phi(x))$ and $g(x, \rho) = \Delta(x, \rho)/\rho[f'(x), \phi(x)]$. Assume E is a normed linear space and S is the level set of f at x_0 in E. In what follows assume $0 < \sigma < \frac{1}{2}$.

THEOREM. *Assume that on S, f is uniformly F-differentiable or that the F-derivative f' exists and is uniformly continuous. Set $x_{k+1} = x_k$, when $[f'(x_k), \phi(x_k)] = 0$; otherwise choose*[1] *ρ_k so that $\sigma < g(x_k, \rho_k) \leqq 1 - \sigma$ when $g(x_k, 1) < \sigma$, or $\rho_k = 1$ when $g(x_k, 1) \geqq \sigma$, and set $x_{k+1} = x_k - \rho_k\phi(x_k)$.*

(a) *If S is bounded or f is bounded below, then $\{f'(x_k)\}$ converges to 0 while $\{f(x_k)\}$ converges downward to a limit, L. If S is compact, then every cluster point of $\{x_k\}$ is a zero of f'. In addition, if $\varphi(x_k) \to 0$ and f' has finitely many zeros, $\{x_k\}$ converges.*

(b) *If S is convex and bounded and f is convex, $L = \inf \{f(x): x \in S\} = \theta$. If, in addition, E is a reflexive Banach space, then every weak cluster point of $\{x_k\}$ minimizes f on E. If E is uniformly convex* (u.c.) *and f is the norm on E, then $\{x_k\}$ converges to a unique minimizer of f.*

(c) *Assume that the Gateaux derivative f' exists on S and satisfies $\mu \| z \|^2 \leqq [f''(x)z, z] \leqq M \| z \|^2$ for all $x \in S, z \in E$, and some $\mu > 0$. Assume S is convex and E is complete. Then $\{x_k\}$ converges to a unique minimizer of f on E.*

The proof of (a) is given in [1]. The proof there is stated for E, a Hilbert space, but the same proof works when E is taken to be a normed linear space. Two comments might be made, however. S bounded and f' uniformly continuous on S imply that f' is bounded on S. (See, e.g., [5, p. 19].) It

[1] If the Gateaux differential f' satisfies $f'(x, h, h) \leqq \|h\|^2\rho_0$ for all h in E, x in S, and some $\rho_0 > 0$, we can choose ρ_k to satisfy $\delta \leqq \rho_k \leqq 2\rho_0 - \delta$ with $0 < \delta \leqq \rho_0$. The method of steepest descent could also be employed, see [9].

follows by employing the mean value theorem that f is bounded below on S. The statements that f is uniformly F-differentiable and that the F-derivative f' is uniformly continuous are equivalent. (See [5, p. 45].)

(b) Given $\epsilon > 0$ choose $z' \in E$ such that $f(z') \leqq \theta + \epsilon/2$. Because f' exists at x_k and f is convex, $f(z') \geqq f(x_k) + [f'(x_k), z' - x_k]$. Since $\{f'(x_k)\} \to 0$ and S is bounded, for all k sufficiently large, $f(x_k) \leqq f(z') + \epsilon/2 \leqq \theta + \epsilon$, showing that $L = \theta$.

If E is reflexive and S is convex, closed, and bounded, then S is weakly compact. Since f is convex, the sets $\{x \in E: f(x) \leqq k\}$ are closed, convex, and weakly closed, for all k. Thus f is weakly lower semicontinuous. If z is a weak cluster point of $\{x_k\}$ then for an appropriate subsequence, $\liminf f(x_k) = L \geqq f(z)$. Assume E is u.c. and f is the norm on E. By [6, p. 113], if $\{x_k\}$ converges weakly to z and $f(x_k) \to z$ then $\{x_k\}$ converges strongly to z. It follows that every weak cluster point of $\{x_k\}$ is a strong cluster point of $\{x_k\}$. Since f' vanishes at every weak cluster point of $\{x_k\}$ and f' vanishes only once by the strict convexity of f, every subsequence of $\{x_k\}$ has the same cluster point z, showing that $\{x_k\}$ converges to z.

(c) The hypotheses of (c) imply that f' is Lipschitz continuous and that the set S is bounded. Otherwise S would contain an unbounded sequence, say $\{z_k\}$. By Taylor's theorem if $u \in S$,

$$f(z_k) \geqq f(u) + \| z_k - u \| \left[(\| z_k - u \|) \frac{\mu}{2} - \| f'(u) \| \right],$$

showing that $f(z_k) \geqq f(x_0)$ for large k, whence S must be bounded. We now show that the sequence $\{x_k\}$ is Cauchy. Again by Taylor's theorem if $s > k$,

$$f(x_s) - f(x_k) \geqq [f'(x_k), x_s - x_k] + \mu \| x_s - x_k \|^2/2.$$

Since S is bounded, $\| x_s - x_k \| \leqq D$, where D is the diameter of S. Thus

$$\| x_s - x_k \|^2 \leqq \frac{2}{\mu} \{ f(x_s) - f(x_k) + D \| f'(x_k) \| \},$$

which shows that $\{x_s\}$ is a Cauchy sequence. By the completeness of E, $\{x_s\}$ has a limit, say z, in E, and $f'(z) = 0$. If z is not unique, then $f'(z_1) = f'(z_2) = 0$, $z_1 \neq z$. Thus

$$f(z_1) - f(z_2) \geqq \frac{\mu}{2} \| z_1 - z_2 \|^2 \leqq f(z_2) - f(z_1),$$

a contradiction. Hence z is unique and is a minimizer of f.

Remark. Useful remarks may be found in [1], [3], and [9].

2. Newtonian steps and acceleration. Suppose that at the given point x_0 the function f satisfies the conditions of the first example, namely

$\phi(x) = f''_{-1}(x_0)f'(x)$, where $f''_{-1}(x_0) = [f''(x_0)]^{-1}$. The corresponding iteration is $x_{n+1} = x_n - \rho_n f''_{-1}(x_0)f'(x_n)$. This algorithm, when $\rho_n \equiv 1$, is known as the "modified" Newton's method. (See [3, p. 259] or [7, p. 696].) In a similar manner if f''_{-1} exists and is "uniformly bounded below" on S, we may define $\phi(x_n) = f''_{-1}(x_n)f'(x_n)$. We shall do this below. It is clear from what has already been said that ϕ satisfies hypotheses of the above theorem. Our object now is to formulate an algorithm using $f''_{-1}(x_n)f'(x_n) = \phi(x_n)$ which will converge at a superlinear rate.

In the following we set $\Delta(x, \rho) = f(x) - f(x - \rho f''_{-1}(x)f'(x))$ and $g(x, \rho) = \Delta(x, \rho)/\rho[f''_{-1}(x)f'(x), f'(x)]$.

THEOREM. *Assume the level set S is a convex subset of a Banach space E. For each x in S assume the F-derivative f'' is continuous on S, $f''(x)$ is onto, $\| f''(x)\| \leqq M$, and $[f''(x)z, z] \geqq m \| z \|^2$ for some $m > 0$ and all z in E. Set $x_{k+1} = x_k - \rho_k f''_{-1}(x_k)f'(x_k)$, where ρ_k is chosen so that for $\theta < \frac{1}{2}$, $0 < \theta \leqq g(x_k, \rho_k) \leqq 1 - \theta$, with $\rho_k = 1$ if possible. Then:*

(a) *there exists a number N such that if $k > N$, then $\rho_k = 1$;*

(b) *there is a unique minimizer of f and the sequence $\{x_k\}$ converges to it faster than any geometric progression.*

Proof. We have for all x in S that $M \| z \|^2 \geqq [f''(x)z, z] \geqq m \| z \|^2$ and $m^{-1} \| y \|^2 \geqq [y, f''_{-1}(x)y] \geqq mM^{-2} \| y \|^2$. Thus if $\phi(x) = f''_{-1}(x)f'(x)$, then $[f'(x), \phi(x)] \geqq mM^{-2} \| f'(x)\|^2$, showing that ϕ satisfies the conditions of the above theorem. Since f'' is bounded on S, f' is Lipschitz continuous, by the mean value theorem. By (c) above, $\{x_k\}$ converges to a unique minimizer of f.

Expand $\Delta(x, \rho)$ to two terms in the Taylor series with remainder $[f''(\xi)h, h]$, where $h = \rho f''_{-1}(x)f'(x)$. Set $f''(\xi) = f''(x) + f''(\xi) - f''(x)$. Then

$$g(x, \rho) = 1 - \frac{\rho}{2} - \rho \frac{[(f''(\xi) - f''(x))f''_{-1}(x)f'(x), f''_{-1}(x)f'(x)]}{2[f'(x), f''_{-1}(x)f'(x)]}$$

$$\geqq 1 - \frac{\rho}{2} - \rho \frac{\| f''(\xi) - f''(x) \| M^2}{2m^3}.$$

Thus

$$\left| g(x, \rho) - 1 + \frac{\rho}{2} \right| \leqq \rho \frac{\| f''(\xi) - f''(x) \| M^2}{2m}.$$

Since $\xi(\rho_k)$ lies between x_k and x_{k+1}, x_0, ξ_0, x_1, $\cdots$ is a Cauchy sequence; and it, together with its limit z, is a compactum. Consequently on this compactum f'' is uniformly continuous, so that $\{\| f''(\xi(\rho_k)) - f''(x)\|\}$ converges to 0, showing that the choice $\rho_k = 1$ is eventually feasible.

To prove (b) we write

$$x_{k+1} - z = x_k - z - \rho_k f''^{-1}(x_k) f'(x_k)$$
$$= x_k - z - \rho_k f''^{-1}(x_k) f''(x_k)(x_k - z)$$
$$+ \rho_k f''^{-1}(x_k)[f''(x_k)(x_k - z) - f'(x_k)].$$

Thus

$$\| x_{k+1} - z \| = \| x_k - z - \rho_k (x_k - z) \|$$
$$+ \rho_k \| f''^{-1}(x_k) \| \cdot \| f''(x_k)(x_k - z) - f'(x_k) \|.$$

Since f' is F-differentiable at x_k,

$$\| f'(z) - f'(x_k) - f''(x_k)(z - x_k) \| < \epsilon \| z - x_k \|.$$

Thus

$$\| x_{k+1} - z \| = (1 - \rho_k) \| x_k - z \| + \rho_k m^{-1} \epsilon \| z - x_k \|.$$

Remark 1. Both sides of the inverse of $f''^{-1}(x)$ are used in the proof.

Remark 2. The analogue of the modified Newton process, namely, choosing $\phi(x) = f''^{-1}(x_0) f'(x)$, or $f''^{-1}(x_k) f'(x)$ with k fixed, will under the hypothesis of the above theorem also generate a sequence converging to a unique minimizer of f. Since

$$\| x_{k+1} - z \| = \| x_k - z - \rho_k f''^{-1}(x_0) f''(z)(x_k - z) \|$$
$$+ \rho_k \| [f''^{-1}(x_0)] \| \epsilon \| x_k - z \| m^{-1}$$

when $\| x_k - z \| < \delta$, the rate of convergence is eventually geometric provided $\| I - \rho_k f''^{-1}(x_0) f''(z) \| < 1$. Since

$$\| I - \rho_k f''^{-1}(x_0) f''(z) \| \leqq 1 - \rho_k + \rho_k \| f''^{-1}(x_0) \| \cdot \| f''(x_0) - f''(z) \|$$

if $\| f''(x_0) - f''(z) \|$ is sufficiently small, $\rho_k \equiv 1$ will generate a sequence converging to z at the rate of a geometric progression. A sufficient condition for the global geometric convergence would be $(M/m) < \frac{1}{2}$, since $\| f''(x) \| \leqq M$ and $\| f''^{-1}(x) \| \leqq m^{-1}$.

Remark 3. Pertinent remarks may be found in [3].

3. Example.

(a) We consider the following problem which arises from a linearized rendezvous problem. See for example [8], [9], and [4]. In [4] this problem is solved in the "dual". We consider here a construction in the "primal". In [8] and [9], we have discussed this problem in the spaces $\mathfrak{L}_1$ and $\mathfrak{L}_2$; we now discuss the problem in $\mathfrak{L}_p$ for $p > 1$. Let $\mathfrak{L}_p$ denote the direct sum of n $L_p[0, 1]$ spaces. Thus a point $x \in \mathfrak{L}_p$ if $x = (x_1, \cdots, x_n)$ and $x_i \in L_p[0, 1]$;

the norm in $\mathcal{L}_p$ will be $\| x \|_p = \left[\int_0^1 | x(t)|^p \, dt\right]^{1/p}$, where $| x(t)| = [\sum_{i=1}^n x_i^2(t)]^{1/2}$. Since

$$\sqrt{n} \max \{x_i(t) : 1 \leqq i \leqq n\} \geqq | x(t)|,$$

$\| x \|_p$ is well defined. Let $\{u^i : 1 \leqq i \leqq m\}$ be a linearly independent set in $\mathcal{L}_p$. Set $1/p + 1/q = 1$. Since $q < p$, u^i is also in $\mathcal{L}_q$. Given numbers α_i, $1 \leqq i \leqq m$, define the affine subspace

$$M = \{x \in \mathcal{L}_p : [u^i, x] = \alpha_i, 1 \leqq i \leqq m\}.$$

We shall consider the problem of minimizing $f(x) = \| x \|_p^p$ on M. The limits $p \to 1$ and $p \to \infty$ correspond to the cases of rendezvous with minimum fuel and minimum thrust amplitude, respectively. In what follows we shall assume for simplicity that $n = 2$. There are no further difficulties in the general case.

We first observe that if the Gateaux differential (G-differential) of f exists, it is given by

$$\begin{aligned} f'(x)h &= p \int_0^1 | x(t) |^{p-2}[x_1(t)h_1(t) + x_2(t)h_2(t)] \, dt \\ &= p \int_0^1 | x(t) |^{p-1} \left[\frac{x_1(t)}{| x(t) |} h_1(t) + \frac{x_2(t)}{| x(t) |} h_2(t)\right] dt \\ &\leqq p \| x \|_p^{p/q}[\| h_1 \|_p + \| h_2 \|_p]. \end{aligned}$$

Here Hölder's inequality has been employed on the function $t \to | x(t) |^{p-1}$ which belongs to $L_q[0, 1]$. We have also used $\| \cdot \|_p$ for the norm in $L_p[0, 1]$. Thus the G-derivative of f exists.

Observe now that if the second G-differential exists, it is given by

$$\begin{aligned} [f''(x)h, k] &= p(p - 2) \int_0^1 | x(t) |^{p-4}(x_1(t)h_1(t) \\ &\quad + x_2(t)h_2(t))(x_1(t)k_1(t) + x_2(t)k_2(t)) \, dt \\ &\quad + p \int_0^1 | x(t) |^{p-2} (k_1(t)h_1(t) + k_2(t)h_2(t)) \, dt \\ &\leqq 2p(p - 2) \int_0^1 | x(t) |^{p-2} | h(t) | \cdot | k(t) | \, dt + p \\ &\quad \int_0^1 | x(t) |^{p-2} | h(t) | \cdot | k(t) | \, dt. \end{aligned}$$

If $x \in L_p$, $y \in L_q$, and $z \in L_r$, and $1/p + 1/q + 1/r = 1$, then

$$\int_0^1 | x(t)y(t)z(t) | \, dt \leqq \| x \|_p \| y \|_q \| z \|_r .$$

Since the function $t \to |\, x(t)\,|^{p-2}$ belongs to $L_{p'}$, where $p' = p/(p-2)$, and $1/p' + 2/p = 1$, it follows that

$$[f''(x)h, k] \leqq (2p^2 - 3p)\, \|\, x\,\|_p^{p-2}\, \|\, h\,\|_p\, \|\, k\,\|_p\,.$$

As before, let S denote the level set of f at x_0, where x_0 will be subsequently chosen in M. Thus if $x \in S$,

$$\|\, x\,\|_p^{p-2} = [f(x)]^{(p-2)/2} \leqq [f(x_0)]^{(p-2)/2},$$

showing that $[f''(x)h, k]$ is uniformly bounded on S, if h and k are confined to the unit sphere. It follows by Taylor's theorem that f' is F-differentiable on S. By the generalized mean value theorem it further follows that f' is Lipschitz continuous on S, and f is uniformly F-differentiable on S, if $p \geqq 2$. The inequality $|\, a\,|^{2r} + |\, b\,|^{2r} - 2\,|\, a\,|^r\,|\, b\,|^r\,[a, b]/|\, a\,|\cdot|\, b\,| \leqq |\, a\,|^{2r-2}M^2(r)\,|\, a - b\,|^2$, where $|\, a\,| > |\, b\,|$ and $M^2(r) < \infty$, and a direct computation show that the F-derivative f' exists and is Lipschitz continuous for all $p > 1$.

We now construct x_0 on M. Let $x_0 = \sum c_j u^j$. Thus x_0 lies on M if and only if $\sum_j c_j[u^i, u^j] = \alpha_i$. We show that the null space of the matrix $\{[u^i, u^j]\}$ consists only of the 0 element so that c_j is uniquely determined. If for some $c_j \not\equiv 0$, $\sum_j c_j[u^i, u^j] = 0$, then

$$\sum_i c_i \sum_j c_j[u^i, u^j] = [\sum c_i u^i, \sum c_j u^j] = 0,$$

contradicting the linear independence of the set $\{u^i\colon 1 \leqq i \leqq m\}$. Let

$$N = \{x \in \mathfrak{L}_p : [u^i, x] = 0,\ i = 1, \cdots, m\}.$$

We now choose h to maximize $[f'(x), h]$ subject to $\|\, h\,\|_p = 1$ and $h \in N$. The maximum is achieved because the sphere meets N in a weakly compact set and the linear function $[f(x), \cdot]$ is weakly continuous. The maximization can be accomplished by the method of Euler multipliers [10], [11]. Let $\varphi(h) = \|\, h\,\|_p^p - 1$ and $\phi_i(h) = [u^i, h]$. Then a necessary condition that h maximize $f'(x, h)$ subject to $\varphi(h) = \phi_i(h) = 0$ is that there exists c_i, $1 \leqq i \leqq m+1$, such that

$$f'(x)k = c_1 p \int_0^1 |\, h(t)\,|^{p-2}\,(h_1(t)k_1(t) + h_2(t)k_2(t))\, dt + \sum_{j=2}^{m+1} c_j[u^{j-1}, k]$$

for all $k \in \mathfrak{L}_p$. It follows that

$$p|\, x(t)\,|^{p-2}x_i(t) = pc_1|\, h(t)\,|^{p-2}h_i(t) + c_2u_i^1(t) + \cdots + c_{m+1}u_i^m(t).$$

Let

$$f_i(t) = (pc_1)^{-1}[p|\, x(t)\,|^{p-2}x_i(t) - c_2u_i^1(t) - \cdots - c_{m+1}u_i^m(t)],$$

and observe that

$$|\, h(t)\,|^{2p-2} = f_1^2(t) + f_2^2(t).$$

Therefore,

$$h_i(t) = [f_1^2(t) + f_2^2(t)^{1/2}]^{q/p} \frac{f_i(t)}{(f_1^2(t) + f_2^2(t))^{1/2}},$$

showing that $h_i \in L_p$. Solve the nonlinear equations $\varphi(h) = 0$, $\phi_i = 0$, $1 \leqq i \leqq m$, for $c_2, \cdots, c_{m+1}$, and replace h by $-h$ if necessary so that $[f'(x), h] > 0$. Because of the strict convexity of $\mathfrak{L}_p$, $[f'(x), \cdot]$ achieves a unique maximum at h.

Moreover the space $\mathfrak{L}_p$ is uniformly convex. This follows by a theorem of Smulian [12], which states that if the norm in a Banach space is uniformly F-differentiable on the unit sphere, then the conjugate space is uniformly convex.

The subspace N is also an $\mathfrak{L}_p$ space. Minimizing $f(x)$ on M is equivalent to minimizing $f(y + x_0)$ on N, with $x = y + x_0$. Clearly the gradient of the function f restricted to N is $h[f'(x), h]$. (See (2) of §1 above.) It follows therefore if $\varphi(x) = h[f'(x), h]$, conditions (a) and (b) of the theorem of §1 are satisfied.

(b) The above processes require that at each cycle a nonlinear system be solved to determine the gradient. This can be circumvented by imbedding the problem into a Hilbert space. Specifically, assume that the components of u^i are bounded and measurable. Let $\mathfrak{L}_2$ denose the direct sum of $L_2[0, 1]$ analogously to the above, and define

$$M' = \{x \in \mathfrak{L}_2 : [u^i, x] = \alpha^i, 1 \leqq i \leqq m\}.$$

Let f be now defined on M'. Since f achieves a minimum on M and $M \subset M'$, f also achieves a minimum on M'. Because M is dense in M' the minima are equal. The gradient of f on M' is merely the restriction of the gradient of f in $\mathfrak{L}_2$ to M' and is obtained by orthogonal projection. See [9]. In general $f'(x)$ does not exist. But if x is bounded and measurable, i.e., $x \in \mathfrak{L}_\infty$, $f'(x) \in \mathfrak{L}_2^*$ and $\nabla f(x) \in \mathfrak{L}_2$. The set S is bounded in $\mathfrak{L}_p$ and this implies S is bounded in $\mathfrak{L}_2$, since $\| x \|_2 \leqq \| x \|_p$ if $p \geqq 2$.

Since f is convex and continuous, S is closed, bounded and convex; furthermore, the derivatives of f are densely defined on S. Assume $x_n \in M'$, $x_n \in \mathfrak{L}_\infty$ and $u^i \in \mathfrak{L}_\infty$, $1 \leqq i \leqq m$. Then x_{n+1} is well defined and is also in $\mathfrak{L}_\infty$. To see this, verify that $\nabla f(x_n) \in \mathfrak{L}_\infty$ and the projection of $\nabla f(x_n)$ on the set $\{x \in \mathfrak{L}_2 : [u^i, x] = 0, 1 \leqq i \leqq m\}$ is also in $\mathfrak{L}_\infty$. Because f is strictly convex it achieves a unique minimum at, say, z. Therefore by the theorem of §1, $\{f(x_n)\}$ converges downward to $f(z)$ and $\{x_n\}$ converges weakly to z.

REFERENCES

[1] A. A. GOLDSTEIN, *On steepest descent*, this Journal, 3 (1965), pp. 147–151.

[2] ———, *On Newton's method*, Document D1-82-0426, Boeing Scientific Research Laboratories, Seattle, 1965.

[3] M. M. VAINBERG, *On the convergence of the method of steepest descents for non-linear equations*, Amer. Math. Soc. Transl., 1, 1 (1960).
[4] L. W. NEUSTADT, *Optimization, a moment problem and nonlinear programming*, this Journal, 2 (1964), pp. 33–53.
[5] M. M. VAINBERG, *Variational Methods for the Study of Non-Linear Operators*, Holden-Day, San Francisco, 1964.
[6] M. DAY, *Normed Linear Spaces*, Academic Press, New York, 1962.
[7] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis in Normed Spaces*, Macmillan, New York, 1964.
[8] A. A. GOLDSTEIN, A. H. GREENE, AND A. T. JOHNSON, *Fuel optimization in orbital rendezvous*, Progress in Astronautics and Aeronautics, vol. 13, Academic Press, New York, 1964, pp. 823–844.
[9] A. A. GOLDSTEIN, *Minimizing functionals on Hilbert space*, Computing Methods in Optimization Problems, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1964, pp. 159–166.
[10] L. A. LUSTERNIK, *On conditional extrema of functionals*. Mat. Sb., 41, 3 (1934), pp. 390–401.
[11] L. A. LUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Ungar, New York, 1961, p. 210.
[12] V. SMULIAN, *Sur la dérivabilité de la norme dans l'espace de Banach*, C.R. (Doklady) Acad. Sci. URSS, 27 (1940), pp. 643–648.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

A MAXIMUM PRINCIPLE OF THE PONTRYAGIN TYPE FOR SYSTEMS DESCRIBED BY NONLINEAR DIFFERENCE EQUATIONS*

HUBERT HALKIN†

1. Introduction. In this paper we consider some optimization problems for systems described by nonlinear difference equations. The present paper is a generalization of Halkin [1]. In [1] it was assumed that the difference equations are linear with respect to the state variables (but not necessarily linear with respect to the control variables). In the present paper we make no assumption whatsoever on the linearity of the difference equations with respect to either the state variables or the control variables. In the present paper however we make the same assumptions, concerning the convexity of some sets, as in [1]. These convexity assumptions, which are stated precisely in §2, are always justified in the case of a system of nonlinear difference equations which approximates a system of nonlinear differential equations (a justification of that statement is given in §5 of the present paper) but they are not necessarily justified in the case of a system of nonlinear difference equations describing a control process which is basically discrete. We remark also that in the present paper we consider more general intial and terminal conditions than in [1].

The present paper should be compared with the papers of Holtzman [2], Rosen [3], Jordan and Polak [4] in which other classes of problems are successfully treated. The reader should realize that [1]–[4] together with the present paper are not the first papers concerned with the optimal control of systems described by difference equations: the same problems have been considered in a vast number of papers and books concerned with the optimization of chemical processes. These papers and books are generally incorrect.

2. Problem statement. In the present paper the state vector will be an element x of a Euclidean space E^n, the control vector will be an element u of a Euclidean space E^r, and the time will assume the discrete values $0, 1, 2, \cdots, k$. The evolution of the system will be described by the difference equations

$$x_{i+1} - x_i = f_i(x_i, u_i), \qquad i = 0, 1, 2, \cdots, k-1. \tag{2.1}$$

* Received by the editors May 17, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Bell Telephone Laboratories, Incorporated, Whippany, New Jersey. Now at Department of Mathematics, University of California, La Jolla, California.

A certain subset $\Omega \subset E^r$ is given and all the control vectors will be required to belong to this set Ω. For every $i = 0, 1, 2, \cdots, k-1$ the vector valued function $f_i(x, u)$ is given and satisfies the following conditions:
(α) the vector valued function $f_i(x, u)$ is defined for all $(x, u) \in E^n \times \Omega$,
(β) for every $u \in \Omega$ the vector valued function $f_i(x, u)$ is twice continuously differentiable with respect to x,
(γ) the function $f_i(x, u)$ and all its first and second partial derivatives with respect to x are uniformly bounded over $A \times \Omega$ for any bounded set $A \subset E^n$,
(δ) the matrix $I + \partial f_i(x, u)/\partial x$ is not singular on $E^n \times \Omega$,
(ϵ) the set $\{f_i(x, u): u \in \Omega\}$ is convex for every $x \in E^n$.

The conditions (α), (β) and (γ) correspond to the usual "smoothness" assumptions. The conditions (δ) and (ϵ) are of another nature: they are always justified in the case of a system of difference equations which approximates a system of differential equations (see [1] and §5 of the present paper), but they are not necessarily justified in the case of a system of difference equations describing a control process which is basically discrete. ††

We shall now define an initial set

$$\{x: \quad h_i(x) = 0, \quad i = 1, 2, \cdots, l\}, \tag{2.2}$$

a terminal set

$$\{x: \quad g_i(x) = 0, \quad i = 1, 2, \cdots, m\}, \tag{2.3}$$

and an objective function $g_0(x)$. The functions $h_1(x)$, $h_2(x)$, $\cdots$, $h_l(x)$, $g_0(x)$, $g_1(x)$, $\cdots$, $g_m(x)$ are given twice continuously differentiable mappings from E^n into E^1 such that for every $x \in E^n$ the vectors

$$\frac{\partial}{\partial x} h_1(x), \frac{\partial}{\partial x} h_2(x), \cdots, \frac{\partial}{\partial x} h_l(x)$$

are linearly independent and the vectors

$$\frac{\partial}{\partial x} g_0(x), \frac{\partial}{\partial x} g_1(x), \cdots, \frac{\partial}{\partial x} g_m(x)$$

are linearly independent.

Two sequences $\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{k-1}$ and $\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_k$ are said to be optimal if they satisfy the conditions

$$h_i(x_0) = 0 \qquad \text{for } i = 1, 2, \cdots, l, \tag{2.4}$$

$$x_{i+1} - x_i = f_i(x_i, u_i) \quad \text{for all } i = 0, 1, 2, \cdots, k-1, \tag{2.5}$$

†† *Added in proof.* J. M. Holtzman has considerably released condition (ϵ) in a forthcoming paper [10].

$$u_i \in \Omega \qquad \text{for all } i = 0, 1, 2, \cdots, k-1, \tag{2.6}$$

$$g_i(x_k) = 0 \qquad \text{for } i = 1, 2, \cdots, m, \tag{2.7}$$

and if $g_0(\hat{x}_k)$ is the maximum value of $g_0(x_k)$ subject to these constraints.

3. Maximum principle. If the sequences $\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{k-1}$ and $\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_k$ are optimal then there exists a sequence of nonzero vectors $\hat{p}_0, \hat{p}_1, \cdots, \hat{p}_k$ such that[1]:

(1) *Maximization of the Hamiltonian.*

$$f_i(\hat{x}_i, \hat{u}_i)\cdot\hat{p}_{i+1} \geqq f_i(\hat{x}_i, u)\cdot\hat{p}_{i+1} \qquad \text{for all } i = 0, 1, 2, \cdots, k-1 \text{ and all } u \in \Omega. \tag{3.1}$$

(2) *Adjoint equations.*

$$\hat{p}_i - \hat{p}_{i+1} = \left(\frac{\partial}{\partial x} f_i(x, \hat{u}_i)|_{x=\hat{x}_i}\right)^T \hat{p}_{i+1} \quad \text{for all } i = 0, 1, 2, \cdots, k-1. \tag{3.2}$$

(3) *Transversality conditions.* There exist real numbers $\alpha_1, \alpha_2, \cdots, \alpha_l, \beta_0, \beta_1, \cdots, \beta_m$ such that

$$\hat{p}_0 = \sum_{i=1}^{l} \alpha_i \frac{\partial}{\partial x} h_i(x)|_{x=\hat{x}_0}, \tag{3.3}$$

$$\hat{p}_k = \sum_{i=0}^{m} \beta_i \frac{\partial}{\partial x} g_i(x)|_{x=\hat{x}_k}, \tag{3.4}$$

$$\beta_0 \geqq 0. \tag{3.5}$$

4. Proof of the maximum principle. The proof given here is similar to the proof of the maximum principle for systems described by differential equations given in [5].

Let us assume that $\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_k ; \hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{k-1}$ is an optimal solution. We shall prove that the maximum principle holds for that optimal solution.

We define the set W of all states x_k corresponding to all sequences $x_0, x_1, \cdots, x_k ; u_0, u_1, \cdots, u_{k-1}$ satisfying (2.4), (2.5), and (2.6). The set W is called the set of reachable states at time k. Next we define the set $S(\hat{x}_k)$ as the set of all states satisfying (2.7) and for which the objective function takes a greater value than at $\hat{x}_k$. Formally we have

$$S(\hat{x}_k) = \{x: \ g_i(x) = 0, \quad i = 1, \cdots, m; \quad g_0(x) > g_0(\hat{x}_k)\}. \tag{4.1}$$

We remark immediately that the sets W and $S(\hat{x}_k)$ are disjoint (i.e., have no point in common). Indeed if the sets W and $S(\hat{x}_k)$ had a point in com-

[1] The scalar product of two vectors a and b is denoted $a\cdot b$.

mon then the solution $\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{k-1}; \hat{x}_0, \hat{x}_1, \cdots, \hat{x}_k$ would not be optimal and we would have a contradiction.

In the case of the linear problem considered in [1] it is easy to prove that the sets W and $S(\hat{x}_k)$ are convex, hence separated[2] since we proved earlier that they are disjoint. When the sets W and $S(\hat{x}_k)$ are separated the proof of the maximum principle is easy, as we shall show at the end of the present section.

For a nonlinear problem of the type considered in this section the sets W and $S(x_k)$ are not necessarily convex, and hence not necessarily separated. The difficulty is turned by considering a certain linearized problem around the solution $\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{k-1}; \hat{x}_0, \hat{x}_1, \cdots, \hat{x}_{k-1}$. This linearized problem is defined as follows:

(α) the functions $h_i(x)$ are replaced by the functions

$$h_i(\hat{x}_0) + \left(\frac{\partial}{\partial x} h_i(x)|_{x=\hat{x}_0}\right)\cdot(x - \hat{x}_0), \tag{4.2}$$

(β) the functions $g_i(x)$ are replaced by the functions

$$g_i(\hat{x}_k) + \left(\frac{\partial}{\partial x} g_i(x)|_{x=\hat{x}_k}\right)\cdot(x - \hat{x}_k), \tag{4.3}$$

(γ) the functions $f_i(x, u)$ are replaced by the functions

$$f_i(\hat{x}_i, u) + \left(\frac{\partial}{\partial x} f(x, \hat{u}_i)|_{x=\hat{x}_i}\right)\cdot(x - \hat{x}_i). \tag{4.4}$$

We note immediately that the sequences $\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{k-1}; \hat{x}_0, \hat{x}_1, \cdots, \hat{x}_k$ constitute also a solution (but not necessarily an optimal solution) for the linearized problem defined above.

We define now the sets $W^+(\hat{x}_k)$ and $S^+(\hat{x}_k)$ in the same way as the sets W and $S(\hat{x}_k)$ defined earlier but with respect to the linearized problem defined above and not with respect to the initial nonlinear problem which was used in the definition of W and $S(\hat{x}_k)$. It is easy to prove that the sets $W^+(\hat{x}_k)$ and $S^+(\hat{x}_k)$ are convex. We shall now state a result which is intuitively obvious but which is nevertheless long to prove (see Appendix D).

LINEARIZATION LEMMA.[3] *If the sets W and $S(\hat{x}_k)$ are disjoint then the sets $W^+(\hat{x}_k)$ and $S^+(\hat{x}_k)$ are separated.*

[2] Two sets A and B of E^n are separated if there exists a hyperplane P such that A is contained in one of the closed halfspaces determined by P and B is contained in the other closed halfspace determined by P. There exist disjoint sets which are not separated and separated sets which are not disjoint.

[3] In §6 we show by an appropriate counterexample that a superficial understanding of the Linearization Lemma can lead to incorrect results.

With the help of this Linearization Lemma we shall now prove the maximum principle.

We have proved earlier that the sets W and $S(\hat{x}_k)$ are disjoint. From the Linearization Lemma we conclude that the convex sets $W^+(\hat{x}_k)$ and $S^+(\hat{x}_k)$ are separated, i.e., there exists a nonzero vector π such that

$$(x - \hat{x}_k)\cdot\pi \geqq 0 \quad \text{for all } x \in S^+(\hat{x}_k), \tag{4.5}$$

$$(x - \hat{x}_k)\cdot\pi \leqq 0 \quad \text{for all } x \in W^+(\hat{x}_k). \tag{4.6}$$

We define the sequence of nonzero vectors $\hat{p}_0, \hat{p}_1, \cdots, \hat{p}_k$ as the solution of the difference equation (3.2) with the terminal condition

$$\hat{p}_k = \pi. \tag{4.7}$$

We conclude by proving that relations (3.1), (3.3), (3.4) and (3.5) are satisfied. Note that (3.2) is satisfied by definition.

For any given i in $\{0, 1, \cdots, k\}$ let W_i^+ be the set of all states reachable at the time i for the linearized system (4.4) from all initial states defined by (4.2) and with all admissible control sequences. We have then $W_k^+ = W^+(\hat{x}_k)$.

We shall first prove that for every $i = 0, 1, \cdots, k$ and all $x \in W_i^+$ we have

$$(x - \hat{x}_i)\cdot\hat{p}_i \leqq 0. \tag{4.8}$$

Indeed let us assume that for some $j \in \{0, 1, \cdots, k\}$, $\tilde{x}_j \in W_j^+$, and $\epsilon > 0$ we have

$$(\tilde{x}_j - \hat{x}_j)\cdot\hat{p}_j = \epsilon > 0, \tag{4.9}$$

and show that we are led to a contradiction. We define $\tilde{x}_{j+1}, \tilde{x}_{j+2}, \cdots, \tilde{x}_k$ by the relations

$$\tilde{x}_{i+1} - \tilde{x}_i = f_i(\hat{x}_i, \hat{u}_i) + \left(\frac{\partial}{\partial x} f(x, \hat{u}_i)|_{x=\hat{x}_i}\right)(\tilde{x}_i - \hat{x}_i), \tag{4.10}$$

$$\text{for } i = j, j+1, \cdots, k-1.$$

From the previous definitions we have immediately

$$(\tilde{x}_i - \hat{x}_i)\cdot\hat{p}_i - (\tilde{x}_{i+1} - \hat{x}_{i+1})\cdot\hat{p}_{i+1} = 0, \tag{4.11}$$

hence

$$(\tilde{x}_k - \hat{x}_k)\cdot\hat{p}_k = \epsilon > 0, \tag{4.12}$$

which contradicts (4.6) since $\hat{p}_k = \pi$ and $\tilde{x}_k \in W^+(\hat{x}_k)$.

We shall now prove (3.1) by contradiction. If there are a $j \in \{0, 1, \cdots,$

$k-1\}$, a $\tilde{u}_j \in \Omega$ and an $\epsilon > 0$ such that

$$f_j(\hat{x}_j, \tilde{u}_j)\cdot\hat{p}_{j+1} = f_j(\hat{x}_j, \hat{u}_j)\cdot\hat{p}_{j+1} + \epsilon, \tag{4.13}$$

then the state $\tilde{x}_{j+1} \in W_{j+1}^{+}$, defined by

$$\tilde{x}_{j+1} - \hat{x}_j = f(\hat{x}_j, \tilde{u}_j), \tag{4.14}$$

is such that

$$(\tilde{x}_{j+1} - \hat{x}_{j+1})\cdot\hat{p}_{i+1} = \epsilon > 0, \tag{4.15}$$

which contradicts (4.8).

We shall now prove (3.3). We have

$$(x - \hat{x}_0)\cdot\hat{p}_0 \leqq 0 \tag{4.16}$$

for all $x \in W_0^{+}$, i.e., for all x such that

$$(x - \hat{x}_0)\cdot\left(\frac{\partial}{\partial x} h_i(x)|_{x=\hat{x}_0}\right) = 0, \qquad i = 1, 2, \cdots, l. \tag{4.17}$$

From (4.16) and (4.17) we obtain furthermore that

$$(x - \hat{x}_0)\cdot\hat{p}_0 = 0 \tag{4.18}$$

for all x satisfying (4.17). From (4.17) and (4.18) we obtain then (3.3).

We shall now prove (3.4). We have

$$\hat{p}_k\cdot(x - \hat{x}_k) \geqq 0 \tag{4.19}$$

for all x such that

$$(x - \hat{x}_k)\cdot\left(\frac{\partial}{\partial x} g_i(x)|_{x=\hat{x}_k}\right) = 0, \qquad i = 1, \cdots, m, \tag{4.20}$$

and

$$(x - \hat{x}_k)\cdot\left(\frac{\partial}{\partial x} g_0(x)|_{x=\hat{x}_k}\right) > 0. \tag{4.21}$$

Relations (4.19), (4.20), and (4.21) imply that

$$\hat{p}_k\cdot(x - \hat{x}_k) = 0 \tag{4.22}$$

for all x such that

$$\left(\frac{\partial}{\partial x} g_i(x)|_{x=\hat{x}_k}\right)\cdot(x - \hat{x}_k) = 0, \qquad i = 0, 1, \cdots, m. \tag{4.23}$$

From (4.22) and (4.23) we obtain then (3.4).

We conclude the proof of the maximum principle by proving (3.5). We

have

$$(x - \hat{x}_k) \cdot \hat{p}_k \geqq 0, \tag{4.24}$$

i.e.,

$$(x - \hat{x}_k) \cdot \left(\sum_{i=0}^{m} \beta_i \frac{\partial}{\partial x} g_i(x)|_{x=\hat{x}_k} \right) \geqq 0 \tag{4.25}$$

for all $x \in S^+(\hat{x}_k)$, i.e., for all x such that

$$(x - \hat{x}_k) \cdot \left(\frac{\partial}{\partial x} g_i(x)|_{x=\hat{x}_k} \right) = 0, \qquad i = 1, \cdots, m, \tag{4.26}$$

and

$$(x - \hat{x}_k) \cdot \left(\frac{\partial}{\partial x} g_0(x)|_{x=\hat{x}_k} \right) > 0. \tag{4.27}$$

The set $S^+(\hat{x}_k)$ is not empty, hence from (4.25), (4.26), and (4.27) we obtain $\beta_0 \geqq 0$.

5. Convexity and relaxed variational problems. The aim of the present section is to show that the convexity requirement,

the set $\{f_i(x, u) : u \in \Omega\}$ is convex

for every $x \in E^n$ and every $i = 0, 1, \cdots, k - 1$,

which was given in the statement of the problem is always acceptable in the case of a system of difference equations which approximates a system of differential equations.

Our claim is based on the theory of relaxed variational problems which is due to L. C. Young, R. V. Gamkrelidze and J. Warga (see, for instance, [7] and [8]). Given a control system

$$\dot{x} = f(x, u, t), \quad u \in \Omega, \tag{I}$$

we introduce its relaxed form,

$$\dot{x} \in \text{convex hull } \{f(x, u, t) : u \in \Omega\}. \tag{II}$$

Under some fairly general conditions Warga has proved that any absolutely continuous solution $x(t)$ of (II) can be uniformly approximated by absolutely continuous solutions of (I).

Let us now introduce a first order approximation of each of the systems (I) and (II). We obtain

$$x(t + h) - x(t) = f(x(t), u(t), t)h, \quad u(t) \in \Omega, \tag{I*}$$

and

$$x(t + h) - x(t) \in \text{convex hull } \{f(x(t), u, t)h : u \in \Omega\}. \tag{II*}$$

The approximation (I*) is perhaps the most "natural" approximation of the given system (I). However the approximation (II*) is much more convenient due to the convexity, and is as accurate since
(i) (II*) is as good an approximation of (II), as (I*) is of (I);
(ii) we know, from the theory of relaxed variational problems, that (I) and (II) are essentially equivalent.

Convexity is the fundamental concept in the theory of optimal control for systems described by differential equations and for systems described by difference equations. In the case of control systems described by differential equations the time, by its evolution on a continuum, has a "convexifying" effect which frees us from the necessity of adding some convexity assumptions to the data of the problem. In the case of control systems described by difference equations the time, by its evolution on a finite set, has no "convexifying" effect and, in order to obtain a maximum principle, we must add some convexity assumptions to the data of the problem.

There is a close relationship between the theory of relaxed variational problems and the convexity of the range of some vector integrals (see [5] and [6]).

The "convexifying" effect of the time, by its evolution on a continuum, is shown most simply by the following theorem [6].

If f is a piecewise continuous function from [0, 1] *into* E^n *and if* $\mathfrak{A}$ *is the set of all subsets of* [0, 1] *which are the union of a finite number of intervals, then the set* $\left\{\int_E f(t)\,dt\colon E \in \mathfrak{A}\right\}$ *is convex.*

In contradistinction let us consider a function g from $\{0, 1, \cdots, k\}$ into E^n and the set P_k of all subsets of $\{0, 1, \cdots, k\}$. The set $\{\sum_{i\in S} g(i)\colon S \in P_k\}$ is *not* convex (unless the function g is identically zero).

6. Linearization Lemma and constraint qualification. In §4 we have stated:

Linearization Lemma. If the sets W and $S(\hat{x}_k)$ are disjoint then the sets $W^+(\hat{x}_k)$ and $S^+(\hat{x}_k)$ are separated.

The aim of the present section is to persuade the reader that the following result is false:

First Naive Linearization Lemma. If the sets W and $S(\hat{x}_k)$ are disjoint then the sets $W^+(\hat{x}_k)$ and $S^+(\hat{x}_k)$ are also disjoint.

A second and equivalent form of the First Naive Linearization Lemma is implicitly accepted in the great majority of the engineering papers devoted to optimal control.[4]

Second Naive Linearization Lemma. The optimal solution of a nonlinear

[4] Unfortunately this last remark applies also to many papers devoted to optimization of systems described by differential equations.

optimization problem is also the optimal solution of the linear optimization problem obtained by linearizing around that solution the given nonlinear problem.

Let us give a counterexample to these Naive Linearization Lemmas. We have $n = k = 2$ and $r = 1$. The two state variables are denoted by x and y and the control variable is denoted by u. We have the constraint $| u | \leqq 1$. The initial manifold is the point $x_0 = y_0 = 0$ and the terminal manifold is the line $y_2 = 2$. The objective function is x_2. The evolution of the system is given by the difference equations

$$\begin{aligned} x_{i+1} - x_i &= u_i, & i &= 0, 1, \\ y_{i+1} - y_i &= 1 - (x_i)^2, & i &= 0, 1. \end{aligned} \tag{6.1}$$

It is an easy matter to visualize the set W, see Fig. 1. The optimal solution is

$$\begin{aligned} &\hat{u}_0 = 0, \quad \hat{u}_1 = +1, \\ &\hat{x}_0 = 0, \quad \hat{x}_1 = 0, \quad \hat{x}_2 = +1, \\ &\hat{y}_0 = 0, \quad \hat{y}_1 = +1, \quad \hat{y}_2 = +2, \end{aligned} \tag{6.2}$$

and $S(1, 2)$ is the set $\{(x, y): y = 2, x > 1\}$. We verify easily that the sets W and $S(1, 2)$ are disjoint.

The linearization of the system (6.1) around the solution (6.2) leads to

$$\begin{aligned} x_{i+1} - x_i &= u_i, & i &= 0, 1, \\ y_{i+1} - y_i &= 1, & i &= 0, 1. \end{aligned} \tag{6.3}$$

We have then

$$\begin{aligned} W^+(1, 2) &= \{(x, y): | x | \leqq 2, y = 2\}, \\ S^+(1, 2) &= \{(x, y): x > 1, y = 2\}. \end{aligned}$$

The sets $W^+(1, 2)$ and $S^+(1, 2)$ are not disjoint but they are separated by the line $y = 2$ which contains both of them.

The reader who is familiar with the mathematical programming literature will recognize immediately the close relationship between the preceding counterexample and the classical example (see [9, p. 229]) showing the importance of constraint qualifications in mathematical programming.

At the end of this section we want to stress the fact that the maximum principle given in §3 is valid even when the Naive Linearization Lemma is not valid. A stronger form of the maximum principle is obtained by replacing the inequality $\beta_0 \geqq 0$ of (3.5) by the strict inequality $\beta_0 > 0$.

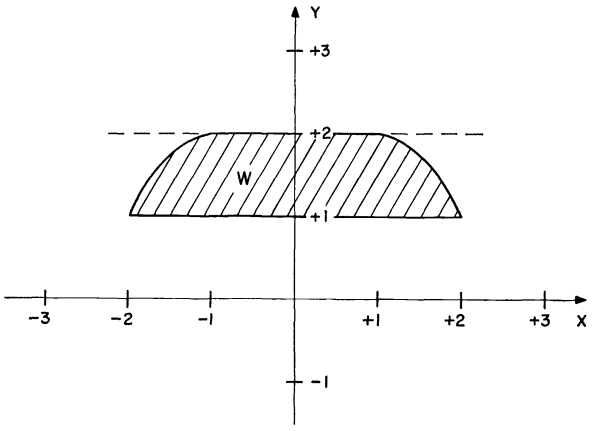

FIG. 1. *The set W for the given example*

This stronger maximum principle is valid only when the Naive Linearization Lemma is valid. A very difficult problem in optimization theory is to determine beforehand if the Naive Linearization Lemma is valid or not. This last question is closely related to the problem of normality in classical calculus of variations. A solution for which there are no vectors $\hat{p}_0, \hat{p}_1, \cdots, \hat{p}_k$ with $\beta_0 > 0$ satisfying the maximum principle is called an abnormal solution. Such a situation indicates that, from an engineering or economic point of view, the problem was ill-formulated.

Appendix A. A simple characterization of nonseparated convex sets. In this appendix we state and prove a simple property of nonseparated convex sets. This result will be used in Appendices B and D.

PROPOSITION A.1. *If K_1 and K_2 are two convex nonseparated subsets of E^n such that $0 \in \overline{K_i}$, $i = 1, 2$, and $0 \notin K_1 \cap K_2$, then there exist an integer q with $1 \leqq q \leqq n$ and $n + 1$ vectors $e_1, e_2, \cdots, e_{n+1}$ such that*

(i) *$e_i \in K_1$, for $i = 1, 2, \cdots, q$,*

(ii) *$e_i \in K_2$, for $i = q + 1, \cdots, n + 1$,*

(iii) *any n vectors among the $n + 1$ vectors $e_1, e_2, \cdots, e_{n+1}$ are linearly independent,*

(iv) *$e_i \cdot e_j$, the scalar product of e_i and e_j, is positive for all i and $j = 1, 2, \cdots, n + 1$,*

(v) *the sets A_1 and A_2 defined by*

$$(A.1) \quad A_1 = \left\{ \sum_{i=1}^{q} \mu_i e_i : \quad \sum_{i=1}^{q} \mu_i < 1, \quad \mu_j > 0 \text{ for } = 1, 2, \cdots, q \right\},$$

$$(A.2) \quad A_2 = \left\{ \sum_{i=q+1}^{n+1} \mu_i e_i : \quad \sum_{i=q+1}^{n+1} \mu_i < 1, \quad \mu_j > 0 \text{for} j = q+1, \cdots n+1 \right\},$$

are nonseparated and satisfy the relations $A_i \subset K_i$, $i = 1, 2$.

The proof of Proposition A.1 follows from the following well-known theorem.

SEPARATION THEOREM. *Two convex subsets* K_1 *and* K_2 *of* E^n *are nonseparated if and only if the two following conditions are satisfied*:
(i) *the smallest linear variety containing* K_1 *and* K_2 *is the entire space* E^n,
(ii) *there exists a point* $e^* \in \text{rint } K_i$, $i = 1, 2$.

By rint K_i we mean the interior of K_i with respect to the smallest linear variety containing K_i.

Proof of Proposition A.1. From the Separation Theorem we know that there exists a point $e^* \in \text{rint } K_i$, $i = 1, 2$. By assumption we have $0 \notin K_1 \cap K_2$ which implies that $e^* \neq 0$. From the Separation Theorem we know also that the smallest linear variety containing $K_1 \cup K_2$ is the entire space E^n. Hence there exist n linearly independent vectors $a_1, a_2, \cdots, a_n$ in $K_1 \cup K_2$. We may always assume that e^* is one of the n vectors $a_1, a_2, \cdots, a_n$ and that for some integer q with $1 \leqq q \leqq n$ we have

$$\begin{aligned} a_i &\in K_1, && \text{for } i = 1, 2, \cdots, q, \\ a_q &= e^*, \\ a_i &\in K_2 && \text{for } i = q, q+1, \cdots, n. \end{aligned}$$

Let P_1 be the smallest linear variety containing $0, a_1, a_2, \cdots, a_q$, and let P_2 be the smallest linear variety containing $0, a_q, \cdots, a_n$. Let $K_i^* = K_i \cap P_i$, $i = 1, 2$. By construction we have $e^* \in \text{rint } K_i^*$, $i = 1, 2$, and the smallest linear variety containing $K_1^* \cup K_2^*$ is the entire space E^n. Hence, by the Separation Theorem, the convex sets K_1^* and K_2^* are nonseparated.

Since $e^* \in \text{rint } K_1^*$, then there exist q linearly independent vectors $e_1, e_2, \cdots, e_q$ in K_1^* such that $e^* \in \text{rint } A_1$, where A_1 is defined by (A.1). Similarly since $e^* \in \text{rint } K_2^*$, then there exist $n - q + 1$ linearly independent vectors $e_{q+1}, \cdots, e_{n+1}$ in K_2^* such that $e^* \in \text{rint } A_2$, where A_2 is defined by (A.2). By construction any n vectors among the $n + 1$ vectors $e_1, e_2, \cdots, e_{n+1}$ are linearly independent and the smallest linear variety containing $A_1 \cup A_2$ is the entire space E^n. Hence, from the Separation Theorem, we know that the sets A_1 and A_2 are nonseparated. There is no loss of generality by assuming that $e_i \cdot e_j > 0$ for all i and $j = 1, 2, \cdots, n + 1$. Indeed for any $\lambda \in (0, 1)$, let $e_i(\lambda) = e^* + \lambda(e_i - e^*)$; for λ small enough we have $e_i(\lambda) \cdot e_j(\lambda) > 0$ for all i and $j = 1, 2, \cdots, n + 1$, and none of the other properties are violated. This concludes the proof of Proposition A.1.

Appendix B. A topological property of nonseparated convex sets. In this appendix we prove a single proposition which plays a fundamental role in the proof of the Linearization Lemma to be given in Appendix D.

PROPOSITION B.1. *Let K_1 and K_2 be two nonseparated convex sets in E^n. We assume that the origin 0 belongs to $\overline{K_1}$, the closure of K_1, and to $\overline{K_2}$, the closure of K_2. Let L be a positive constant. We are given a continuous mapping φ_1 from K_1 into E^n and a continuous mapping φ_2 from K_2 into E^n. We assume that for $i = 1, 2$ we have*

$$|\varphi_1(e) - e| \leqq L|e|^2 \qquad \text{for all } e \in K_i. \tag{B.1}$$

Then the set $\varphi_1(K_1) \cap \varphi_2(K_2)$ is not empty.

Before proving Proposition B.1 we shall state and prove Proposition B.2 which is a particular case of Proposition B.1. (One of the given convex sets is merely a one-dimensional linear segment and the mapping corresponding to this linear segment is the identity mapping.)

PROPOSITION B.2. *Let K be a convex set in E^n. We assume that the origin 0 belongs to $\bar{K}$, the closure of K, and that there exists a point x interior to the set K. Let L be a positive constant. We are given a continuous mapping φ from K into E^n such that*

$$|\varphi(e) - e| \leqq L|e|^2 \qquad \text{for all } e \in K. \tag{B.2}$$

Then there are an $\alpha > 0$ and a point y interior to the set K such that

$$\varphi(y) = \alpha x. \tag{B.3}$$

Proof of Proposition B.2. Since x is an interior point of K we know that there exists an $\epsilon > 0$ with $N(x, \epsilon) \subset K$. By $N(x, \epsilon)$ we mean the set $\{x^*: |x - x^*| \leqq \epsilon\}$. Let α be a positive constant such that

$$L\alpha^2(|x| + \epsilon)^2 \leqq \frac{\alpha\epsilon}{4} \tag{B.4}$$

and

$$\alpha \leqq 1. \tag{B.5}$$

Since the set K is convex, $0 \in \bar{K}$, and $N(x, \epsilon) \subset K$, we have then immediately $N(\alpha x, \alpha\epsilon/2) \subset K$. Let $h(e) = e - \varphi(e) + \alpha x$. The function h is continuous and maps $N(\alpha x, \alpha\epsilon/2)$ into itself since

$$|h(e) - \alpha x| = |e - \varphi(e)| \leqq L\alpha^2(|x| + \epsilon)^2 \leqq \frac{\alpha\epsilon}{4} \tag{B.6}$$

for all $e \in N(\alpha x, \alpha\epsilon/2)$. Hence, by Brouwer's fixed point theorem, there is a $y \in N(\alpha x, \alpha\epsilon/2)$ such that $h(y) = y$, i.e., $\varphi(y) = \alpha x$. The point y belongs to the interior of the set K and Proposition B.2 is proved.

Proof of Proposition B.1. The two convex sets K_1 and K_2 are nonsepa-

rated and we have $0 \in \overline{K_i}$, $i = 1, 2$. If $0 \in K_1 \cap K_2$, then Proposition B.1 is trivially satisfied, since $\varphi_1(0) = \varphi_2(0) = 0$ and $0 \in \varphi_1(K_1) \cap \varphi_2(K_2)$. Let us assume that $0 \notin K_1 \cap K_2$. Then, from Proposition A.1, there exist an integer q with $1 \leqq q \leqq n$ and $n + 1$ vectors $e_1, e_2, \cdots, e_{n+1}$ such that conditions (i)–(v) in Proposition A.1 are satisfied. From now to the end of the proof we shall restrict our attention to the sets A_1 and A_2 and prove that

$$\varphi_1(A_1) \cap \varphi_2(A_2) \tag{B.7}$$

is not empty which a fortiori implies that $\varphi_1(K_1) \cap \varphi_2(K_2)$ is not empty.

Since the sets A_1 and A_2 are nonseparated then there exists a point

$$e^* \in \operatorname{rint} A_i, \quad i = 1, 2. \tag{B.8}$$

In other words there exist $n + 1$ positive numbers $\lambda_1, \lambda_2, \cdots, \lambda_{n+1}$ such that

$$e^* = \sum_{i=1}^{q} \lambda_i e_i, \tag{B.9}$$

$$e^* = \sum_{i=q+1}^{n+1} \lambda_i e_i, \tag{B.10}$$

$$\sum_{i=1}^{q} \lambda_i < 1, \tag{B.11}$$

$$\sum_{i=q+1}^{n+1} \lambda_i < 1. \tag{B.12}$$

Let $x_1, x_2, \cdots, x_{n+1}$ be vectors in E^{n+1} determined by

$$x_i = \left(e_i, \frac{1}{q\lambda_i}\right) \qquad \text{for } i = 1, 2, \cdots, q, \tag{B.13}$$

$$x_i = \left(-e_i, \frac{1}{(n + 1 - q)\lambda_i}\right) \qquad \text{for } i = q + 1, \cdots, n + 1. \tag{B.14}$$

It is easy to prove that the vectors $x_1, x_2, \cdots, x_{n+1}$ are linearly independent and that

$$\sum_{i=1}^{n+1} \lambda_i x_i = (0, 0, \cdots, 0, 2). \tag{B.15}$$

Let A be the subset of E^{n+1} defined by

$$A = \left\{ \sum_{i=1}^{n+1} \mu_i x_i : \sum_{i=1}^{n+1} \mu_i < 1, \quad \mu_j > 0 \text{ for } j = 1, 2, \cdots, n + 1 \right\}, \tag{B.16}$$

and let $x^* = (0, 0, \cdots, 0, 1)$. We have $x^* \in \operatorname{rint} A$.

Let φ be a continuous mapping from A into E^{n+1} defined by

$$\varphi\left(\sum_{i=1}^{n+1} \mu_i x_i\right) = \left(\varphi_1\left(\sum_{i=1}^{q} \mu_i e_i\right), \sum_{i=1}^{q} \frac{\mu_i}{q\lambda_i}\right) + \left(\varphi_2\left(\sum_{i=q+1}^{n+1} \mu_i e_i\right), \sum_{i=q+1}^{n+1} \frac{\mu_i}{(n+1-q)\lambda_i}\right). \tag{B.17}$$

The mapping φ is well defined since the representation $\sum_{i=1}^{n+1} \mu_i x_i$ is unique (the vectors $x_1, x_2, \cdots, x_{n+1}$ are linearly independent) and since $\sum_{i=1}^{q} \mu_i e_i \in A_1$ and $\sum_{i=q+1}^{n+1} \mu_i e_i \in A_2$. We have

$$\varphi\left(\sum_{i=1}^{n+1} \mu_i x_i\right) - \sum_{i=1}^{n+1} u_i x_i = \left(\left(\varphi_1\left(\sum_{i=1}^{q} \mu_i e_i\right) - \sum_{i=1}^{q} \mu_i e_i\right) - \left(\varphi_2\left(\sum_{i=q+1}^{n+1} \mu_i e_i\right) - \sum_{i=q+1}^{n+1} \mu_i e_i\right), 0\right). \tag{B.18}$$

Hence

$$\begin{aligned} \left|\varphi\left(\sum_{i=1}^{n+1} \mu_i x_i\right) - \sum_{i=1}^{n+1} \mu_i x_i\right| &\leqq \left|\varphi_1\left(\sum_{i=1}^{q} \mu_i e_i\right) - \sum_{i=1}^{q} \mu_i e_i\right| \\ &\quad + \left|\varphi_2\left(\sum_{i=q+1}^{n+1} \mu_i e_i\right) - \sum_{i=q+1}^{n+1} \mu_i e_i\right| \\ &\leqq L\left|\sum_{i=1}^{q} \mu_i e_i\right|^2 + L\left|\sum_{i=q+1}^{n+1} \mu_i e_i\right|^2 \\ &\leqq L \max_{j=1,\cdots,n+1} |e_j|^2 \sum_{i=1}^{n+1} |\mu_i|^2. \end{aligned} \tag{B.19}$$

Since the vectors $x_1, x_2, \cdots, x_{n+1}$ are linearly independent there exists a constant $N < +\infty$ such that

$$\sum_{i=1}^{n+1} |\mu_i|^2 \leqq N\left|\sum_{i=1}^{n+1} \mu_i x_i\right|^2 \tag{B.20}$$

for all $\mu_1, \mu_2, \cdots, \mu_{n+1}$. From (B.19) and (B.20) we obtain

$$\left|\varphi\left(\sum_{i=1}^{n+1} \mu_i x_i\right) - \sum_{i=1}^{n+1} \mu_i x_i\right| \leqq L^*\left|\sum_{i=1}^{n+1} \mu_i x_i\right|^2, \tag{B.21}$$

where

$$L^* = L \max_{j=1,\cdots,n+1} |e_j|^2 N. \tag{B.22}$$

The relation (B.21) can be written

$$|\varphi(e) - e| \leqq L^* |e|^2 \qquad \text{for all } e \in A. \tag{B.23}$$

By Proposition B.2 there are an $\tilde{x} \in \text{rint } A$ and an $\alpha > 0$ such that

$$\varphi(\tilde{x}) = \alpha x^* = (0, 0, \cdots, 0, \alpha). \tag{B.24}$$

We have then $\tilde{x} = \sum_{i=1}^{n+1} \nu_i x_i$, for some $\nu_1, \nu_2, \cdots, \nu_{n+1}$ such that $\sum_{i=1}^{n+1} \nu_i < +1$, and $\nu_j > 0$ for $j = 1, 2, \cdots, n+1$. This implies that

$$\begin{aligned}(0, 0, \cdots, 0, \alpha) = \varphi(\tilde{x}) &= \varphi\left(\sum_{i=1}^{n+1} \nu_i x_i\right) \\ &= \left(\varphi_1\left(\sum_{i=1}^{q} \nu_i e_i\right), \sum_{i=1}^{q} \frac{\nu_i}{q\lambda_i}\right) \\ &\quad + \left(-\varphi_2\left(\sum_{i=q+1}^{n+1} \nu_i e_i\right), \sum_{i=q+1}^{n+1} \frac{\nu_i}{(n+1-q)\lambda_i}\right)\end{aligned} \tag{B.25}$$

and, in particular, that

$$\varphi_1\left(\sum_{i=1}^{q} \nu_i e_i\right) - \varphi_2\left(\sum_{i=q+1}^{n+1} \nu_i e_i\right) = 0. \tag{B.26}$$

Let $e^1 = \sum_{i=1}^{q} \nu_i e_i$ and $e^2 = \sum_{i=q+1}^{n+1} \nu_i e_i$. We have then $\varphi_1(e^1) = \varphi_2(e^2)$ with $e^1 \in A_1$ and $e^2 \in A_2$. This concludes the proof of Proposition B.1.

Appendix C. A Gronwall inequality for difference equations. In this appendix we prove a single result.

PROPOSITION C.1. *Let L, $\alpha_0, \alpha_1, \cdots, \alpha_k$, $\beta_0, \beta_1, \cdots, \beta_{k-1}$ be real numbers such that*

$$\alpha_{i+1} - \alpha_i \leqq L\alpha_i + \beta_i, \qquad i = 0, 1, 2, \cdots, k-1, \tag{C.1}$$

$$L \geqq 0, \tag{C.2}$$

$$\beta_i \geqq 0, \qquad i = 0, 1, 2, \cdots, k-1, \tag{C.3}$$

$$\alpha_i \geqq 0, \qquad i = 0, 1, 2, \cdots, k. \tag{C.4}$$

Then

$$\alpha_i \leqq (1+L)^i \left(\alpha_0 + \sum_{j=0}^{i-1} (1+L)^{-j-1}\beta_j\right), \qquad i = 0, 1, 2, \cdots, k. \tag{C.5}$$

Proof of Proposition C.1. We prove (C.5) by induction. For $i = 0$ we have identically $\alpha_0 \leqq \alpha_0$. Let us assume that (C.5) is true for $i = 0, 1, 2, \cdots, \nu$ and prove it for $i = \nu + 1$. We have

$$\alpha_{\nu+1} - \alpha_\nu \leqq L\alpha_\nu + \beta_\nu, \tag{C.6}$$

$$\alpha_{\nu+1} \leqq (1+L)\alpha_\nu + \beta_\nu, \tag{C.7}$$

and

$$\alpha_\nu \leqq (1+L)^\nu \left(\alpha_0 + \sum_{j=0}^{\nu-1} (1+L)^{-j-1}\beta_j\right). \tag{C.8}$$

From (C.7) and (C.8) we obtain

$$\begin{aligned}\alpha_{\nu+1} &\leqq (1+L)^{\nu+1}\left(\alpha_0 + \sum_{j=0}^{\nu-1}(1+L)^{-j-1}\beta_j\right) + \beta_\nu \\ &\leqq (1+L)^{\nu+1}\left(\alpha_0 + \sum_{j=0}^{\nu-1}(1+L)^{-j-1}\beta_j + (1+L)^{-\nu-1}\beta_\nu\right) \qquad \text{(C.9)} \\ &\leqq (1+L)^{\nu+1}\left(\alpha_0 + \sum_{j=0}^{\nu}(1+L)^{-j-1}\beta_j\right).\end{aligned}$$

This concludes the proof of Proposition C.1.

Appendix D. Proof of the Linearization Lemma. In this appendix we prove the Linearization Lemma which has been stated in §4 as follows.

LINEARIZATION LEMMA. *If the sets W and $S(\hat{x}_k)$ are disjoint then the sets $W^+(\hat{x}_k)$ and $S^+(\hat{x}_k)$ are separated.*

Proof of the Linearization Lemma. We shall assume that the convex sets $W^+(\hat{x}_k)$ and $S^+(\hat{x}_k)$ are nonseparated and show that this implies that the sets W and $S(\hat{x}_k)$ are not disjoint.

There is no loss of generality in assuming[5] that

$$\hat{x}_0 = \hat{x}_1 = \cdots = \hat{x}_k = 0. \qquad \text{(D.1)}$$

We shall have accordingly

$$h_i(0) = 0, \qquad i = 1, 2, \cdots, l, \qquad \text{(D.2)}$$

and

$$g_i(0) = 0, \qquad i = 1, 2, \cdots, m. \qquad \text{(D.3)}$$

In order to simplify the notation we shall write W^+, S^+ and S instead of $W^+(\hat{x}_k)$, $S^+(\hat{x}_k)$ and $S(\hat{x}_k)$.

From the previous definitions and assumptions we know that the sets W^+ and S^+ are convex and nonseparated, that 0 belongs to the closure of W^+ and to the closure of S^+, and that $0 \notin S^+$. Hence, from Proposition A.1, there exist an integer q with $1 \leqq q \leqq n$ and $n+1$ vectors $e_1, e_2, \cdots, e_{n+1}$ such that

(i) $e_i \in W^+$, for $i = 1, 2, \cdots, q$,

(ii) $e_i \in S^+$, for $i = q+1, \cdots, n+1$,

(iii) any n vectors among the $n+1$ vectors $e_1, e_2, \cdots, e_{n+1}$ are linearly independent,

(iv) $e_i \cdot e_j > 0$ for all i and $j = 1, 2, \cdots, n+1$,

(v) the sets A_1 and A_2 defined by (A.1) and (A.2) are nonseparated and satisfy the relations $A_1 \subset W^+$ and $A_2 \subset S^+$.

From now to the end of the proof we shall restrict our attention to the sets A_1 and A_2.

[5] Define a new state variable y by the time varying transformation $y = x - \hat{x}_1$.

We shall construct a continuous mapping φ_1 from A_1 into W and a continuous mapping φ_2 from A_2 into S such that for some L_1 and $L_2 < +\infty$ we have for $i = 1$ and 2,

$$|\varphi_i(x) - x| \leqq L_i |x|^2 \qquad \text{for all } x \in A_i, \tag{D.4}$$

i.e.,

$$|\varphi_i(x) - x| \leqq L |x|^2, \qquad \text{for all } x \in A_i, \tag{D.5}$$

where $L = \max\{L_1, L_2\}$. From the last two relations and from Proposition B.1 we conclude that the sets $\varphi_1(A_1)$ and $\varphi_2(A_2)$ are not disjoint which implies a fortiori that the sets W and S are not disjoint. This contradiction will then conclude the proof of the Linearization Lemma.

The second of the relations (D.4), for $i = 2$, is immediate since we have assumed that the functions $g_0(x)$, $g_1(x)$, $\cdots$, $g_m(x)$ defining the sets S^+ and S are *twice* continuously differentiable with respect to x.

To simplify the notation we shall sometimes write e_0 for 0, $u_j(e_0)$ and $u_j(0)$ for $\hat{u}_j$, and $x_i^+(e_0)$ for 0.

By assumption we have $e_i \in W^+$ for $i = 1, 2, \cdots, q$. In other words, the point e_i is reachable for the linearized problem defined in §4. Let us denote by $x_0^+(e_i)$, $x_1^+(e_i)$, $\cdots$, $x_k^+(e_i)$ a sequence of states and by $u_0^+(e_i)$, $u_1^+(e_i)$, $\cdots$, $u_{k-1}^+(e_i)$ a sequence of controls leading to the point e_i for this linearized problem. For $i = 1, 2, \cdots, q$ we have then

$$\left(\frac{\partial}{\partial x} h_j(x) \,|_{x=0}\right) \cdot x_0^+(e_i) = 0 \quad \text{for} \quad j = 1, 2, \cdots, l, \tag{D.6}$$

$$\begin{aligned} x_{j+1}^+(e_i) - x_j^+(e_i) &= f(0, u_j(e_i)) \\ &+ \left(\frac{\partial}{\partial x} f(x, u_j(0)) \,|_{x=0}\right) x_j^+(e_i) \quad \text{for} \quad j = 0, 1, 2, \cdots, k-1, \end{aligned} \tag{D.7}$$

$$x_k^+(e_i) = e_i. \tag{D.8}$$

Let D^+ be the convex hull of $\{x_0^+(e_0) = 0, x_0^+(e_1), \cdots, x_0^+(e_q)\}$. Since we have assumed that the functions $h_1(x)$, $h_2(x)$, $\cdots$, $h_l(x)$ are *twice* continuously differentiable with respect to x then there exist a continuous mapping $\psi(x)$ from

$$\left\{y: \quad \left(\frac{\partial}{\partial x} h_j(x) \,|_{x=0}\right) \cdot y = 0, \quad j = 1, \cdots, l\right\} \tag{D.9}$$

into

$$\{y: \quad h_j(y) = 0, \quad j = 1, \cdots, l\} \tag{D.10}$$

and a constant $L_3 < +\infty$ such that

$$|\psi(x) - x| \leqq L_3 |x|^2 \qquad \text{for all } x \in D^+. \tag{D.11}$$

Let $D = \psi(D^+)$.

For any $a \in A_1$ let $\lambda(a) = (\lambda_0(a), \lambda_1(a), \lambda_2(a), \cdots, \lambda_q(a))$ be the barycentric coordinates of a with respect to the points $e_0 = 0$, e_1, e_2, $\cdots$, e_q. In other words we have

$$a = \sum_{j=0}^{q} \lambda_j(a) e_j, \quad \sum_{j=0}^{q} \lambda_j(a) = 1. \tag{D.12}$$

By assumption we have $a \in W^+$ for any $a \in A_1$. In other words, any point a in A_1 is reachable for the linearized problem defined in §4. For any $a \in A$ we shall define below a sequence of states $x_0^+(a), x_1^+(a), \cdots, x_k^+(a)$ leading to the point a for this linearized problem:

$$x_0^+(a) = \sum_{j=0}^{q} \lambda_j(a) x_0^+(e_j), \tag{D.13}$$

$$\begin{aligned} x_{i+1}^+(a) - x_i^+(a) &= \sum_{j=0}^{q} \lambda_j(a) f_i(0, u_i(e_j)) \\ &+ \left(\frac{\partial}{\partial x} f_i(x, u_i(0))|_{x=0} \right) x_i^+(a) \quad \text{for} \quad i = 0, 1, 2, \cdots, k-1. \end{aligned} \tag{D.14}$$

The definitions (D.13) and (D.14) are compatible with the definitions (D.7) and (D.8), and for $i = 0, 1, 2, \cdots, k$ we have

$$x_i^+(a) = \sum_{j=0}^{q} \lambda_j(a) x_i^+(e_j). \tag{D.15}$$

We have assumed that the set $\{f_i(x, u) : u \in \Omega\}$ is convex for every $i = 0, 1, \cdots, k-1$ and every $x \in E^n$. Hence for any $a \in A_1$, there exists a control sequence $u_0^+(a), u_1^+(a), \cdots, u_{k-1}^+(a)$ such that

$$f_i(0, u_i^+(a)) = \sum_{j=0}^{q} \lambda_j(a) f_i(0, u_i(e_j)) \quad \text{for } i = 0, 1, 2, \cdots, k-1. \tag{D.16}$$

Consequently the sequence $x_0^+(a), x_1^+(a), \cdots, x_k^+(a)$ defined by (D.13) and (D.14) corresponds to some admissible control sequence $u_0^+(a), u_1^+(a), \cdots, u_{k-1}^+(a)$ for the linearized problem defined in §4.

We shall now define for every $a \in A_1$ a sequence of states $x_0(a), x_1(a), \cdots, x_k(a)$ which forms a solution for the original nonlinear difference equations. The initial state $x_0(a)$ is defined by

$$x_0(a) = \psi(x_0^+(a)),$$

and the states $x_1(a), x_2(a), \cdots, x_k(a)$ are defined inductively by

$$\begin{aligned} x_{i+1}(a) - x_i(a) \\ = \sum_{j=0}^{q} \lambda_j(a) f_i(x_i(a), u_i(e_j)) \quad \text{for } i = 0, 1, 2, \cdots, k-1. \end{aligned} \tag{D.17}$$

Again from the assumption that the set $\{f_i(x, u): u \in \Omega\}$ is convex for every $i = 0, 1, \cdots, k-1$ and every $x \in E^n$ we know that for any $a \in A_1$ there exists a control sequence $u_0(a), u_1(a), \cdots, u_{k-1}(a)$ such that

$$f_i(x_i(a), u_i(a)) = \sum_{j=0}^{q} \lambda_j(a) f_i(x_i(a), u_i(e_j)) \quad \text{for } i = 0, 1, 2, \cdots, k-1. \tag{D.18}$$

Consequently the sequence $x_0(a), x_1(a), \cdots, x_k(a)$ defined by (D.16) and (D.17) corresponds, for the original nonlinear equations, to some admissible control sequence $u_0(a), u_1(a), \cdots, u_{k-1}(a)$. Moreover we have

$$x_i(0) = \hat{x}_i = 0 \qquad \text{for } i = 0, 1, 2, \cdots, k. \tag{D.19}$$

The mapping φ_1 from A_1 into W is defined as the mapping from $x_k^+(a) = a$ into $x_k(a)$. In other words we define the mapping φ_1 by the relation

$$\varphi_1(a) = x_k(a) \qquad \text{for all } a \in A_1. \tag{D.20}$$

We shall now prove that $x_k(a)$ is continuous with respect to a over A_1 and that there exists an $L_1 < +\infty$ such that

$$| x_k(a) - a | \leqq L_1 | a |^2 \qquad \text{for all } a \in A_1. \tag{D.21}$$

Let $M = \sup_{a \in A} | x_0(a) - \hat{x}_0 |$. We have $M < +\infty$. Let $H \subset E^n$ be the set of all states which are reachable at some time $i = 0, 1, \cdots, k$ with some admissible control function and from some point x at $i = 0$ satisfying the condition $| x - \hat{x}_0 | \leqq M$. The set H is bounded. Let N be such that

$$| f_i(x', u) - f_i(x'', u) | \leqq N | x' - x'' | \quad \text{for all } x' \text{ and } x'' \in H, \text{ all } i = 0, 1, \cdots, k-1, \text{ and all } u \in \Omega, \tag{D.22}$$

$$| f_i(x, u) | \leqq N \quad \text{for all } x \in H, \text{ all } i = 0, 1, \cdots, k-1, \text{ and all } u \in \Omega. \tag{D.23}$$

From the initial assumptions relative to the set Ω and to the functions $f_i(x, u)$ we know that the constant $N < +\infty$ exists.

Let us first prove that the function $x_k(a)$ is continuous with respect to a over A_1. For any $i = 0, 1, 2, \cdots, k-1$ we have

$$| x_{i+1}(a') - x_{i+1}(a'') | - | x_i(a') - x_i(a'') | \leqq | (x_{i+1}(a') - x_i(a')) - (x_{i+1}(a'') - x_i(a'')) | \tag{D.24}$$

$$\leqq \left| \sum_{j=0}^{q} \lambda_j(a') f_i(x_i(a'), u_i(e_j)) - \sum_{j=0}^{q} \lambda_j(a'') f_i(x_i(a''), u_i(e_j)) \right| \tag{D.25}$$

$$\leqq \left| \sum_{j=0}^{q} \lambda_j(a') f_i(x_i(a'), u_i(e_j)) - \sum_{j=0}^{q} \lambda_j(a') f_i(x_i(a''), u_i(e)) \right| \tag{D.26}$$
$$+ \left| \sum_{j=0}^{q} \lambda_j(a') f_i(x_i(a''), u_i(e_j)) - \sum_{j=0}^{q} \lambda_j(a'') f_i(x_i(a''), u_i(e_j)) \right|$$

$$\leqq N \mid x_i(a') - x_i(a'') \mid + N \sum_{j=0}^{q} \mid \lambda_j(a') - \lambda_j(a'') \mid. \tag{D.27}$$

From Proposition C.1 (Gronwall's inequality for difference equations) we have then

$$\mid x_k(a') - x_k(a'') \mid \tag{D.28}$$
$$\leqq (1 + N)^k \left(x_0(a') - x_0(a'') + kN \sum_{j=0}^{q} \mid \lambda_j(a') - \lambda_j(a'') \mid \right).$$

We know already that $x_0(a)$ and $\lambda_j(a)$, $j = 0, 1, 2, \cdots, q$, are continuous functions of a over A_1. Hence from (D.28) it follows that $x_k(a)$ is a continuous function of a over A_1.

It remains to prove that there exists an $L_1 < +\infty$ such that

$$\mid x_k(a) - a \mid \leqq L_1 \mid a \mid^2 \qquad \text{for all } a \in A_1. \tag{D.29}$$

From the previous definitions we have immediately

$$x_k(0) = 0. \tag{D.30}$$

For every $a \in A_1$ with $a \neq 0$ let us consider the vector $x_k(\epsilon a/\mid a \mid)$ as a function of ϵ. We shall prove below that:

(i) $\dfrac{\partial^2}{\partial \epsilon^2} x_k \left(\epsilon \dfrac{a}{\mid a \mid} \right)$ is well defined for any $a \in A_1$ with $a \neq 0$ and for any $\epsilon \in [0, \mid a \mid]$; moreover there exists a constant $L_1 < +\infty$ such that for any $a \in A_1$ with $a \neq 0$ and for any $\epsilon \in [0, \mid a \mid]$ we have

$$\left| \frac{\partial^2}{\partial \epsilon^2} x_k \left(\epsilon \frac{a}{\mid a \mid} \right) \right| \leqq L_1; \tag{D.31}$$

(ii) for every $a \in A_1$ with $a \neq 0$ we have

$$\frac{\partial}{\partial \epsilon} x_k \left(\epsilon \frac{a}{\mid a \mid} \right) \Bigg|_{\epsilon=0} = \frac{a}{\mid a \mid}. \tag{D.32}$$

From (D.30), (D.31), and (D.32) we obtain the desired result (D.29).

We conclude the proof of the Linearization Lemma by proving (D.31) and (D.32). The proof of (D.31) is an immediate consequence of the assumptions that the functions $h_i(x)$ and $f_i(x, u)$ are twice continuously differentiable with respect to x. The relation (D.32) can be written equiva-

lently

$$\frac{\partial}{\partial \epsilon} x_k(\epsilon a)|_{\epsilon=0} = a \quad \text{for every } a \in A_1 . \tag{D.33}$$

For every $a \in A_1$ and every $i = 0, 1, \cdots, k$ we define $y_i(a)$ by the relation

$$y_i(a) = \frac{\partial}{\partial \epsilon} x_i(\epsilon a) \, |_{\epsilon=0} . \tag{D.34}$$

We have immediately

$$y_0(a) = \frac{\partial}{\partial \epsilon} \psi \left(\sum_{j=0}^{q} \lambda_j(\epsilon a) x_0^+(e_j) \right) \Bigg|_{\epsilon=0} \tag{D.35}$$

$$= \sum_{j=0}^{q} \lambda_i(a) x_0^+(e_j) \tag{D.36}$$

$$= x_0^+(a) \tag{D.37}$$

and

$$y_{i+1}(a) - y_i(a) = \frac{\partial}{\partial \epsilon} \sum_{j=0}^{q} \lambda_j(\epsilon a) f_i\big(x_i(\epsilon a), u_i(e_j)\big) \, |_{\epsilon=0} \tag{D.38}$$

$$= \sum_{j=0}^{q} \lambda_j(a) f_i\big(0, u_i(e_j)\big)$$

$$+ \left(\frac{\partial}{\partial x} f_i\big(x, u_i(0)\big)|_{x=0} \right) y_i(a) \quad \text{for} \quad i = 0, 1, 2, \cdots, k-1. \tag{D.39}$$

By comparing (D.14), (D.37), and (D.39) we see that for all $a \in A_1$ and all $i = 0, 1, 2, \cdots, k$ we have

$$y_i(a) = x_i^+(a), \tag{D.40}$$

in particular we have

$$x_k^+(a) = y_k(a). \tag{D.41}$$

By definition we have also

$$x_k^+(a) = a \tag{D.42}$$

and

$$y_k(a) = \frac{\partial}{\partial \epsilon} x_k(\epsilon a) \, |_{\epsilon=0} . \tag{D.43}$$

From (D.41), (D.42), and (D.43) we obtain the desired result that

$$a = \frac{\partial}{\partial \epsilon} x_k(\epsilon a)|_{\epsilon=0} . \tag{D.44}$$

This concludes the proof of the Linearization Lemma.

Acknowledgment. This research was partially supported by the Advanced Research Projects Agency (Project DEFENDER), United States Army Research Office—Durham under Contract DA-31-124-ARO-D-257 and the Air Force Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grant 1039-66.

REFERENCES

[1] H. HALKIN, *Optimal control for systems described by difference equations*, Advances in Control Systems: Theory and Applications, Academic Press, New York, 1964, pp. 173–196.

[2] J. M. HOLTZMAN, *Convexity and the maximum principle for discrete systems*, to appear.

[3] J. B. ROSEN, *Optimal control and convex programing*, Tech. Rpt. 547, Mathematics Research Center, University of Wisconsin, Madison, 1965.

[4] B. W. JORDAN AND E. POLAK, *Theory of a class of discrete optimal control systems*, J. Electronics Control, 17 (1964), pp. 697–713.

[5] H. HALKIN, *On the necessary condition for optimal control of nonlinear systems*, J. Analyse Math., 12 (1964), pp. 1–82.

[6] ———, *Some further generalizations of a theorem of Lyapounov*, Arch. Rational Mech. Anal., 17 (1964), pp. 272–277.

[7] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

[8] ———, *Necessary conditions for minimum in relaxed variational problems*, Ibid., 4 (1962), pp. 129–145.

[9] C. BERGE, *Topological Spaces*, Macmillan, New York, 1963.

[10] J. M. HOLTZMAN, *On the maximum principle for nonlinear discrete-time systems*, IEEE Trans. Automatic Control, (1966), to appear.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

A CLASS OF ITERATIVE PROCEDURES FOR LINEAR INEQUALITIES*

YU-CHI HO AND R. L. KASHYAP†

1. Introduction and notations. In this paper we are concerned with the problem of finding α (m-vector) such that $A\alpha > 0$, where A is a given $N \times m$ matrix and $N > m$. This problem is fundamental in mathematical programming, switching theory, and pattern classification. We shall demonstrate a class of exponentially convergent and finite iterative procedures for solving this problem.

Matrices will be denoted by upper case letters; vectors, lower case letters. The subscript i will indicate the iteration number. $\|M\|$ denotes the spectral norm of M, and $M > 0$ means that M is positive definite and symmetric.

2. The algorithm. It is clear that the problem in §1 can be restated as: Find α (m-vector) and β (N-vector) such that $A\alpha - \beta = 0$ and $\beta > 0$. We shall see later that the introduction of the vector β as additional variables plays a crucial role in the convergence rate of the algorithm without any appreciable increase in computational complexity. Let us define

$$y = A\alpha - \beta \tag{1}$$

and ρ as a scalar constant, S as an $m \times m$ symmetric matrix to be specified later.

PROPOSITION. *The algorithm*

$$\alpha_{i+1} = \alpha_i + \rho SA^T|y_i|, \quad \alpha_0 \textit{ arbitrary}, \tag{2}$$

$$\beta_{i+1} = \beta_i + (y_i + |y_i|), \quad \beta_0 > 0 \textit{ but arbitrary otherwise}, \tag{3}$$

converges to the solution of the problem in a finite number of steps provided a solution exists.

3. Proof of convergence. To show convergence, we must demonstrate that (2) and (3) imply that $\lim_{i\to\infty} y_i = 0$. From (2) and (3), we have

$$y_{i+1} = (\rho ASA^T - I)|y_i|. \tag{4}$$

Consider a Lyapunov function $V(y_i) = \|y_i\|^2$; then

* Received by the editors (in summary form) June 3, 1965, and in complete form September 9, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 16, 1965.

† Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts. This work was supported in part by the Radiation Center, Honeywell, Incorporated, and by the Joint Electronics Program under Contract NONR 1866(16).

$$\begin{aligned}\Delta V(y_i) &\triangleq V(y_{i+1}) - V(y_i)\\ &= |y_i|^T(\rho^2 ASA^TASA^T - 2\rho ASA^T)|y_i|\\ &= |y_i|^TA[\rho^2 SA^TAS - 2\rho S]A^T|y_i|.\end{aligned} \tag{5}$$

By a well-known theorem of Ky Fan [1] we know that $A^T|y_i| \neq 0$ for all $y_i \neq 0$ if a solution exists to the problem $A\alpha > 0$. Hence, in (5) we need only to choose ρ and S appropriately to insure

$$[\rho^2 SA^TAS - 2\rho S] < 0. \tag{6}$$

Case 1.* $(A^TA)^{-1} = S$, $0 < \rho < 2$. In this case, (6) becomes

$$[\rho^2 SA^TAS - 2\rho S] = \rho(\rho - 2)(A^TA)^{-1} < 0. \tag{7}$$

Case 2. $S = I$, $0 < \rho < (\|A^TA\|)^{-1}$. As before, we have for (6),

$$\rho[\rho A^TA - 2I] < 0, \tag{8}$$

which is negative definite by definition of ρ.

Case 3. $S = (2\,\|A^TA\|\,I - A^TA)/\|A^TA\|^2$, $0 < \rho < 2$. In this case, (6) reduces to

$$\begin{aligned}[\rho^2 SA^TAS - 2\rho S] = \rho\Bigg[&(\rho - 2)S - \rho\left(\frac{A^TA}{\|A^TA\|} - I\right)\\ &\cdot S\left(\frac{A^TA}{\|A^TA\|} - I\right)\Bigg],\end{aligned} \tag{9}$$†

which is negative definite in view of the fact that $S > 0$.

In (3) if we let $\beta_0{}^T = [1, 1, \cdots, 1]$, then every component of β_i is greater than one for all i. Thus $|y_i| < 1$ implies that $A\alpha_i > 0$. Since y converges to zero in infinite time, it follows that y must enter the unit cube which represents solutions of $A\alpha > 0$ in finite time. Our proposition is proved.

Remark. Computationally, Case 1 can be expected to have the fastest convergence rate. However, it is necessary to invert an $m \times m$ matrix *once* per problem. Case 2 is simplest but probably slowest in convergence.

* The case where A is not of maximal rank will be treated in the Appendix.

† Define matrix L by $A^TA = \|A^TA\|[I + L]$. Then $S = [I - L]/\|A^TA\|$ and

$$\begin{aligned}[\rho^2 SA^TAS - 2\rho S] &= \rho[(\rho - 2)S + \rho(SA^TAS - S)]\\ &= \rho[(\rho - 2)S + \rho\{(I - L)(I + L)(I - L) - (I - L)\}/\|A^TA\|]\\ &= \rho[(\rho - 2)S - \rho L(I - L)L/\|A^TA\|],\end{aligned}$$

which is equal to the right-hand side of (9).

Case 3 occupies a place in between Cases 1 and 2 in complexity. The spectral norm, $\|A^TA\|$, can be easily computed by the power method.

The scalar ρ, instead of being treated as a constant, can be chosen suitably at each stage so that $-\Delta V(y_i)$ is a maximum. The optimal value of ρ at the ith stage is found to be

$$\rho_i = \frac{|\, y_i \,|^T ASA^T |\, y_i \,|}{|\, y_i \,|^T ASA^T ASA \,|\, y_i \,|}.$$

In particular, for Case 1, $\rho_i = 1$ for all i.

4. Comparison with other algorithms. Let $\beta_0{}^T = [1, 1, \cdots, 1]$. Then the Novikoff procedure [2] for solving the problem can be stated essentially as

$$\alpha_{i+1} = \alpha_i + \rho A^T[\operatorname{sgn}(|A\alpha_i - \beta_0|) - \operatorname{sgn}(A\alpha_i - \beta_0)]. \tag{10}$$

The Agmon-Mays procedure [3], [5] becomes

$$\alpha_{i+1} = \alpha_i + \rho A^T[|A\alpha_i - \beta_0| - (A\alpha_i - \beta_0)] \tag{11}$$

and finally, the Wong-Eisenberg procedure [4],

$$A\alpha_{i+1} = A\alpha_i + \rho A(A^TA)^{-1}A^T[\beta_0 - \operatorname{sgn}(A\alpha_i)], \tag{12}$$

which are variants of Cases 2 and 1, respectively. The key difference in (10)–(12) is the fact that β is *not* treated as a variable but held constant. This difference apparently accounts for the high convergence rate experimentally observed for our procedure [6]. Heuristically, a variable vector β allows those row constraints which are not satisfied to have more weight in the iterative procedure (cf. (2)).

Appendix. Modification of Case 1 when A is not of maximal rank. Let $A^\dagger$ be the Penrose generalized inverse of A. We modify (2) and (3) to

$$\alpha_i = A^\dagger\beta_i, \tag{A-1}$$

$$\beta_{i+1} = \beta_i + \rho[y_i + |y_i|], \quad \beta_0{}^T = [1, \cdots, 1], \tag{A-2}$$

which lead to

$$y_{i+1} = y_i + \rho(AA^\dagger - I)(y_i + |y_i|). \tag{A-3}$$

The Lyapunov function $V(y_i) = \|y_i\|^2$ implies

$$\Delta V(y_i) = 2\rho y_i{}^T(AA^\dagger - I)(y_i + |y_i|) + \rho^2\|y_i + |y_i|\,\|^2_{I-AA^\dagger}, \tag{A-4}$$

which reduces to

$$\Delta V(y_i) = -\|y_i + |y_i|\,\|^2_{[\rho^2 AA^\dagger + (\rho - \rho^2)I]} \tag{A-5}$$

by virtue of the facts

$$(y + |y|)^T(y - |y|) = 0, \qquad A^Ty = A^TAA^\dagger\beta - A^T\beta = 0. \tag{A-6}$$

The matrix $[\rho AA^{\dagger} + (\rho - \rho^2)I]$ is positive definite for $0 < \rho < 1$.

Furthermore, all components of y_i cannot be simultaneously negative. For otherwise, let α^* be a solution of $A\alpha > 0$, then

$$0 = y^T A\alpha^* = y^T b < 0 \tag{A-7}$$

represents a contradiction. Consequently, $\Delta V(y_i)$ is again negative definite. We have convergence once again.

REFERENCES

[1] KY FAN, *On systems of linear inequalities*, Linear Inequalities and Related Systems, Kuhn and Tucker, eds., Annals of Mathematics Studies 38, Princeton University Press, Princeton, 1956.

[2] A. NOVIKOFF, *On convergence proofs for perceptions*, Proceedings of Symposium on Mathematical Theory of Automata, vol. XII, Polytechnic Institute of Brooklyn, 1963, pp. 615–622.

[3] S. AGMON, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1956), pp. 382–392.

[4] E. WONG AND E. EISENBERG, *Iterative synthesis of threshold functions*, J. Math. Anal. Appl., to appear.

[5] C. H. MAYS, *Effect of adaptation parameters on convergence time and tolerance for adaptive threshold elements*, IEEE Trans. Electronic Computers, EC-13 (1964), pp. 465–468; (see also Stanford Electronics Laboratory Report SEL-63-027, TR 1557-1, 1963).

[6] Y. C. HO AND R. L. KASHYAP, *An algorithm for linear inequalities and its applications*, IEEE Trans. Electronic Computers, EC-14 (1965), pp. 683–688

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

AN OPTIMAL PROCEDURE FOR AN N-STAGE LEARNING PROCESS*

W. KARUSH AND R. E. DEAR†

1. Introduction. In this paper we deal with a learning process, or experiment, that involves a finite number N of trials (N arbitrary), where each trial consists of the presentation of a single (stimulus) item chosen out of a given set of n items, $n \leqq N$. In each trial, the subject responds to the presented item, either correctly or incorrectly, and following this, undergoes a reinforcement or corrective action that allows him to improve his state of learning with respect to that item; the response and change of state are taken to occur probabilistically. The sequence of item presentations is under the control of the experimenter, and he is free to follow any strategy of presentation he chooses; a strategy is a procedure that specifies a definite item for each trial, the specification being contingent on the history of presentations and responses up to that trial. We are concerned with the expected level of learning reached by a subject with respect to all n items at the end of the experiment. To measure the terminal level of learning, we define a risk for each strategy, this being the expectation of a function that assigns a numerical loss b_k to the terminal event of being in the unlearned state with respect to exactly k items, $k = 0, 1, 2, \cdots, n$. We then use Bayes' criterion to define an optimal strategy as one that minimizes the risk.

In this paper, we assume a mathematical model of learning that is based upon the so-called single-element model of the stimulus-sampling theory of learning (see [1]). Using this model, we formulate the problem of determining optimal strategies as a type of dynamic programming problem involving branching (correct or incorrect response) at each node (item presentation), which may be viewed in some respects as a generalization of Bellman's "gold-mining" problem [2]. We show that a certain simple, intuitively appealing decision rule is the correct rule for generating optimal solutions.

To describe the decision rule we introduce the vector of probabilities $\mathfrak{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_n)$, where λ_i is the probability of the learned state with respect to item i, or, as we shall say, the probability of "knowing" item i. An initial vector $\mathfrak{\lambda}^0$ is assumed at the outset of the process, and the subsequent values of $\mathfrak{\lambda}$ are computed trial-by-trial on the basis of the model of

* Received by the editors July 2, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Research and Technology Division, System Development Corporation, 2500 Colorado Avenue, Santa Monica, California.

learning. The present value depends upon the immediately preceding $\mathfrak{A}$, the item last presented, and the response elicited. The decision rule is then the following (Theorem 1): present an item j for which the current probability λ_j is least among all the components λ_i. Another way to express this result is: the strategy of local optimization whereby the item presented is that minimizing the risk if the present trial were the last (i.e., looking ahead one step) is in fact a strategy of global optimization.

In addition to assuming dichotomous responses, the single-element model we use also assumes just two possible states, learned or unlearned, with respect to a given item. The state is not observable; what is observable is the response when the item is presented in a trial. The learned state is characterized by the probability of a correct response being 1, and the unlearned by this probability having a value γ, $0 < \gamma < 1$, called the "guessing" probability. The existence of a positive guessing probability means that the test on an item is not perfect in separating unlearned subjects from learned ones (although, as we shall see, it does imply that incorrect responders are necessarily unlearned). It is this aspect of the model that gives content to the problem of an optimal strategy; if the test separated perfectly, then we would never re-present an item to a subject who had once given a correct response to that item.

The remaining parameters of the model, other than the initial vector $\mathfrak{A}^0$ and the guessing probability γ, are the "learning rates" θ_1 and θ_0, which have the following meanings: given that the subject responded correctly in a trial, θ_1 is the probability of a transition from the unlearned to the learned state with respect to the item presented (as a result of the corrective action during the trial); θ_0 is this transition probability given that the subject responded incorrectly. In general, to apply the above decision rule for an optimal strategy to a particular subject we must know the values of the parameters $\mathfrak{A}^0$, γ, θ_1, θ_0 for the subject. The determination of these values raises practical difficulties, particularly the determination of the learning rates, which can be expected to vary significantly as individual attributes. We are rescued from this difficulty by the fact that typical practice is to assume that the initial probabilities λ_i^0 are all 0; under this assumption (and certain mild restrictions) there is a fixed strategy independent of θ_1, θ_0 (and γ), which is optimal for arbitrary values of these parameters. This strategy is given by a counting rule depending only on the responses of the subject (see Theorem 2).

The present work grew out of some laboratory experiments that were designed to test the stimulus-sampling theory with human subjects in paired-associate learning. The experiments were such as to allow the assumption $\theta_1 = \theta_0$; also, $\mathfrak{A}^0$ was taken as zero. The intention was to compare the performance of subjects when taught by a strategy that was optimal

(if the model held) with their performance when a strategy of random presentations was used. This approach required knowing what an optimal strategy was, and it had been conjectured and assumed that the decision rule described above gave an optimal strategy. Our results verify this conjecture. It might be mentioned that the evidence derived from the experiments was inconclusive in validating the learning model. The assumed single-element model is too simple to explain the learning actually taking place in these experiments, and it seems that an extension of the model to include a forgetting rate might be called for. We have not been successful so far in treating this extension and shall not enter into a discussion of it here.

We conclude this introduction by a reference to the treatment of the "two-armed bandit" problem given by Feldman [3]. He establishes an optimal strategy for the sequence in which one should play the two arms of the machine when the arms have different and unknown probabilities of winning. Although the problem is formally quite different from ours, there is an analogy with our two-item case with respect to the methods of proof.

2. Model of learning. Consider a single item. As mentioned above, there are two possible states with respect to the item, learned or unlearned. The state of a subject is unobservable, but his response, correct or incorrect, upon presentation of the item in a trial is observable. For a given trial, we assume that the probability of a correct response depends only upon his state at the outset of the trial and is given by the conditional probabilities

$$\begin{aligned} \text{prob (correct} \mid \text{learned)} &= 1, \\ \text{prob (correct} \mid \text{unlearned)} &= \gamma, \qquad 0 < \gamma < 1. \end{aligned}$$

The parameter γ is the "guessing" probability; using the notation

$$\gamma^* = 1 - \gamma, \tag{1}$$

we note that γ^* is the probability of an incorrect response in the unlearned state.

As a result of the corrective action that is applied during a trial, following the response of the subject, the subject may "learn" the item, i.e., make the transition from the unlearned to the learned state with respect to the presented item. The nature or "strength" of the corrective action is allowed to differ with the response to the item, and consequently the transition probability is allowed to depend on which response is forthcoming. We assume that the probability of a transition to the learned state depends only upon the state at the outset of the trial and the response during the trial, and we take

$$\text{prob (learned | unlearned)} = \begin{cases} \theta_0 \text{ given an incorrect response,} \\ \theta_1 \text{ given a correct response,} \end{cases}$$

$$0 < \theta_0 < 1, \; 0 \leqq \theta_1 < 1,$$

$$\text{prob (learned | learned)} = 1.$$

(In the second case, when the subject "knows" the item at the outset of the trial, we need not distinguish between the two responses, since he then responds correctly with probability 1.) The parameters θ_0, θ_1 are the "learning rates." For later convenience we introduce

$$\bar{\theta} = \gamma\theta_1 + \gamma^*\theta_0 .$$

In contrast to incremental models of learning, the stimulus-sampling model we use is referred to as an all-or-none model of learning—there are only two states, the jump from the unlearned to the learned state is made in a single trial, and once in the learned state the subject thereafter remains in that state.[1]

Suppose the subject enters a trial with probability λ of knowing the item presented. We think of his response as taking him along either one of two branches, the correct-response or incorrect-response branch. In accordance with the definition of γ, he will follow the correct branch with probability $\lambda + \gamma\lambda^*$ and the incorrect one with probability $\gamma^*\lambda^*$. We now compute the probability of the learned state at the end of the trial, conditioned on his response. Consider the case when he responds correctly. There are then two ways he might wind up in the learned state. First, he might have been in the learned state at the outset (then necessarily responded correctly and remained in the learned state); this occurs with probability λ. Second, he might have been in the unlearned state, responded correctly, and then made the transition to the learned state; this occurs with probability $\theta_1\gamma\lambda^*$. Since a correct response occurs with probability $\lambda + \gamma\lambda^*$, we find that the required probability, given a correct response, is

$$\lambda' = \frac{\lambda + \gamma\theta_1\lambda^*}{\lambda + \gamma\lambda^*} ; \tag{2}$$

λ^* is defined according to (1). Similarly, by considering an incorrect response, we find that the a posteriori probability of the learned state, given an incorrect response, is θ_0.

If the same item were presented in the next trial, there would be two branches emerging from each of the terminal nodes above, making four

[1] In a "forgetting" model we would allow transitions from the learned to the unlearned state during a trial for items that were not presented in that trial.

terminal nodes in all, or equivalently, making four two-branch paths beginning at the initial node. The four paths are correct-correct, correct-incorrect, incorrect-correct, and incorrect-incorrect; the corresponding probabilities of being traversed are $(\lambda + \gamma\lambda^*)(\lambda' + \gamma(\lambda')^*)$, $(\lambda + \gamma\lambda^*)\cdot\gamma^*(\lambda')^*$, $(\gamma^*\lambda^*)(\theta_0 + \gamma\theta_0^*)$, and $(\gamma^*\lambda^*)(\gamma^*\theta_0^*)$.

We digress briefly to note some immediate properties of the mapping from λ to λ', $0 \leqq \lambda \leqq 1$, given by (2). If $\lambda = 0$ or 1, then $\lambda' = \theta_1$ or 1, respectively; when $0 < \theta_1 < 1$, λ' is strictly increasing with λ and we have $\lambda < \lambda' < 1$ for λ in the range $0 \leqq \lambda < 1$.

Now let us turn to n items, $i = 1, 2, \cdots, n$, $n \geqq 2$. We suppose that the parameters γ, θ_0, θ_1 which describe the single-trial response and transition probabilities for an individual item are the same for all items. Also, we assume that response and learning for any item is independent of the state with respect to any other item. Let $\mathfrak{\lambda}^0 = (\lambda_1^0, \lambda_2^0, \cdots, \lambda_n^0)$ denote the vector of initial probabilities, i.e., let

λ_i^0 be the probability of being in the learned state with respect to item i at the outset of the experiment, $\quad i = 1, 2, \cdots, n.$

Let $\mathfrak{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_n)$ be the corresponding vector at the beginning of any trial and suppose that item i is presented during the trial. Then $\lambda_i + \gamma\lambda_i^*$ and $\gamma^*\lambda_i^*$ are the probabilities of a correct and incorrect response respectively, and the vectors of probabilities at the end of the trial are

$$(\lambda_1, \cdots, \lambda_i', \cdots, \lambda_n) \text{ for the correct branch,}$$

$$(\lambda_1, \cdots, \theta_0, \cdots, \lambda_n) \text{ for the incorrect branch.}$$

These become the initial vectors for the next trial. Suppose item j, $j \neq i$, is presented in the next trial to correct responders and item k, $k \neq i, j$, to incorrect responders. There are then four possible response paths of two branches each; these are (correct on i, correct on j), (correct on i, incorrect on j), (incorrect on i, correct on k), (incorrect on i, incorrect on k) The respective probabilities of occurrence are $(\lambda_i + \gamma\lambda_i^*)(\lambda_j + \gamma\lambda_j^*)$, $(\lambda_i + \gamma\lambda_i^*)(\gamma^*\lambda_j^*)$, $(\gamma^*\lambda_i^*)(\lambda_k + \gamma\lambda_k^*)$, $(\gamma^*\lambda_i^*)(\gamma^*\lambda_k^*)$, and the respective probabilities of knowing the items i, j, k are $(\lambda_i', \lambda_j', \lambda_k)$, $(\lambda_i', \theta_0, \lambda_k)$, $(\theta_0, \lambda_j, \lambda_k')$, $(\theta_0, \lambda_j, \theta_0)$; the probabilities for the other items are unchanged.

We may visualize an N-trial experiment as a tree structure with a single initial node and 2^N terminal nodes. The tree contains $2^N - 1$ $(= 1 + 2 + 2^2 + \cdots + 2^{N-1})$ nonterminal nodes in all, and at each of these, one of the n items is to be chosen; a branch emerging from a node indicates the response to the item chosen for that node. A presentation strategy is given when the item to be used at each nonterminal node is given. There are

n^{2^N-1} strategies in all. Each terminal node of the tree of a strategy determines a unique path from the initial node to the given terminal node, and this corresponds to a definite sequence of alternating responses and item presentations. The probability of reaching the terminal node is well defined as the product of the probabilities of occurrence of the individual response branches making up the path to the node. For later use, we let

$$q(h) \text{ be the probability of reaching terminal node } h,$$

where the terminal nodes are indexed in some order by $h = 1, 2, \cdots, 2^N$.

3. Loss function and risk. Since a subject may be in either of two states with respect to an individual item at the end of the experiment, he may occupy any one of 2^n possible terminal joint states with respect to all n items. However, we wish to give equal weight to the various items in assessing the terminal effects of a strategy of presentation and so we distinguish only the following $n + 1$ terminal joint states:

T_k is the event of being in the unlearned state with respect to exactly k items, $\quad k = 0, 1, 2, \cdots, n.$

We assign a numerical loss b_k to the event T_k, and, since the greater k the less desirable the event, we assume

$$b_0 \leqq b_1 \leqq b_2 \leqq \cdots \leqq b_n \quad \text{with} \quad b_0 = 0, b_1 = 1. \tag{3}$$

The normalization $b_0 = 0$, $b_1 = 1$ is assumed merely as a notational convenience.

Consider a terminal node with vector of probabilities $\mathfrak{y} = (\mu_1, \mu_2, \cdots, \mu_n)$ of knowing the n items. Let $p_k = p_k(\mathfrak{y})$ be the probability of T_k. The generating polynomial,

$$f(t) = (\mu_1 + t\mu_1^*)\cdot(\mu_2 + t\mu_2^*)\cdots(\mu_n + t\mu_n^*) = \sum_{k=0}^{n} p_k t^k, \tag{4}$$

gives the p_k in terms of the μ_i. The loss associated with the terminal vector $\mathfrak{y}$ is then taken as

$$L(\mathfrak{y}) = p_1 + b_2 p_2 + \cdots + b_n p_n. \tag{5}$$

Consider a strategy S_N with its terminal nodes $h = 1, 2, \cdots, 2^N$ and associated terminal vectors $\mathfrak{y}(h)$; we define its risk as the expected value

$$R_N(\mathfrak{x}^0; S_N) = \sum_h q(h) L(\mathfrak{y}(h)); \tag{6}$$

here, for simplicity, we have omitted the parameters γ, θ_0, θ_1 from the argument of the risk. An optimal strategy for given $\mathfrak{x}^0$ is one that minimizes

(6); its risk is

$$\rho_N(\mathfrak{A}^0) = \min_{S_N} R_N(\mathfrak{A}^0; S_N). \tag{7}$$

Two special loss functions are of practical interest. One occurs when we take $b_j = 1, j = 1, 2, \cdots, n$. Then (5) becomes

$$L = \sum_{j=1}^{n} p_j = 1 - p_0 = 1 - \mu_1\mu_2 \cdots \mu_n .$$

Thus, in this case, an optimal strategy is one that maximizes the expected product of the terminal probabilities, i.e., that maximizes the probability of knowing all n items. The other special case occurs when $b_j = j$. Then (5) becomes, with the help of (4),

$$L = \sum_{j=1}^{n} jp_j = \frac{df}{dt}\bigg|_{t=1} = \sum_{i=1}^{n} \mu_i^* = n - \sum_{i=1}^{n} \mu_i .$$

In this case, an optimal strategy is one that maximizes the expected sum of the terminal probabilities of knowing the items. Our result on optimal strategies is valid for the general case (3) and hence for each of these special cases.

4. Minimum risk function. Define ρ_0 by $\rho_0(\mathfrak{A}) = L(\mathfrak{A})$, where L is given by (5) and the p_j by (4) with $\mathfrak{y}$ replaced by $\mathfrak{A}$. This, together with (7), defines the minimum risk function ρ_N for all $N \geqq 0$. For arbitrary initial $\mathfrak{A}$, let

$\rho_N^{\,i}(\mathfrak{A})$ be the minimum risk relative to all strategies that use item i in the first trial,

and

$\rho_N^{\,ij}(\mathfrak{A})$ be the minimum risk relative to all strategies that use item i in the first trial and item j in the second.

(Notice that the second type of strategy is special in that the same item is used in the second trial regardless of the response in the first trial.) In this section we shall derive some properties of these minimum risk functions.

By the symmetry of the problem with respect to the n items, we have

$$\rho_N(\lambda_1, \lambda_2, \cdots, \lambda_n) = \rho_N(\lambda_{i_1}, \lambda_{i_2}, \cdots, \lambda_{i_n}), \qquad N \geqq 0,$$

for any permutation $\pi = (i_1, i_2, \cdots, i_n)$ of $(1, 2, \cdots, n)$. In fact, let S_N be any strategy and let S_N^{π} be the strategy obtained from S by substituting item i_j for j everywhere in the strategy tree of S_N. Then

$$R_N(\lambda_1, \lambda_2, \cdots, \lambda_n; S_N) = R_N(\lambda_{i_1}, \lambda_{i_2}, \cdots, \lambda_{i_n}; S_N^{\pi}),$$

so that if $\bar{S}_N$ minimizes the risk for initial values $\mathfrak{A}$, then $\bar{S}_N^{\pi}$ minimizes it

for the permuted initial values. The asserted symmetry of ρ_N follows from this. Two other consequences of symmetry are $\rho_N(\lambda, \lambda, \cdots, \lambda) = \rho_N{}^i(\lambda, \lambda, \cdots, \lambda)$ for every i, and

$$\rho_N{}^i(\boldsymbol{\lambda}) = \rho_N{}^j(\boldsymbol{\lambda}) \quad \text{when} \quad \lambda_i = \lambda_j . \tag{8}$$

We next state the basic recursive relation of this paper. Before doing this, we introduce the notational device of using dummy variables $x_1, x_2, \cdots, x_n$ to signify the successive places in the argument of ρ_N and other functions of $\boldsymbol{\lambda}$ (e.g., ρ_N at $x_i = \theta$, $x_j = \lambda_j$ $(j \neq i)$ stands for $\rho_N(\lambda_1, \cdots, \lambda_{i-1}, \theta, \lambda_{i+1}, \cdots, \lambda_n)$). The recursive relation is that for any i,

$$\rho^i_{N+1}(\boldsymbol{\lambda}) = (\lambda_i + \gamma\lambda_i{}^*)\rho_N\,|_{x_i=\lambda_{i'}} + \gamma^*\lambda_i{}^*\rho_N\,|_{x_i=\theta_0}, \qquad N \geqq 0; \tag{9}$$

this follows from the branching process described earlier and the definition of the minimum risk. Note that on the right side of (9) we show only the changes in argument, a practice that we shall follow. Equation (9), together with

$$\rho_{N+1} = \min_i [\rho^i_{N+1}],$$

characterizes an optimal strategy; this shows that our problem is one in dynamic programming. Observe that the recursion (9) remains valid if we adjoin the superscript j everywhere, i.e.,

$$\rho^{ij}_{N+1}(\boldsymbol{\lambda}) = (\lambda_i + \gamma\lambda_i{}^*)\rho_N{}^j\,|_{x_i=\lambda_{i'}} + \gamma^*\lambda_i{}^*\rho_N{}^j\,|_{x_i=\theta_0} .$$

Our work depends in an essential way on the following "commutative" property:[2] for any i, j,

$$\rho^{ij}_{N+2}(\boldsymbol{\lambda}) = \rho^{ji}_{N+2}(\boldsymbol{\lambda}), \qquad N \geqq 0. \tag{10}$$

In the following proof, we suppress the constant arguments λ_k, $k \neq i, j$; also, we may suppose $i \neq j$:

$$\begin{aligned}
\rho^{ij}_{N+2}(\lambda_i, \lambda_j) &= (\lambda_i + \gamma\lambda_i{}^*)\rho^j_{N+1}(\lambda_i', \lambda_j) + \gamma^*\lambda_i{}^*\rho^j_{N+1}(\theta_0, \lambda_j) \\
&= (\lambda_i + \gamma\lambda_i{}^*)\,[(\lambda_j + \gamma\lambda_j{}^*)\rho_N(\lambda_i', \lambda_j') + \gamma^*\lambda_j{}^*\rho_N(\lambda_i', \theta_0)] \\
&\quad + \gamma^*\lambda_i{}^*[(\lambda_j + \gamma\lambda_j{}^*)\rho_N(\theta_0, \lambda_j') + \gamma^*\lambda_j{}^*\rho_N(\theta_0, \theta_0)] \\
&= (\lambda_j + \gamma\lambda_j{}^*)\,[(\lambda_i + \gamma\lambda_i{}^*)\rho_N(\lambda_i', \lambda_j') + \gamma^*\lambda_i{}^*\rho_N(\theta_0, \lambda_j')] \\
&\quad + \gamma^*\lambda_j{}^*[(\lambda_i + \gamma\lambda_i{}^*)\rho_N(\lambda_i', \theta_0) + \gamma^*\lambda_i{}^*\rho_N(\theta_0, \theta_0)] \\
&= (\lambda_j + \gamma\lambda_j{}^*)\rho^i_{N+1}(\lambda_i, \lambda_j') + \gamma^*\lambda_j{}^*\rho^i_{N+1}(\lambda_i, \theta_0) \\
&= \rho^{ji}_{N+2}(\lambda_i, \lambda_j).
\end{aligned}$$

[2] We wish to mention that it was R. E. Bellman who, in the early stages of our work, first pointed out to one of us in a conversation the significance of this property in multistage decision making.

5. Difference function. We define the differences

$$d_N^{ij} = \rho_N{}^i - \rho_N{}^i, \qquad d_N{}^j = \rho_N{}^j - \rho_N, \qquad N \geqq 0, \tag{11}$$

where the argument is λ everywhere. We begin our study of these functions by an explicit computation in the one-step case, $N = 1$.

Let λ_i, λ_{ij} denote the vectors of $n - 1$ and $n - 2$ components, respectively, derived from λ by deletion of λ_i and λ_i, λ_j, $i \neq j$, respectively. Then the quantities $p_k(\lambda_i)$, $k = 0, 1, \cdots, n - 1$, and $p_k(\lambda_{ij})$, $k = 0, 1, \cdots, n - 2$, are well defined, being the probability of not knowing exactly k items exclusive of item i and exclusive of both items i and j, respectively; these probabilities are given as the coefficients of the polynomial $f(t)$ in (4) with the factors $(\lambda_i + t\lambda_i^*)$, and $(\lambda_i + t\lambda_i^*)$, $(\lambda_j + t\lambda_j^*)$ deleted, respectively. We have

$$\begin{aligned} \lambda_i^* p_{k-1}(\lambda_i) + \lambda_i p_k(\lambda_i) &= p_k(\lambda), \qquad k = 1, 2, \cdots, n - 1, \\ \lambda_i^* p_{n-1}(\lambda_i) &= p_n(\lambda). \end{aligned} \tag{12}$$

These equations simply compute the probability of the state T_k in terms of the two alternatives of not knowing or knowing item i; they also follow from (4). Now let

$$F_i = \sum_{k=0}^{n-1} b_{k+1} p_k(\lambda_i), \qquad G_i = \sum_{k=0}^{n-1} b_k p_k(\lambda_i);$$

these are, respectively, the expected losses conditioned on not knowing or knowing item i. We have

$$L(\lambda) = \lambda_i^* F_i + \lambda_i G_i,$$

which can be seen directly in terms of conditional expectations or can be verified by calculation using (12). Using this, as well as (9) with $N = 0$ and the definitions of λ' and $\bar{\theta}$, we find that

$$\begin{aligned} \rho_1{}^i &= (\lambda_i + \gamma\lambda_i^*)L\,|_{x_i=\lambda_i'} + \gamma^*\lambda_i^* L\,|_{x_i=\theta_0} \\ &= (\lambda_i + \gamma\lambda_i^*)((\lambda_i')^* F_i + \lambda_i' G_i) + \gamma^*\lambda_i^*(\theta_0^* F_i + \theta_0 G_i) \\ &= (1 - \bar{\theta})\lambda_i^* F_i + (\lambda_i + \bar{\theta}\lambda_i^*)G_i \\ &= \bar{\theta}^*\lambda_i^* F_i + \bar{\theta}^*\lambda_i G_i + \bar{\theta} G_i \\ &= \bar{\theta}^* L + \bar{\theta} G_i. \end{aligned}$$

Next, we show that

$$G_i = \sum_{k=0}^{n-2} b_{k+1} p_k(\lambda_{ij}) - \lambda_j \sum_{k=0}^{n-2} (b_{k+1} - b_k) p_k(\lambda_{ij}).$$

To do this, we may use the equations

$$p_k(\mathfrak{A}_i) = \lambda_j^* p_{k-1}(\mathfrak{A}_{ij}) + \lambda_j p_k(\mathfrak{A}_{ij}), \quad k = 1, 2, \cdots, n-2,$$
$$p_{n-1}(\mathfrak{A}_i) = \lambda_j^* p_{n-2}(\mathfrak{A}_{ij}),$$

which are analogs of (12); substitution of these relations into the expression defining G_i yields the desired formula. By forming the difference ${\rho_1}^j - {\rho_1}^i$ we obtain the following result.

LEMMA 1. *For any* $\mathfrak{A}$ *and any* $i \neq j$, *we have*

$$d_1^{ij} = K_{ij}\bar{\theta}(\lambda_j - \lambda_i), \tag{13}$$

where

$$K_{ij} = \sum_{k=0}^{n-2} (b_{k+1} - b_k) p_k(\mathfrak{A}_{ij}) \geqq 0. \tag{14}$$

The last inequality is strict, i.e., $K_{ij} > 0$, *in case* (a) *either* (3) *is a chain of strict inequalities, or* (b) $\lambda_k > 0$ *for all* $k \neq i, j$.

Proof. In view of the preceding discussion it remains to justify only the latter statement of the lemma. When (3) is a strict chain, every factor $b_{k+1} - b_k$ in (14) is positive; since the nonnegative factors p_k in (14) must sum to 1 by their definition as probabilities, it follows that K_{ij} is positive. When $\lambda_k > 0$, $k \neq i, j$, then the first term in (14), namely,

$$p_0(\mathfrak{A}_{ij}) = \prod_{\substack{k \\ k \neq i,j}} \lambda_k ,$$

is positive, and we have the same conclusion.

The preceding proof of Lemma 1 was carried out with the implicit assumption that $n \geqq 3$; the same proof applies when $n = 2$ with obvious notational interpretations (e.g., $p_k(\mathfrak{A}_{ij})$ is defined only for $k = 0$ and has the value 1). Notice that this lemma provides a description of the strategy that is locally optimal in the sense of always minimizing the risk relative to the outcome of the current trial—namely, it is the strategy that selects, in the current trial, an item that has a least probability λ_i of the learned state at the outset of the trial.

We next derive a recursion relation for the general difference d_N^{ij}. Starting with the recursion (9) we have

$$\begin{aligned} \rho_{N+1}^i &= (\lambda_i + \gamma\lambda_i^*)\rho_N^j|_{x_i=\lambda_{i'}} + \gamma^*\lambda_i^*\rho_N^j|_{x_i=\theta_0} \\ &\quad - (\lambda_i + \gamma\lambda_i^*)d_N^j|_{x_i=\lambda_{i'}} - \gamma^*\lambda_i^* d_N^j|_{x_i=\theta_0} \\ &= \rho_{N+1}^{ij} - (\lambda_i + \gamma\lambda_i^*)d_N^j|_{x_i=\lambda_{i'}} - \gamma^*\lambda_i^* d_N^j|_{x_i=\theta_0} . \end{aligned}$$

Interchanging i and j gives a similar expression for ρ_{N+1}^j. Then, using the

commutative property (10), we find

$$d_{N+1}^{ij}(\mathfrak{A}) = (\lambda_i + \gamma\lambda_i^*)d_N{}^j|_{x_i=\lambda_{i'}} + \gamma^*\lambda_i^* d_N{}^j|_{x_i=\theta_0} - (\lambda_j + \gamma\lambda_j^*)d_N{}^i|_{x_j=\lambda_{j'}} - \gamma^*\lambda_j^* d_N{}^i|_{x_j=\theta_0}. \tag{15}$$

This is the formula we require.

6. Properties of difference function. Our main results are based on properties of the difference functions (11), which we develop in this section. We remark that in the forthcoming we will be using the term "increasing" in the following sense: $g(t)$, say, is increasing in case $t_1 < t_2$ implies $g(t_1) \leqq g(t_2)$ (i.e., increasing means nondecreasing).

Property P_N. Property P_N holds at $\mathfrak{A}$ in case the following is true: for any j, $j = 1, 2, \cdots, n$, let σ and $i \neq j$ satisfy

$$\lambda_i = \sigma = \min_{k\neq j}[\lambda_k]; \tag{16}$$

let λ_j be replaced by the running variable t in $d_N{}^{ij}$ and consider the resulting difference as a function $d_N{}^{ij}(t)$ of t; then

$$d_N{}^{ij}(t) \quad \text{is increasing on} \quad \sigma \leqq t \leqq 1. \tag{17}$$

Property $P_N{}^+$. Same as property P_N with "increasing" replaced by "strictly increasing" in (17).

LEMMA 2. *Let* P_N *hold at a given* $\mathfrak{A}$ *and let* i *be such that* $\lambda_i = \min_k[\lambda_k]$. *Then* $\rho_N(\mathfrak{A}) = \rho_N{}^i(\mathfrak{A})$.

Proof. Let $\mathfrak{A}$ and i be as described. For arbitrary $j \neq i$, $\lambda_j \geqq \lambda_i$, property P_N gives

$$d_N{}^{ij}(\mathfrak{A}) \geqq d_N{}^{ij}|_{x_j=\lambda_i}.$$

But the right-hand term vanishes by (8). Thus, $\rho_N{}^j \geqq \rho_N{}^i$, which establishes the lemma.

Our aim is to show that property P_N holds for all $\mathfrak{A}$. This is done by induction on N in the next two lemmas.

LEMMA 3. *Suppose* N *is such that* P_N *holds for all* $\mathfrak{A}$. *For each* $\mathfrak{A}$, *let* $d_N{}^j(t)$ *denote the (nonnegative) difference* $d_N{}^j = \rho_N{}^j - \rho_N$ *regarded as a function of* $\lambda_j = t$ *alone. Then* $d_N{}^j(t)$ *is an increasing function on* $0 \leqq t \leqq 1$.

Proof. Given j, let σ and $i \neq j$ satisfy (16). By Lemma 2, $\rho_N = \rho_N{}^j$ on $0 \leqq t \leqq \sigma$, $\rho_N = \rho_N{}^i$ on $\sigma \leqq t \leqq 1$. Thus, $d_N{}^j(t) = 0$ or $d_N{}^{ij}(t)$ on these respective ranges. The lemma now follows from (17).

LEMMA 4. *For every* $N \geqq 1$, *property* P_N *holds for all* $\mathfrak{A}$.

Proof. The proof is by induction. The lemma is valid for $N = 1$ by Lemma 1. Now consider any N for which the lemma holds. Given $\mathfrak{A}$ and j, let σ and i satisfy (16). Consider the recursion relation (15). By the induc-

tive assumption and Lemma 2, the third term on the right vanishes for $\lambda_j \geqq \sigma$ (since $\lambda_j \leqq \lambda_j'$); thus

$$d_{N+1}^{ij}(t) = \gamma^*(t-1)d_N{}^i|_{\substack{x_i=\sigma \\ x_j=\theta_0}} + (\sigma + \gamma\sigma^*)d_N{}^j(t)|_{x_i=\sigma'} + \gamma^*\sigma^* d_N{}^j(t)|_{x_i=\theta_0} \quad \text{for } \sigma \leqq t \leqq 1. \tag{18}$$

On the right, the first term is increasing in t by its explicit form while the second and third terms are increasing by Lemma 3. Thus, the left side is increasing on $\sigma \leqq t \leqq 1$, which establishes P_{N+1}. This concludes the proof.

Lemmas 3 and 4 have their counterparts with respect to the stronger property $P_N{}^+$.

LEMMA 5. *Let $P_N{}^+$ hold at $\mathfrak{A}$. Then for each j, $d_N{}^j(t)$ vanishes on $0 \leqq t \leqq \sigma$ and is strictly increasing on $\sigma \leqq t \leqq 1$; here σ is defined by* (16).

Proof. By Lemma 4, we see that the conclusion of Lemma 2 always holds. From this, we may verify that the proof of Lemma 3 carries over to the present lemma.

LEMMA 6. *For every $N \geqq 1$, property $P_N{}^+$ holds for all $\mathfrak{A}$ with $\lambda_i \geqq \theta_0$, $i = 1, 2, \cdots, n$.*

Proof. The lemma is valid for $N = 1$ by Lemma 1 (we have $K_{ij} > 0$ in Lemma 1, since $\lambda_k \geqq \theta_0 > 0$ for all k). Now let the lemma hold for a particular N and consider $N + 1$. Select any $\mathfrak{A}$ with all $\lambda_k \geqq \theta_0$ and any j; choose i and σ to satisfy (16). We have $\sigma \geqq \theta_0$. By Lemmas 4 and 2, we deduce (18). The first term on the right in (18) is increasing by explicit form, and the second term is increasing by Lemma 3. The third term may be written $\gamma^*\sigma^* d_N{}^{ij}(t)$ for $\theta_0 \leqq t \leqq 1$ where the difference is evaluated for original components $\lambda_k \geqq \theta_0$, new component $\lambda_i = \theta_0$, and $\lambda_j = t$. By the inductive assumption, this term is strictly increasing on $\theta_0 \leqq t \leqq 1$. Thus, the left side of (18) is strictly increasing on $\sigma \leqq t \leqq 1$. This completes the proof.

7. Optimal strategies. We are now in a position to establish our results on optimal strategies. The first theorem is an immediate consequence of Lemmas 4 and 2 and the recursion relation (9).

THEOREM 1. *Consider an N-trial experiment with arbitrary N and arbitrary initial probabilities $\mathfrak{A}^0 = (\lambda_1{}^0, \lambda_2{}^0, \cdots, \lambda_n{}^0)$ of being in the learned state relative to the items $i = 1, 2, \cdots, n$. Then an optimal strategy is determined by the rule of presenting in any trial an item for which the probability of being in the learned state at the outset of the trial is least among all items.*

The next theorem is concerned with the question of characterizing optimal strategies, while Theorem 1 deals only with a sufficient condition for an optimal strategy. For simplicity of statement we consider only the most important case in which the initial probabilities $\lambda_i{}^0$ are all zero; in this case,

the decision procedure can be simplified and expressed in terms of a simple counting rule, as we shall see.

From Theorem 1 we know that if we present the n items in sequence (in any order whatever) in the first n steps, regardless of the subject's responses, then we are initiating a possible optimal strategy. From a practical point of view it is natural to limit consideration to such strategies, and we shall restrict ourselves to this class of strategies in the next theorem. We impose the further minor modification of taking the learning rate to be the larger value θ_0 in these initial steps regardless of the response of the subject. In terms of the description of the model given in the Introduction, this amounts to the reasonable procedure of applying the "stronger" corrective action the first time that an item is presented, regardless of response, and only discriminating between corrective actions thereafter.

THEOREM 2. *Consider an N-trial experiment, N arbitrary, with initial probabilities λ_i^0 all zero. Consider strategies that present the n items in (an arbitrary) sequence in the first n trials, and assume that in these (but only in these) trials the learning rate is θ_0 regardless of the response. A strategy is optimal in this class if and only if it conforms to the following rule: beginning with trial $n + 1$ associate with each item a count whose initial value is 0; in any trial choose an item for presentation whose count is minimal at the outset of the trial; at the end of a trial, increase the count of the presented item by 1 if the response is correct but set it back to 0 if the response is incorrect.*

Proof. Consider any one of the first n trials, and let item i be presented in the trial. We have $\lambda_i^0 = 0$ at the outset of the trial, by assumption. This probability will increase to θ_0 at the end of the trial regardless of the response. That is to say, if the response is correct, then the new value is given by (2) with $\lambda = 0$ and with θ_0 in place of θ_1, and this yields θ_0; if the response is incorrect, then the new value becomes θ_0 directly by the single-trial branching process. Thus, at the end of the first n trials we have the probability vector $\mathfrak{\lambda} = (\theta_0, \theta_0, \cdots, \theta_0)$ regardless of the pattern of responses (the tree of the process reduces to a simple linear chain for the first n steps).

We may now see that a continuation of such a strategy is optimal if and only if it conforms to the rule of Theorem 1; for from trial $n + 1$ on, we have all $\lambda_i \geqq \theta_0$ and, hence, by Lemma 6 we have property P_M^+ holding for any number of trials M remaining to the end of the experiment. The strict monotonicity of $d_M{}^j(t)$ in Lemma 5 then shows that the rule in question is necessary for optimality, as well as sufficient.

To establish the counting rule, we define $\theta_0^{(k)}$, $k = 0, 1, 2, \cdots$, inductively by

$$\theta_0^{(0)} = \theta_0, \qquad \theta_0^{(k+1)} = (\theta_0^{(k)})',$$

where λ' is defined by (2). At the outset of any trial from trial $n + 1$ onward, each component of the vector $\mathfrak{A}$ will have the form $\theta_0^{(k)}$ for some k. Now interpret k as the count of an item. The rule of Theorem 2 for changing counts from trial to trial is then the correct rule for computing the probability of the learned state at the end of a trial, in accordance with the single-trial branching process; further, the rule for selection of an item is the same as the rule of Theorem 1 because $\theta_0^{(k)}$ is a strictly increasing function of k. This completes the proof of the theorem.

To apply the decision rule of Theorem 1 to a given subject, it is necessary to know, among other things, the values of the learning rates θ_0 and θ_1 for that subject. Since these parameters may vary significantly from subject to subject, and it cannot be reasonably assumed in practice that they will be known or can be readily determined, it becomes important to consider strategies that are independent of θ_0 and θ_1. A natural way to state the problem of an optimal strategy in this case is the following. Let $R_N(\mathfrak{A}^0, \theta_0, \theta_1 ; S_N)$ denote the risk of a strategy S_N, where we exhibit explicitly the dependence on the model parameters (but omit the guessing probability γ for simplicity). What Theorem 1 does is to construct for each $\mathfrak{A}^0$, θ_0, θ_1 a strategy $\bar{S}_N(\lambda^0, \theta_0, \theta_1)$ that minimizes R_N. Suppose that a probability distribution is given over pairs of values (θ_0, θ_1), and consider the expectation $r_N(\mathfrak{A}^0; S_N)$ of R_N with respect to these variables. We view the new problem as the one that attempts to minimize r_N relative to S_N.

The latter problem seems to be a difficult one to solve. Fortunately, in the most important case of $\mathfrak{A}^0 = 0$, the problem can be handled (for strategies S_N that are initiated as in Theorem 2). Observe that the counting rule for an optimal strategy given in Theorem 2 specifies a strategy tree that is *independent of the learning rates*; it gives a fixed $\bar{S}_N$ such that for any S_N,

$$R_N(\theta_0, \theta_1 ; \bar{S}_N) \leqq R_N(\theta_0, \theta_1 ; S_N).$$

(We have omitted the argument $\mathfrak{A}^0 = 0$.) Thus, the strategy $\bar{S}_N$ does more than minimize the expected risk r_N ; it minimizes the risk R_N uniformly in the parameters θ_0 and θ_1. We conclude by noting that $\bar{S}_N$ is also independent of γ and hence minimizes the risk uniformly in γ as well as θ_0 and θ_1.

REFERENCES

[1] R. C. Atkinson and W. K. Estes, *Stimulus sampling theory*, Handbook of Mathematical Psychology, II, R. R. Bush, E. Galanter, and R. D. Luce, eds., John Wiley, New York, 1963.

[2] R. E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[3] D. Feldman, *Contributions to the "two-armed bandit" problem*, Ann. Math. Statist., 33 (1962), pp. 847–856.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

THE DUOPLEX METHOD IN NONLINEAR PROGRAMMING*

H. P. KÜNZI†

1. Introduction. In a previous paper [1] H. Tzschach and the author discussed the duoplex algorithm in linear programming. It is a method which is convenient to use in linear programming when the number of restrictions is large.

The first step in the duoplex method is to try and come as close as possible to the optimum point. This is achieved by determining the restriction whose normal, which is pointing into the convex region, has the largest angle with the gradient of the linear objective function. This will be referred to as the main restriction. In the two-dimensional case, it is easy to show that the optimum lies on the main restriction. This is not necessarily so for higher dimensions, that is, when $n > 2$. (See Fig. 1.)

The second step in the duoplex method consists of determining the optimum point using a method which is similar to the simplex one.

When the optimum lies on the main restriction, it can be determined very quickly. However, if it does not—which can happen if the main restriction is redundant—the duoplex algorithm is still useful.

It is worth noting that in the duoplex method, it is not required to operate always with permissible solutions, that is, departure from the feasible region is permitted. The multiphase method for determining a feasible solution plays an important role in the second step. See also [2].

The duoplex method referred to has been programmed on an electronic computer and a few hundred examples have been solved. In the majority of cases a considerable reduction in computation time has been noted compared to the standard methods.

As previously mentioned, the duoplex method is to be recommended when the linear programming problem has a large number of restrictions. In such a case the first step is especially economical.

Due to this reason, we consider extending the duoplex method to nonlinear convex programming. This we achieve by linearizing the nonlinear expressions (objective function and/or restrictions) using a specially convenient method and then solving the problem with duoplex. In many cases, as a consequence of this linearization, a very large number of linear restrictions result, so that if a standard method is used it becomes tedious to obtain the optimum.

* Received by the editors June 25, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Rechenzentrum der Universität Zürich, Zürich, Switzerland.

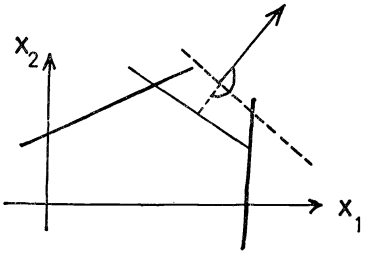

FIG. 1

In this paper some nonlinear duoplex cases are considered and are discussed in detail. The ideas concerning the linear duoplex are based on the work previously cited and knowledge of which is a prerequisite for what follows.

2. The λ-algorithm and the duoplex method. In this section we limit ourselves to problems which lead to separable functions, and to perform the linearization we will use the well known algorithm which is basically based on the Charnes and Lemke's methods [3]. Compare also the description by Hadley in his work on nonlinear and dynamic programming [4]. Starting with the nonlinear optimization problem:

We maximize

$$z = \sum_{j=1}^{n} a_j(x_j), \tag{2.1}$$

with respect to the restrictions

$$\sum_{j=1}^{n} a_{ij}(x_j) \leqq a_{i0}, \qquad i = 1, \cdots, m, \tag{2.2}$$

and

$$x_j \geqq 0, \qquad j = 1, \cdots, n. \tag{2.3}$$

As we mentioned, we require that all the functions that occur be separable, and to have a single optimum we require these functions to be convex (or concave) and also to be differentiable.

As an illustration, let $f(x)$ be such a function (compare Fig. 2) with x varying in the interval $0 \leqq x \leqq a$. This interval is to be divided into $\nu + 1$ points x_k. Then we determine $f_k = f(x_k)$ and consider for each k the connection between

$$(x_k, f_k) \quad \text{and} \quad (x_{k+1}, f_{k+1}).$$

Thus we obtain the dashed lines $\hat{f}(x)$ shown in Fig. 2.

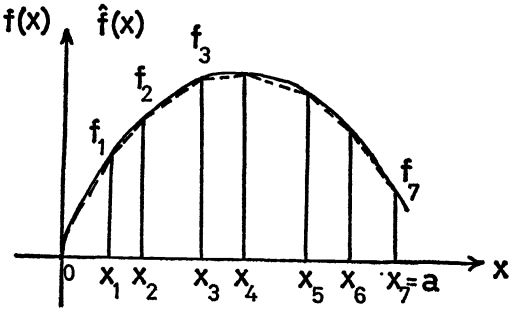

FIG. 2

If x lies within the interval $x_k \leqq x \leqq x_{k+1}$, then the following expression for the linear approximation is valid:

$$\hat{f}(x) = f_k + \frac{f_{k+1} - f_k}{x_{k+1} - x_k}(x - x_k). \tag{2.4}$$

Returning to the original problem (2.1)–(2.3) and linearizing the functions $a_j(x_j)$ and $a_{ij}(x_j)$ by subdividing each interval in which x_j varies into a number of points x_{kj}, the problem becomes the linear programming one:

We maximize

$$\hat{z} = \sum_{j=1}^{n} \hat{a}_j(x_j), \tag{2.5}$$

with respect to

$$\sum_{j=1}^{n} \hat{a}_{ij}(x_j) \leqq a_{i0}, \qquad i = 1, \cdots, m, \tag{2.6}$$

and

$$x_j \geqq 0, \qquad j = 1, \cdots, n. \tag{2.7}$$

Returning to Fig. 2 we recognize for every x in the interval $x_k \leqq x \leqq x_{k+1}$ that

$$x = \lambda x_{k+1} + (1 - \lambda)x_k, \quad \text{for} \quad 0 \leqq \lambda \leqq 1,$$

and furthermore

$$x - x_k = \lambda(x_{k+1} - x_k).$$

Similarly,

$$\hat{f}(x) = \lambda f_{k+1} + (1 - \lambda)f_k.$$

Substituting

$$\lambda = \lambda_{k+1} \quad \text{and} \quad (1 - \lambda) = \lambda_k ,$$

the following unique representation for $x_k \leqq x \leqq x_{k+1}$ is obtained:

$$\begin{aligned} x &= \lambda_k x_k + \lambda_{k+1} x_{k+1} , \\ \hat{f}(x) &= \lambda_k f_k + \lambda_{k+1} f_{k+1} , \end{aligned} \tag{2.8}$$

with

$$\lambda_k + \lambda_{k+1} = 1, \quad \lambda_k \geqq 0, \quad \lambda_{k+1} \geqq 0.$$

Applying the ideas of (2.8) to each variable x_j—this is performed by subdividing the appropriate interval into p_j subintervals having the points x_{kj}—then for each x_j we can write:

$$\begin{aligned} x_j &= \sum_{k=0}^{p_j} \lambda_{kj} x_{kj} , \\ \sum_{k=0}^{p_j} \lambda_{kj} &= 1, \quad \lambda_{kj} \geqq 0, \quad \text{for all } k, j, \end{aligned} \tag{2.9}$$

where for a given j no more than two λ_{kj} must be positive. Such λ_{kj} must in addition be adjacent to each other.

Furthermore, by substituting

$$a_{ij}(x_k) = a_{kij} \quad \text{and} \quad a_j(x_k) = a_{kj} ,$$

problem (2.5)-(2.7) can be formulated in terms of the new variables λ_{kj} as follows

$$\max \hat{z} = \sum_{j=1}^{n} \sum_{k=0}^{p_j} a_{kj} \cdot \lambda_{kj} \tag{2.10}$$

with respect to

$$\sum_{j=1}^{n} \sum_{k=0}^{p_j} a_{kij} \cdot \lambda_{kj} \leqq a_{i0} , \qquad i = 1, \cdots, m, \tag{2.11}$$

$$\sum_{k=0}^{p_j} \lambda_{kj} = 1, \qquad j = 1, \cdots, n, \tag{2.12}$$

and

$$\lambda_{kj} \geqq 0, \text{ for all } k, j. \tag{2.13}$$

It can be proved (compare [4, p. 124]) that, assuming concavity or convexity respectively of $a_j(x_j)$ and $a_{ij}(x_j)$, it is not necessary that in the optimum solution of the above problem for each j, not more than two λ_{kj} be positive and adjacent to one another.

Consequently (2.10)–(2.13) produce a linear optimization problem in $m + n$ restrictions and $(\sum_j p_j) + n$ variables. Thus we have often a problem with a large number of restrictions so that the duoplex method can be profitably used.

In many cases it is more appropriate in problem (2.10)-(2.13) first to go from the primary to the dual because in general the number of variables in the primary increases more than the number of restrictions, so that in the dual the number of restrictions will be even larger. In the following work we will remain in the primary and we leave it to the reader to make analogous reflections about what happens in the dual. In certain cases this can result in further reduction in the computation work.

Some numerical examples will be now presented, in which first linearization using the λ-algorithm is made, then finally the duoplex method is applied to determine the optimum.

3. Three duoplex examples.

Example 1. $n = 2, m = 3;$

maximize $z = x_1,$

under the restrictions

(1) $-x_1^2 - x_2^2 + 25 \geqq 0,$

(2) $-x_1^2 + x_2 + 23 \geqq 0,$

(3) $-x_1 + x_2 + 6 \geqq 0$ (this restriction is redundant!),

$$x_1, x_2 \geqq 0.$$

(Compare Fig. 3.)

For the linearization x_1 and x_2 are taken to be

$$0 \leqq x_1 < 6,$$

$$0 \leqq x_2 < 5,$$

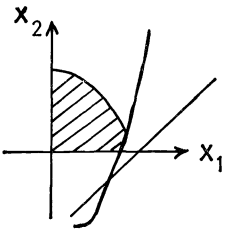

FIG. 3

and as support points in the intervals obtained we take:

$x_1 =$	$x_2 =$
0	0
0.5	0.5
1.0	1.0
1.5	1.5
2.0	2.0
3.0	3.0
4.0	4.0
5.0	5.0
6.0	

This results in 17 variables λ_ν, $\nu = 1, \cdots, 17$, and 2 additional restrictions:

$$\sum_{\nu=1}^{9} \lambda_\nu = 1, \tag{4}$$

$$\sum_{\nu=10}^{17} \lambda_\nu = 1, \qquad \lambda_\nu \geqq 0, \quad \nu = 1, \cdots, 17. \tag{5}$$

Linearized and expressed in terms of the variable λ_ν, the problem becomes:

$$x_1 = 0.5\lambda_2 + 1.0\lambda_3 + 1.5\lambda_4 + 2.0\lambda_5 + 3.0\lambda_6 + 4.0\lambda_7 + 5.0\lambda_8 + 6.0\lambda_9,$$
$$x_2 = 0.5\lambda_{10} + 1.0\lambda_{11} + 1.5\lambda_{12} + 2.0\lambda_{13} + 3.0\lambda_{14} + 4.0\lambda_{15} + 5.0\lambda_{16} + 6.0\lambda_{17},$$

and the objective function z to be maximized:

$$z = 0.5\lambda_2 + \lambda_3 + 1.5\lambda_4 + 2\lambda_5 + 3\lambda_6 + 4\lambda_7 + 5\lambda_8 + 6\lambda_9.$$

The coefficient matrix corresponding to the above five restrictions is (3 inequalities and 2 equations $=0$)

$$\begin{bmatrix}
0 & -.25 & -1 & -2.15 & -4 & -9 & -16 & -25 & -36 & 0 & -.25 & -1 & -2.25 & -4 & -9 & -16 & -25 & 25 \\
0 & -.25 & -1 & -2.15 & -4 & -9 & -16 & -25 & -36 & 0 & .5 & 1 & 1.5 & 2 & 3 & 4 & 5 & 23 \\
0 & -.5 & -1 & -1.5 & -2 & -3 & -4 & -5 & -6 & 0 & .5 & 1 & 1.5 & 2 & 3 & 4 & 5 & 6 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1
\end{bmatrix}.$$

The solution is obtained using the duoplex method in 11 iterations and is:

$$x_1 = 4.889, \quad x_2 = 1.000, \quad \text{and} \quad z = 4.889.$$

The actual optimum is:

$$x_1 = 4.898, \quad x_2 = 1.000, \quad \text{and} \quad z = 4.898.$$

Example 2. $n = 2$, $m = 3$;

maximize $z = x_2$,

under the restrictions

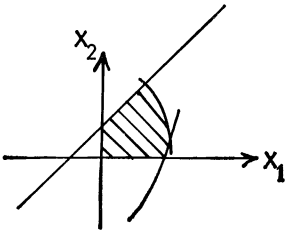

FIG. 4

(1) (as in Example 1),

(2) (as in Example 1),

(3) $x_1 - x_2 + 2 \geqq 0$. (Compare Fig. 4.)

In this example there are no redundant restrictions. The procedure is as in Example 1 and the solution is obtained in 10 iterations using the duoplex method:

$$x_1 = 2.357, \quad x_2 = 4.357, \quad \text{and} \quad z = 4.357.$$

The actual optimum is:

$$x_1 = 2.391, \quad x_2 = 4.391, \quad \text{and} \quad z = 4.391.$$

Example 3. $n = 3, m = 4$;

$$\text{maximize } z = x_1 + 0.5x_3,$$

with the restrictions:

(1) $-x_1^2 - x_2^2 + 25 \geqq 0,$

(2) $-x_1^2 + x_2 + 23 \geqq 0,$

(3) $x_1 - x_2 + 2 \geqq 0,$

(4) $-x_1 - x_2 - x_3 + 16 \geqq 0,$

$$x_1, x_2, x_3 \geqq 0.$$

(Compare Fig. 5.)

Since the variable x_3 appears in a linear form, it does not require any linearization. The following limits are taken for variables x_1 and x_2 :

$$0 \leqq x_1 \leqq 6, \quad 0 \leqq x_2 \leqq 8.$$

Using the same interval subdivision as in Example 1 we obtain:

$$x_1 \rightarrow 9 \quad \text{variables: } \lambda_1 \text{ to } \lambda_9,$$

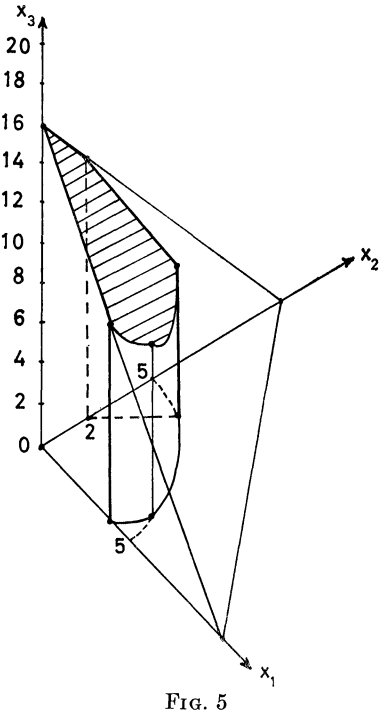

FIG. 5

$x_2 \longrightarrow$ 11 variables: λ_{10} to λ_{20},

$x_3 \longrightarrow$ 1 variable: λ_{21}.

Analogous to Example 1 we obtain as restrictions 4 inequalities and 2 equations and as objective function to be maximized:

$$z = 0.5\lambda_2 - \lambda_3 + 1.5\lambda_4 + 2\lambda_5 + 3\lambda_6 + 4\lambda_7 + 5\lambda_8 + 6\lambda_9 + 0.5\lambda_{21}.$$

As a solution we obtain in 11 iterations:

$$x_1 = 4.778, \quad x_2 = 0.0, \quad x_3 = 11.222, \quad \text{and} \quad z = 10.389.$$

The actual optimum is:

$$x_1 = 4.796, \quad x_2 = 0.0, \quad x_3 = 11.204, \quad \text{and} \quad z = 10.398.$$

To show that there is no disadvantage in exceeding the limit x_p, Example 2 is solved once more using a larger interval for x_2.

Example 4. As in the second, except that $0 \leqq x_2 \leqq 8$, i.e.,

$$x_2 = 0\lambda_{10} + 0.5\lambda_{11} + \lambda_{12} + 1.5\lambda_{13} + 2\lambda_{14} + 3\lambda_{15} + 4\lambda_{16} + 5\lambda_{17} + 6\lambda_{18} + 7\lambda_{19} + 8\lambda_{20}.$$

20 variables are obtained instead of 17.

Using the duoplex method the same solution as in Example 2 is obtained after 12 *iterations*. It is to be noted that the new variables λ_{18}, λ_{19}, λ_{20} may appear in the solution of the fourth problem, however they have no effect on the numerical value of x_r, that is, on the accuracy of the solution.

Example 5. In this example two solutions having different limits for x_2 are compared. The objective function is

$$z = -2x_1 + x_2 ;$$

the restrictions are as in Example 2.

The intervals are

(a) $0 \leqq x_2 \leqq 5$, as in Example 2,

(b) $0 \leqq x_2 \leqq 8$, as in Example 4.

As a solution we obtain after 7 steps in (a) and after 4 steps in (b):

$x_1 = 0.0$, $x_1 = x_1(\lambda_1)$ with (a) and (b);

$x_2 = 2.0$, (a) $x_2 = x_2(\lambda_{13}, \lambda_{17})$;

$z = 2.0$. (b) $x_2 = x_2(\lambda_{10}, \lambda_{19})$.

The exact solution coincides with the above one!

REFERENCES

[1] H. P. KÜNZI AND H. TZSCHACH, *The duoplex algorithm*, Numer. Math., 7 (1965), pp. 222–225.

[2] H. P. KÜNZI, *Die Simplexmethode zur Bestimmung einer Ausgangslösung bei bestimmten linearen Programmen*, Unternehmensforschung, 2 (1958).

[3] A. CHARNES AND C. LEMKE, *Minimisation of nonlinear separable convex functionals*, Naval Res. Logist. Quart., 1 (1954), pp. 301–312.

[4] A. HADLEY, *Nonlinear and Dynamic Programming*, Addison-Wesley, Reading, Massachusetts, 1964.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

SUFFICIENT CONDITIONS FOR THE OPTIMAL CONTROL OF NONLINEAR SYSTEMS*

O. L. MANGASARIAN†

Abstract. It is well known that Pontryagin's maximum principle furnishes necessary conditions for the optimality of the control of a dynamic system. In the present work sufficient conditions for the optimality of the control of a nonlinear system with state and control variable constraints and with fixed initial and terminal times are given. These conditions are essentially Pontryagin's necessary conditions for the same problem, plus some convexity, negativity and strict negativity conditions. The present sufficient conditions subsume the recent results of Lee, wherein sufficient conditions for the optimality of a system, linear in the state variables, were given.

1. Introduction. Consider the following problem in optimal control: given an initial time t^0 and a terminal time t^1, find vector functions $u(t)$ and $x(t)$ that will minimize the functional[1]

$$I(u, x) = \int_{t^0}^{t^1} \phi(t, x(t), u(t))\, dt + \theta(x(t^0), x(t^1)), \tag{1.1}$$

subject to the differential equations

$$\dot{x} = g(t, x, u), \tag{1.2}$$

the constraints

$$h(t, x(t), u(t)) \leqq 0, \tag{1.3}$$

the initial conditions

$$p(x(t^0)) \leqq 0, \tag{1.4}$$

and the terminal conditions

$$q(x(t^1)) \leqq 0. \tag{1.5}$$

Here, x is an n-dimensional state vector, u is an m-dimensional control vector, h is a k-dimensional vector of constraints, p is an l^0-dimensional vector of initial conditions and q is an l^1-dimensional vector of terminal

* Received by the editors May 25, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 16, 1965.

† Shell Development Company, Emeryville, California.

[1] Throughout this work, we shall assume that $u(t)$ is continuous in $t \in [t^0, t^1]$ except for a finite number of jump discontinuities and shall consider only continuous solutions $x(t)$ of (1.2), [3, p. 12]. Equations (1.2) need not be satisfied at the points of discontinuity of $u(t)$. Similarly, other differential equations (2.1), (2.31) and (2.36) need not be satisfied at the points of discontinuity of $u(t)$.

conditions. Various differentiability conditions will be imposed on these functions subsequently.

The main results of this work are Theorems 1 and 2 which give sufficient conditions for optimality. Theorem 1 gives sufficient optimality conditions for the above problem as it stands, while Theorem 2 gives sufficient optimality conditions for the above problem for the "separable" case when

$$\phi(t, x, u) = \phi_1(t, x) + \phi_2(t, u), \tag{1.6}$$

$$g(t, x, u) = g_1(t, x) + g_2(t, u), \tag{1.7}$$

and

$$h(t, x, u) = \begin{bmatrix} h_1(t, x) \\ h_2(t, u) \end{bmatrix} \leqq 0. \tag{1.8a, 1.8b}$$

Essentially, the present sufficient conditions are Pontryagin's conditions [3] plus some convexity conditions on ϕ, θ, g, h, p and q, strict negativity of the adjoint variable associated with ϕ, and negativity of the adjoint variables associated with the differential equations (1.2).

The sufficient conditions given in Theorem 2, subsume the recent results of Lee [2], which in turn subsume the sufficient conditions given by Rozonoér [5, Part I, Theorem 2]. Essentially, Lee considers the separable case where $g_1(t, x)$ is linear in x, and with no state variable constraints. Rozonoér considers the same case but with an objective function that depends only on a linear combination of $x(t^1)$. Rosen [4] has also recently given somewhat different sufficient conditions for optimal control by utilizing an integral representation of $x(t)$ in terms of $u(t)$.

It should be remarked here that the present sufficient conditions were obtained in the same spirit as that of the Kuhn-Tucker sufficient conditions for mathematical programming [1]. In the present sufficient conditions, the Euler conditions play the same role as the gradients in the Kuhn-Tucker conditions. Furthermore, the adjoint variables associated with the differential equations (1.2) turn out to be precisely the negative of the Lagrange multipliers associated with (1.2). In particular we shall have time-dependent multipliers $v(t)$ and $w(t)$ associated with (1.2) and (1.3) and fixed multipliers r and s associated with (1.4) and (1.5).

Vector notation will generally be used. Vectors will be denoted by single letters. In general, subscripts will be used to denote components or groups of components, superscripts will be used to distinguish vectors. A vector will be either a column or a row vector, as it will be clear from the context how the vector is to be considered. Thus we shall write the inner product of two vectors x and y simply as xy. The partial differential operator

$$\left[\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \cdots, \frac{\partial}{\partial x_n}\right]$$

will be denoted by ∇_x, and similarly for ∇_u. The dimensionality of some vectors will not be stated explicitly, it being clear from the context. Also, when we say that a vector function is convex we mean that every component is convex, and similarly for other properties.

2. The sufficient conditions. We shall start by giving sufficient conditions for the nonseparable case (1.1) through (1.5). These conditions were arrived at by introducing Lagrange multipliers $v(t)$, $w(t)$, r, and s for the relations (1.2) to (1.5) respectively, appending these relations and multipliers to the functional (1.1), and then imposing Kuhn-Tucker type conditions [1] to obtain the following sufficient conditions for optimality.

THEOREM 1. *Let $\phi(t, x, u)$ and each component of $g(t, x, u)$ and $h(t, x, u)$ be differentiable and convex in the variables (x, u) for $t \in [t^0, t^1]$, let each component of $p(x(t^0))$ and $q(x(t^1))$ be differentiable and convex in $x(t^0)$ and $x(t^1)$, respectively, and let $\theta(x(t^0), x(t^1))$ be differentiable and convex in $(x(t^0), x(t^1))$. If there exist vectors $\bar{u}(t)$, $\bar{x}(t)$, $\bar{v}(t)$, $\bar{w}(t)$, $\bar{r}$, and $\bar{s}$ satisfying the relations (1.2) through (1.5),[2] with $\bar{x}(t)$, $\bar{v}(t)$ continuous and $\bar{w}(t)$ integrable and such that:*

$$\nabla_x\phi(t, \bar{x}, \bar{u}) + \nabla_x\bar{v}g(t, \bar{x}, \bar{u}) + \nabla_x\bar{w}h(t, \bar{x}, \bar{u}) + \dot{\bar{v}}(t) = 0, \tag{2.1}$$

$$\nabla_u\phi(t, \bar{x}, \bar{u}) + \nabla_u\bar{v}g(t, \bar{x}, \bar{u}) + \nabla_u\bar{w}h(t, \bar{x}, \bar{u}) = 0, \tag{2.2}$$

$$\nabla_{x(t^0)}\theta(\bar{x}(t^0), \bar{x}(t^1)) + \nabla_{x(t^0)}\bar{r}p(\bar{x}(t^0)) + \bar{v}(t^0) = 0, \tag{2.3}$$

$$\nabla_{x(t^1)}\theta(\bar{x}(t^0), \bar{x}(t^1)) + \nabla_{x(t^1)}\bar{s}q(\bar{x}(t^1)) - \bar{v}(t^1) = 0, \tag{2.4}$$

$$\bar{r} \geqq 0, \tag{2.5}$$

$$\bar{r}p(\bar{x}(t^0)) = 0, \tag{2.6}$$

$$\bar{s} \geqq 0, \tag{2.7}$$

$$\bar{s}q(\bar{x}(t^1)) = 0, \tag{2.8}$$

$$\bar{w}(t) \geqq 0, \tag{2.9}$$

$$\bar{w}(t)h(t, \bar{x}(t), \bar{u}(t)) = 0, \tag{2.10}$$

$$\bar{v}(t) \geqq 0, \tag{2.11}[3]$$

then $\bar{u}(t)$, $\bar{x}(t)$ will minimize the functional (1.1) subject to the conditions (1.2) through (1.5). Condition (2.11) need hold only for those components of $g(t, x, u)$ that are nonlinear in x or u or both.

[2] It is understood here and elsewhere that all relations involving t must be satisfied for all t in $[t^0, t^1]$, and that the convexity and differentiability assumptions hold over the entire space over which the functions are defined. However, (2.1) need not be satisfied at discontinuities of $u(t)$.

[3] Condition (2.11) and the convexity of g may be replaced by the weaker requirement that $\bar{v}g$ be convex in (x, u). I am indebted to J. B. Rosen for this observation.

Proof. For simplicity we shall denote $\phi(t, \bar{x}, \bar{u})$ by $\bar{\phi}$ and $\phi(t, x, u)$ by ϕ, and similarly for θ, g, h, p and q. Let $\bar{u}(t)$, $\bar{x}(t)$, $\bar{v}(t)$, $\bar{w}(t)$, $\bar{r}$ and $\bar{s}$ satisfy $(\overline{1.2})$ to $(\overline{1.5})$[4] and (2.1) to (2.11). Let $u(t)$ and $x(t)$ satisfy (1.2) to (1.5). We shall prove that

$$I(u, x) \geqq I(\bar{u}, \bar{x}).$$

We shall now write a string of equalities and inequalities that will prove this result. Explanation of the less obvious equalities and inequalities is found directly below the string.

$$I(u, x) - I(\bar{u}, \bar{x}) = \int_{t^0}^{t^1} (\phi - \bar{\phi})\, dt + \theta - \bar{\theta}$$

$$\text{(a)} \quad \geqq \int_{t^0}^{t^1} [(x - \bar{x})\nabla_x \bar{\phi} + (u - \bar{u})\nabla_u \bar{\phi}]\, dt + (x(t^0) - \bar{x}(t^0))\nabla_{x(t^0)} \bar{\theta} + (x(t^1) - \bar{x}(t^1))\nabla_{x(t^1)} \bar{\theta}$$

$$\text{(b)} \quad = \int_{t^0}^{t^1} [-(x - \bar{x})(\nabla_x \bar{v}\bar{g} + \nabla_x \bar{w}\bar{h} + \dot{\bar{v}}) - (u - \bar{u})(\nabla_u \bar{v}\bar{g} + \nabla_u \bar{w}\bar{h})]\, dt - (x(t^0) - \bar{x}(t^0)) \cdot (\nabla_{x(t^0)} \bar{r}\bar{p} + \bar{v}(t^0)) - (x(t^1) - \bar{x}(t^1))(\nabla_{x(t^1)} \bar{s}\bar{q} - \bar{v}(t^1))$$

$$\text{(c)} \quad = \int_{t^0}^{t^1} [-(x - \bar{x})(\nabla_x \bar{v}\bar{g} + \nabla_x \bar{w}\bar{h}) + (g - \bar{g})\bar{v} - (u - \bar{u}) \cdot (\nabla_u \bar{v}\bar{g} + \nabla_u \bar{w}\bar{h})]\, dt - (x(t^0) - \bar{x}(t^0))\nabla_{x(t^0)} \bar{r}\bar{p} - (x(t^1) - \bar{x}(t^1))\nabla_{x(t^1)} \bar{s}\bar{q}$$

$$\text{(d)} \quad \geqq \int_{t^0}^{t^1} [\bar{v}\bar{g} - \bar{v}g + \bar{w}\bar{h} - \bar{w}h + (g - \bar{g})\bar{v}]\, dt - (x(t^0) - \bar{x}(t^0))\nabla_{x(t^0)} \bar{r}\bar{p} - (x(t^1) - \bar{x}(t^1))\nabla_{x(t^1)} \bar{s}\bar{q}$$

$$\text{(e)} \quad \geqq -(x(t^0) - \bar{x}(t^0))\nabla_{x(t^0)} \bar{r}\bar{p} - (x(t^1) - \bar{x}(t^1))\nabla_{x(t^1)} \bar{s}\bar{q}$$

$$\text{(f)} \quad \geqq \bar{r}\bar{p} - \bar{r}p + \bar{s}\bar{q} - \bar{s}q$$

$$\text{(g)} \quad \geqq 0.$$

The above relations hold:

(a) by the differentiability and convexity of ϕ and θ;

(b) by (2.1), (2.2), (2.3) and (2.4);

[4] For subsequent reference in the proof, it is clearer to refer to relations (1.2) to (1.5) by "$\overline{(1.2)}$ to $\overline{(1.5)}$" when they are satisfied by the barred quantities.

(c) by integration by parts, (1.2), ($\overline{1.2}$) and continuity of $x(t)$, $\bar{x}(t)$ and $\bar{v}(t)$;

(d) by the differentiability and convexity of g and h, and (2.9) and (2.11), (note that this is the only step in the proof where (2.11) is used—note also that if a component of $g(t, x, u)$ is linear in (x, u), then (2.11) is *not* needed for that component of $g(t, x, u)$ in order that this step go through, i.e., in order that the last sentence hold in Theorem 1);

(e) by (2.10), (2.9) and (1.3);

(f) by the convexity and differentiability of p and q, and by (2.5) and (2.7);

(g) by (2.6), (2.8), (2.5), (1.4), (2.7) and (1.5).

It should be remarked that the initial and terminal conditions (1.4) and (1.5) are of sufficiently general form to include all types of linear[5] equalities. We simply write an equality as two inequalities. Thus the inequalities.

$$x(t^0) - x^0 \leqq 0,$$

$$-x(t^0) + x^0 \leqq 0,$$

imply the equality

$$x(t^0) = x^0.$$

With initial conditions of this type, and terminal conditions of the type

$$x(t^1) = x^1,$$

considerable simplifications can be achieved in the sufficient conditions of Theorem 1. We state these results as the following.

COROLLARY 1.

(a) *If the initial conditions* (1.4) *are replaced by*

$$x(t^0) = x^0, \tag{1.4'}$$

then Theorem 1 *holds with conditions* (2.3), (2.5), (2.6), *and the vector* $\bar{r}$, *all deleted*;

(b) *if the terminal conditions* (1.5) *are replaced by*

$$x(t^1) = x^1, \tag{1.5'}$$

then Theorem 1 *holds with conditions* (2.4), (2.7), (2.8), *and the vector* $\bar{s}$, *all deleted*;

(c) *if the initial and terminal conditions* (1.4) *and* (1.5) *are replaced by* (1.4′) *and* (1.5′) *respectively, then Theorem* 1 *holds with conditions* (2.3), (2.4), (2.5), (2.6), (2.7), (2.8), *and the vectors* $\bar{r}$ *and* $\bar{s}$, *all deleted.*

[5] The convexity requirement on p and q restricts us to only *linear* equalities.

We shall only indicate how part (a) of Corollary 1 follows from Theorem 1. Let

$$p(x(t^0)) = \begin{bmatrix} x(t^0) - x^0 \\ -x(t^0) + x^0 \end{bmatrix}.$$

Then (2.3) becomes

$$\nabla_{x(t^0)}\bar{\theta} + \bar{r}_1 - \bar{r}_2 + \bar{v}(t^0) = 0,$$

or

$$\nabla_{x(t^0)}\bar{\theta} + \bar{v}(t^0) = \bar{r}_2 - \bar{r}_1 . \tag{2.5'}$$

Since any number can be expressed as the difference of two nonnegative numbers, (2.3) and (2.5) can be automatically satisfied by picking nonnegative $\bar{r}_2$ and $\bar{r}_1$ such that (2.5') is satisfied. Relation (2.6) is automatically satisfied because $p(\bar{x}(t^0)) = 0$.

It is also possible to show that the conditions of Corollary 1(c) imply Pontryagin's maximum principle for the fixed-time case. Pontryagin [3, pp. 298–299][6] considers the problem of minimizing

$$I(u, x) = \int_{t^0}^{t^1} \phi(x, u)\, dt, \tag{2.12}$$

subject to:

$$\dot{x} = g(x, u), \tag{2.13}$$

$$\left.\begin{aligned} h_1(x, u) &\leqq 0 \\ h_2(u) &\leqq 0 \end{aligned}\right\} h(x, u) \leqq 0, \tag{2.14, 2.15}$$

$$x(t^0) = x^0, \tag{2.16}$$

and

$$x(t^1) = x^1, \tag{2.17}$$

where h_1 and h_2 are vector constraints. Pontryagin's maximum principle for this problem asserts that if $\bar{u}(t)$ and $\bar{x}(t)$ solve the above problem, then there exist a scalar $\bar{\psi}_0(t)$ and a vector $\bar{\psi}(t)$, not both zero, and vectors $\bar{\lambda}(t)$ and $\bar{\nu}(t)$ such that $\bar{u}(t)$, $\bar{x}(t)$, $\bar{\psi}_0(t)$, $\bar{\psi}(t)$, $\bar{\lambda}(t)$ and $\bar{\nu}(t)$ satisfy (2.13) to (2.17), and;

$$\nabla_x\bar{\psi}_0\phi(\bar{x}, \bar{u}) + \nabla_x\bar{\psi}g(\bar{x}, \bar{u}) - \nabla_x\bar{\lambda}h_1(\bar{x}, \bar{u}) + \dot{\bar{\psi}} = 0, \tag{2.18}$$

$$\nabla_u\bar{\psi}_0\phi(\bar{x}, \bar{u}) + \nabla_u\bar{\psi}g(\bar{x}, \bar{u}) - \nabla_u\bar{\lambda}h_1(\bar{x}, \bar{u}) - \nabla_u\bar{\nu}h_2(\bar{u}) = 0, \tag{2.19}$$

[6] We have changed Pontryagin's problem to a fixed-time problem and have omitted the variable-time condition that the maximum of the Hamiltonian be zero.

$$\psi_0(t) = \text{const.} \leqq 0, \tag{2.20}$$

$$\bar{\psi}_0\phi(\bar{x}, \bar{u}) + \bar{\psi}g(\bar{x}, \bar{u}) \geqq \bar{\psi}_0\phi(\bar{x}, u) + \bar{\psi}g(\bar{x}, u) \tag{2.21}$$

for all u satisfying $h_1(\bar{x}, u) \leqq 0$ and $h_2(u) \leqq 0$. Condition (2.21) is the "maximum" condition.

From Corollary 1(c), the sufficient conditions at our disposal are that $\bar{u}(t)$, $\bar{x}(t)$ and some $\bar{v}(t)$ and $\bar{w}(t)$ satisfy (2.13) to (2.17) and (2.1), (2.2), (2.9), (2.10) and (2.11).[7] It is easy to see that if we set

$$\bar{\psi}^0 = -1, \tag{2.22}$$

$$\bar{\psi} = -\bar{v}, \tag{2.23}$$

$$\begin{bmatrix} \bar{\lambda} \\ \bar{\nu} \end{bmatrix} = \bar{w}, \tag{2.24}$$

then (2.1) implies (2.18) and (2.2) implies (2.19). It is obvious that (2.20) is implied by $\psi^0 = -1$. We shall now show that (2.9), (2.10), (2.11) and (2.2) imply the maximum condition (2.21). We have

$$\bar{\psi}_0\phi(\bar{x}, u) + \bar{\psi}g(\bar{x}, u)$$

$$= -\phi(\bar{x}, u) - \bar{v}g(\bar{x}, u) \tag{2.24}$$

$$\leqq -\phi(\bar{x}, \bar{u}) - \bar{v}g(\bar{x}, \bar{u}) - (u - \bar{u})(\nabla_u\phi(\bar{x}, \bar{u}) + \nabla_u\bar{v}g(\bar{x}, \bar{u})) \tag{2.25}$$

$$= -\phi(\bar{x}, \bar{u}) - \bar{v}g(\bar{x}, \bar{u}) + (u - \bar{u})\nabla_u\bar{w}h(\bar{x}, \bar{u}) \tag{2.26}$$

$$\leqq -\phi(\bar{x}, \bar{u}) - \bar{v}g(\bar{x}, \bar{u}) + \bar{w}h(\bar{x}, u) - \bar{w}h(\bar{x}, \bar{u}) \tag{2.27}$$

$$\leqq -\phi(\bar{x}, \bar{u}) - \bar{v}g(\bar{x}, \bar{u}) \qquad (\text{for } h(\bar{x}, u) \leqq 0) \tag{2.28}$$

$$= \bar{\psi}_0\phi(\bar{x}, \bar{u}) + \bar{\psi}g(\bar{x}, \bar{u}) \qquad (\text{for } h(\bar{x}, u) \leqq 0), \tag{2.29}$$

where (2.24) follows from (2.22) and (2.23), (2.25) from the convexity of $\phi(x, u)$ and $g(x, u)$ in u and (2.11),[8] (2.26) from (2.2), (2.27) from the convexity of $h(x, u)$ in u and (2.9), (2.28) from (2.9) and (2.10), and finally (2.29) from (2.22) and (2.23). Hence the maximum condition (2.21) is established.

We consider now the "separable" case where the state and control vectors enter into separate functions. For this case we shall obtain somewhat stronger sufficient conditions than those in Theorem 1. Furthermore, these conditions will, essentially, be identical to Pontryagin's conditions, with additional requirements of convexity, negativity and strict negativity. We now state the separable problem.

[7] Condition (2.11) is needed only if $g(x, u)$ is nonlinear.

[8] Note if $g(x, u)$ were linear, (2.11) would not be needed.

Given an initial time t^0 and a terminal time t^1, find vector functions $u(t)$ and $x(t)$ that will minimize the functional

$$(2.30)\quad I(u, x) = \int_{t^0}^{t^1} (\phi_1(t, x(t)) + \phi_2(t, u(t)))\, dt + \theta(x(t^0), x(t^1)),$$

subject to the differential equations

$$(2.31)\quad \dot{x} = g_1(t, x) + g_2(t, u),$$

the state-vector constraints

$$(2.32)\quad h(t, x) \leqq 0,$$

the control-vector constraints[9]

$$(2.33)\quad u(t) \in \Omega(t) \subset E^m,$$

the initial conditions

$$(2.34)\quad p(x(t^0)) \leqq 0,$$

and the terminal conditions

$$(2.35)\quad q(x(t^1)) \leqq 0.$$

THEOREM 2. *Let $\phi_1(t, x)$ and each component of $g_1(t, x)$ and $h(t, x)$ be differentiable and convex in x for $t \in [t^0, t^1]$, let each component of $p(x(t^0))$ and $q(x(t^1))$ be differentiable and convex in $x(t^0)$ and $x(t^1)$, respectively, and let $\theta(x(t^0), x(t^1))$ be differentiable and convex in $(x(t^0), \bar{x}(t^1))$. If there exist vectors $\bar{u}(t)$, $\bar{x}(t)$, $\bar{v}(t)$, $\bar{w}(t)$, $\bar{r}$ and $\bar{s}$ with $\bar{x}(t)$, $\bar{v}(t)$ continuous and $\bar{w}(t)$ integrable, satisfying* (2.31) *to* (2.35) *and*

$$(2.36)\quad \nabla_x\phi_1(t, \bar{x}) + \nabla_x\bar{v}g_1(t, \bar{x}) + \nabla_x\bar{w}h(t, \bar{x}) + \dot{\bar{v}}(t) = 0,$$

$$(2.37)^{10}\quad \phi_2(t, u) + \bar{v}g_2(t, u) \geqq \phi_2(t, \bar{u}) + \bar{v}g_2(t, \bar{u}) \text{ for all } u(t) \in \Omega(t),$$

$$(2.38)\quad \nabla_{x(t^0)}\theta(\bar{x}(t^0), \bar{x}(t^1)) + \nabla_{x(t^0)}\bar{r}p(\bar{x}(t^0)) + \bar{v}(t^0) = 0,$$

$$(2.39)\quad \nabla_{x(t^1)}\theta(\bar{x}(t^0), \bar{x}(t^1)) + \nabla_{x(t^1)}\bar{s}q(\bar{x}(t^1)) - \bar{v}(t^1) = 0,$$

$$(2.40)\quad \bar{r} \geqq 0,$$

$$(2.41)\quad \bar{r}p(\bar{x}(t^0)) = 0,$$

$$(2.42)\quad \bar{s} \geqq 0,$$

$$(2.43)\quad \bar{s}q(\bar{x}(t^1)) = 0,$$

[9] $\Omega(t)$ is an arbitrary time-dependent set in the m-dimensional Euclidean space E^m.

[10] This is the "maximum" condition if we set $\bar{\psi} = -\bar{v}$ and $\bar{\psi}_0 = -1$.

$$\bar{w}(t) \geqq 0, \tag{2.44}$$

$$\bar{w}(t)h(t, \bar{x}) = 0, \tag{2.45}$$

$$\bar{v}(t) \geqq 0, \tag{2.46}$$[11]

then $\bar{u}(t)$, $\bar{x}(t)$ *will minimize the functional* (2.30) *subject to* (2.31) *to* (2.35). *Condition* (2.46) *need hold only for those components of* $g_1(t, x)$ *that are nonlinear in* x.

Proof. The proof is similar to that of Theorem 1. Let $\bar{u}(t)$, $\bar{x}(t)$, $\bar{v}(t)$, $\bar{w}(t)$, $\bar{r}$ and $\bar{s}$ satisfy (2.31) to (2.35) and (2.36) to (2.46). Let $u(t)$ and $x(t)$ satisfy (2.31) to (2.35). We shall prove that

$$I(u, x) \geqq I(\bar{u}, \bar{x}).$$

We have

$$I(u, x) - I(\bar{u}, \bar{x}) = \int_{t^0}^{t^1} (\phi_1 - \bar{\phi}_1 + \phi_2 - \bar{\phi}_2)\, dt + \theta - \bar{\theta}$$

$$\text{(a)} \quad \geqq \int_{t^0}^{t^1} ((x - \bar{x})\nabla_x \bar{\phi}_1 + \phi_2 - \bar{\phi}_2)\, dt + (x(t^0) - \bar{x}(t^0))\nabla_{x(t^0)} \bar{\theta} + (x(t_1) - \bar{x}(t^1))\nabla_{x(t^1)} \bar{\theta}$$

$$\text{(b)} \quad \geqq \int_{t^0}^{t^1} (-(x - \bar{x})(\nabla_x \bar{v}\bar{g}_1 + \nabla_x \bar{w}\bar{h} + \dot{\bar{v}}) + \bar{v}\bar{g}_2 - \bar{v}g_2)\, dt - (x(t^0) - \bar{x}(t^0))(\nabla_{x(t^0)} \bar{r}\bar{p} + \bar{v}(t^0)) - (x(t^1) - \bar{x}(t^1))(\nabla_{x(t^1)} \bar{s}\bar{q} - \bar{v}(t^1))$$

$$\text{(c)} \quad = \int_{t^0}^{t^1} (-(x - \bar{x})(\nabla_x \bar{v}\bar{g}_1 + \nabla_x \bar{w}\bar{h}) + \bar{v}g_1 + \bar{v}g_2 - \bar{v}\bar{g}_1 - \bar{v}\bar{g}_2 + \bar{v}\bar{g}_2 - \bar{v}g_2)\, dt - (x(t^0) - \bar{x}(t^0))\nabla_{x(t^0)} \bar{r}\bar{p} - (x(t^1) - \bar{x}(t^1))\nabla_{x(t^1)} \bar{s}\bar{q}$$

$$\text{(d)} \quad \geqq \int_{t^0}^{t^1} (\bar{w}\bar{h} - \bar{w}h)\, dt + \bar{r}\bar{p} - \bar{r}p + \bar{s}\bar{q} - \bar{s}q$$

$$\text{(e)} \quad \geqq 0$$

The above relations hold:

(a) by the differentiability and convexity of ϕ_1 and θ;

(b) by (2.36), (2.37), (2.33), (2.38) and (2.39);

(c) by integration by parts, (2.31), (2.31) and continuity of $x(t)$, $\bar{x}(t)$ and $\bar{v}(t)$;

(d) by the differentiability and convexity of g_1, h, p and q, and by (2.40),

[11] Condition (2.46) and the convexity of g_1 may be replaced by the weaker requirement that $\bar{v}g_1$ be convex in x.

(2.42), (2.44) and (2.46) (note that this is the only step in the proof where (2.46) is used—note also that if a component of $g_1(t, x)$ is linear in x, then (2.46) is *not* needed for that component of $g_1(t, x)$ in order that this step go through, i.e., in order that the last sentence hold in Theorem 2);

(e) by (2.45), (2.44), (2.32), (2.41), (2.40), (2.34), (2.43), (2.42) and (2.35).

Again here considerable simplification in Theorem 2 can be achieved if the initial and terminal conditions (2.34) and (2.35) are replaced by

$$x(t^0) = x^0 \tag{2.34'}$$

and

$$x(t^1) = x^1. \tag{2.35'}$$

We have then the following.

COROLLARY 1.

(a) *If the initial conditions* (2.34) *are replaced by* (2.34′) *then Theorem* 2 *holds with conditions* (2.38), (2.40), (2.41) *and the vector* $\bar{r}$, *all deleted*;

(b) *if the terminal conditions* (2.35) *are replaced by* (2.35′) *then Theorem* 2 *holds with conditions* (2.39), (2.42), (2.43) *and the vector* $\bar{s}$, *all deleted*;

(c) *if the initial and terminal conditions* (2.34) *and* (2.35) *are replaced by* (2.34′) *and* (2.35′) *then Theorem* 2 *holds with conditions* (2.38) *through* (2.43), *and the vectors* $\bar{r}$ *and* $\bar{s}$, *all deleted.*

Corollary 1(a) above subsumes the sufficient conditions given by Lee [2]. Lee considers the case where $g_1(t, x) = A(t)x$ and with no state variable constraints.

It is quite straightforward to obtain a sufficient version of Pontryagin's maximum principle for fixed-time [3, Theorem 6, p. 67] from Corollary 1(c). For this purpose we consider a simpler version of the separable case as follows. Find $u(t)$ and $x(t)$ that will minimize

$$I(u, x) = \int_{t^0}^{t^1} (\phi_1(t, x) + \phi_2(t, u))\, dt, \tag{2.47}$$

subject to:

$$\dot{x} = g_1(t, x) + g_2(t, u), \tag{2.48}$$

$$u \in \Omega \subset E^m, \tag{2.49}$$

$$x(t^0) = x^0, \tag{2.50}$$

$$x(t^1) = x^1. \tag{2.51}$$

For this case, the following corollary follows from Corollary 1(c) to Theorem 2.

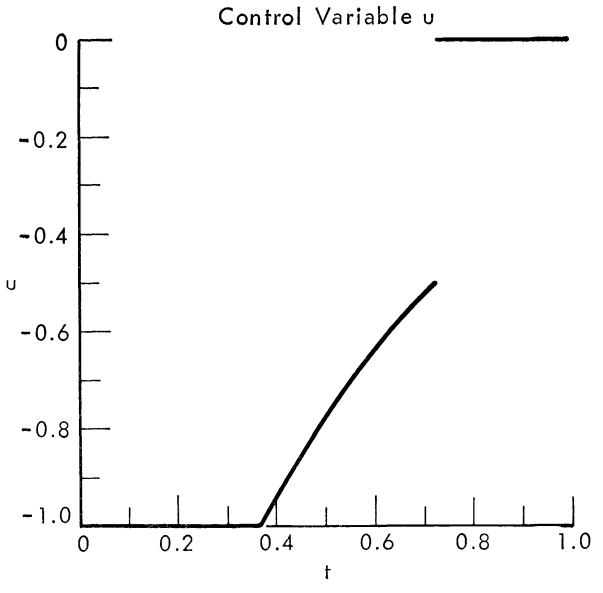

FIG. 1

COROLLARY 2. *Let* $\phi_1(t, x)$ *and each component of* $g_1(t, x)$ *be differentiable and convex in* x *for* $t \in [t^0, t^1]$. *If there exist vectors* $\bar{u}(t)$, $\bar{x}(t)$, $\bar{\psi}(t)$ *satisfying the relations* (2.48) *to* (2.51), *with* $\bar{x}(t)$ *and* $\bar{v}(t)$ *continuous and such that*

$$-\nabla_x \phi_1(t, \bar{x}) + \nabla_x \bar{\psi} g_1(t, \bar{x}) + \dot{\bar{\psi}}(t) = 0, \tag{2.52}$$

$$-\phi_2(t, \bar{u}) + \bar{\psi} g_2(t, \bar{u}) \geqq -\phi_2(t, u) + \bar{\psi} g_2(t, u) \text{ for all } u \in \Omega, \tag{2.53}$$

$$\bar{\psi}(t) \leqq 0, \tag{2.54}$$

then $\bar{u}(t)$, $\bar{x}(t)$ *will minimize the functional* (2.47) *subject to the conditions* (2.48) *to* (2.51). *Condition* (2.54) *is needed only if* $g_1(t, x)$ *is nonlinear in* x.[12]

This corollary follows directly from Corollary 1(c) by setting $\bar{\psi} = -\bar{v}$ and suppressing the constraint $h(t, x) \leqq 0$ and the vector $\bar{w}$. The conditions of Corollary 2 are identical with Pontryagin's conditions [3, Theorem 6, p. 67] plus the additional requirements that $\psi_0 = -1$, that (2.54) holds for nonlinear $g_1(t, x)$, and that $\phi_1(t, x)$ and $g_1(t, x)$ are convex.

3. Numerical example. Consider the following numerical example. Given initial time $t^0 = 0$ and terminal time $t^1 = 1$, find scalar functions $u(t)$ and $x(t)$ that will minimize the functional

[12] Condition (2.54) and the convexity of g_1 may be replaced by the weaker requirement that $-\bar{\psi} g_1$ be convex in x.

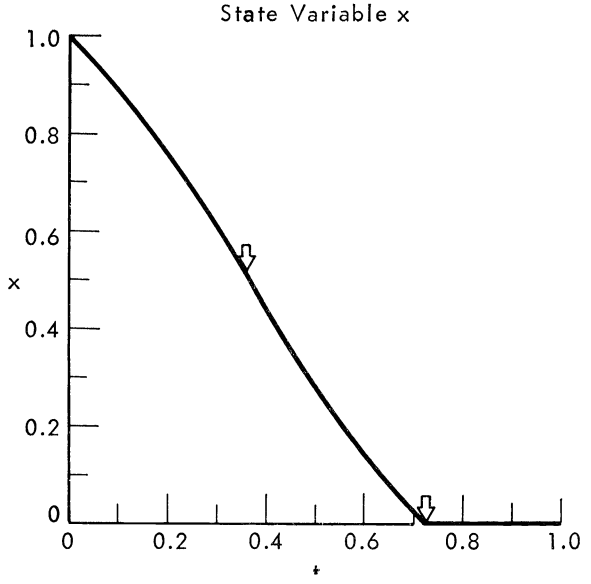

FIG. 2

$$\int_0^1 x(t)\, dt,$$

subject to the differential equation

$$\dot{x} = x^2 + 2u,$$

the constraints

$$-x - u - 0.5 \leqq 0,$$

$$-x \leqq 0,$$

$$-u - 1 \leqq 0,$$

and the initial condition

$$x(0) = 1.$$

It is obvious that the above problem satisfies the conditions of Theorem 1. The optimality conditions of that theorem require that x and u satisfy the above differential equation, constraints, initial condition and:

$$1 + 2vx - w_1 - w_2 + \dot{v} = 0;$$

$$2v - w_1 - w_3 = 0;$$

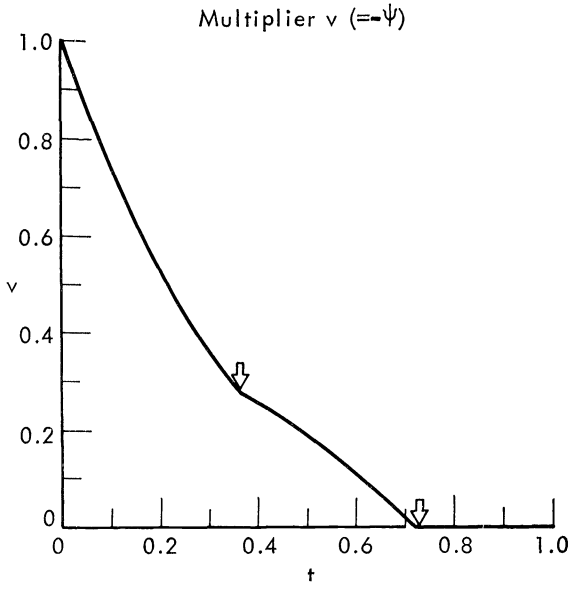

FIG. 3

$$v(1) = 0;$$

$$w_1 \geqq 0,\ w_2 \geqq 0,\ w_3 \geqq 0;$$

$$w_1(x + u + 0.5) = 0, \qquad w_2 x = 0, \qquad w_3(u + 1) = 0;^{13}$$

$$v \geqq 0;$$

$$x(t+) = x(t-);$$

$$v(t+) = v(t-).$$

The following solution satisfies the above sufficient conditions and hence is optimal:

(i) for $0 \leqq t \leqq 0.3619$,

$$u = -1,$$

$$t = 0.6232 - 0.3536 \log \frac{1.414 + x}{1.414 - x},$$

$$v = \frac{x}{2 - x^2}, \quad w_1 = w_2 = 0, \quad w_3 = 2v;$$

[13] These conditions are equivalent to (2.10) in view of (2.9) and (1.3).

(ii) for $0.3619 < t \leqq 0.7239$,

$$u = -x - 0.5,$$

$$x = 1 + 1.414 \tanh 1.414(0.1006 - t),$$

$$v = \frac{x}{1 + 2x - x^2}, \quad w_1 = 2v, \quad w_2 = w_3 = 0;$$

(iii) for $0.7239 < t \leqq 1$,

$$u = 0,$$

$$x = 0,$$

$$v = 0, \qquad w_1 = 0, \qquad w_2 = 1, \qquad w_3 = 0.$$

Figs. 1, 2, and 3 depict, respectively, the control variable $u(t)$, the state variable $x(t)$, and the multiplier $v(t)$, all as functions of time t.

REFERENCES

[1] H. W. Kuhn and A. W. Tucker, *Nonlinear programming*, Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1951, pp. 481–492.

[2] E. B. Lee, *A sufficient condition in the theory of optimal control*, this Journal, 1 (1963), pp. 241–245.

[3] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[4] J. B. Rosen, *Sufficient conditions for optimal control of convex processes*, Tech. Rep. CS7, Stanford University, 1964.

[5] L. I. Rozonoér, *The L. S. Pontryagin maximum principle in the theory of optimal systems*, I, II, III, Automation and Remote Control, 20 (1959), pp. 1288–1302, 1405–1421, 1517–1532.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

QUADRATIC PROGRAMMING IN MECHANICS: DYNAMICS OF ONE-SIDED CONSTRAINTS*

J. J. MOREAU†

1. Let S be a frictionless mechanical system with n degrees of freedom; we denote by $q_1, q_2, \cdots, q_n$ the generalized coordinates, representing the point q of a configuration space. A finite family of one-sided constraints is imposed on the system; the kinematic effect of these constraints is expressed by the conditions (assumed compatible)

$$f_\alpha(q, t) \geqq 0, \qquad \alpha \in I, \text{ finite set of indexes.} \tag{1}$$

For instance, some solid parts of the system may be in contact or become detached but they can never overlap. These constraints are frictionless, i.e., as long as the equalities hold in (1), the motion of the system is governed by Lagrange's equations with multipliers λ_α, $\alpha \in I$. The mechanical meaning of these multipliers is to describe the reaction forces associated with possible contacts and, conventionally, we have

$$\lambda_\alpha \geqq 0, \tag{2}$$

i.e., the force of reaction is directed towards the region defined by (1) and

$$\lambda_\alpha f_\alpha(q, t) = 0, \qquad \text{for all } \alpha \in I, \tag{3}$$

i.e., as soon as a contact ceases, the corresponding reaction becomes zero.

The set of the active forces experienced by the system is described by its covariant components Q^i (continuous functions of q, t) relative to the coordinates (q_i).

The kinetic energy is expressed as

$$T(q, \dot{q}, t) = \tfrac{1}{2}\sum_{i,k} a^{ik}(q, t)\dot{q}_i\dot{q}_k + \sum_i b^i(q, t)\dot{q}_i + c(q, t). \tag{4}$$

We shall always assume that the considered configuration is regular with respect to the coordinates (q_i) so that the quadratic part of this expression is positive definite.

It is usual to study such a mechanical system by starting with the tentative hypothesis that all the contacts $f_\alpha = 0$ are present at any instant. Then, by putting $\partial f_\alpha/\partial q_i = u_\alpha{}^i$, the n differential equations of Lagrange

* Received by the editors June 21, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Mathématiques, Faculté des Sciences, Université de Montpellier, Montpellier, France.

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right) - \frac{\partial T}{\partial q_i} = Q^i + \sum_{\alpha \in I} \lambda_\alpha \, u_\alpha{}^i, \tag{5}$$

together with the vanishing of f_α (for every $\alpha \in I$), determine the functions $q_i(t)$ and $\lambda_\alpha(t)$.

As long as the values λ_α so calculated are all nonnegative, the initial hypothesis of permanent contacts is accepted. When, on the contrary, some of the λ_α become negative, the hypothesis is rejected: some of the contacts must cease. But, as Delassus [1] pointed out, the contacts f_α which cease are not necessarily those for which the above computation gives a negative λ_α (simple counterexamples may be formulated). Delassus' arguments towards a correct solution were rather intricate; actually the author has proved [4] that the determination of the acceleration (i.e., the second derivatives $\ddot{q}_i$) is governed by a *generalization of Gauss' variational principle*; this leads to a typical quadratic programming procedure. An extremal principle also holds which characterizes the values of the one-sided reactions (i.e., the λ_α), independently of the accelerations: this leads to a quadratic programming problem *dual* to the preceding one.

2. Our problem may be expressed in the following manner.

For $t = t_0$, the configuration q (i.e., the values of the $q_i(t_0)$) and the velocity $\dot{q}$ (i.e., the values of the derivatives $\dot{q}_i(t_0)$) are given. These data are assumed compatible with the contacts $f_\alpha = 0$ for $\alpha \in K \subset I$; that means that

$$\left(\frac{df_\alpha}{dt}\right)_{t=t_0} = \sum_i u_\alpha{}^i \, \dot{q}_i + \frac{\partial f_\alpha}{\partial t} = 0, \qquad \text{for all } \alpha \in K, \tag{6}$$

while $f_\alpha > 0$ for $\alpha \notin K$. The question is to find the *state of acceleration* after t_0, i. e., the right-limits $\ddot{q}_i(t_0 + 0)$.

By continuity, for $\alpha \notin K$, we have $f_\alpha > 0$ during an interval $(t_0, t_0 + \epsilon)$ so that the corresponding contact does not intervene. For $\alpha \in K$, on the contrary, the conditions (1), together with (6), yield

$$\left(\frac{d^2 f_\alpha}{dt^2}\right)_{t=t_0} = \sum_i u_\alpha{}^i \ddot{q}_i - s_\alpha \geqq 0, \tag{7}$$

where s_α is a known quantity. Using the energy expression (4), Lagrange's equations, analogous to (5), may be written

$$\sum_k a^{ik} \ddot{q}_k = z^i + \sum_{\alpha \in K} \lambda_\alpha u_\alpha{}^i, \tag{8}$$

where z^i denotes known quantities. The λ_α are nonnegative by virtue of (2), and (3) yields

$$\lambda_\alpha \left[\sum_i u_\alpha{}^i \ddot{q}_i - s_\alpha\right] = 0. \tag{9}$$

We can prove that *the conditions* (2), (7), (8), (9) *define one and only one set of values for the unknowns* $\ddot{q}_i$, $i = 1, 2, \cdots, n$, *and* λ_α, $\alpha \in K$; this solution possesses the following variational characterization: in the $\mathbf{R}^n$-space of $\ddot{q}$, the inequalities (7) define a closed convex polyhedral region $\mathcal{C}$ (non-empty, since the set of inequalities (1) is assumed to permit a motion). One proves that *the above solution corresponds to the unique point* $\ddot{q}$ *of* $\mathcal{C}$ *where the function*

$$G = \tfrac{1}{2} \sum_{i,k} a^{ik} \ddot{q}_i \ddot{q}_k - \sum_i z^i \ddot{q}_i \tag{10}$$

attains its minimum. The proof may be derived from Kuhn and Tucker's theory of multipliers in nonlinear programming. A direct derivation may also be found in [4].

On the other hand, *Gauss' principle* (of "least deviation") may be formulated, for the classical case of two-sided differentiable constraints, in the following way: given the configuration and the velocity state of such a classical system $\mathcal{S}$ at an instant t_0, the resulting acceleration state is, among all the acceleration states compatible with these data and with the constraints, that one which confers its minimum to the "Appell function"

$$\mathfrak{A} = \tfrac{1}{2} \int_{\mathcal{S}} \boldsymbol{\Gamma}^2 \, dm - \int_{\mathcal{S}} \boldsymbol{\Gamma} \cdot \mathbf{dF},$$

where $\boldsymbol{\Gamma}$ denotes the acceleration of the generic element of $\mathcal{S}$, dm is the mass measure defined on $\mathcal{S}$, while the vectorial measure $\mathbf{dF}$ represents the active forces experienced by $\mathcal{S}$. Since it happens that, for an arbitrary motion defined by some $q_i(t)$, the function $\mathfrak{A}$ has exactly the expression G written in (10) (disregarding an additive constant), the variational characterization given above for the solution of our problem means that *Gauss' principle is still valid for systems with one-sided frictionless constraints.*

3. In order to deal with *duality*, it is useful to introduce additional geometrical terminology. Let (e_i), $i = 1, 2, \cdots, n$, represent a base in an n-dimensional linear space E and let (e^i) be the dual base in the dual space E'; we denote by $\langle , \rangle$ the duality bilinear form. The symmetric positive regular matrix a^{ik} represents, relative to these bases, a one-to-one linear mapping A of E' onto E. We provide E with an Euclidean metric by defining, for every pair $x \in E$, $y \in E$, the scalar product

$$(x \mid y) = \langle x, A^{-1}(y) \rangle = \langle y, A^{-1}(x) \rangle.$$

Let us put

$$\ddot{q} = \sum_i \ddot{q}_i e^i \in E',$$

$$u_\alpha = \sum_i u_\alpha{}^i e_i \in E,$$

$$z = \sum_i z^i e_i \in E,$$

so that the system of Lagrange's equations (8) is written as an equation in E,

$$A(\ddot{q}) = z + \sum_{\alpha \in K} \lambda_\alpha u_\alpha . \tag{11}$$

Instead of $\ddot{q} \in E'$, we now introduce the new unknown $x = A(\ddot{q}) \in E$, so that (11) becomes

$$x - \sum_{\alpha \in K} \lambda_\alpha u_\alpha = z. \tag{12}$$

The inequalities (7) are rewritten as

$$(u_\alpha \mid x) - s_\alpha \geqq 0, \qquad \text{for all} \quad \alpha \in K, \tag{13}$$

defining thereby a closed convex polyhedral region C in E.

Then the variational characterization stated above is formulated with regard to the Euclidean metric of the space E: *the solution x is, in C, the nearest point from the known point z.*

4. We are now prepared to invoke the author's *duality-decomposition theorem* on quadratic programming (cf. [3], [8]). This theorem was derived for the more general case of infinite-dimensional Hilbert spaces in connection with problems of unilaterality in the mechanics of continua. In contrast with other duality treatments, the elements of a pair of dual problems belong to the same self-dual (Hilbert) space, so that they may be added together.

Let us first recall Fenchel's [2] concept of *conjugate convex functions* (slightly modified by the author, in order to accept $+\infty$ as a value for such functions): we denote by $\Gamma_0(E)$ the totality of the functions everywhere defined in E, taking their values in $(-\infty, +\infty]$, which are convex, lower semicontinuous, and other than the constant $+\infty$. For instance, given a nonempty subset P of E, the *indicatrix function*

$$\psi_P(x) = \begin{cases} 0 & \text{if} \quad x \in P, \\ +\infty & \text{if} \quad x \notin P, \end{cases}$$

belongs to $\Gamma_0(E)$ if and only if P is closed and convex. Now one easily proves that a one-to-one involutory mapping of $\Gamma_0(E)$ onto itself is defined by associating to any $f \in \Gamma_0(E)$ its *conjugate* or *dual* function

$$g(y) = \sup_{x \in E} [(x \mid y) - f(x)]. \tag{14}$$

In other words, g is the smallest element in the set of functions for which

$$f(x) + g(y) \geqq (x \mid y) \tag{15}$$

for every x and y in E. The points x and y are called *conjugate*, relative to the pair of dual functions (f, g), if the equality holds in (15).

For any $z \in E$ and $f \in \Gamma_0(E)$ we denote by $\text{prox}_f z$ (proximal point of z with regard to the function f) the point where the function

$$u \to \tfrac{1}{2} \| z - u \|^2 + f(u)$$

attains its minimum (existence and uniqueness of this point are assured): specifically, if f is the indicatrix function of a closed convex set C, $\text{prox}_f z$ is the nearest point from z which lies in C, denoted by $\text{proj}_C z$.

Then our duality-decomposition theorem may be stated in the following form: *If f and g are dual functions, every $z \in E$ equals the sum of $x = \text{prox}_f z$ and $y = \text{prox}_g z$; the points x and y are conjugate relative to (f, g) and they embody the unique decomposition of z into a sum of two such terms.*

A particularly interesting case occurs when f and g are the indicatrices of two *mutually polar closed convex cones* P and Q, i. e.,

$$Q = \{y \in E: (x \mid y) \leqq 0 \quad \text{for every} \quad x \in P\}$$

(and conversely). Here the theorem gives: *Every $z \in E$ equals the sum of $x = \text{proj}_P z$ and $y = \text{proj}_Q z$; the elements x and y are orthogonal and embody the unique decomposition of z into a sum of two orthogonal elements respectively belonging to P and Q.* This result may be regarded as a generalization of the classical decomposition of E into the direct sum of two orthogonal complementary subspaces.

5. Returning to our mechanical problem, let us take as f the indicatrix of the set C defined by (13). The dual function

$$g(y) = \sup_{x \in E} [(x \mid y) - f(x)] = \sup_{x \in C} (x \mid y)$$

is the *support function* of C. Our generalization of Gauss' principle means that the unknown x defined in §3 has the value $x = \text{prox}_f z$. Then, by the duality-decomposition theorem, (12) leads to a *variational characterization of the (abstract) reaction exerted by the system against its set of one-sided constraints, i.e., the term*

$$-\sum_{\alpha \in K} \lambda_\alpha u_\alpha \in E.$$

This term equals the proximal point $\text{prox}_g z$.

Incidentally, we may note that, in the present case, the set C is a (non-homogeneous) cone with vertex at the point x_0 which would be found for x, in the case where the system underwent the *two-sided* constraints $f_\alpha = 0$, $\alpha \in K$. That leads to an alternate characterization of $\text{prox}_g z$: it is the nearest point from $z - x_0$ in the convex polyhedral homogeneous cone C' gener-

ated by the $-u_\alpha$, $\alpha \in K$. Actually, $z - x_0$ is the value found for the reaction in this hypothetical case of two-sided constraints; in that sense, it can be said that the motion in the presence of one-sided constraints takes place in such a way that the one-sided reactions differ the least from the reactions corresponding to the two-sided case.

6. In conclusion, we hope that such a theory may prove useful in studying the dynamical response of mechanical transmissions affected by *looseness*. The author's main concern in mechanics is with the infinite-dimensional cases appearing in the mechanics of continua, e.g., inception of cavitation in a liquid flow (cf. [6], [7]). In this connection, conjugate convex functions in topological linear spaces, more general than Hilbert's, have been intensively studied for three years, together with various related notions such as *subdifferentiability*, *inf-convolution* (see, e.g., [5]).

REFERENCES

[1] E. Delassus, *Sur les liaisons unilatérales*, Ann. Sci. École Norm. Sup., 34 (1917), pp. 95–179.

[2] W. Fenchel, *On conjugate convex functions*, Canad. J. Math., 1 (1949), pp. 73–77.

[3] J. J. Moreau, *Fonctions convexes duales et points proximaux dans un espace hilbertien*, C. R. Acad. Sci. Paris, 255 (1962), pp. 2857–2899.

[4] ———, *Les liaisons unilatérales et le principe de Gauss*, Ibid., 256 (1963), pp. 871–874.

[5] ———, *Théorèmes "inf-sup"*, Ibid., 258 (1964), pp. 2720–2722.

[6] ———, *Sur la naissance de la cavitation dans une conduite*, Ibid., 259 (1964), pp. 3948–3950.

[7] ———, *One-sided constraints in hydrodynamics*, Nonlinear Programming, a Course, J. Abadie, ed., North-Holland, Amsterdam, to appear.

[8] ———, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math., to appear.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

DUALITY IN DYNAMIC OPTIMIZATION*

R. PALLU DE LA BARRIÈRE†

1. Statement of the problem. The problem we shall consider is the following.

(I) *Find the function x minimizing*

$$f(x) = \frac{1}{2}\int_0^1 \int_0^1 A(t, \tau)x(t)x(\tau)\, dt\, d\tau + \int_0^1 b(t)x(t)\, dt,$$

subject to the constraints

$$0 \leqq x(t) \leqq 1.$$

It is assumed that the two functions x and b belong to the space L^2 of square-integrable functions defined on the interval [0, 1]. The function A is assumed to be a continuous positive-definite kernel. The scalar product of two elements y, z of L^2 will be denoted by $\langle y, z\rangle$:

$$\langle y, z\rangle = \int_0^1 y(t)z(t)\, dt.$$

The operator transforming the function x into the function y, defined by

$$y(t) = \int_0^1 A(t, \tau)x(\tau)\, d\tau,$$

will be denoted by $\mathbf{A}$. Consequently, the problem (I) can be written in the condensed form:

Find x minimizing

$$f(x) = \tfrac{1}{2}\langle x, \mathbf{A}x\rangle + \langle b, x\rangle,$$

subject to the constraints $x \in L^2$, $0 \leqq x \leqq 1$.

Remark. From the positive-definiteness of A, it follows that f is a convex function. On the other hand, let us note that the feasible set Δ_1 (defined by $x \in L^2$ and $0 \leqq x \leqq 1$) is convex.

This problem occurs in problems of statistical optimization (see, for instance, [2]).

* Received by the editors June 1, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Laboratoire d'Automatique Théorique, Faculté des Sciences, Université de Caen, Caen, France. This work was done under the author's direction by G. Borget and M. Valadier with the support of Electricité de France and Délégation Générale à la Recherche Scientifique et Technique under Contract No. 63 FR 199.

2. Existence of the solution. Let us denote by L^1 the space of all integrable functions defined on [0, 1] and by L^∞ the space of all bounded measurable functions defined on [0, 1] with the norm

$$\|u\|_\infty = \text{ess. sup } |u(t)|.$$

The feasible set Δ_1 can be considered as the closed ball of center 1/2 and radius 1/2 in L^∞. As L^∞ is the dual space of L^1, Δ_1 is weakly-star compact in L^∞ and a fortiori weakly compact in L^2.

On the other hand, $\mathbf{A}$ is a completely continuous operator and therefore, if $x_n \to x$ weakly, then $\mathbf{A}x_n \to \mathbf{A}x$ strongly and $\langle x_n, \mathbf{A}x_n\rangle \to \langle x, \mathbf{A}x\rangle$. Finally, f is weakly continuous on a weakly compact set. Therefore f attains its lower bound and problem (I) has a solution.

3. Conditions for optimality. An easy computation gives the following value for the gradient of f:

$$\nabla f(x) = \mathbf{A}x + b.$$

This gradient is an element of L^2.

Now it can be proved[1] that each of the following conditions is necessary and sufficient for $\tilde{x}$ to be optimal for problem (I):

$$(1) \qquad \begin{cases} \tilde{x}(t) = 0 & \Rightarrow (\mathbf{A}\tilde{x} + b)(t) \geqq 0 \text{ (a.e.)}, \\ \tilde{x}(t) = 1 & \Rightarrow (\mathbf{A}\tilde{x} + b)(t) \leqq 0 \text{ (a.e.)}, \\ 0 < \tilde{x}(t) < 1 & \Rightarrow (\mathbf{A}\tilde{x} + b)(t) = 0 \text{ (a.e.)}. \end{cases}$$

(2) There exist two functions (called multipliers) $\tilde{\lambda}$ and $\tilde{\mu}$, belonging to L^2, such that

$$\mathbf{A}\tilde{x} + b = \tilde{\lambda} - \tilde{\mu},\ \tilde{\lambda} \geqq 0,\ \tilde{\mu} \geqq 0,$$

$$\tilde{x}(t) > 0 \Rightarrow \tilde{\lambda}(t) = 0 \text{ (a.e.)},$$

$$\tilde{x}(t) < 1 \Rightarrow \tilde{\mu}(t) = 0 \text{ (a.e.)}.$$

The two last conditions can be summarized as follows:

$$\langle \tilde{\lambda}, \tilde{x}\rangle + \langle \tilde{\mu}, 1 - \tilde{x}\rangle = 0.$$

(3) There exist two functions $\tilde{\lambda}$ and $\tilde{\mu}$, belonging to L^2, such that

$$\Phi(\tilde{x}; \lambda, \mu) \leqq \Phi(\tilde{x}; \tilde{\lambda}, \tilde{\mu}) \leqq \Phi(x; \tilde{\lambda}, \tilde{\mu}),$$

for all $x, \lambda, \mu \in L^2$, $\lambda \geqq 0$, $\mu \geqq 0$,

[1] The proof has been carried out by direct methods, without using the general results of Hurwicz and Uzawa [1].

where Φ is the function defined as follows:

$$\Phi(x; \lambda, \mu) = f(x) - \langle \lambda, x \rangle - \langle \mu, 1 - x \rangle.$$

The condition (3) is the saddle point form of the condition for optimality.

Remark. If $\tilde{x}$ is optimal and if $\tilde{\lambda}$ and $\tilde{\mu}$ satisfy (1) or (2), then we have $\tilde{\lambda}(t)\tilde{\mu}(t) = 0$ and therefore $\tilde{\lambda} = \mathcal{P}^+(\mathbf{A}x + b)$, $\tilde{\mu} = \mathcal{P}^-(\mathbf{A}x + b)$, where $\mathcal{P}^+(u)$ and $\mathcal{P}^-(u)$ denote respectively the positive part and the negative part of u.

4. Duality theorems. A duality theorem was given by Dorn [3] for quadratic programming. It has been generalized by Wolfe [6] for convex programming. We shall rather recall the formulation of Wolfe, though the problem is quadratic.

Let us consider the problem:

(A) *Find* $x \in \mathbf{R}^n$ *minimizing* $f(x)$,

subject to the constraints $g^i(x) \geqq 0$, $i = 1, \cdots, m$,

where f and g^i are continuously differentiable functions, f is convex and g^i is concave.

Following Wolfe, we introduce the "dual" problem:

(B) *Find* $x \in \mathbf{R}^n$ *and* $\lambda \in \mathbf{R}^m$ *maximizing*

$$\Phi(x, \lambda) = f(x) - \sum_{i=1}^{m} g^i(x),$$

subject to the constraints

$$\lambda_i \geqq 0, \quad \nabla f(x) = \sum_{i=1}^{m} \lambda_i \nabla g^i(x).$$

Let Δ_1 and Δ_2 be the feasible sets respectively for problem (A) and problem (B). Then the theorem of Wolfe states, under some conditions of regularity, that if $\tilde{x}$ is optimal for problem (A), there exists $\tilde{\lambda}$ such that the couple $\tilde{x}$, $\tilde{\lambda}$ is optimal for problem (B).

Analogously, the following "dual" problem can be associated to problem (I).

(II) *Find* $x, \lambda, \mu \in L^2$ *maximizing*

$$\Phi(x; \lambda, \mu) = f(x) - \langle \lambda, x \rangle - \langle \mu, 1 - x \rangle,$$

subject to the constraints

$$\lambda, \mu \geqq 0, \qquad \nabla f(x) = \lambda - \mu.$$

It has been proved that if $\tilde{x}$ is optimal for problem (I) and if $\tilde{\lambda}$ and $\tilde{\mu}$ are the corresponding multipliers, then the triplet $(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$ is optimal for

problem (II). It can also be proved that if $(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$ is optimal for problem (II), then $\tilde{x}$ is optimal for problem (I).

Now let us define the function $\hat{\Phi}$ by

$$\hat{\Phi}(x) = \max_{\substack{\lambda,\mu \geqq 0 \\ \lambda - \mu = \nabla f(x)}} \Phi(x; \lambda, \mu).$$

It can easily be verified that the maximum is reached for

$$\lambda = \wp^+(\mathbf{A}x + b), \quad \mu = \wp^-(\mathbf{A}x + b),$$

and therefore we have

$$\hat{\Phi}(x) = -\tfrac{1}{2}\langle x, \mathbf{A}x\rangle - \langle \wp^-(\mathbf{A}x + b), 1\rangle.$$

A necessary and sufficient condition for $\tilde{x}$ to be optimal for problem (I) is that $\tilde{x}$ maximize $\hat{\Phi}$ (without constraints).

5. A differentiability problem. When examining if $\hat{\Phi}$ is differentiable or not, we meet the following problem: is the mapping $\wp^+$ differentiable? The following result has been proved by M. Valadier.

THEOREM. *Let μ be a finite measure, and $L^p(p \geqq 1)$ the space of all measurable functions x such that $\int |x|^p \cdot \mu < +\infty$. Assume that $x_0 \in L^p(p > 1)$ and that the set $\{t \mid x_0(t) = 0\}$ is of measure 0.*

(i) *The mapping $\wp^+$ from L^p to L^q with $1 \leqq q < p$ is differentiable at x_0, and its derivative is the operator*

$$h \longrightarrow 1_{\{x_0>0\}} h,$$

where $1_{\{x_0>0\}}$ is the characteristic function of the set $\{t \mid x_0(t) > 0\}$.

(ii) *The function $x \longrightarrow \int \wp^+(x) \cdot \mu$ is differentiable at x_0, and its derivative is the function $h \longrightarrow \int_{\{x_0>0\}} h \cdot \mu$.*

By application of this theorem we see that $\hat{\Phi}$ is differentiable at every point x such that $\mathbf{A}x + b$ (i.e., the gradient of f) is almost everywhere different from 0.

6. Computational aspects. We can propose two methods for solving numerically problem (I).

(i) The generalization of the Wolfe and Frank algorithm [4]. It has been proved by M. Valadier that this algorithm is valid in a Banach space for a weakly continuous function and a convex weakly compact feasible set [5].

(ii) The use of the theorem of duality, i.e., the maximization of the

function $\hat{\Phi}$. The results of the numerical experiences will be reported elsewhere.

REFERENCES

[1] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-linear Programming*, Stanford University Press, Stanford, California, 1958.

[2] W. Chervin, *Synthesis of optimal control program with taking account of random errors of its realization*, Avtomat. i Telemeh., 26 (1965), pp. 235–244.

[3] W. S. Dorn, *Duality in quadratic programming*, Quart. Appl. Math., 18 (1960), pp. 155–164.

[4] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[5] M. Valadier, *Généralisation d'un algorithme de Frank et Wolfe*, to appear in Rev. Française Recherche Opérationnelle, 36.

[6] P. Wolfe, *A duality theorem for non-linear programming*, Quart. Appl. Math., 19 (1961), pp. 239–244.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

DUALITY AND A DECOMPOSITION TECHNIQUE*

J. D. PEARSON†

1. Introduction. In 1927, K. O. Friedrichs demonstrated that a convex variational problem could be Legendre transformed into an equivalent concave variational problem [1]. This work was reported by R. Courant and has since been rediscovered [2], [3], [4], [5], spurred principally by the development of duality principles in convex programming [6]. However, convex programming techniques can also be extended by use of a function space treatment of the variational problem [7].

The primal and dual problems discussed here are related by the fact that finding the lowest point on the graph of a convex function is equivalent to finding the highest tangent plane underneath the graph. Use of the Legendre transform enables a symmetric treatment of both problems [8]. However, the complete eliminations required by the transformation are unnecessary and result in the simpler but unsymmetric primal and dual problems reported here. The associated composite program is largely trivial for control type problems, because of the preponderance of equality constraints.

The principle contribution of this paper is to present a decomposition technique by which a convex control programming problem having "coupled subsystem" constraints can be decomposed into smaller subproblems. Coordinating these subproblems is shown to be the dual problem. This technique is based on ideas due to Dantzig [9], Mesarovic [10], and Lasdon [11], and is aimed at a theory of multilevel or hierarchical control originated by Mesarovic.

Notation. A vector has scalar components $(y^1, y^2, \cdots, y^n)$ and vector components $(y_1, y_2, \cdots, y_N)$. x' denotes dx/dt, f_x denotes $\partial f/\partial x$; $y\,|_0 \equiv y(0)$, $y\,|^{t_1} \equiv y(t_1)$. $\langle x, y\rangle = x^T y$ shows the inner product and transpose notation. $f(x)$ is convex if $f(x_2) \geqq f(x_1) + \langle f_{x_1}, x_2 - x_1\rangle$. x^0 denotes optimal x value and $f(x^0) \equiv f^0$.

2. Primal, dual and composite programs. Let f, g, R be convex, twice differentiable functions of y, m, where y, $p \in E^n$, $m \in E^r$, $q \in E^s$, for $t \in [0, t_1]$. Suppose that f has strict convexity in y and m.

* Received by the editors June 17, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Systems Research Center, Case Institute of Technology, University Circle, Cleveland, Ohio. This research was supported in part by Contracts NSF GP 3118 and Nonr-1141(12).

PRIMAL PROBLEM.

$$\text{Minimize } v = g(y, t_1) + \int_0^{t_1} f(y, m)\, dt,$$

$$\text{subject to } y' = Ay + Bm, \quad y(0) = y_0, \tag{2.1}$$

$$R^i(y, m) \leqq 0, \quad i = 1, 2, \cdots, s. \tag{2.2}$$

R is supposed to satisfy a constraint condition [12, p. 148].

DUAL PROBLEM.

$$\text{Maximize } \omega = g^*(y, t_1) + \langle p(0), y_0 \rangle + \int_0^{t_1} f^*(y, m, p, q)\, dt,$$

$$\text{subject to } p' + A^T p + h_y = 0, \quad p(t_1) = g_y, \tag{2.3}$$

$$h_m + B^T p = 0, \tag{2.4}$$

$$q \geqq 0, \tag{2.5}$$

where, by definition,

$$\begin{aligned} g^* &= g - \langle g_y , y \rangle, \\ f^* &= h - \langle h_y , y \rangle - \langle h_m , m \rangle, \\ h &= f + \langle q, R \rangle. \end{aligned} \tag{2.6}$$

Assume that the differential constraints (2.1), (2.3) are instantaneously controllable for all t such that both variational problems are normal.

COMPOSITE PROBLEM. Minimize $c(m, q)$ such that

$$c(m, q) = \langle g_y , y \rangle \,|^{t_1} - \langle p, y_0 \rangle \,|_0 + \int_0^{t_1} [\langle h_y , y \rangle + \langle h_m , m \rangle - \langle q, R \rangle]\, dt,$$

$$\text{subject to } y' = Ay + Bm, \quad y(0) = y_0,$$

$$R(y, m) \leqq 0,$$

$$p' + A^T p + h_y = 0, \quad p(t_1) = g_y,$$

$$h_m + B^T p = 0,$$

$$q \geqq 0.$$

The composite problem is the difference between the primal and dual objectives subject to all the constraints (2.1)–(2.5).

Let (y^0, m^0, p^0, q^0) be the unique solution to (2.1)–(2.5) together with

$$\langle q, R\rangle = 0. \tag{2.7}$$

This is the extremal curve for all three problems since these equations are the first order necessary conditions for the problems [12, Theorem 2, p. 155].

PROPOSITION 1. *Let y, m, p, q be any other solution to the composite constraints* (2.1)-(2.5). *Then*

$$c(m, q) \geqq c(m^0, q^0) = 0. \tag{2.8}$$

Proof. Application of the constraint equations (2.1)–(2.5) yields

$$\int_0^{t_1} [\langle h_y , y\rangle + \langle h_m , m\rangle]\, dt = \langle p, y_0\rangle |_0 - \langle y, g_y\rangle |^{t_1}, \tag{2.9}$$

whence

$$c(m, q) = -\int_0^{t_1} \langle q, R\rangle\, dt.$$

However since $\langle q, R\rangle \leqq 0$ by (2.2) and (2.5), then

$$c(m, q) \geqq 0 = c(m^0, q^0)$$

from (2.7) and the definition of (y^0, m^0, p^0, q^0).

Define

$$L(y,m,p,q) = \langle e_0, y - y_0\rangle|_0 + g + \int_0^{t_1} [f + \langle p, Ay + Bm - y'\rangle + \langle q, R\rangle]\, dt, \tag{2.10}$$

where e_0, p, q are multipliers with e_0 determined by $y(0) = y_0$.

PROPOSITION 2. *With y^0, m^0, p^0, q^0 defined as before, it follows that subject to the composite constraints* (2.1)–(2.5) *and any y, $m \neq y^0$, m^0,*

$$L(y, m, p^0, q^0) > L(y^0, m^0, p^0, q^0) = v^0, \tag{2.11}$$

where v^0 is the optimal minimal value of the primal objective.

Proof. This follows since the integrand of (2.10) has strict convexity in (y, m) by definition of f. Using (2.3) and (2.4) and using (2.6) to define h,

$$\begin{aligned} &L(y, m, p^0, q^0) - L(y^0, m^0, p^0, q^0) \\ &> \langle g_y{}^0, y - y^0\rangle|^{t_1} + \langle e_0, y - y^0\rangle|_0 + \int_0^{t_1} [\langle h_y{}^0, y - y^0\rangle + \langle h_m{}^0, m - m^0\rangle \\ &\qquad + \langle A^T p^0, p - p^0\rangle + \langle B^T p^0, m - m^0\rangle + \langle p^0, y' - y'^0\rangle]\, dt \\ &= \langle g_y{}^0 - p^0, y - y^0\rangle|^{t_1} + \langle p^0 + e_0, y - y^0\rangle|_0 + \int_0^{t_1} [\langle h_y{}^0 + A^T p^0 + p'^0, y - y^0\rangle \end{aligned}$$

$$+ \langle h_m{}^0 + B^T p^0, m - m^0 \rangle] \, dt$$

$$= 0.$$

The right hand side of (2.11) is completed using (2.1) and (2.7) in (2.10).

COROLLARY 1. *Using Proposition* 2 *it follows that if* (y, m) *satisfies* (2.1), (2.2) *and* (2.7), *then*

$$\nu > \nu^0. \tag{2.12}$$

Define

$$M(y, m, p, q, y^*, m^*)$$

$$= \langle p, y_0 \rangle|_0 + g^* + \langle e_1, g_y - p \rangle|^{t_1} \tag{2.13}$$

$$+ \int_0^{t_1} [f^* + \langle y^*, A^T p + h_y + p' \rangle + \langle m^*, B^T p + h_m \rangle] \, dt,$$

where e_1, y^*, m^* are multipliers for the dual variational problem with e_1 determined by $p(t_1) - g_y = 0$.

PROPOSITION 3. *With* y^0, m^0, p^0, q^0 *defined as before, it follows that subject to the composite constraints* (2.1)–(2.5) *with* $(y, m, p, q) \neq (y^0, m^0, p^0, q^0)$,

$$\omega^0 = M(y^0, m^0, p^0, q^0, y^0 m^0) > M(y, m, p, q, y^0, m^0). \tag{2.14}$$

Proof. Using the convexity of f and g along the optimal solution y^0, m^0, p^0, q^0, then choosing $y^* = y^0, m^* = m^0$, where y^*, m^* satisfy (2.1), (2.2), it follows that for any y, m, p, $q \geqq 0$,

$$M(y^0, m^0, p^0, q^0, y^*, m^*) > M(y, m, p, q, y^*, m^*). \tag{2.15}$$

Eliminating y^*, m^* establishes the right hand side of (2.14), while the left hand side follows by enforcing (2.3) and (2.4).

COROLLARY 2. *Using Proposition* 3 *it follows that for* y, m, p, q *satisfying* (2.3), (2.4) *and* (2.5), *then*

$$\omega^0 > \omega. \tag{2.16}$$

COROLLARY 3. *Using Propositions* 1, 2, 3 *and* (2.9) *in the definition of* ω, *the dual objective, then*

$$\nu^0 = \omega^0,$$

whence

$$\nu > \nu^0 = \omega^0 > \omega$$

for y, m, p, q *satisfying appropriate constraints among* (2.1)–(2.5) *and* (2.7).

Extensions. A certain amount of trade off is possible between the allowable nonlinearities.

(i) Use of the Weierstrass or Clebsch condition gives convexity with respect to m when

$$y' = Ay + \phi(m),$$
$$R^i(y, m^i) \leqq 0.$$

(ii) Generalization to the case

$$y' = G(y, m)$$

is possible if G is convex and restrictions

$$G_{iyj} > 0$$

are imposed to give $p(t) \geqq 0$ [7].

3. Quadratic control programming. The quadratic control programming problem is defined as follows: given

$$y' = Ay + Bm, \quad y(0) = y_0, \quad Cy + Dm \leqq 0,$$

minimize

$$v = \frac{1}{2}\langle y, Py\rangle|^{t_1} + \frac{1}{2}\int_0^{t_1} [\langle y, Qy\rangle + \langle m, Rm\rangle]\, dt.$$

The dual problem is readily found using §2; given

$$p' = -A^T p - Qy - C^T q, \quad p(t_1) = Py(t_1),$$
$$q \geqq 0, \quad Rm + B^T p + D^T q = 0,$$

maximize

$$\omega = \langle y_0, p(0)\rangle - \frac{1}{2}\langle y, Py\rangle|^{t_1} - \frac{1}{2}\int_0^{t_1} [\langle y, Qy\rangle + \langle m, Rm\rangle]\, dt.$$

Both problems can be solved by expansion techniques resulting in Riccati equations. Either of these problems can be recast as a best approximation problem, thus relating the "Kalman dual" to quadratic programming as will be described elsewhere.

4. Decomposition and coordination problems as duals. Suppose that the previous primal control problem is such that

(i) the constraints represent a collection of N interconnected subsystems,

(ii) the objective functional is separable into N objective functions.

This section demonstrates that this primal problem can be decomposed into N independent parametric subproblems. The subproblem solutions

can be coordinated by adjusting the parameters so as to solve the dual control problem, and hence the original problem. In a limited sense decomposition and coordination are dual techniques. Let f_i, g_i, R_i be convex twice differentiable functions of y_i, x_i, m_i, and let f_i be strictly convex in y_i, x_i, m_i for $t \in [0, t_1]$.

PRIMAL CONTROL PROBLEM. Minimize

$$v = \sum_i^N \left\{ g_i(y_i, x_i) + \int_0^{t_1} f_i(y_i, m_i, x_i)\, dt \right\},$$

$$y_i' = A_i y_i + B_i m_i + C_i x_i, \qquad y_i(0) = T_i, \tag{4.1}$$

$$R_i^j(y_i, m_i, x_i) \leqq 0, \qquad j = 1, 2, \cdots, s_i, \tag{4.2}$$

$$x_i = \sum_{j \neq i}^N N_{ij} y_j, \qquad i = 1, 2, \cdots, N. \tag{4.3}$$

Equation (4.3) represents an interconnection constraint where N_{ij} is an incidence matrix connecting some components of y_j to x_i.

DUAL CONTROL PROBLEM. Maximize

$$\omega = \sum_i \left\{ g_i^* + \langle p_i(0), c_i \rangle + \int_0^{t_1} f_i^*\, dt \right\},$$

$$p_i' + A_i^T p_i + h_{iy_i} + \sum_{j \neq i}^N N_{ij}^T \pi_j = 0, \qquad p_i(t_1) = g_{iy_i}, \tag{4.4}$$

$$B_i^T p_i + h_{im_i} = 0, \tag{4.5}$$

$$h_{ix_i} + C_i^T p_i - \pi_i = 0, \tag{4.6}$$

$$q_i \geqq 0, \qquad i = 1, 2, \cdots, N, \tag{4.7}$$

where

$$\begin{aligned} h_i &= f_i + \langle q_i, R_i \rangle, \\ f_i^* &= h_i - \langle h_{iy_i}, y_i \rangle - \langle h_{ix_i}, x_i \rangle - \langle h_{im_i}, m_i \rangle, \\ g_i^* &= g_i - \langle g_{iy_i}, y_i \rangle - \langle g_{ix_i}, x_i \rangle. \end{aligned} \tag{4.8}$$

The complementary slackness equation has the form

$$\langle q_i, R_i \rangle = 0. \tag{4.9}$$

The decomposition technique follows immediately because of the introduction of the redundant variables x_i which form interconnections between subsystems.

Associated with the primal optimal solution y_i^0, x_i^0, m_i^0, p_i^0, π_i^0, q_i^0, which satisfies the composite constraints (4.1)–(4.9), minimizes v, and maximizes

ω, a Lagrangian $\mathfrak{L}$ can be defined,

$$\mathfrak{L}^0 = \sum_i^N \Big\{ f_i^0 + \langle q_i^0, R_i^0 \rangle + \langle p_i^0, A_i y_i^0 + B_i m_i^0 + C_i x_i^0 - y_i'^0 + \langle \pi_i^0, \sum_{j \neq i}^N N_{ij} y_j^0 - x_i^0 \rangle \Big\} \tag{4.10}$$

$$= \sum_i^N \Big\{ f_i + \langle q_i^0, R_i^0 \rangle + \langle p_i^0, A_i y_i^0 + B_i m_i^0 + C_i x_i^0 - y_i'^0 \rangle + \langle y_i^0, \sum_{j \neq i}^N N_{ij}^T \pi_j^0 \rangle - \langle \pi_i^0, x_i^0 \rangle \Big\}. \tag{4.11}$$

As a consequence of the interconnection constraint in this Lagrangian, rearrangement in the manner of (4.11) reveals that it can equally well be a sum of N sub-Lagrangians corresponding to N parametric subproblems.

The latter, parametric subproblems are defined as:

PARAMETRIC SUBPROBLEMS. Minimize

$$v_i(\pi) = g_i + \int_0^{t_1} \Big[f_i + \langle y_i, \sum_{j \neq i}^N N_{ij}^T \pi_j \rangle - \langle \pi_i, x_i \rangle \Big] dt, \tag{4.12}$$

$$\text{subject to } y_i' = A_i y_i + B_i m_i + C_i x_i, \quad y_i(0) = c_i,$$

$$R_i(y_i, m_i, x_i) \leqq 0.$$

The input variables x_i, m_i are to be chosen independently for all i.

By virtue of the strict convexity of $f_i(y_i, x_i, m_i)$ it can be seen that for any given $\pi(t)$, $0 \leqq t \leqq t_1$, a unique minimizing solution exists if the primal solution exists, denoted by $y_i^0(\pi)$, $x_i^0(\pi)$, $m_i^0(\pi)$, $p_i^0(\pi)$, $q_i^0(\pi)$ on $[0, t_1]$. Associated with this solution is a sub-Lagrangian function,

$$\mathfrak{L}_i = \Big\{ f_i + \langle q_i, R_i \rangle + \langle y_i, \sum_{j \neq i}^N N_{ij}^T \pi_j \rangle - \langle \pi_i, x_i \rangle + \langle p_i, A_i y_i + B_i m_i + C_i x_i - y_i' \rangle \Big\}. \tag{4.13}$$

First order necessary conditions are satisfied for $i = 1, 2, \cdots, N$, comprising for a given π, equations (4.1)–(4.9) with the *exception* in general of (4.3), i.e.,

$$x_i^0(\pi) \neq \sum_{j \neq i}^N N_{ij} y_j^0(\pi), \tag{4.14}$$

if $\pi \neq \pi^0$. However, it follows from comparison of (4.13) and (4.10) or (4.11) that $\pi = \pi^0$ exists such that for all i,

$$\lim_{\pi \to \pi^0} \{y_i^0(\pi)\} = y_i^0(\pi^0) = y_i^0, \quad \text{etc.},$$

thus satisfying (4.14). Furthermore, this optimal π^0 clearly maximizes the dual functional ω, since if $\pi \neq \pi^0$, (4.14) indicates that the first order necessary conditions for the dual, i.e., (4.1)–(4.7), are *not* all satisfied, i.e.,

$$\omega^0 > \omega(y^0(\pi), x^0(\pi), m^0(\pi), p^0(\pi), q^0(\pi)), \quad \pi \neq \pi^0. \tag{4.15}$$

To summarize then:

PROPOSITION 4. *If the primal integrated problem has a solution, then*
(i) *subsolutions to the N parametric subproblems exist for all continuous functions* $\pi(t)$,
(ii) *an optimal* $\pi^0(t)$ *exists which causes the N sets of subproblem solutions to satisfy the composite constraints,*
(iii) *the optimal* $\pi^0(t)$ *causes the parametric subproblem solutions to maximize the dual functional* ω.

In the parlance of multilevel control theory it could be said that this decomposition is "two-level" such that the "first level activity" comprises N parametric subproblems which are directed from a "second level" by manipulating π to maximize the dual functional ω.

A gradient scheme of adjustment will achieve the coordination required. although there are superior ways of doing so.

Let $\pi_0(t)$, $0 \leqq t \leqq t_1$, be an initial guess of $\pi^0(t)$ and consider $\pi(t, \sigma)$ for $0 \leqq \sigma \leqq \infty$ such that (by definition)

$$\pi(t, 0) = \pi_0(t),$$

and

$$\frac{d\pi}{d\sigma}(t, \sigma) \in C', \qquad 0 \leqq t \leqq t_1, \qquad 0 \leqq \sigma \leqq \infty.$$

Since from the strict convexity of f_i, $y^0(\pi)$, $x^0(\pi)$, $m^0(\pi)$, $p^0(\pi)$, $q^0(\pi)$ are continuous functions of π, the derivative of ω with respect to σ, evaluated along the trajectories $y^0(\pi)$, $x^0(\pi)$, $m^0(\pi)$, $p^0(\pi)$, $q^0(\pi)$ of the parametric subproblems as they vary with σ, is easily found from §2 to be

$$\frac{d\omega}{d\sigma} = \sum_{i=1}^{N} \int_0^{t_1} \langle \mathfrak{L}^0_{\pi_i}, \frac{d\pi_i}{d\sigma} \rangle \, dt \geqq 0, \tag{4.16}$$

subject to

$$\frac{d\pi_i}{d\sigma} = Q_i(t)\mathfrak{L}^0_{\pi_i} = Q_i(t)\left(\sum_{j \neq i}^{N} N_{ij} y_j^0(\pi) - x_i^0(\pi)\right) \in C', \tag{4.17}$$

with $Q_i(t)$ being any positive definite continuous matrix. Equality occurs uniquely when $\pi = \pi^0$, (4.14) is satisfied and ω is maximal at ω^0.

Thus subject to (4.17), a gradient coordination rule, ω is monotonically increased as $\sigma \to \infty$ and is bounded above by ω^0.

PROPOSITION 5. *The rule*

$$\frac{d\pi_i}{d\sigma} = Q_i\left(\sum_{j\neq i}^{N} N_{ij}\, y_j{}^0(\pi) - x_i{}^0(\pi)\right), \qquad i = 1, 2 \cdots, N,$$

with

$$\pi(t, 0) = \pi_0(t),$$

will coordinate the N subproblems into maximizing the dual functional and thus solving the original control problem.

Extensions. Convexity with respect to x_i is not necessary if each subproblem is recognized to be singular with respect to x_i on $[0, t_1]$.

5. Conclusions. Sufficient conditions for a primal control programming problem to have a minimum also provide a maximum for a dual problem.

A particular application of the reciprocal nature of the primal-dual constraints is to generate a two-level decomposition and coordination procedure which enables the solution of N medium size convex control problems which together might comprise a prohibitively large integrated problem.

6. Acknowledgment. It is a pleasure to acknowledge discussion with my colleagues, Mihajlo Mesarovic, Sanjoy Mitter, Irving Lefkowitz, Leon Lasdon, and Cole Brosilow.

REFERENCES

[1] R. Courant and D. Hilbert, *Methods of Mathematical Physics I*, Interscience, New York, 1943, p. 233.

[2] R. Bellman, *Quasi-linearization and upper and lower bounds for variational problems*, Quart. Appl. Math., 19(1962), pp. 349–350.

[3] J. D. Pearson, *A successive approximation method*, Proceedings of Fourth Joint Automatic Control Conference, Minneapolis, 1963.

[4] M. A. Hanson, *Bounds for functionally convex optimal control problems*, J. Math. Anal. Appl., 8(1964), pp. 84–89.

[5] R. J. Ringlee, *Bounds for convex variational programming problems arising in power system scheduling and control*, Proceedings of Fifth Joint Automatic Control Conference, Stanford, 1964.

[6] J. B. Dennis, *Mathematical Programming and Electrical Networks*, John Wiley, New York, 1959.

[7] S. K. Mitter, Ph.D. thesis, University of London, 1965.

[8] S. Karlin, *Mathematical Methods and Theory in Games, Programming and Economics*, Vol. I, Addison-Wesley, Reading, Massachusetts, 1959, p. 218.

[9] G. Dantzig and P. Wolfe, *Decomposition principle for linear programs*, Operations Res., 8(1960), pp. 101–111.

[10] M. D. Mesarovic, *A general systems approach to organizational theory*, Systems Research Center Rpt. 2-A-61-2, 1961.

[11] L. S. Lasdon, *A multilevel technique for optimization*, Systems Research Center Rpt. 50-C-64-19, 1964.

[12] L. D. Berkovitz, *Variational methods in problems of control and programming*, J. Math. Anal. Appl., 3(1961), pp. 145–169.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

DECOMPOSITION OF LARGE-SCALE SYSTEMS*

P. VARAIYA†

1. Introduction. A considerable amount of effort has been devoted in recent years to develop decomposition techniques for the solution of large problems in mathematical programming. In all these cases the complete problem can be represented as a number of small subproblems tied together by coupling constraint equations or coupling variables. The various techniques make use of this structure and differ in the classes of problems that they can deal with.

The purpose of this paper is to present a slight modification of the usual problem in nonlinear programming. We will call this modified problem P. Next we will state, without proof, the theorem which yields conditions (named CP) for the solution of P. The conditions CP are very similar to the results of Kuhn and Tucker [1] and, in fact, they can be obtained by a parallel proof. Finally, we will show how different specifications of P give rise to different classes of "decomposition problems". In each case, CP will yield "existence results" for resolving the decomposition problem, and in most cases we will present computational techniques.

2. Statement of P. Consider

$$\text{(P)} \qquad \max \{f(x) \mid Ax \in \Omega, x \in T\},$$

where $x \in E^n$, $\Omega \subseteq E^m$ and $T \subseteq E^n$ are closed convex sets, A is an $m \times n$ matrix with full rank and f is a real-valued, concave, differentiable function of x. In order to facilitate the statement of CP we adopt the following notation.

DEFINITION. *Let $K \subseteq E^l$ be a convex set and let $\underline{k} \in K$. By the polar generated by K at $\underline{k}$ we mean the set*

$$P(K, \underline{k}) = \{v \in E^l \mid \langle v, k\rangle \leqq \langle v, \underline{k}\rangle \forall k \in K\}.$$

THEOREM 1. *The vector $\underline{x} \in \Omega' \cap T$ is a solution of* P *if and only if*

$$(1) \qquad f'(\underline{x}) \in P(\Omega' \cap T, \underline{x}),$$

where $\Omega' = \{x \mid Ax \in \Omega\}$ and $f'(x)$ is the derivative of f at x.

* Received by the editors May 25, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Electronics Research Laboratory, University of California, Berkeley, California. This research was supported by the National Aeronautics and Space Administration under Grant NsG-354 (S-2).

Assumptions A1, A2 and A3 will be made throughout the remainder of this paper.

A1. $$P(\Omega' \cap T, \underline{x}) = P(\Omega', \underline{x}) + P(T, \underline{x}).$$

A2. $$P(\Omega', \underline{x}) = \overline{A^T[P(\Omega, \underline{x})]},$$

where the overbar denotes the closure of the set underneath.

A3. $A^T[P(\Omega, \underline{x})]$ is a closed set.

Finally, combining (1), A1, A2 and A3, we have CP. The vector $\underline{x} \in \Omega' \cap T$ is a solution of P if and only if

(CP) $$f'(\underline{x}) \in A^T[P(\Omega, A\underline{x})] + P(T, \underline{x}).$$

Remark. Theorem 1 can be usefully generalized to the case where f is an arbitrary differentiable function, A is an arbitrary differentiable mapping from E^n to E^m, and T is an arbitrary set. Equation (1) then becomes only a necessary condition, unless appropriate convexity restrictions are imposed. A sufficient condition for (2) to be valid is a suitable generalization of the Kuhn-Tucker constraint qualification, or the weak constraint of Arrow et al. For details and proofs of these results, see [2]. For presentation here we are limited to the simpler case shown above.

3. Classes of decomposition problems and computational techniques.

3.1. We specialize P to

(3) $$\max_x \{\langle c, x\rangle \mid Ax \in \Omega\},$$

where A is an $m \times n$ matrix with $n > m$. Let x be a solution of (3) and let $Ax = \underline{y}$. We can assume that there is an invertible submatrix $\underline{A}$ of A such that $x = (\underline{x}, 0)$ where $\underline{x} = (\underline{A})^{-1}\underline{y}$. If $u = (\underline{A}^T)^{-1}\underline{c}$, then we must have $A^T u = c$. By CP, x is a solution of (3) if and only if $c \in A^T P(\Omega, \underline{y})$, if and only if

(4) $$u \in P(\Omega, \underline{y}).$$

Now consider problems (4a) and (4b). Let $\underline{y} \in \Omega$ be a fixed vector. Let $x = (\underline{x}, 0)$ be a solution of

(4a) $$\max_x \{\langle c, x\rangle \mid Ax = \underline{y}\},$$

with $\underline{x} = (\underline{A})^{-1}\underline{y}$, and let $u = (\underline{A}^T)^{-1}\underline{c}$ be a shadow price vector. Now consider

(4b) $$\max_y \{\langle u, y\rangle \mid y \in \Omega\}.$$

Then by CP, $\underline{y}$ is a solution of (4b) if and only if

$$u \in P(\Omega, \underline{y}).$$

Combining the previous facts we have the following.

THEOREM 2. $x = (\underline{x}, 0)$ *is a solution of* (3) *if and only if* x *and* y *are solutions of* (4a) *and* (4b), *respectively.*

We can also give the following computational algorithm:

Step 1. Select $y \in \Omega$. Construct and solve (4a). Obtain $\underline{x}$ and u.

Step 2. Construct and solve (4b). Obtain the solution $\underline{y}'$ to (4b).

Step 3. If $\langle u, \underline{y}\rangle = \langle u, \underline{y}'\rangle$, stop. If $\langle u, \underline{y}\rangle < \langle u, \underline{y}'\rangle$, go to step 1 with $\underline{y}$ replaced by $\underline{y}'$.

Remark 1. Since there are a finite number of invertible submatrices of A, there will be only a finite number of cycles. However, each step 2 may involve nonfinite procedures.

Remark 2. At each step in the process we get feasible solutions.

3.2. Now consider the following special case of P:

$$\max \{\langle c, x\rangle \mid Ax \in \Omega, x \geqq 0\}, \tag{5}$$

where again A is an $m \times n$ matrix with $n > m$. Let x be a solution of (5) and let $Ax = \underline{y}$. Clearly, x also solves

$$\max_{x} \{\langle c, x\rangle \mid Ax = \underline{y}, x \geqq 0\}. \tag{5a}$$

For convenience, we shall assume that the solutions to problems of the form (5a) are nondegenerate. It is clear now that x can be assumed to have the form $x = (\underline{x}, 0)$, where $\underline{x} = (\underline{A})^{-1}\underline{y}$. Let $u = (\underline{A}^T)^{-1}\underline{c}$ be the shadow price vector. Then

$$c = A^T u + v, \tag{6}$$

with $v \leqq 0$, $\underline{v} = 0$. Moreover since x solves (5) we have by CP that

$$c \in A^T[P(\Omega, \underline{y})] + P(E^{n+}, \underline{x}), \tag{7}$$

where E^{n+} is the nonnegative orthant of E^n. Comparing (6) and (7), we have, by the nondegeneracy assumption,

$$u \in P(\Omega, \underline{y}). \tag{8}$$

Now consider

$$\max_{y} \{\langle u, y\rangle \mid y \in \Omega, (\underline{A})^{-1}y \geqq 0\}. \tag{5b}$$

Since (8) holds, it is clear by CP that y solves (5b). Conversely, suppose that $x = (\underline{x}, 0)$ and $\underline{y} = Ax$ solve (5a) and (5b), respectively. Equation

(6) is still valid with $\underline{v} = 0$ and $v \leqq 0$. Also by CP, the fact that $\underline{y}$ solves (5b) implies that

$$u \in P(\Omega, \underline{y}) + (\underline{A}^T)^{-1}[P(E^{m+}, (\underline{A})^{-1}\underline{y})].$$

Let $u_0 \in P(\Omega, \underline{y})$ and $v_0 \in P(E^{m+}, \underline{x})$ be such that

$$u = u_0 + (\underline{A}^T)^{-1}\underline{v}_0 .$$

By the nondegeneracy assumption we must have $v_0 = 0$. Combining (9) with (6), we see that (7) is satisfied so that by CP, $x = (\underline{x}, 0)$ solves (5). We have proved the next result.

THEOREM 3. *Under the nondegeneracy assumption, $x = (\underline{x}, 0)$ is a solution of* (5) *if and only if x and y are solutions of* (5a) *and* (5b), *respectively.*

Remark 1. The computational technique given in §3.1, after replacing (4a) and (4b) by (5a) and (5b), works for problem (5).

Remark 2. If we admit degenerate solutions to (5a), the technique given above is not valid without revision. The "only if" part of Theorem 3 still holds, but the "if" part does not. It appears that the alternate shadow price vector should be taken into account.

3.3. This time we consider a problem which is very similar to Rosen's convex partition programming problem [3],

$$\max_{x,y} \{\langle c, x\rangle \mid A^T x \geqq y \text{ for some } y \in \Omega'\}, \tag{9}$$

where A is an $m \times n$ matrix with $n > m$, $\Omega' \supseteq E^n$ is a set such that the set

$$\Omega = \{w + y \mid w \geqq 0, y \in \Omega'\}$$

is closed, and convex.

By CP, a feasible vector $\underline{x}$ is a solution of (9) if and only if

$$c \in A[P(\Omega, A^T\underline{x})]. \tag{10}$$

Let $y_0 \in \Omega'$ be a vector such that $A^T\underline{x} \geqq y_0$. Clearly $\underline{x}$ solves

$$\max_x \{\langle c, x\rangle \mid A^T x \geqq y_0\}. \tag{9a}$$

We may assume then, that $\underline{x} = (\underline{A}^T)^{-1}y_0$, where $y_0 = (y_0, y_0')$, and that the vector $u = (\underline{u}, 0) \leqq 0$, where

$$\underline{u} = (\underline{A})^{-1}c. \tag{11}$$

Next consider

$$\max_y \{\langle u, y\rangle \mid y \in \Omega, Q^T\underline{y} \geqq y'\}, \tag{9b}$$

where $\underline{A} = [\underline{A} \mid B]$, $Q = (\underline{A})^{-1}B$, and $y = (\underline{y}, y')$.

LEMMA. *If* $\underline{x}$ *solves* (9), *then* $\underline{x}$ *solves* (9a) *and* y_0 *solves* (9b).

Proof. Only the second assertion needs proof. Clearly y_0 is feasible. Let y be any feasible vector. Then

$$\begin{aligned}\langle u, y\rangle - \langle u, y_0\rangle &= \langle \underline{u}, \underline{y}\rangle - \langle \underline{u}, \underline{y}_0\rangle \\ &= \langle c, x\rangle - \langle c, \underline{x}\rangle,\end{aligned}$$

where $x = (\underline{A}^T)^{-1}y$ and $\underline{x} = (\underline{A}^T)^{-1}\underline{y}_0$. Clearly x is feasible for (9) so that $\langle c, x\rangle - \langle c, \underline{x}\rangle \leqq 0$.

We will now prove the converse result. Suppose we are given $\underline{x}$ and y_0 with $y_0 \in \Omega'$ such that $\underline{x}$ solves (9a) and y_0 solves (9b). The second postulate implies by CP that there are vectors u_0, v' such that $u_0 \in P(\Omega, y_0)$, $v' \leqq 0$, and

$$\langle Q^T\underline{y}_0 - y_0', v'\rangle = 0$$

and

$$u = u_0 + w, \tag{12}$$

where $\underline{w} = Qv'$ and $w' = -v'$. Combining (12) with (11) we have

$$Au = Au_0 = v. \tag{13}$$

In order to prove that (10) holds, it remains to show that $u_0 \in P(\Omega, A^T\underline{x})$. From the definition of the polar it suffices to prove that

$$\langle u_0, y_0\rangle = \langle u_0, A^T\underline{x}\rangle. \tag{14}$$

The following chain of equalities yield (14).

$$\begin{aligned}\langle u_0, y_0\rangle &= \langle u - w, y_0\rangle = \langle \underline{u}, \underline{y}_0\rangle - \langle Q^T\underline{y}_0 - y_0', v'\rangle \\ &= \langle \underline{y}, \underline{A}^T\underline{x}\rangle = \langle Au, \underline{x}\rangle = \langle Au_0, \underline{x}\rangle.\end{aligned}$$

We have thus proved Theorem 4.

THEOREM 4. *The vector* $\underline{x}$ *solves* (9) *if and only if* $\underline{x}$ *and* y_0 *solve* (9a) *and* (9b), *respectively.*

Remark. A computational algorithm similar to the ones suggested in §3.2 and §3.3 can be employed for this case also.

3.4. The last class of "decomposition problems" that we deal with here is a generalization of a technique due to Lasdon [4]. Consider the following special case of P,

$$\max_x \{f(x) \mid Ax = b, x \in T\}. \tag{15}$$

By CP, a feasible vector $\underline{x}$ solves (15) if and only if there exists a vector $\underline{y}$ such that for $\underline{u} = A^T\underline{y}$,

$$f'(\underline{x}) + \underline{u} \in P(T, \underline{x}). \tag{16}$$

Now consider

$$\max_{u} \{f(x) + \langle u, x\rangle \mid x \in T\}. \tag{15a}$$

Again by CP, a vector x solves (15a) if and only if

$$x \in T \text{ and } f'(x) + u \in P(T, x). \tag{17}$$

Comparing (16) and (17) we have the following "existence" theorem. We will assume that both (15) and (15a) have solutions.

THEOREM 5. (a) *If $\underline{x}$ solves* (15), *then $\underline{x}$ solves* (15a) *for $u = \underline{u}$.*
(b) *Conversely, if $u = \underline{u}$ in* (15a), *then there is a solution of* (15a) *which also solves* (15). *Moreover, this is the case if and only if the solution x of* (15a) *satisfies $Ax = b$.*

If we assume that the function f is strictly concave, the solutions will be unique so that if we can determine $\underline{u}$, then the solutions of (15) and (15a) are the same. Under certain circumstances, $\underline{u}$ can be obtained as follows.

Suppose $T = \{x \mid h(x) \geqq 0\}$, where h is a concave, vector-valued function satisfying the Kuhn-Tucker constraint qualifications. Let $u(t)$ be any vector and let $x(t)$ be the solution of (15a) with $u = u(t)$. Let $e(t) = Ax(t) - b$. Now change $u(t)$ according to the differential equation

$$\frac{du}{dt} = -A^T e(t). \tag{18}$$

Then, if T is compact, the following theorem can be proved.

THEOREM 6. *For any initial condition $u(0)$, the solution $u(t)$ of* (18) *converges to the required vector $\underline{u}$ and the solution $x(t)$ of* (15a) *converges to the solution $\underline{x}$ of* (15).

The proof of Theorem 6, although straightforward, is quite long and is therefore omitted. The interested reader is referred to Varaiya [5].

Acknowledgment. I would like to acknowledge the valuable guidance of Professor L. A. Zadeh, my research adviser, throughout the course of this research.

REFERENCES

[1] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability, University of California, Berkeley, 1950, pp. 481–492.
[2] P. VARAIYA, Ph.D. dissertation, University of California, Berkeley, 1965.
[3] J. B. ROSEN, *Convex partition programming*, Recent Advances in Mathematical Programming, P. Wolfe and R. L. Graves, eds., McGraw-Hill, New York, 1963, pp. 159–176.
[4] L. S. LASDON, *A multi-level technique for optimization*, Systems Research Center, Case Institute of Technology, Cleveland, Ohio, SRC 50-C-64-19, 1964.
[5] P. VARAIYA, *A decomposition technique for nonlinear programming* (to be published as an IBM Research Report).

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

PROGRAMMING UNDER UNCERTAINTY AND STOCHASTIC OPTIMAL CONTROL*

RICHARD VAN SLYKE† AND ROGER WETS‡

1. Introduction. Most optimization models (programming models, optimal control models, etc.) assume that the model's parameters (coefficients, functions, etc.) are well specified, either as best estimates, or by their expected values, and so on. In reality, however, these quantities are subject to uncertain or random variations of various kinds due to noise, component failure, unexpected demands, etc. Such discrepancies between reality and model can be reduced by assuming that all or some of the parameters are random variables with known probability distribution function.

Unfortunately, the complexity of such models, and of their solution, increases rapidly with the "amount" of uncertainty present in the problem. Nonetheless, different approaches and different techniques have given us some grip on a certain class of problems, for which there exist now "efficient" solution methods.

1a. Programming under uncertainty. In 1955 Dantzig formulated the two-stage linear program under uncertainty model [2]. The theory was furthered by Dantzig and Madansky [3], Madansky [5], Wets [8], and some special cases were investigated by Williams [9], [10] and Wets [7].

The *standard form* of a programming under uncertainty problem reads:

$$\begin{aligned} \text{Minimize} \quad & z(x) = cx + E_\xi\{\min qy\}, \\ \text{subject to} \quad & Ax = b, \\ & Tx + My = \xi, \quad \xi \quad \text{on} \quad (\Xi, \Im, F), \\ & x \geqq 0, \quad y \geqq 0, \end{aligned}$$

where A, T, and M are fixed matrices, c, q, b are constant vectors, x and y are variables, and ξ is a random vector defined on the probability space $(\Xi, \Im, F)$. The only random parameter present in this problem is ξ. The decision process described by this model is a two-stage process in which one

* Received by the editors June 30, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965. This research was partially supported by the Office of Naval Research under Contract NONR-222(83) with the Operations Research Center, University of California, Berkeley, California.

† University of California, Berkeley, California.

‡ Mathematics Research Laboratory, Boeing Scientific Research Laboratories, Seattle, Washington.

first selects x, then observes ξ and finally selects y so as to satisfy the constraints of the problem. The decision process is thus divided into two parts, but only the first one is of interest since once x is selected and ξ is observed, finding inf qy subject to $My = Tx - \xi$, $y \geqq 0$, is a deterministic problem. One procedure to solve such a problem is to exhibit a deterministic problem, whose set of optimal solutions is identical to the set of optimal solutions of our original problem. In general, such a deterministic problem exists, and it is shown that it has the form of a convex program. To find an explicit expression for this *equivalent* convex program is not always trivial, but it is possible to do so for an important class of problems [7].

1b. Sequential decision processes and stochastic optimization. It is not difficult to see that the two-stage programming under uncertainty problem can be generalized to an n-stage decision process where we have a sequence of decisions, observance of the behavior of the system and new decisions (corrective action). This idea is not new but literally illustrated by dynamic programming. Many stochastic optimization problems fall naturally in this framework, even if sometimes the concept of decision stage may only be a mathematical fiction, see [8, II.A].

1c. The stochastic optimal control problem. Usually, the stochastic optimal control problem is also formulated in the framework of a sequential decision process. But rather than dealing with a finite number of stages, it is assumed that the corrective actions are taken at every instant, i.e., at an infinite number of stages. To see this, it suffices to remark that a solution (control) for such a problem is not only expressed as a function of time, but also as a function of the actual state of the system [1], [4]. The observed state of the system consists then of the space-state determined by the control function affected by the interference of a random (noise) process.

In order to obtain an explicit expression for the solution of such a problem, or to find an algorithmic procedure leading to the solution, different assumptions have been made, explicitly or implicitly in the formulation of the problem. From a practical point of view, probably one of the weakest assumptions one could make is to assume that the number of corrective actions is finite, either at fixed time intervals or at some time intervals to be determined by the control system itself.

An n-stage control system can be described as follows: Let $x(t)$ describe the space state obtained by controlling the system with $u_1(t)$ for $0 \leqq t \leqq t_1$. Let $y(t)$ be the observed state of the process, i.e., $y(t) = x(t) + \xi(t)$, where $\xi(t)$ is a random (noise) function. If $u_2(t)$ is the second stage control for $t_1 \leqq t \leqq t_2$, we have $u_2(t) = \phi(t, y_1(t))$ or $\hat{\phi}(t, u_1(t), \xi(t_1))$ and similarly for $t_2 \leqq t \leqq t_3$, we have $u_3(t) = \psi(t, y(t_2))$ or $\hat{\psi}(t, u_1(t), \xi(t_1), u_2(t), \xi(t_2))$, and so on.

This structure is underlying our approach to the stochastic optimal control problem. We develop the theory for a two-stage system but the generalization to an n-stage process presents no mathematical difficulty. In §2 we derive the deterministic equivalent of the stochastic problem. §3 is devoted to a duality theory for this class of problems and its relation to the maximum principle. A projected paper will deal with the applications of the theoretical results obtained here to specific control problems.

2. The equivalent convex program.

2a. The problem. The *standard form* of the problem to be considered in this paper is:

$$\begin{aligned} (1) \qquad & \text{Find} \quad \inf z(u) = c(u) + E_\xi\{\inf q(v[\xi])\}, \\ & \text{subject to} \qquad A(u) \qquad\qquad = b, \\ & \qquad\qquad\qquad T(u, \xi) + W(v[\xi]) = d, \end{aligned}$$

where u is restricted to lie in some closed convex subset U of a Banach space $\mathfrak{U}$ and $v[\xi]$ must belong to a closed convex subset V of a Banach space $\mathfrak{V}$ for each ξ; b and d are points in $\mathfrak{R}^m$ and $\mathfrak{R}^{\bar{m}}$, respectively; ξ is a random variable defined on a probability space $(\Xi, \mathfrak{F}, F)$, (note that $v: \Xi \to \mathfrak{V}$); c and q are continuous convex functionals on $\mathfrak{U}$ and $\mathfrak{V}$, respectively; A, T, W are continuous linear operators such that $A: \mathfrak{U} \to \mathfrak{R}^m$, $T: \mathfrak{U} \times \Xi \to \mathfrak{R}^{\bar{m}}$, $W: \mathfrak{V} \to \mathfrak{R}^{\bar{m}}$; the operator E_ξ stands for expectation of the infimum of $q(v[\xi])$ with respect to ξ.

The process described by Problem (1) can be interpreted as follows: We first select a point in $\mathfrak{U}$, satisfying the constraints $A(u) = b$ and $u \in U$, say $\hat{u}$; we then observe the random event, say $\hat{\xi}$, and we are finally allowed to pick a point of $\mathfrak{V}$ such that $v \in V$, $W(v) + T(\hat{u}, \hat{\xi}) = 0$ and $q(v)$ is minimum. The decision process is thus divided into two stages. The second-stage decision is taken, when no uncertainties are left in the problem, i.e., when the random variable has been observed. This second stage is not our immediate interest here. Our primary interest is to find a *feasible* u which minimizes our total cost. Not only does our objective function take into account the immediate cost, $c(u)$, but also a weighted average of the cost of all the optimal second-stage decision a given u may lead to.

For the sake of simplicity, we shall assume that $(\Xi, \mathfrak{F}, F)$ is the probability space induced in $\mathfrak{R}^{\bar{m}}$. Ξ is a subset of $\mathfrak{R}^{\bar{m}}$, F is a probability measure generated by a distribution function also denoted by F and $\mathfrak{F}$ is the completion for F of the Borel algebra in $\mathfrak{R}^{\bar{m}}$. We shall assume that Ξ is convex. If this is not the case, we then replace it by its convex hull which we will also denote by Ξ and fill up $\mathfrak{F}$ with the appropriate sets of measure zero. Without loss of generality, we can assume that Ξ is of full dimension. If not,

we can change Problem (1) so as to include the deterministic second-stage constraints into the set of fixed first-stage constraints. Then, our new Ξ has full dimension. The probability distribution function F is continuous, discrete, or a mixture of both.

In view of the interpretation given to (1), it is easy to see that the second-stage decision (control) variable v is a function of the observed state of the system, viz., $d - T(u, \xi)$, and in particular a function of the random variable ξ. Thus, v is itself a random variable. This fact is expressed by our notation $v[\xi]$. Moreover, we do not make any assumptions on v as a function of ξ, e.g., as to its measurability. Since by the nature of the model it is "calculated" only for the value of ξ which actually occurs. We will, however, show that $E_\xi\{\inf q(v[\xi])\}$ makes sense.

In what follows we show that there exists an equivalent problem to (1), i.e., a problem with the same set of optimal solutions as (1), that can be expressed as the minimization of a convex functional on a convex set.

2b. The second stage problem. Once u is selected and ξ is observed, the second stage problem

$$\begin{aligned} &\text{Find} \quad \inf q(v), \\ &\text{subject to} \quad W(v) = d - T(u, \xi), \\ &\qquad\qquad v \in V, \end{aligned} \tag{2}$$

becomes a deterministic problem. Let

$$V(u, \xi) = \{v \mid W(v) + T(u, \xi) = d, v \in V\} \tag{3}$$

be the set of feasible solutions for (2), and let

$$Q(u, \xi) = \inf \{q(v) \mid v \in V(u, \xi)\} \tag{4}$$

be the functional describing the range of the infimum of $q(v)$ as a function of u and ξ. As we shall see later, we may restrict ourselves to the case where $V(u, \xi)$ is nonempty. The set $V(u, \xi)$ is convex and closed, but not necessarily compact. Thus, the functional $q(v)$ may fail to achieve its minimum on $V(u, \xi)$. We shall assume that $q(v)$ possesses finite infimum on $V(u, \xi)$. Such a condition is not very restrictive, because if for some u, $Q(u, \xi) \equiv -\infty$ for all ξ in Ξ, then $z(u) = -\infty$ and Problem (1) is of no interest. Moreover, if for some u, $Q(u, \xi) = -\infty$ for a proper subset of Ξ, we could still hope that this set would have measure zero, and our problem could have a meaningful solution. But it is not the case, since we shall show that if $Q(u, \xi) = -\infty$ for some ξ in Ξ, then $Q(u, \xi) \equiv -\infty$ for all ξ in Ξ. To do so, we need the following results.

PROPOSITION 1. *Fix u and let $V(u, \xi) \neq \varnothing$ for all ξ in Ξ. Then $Q(u, \xi)$ is a convex function in ξ on Ξ.*

Proof. For $\epsilon > 0$, we say that v determines an ϵ-inf of $q(v)$ on $V(u, \xi)$ if $v \in V(u, \xi)$ and $q(v) \leqq Q(u, \xi) + \epsilon$.

First, we shall assume that $Q(u, \xi) > -\infty$ for all ξ in Ξ. Let $\xi_0, \xi_1 \in \Xi$; then $\lambda\xi_0 + (1 - \lambda)\xi_1 = \xi_\lambda \in \Xi$ for $\lambda \in [0, 1]$. Let v_0 and v_1 determine ϵ-inf on $V(u, \xi_0)$ and $V(u, \xi_1)$, respectively. By the convexity of V and linearity of the operators W and T,

$$\lambda v_0 + (1 - \lambda)v_1 \in V(u, \xi_\lambda) \quad \text{for} \quad \lambda \in [0, 1].$$

Then

$$Q(u, \xi_\lambda) \leqq q(\lambda v_0 + (1 - \lambda)v_1);$$

also, by the convexity of the functional q,

$$q(\lambda v_0 + (1 - \lambda)v_1) \leqq \lambda q(v_0) + (1 - \lambda)q(v_1);$$

and since v_0 and v_1 determine ϵ-inf, we have

$$\lambda q(v_0) + (1 - \lambda)q(v_1) \leqq \lambda Q(u, \xi_0) + (1 - \lambda)Q(u, \xi_1) + \epsilon,$$

i.e.,

$$Q(u, \xi_\lambda) \leqq \lambda Q(u, \xi_0) + (1 - \lambda)Q(u, \xi_1) + \epsilon.$$

Since the above inequality holds for any ϵ, arbitrarily close to zero, we obtain

$$Q(u, \xi_\lambda) \leqq \lambda Q(u, \xi_0) + (1 - \lambda)Q(u, \xi_1).$$

Let us now consider the case where $Q(u, \xi)$ is not finite for all ξ in Ξ. Without loss of generality, we can assume that $Q(u, \xi_0) = -\infty$. If $Q(u, \xi_0) = -\infty$, then for all N arbitrarily large, there exists $v_0 \in V(u, \xi_0)$ such that $q(v_0) \leqq -N$. But

$$Q(u, \xi_\lambda) \leqq q(\lambda v_0 + (1 - \lambda)v_1);$$

and by convexity of the functional q,

$$q(\lambda v_0 + (1 - \lambda)v_1) \leqq \lambda q(v_0) + (1 - \lambda)q(v_1);$$

and since $Q(u, \xi_0) = -\infty$, there exists v_0 such that

$$\lambda q(v_0) + (1 - \lambda)q(v_1) \leqq -N$$

for any N; thus $Q(u, \xi_\lambda) \leqq -N$, i.e., $Q(u, \xi_\lambda) = -\infty$ for $\lambda \in (0, 1)$.

This implies that if there exists some ξ in Ξ such that $Q(u, \xi)$ has no lower bound, then $Q(u, \xi) \equiv -\infty$ for every ξ in the interior of Ξ and $Q(u, \xi)$ may be different from $-\infty$ at most on the boundaries of Ξ.

PROPOSITION 2. *If for a fixed u, $V(u, \xi) \neq \varnothing$ for all ξ in Ξ and at least one of the four following assumptions is satisfied*:

(i) *$q(v)$ is linear and V is a convex polyhedral subset of $\mathfrak{R}^n$,*
(ii) *V is compact,*
(iii) *$q(v)$ is weakly continuous on V and V is weakly compact,*
(iv) *Ξ is open,*
then $Q(u, \xi)$ is continuous in ξ on Ξ.

Proof. Since $Q(u, \xi)$ is convex, $Q(u, \xi)$ is continuous on the interior of Ξ (this proves the proposition under assumption (iv)). Thus, the only case of interest is when ξ is on the boundary of Ξ. The proposition under assumption (i) is proved in [8]. We limit ourselves to (ii) and (iii).

Let $\xi_0 \in \delta\Xi$ and $\xi_i \to \xi_0$, where each ξ_i belongs to the interior of Ξ. Under either (ii) or (iii) there exists a subsequence v^{ik} such that $q(v^{ik}) \to q(v^0)$ for some v^0 in V and such that $W(v^0) + T(u, \xi_0) = d$, where v^i is an ϵ-inf corresponding to ξ_i. Hence, $\lim_{i\to\infty} Q(u, \xi_i) \geqq \{\inf q(v) \mid v \in V(u, \lim_{i\to\infty} \xi_i)\} = Q(u, \xi_0)$. On the other hand, by the convexity of $Q(u, \xi)$, we have

$$Q(u, \xi_0) \geqq \lim_{k\to\infty} Q(u, \xi_k),$$

thus

$$\lim_{k\to\infty} Q(u, \xi_k) = Q(u, \xi^0).$$

Remark. The conditions (i), (ii), (iii) or (iv) are sufficient conditions to ensure the continuity of $Q(u, \xi)$. They are not necessary. In general, however, $Q(u, \xi)$ may fail to be continuous in ξ, as is shown by the following example, where $\mathcal{U}$ is of finite dimension. Let

$$V = \mathfrak{R}_+^2, \quad \Xi = [0, 1],$$

$$q(v) = q(x, y) = -\min(|\sqrt{xy}|, 1), \quad d = 0,$$

$$W(v) = x \quad \text{and} \quad T(u, \xi) = -\xi.$$

It is easy to see that $Q(u, \xi) = -1$ if $\xi \neq 0$ and $Q(u, 0) = 0$. Hence, $Q(u, \xi)$ is not continuous at $\xi = 0$.

COROLLARY. *For a fixed u, let $V(u, \xi) \neq \varnothing$ for all ξ in Ξ. If $Q(u, \xi) = -\infty$ for some ξ in Ξ and at least one of the conditions* (i), (ii), (iii) *or* (iv) *of Proposition 2 is satisfied, then $Q(u, \xi) \equiv -\infty$ for all ξ in Ξ.*

In what follows, we shall assume that either Ξ is open—or it can be redefined so that it is open—or that at least one of the conditions (i), (ii), or (iii) of Proposition 2 holds.

2c. The solution set. A fixed u and an observed ξ determine $Q(u, \xi)$ uniquely; then our only decision variable is u. It is in this context that we examine the solution set of Problem (1). Nonetheless, the second-stage decisions affect our first-stage decision, not only by the values assumed by $Q(u, \xi)$, but also by the restriction that we have to limit our set of ad-

missible first-stage decisions to those for which there exists a feasible second-stage decision.

DEFINITION. u is a *feasible solution to* (1), if $A(u) = b$, $u \in U$, and if the feasibility of Problem (2) is independent of the value assumed by ξ in Ξ. Let K be the set of feasible solutions to (1). Let

$$K_1 = \{u \in \mathfrak{U} \mid A(u) = b\} \cap U$$

be the set determined by the *fixed constraints.*

PROPOSITION 3. *K_1 is a closed convex subset of $\mathfrak{U}$.*

Proof. By linearity and continuity of the operator A and convexity of the closed set U.

Let $K_2 = \{u \in \mathfrak{U} \mid \xi \in \Xi,\ V(u, \xi) \neq \varnothing\}$ be the set representing the *induced constraints.* By induced, we mean that the set K_2 is determined by a condition to be satisfied at some later time, viz., the second-stage problem must be feasible for all ξ in Ξ.

Let $K_{2\xi} = \{u \in \mathfrak{U} \mid V(u, \xi) \neq \varnothing\}$; then $K_2 = \bigcap_{\xi \in \Xi} K_{2\xi}$. By the linearity of the operators W and T and convexity of V, $K_{2\xi}$ is convex. Thus:

PROPOSITION 4. *K_2 is a convex subset of $\mathfrak{U}$.*

Note that introducing the appropriate sets of measure zero, in order to replace the original Ξ by its convex hull, does not change the set K_2. Let $\tilde{\Xi}$ be the original probability space and let Ξ be its convex hull. Let $\tilde{K}_2 = \bigcap_{\xi \in \tilde{\Xi}} K_{2\xi}$ and K_2 as above. Obviously, $\tilde{K}_2 \supset K_2$ since the intersection is taken over a smaller index set. Thus, it suffices to show that $u \in \tilde{K}_2$ implies that $u \in K_2$. If $u \in \tilde{K}_2$, then $u \in K_{2\xi}$ for all ξ in $\tilde{\Xi}$, and $V(u, \xi) \neq \varnothing$ for all ξ in $\tilde{\Xi}$. If $\hat{\xi} \in \Xi$, but not to $\tilde{\Xi}$, then there exist $\xi_1, \cdots, \xi_{m+1}$ in $\tilde{\Xi}$ such that $\hat{\xi} = \sum_{i=1}^{m+1} \lambda_i \xi_i$ for some λ_i, $i = 1, \cdots, m+1$, such that $\sum_{i=1}^{m+1} \lambda_i = 1$. By linearity of W and T and since $V(u, \xi_i)$ is nonempty for $i = 1, \cdots, m$, so is $V(u, \hat{\xi})$. Thus, $V(u, \xi) \neq \varnothing$ for all ξ in Ξ, i.e., $u \in K_2$.

PROPOSITION 5. *The set K of feasible solutions to Problem* (1) *is a convex subset of $\mathfrak{U}$.*

Proof. $K = K_1 \cap K_2$.

2d. The objective function. To show that (1) can be reduced to an equivalent convex program, it now suffices to show that $z(u)$—the objective function of Problem (1)—is a convex function in u on K. Remark that $u \in K$ implies that $V(u, \xi)$ is nonempty for all ξ in Ξ.

PROPOSITION 6. *$Q(u, \xi)$ is convex in u on K.*

Proof. Fix ξ and take $u_0, u_1 \in K$; then $\lambda u_0 + (1 - \lambda)u_1 = u_\lambda \in K$. Since we assumed that $Q(u, \xi) > -\infty$, there exist v_0 and v_1 which determine ϵ-inf of $q(v)$ on $V(u_0, \xi)$ and $V(u_1, \xi)$, respectively. Also, by convexity of V and linearity of W and T, $\lambda v_0 + (1 - \lambda)v_1 \in V(u_\lambda, \xi)$. Then

$$Q(u_\lambda, \xi) \leqq q(\lambda v_0 + (1 - \lambda)v_1) \leqq \lambda q(v_0) + (1 - \lambda)q(v_1)$$
$$\leqq \lambda Q(u_0, \xi) + (1 - \lambda)Q(u_1, \xi) + \epsilon.$$

Since this inequality holds for all ϵ, we have

$$Q(u_\lambda, \xi) \leqq \lambda Q(u_0, \xi) + (1 - \lambda)Q(u_1, \xi) \quad \text{for} \quad \lambda \in [0, 1].$$

PROPOSITION 7. *Let* $Q(u) = E_\xi\{Q(u, \xi)\}$. *Then* $Q(u)$ *is convex in* u *on* K.

Proof. The function $Q(u, \xi)$ is continuous, thus Lebesgue measurable. But F is a Lebesgue-Stieltjes measure and $\mathfrak{F}$ contains the Borel algebra; thus $Q(u, \xi)$ is also F-measurable. Since F determines a positive measure, $Q(u)$ is the result of a weighted positive linear combination of convex functions. Thus $Q(u)$ is convex.

Since $c(u)$ is convex, we have shown that there exists an equivalent convex program to Problem (1), viz.,

$$\begin{aligned} &\text{Find} \quad \inf z(u) = c(u) + Q(u), \\ &\text{subject to} \quad u \in K, \end{aligned} \tag{5}$$

where no random elements are present any longer. Nonetheless, two main difficulties remain to be solved before one can use efficiently the techniques available for convex programs, namely, depending on the structure of the different operators of the original problem, to find an explicit expression for $Q(u)$ and the set K may be a major undertaking. As we shall show in a forthcoming paper, a certain interesting class of problems allows us to express $Q(u)$ and K explicitly, with relatively little effort.

3. Duality.

3a. The dual problem. Solution methods for any particular problem of the form (1) depend strongly on the form of the operators involved. However, as was the case in linear programming under uncertainty [8], there is a duality theory which plays a crucial role.

The second-stage problem (once u is selected and ξ is observed),

$$\begin{aligned} &\text{find} \quad \inf q(v), \\ &\text{subject to} \quad W(v) = d - T(u, \xi), \\ &\qquad\qquad\qquad v \in V, \end{aligned}$$

and the equivalent convex program,

$$\begin{aligned} &\text{find} \quad \inf c(u) + Q(u), \\ &\text{subject to} \quad A(u) = b, \\ &\qquad\qquad\qquad u \in K_2 \cap U, \end{aligned}$$

are in the same form. To develop the duality theory for this class of problems, it suffices to consider the following simple problem.

$$\begin{aligned} &\text{find} \quad \inf c(u) \\ (6) \qquad &\text{subject to} \quad A(u) = b, \\ &\qquad\qquad u \in U \subset \mathfrak{U}, \end{aligned}$$

where b, $c(u)$, U and $\mathfrak{U}$ are as defined in the previous section. We remember in particular that U is closed and convex and A is a continuous linear operator with range in $\mathfrak{R}^m$.

Let

$$C = \{p = (p_0, p_1 \cdots, p_m) \mid p_0 \geqq c(u), (p_1, \cdots, p_m) = A(u) - b, u \in U\}$$

and

$$\mathfrak{C} = \{(p_1, \cdots, p_m) \mid (p_1, \cdots, p_m) = A(u) - b, u \in U\}.$$

LEMMA 1. *C is convex.*

Proof. Suppose $p^1, p^2 \in C$ and suppose further that $u^1, u^2 \in U$ satisfy

$$p_0{}^i \geqq c(u^i), \quad (p_1{}^i, \cdots, p_m{}^i) = A(u^i) - b, \qquad i = 1, 2.$$

Let $p^\lambda = \lambda p^1 + (1 - \lambda)p^2$ and $u^\lambda = u^1 + (1 - \lambda)u^2$ for $\lambda \in [0, 1]$. Then

$$\begin{aligned} &\lambda(p_1{}^1, \cdots, p_m{}^1) + (1 - \lambda)(p_1{}^2, \cdots, p_m{}^2) \\ &\qquad = \lambda(A(u^1) - b) + (1 - \lambda)(A(u^2) - b) = A(u^\lambda) - b. \end{aligned}$$

Also, by convexity of the functional c we have

$$c(u^\lambda) \leqq \lambda c(u^1) + (1 - \lambda)c(u^2) \leqq \lambda p^1 + (1 - \lambda)p^2.$$

Unfortunately, it is not true, in general, that C is closed. Consider (6) with

$$\mathfrak{U} = l_2 = \{(u_i) \mid \sum u_i{}^2 < \infty\}, \qquad A(u) = u_1,$$

$$c(u) = \sum_{i=2}^{\infty} \frac{1}{2^i} u_i{}^2, \qquad U = \left\{u \;\middle|\; \sum_{i=2}^{\infty} u_i{}^2 = 1\right\}.$$

For any b in $\mathfrak{C}$, $\inf c(u) = 0$, but there exists no feasible u such that $c(u) = 0$. In particular, let u^i, $i = 2, \cdots$, be given by $u_j{}^i = \delta_{ij}$. Then $A(u^i) = 0$ and $c(u^i) = 1/2^i$, $i = 2, \cdots$. Thus, we have $p^i = (p_0{}^i, p_1{}^i) \geqq (1/2^i, 0) \to (0, 0) = p^0$, where $p^0 \notin C$ and each p^i does.

However, we will need $\mathfrak{C}$ closed. This will be the case when U is weakly sequentially compact or if it is a convex polyhedral subset of a finite Euclidean space. In general, we see that we are essentially seeking to find

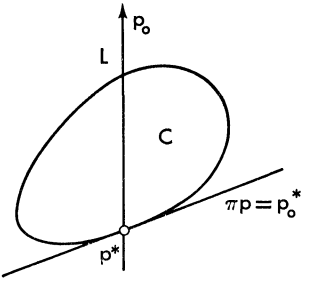

FIG. 1

the "lowest" point of C on the p_0 axis. That is, we can reformulate (6) as follows:

$$\begin{aligned} &\text{find} && \inf p_0, \\ &\text{subject to} && p \in L \cap C, \end{aligned} \tag{7}$$

where $L = \{(p_0, p_1, \cdots, p_m) | \, p_i = 0, i = 1, \cdots, m\}$. Problem (7) has the very natural dual,

$$\begin{aligned} &\text{find} && \sup \mu, \\ &\text{subject to} && \pi_0 = 1, \\ & && \pi p - \mu \geqq 0 \quad \text{for all} \quad p \quad \text{in} \quad C, \end{aligned} \tag{8}$$

where μ is a scalar, and π_0 is the first component of the $(m + 1)$-dimensional vector π.

If we think of $\pi p - \mu = 0$ as defining a hyperplane in $\mathfrak{R}^m$, then there is a one-to-one correspondence between feasible solutions of (8) and nonvertical supporting hyperplanes which are "below" the set C, in the sense that increasing p_0 means up.

Immediately, we have:

PROPOSITION 8 (*Weak duality*). $p_0 \geqq \mu$ *for all feasible solutions to* (7) *and* (8).

Proof. $\pi p - \mu \geqq 0$ by (8). Since p is feasible for (7), then $p_1 = \cdots = p_m = 0$ and hence $\pi_0 p_0 - \mu \geqq 0$. But $\pi_0 = 1$.

We now prove the following intuitively obvious duality theorem:

THEOREM 1 (*Strong duality*). *If the projection* $\mathfrak{C}$ *of* C *with respect to* p_0 *is closed, exactly one of the following occurs*:

(a) (7) *and* (8) *both admit feasible solutions, in which case* $\inf p_0 = \sup \mu$,

(b) (7) *is feasible and* (8) *is not, in which case* $\inf p_0 = -\infty$,

(c) (8) *is feasible and* (7) *is not, in which case* $\sup \mu = +\infty$,

(d) *neither* (7) *nor* (8) *is feasible.*

Proof. (a) By Proposition 8, $\inf p_0$ and $\sup \mu$ are finite. Let $\mu^* = \inf \{p_0 \mid p \in L \cap C\} > -\infty$. Clearly there exists p^* which belongs to $\tilde{C} \cap L$ such that $\mu^* = p_0^*$ and p^* is a boundary point of

$$\tilde{C} = \{(p_0, \cdots, p_m) \mid p_i = p_i', \ i = 1, \cdots, m, \ p_0 \geqq p_0', \ p' \in C\}.$$

Hence, there exists a supporting hyperplane to $\tilde{C}$ at p^*. Let it be defined by $\hat{\pi}p - \hat{\mu} \geqq 0$. Clearly $\hat{\pi}_0 \geqq 0$. If $\hat{\pi}_0 > 0$, division by $\hat{\pi}_0$ yields $\pi p - \mu \geqq 0$, where

$$\pi = \frac{1}{\hat{\pi}_0}(\hat{\pi}), \qquad \mu = \frac{\hat{\mu}}{\hat{\pi}_0}$$

for all $p \in C$. Since $\pi p^* - \mu = 0$ implies $p_0^* = \mu^* = \mu$, (π, μ) is optimal for (8) and $\inf p_0 = \sup \mu$. If for every supporting hyperplane of $\tilde{C}$ at p^*, we have that $\hat{\pi}_0 = 0$, a somewhat more complicated construction is necessary. Let $\epsilon > 0$ be arbitrary, and let $\tilde{p} = (p_0^* - \epsilon, 0, \cdots, 0)$; then $\tilde{p} \notin \tilde{C}$. Hence there exists a hyperplane separating strictly $\tilde{p}$ and $\tilde{C}$, i.e., there exist $\tilde{\pi}$, $\tilde{\mu}$ such that $\tilde{\pi}\tilde{p} - \tilde{\mu} < 0$ and $\tilde{\pi}p - \tilde{\mu} \geqq 0$ for all p in $\tilde{C}$. In particular, $\pi p^* - \mu \geqq 0$ but

$$\tilde{\pi}\tilde{p} = \tilde{\pi}_0\tilde{p}_0 = \tilde{\pi}_0 (p_0^* - \epsilon) < \tilde{\pi}_0 p_0^*$$

implies that $\tilde{\pi}_0 > 0$. Letting

$$\pi = \frac{1}{\tilde{\pi}_0}\tilde{\pi} \quad \text{and} \quad \mu = \tilde{p}_0 = p_0^* - \epsilon,$$

we have a feasible solution to (8) with $\mu = p_0^* - \epsilon$. Since ϵ is arbitrary, $\sup \mu = \inf p_0 = p_0^*$.

(b) If $\inf p_0 > -\infty$, a feasible solution to (8) exists by the same construction as in (a), but by hypothesis no feasible solution to (8) exists, therefore $\inf p_0 = -\infty$.

(c) Suppose $\sup \mu < +\infty$, then let $\tilde{\mu} = \sup \{\mu \mid \pi_0 = 1, \pi p - \mu \geqq 0$ for all $p \in C\}$ and let $\tilde{p} = (\tilde{\mu}, 0, \cdots, 0)$. We now establish that $\tilde{p} \notin \tilde{C}$. In fact, $\tilde{p} \notin \tilde{C}$, for if it did, $(\tilde{p}_1, \cdots, \tilde{p}_m) = (0, \cdots, 0)$ would belong to $\bar{\mathfrak{C}} = \mathfrak{C}$. But then $(c(0), 0, \cdots, 0)$ would be a feasible point for (7), which is assumed infeasible. Hence, $\tilde{p} \notin \tilde{C}$. Therefore, there is a hyperplane separating strictly $\tilde{p}$ and $\tilde{C}$ determined by, say, $\tilde{\pi}$, $\tilde{\mu}$, and such that

$$\tilde{\pi}\tilde{p} < \inf \{\tilde{\pi}p \mid p \in \tilde{C}\} \leqq \inf \{\tilde{\pi}p \mid p \in C\}.$$

By definition of $\tilde{C}$ and since C is nonempty, we have that $\tilde{\pi}_0 \geqq 0$. If $\tilde{\pi}_0 > 0$, let π, μ be given by

$$\pi = \frac{1}{\tilde{\pi}_0} \cdot \tilde{\pi} \quad \text{and} \quad \frac{\tilde{\pi}\tilde{p}}{\tilde{\pi}_0} = \tilde{\mu} < \mu \leqq \frac{1}{\tilde{\pi}_0} \inf \{\tilde{\pi}p \mid p \in C\};$$

then π, μ is a feasible solution to (8) with $\mu > \tilde{\mu}$, which contradicts the definition of $\tilde{\mu}$. Suppose now that $\tilde{\pi}_0 = 0$. Then $\tilde{\pi}\tilde{p} = 0$, hence $\tilde{\pi}p > \delta > 0$ for all $p \in \tilde{C}$. Let $\bar{\pi}$, $\bar{\mu}$ be any feasible solution to (8), then $\pi = (\bar{\pi} + \lambda\tilde{\pi})$, $\mu = (\bar{\mu} + \lambda\bar{\mu})$ is also feasible for any λ. Taking any $\lambda > 0$ contradicts the definition of $\tilde{\mu}$.

COROLLARY 1. *If p^* and π^*, μ^* are respectively feasible for* (7) *and* (8), *they are optimal if and only if*

$$\pi^* p^* = \mu^*.$$

Proof. $(\pi^* p^*) = \langle (1, \pi_1^*, \cdots, \pi_m^*), (p_0^*, 0, \cdots, 0) \rangle = p_0^*$, hence $p_0^* = \mu^*$; but $p_0 \geqq \mu$ for all feasible solutions of (7) and (8) by Proposition 8. In particular, $p_0^* \geqq \sup \mu$. Since $\mu^* = p_0^*$, (π^*, μ^*) is optimal. Conversely, inf $p_0 \geqq \mu^*$ and since $p_0^* = \pi^*$, p^* is optimal. On the other hand, if p^* and π^*, μ^* are optimal, i.e., they achieve the infimum and supremum in (7) and (8), respectively, then by Theorem 1 they must satisfy $\pi^* p^* = \mu^*$.

COROLLARY 2 (*Pre-maximum principle*). *If p^* is optimal for* (7), *there exists a π^* such that*

$$\pi^* p^* = \min \{\pi^* p \mid p \in C\}.$$

Proof. Clearly p^* is a boundary point of C. Then there is a supporting hyperplane $\pi^* p - \mu \geqq 0$ for all $p \in C$ with $\pi^* p^* - \mu = 0$.

Remark. If in Corollary 2, $\pi_0^* > 0$, then π^*/π_0^*, p_0^* determine an optimal solution to (7); however, this need not be the case (see Fig. 2).

COROLLARY 3. *If C is closed and the infimum in* (7) *exists and is finite, then the infimum is attained for a feasible solution.*

COROLLARY 4. *If L intersects the relative interior of C and* (8) *is feasible, the supremum in* (8) *is attained.*

Proof. See [6].

3b. Special cases. We now apply the results of the last section to the

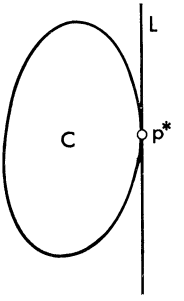

FIG. 2

original problem (6) and examine some special cases. We first interpret the dual problem (8). For the special case (see Fig. 1), it is equivalent to:

$$\begin{aligned} &\text{find} \quad \sup \mu \\ &\text{such that} \quad c(u) - \pi[A(u) - b] \geqq \mu \quad \text{for all} \quad u \in U. \end{aligned} \tag{9}$$

An easy lemma is:

LEMMA 2. *If* $c(u) = c \cdot u$ *is linear and* U *is a cone, then for any feasible* π, μ *for* (9), *we have*

$$[c - \pi A](u) \geqq 0 \quad \textit{for all} \quad u \in U.$$

(a) *Application to linear programs.* Consider the linear program:

$$\min cx \quad \text{such that} \quad Ax = b, \qquad x \geqq 0.$$

By direct application of (9), its "dual" is:

$$\begin{aligned} &\text{find} \quad \sup \mu \\ &\text{such that} \quad cx - \pi[Ax - b] \geqq \mu \quad \text{for all} \quad x \geqq 0. \end{aligned} \tag{10}$$

Since the set of nonnegative x is a cone, Lemma 2 applies and we have $[c - \pi A]x \geqq 0$ for all $x \geqq 0$. Further, by taking $x = I_i$, $i = 1, \cdots, m$, where I_i is the ith unit vector, we obtain

$$c - \pi A \geqq 0 \quad \text{for any feasible} \quad \pi.$$

Rearranging (10) we obtain $(c - \pi A)x + \pi b \geqq \mu$. Clearly for a given π, the largest μ is given by

$$\mu - \pi b = \inf_{x \geqq 0} (c - \pi A)x = 0.$$

Hence, the dual problem becomes

$$\begin{aligned} &\sup \pi b \\ &\text{such that} \quad c - \pi A \geqq 0. \end{aligned} \tag{11}$$

To obtain the usual duality theorem for linear programming from Theorem 1 it suffices to observe that if $\sup \pi b < \infty$, then it is attained by some feasible π, and similarly for the primal objective.

(b) *Application to linear control problem.* Consider a dynamical system, the evolution of which is described by the ordinary linear differential equations

$$\frac{dx(t)}{dt} = A(t)x(t) + u(t)$$

on the time interval $[0, T]$, where $x(t) = (x_0(t), \cdots, x_n(t))$, $A(t)$ is

an $(n+1) \times (n+1)$ matrix of continuous functions and $u(t) = (u_0(t), \cdots, u_n(t))$ is a vector of controls. For simplicity, we assume that $u \in U = \{u \mid u(t) \in \Omega,\ 0 \leqq t \leqq T$, and u is measurable and bounded$\}$, where Ω is a closed convex subset of $\mathfrak{R}^{n+1}$. Further we assume that

$$x_i(0) = x_i^0, \quad i = 0, \cdots, n,$$

and

$$x_i(T) = x_i^T, \quad i = 1, \cdots, n.$$

The value of $x_0(T)$ is not prescribed and the problem is to minimize $x_0(T)$ over all $x(t)$ and $u(t)$ satisfying the above relations.

As is well known,

$$x(T) = Y(T)x^0 + \int_0^1 Y(T)Y^{-1}(s)u(s)\, ds,$$

where $Y(T)$ is an $(n+1) \times (n+1)$ matrix of functions satisfying the adjoint equation:

$$Y(t) = -Y(t)A(t), \quad Y(0) = I.$$

It is easily seen that

$$S_T = \left\{x(T) \mid x(T) = Y(T)x(0) + \int_0^T Y(T)Y^{-1}(s)u(s)\, ds, \quad u \in U\right\}$$

is convex. Thus, the duality theory previously developed can be used.

If $x^*(t), u^*(t)$ is an optimal solution to the control problem, we can apply Corollary 2 yielding the existence of $\pi = (\pi_0, \cdots, \pi_{m+1})$ such that

$$\pi x^*(t) = \min \{\pi x \mid x \in S_T\}.$$

Using the particular form of the description of the set S_T, we have

$$\pi\left[Y(t)x^0 + \int_0^T Y(T)Y^{-1}(s)u^*(s)\, ds\right] \leqq \pi Y(T)x^0 + \int_0^T \pi Y(T)Y^{-1}(s)u(s)\, ds$$

for all $u \in \Omega$. Thus

$$0 \leqq \int \pi Y(T)Y^{-1}(s)[u(s) - u^*(s)]\, ds.$$

If we define $\Pi(t) = \pi Y(T)Y^{-1}(s)$, it is easily seen [6] that $u^*(s) = \min_{u \in \Omega} \Pi(s)u$ and $\Pi(t)$ is a vector solution of the adjoint equation, which is equivalent to the maximum principle for this problem.

REFERENCES

[1] A. BALAKRISHNAN, *Optimal control problems in Banach spaces*, this Journal, 3(1965), pp. 152–180.

[2] G. B. DANTZIG, *Linear programming under uncertainty*, Management Sci., 1(1955), pp. 197–206.

[3] G. B. DANTZIG AND A. MADANSKY, *On the solution of two-stage linear programs under uncertainty*, Proceedings of the Fourth Symposium on Mathematical Statistics and Probability, vol. 1, University of California, Berkeley, 1961.

[4] H. KUSHNER, *On the stochastic maximum principle: Fixed time of control*, RIAS Report, Baltimore, 1963.

[5] A. MADANSKY, *Dual variables in two-stage linear programming under uncertainty*, J. Math. Anal. Appl., 6(1963), pp. 98–108.

[6] R. VAN SLYKE, *Mathematical programming and optimal control*, Ph. D. thesis, University of California, Berkeley, 1965.

[7] R. WETS, *Programming under uncertainty: The complete problem*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, to appear.

[8] ——, *Programming under uncertainty: The equivalent convex problem*, J. Soc. Indust. Appl. Math., 14 (1966), to appear.

[9] A. WILLIAMS, *A stochastic transportation problem*, Operations Res., 11(1963), pp. 759–770.

[10] ——, *On stochastic linear programming*, J. Soc. Indust. Appl. Math., 13(1965), pp. 927–940.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

NONLINEAR PROGRAMMING: A NUMERICAL SURVEY*

G. ZOUTENDIJK†

1. Introduction. In this paper some of the existing general nonlinear programming methods will be reviewed with special emphasis on their numerical aspects. A few modifications will be suggested. In addition two new methods will be described which to a certain extent combine the advantages of the other methods.

The problem we will consider is to find a (local) maximum, if it exists, of a function $f(x)$ of the vector $x \in E_n$ on a closed and connected set R in n-space:

$$\max \{f(x) \mid x \in R\}. \tag{1}$$

In most applications R will be of the form $R = \{x \mid f_i(x) \leqq 0, i \in I\}$, I being a finite set.[1] Some methods can also be applied to problems with an infinite number of constraints.

The functions $f(x)$ and $f_i(x)$ will be assumed to be differentiable with continuous partial derivatives.

The question which method is best cannot be answered in this general form. To a considerable extent this will depend on the structure of the problem, e.g., on the degree of nonlinearity. Moreover what is best? The method that is fastest in a series of tests? The method that can solve larger problems? The method that is most accurate? The method that has the simplest computer program? There obviously is no definite answer. If there are many constraining inequalities and if most of them are linear, then any method that is a direct extension of the simplex method for linear programming [5], [10] is to be preferred. If, in addition, the variables are separable, Miller's simplex method for local separable programming [17] is the obvious one to apply. Since, however, this method imposes a severe restriction on the form of the nonlinearities, it cannot be considered as a general nonlinear programming method and will therefore not be discussed in this paper.

There are two other types of methods to which we will pay no attention in this paper, viz., the combinatorial methods and the decomposition/partitioning methods. For instance, by a systematic and sophisticated

* Received by the editors June 17, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Centraal Reken-Instituut der Rijksuniversiteit te Leiden, Leiden, The Netherlands.

[1] R also has to satisfy a regularity condition to exclude "cusps". See [16].

search through all possible combinations of equalities out of the (finite) set of inequalities we would reduce the programming problem to a finite number of classical Lagrange type maximization problems. This, together with a numerical technique for solving these classical problems, would form a general nonlinear programming method. The combinatorial nature of such a method would restrict its applicability to problems with relatively few constraints, however. The decomposition and partitioning methods [2], [19], [20], [23] will probably be of considerable practical importance. Since they are still under development and since they reduce the original problem to a sequence of smaller and simpler nonlinear programming problems to the solution of which a general method has to be applied, we will not take them into account in this paper.

In §2 some existing methods will be reviewed. In §3 two new methods will be described which we will call the MIP and the MFD method (for modified interior point and modified feasible directions). These new methods are thought to have some important numerical advantages over other methods.

2. Review of some existing general nonlinear programming methods.

2.1. Dual methods. In this category we find the cutting plane method developed by Cheney and Goldstein [3] and independently by Kelley [15]. A slightly different version was suggested by Hartley and Hocking [13]. The cutting plane method is as follows.

(a) Add the relation $-f(x) + x_0 \leqq 0$ to the constraint set and maximize the value of the new variable x_0 instead of the function $f(x)$, which is obviously equivalent to maximizing $f(x)$. Writing $f_0(x)$ for $f(x)$ and adding the index 0 to I the problem is still of the form (1) but now with a (simple) linear objective function. The algorithm works for a general linear objective function $p^T x$, however.

(b) Start with some initial point x^0 (x^0 need not be feasible) and solve the "linearized" problem

$$\max \{p^T x \mid \nabla f_i(x^0)^T x \leqq \nabla f_i(x^0)^T x^0 - f(x^0),\ i \in I\}, \tag{2}$$

where T denotes the transpose and ∇f_i denotes the gradient vector of the function f_i. A possible infinite solution can be prevented by adding constraints of the type $| x_j | \leqq \alpha$, α being a number chosen so large that no x_j will equal $\pm\alpha$ in the optimum solution of (1). Let x^1 be the solution of (2).

(c) Suppose $x^0, x^1, \cdots, x^k$ have already been calculated. If for every i, $f_i(x^k) \leqq \epsilon$, where ϵ is a predetermined small positive number, we will stop; otherwise we will choose i_k in such a way that $f_{i_k}(x^k) = \max_i f_i(x^k)$ (greatest infeasibility), relinearize the i_kth constraint with respect to x^k and add the new linear relation

$$\nabla f_{i_k}(x^k)^T x \leqq \nabla f_{i_k}(x^k)^T x^k - f_{i_k}(x^k)$$

to our linear subproblem.

(d) Solve the linear subproblem again (starting from the old solution and using the dual simplex method). This will give a new point x^{k+1}. Repeat (c) and (d), if necessary.

A convergence proof can easily be given if $-f$ and all f_i are convex. This is based on the following observations: the feasible region of the linearized problem (2) contains R; each new linearization cuts off the old optimum but never part of R; the x^k form a sequence with nonincreasing values for $f(x)$ exceeding the maximum of $f(x)$ on R; and, with the choice of the next constraint to be relinearized as in rule (c), any point of accumulation of the sequence x^k will belong to R and will therefore be a maximum of $f(x)$ on R. The proof makes an essential use of convexity properties. Indeed, the method will not work in nonconvex problems. This is a serious drawback since most practical problems are not convex, while a test on convexity is a time consuming numerical procedure. The dual methods are thus of limited practical value and have only been described to simplify the discussion of other methods. In relation to these dual methods several questions of a computational nature can be discussed.

Is it worthwhile to start with a complete linearization (2) or would it be better to start with linear constraints and upper bounds only and add linearizations gradually?

Would it not be better to go cyclically through the nonlinear constraints accepting for relinearization the first one with $f_i(x^k) > \epsilon_k$, where $\epsilon_k \geqq \epsilon$ is gradually reduced? This would increase the number of steps but reduce the amount of work per step.

Is it advisable to solve each linear subproblem completely or could we stop after a few or even one iteration in the subproblem? (Making only one iteration would lead to the method due to Hocking and Hartley.) Probably it is best to solve the linear problems completely.

Is it necessary or recommendable to retain all old linearizations, even those which are no longer active in the linear subproblem?

Different answers to these questions will lead to different computational variants. For all these variants the following advantages (A) and disadvantages (D) will probably hold (compared to other methods).

A. 1. Direct extension of the simplex method (therefore efficient for convex programs which are nearly linear).
   2. Relatively little work per step.
   3. Simple computer program.
   4. Some problems with an infinite number of constraints can be solved (e.g., the Chebyshev approximation problem).

D. 1. Methods cannot be applied to nonconvex problems.

2. Intermediate solutions are not feasible.
3. Rather slow convergence, especially if the maximum is not in a vertex.
4. Linear subproblems will consist of near-dependent constraints which could lead to serious rounding-off errors, especially if the maximum is not in a vertex.
5. Rather inefficient for problems with linear constraints and a nonlinear, possibly quadratic, objective function.[2]

2.2. Small step gradient methods. The only method in this class which we will discuss, is the so called method of approximation programming (MAP) developed by Griffith and Stewart [11] which has been successfully applied to many nonlinear programming problems. The method has some relation to the cutting plane method and proceeds according to the following rules.

(a) Start with some[3] feasible initial point x^0.

(b) Suppose $x^0, x^1, \cdots, x^k$ have already been calculated. Linearize all constraints with respect to x^k and solve the linear subproblem

$$\max \{\nabla f(x^k)^T x \mid \nabla f_i(x^k)^T x \leqq \nabla f_i(x^k)^T x^k - f(x^k),\ i \in I; \quad | x_j - x_j^k | \leqq \delta_k \quad \text{for all} \quad j\}, \tag{3}$$

where $\delta_k > 0$ is a small number which prevents large steps.

(c) Repeat (b) with gradually decreasing "stepsize" δ_k until the improvement in value becomes sufficiently small and the infeasibilities in x^k are acceptable.

In MAP a complete relinearization of the problem takes place at each step. Hence, no old information is retained. This, together with a predetermined small stepsize, makes it possible to apply MAP to nonconvex

[2] Wolfe [22] describes an accelerating device for problems with linear constraints resulting in a finite quadratic programming method. The computational value of his suggestion is unknown but is admitted of doubt.

[3] A simple trick can be applied if such a point is not available. We replace the original problem (1) by the equivalent problem

$$\max \{f(x) - \mu\xi \mid f_i(x) - \rho_i\xi \leqq 0, \quad \rho_i \geqq 0,\ i \in I\},$$

where μ is a large number, ξ is an additional variable and $\rho_i = 1$ if for the (infeasible) estimate x^0, $f_i(x^0) \geqq 0$ will hold, $\rho_i = 0$ if $f_i(x^0) < 0$. It is clear that ξ^0 can be chosen such that x^0, ξ^0 is feasible in the modified problem. It can be proved [24, Theorem 2, p. 67] that for μ sufficiently large the modified problem will have the same solution as the original one. Note that the modified problem will always have an interior point. The same trick can be applied if there are nonlinear equality constraints. Suppose $f_i(x) = 0$ has to hold and $f_i(x^0) > 0$. Then replace the equation by the two inequalities $f_i(x) - \xi \leqq 0$ and $-f_i(x) \leqq 0$. We must be careful, however: a local minimum of the original problem may be a local maximum of the modified problem!

problems without any modification. No proof of convergence has ever been published. It is likely, however, that the choice of the δ_k can be formalized in such a way that convergence can be proved, at least for convex programming problems. We will obtain a sequence of near-feasible points. Due to the many small steps needed and the complete relinearization at each step, MAP will not be very efficient if no prior knowledge of the problem is available. In many practical problems, however, the present way of doing things is a very good starting point, which by applying MAP will be improved.

The advantages and disadvantages of MAP are listed as follows:

A. 1. Works for nonconvex problems.
   2. Rather simple computer program.
   3. More accurate results can be expected.
   4. Simplex method used as subroutine.
   5. Intermediate solutions are (near) feasible.

D. 1. No formal way to determine the stepsize (this probably is only a matter of mathematical elegance since there have been no difficulties in a large number of practical problems).
   2. Many small steps, hence slow convergence, especially if the starting point is arbitrary.
   3. Complete relinearization, hence more work, at each step (the linear subproblem will have to start with a preinversion while in the dual methods the existing dual inverse can immediately be used to price out the added (dual) column).
   4. Inefficient for quadratic, for nearly linear programs and for unconstrained problems.
   5. No extension to problems with an infinite number of constraints.

2.3. Large step gradient methods, also called methods of feasible directions. A great number of methods, described in [24], belong to this class. Well-known is Rosen's gradient projection method [18] which has successfully been applied to many problems with a nonlinear objective function and linear constraints. Another method in this class is the one described in a paper by Frank and Wolfe [9], which has been reinvented recently as a method of solution for certain control optimization problems.

Any method of feasible directions proceeds according to the following rules.

(a) Start with some feasible initial point x^0 (no restriction, see the remark in §2.2). Suppose $x^0, x^1, \cdots, x^{k-1} \in R$ have already been calculated.

(b) Then determine a usable feasible direction in x^{k-1}, i.e., a direction s^{k-1} with the property that a $\bar{\lambda} > 0$ exists such that for all λ, $0 < \lambda \leqq \bar{\lambda}$, $x^{k-1} + \lambda s^{k-1} \in R$ and $f(x^{k-1} + \lambda s^{k-1}) > f(x^{k-1})$ hold.

(c) Determine the steplength λ_{k-1} by solving the one-dimensional maximum problem in λ,

$$\max \{f(x^{k-1} + \lambda s^{k-1}) \mid x^{k-1} + \lambda s^{k-1} \in R\}.$$

The direction finding problems are easy to formulate in the case of linear constraints of the form $a_i{}^T x \leqq b_i$. If the present solution is $\bar{x}$, then require:
(a) $a_i{}^T s \leqq 0$ for $i \in I(\bar{x}) = \{i \mid a_i{}^T \bar{x} = b_i\}$,
(b) a normalization like $s^T s \leqq 1$, $-1 \leqq s_j \leqq 1$ for all j, $\sum |s_j| \leqq 1$, etc.,
(c) $\nabla f(\bar{x})^T s$ to be maximized.
Different normalizations will lead to different computational procedures. A linear normalization will lead to a sequence of linear subproblems to which a variant of the simplex method can be applied.

In the case of a nonlinear constraint for which $f_i(\bar{x}) = 0$ we must require $\nabla f_i(\bar{x})^T s < 0$ instead of $\leqq 0$. This can be satisfied by adding an additional variable σ and by requiring
(a) $\nabla f_i(\bar{x})^T s + \theta_i \sigma \leqq 0$ if $i \in I(\bar{x})$, where $\theta_i = 0$ if f_i is linear, $\theta_i > 0$ if f_i is nonlinear;
(b) $-\nabla f(\bar{x})^T s + \sigma \leqq 0$;
(c) normalization;
(d) σ to be maximized.

To avoid so-called zigzagging, a common feature in all gradient methods, and to guarantee or speed up convergence, additional requirements can be added to the direction finding problems. They have been described in [24] and result in a number of finite methods for the quadratic programming problem which have wrongly received considerably less attention than the methods due to Beale [1] and Wolfe [21].

Several questions of a computational nature may arise in relation to the methods of feasible directions:

Which is the "best" normalization to use?

What is the most efficient way to determine the steplength?

When to add and when to drop antizigzagging requirements?

Which "pushing off" factors θ_i to use?

These questions can only be answered computationally. The considerable amount of computer programming needed hereto will probably be justified by the ultimate result: an efficient nonlinear programming method.

These methods have the following advantages and disadvantages:

A. 1. Applicable to nonconvex problems.
2. Faster convergence, especially if the pushing off factor is properly chosen and antizigzagging precautions are taken.
3. Reduce to an efficient linear programming method in the linear case (only if a linear normalization is used).
4. Finite for quadratic programming problems.

5. Intermediate solutions feasible.
6. More accurate results can be expected in the case of the maximum not being at a vertex.
7. Extension to certain problems with an infinite number of constraints possible.

D. 1. Determination of the steplength needed, resulting in more work per step.
2. Complicated computer program.
3. Upper bound for further increase in value in convex programs cannot easily be obtained (except in what has been called procedure P2 in [24]).

2.4. Interior point methods. In this class of methods the principle is to keep away from the boundary of the feasible region. A sequence of interior points is constructed converging to a maximum of $f(x)$ on R. There are two variants.

(a) *Huard's method* [14]. This essentially is a method to find an interior point. Start with $x^0 \in R_0$ (the interior of R; if R_0 is empty the problem can be modified, see footnote 3). Find $x^1 \in R_0 \cap \{x \mid f(x) > f(x^0)\} = R_1$. Next find $x^2 \in R_1 \cap \{x \mid f(x) > f(x^1)\} = R_2$, etc. The method to find an interior point can be devised in such a way that the sequence of points x^k converges to a maximum of $f(x)$ on R.

(b) *The SUMT method developed by Fiacco and McCormick (sequential unconstrained minimization technique,* see [6], [7] and [8]). In this method the nonlinear programming problem (1) is solved through a sequence of unconstrained maximization problems of the form:

$$\text{(4)} \qquad \text{maximize } f(x) - \rho g\{f_1(x), \cdots, f_m(x)\}, \qquad \rho > 0 \text{ fixed},$$

where $g(y_1, \cdots, y_m)$ is defined and bounded below for all $y_i < 0$ and

$$\lim_{y_i \uparrow 0} g(y_1, \cdots, y_m) = +\infty \qquad \text{for all } i.$$

Fiacco and McCormick have taken

$$g(y_1, \cdots, y_m) = -\sum_{i=1}^{m} \frac{1}{y_i} = -\sum_{i=1}^{m} \frac{1}{f_i(x)},$$

but other choices are possible, e.g.,

$$g(y_1, \cdots, y_m) = -\sum_{i=1}^{m} \log(-y_i).$$

Starting with some interior point $x^0 \in R_0$ they solve (4) for $\rho = \rho_0 > 0$, leading to $x^1 \in R_0$. This is repeated for $\rho = \rho_1 < \rho_0$ and further values of ρ: $\rho_k < \rho_{k-1}$, $\rho_k \to 0$. It is supposed that an efficient algorithm for uncon-

strained maximization is available which uses the maximum of the previous step as a starting point for the next maximization. (See [4] for this.) Newton's method is chosen, so that the matrix of second partial derivatives of the function (4) has to be calculated. To speed up convergence an extrapolation device is included in the method.

Extension of the method to problems containing nonlinear equations is possible. If, in addition to the inequalities $f_i(x) \leqq 0$, we have a number of equations $g_i(x) = 0$, $i = 1, \cdots, p$, then the sequence of unconstrained problems to be solved is of the form

$$(5) \qquad \text{maximize } f(x) + \rho \sum_{i=1}^{m} \frac{1}{f_i(x)} - \rho^{-1/2} \sum_{i=1}^{p} \{g_i(x)\}^2, \qquad \rho > 0.$$

Again starting with $x^0 \in R_0 = \{x \mid f_i(x) < 0\}$, a sequence of problems of type (5) is to be solved for $\rho = \rho_1, \rho_2, \cdots$ with $\rho_k < \rho_{k-1}$ and $\rho_k \to 0$ (see [8, pp. 1–3]).

It is also possible to avoid using a parameter ρ, at least if there are no nonlinear equations. The procedure (see [8, p. 5]) is then as follows.

(a) Start with $x^0 \in R_0$. Suppose $x^0, x^1, \cdots, x^{k-1} \in R_0$ have already been determined.

(b) Then

$$\text{maximize } g(x, x^{k-1}) = \frac{1}{-f(x) + f(x^{k-1})} + \sum_{i=1}^{m} \frac{1}{f_i(x)}$$

starting with a point $x = x^{k-1} + \lambda \nabla f(x^{k-1})$, $\lambda > 0$ so small that

$$x \in R_0 \cap \{x \mid f(x) > f(x^{k-1})\}.$$

This results in x^k.

(c) Repeat (b) for x^k, etc. This procedure obviously belongs to class (a), described by Huard [14].

Several questions of a computational nature may be studied in relation to the interior point methods.

What is the best choice for the function g in (4)?

What is the computationally most efficient way of solving the unconstrained problems?

Is it necessary to solve the unconstrained problems completely or can we stop after using a gradient method for a number of steps or maybe for one step?

What is the best way of decreasing ρ? Here we have to compromise between speeding up convergence and increasing accuracy.

This method has the following advantages and disadvantages.

A. 1. Applicable to nonconvex problems, including those with nonlinear constraints.

2. Very efficient for unconstrained problems as well as for problems with a few highly nonlinear constraints.
3. Good convergence can be expected if the ρ_k are well chosen and an extrapolation device is used.
4. Intermediate solutions feasible.
5. Relatively simple computer program.

D. 1. A special structure of the constraints (linearity or near linearity, partial nonlinearity, etc.) is destroyed; even constraints like $x_j \geqq 0$ are not dealt with in a special simple way.
2. Not finite for linear or quadratic programs.
3. Much work per step (solution of a complete unconstrained problem, whereas the other methods require a few additional simplex iterations in the linear subproblems).
4. Rounding-off problems may sometimes arise ($\rho \to 0$ and simultaneously $g \to \infty$). Practical experiments are very promising in this respect, however.
5. No upper bound for the value of the objective function available.
6. Problems with an infinite number of constraints cannot be solved with the methods in their present form.[4]

A summary of the properties of the methods discussed can be found in the appendix.

3. New methods. Of the four methods considered in §2, the first two (cutting plane and MAP) are direct extensions of the simplex method; SUMT has no relation with the simplex method but works with a sequence of unconstrained maximization problems, while most of the methods of feasible directions, though working with linear subproblems, also make use of techniques originally developed for unconstrained problems (step-length determination, conjugated directions). The SUMT method is

[4] The obvious extension of the SUMT method to (some) problems with an infinite number of constraints, $\max \{f(x) \mid f(x, t) \leqq 0,\ t \in T\}$, is to consider $\lim_{\rho \to 0} \max_x g(x, \rho)$, where

$$g(x, \rho) = f(x) + \rho \int_T \frac{dt}{f(x, t)}$$

or

$$g(x, \rho) = f(x) + \rho \int_T \log \mid f(x, t) \mid dt.$$

It is not known to the author under which conditions convergence can be proved for this method. In particular, it is not known whether the Chebyshev approximation problem can be solved in this way.

obviously not suited to linear or nearly linear problems. It can, however, easily be modified in such a way that the linear constraints are not included in the function to be maximized.

Suppose the problem is of the form

$$\max \{f(x) \mid f_i(x) \leqq 0, i \in I_1, \quad l_i(x) \leqq 0, i \in I_2\}, \tag{6}$$

where I_1 and I_2 are finite sets of indices and the $l_i(x)$ are linear constraints (including those of the form $x_j \geqq 0$, if any). The modified method, already described by Fiacco and McCormick in [8], will then work through a sequence of linearly constrained maximization problems of the form

$$\max \left\{ f(x) + \rho \sum_{i \in I_1} \frac{1}{f_i(x)} \;\middle|\; l_i(x) \leqq 0, i \in I_2 \right\}. \tag{7}$$

These problems having a highly nonlinear objective function and linear constraints can then be solved using a method of feasible directions. Fiacco and McCormick have used Rosen's gradient projection method [18], for which an efficient computer program is available. For a 100 variable problem they report a decrease in computer time from 13 to 6 minutes by treating the nonnegativity requirements separately. A further decrease may be expected if the other linear constraints are also treated separately and if a method of feasible directions is selected using a linear normalization.

It may be worthwhile to make only one or perhaps a few steps in the linearly constrained subproblem. A typical step in the corresponding modified interior point method would then be:

(a) calculate

$$\nabla g^k = \nabla g(x^k, \rho_k) = \nabla f(x^k) - \rho_k \sum_{i \in I_1} \frac{\nabla f_i(x^k)}{\{f_i(x^k)\}^2};$$

(b) solve the direction finding problem

$$\max \{(\nabla g^k)^T s \mid a_i{}^T s \leqq 0 \text{ if } i \in I_2(x^k), \text{ normalization}\},$$

where $l_i(x) = a_i{}^T x - b_i$; this will give the vector s^k;

(c) determine $x^{k+1} = x^k + \lambda_k s^k$, where λ_k is given by

$$g(x^k + \lambda_k s^k, \rho_k) = \max_{\lambda \geqq 0} \{g(x^k + \lambda s^k, \rho_k) \mid l_i(x^k + \lambda s^k) \leqq 0, i \in I_2\};$$

(d) determine ρ_{k+1} and repeat.

Thus far, convergence has not been proved for this procedure. Numerical details concerning the best way to decrease ρ, the necessity to use antizigzagging precautions, etc., should be investigated. The so-obtained modified interior point method (MIP) has all the advantages A.1–5 mentioned in

§2.4 but not the disadvantages D.1–3. Therefore, it looks very promising, especially for problems with many nonlinear constraints.

Another new method, to be called the MFD method (modified feasible directions), makes extensive use of the linearization technique as applied in the cutting plane method, MAP, and a particular method of feasible directions (procedure P2 in [24]). At each step a feasible direction and a steplength are determined; but in addition a sequence of feasible points interior with respect to the nonlinear constraints is obtained which converges to a (local) maximum of $f(x)$ on R, while for convex programs another sequence of nonfeasible points converges to the same maximum and gives an upper bound for the value of the objective function.

We shall first describe the method for convex programs with a linear objective function and then mention the modifications for more general nonlinear programs. The problem is supposed to be of the form (6).

1. *$f(x)$ linear, all f_i convex.* The procedure is as follows.

(a) Start with $x^0 \in R' = \{x \mid f_i(x) < 0,\ i \in I_1,\ l_i(x) \leqq 0,\ i \in I_2\}$ (see footnote 3 if such a point is not immediately available or does not exist). Start with the linear subproblem

$$L_0 = \max \{f(x) \mid l_i(x) \leqq 0, i \in I_2, \mid x_j \mid \leqq \alpha\},$$

with α chosen sufficiently large. Suppose we have already determined:

a sequence $x^h \in R'$, $h = 0, 1, \cdots, k$ ("interior" feasible points),
a sequence $y^h \in E_n - R$, $h = 0, 1, \cdots, k-1$ (infeasible points),
a sequence $z^h \in R - R_0$, $h = 0, 1, \cdots, k-1$ (boundary points),
a linear subproblem L_k.

We then perform the following calculations.

(b) Solve the linear subproblem L_k giving the solution $y^k \in E_n - R$.

(c) Determine z^k by $z^k = x^k + \lambda_k(y^k - x^k)$, where $\lambda_k = \max \{\lambda \mid x^k + \lambda(y^k - x^k) \in R\}$, $z^k \in R - R_0$. Let $f_i(z^k) = 0$ for $i \in I_1{}^k \subset I_1$.

(d) Linearize for $i \in I_1{}^k$ the constraints $f_i(x) \leqq 0$ with respect to z^k giving the linearized constraints (tangent planes)

$$\nabla f_i(z^k)^T x \leqq \nabla f_i(z^k)^T z^k, \qquad i \in I_1{}^k.$$

Add these relations to the constraints of L_k giving the linear subproblem L_{k+1}.

(e) Calculate $x^{k+1} = \tau x^k + (1 - \tau)z^k$, $0 < \tau < 1$, e.g., $\tau = \frac{1}{2}$, $x^{k+1} \in R'$. Repeat (b)–(e) for $k + 1$ ($k = k + 1$).

We so obtain:

a sequence of feasible points x^k with increasing values, $f(x^{k+1}) > f(x^k)$, converging to a maximum of $f(x)$ on R;

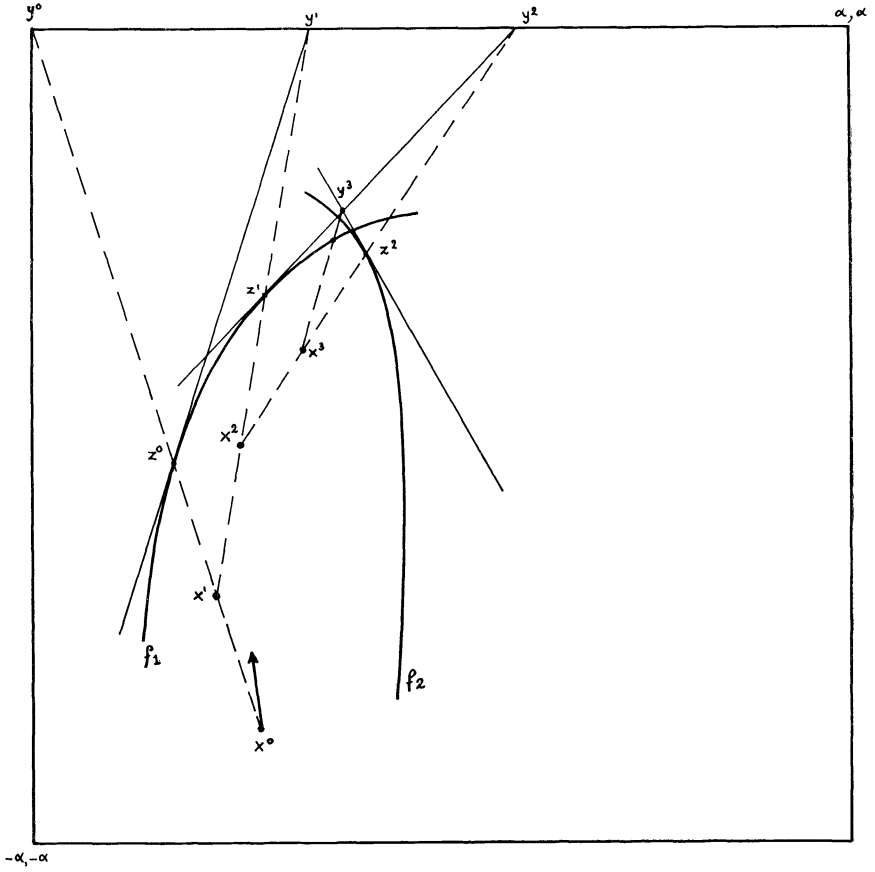

FIG. 1. *Linear objective function, two nonlinear convex constraints,* $\tau = \frac{1}{2}$

a sequence of infeasible points y^k with nonincreasing values, $f(y^{k+1}) \leqq f(y^k)$, converging to a maximum of $f(x)$ on R and at each step giving an upper bound for the maximum value;

a sequence of feasible points z^k giving at each step a lower bound for the maximum value;

an upper bound for the possible further increase in value, $f(y^k) - f(z^k)$.

If this expression is less than some predetermined ϵ we will stop. The convergence proof is equivalent to the one given for procedure P2 in [24, p. 78]. A geometrical illustration is given in Fig. 1 and Fig. 2.

2. $f(x)$ *nonlinear, all* f_i *convex.* If f is concave we could add $-f(x) + x_0 \leqq 0$ to the constraints and maximize x_0. Probably it is better for concave as well as more general $f(x)$ to perform the steplength procedure

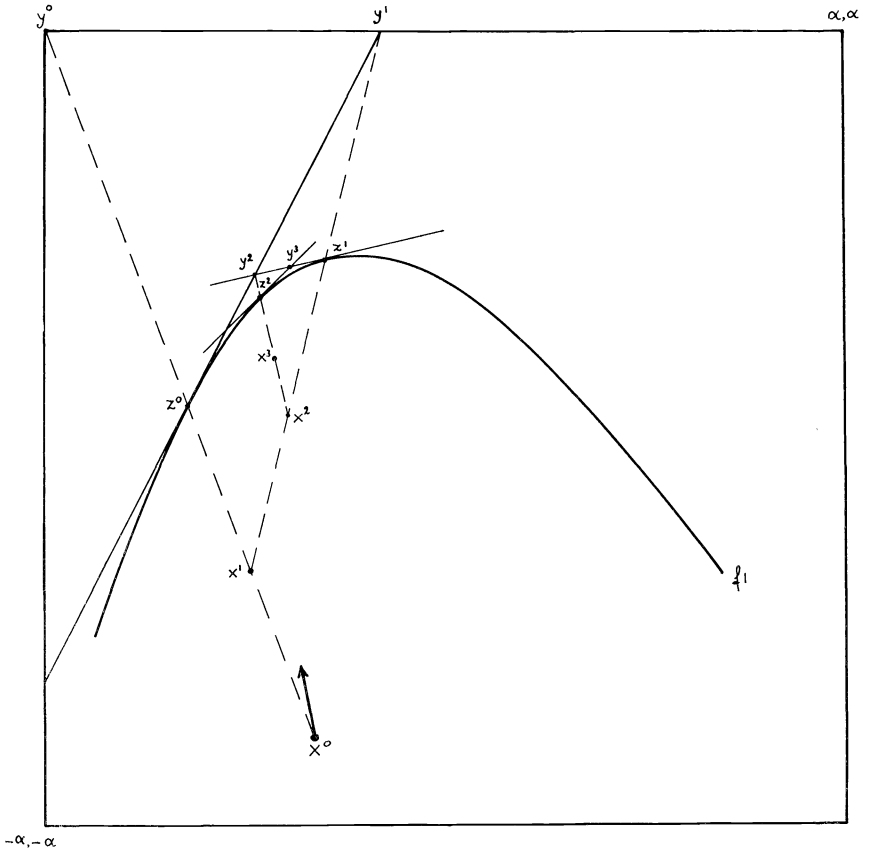

FIG. 2. *Linear objective function, one nonlinear convex constraint,* $\tau = \frac{1}{2}$

as in the methods of feasible directions, so that an interior maximum is possible on the line connecting x^k to y^k. This point, if feasible, will be taken as x^{k+1} and (d) and (e) need not be performed. The linear function to be maximized in the subproblem will then be $\nabla f(x^k)^T x$.

As in the original methods of feasible directions convergence can be speeded up by using the principle of conjugated directions, i.e., by temporarily adding a relation of the form

$$\{\nabla f(x^{k+1}) - \nabla f(x^k)\}^T (x - x^{k+1}) = 0,$$

when x^{k+1} has been determined as an interior maximum of $f(x)$ on the line connecting x^k to y^k.

3. *Nonconvex problems of type* (1). The only additional rule we have to

obey is that, if for some value of i, $f_i(y^k) < 0$, then any active linearization of $f_i(x)$ in the linear subproblem (i.e., a constraint in the subproblem which determines the solution y^k) should be taken out in the next subproblem. By applying this rule no part of the feasible region will ever be definitely cut off. The test corresponding to this rule need not be carried out at each step for all i provided each value of i is reconsidered from time to time.

A geometrical illustration of a nonconvex problem, where the above mentioned rule is applied at each step, is given in Fig. 3.

Nonlinear equations can be handled with the MFD method by using the trick mentioned in footnote 3. The MFD method can be extended to problems with an infinite number of constraints such as the Chebyshev approximation problem. A requirement is that for all $x \in R = \{x \mid f(x, t)$

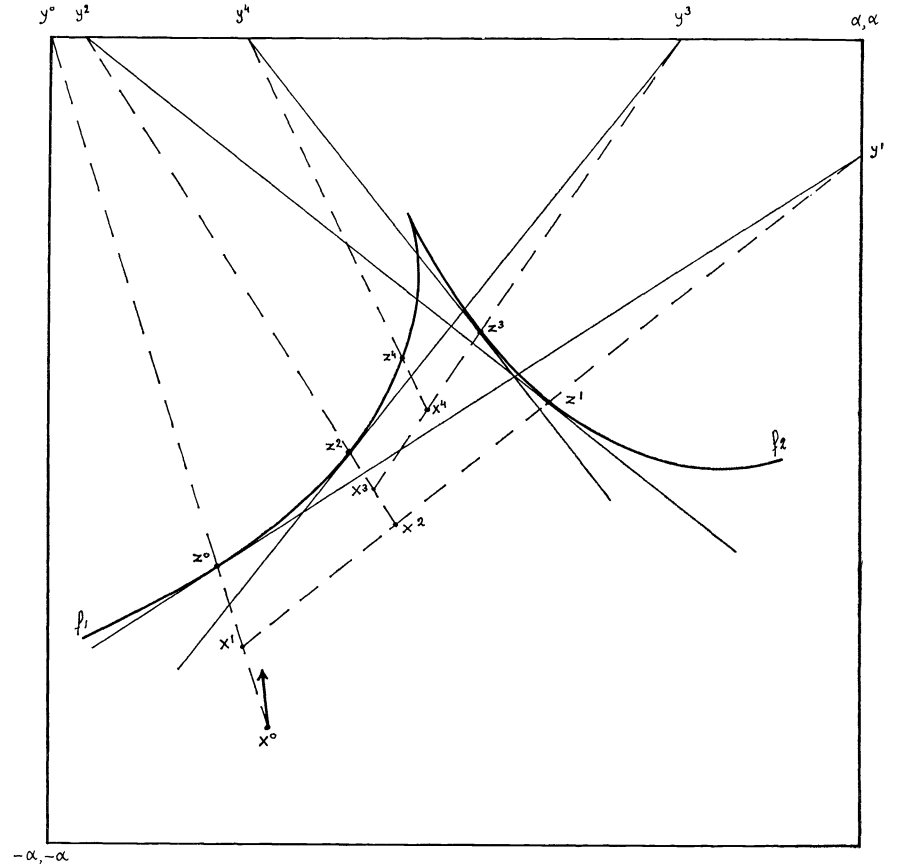

FIG. 3. *Linear objective function, two nonlinear nonconvex constraints,* $\tau = \frac{1}{2}$

$\leqq 0, t \in T\}$, we have that $\{t \in T \mid f(x, t) = 0\}$ is finite. The determination of z^k will involve the solution of problems of the type $\min \{\lambda(t) \mid t \in T\}$, where

$$\lambda(t) = \max_{\lambda} \{f(x + \lambda(y - x), t) \leqq 0\}.$$

The minimization problem is a nonlinear programming problem itself. In the Chebyshev approximation problem $\lambda(t)$ can be explicitly determined.

A number of questions of a computational nature will need further investigation.

Is it worthwhile to remove nonactive linearizations from the linear subproblems and, if so, how often should this be done?

What is the best value of τ? Should it be changed or fixed once and for all?[5]

When should the additional linear relations based on the principle of conjugated directions be dropped from the subproblem?

How often should we carry out the test mentioned under point 3 for the nonconvex problem?

What is the best policy with respect to the μ, mentioned in footnote 3, which has to be introduced if no x^0 is immediately available; for instance, if there are nonlinear constraints? Should we start with a large value of μ or should we gradually increase μ?

It is to be expected that the MIP and the MFD methods will be the best methods for the general nonlinear programming problem. Which of the two methods is to be preferred will probably depend on the nature of the problem, particularly on the amount of nonlinearity and the activity of the constraints.

For problems with linear constraints and a nonlinear objective function both methods are equivalent to normal methods of feasible directions. They only differ in the way they handle the nonlinear constraints. In MFD the nonlinear constraints are supposed to be sufficiently linear, so that successive linearizations lead to rapid convergence; in MIP they are added to the objective function. Hence, if the assumption of near linearity is valid, the MFD method is the best, but if the problem has very little resemblance to the linear programming problem the MIP method should be selected. Nonlinear scheduling problems mostly belong to the first class, many design optimization problems to the second class. A summary of some properties of the MIP and MFD methods is also given in the appendix.

[5] One could try to determine z^k by maximizing $f(x)$ in the triangle determined by x^k, y^k, and z^{k-1} under the condition that the maximum should be feasible. x^{k+1} could then be chosen on the line connecting x^k and z^k. This would lead to a sequence of two-dimensional nonlinear programming problems, which can only be easily solved in special cases.

Appendix. Some features of various methods.

	Dual methods	Small step gradient methods	Large step gradient methods	Interior point methods	MIP method	MFD method
Nonconvex programs	N	Y	Y	Y	Y	Y
Finite for linear programs	Y	N[a]	Y	N	Y	Y
Finite for quadratic programs	N[b]	N	Y[c]	N	Y	Y
Efficient for linearly constrained programs	N	N	Y	N	Y	Y
Efficient for problems with only a few nonlinearities	Y	N	Y	N	Y[d]	Y[d]
Efficient for unconstrained programs	N	N	Y[e]	Y[e]	Y[e]	Y[e]
Infinite number of constrains possible[f]	Y	N	Y	N[g]	N[g]	Y
Intermediate solution (near-)feasible	N	Y	Y	Y	Y	Y
Upper bound in convex programs	Y	N	N[h]	N	N	Y
Speed of convergence[i]	3	3	2	2	2	1
Amount of work per step	1	2	3	3	2	3
Accuracy of calculations[i]	3	2	2	2	2	2
Simplicity of computer program	1	1	3	2	2	3

Y and N are Yes and No answers to the question.

1 = favorable; 2 = reasonable; 3 = unfavorable (all in comparison to other methods).

Remarks.

[a] A simple change in an MAP computer program will make it finite for linear programs but it will then no longer be a small step gradient method.

[b] With the accelerating device the cutting plane method is finite for quadratic programs.

[c] Not all large step gradient methods are finite, for example the gradient projection method is not.

[d] The MFD method will be more efficient than the MIP method.

[e] The interior point methods will be more efficient than the methods of feasible directions.

[f] The class of solvable problems is restricted.

[g] A modification of the (modified) interior point methods may work (see footnote 4).

[h] In a few methods of feasible directions an upper bound can be obtained at relatively little cost.

[i] Needs further computational study.

REFERENCES

[1] E. M. L. Beale, *On quadratic programming*, Naval Res. Logist. Quart., 6 (1959), pp. 227–243.

[2] J. F. Benders, *Partitioning procedures for solving mixed-variables programming problems*, Numer. Math., 4 (1962), pp. 238–252.

[3] E. W. Cheney and A. A. Goldstein, *Newton's method for convex programming and Tchebycheff approximation*, Ibid., 1 (1959), pp. 253–268.

[4] J. B. CROCKETT AND H. CHERNOFF, *Gradient methods of maximization*, Pacific J. Math., 5 (1955), pp. 33–50.

[5] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, 1963.

[6] A. V. FIACCO AND G. P. MCCORMICK, *Programming under nonlinear constraints by unconstrained minimization: a primal-dual method*, RAC-TP-96, Research Analysis Corporation, McLean, Virginia, 1963.

[7] ———, *The sequential unconstrained minimization technique for nonlinear programming, a primal-dual method*, Management Sci., 10 (1964), pp. 360–366.

[8] ———, *Extensions of the sequential unconstrained minimization technique (SUMT) for nonlinear programming*, presented at the American Meeting of the Institute of Management Science, San Francisco, 1965.

[9] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[10] S. I. GASS, *Linear Programming, Methods and Applications*, McGraw-Hill, New York, 1964.

[11] R. E. GRIFFITH AND R. A. STEWART, *A nonlinear programming technique for the optimization of continuous processing systems*, Management Sci., 7 (1961), pp. 379–392.

[12] G. HADLEY, *Nonlinear and Dynamic Programming*, Addison-Wesley, Reading, Massachusetts, 1964.

[13] H. O. HARTLEY AND R. R. HOCKING, *Convex programming by tangential approximation*, Management Sci., 9 (1963), pp. 600–612.

[14] P. HUARD, *Résolution de programmes mathématiques à constraintes non linéaires par la méthode des centres*, Note Electricité de France HR 5690/3/317, 1964.

[15] J. E. KELLEY, JR., *The cutting-plane method for solving convex programs*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 703–712.

[16] H. W. KUHN AND A. W. TUCKER, *Non-linear programming*, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, California, 1950, pp. 481–492.

[17] C. E. MILLER, *The simplex method for local separable programming*, Recent Advances in Mathematical Programming, R. L. Graves and P. Wolfe, eds., McGraw-Hill, New York, 1963, pp. 89–100.

[18] J. B. ROSEN, *The gradient projection method for nonlinear programming, Part I, Linear constraints*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 181–217.

[19] ———, *Convex partitioning programming*, Recent Advances in Mathematical Programming, R. L. Graves and P. Wolfe, eds., McGraw-Hill, New York, 1963, pp. 159–176.

[20] J. B. ROSEN AND J. C. ORNEA, *Solution of nonlinear programming problems by partitioning*, Management Sci., 10 (1963), pp. 160–173.

[21] P. WOLFE, *The simplex method for quadratic programming*, Econometrica, 27 (1959), pp. 382–398.

[22] ———, *Accelerating the cutting plane method for nonlinear programming*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 481–488.

[23] ———, *Methods of nonlinear programming*, Recent Advances in Mathematical Programming, R. L. Graves and P. Wolfe, eds., McGraw-Hill, New York, 1963, pp. 67–86.

[24] G. ZOUTENDIJK, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

ON THE PROBABILITY DISTRIBUTION OF THE OPTIMUM OF A RANDOM LINEAR PROGRAM*

ANDRÁS PRÉKOPA†

1. Introduction. In the present paper we shall consider linear programming problems

$$\begin{aligned} \mu &= \max c'x, \\ Ax &= b, \quad x \geqq 0, \end{aligned} \tag{1.1}$$

where A is an $m \times n$ matrix, c and x are n-dimensional vectors, and b is an m-dimensional vector. We shall suppose that A, b, c have random elements and components, respectively. As μ is a function of the variables in A, b and c,

$$\mu = \mu(A, b, c), \tag{1.2}$$

it is also a random variable and its probability distribution is what we are interested in. This problem is of basic importance and is conceivable as a stochastic sensitivity analysis of a linear programming model. The question how the transformation $A, b, c \to \mu$ operates under the presence of random influences in A, b, and c does not play just the role of a sensitivity analysis, however. In fact, in A, b, c we may have not just small random disturbances but random variables of significant variation.

The problem in its general form has been considered by Tintner [2], [3], and Babbar [1]. In these papers it is supposed that the random variation does not change the optimal basis in the sense that the subscripts of the optimal basis vectors remain the same for all possible values of A, b, c. Thus finding the probability distribution of μ is equivalent to finding the probability distribution of an (also random) linear functional defined over the random solution of a set of linear equations. In this respect it is also possible to proceed in two different ways: either to develop μ into a finite Taylor series and use the leading, linear terms as an approximation to μ and obtain its probability distribution, or to consider the components of the solution as fractions of random determinants, approximate their distributions by the normal law and again approximate by the normal law the fraction of two normally distributed variables. This method has the handicap that it produces sophisticated approximation formulas.

* Received by the editors July 12, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Eötvös L. University, Budapest, Hungary, and the Mathematical Institute of the Hungarian Academy of Sciences.

In §§2, 3, 4, we consider systems of linear equations, the probability distribution of a random linear functional defined over the solutions and apply this theory for our original problem concerning random linear programs. Our approximation formulas for the characteristics of μ, especially for the dispersion, will be particularly simple, as simple as possible in this general formulation of the problem from the point of view of practical application, involving the primal and dual optimal solutions of the linear programming problem carried out with the expectations and the covariances of the present random variables. We express our statements in limit theorems and list carefully all mathematical assumptions. Our results are formulated in general, containing the essential features of the problem and allowing the possibility of specialization when facing a particular problem.

The results of the present paper, however, apply to the case when the random elements in the technology matrix keep the optimal basis (basis subscripts), obtained by computing with the expectations, with a high probability. To more general questions we return in subsequent papers.

Since in the present approach the principal aim is to reduce μ to a sum of random variables, the asymptotic normality of this sum will be supposed. In the particular cases where the sum in question contains an increasing number of independent random variables, e.g., A has independent elements or it is enough if its rows or columns are independent, the limit distribution theory of sums of double sequences of independent random variables can be applied. (See [19].) If independence does not occur then we may suppose the joint normality of the random variables in A, b, c which is sufficient, or suppose simply the sum in question to be normally distributed; but there is no detailed general theory of the limit distributions of sums of double sequences of dependent random variables. On the other hand, any particular problem reveals, in some specific way, how the random elements intervene, from the knowledge of which we may assume the normality of the approximating sum.

2. Systems of random linear equations. Consider the following system of linear equations:

$$\sum_{k=1}^{m} a_{ik}x_k = b_i, \qquad i = 1, \cdots, m. \tag{2.1}$$

Let us denote by B the matrix of the equations and by b the vector consisting of the b_i's as components and let us introduce an m-dimensional vector c. If B is nonsingular then (2.1) has a unique solution $B^{-1}b$. Suppose now that B, c, b are all random and that all elements and components have finite variances. We are interested in the probability distribution of the

functional

$$\mu = c'Rb, \quad \text{where } R = B^{-1} \tag{2.2}$$

and where the prime denotes transpose. In order to avoid complications in the notation we shall suppose that B is independent of the couple c, b and that c, b are also independent of each other. This assumption does not play any significant role here. We shall denote by $a_1, \cdots, a_m$ the columns of B and by D_{ik} the cross covariance matrix of a_i and a_k, i.e.,

$$D_{ik} = \mathrm{E}[(a_i - a_i^{(0)})'(a_k - a_k^{(0)})], \quad i, k = 1, \cdots, m, \tag{2.3}$$

where $a_i^{(0)} = \mathrm{E}(a_i)$, $i = 1, \cdots, m$, and E is the expectation operator. The expectations of B, b, c, a_{ik}, b_i, c_j will be denoted by B_0, b_0, c_0, $a_{ik}^{(0)}$, $b_i^{(0)}$, $c_j^{(0)}$, respectively. R_0 will denote B_0^{-1}. The covariance matrices of γ and β will be denoted by C and F, respectively.

Disregarding for a while the random nature of our quantities, we shall give the finite Taylor expansion of μ around the expectations, as far as the second order terms. It can be done by using a formula well-known in matrix theory stating that if the inverse of a nonsingular square matrix is R and we modify the original matrix by adding ξ to the element in the ith row and kth column then the inverse of the modified matrix will be

$$R - \frac{\xi}{1 + r_{ki}\,\xi} \begin{pmatrix} r_{1i} \\ \vdots \\ r_{mi} \end{pmatrix} (r_{k1} \cdots r_{km}). \tag{2.4}$$

Hence if we change B in the manner described above and consider the change in the functional then we obtain

$$\mu(a_{ik} + \xi) - \mu(a_{ik}) = -\frac{\xi}{1 + r_{ki}\,\xi}\, y_i\, x_k\,, \tag{2.5}$$

where

$$y_i = \sum_{j=1}^{m} c_j r_{ji}\,.$$

From this it follows that

$$\frac{\partial \mu}{\partial a_{ik}} = y_i\, x_k\,, \qquad i, k = 1, \cdots, m. \tag{2.6}$$

By a double application of (2.4) we get

$$\mu(a_{ik} + \xi) + \mu(a_{ik} - \xi) - 2\mu(a_{ik}) = \xi^2 y_i\, x_k\, \frac{r_{ki}}{1 - \xi^2 r_{ki}^2}\,, \tag{2.7}$$

hence

$$\frac{\partial^2 \mu}{\partial a_{ik}^2} = \lim_{\xi \to 0} \frac{\mu(a_{ik} + \xi) + \mu(a_{ik} - \xi) - 2\mu(a_{ik})}{\xi^2} \tag{2.8}$$

$$= y_i x_k r_{ki}, \qquad i, k = 1, \cdots, m.$$

We can determine similarly the mixed partial derivatives. The result is the following

$$\frac{\partial^2 \mu}{\partial a_{ik} \partial a_{pq}} = \lim_{\substack{\xi \to 0 \\ \eta \to 0}} \frac{\begin{matrix}\mu(a_{ik} + \xi, a_{pq} + \eta) - \mu(a_{ik} + \xi, a_{pq}) \\ - \mu(a_{ik}, a_{pq} + \eta) + \mu(a_{ik}, a_{pq})\end{matrix}}{\xi\eta} \tag{2.9}$$

$$= x_k y_p r_{qi} + x_q y_i r_{kp}, \quad \text{for} \quad |i - p| + |k - q| > 0.$$

Let us finally mention the derivatives where c_i and b_i are involved:

$$\frac{\partial \mu}{\partial c_j} = x_j, \qquad \frac{\partial \mu}{\partial b_j} = y_j, \qquad \frac{\partial^2 \mu}{\partial c_i \partial b_j} = r_{ij}, \tag{2.10}$$

$$\frac{\partial^2 \mu}{\partial a_{ik} \partial c_j} = x_k r_{ji}, \qquad \frac{\partial^2 \mu}{\partial a_{ik} \partial b_j} = y_i r_{kj}, \qquad i, j, k = 1, \cdots, m.$$

Let us introduce the notations

$$\begin{aligned} y' &= c'R, \quad y_0' = c_0'R_0, \\ x &= Rb, \quad x_0 = R_0 b_0, \end{aligned} \tag{2.11}$$

and denote by $x_i^{(0)}$, $y_j^{(0)}$ the components of x_0, y_0, respectively. Furthermore

$$\begin{aligned} a_{ik} - a_{ik}^{(0)} &= \xi_{ik}, \quad B - B_0 = \Xi, \\ c_i - c_i^{(0)} &= \gamma_i, \quad c - c_0 = \gamma, \\ b_i - b_i^{(0)} &= \beta_i, \quad b - b_0 = \beta, \qquad i, k = 1, \cdots, m. \end{aligned} \tag{2.12}$$

With the aid of these notations the desired Taylor expansion is the following:

$$\begin{aligned} \mu = \mu_0 &- \sum_{i,k=1}^{m} y_i^{(0)} \xi_{ik} x_k^{(0)} + \sum_{i=1}^{m} y_i^{(0)} \gamma_i + \sum_{k=1}^{m} x_k^{(0)} \beta_k \\ &+ \frac{1}{2} \sum_{i,k=1}^{m} y_i \xi_{ik}^2 r_{ki} x_k + \sum_{i,k,j=1}^{m} x_k r_{ji} \xi_{ik} \gamma_j \\ &+ \sum_{i,k,j=1}^{m} y_i r_{kj} \xi_{ik} \beta_j + \sum_{|i-p|+|k-q|>0} (x_k y_p r_{qi} + x_q y_i r_{kp}) \xi_{ik} \xi_{pq}, \end{aligned} \tag{2.13}$$

or in a concise form,

$$\mu - \mu_0 = -y_0'\Xi x_0 + y_0'\gamma + x_0'\beta + \rho, \tag{2.14}$$

where the error term ρ is given by

$$\rho = -\tfrac{3}{2}\, x'\Xi_2 y + 2y'\Xi R\Xi x + \gamma' R\Xi x + y'\Xi R\beta, \tag{2.15}$$

and Ξ_2 is the matrix consisting of the entries $\xi_{ik}^2 r_{ki}$. In the above development of the error term, x, y, and R, which are functions of b, c, B, are taken at a point $b_i^{(0)} + \vartheta\beta_i$, $c_i^{(0)} + \vartheta\gamma_i$, $a_{ij}^{(0)} + \vartheta\xi_{ij}$, where $0 < \vartheta < 1$.

The leading term in (2.14) has expectation 0 and variance

$$\sigma^2 = \sum_{i,k=1}^{m} x_i^{(0)} y_0' D_{ik} y_0 x_k^{(0)} + y_0' C y_0 + x_0' F x_0 . \tag{2.16}$$

If the columns of B are independent random vectors, as it can be supposed in some practical cases, then σ^2 reduces to

$$\sigma^2 = \sum_{k=1}^{m} (x_k^{(0)})^2 y_0' D_{kk} y_0 + y_0' C y_0 + x_0' F x_0 , \tag{2.17}$$

which further reduces if all elements of B are independent. If also c and b have independent components then we have

$$\sigma^2 = \sum_{i,k=1}^{m} (y_i^{(0)})^2 \sigma_{ik}^2 (x_k^{(0)})^2 + \sum_{i=1}^{m} (y_i^{(0)})^2 s_i^2 + \sum_{k=1}^{m} (x_k^{(0)})^2 t_k^2, \tag{2.18}$$

where

$$\sigma_{ik}^2 = \mathrm{E}(\xi_{ik}^2), \quad s_i^2 = \mathrm{E}(\gamma_i^2), \quad t_k^2 = \mathrm{E}(\beta_k^2), \qquad i, k = 1, \cdots, m.$$

In the next sections we shall give sufficient conditions under which $(\mu - \mu_0)/\sigma$ has an asymptotic normal distribution. This will mean from the practical point of view that μ has an asymptotic normal distribution with expectation μ_0, i.e., the value of the functional belonging to the expectations of all values involved, and variance given by (2.16) which may specialize.

3. Limit distribution theorems for random linear equations. First we prove a lemma.

LEMMA. *Let H_1, H_2, $\cdots$ be a sequence of events with the property that*

$$\lim_{N\to\infty} \mathrm{P}(H_N) = 1.$$

Let further η_N and ζ_N be two sequences of random variables, where η_N has a limit distribution, i.e.,

$$\lim_{N\to\infty} \mathrm{P}(\eta_N < x) = G(x)$$

at every point of continuity of $G(x)$ *and* ζ_N *tends stochastically to* 0 (*in symbols,* $\zeta_N \Rightarrow 0$), *i.e.*,

$$\lim_{N\to\infty} P(|\zeta_N| > \epsilon) = 0, \qquad \text{for every } \epsilon > 0$$

(ζ_N *has a degenerate limit distribution*). *Under these conditions*

$$\lim_{N\to\infty} P(\eta_N + \zeta_N \mid H_N) = G(x)$$

at every point of continuity of $G(x)$.

Since this lemma is essentially Cramér's lemma (see [18, p. 254]) expressed in a slightly modified form, we omit the proof.

In order to obtain limit theorems we can proceed in two directions. We may keep m, the size of the system of equations, fixed, while the random disturbances have a slowing down tendency. This is the case when, for example, the random disturbances are due to some inaccuracy in the measurements of the data which shows a decreasing tendency upon using more data or, in other terms, a larger sample. The other possibility is to increase m. In this case we shall also suppose implicitly that the random elements are small as compared to the expectations, but for convergence to the normal distribution the increasing size of the matrix B contributes also. First we formulate two theorems in general forms.

THEOREM 1. *If a function* $f(z_1, z_2, \cdots, z_k)$ *has continuous second order derivatives in some convex neighborhood of the point* $(z_1^{(0)}, z_2^{(0)}, \cdots, z_k^{(0)})$, *where* k *is fixed, and if for a sequence of random vectors* $(\xi_1^{(N)}, \xi_2^{(N)}, \cdots, \xi_k^{(N)})$ *with* $E(\xi_i^{(N)}) = 0$, $i = 1, \cdots, k$, $N = 1, 2, \cdots$, *the following conditions are satisfied*:

(1) $$\xi_i^{(N)} \Rightarrow 0 \text{ if } N \to \infty, \quad i = 1, \cdots k,$$

(2) $$\lim_{N\to\infty} P\left(\frac{1}{\sigma_N}\sum_{i=1}^{k} \frac{\partial f^{(0)}}{\partial z_i}\,\xi_i^{(N)} < x\right) = G(x),$$

at every point of continuity of $G(x)$, *where* $\partial f^{(0)}/\partial z_i$ *means the derivative* $\partial f/\partial z_i$ *taken at* $(z_1^{(0)}, z_2^{(0)}, \cdots, z_k^{(0)})$ *and* σ_N *is the dispersion of*

$$\sum_{i=1}^{k} (\partial f^{(0)}/\partial z_i)\xi_i^{(N)},$$

(3) $$\frac{1}{\sigma_N}\sum_{i,j=1}^{k} \frac{\partial^2 f^{(1)}}{\partial z_i\,\partial z_j}\,\xi_i^{(N)}\,\xi_j^{(N)} \Rightarrow 0 \text{ if } N \to \infty,$$

where the superscript (1) *means that the derivative is taken at an arbitrary point of the convex domain and this point may also vary with* N, *then we have*

$$\lim_{N\to\infty} P\left\{\frac{1}{\sigma_N}\left[f(z_1^{(0)} + \xi_1^{(N)}, \cdots, z_k^{(0)} + \xi_k^{(N)}) - f(z_1^{(0)}, \cdots, z_k^{(N)})\right] < x\right\} = G(x).$$

Proof. Let H_N denote the event that $(z_1^{(0)} + \xi_1^{(N)}, \cdots, z_k^{(0)} + \xi_k^{(N)})$ is in that neighborhood of $(z_1^{(0)}, \cdots, z_k^{(0)})$ where f has continuous second order derivatives. In this case,

$$f(z_1^{(0)} + \xi_1^{(N)}, \cdots, z_k^{(0)} + \xi_k^{(N)}) - f(z_1^{(0)}, \cdots, z_k^{(0)})$$
$$= \sum_{i=1}^{k} \frac{\partial f^{(0)}}{\partial z_i} \xi_i^{(N)} + \frac{1}{2} \sum_{i,j=1}^{k} \frac{\partial^2 f^{(1)}}{\partial z_i \partial z_j} \xi_i^{(N)} \xi_j^{(N)}, \tag{3.1}$$

where $\partial^2 f^{(1)}/\partial z_i \partial z_j$ is the second order derivative taken at $(z_1^{(0)} + \vartheta\xi_1^{(N)}, \cdots, z_k^{(0)} + \vartheta\xi_k^{(N)})$, $0 < \vartheta < 1$. According to (1), $\lim_{N\to\infty} \mathrm{P}(H_N) = 1$. Let us divide by σ_N on both sides in (3.1). Then the second term on the right-hand side tends stochastically to 0 according to (3). Let us denote this term by ζ_N and the first term by η_N. Then a direct application of the lemma completes the proof.

Condition (3) is clearly fulfilled if

$$\frac{1}{\sigma_N} \xi_i^{(N)} \xi_j^{(N)} \Rightarrow 0, \qquad i, j = 1, \cdots, k.$$

Before stating Theorem 2, we mention the notion of a star domain. An open domain K around and containing a point $(z_1, \cdots, z_k)$ in the k-dimensional space is called a *star domain* if the intersection of K with any ray $(z_1 + t\xi_1, \cdots, z_k + t\xi_k)$, $t > 0$, is an open interval. This may contain, in particular, every point of the ray. The point $(z_1, \cdots, z_k)$ is called the *seed* of the domain. This notion will be important to extend the possibility of the Taylor-series expansion around the given point as large as possible.

For later purposes we introduce a notion, that of a *maximal star domain around a nonsingular matrix* B_0, which by definition consists of all matrices of the form

$$B_0 + t\Xi,$$

where for any given Ξ, t runs continuously from 0 until the sum becomes singular. That singular matrix is excluded, however.

THEOREM 2. *Suppose that we have a sequence of functions of an increasing number of variables* $f_N(z_1, z_2, \cdots, z_{k_N})$ *where* $k_N \to \infty$ *as* $N \to \infty$, *and each* f_N *has a neighborhood, a star domain around a point* $(z_{N1}^{(0)}, \cdots, z_{Nk_N}^{(0)})$ *where its second order derivatives exist and are continuous. Suppose furthermore that we have a double sequence of random variables* $\xi_1^{(N)}, \xi_2^{(N)}, \cdots, \xi_{k_N}^{(N)}$, *with expectations* 0 *and finite variances, satisfying the following conditions*

(1) $$\lim_{N\to\infty} \mathrm{P}\{(z_{N1}^{(0)} + \xi_1^{(N)}, \cdots, z_{Nk_N}^{(0)} + \xi_{k_N}^{(N)}) \in K_N\} = 1,$$

where K_N *is the abovementioned neighborhood,*

(2) $$\lim_{N\to\infty} \mathrm{P}\left(\frac{1}{\sigma_N} \sum_{i=1}^{k_N} \frac{\partial f_N^{(0)}}{\partial z_i} \xi_i^{(N)} < x\right) = G(x)$$

at every point of continuity of $G(x)$ ($\partial f_N^{(0)}/\partial z_i$ *and* $\partial^2 f_N^{(1)}/\partial z_i \partial z_j$ *have the same meaning as in Theorem* 1),

$$\frac{1}{\sigma_N} \sum_{i,j=1}^{k_N} \frac{\partial^2 f_N^{(1)}}{\partial z_i \, \partial z_j} \xi_i^{(N)} \xi_j^{(N)} \Rightarrow 0 \quad \text{if} \quad N \to \infty. \tag{3}$$

Then

$$\lim_{N\to\infty} \mathrm{P}\left\{\frac{1}{\sigma_N}[f_N(z_{N1}^{(0)} + \xi_1^{(N)}, \cdots, z_{Nk_N}^{(0)} + \xi_{k_N}^{(N)}) - f(z_{N1}^{(0)}, \cdots, z_{Nk_N}^{(0)})] < x\right\} = G(x)$$

at every point of continuity of $G(x)$.

The proof is similar to that of Theorem 1.

It is worth mentioning that the fulfillment of condition (1) in Theorem 2 may be the cause of the slowing down tendency of random elements or the increase of K_N or both.

In both theorems we used the same idea Cramér used when proving the asymptotic normality of functions of moments (see [19, pp. 366–367], noting that in that case the number of variables is fixed). We can apply these theorems for random linear equations. In the following theorem we shall omit the subscript N which would refer to the fact that we have a sequence of random elements. Thus all our previous notations concerning random equations are applicable.

THEOREM 3. *Suppose that* m, B_0, c_0, b_0 *are fixed and that* B_0 *is nonsingular and introduce the following conditions:*

$$\sigma_{ik} \to 0, \quad i,k = 1, \cdots, m, \tag{1}$$

$$\mathrm{P}\left\{\frac{1}{\sigma}[-y_0' \, \Xi \, x_0 + y_0' \, \gamma + x_0' \beta] < x\right\} \to \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-u^2/2} \, du, \quad \text{for every } x, \ -\infty < x < \infty, \tag{2}$$

$$\frac{\rho}{\sigma} \Rightarrow 0. \tag{3}$$

Then for every x,

$$\mathrm{P}\left(\frac{\mu - c_0' \, R_0 \, b_0}{\sigma} < x\right) \to \Phi(x). \tag{3.2}$$

Proof. Theorem 3 is an immediate consequence of Theorem 1 applied to the function $\mu = \mu(A, b, c)$ of $m^2 + 2m$ variables, (A_0, b_0, c_0) as the point around which the Taylor-series development is taken, and Ξ, γ, β as the sequence of $(m^2 + 2m)$-dimensional random vectors. We just have to mention that condition (1) in Theorem 1 is ensured by condition (1)

of Theorem 3. Various consequences of this theorem can be derived. Among them we mention the simplest.

COROLLARY. *Suppose that the* $m^2 + 2m$ *random variables in* Ξ, γ *and* β *have a normal joint distribution and*

$$\sigma_{\max} \to 0, \quad \frac{\sigma_{\max}^2}{\sigma} \to 0, \qquad \text{where } \sigma_{\max} = \max(\sigma_{ik}, s_i, t_k).$$

Then (3.2) *holds.*

Proof. All that we have to verify is the fulfillment of (3) in Theorem 3. If we look at the detailed expression of ρ given by the last terms in (2.13), we see that, separately, each term of that sum divided by σ converges stochastically to 0. In fact, considering the quadratic terms ξ_{ik}^2/σ we see by the Markov inequality that

$$P\left(\frac{\xi_{ik}^2}{\sigma} > \epsilon\right) \leqq \frac{E(\xi_{ik}^2)}{\sigma} \leqq \frac{\sigma_{\max}^2}{\sigma} \to 0.$$

For all other terms the Chebyshev inequality can be applied.

THEOREM 4. *Consider a sequence of matrices and vectors* B_0, c_0, b_0, *and a corresponding random sequence* Ξ, γ, β (*the subscripts are omitted*), *where* m, *the size of the matrices* (*equal to the dimension of the vectors*), *tends to infinity. Suppose that all* B_0 *matrices are nonsingular. To every* B_0 *in the sequence there corresponds a maximal star domain* K *where* $B_0 + \Xi$ *is nonsingular and the Taylor expansion around* B_0 *applies. Suppose that*

$$P(B_0 + \Xi \in K) \to 1, \tag{1}$$

$$P\left\{\frac{1}{\sigma}(-y_0'\Xi x_0 + x_0'\gamma + y_0'\beta) < x\right\} \to \Phi(x), \quad -\infty < x < \infty, \tag{2}$$

$$\frac{\rho}{\sigma} \Rightarrow 0. \tag{3}$$

Under these conditions‡

$$P\left\{\frac{1}{\sigma}(\mu - c_0'R_0 b_0) < x\right\} \to \Phi(x), \quad -\infty < x < \infty.$$

The proof of this theorem is similar to that of Theorem 3. Analyzing the conditions here, (1) and (2) are realistic as the size of the matrix increases. The crucial point is condition (3) which may very easily fail to hold. In fact, first of all, the fourth term containing the squares ξ_{ik}^2 may not be negligible as compared to σ. It does not have, in general, expectation 0 even

‡ Instead of $\Phi(x)$ we may suppose some other distribution function too.

in the case where all random variables are independent. It seems, therefore, advisable to attach the sum

$$\frac{1}{2}\sum_{i,k=1}^{m} y_i^{(0)}\xi_{ik}^2 r_{ki}^{(0)} x_k^{(0)}$$

to the leading term, changing it into

$$-\sum_{i,k=1}^{m} y_i^{(0)}\xi_{ik}(1 - \tfrac{1}{2}\xi_{ik}r_{ki})x_k^{(0)} + y_0'\gamma + x_0'\beta. \tag{3.3}$$

We may then approximate the distribution of μ by a normal distribution with the expectation

$$c_0'R_0b_0 + \frac{1}{2}\sum_{i,k=1}^{m} y_i^{(0)}\sigma_{ik}^2 r_{ki}^{(0)} x_k^{(0)} \tag{3.4}$$

and variance (2.16), where D_{ik} has to be replaced by

$$D_{ik} - T_iD_{ik}T_k - T_iD_{ik} - D_{ik}T_k \tag{3.5}$$

and T_i is a diagonal matrix consisting of elements $r_{1i}^{(0)}, \cdots, r_{mi}^{(0)}$ in the diagonal. The same sum that we added to the leading term has to be subtracted from the remainder and it is more realistic to say that the new remainder divided by the dispersion of the new leading term tends stochastically to 0.

4. Application to random linear programs. Consider the linear programming problem

$$\mu_0 = \max c_0'x, \tag{4.1}$$

subject to the conditions

$$A_0x = b_0, \quad x \geqq 0, \tag{4.2}$$

and suppose that it has a unique optimal basis B_0 which, for the sake of simplicity, we suppose to be the set of vectors $a_1^{(0)}, \cdots, a_m^{(0)}$. We also suppose that A_0 has rank m. Consider also the problem

$$\mu = \max c'x, \tag{4.3}$$

subject to the conditions

$$Ax = b, \quad x \geqq 0, \tag{4.4}$$

where A, b, c have random elements, components, respectively. In these problems we apply the same notations as those used concerning random equations in §§2, 3, but we observe that A has mn elements and c has n components.

We suppose also that B_0 is nondegenerate. There is then a neighborhood of A_0, b_0, c_0 in which the problem (4.3)–(4.4) will preserve the subscripts of the optimal basis. Keeping m and n fixed, consider a sequence of random matrices, vectors A, b, c, respectively. If we suppose that

$$\sigma_{\max} \to 0, \tag{4.5}$$

where $\sigma_{\max} = \max(\sigma_{ik}, t_i, s_k, i = 1, \cdots, m; k = 1, \cdots, n)$, the probability that $B = (a_1, \cdots, a_m)$ will be the optimal basis to problem (4.3)–(4.4) tends to 1. Hence, according to our lemma, $(\mu - \mu_0)/\sigma$ will have the same asymptotic probability distribution unconditionally or conditionally, given that B is optimal. If, furthermore, conditions (2) and (3) are also satisfied in Theorem 3, where all quantities, vectors, matrices are taken from the random equation $Bx = b$, and c means the vector consisting of the first m components of that used in (4.3), then we may state the following.

THEOREM 5. *The optimum value μ of the random programming problem* (4.3)–(4.4) *has an asymptotic normal distribution with expectation μ_0, the optimum of the program taken with the expectations in each place, and variance* (2.16), *where x_0, y_0 are the primal and dual optimal solutions of the first problem; more exactly, x_0 is a part of the primal optimal solution consisting of the basic components. The meaning of D_{ik}, C, F remains unchanged. Asymptotic normality means that the probability that $(\mu - \mu_0)/\sigma < x$ tends to $\Phi(x)$.*

It is seen from these that the present approach gives a particularly simple result which is very advantageous from the practical point of view because in the characteristics of the random variable μ such vectors and matrices appear as the primal and dual optimal solutions x_0, y_0 and D_{ik}, C and F, the covariance matrices of the random variables involved.

We may also apply Theorem 4 by considering a sequence of programming problems, where $m \to \infty$, $n \to \infty$. Here we suppose that at each problem with A_0, b_0, c_0 there is a unique finite, nondegenerated optimum and the probability that the optimal basis has the same column subscripts in problem (4.1)–(4.2) and in (4.3)–(4.4) tends to 1. Then if we take into account our lemma, the results of Theorem 4 are applicable, where x_0 and y_0 have the same meaning as before.

One practical conclusion of these results is the following: if for some reason we solve the linear programming problem with the expectations, e.g., with predicted prices and predicted technology coefficients, but we have information about their random variation, then we may set up confidence limits for the optimum value which would have been the result if we had programmed with the particular realization of the random data in A, b, and c.

REFERENCES

[1] M. M. Babbar, *Distributions of solutions of a set of linear equations (with an application to linear programming)*, J. Amer. Statist. Assoc., 50 (1955), pp. 854–869.

[2] G. Tintner, *Stochastic linear programming with applications to agricultural economics*, Second Symposium on Linear Programming, vol. 1, National Bureau of Standards, Washington, D. C., 1955, pp. 197–227.

[3] ———, *Les programmes linéaires stochastiques*, Revue d'Économie Politique, 67 (1957), pp. 208–215.

[4] ———, *A note on stochastic linear programming*, Econometrica, 28 (1960), pp. 490–495.

[5] H. M. Wagner, *On the distribution of solutions in linear programming problems*, J. Amer. Statist. Assoc., 53 (1958), pp. 161–163.

[6] J. V. Talacko, *On stochastic linear inequalities*, Trabajos Estadíst., 10 (1959), pp. 89–112.

[7] A. Madansky, *Inequalities for stochastic linear programming problems*, Management Sci., 6 (1960), pp. 197–204.

[8] S. Vajda, *Inequalities in stochastic linear programming*, Bull. Inst. Internat. Statist., 36 (1958), pp. 357–363.

[9] H. W. Kuhn and R. E. Quandt, *An experimental study of the simplex method*, Proc. Symp. Appl. Math., 15 (1963), pp. 107–124.

[10] A. T. Lonseth, *Systems of linear equations with coefficients subject to error*, Ann. Math. Statist., 13 (1942), pp. 332–337.

[11] ———, *On relative errors in systems of linear equations*, Ibid., 15 (1944), pp. 323–325.

[12] ———, *The propagation of errors in linear problems*, Trans. Amer. Math. Soc., 62 (1947), pp. 193–212.

[13] G. E. P. Box and J. S. Hunter, *A confidence region for the solution of a set of simultaneous equations with an application to experimental design*, Biometrika, 41 (1954), pp. 190–199.

[14] R. E. Quandt, *Probabilistic errors in the Leontief system*, Naval Res. Logist. Quart., 5 (1958), pp. 155–170.

[15] H. D. Mills, *Marginal values of matrix games and linear programmes*, Linear Inequalities and Related Systems, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, 1956, pp. 183–193.

[16] A. C. Williams, *Marginal values in linear programming*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 82–94.

[17] E. Bodewig, *Matrix Calculus*, North-Holland, Amsterdam, 1956.

[18] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.

[19] B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley, Reading, Massachusetts, 1954.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

ITERATIVE SOLUTION OF NONLINEAR OPTIMAL CONTROL PROBLEMS*

J. B. ROSEN†

Abstract. The solution of nonlinear, state-constrained, discrete optimal control problems by mathematical programming methods is described. The iterative solution consists essentially of Newton's method with a convex (or linear) programming problem solved at each iteration. Global convergence of the iterative method is demonstrated provided a convexity and constraint set condition are both satisfied. The computational solution of nonlinear equation control problems makes use of a previously developed method for state-constrained linear equation problems. The solution method for nonlinear problems is illustrated by means of two numerical examples.

1. Introduction. The optimal control problem considered here is a rather general type of discrete problem. We wish to minimize a convex function of the state and control vectors, where the control vectors must lie in a specified convex set. In addition the state vectors must also satisfy specified constraints at each discrete time, as well as initial and terminal conditions. Furthermore, the system dynamics may be given by a nonlinear recursion relation provided that the nonlinearity is convex in an appropriate way. A discrete system of the type considered here may represent a process which is actually discrete (see, for example, [3], [1]), or it may be obtained from a finite difference approximation to a continuous system in which we wish to minimize a convex functional. Such an approximation is *always* required when a numerical integration, using a digital computer, is part of the solution process.

The purpose of this report is to describe a computational method for solving this general type of discrete problem, and to show by means of the relevant theorems that the method will always work when the appropriate assumptions are satisfied. The method is an iterative procedure that determines a sequence of admissible trajectories (state and control vectors satisfying all constraints); the sequence converging to an admissible trajectory that satisfies the necessary conditions for optimality. The method has been used to obtain numerical solutions to several small nonlinear test problems. In addition to showing that it is not difficult to implement the

* Received by the editors June 28, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† Computer Sciences Department and Mathematics Research Center, University of Wisconsin, Madison, Wisconsin. This research was sponsored in part by the National Aeronautics and Space Administration under Research Grant NGR-50-002-028 and in part by the Mathematics Research Center under Contract No. DA-11-022-ORD-2059.

scheme described here, these numerical results show that, at least for the test problems considered, the number of iterations required is small.

In a previous publication [14] a statement of the Kuhn-Tucker conditions was given for the nonlinear state-constrained problem considered here. A computational procedure for systems described by linear recursion relations was also given based on a convex (or linear) programming computer code. Numerical results described there show that this computational procedure is efficient for typical linear systems. The method described in the present paper takes advantage of this efficiency by solving a sequence of such linear problems. From this point of view the method of the present report may be thought of as Newton's method (see, for example, [9]) with a convex (or linear) programming problem solved at each iteration. The use of various forms of Newton's method for the numerical solution of optimal control problems has been proposed in a number of earlier publications [4], [6], [10], [12]. The two important differences between the method described here and these earlier proposals are that (1) in the present method *global* convergence is assured when a convexity and constraint set condition are both satisfied, and (2) large changes in both the control and state vectors may take place at each iteration until these vectors are close to their limiting values, thereby greatly accelerating convergence during the early states. The limiting convergence rate is quadratic, as expected in Newton's method.

Another way of looking at this method for nonlinear problems is that at each iteration we get an admissible and optimal trajectory which satisfies a linear recursion relation which differs to some extent from the true nonlinear recursion relation. At each iteration the amount by which the linearization is in error decreases, so that in the limit the trajectory obtained is an *optimal* solution to the *linearized* problem obtained by linearizing about the limiting trajectory. Since it is the recursion relation which is linearized, the limiting trajectory is the optimal solution to a control problem described by linear recursion relations. It therefore follows that for the class of discrete nonlinear problems considered, the optimal solution has the properties of a solution to a discrete problem with linear recursion relations.

The requirement that the state vectors satisfy specified constraints usually increases the difficulty of the optimal control problem (see, for example, [5] and [13, Chap. 6]). In the approach used here to solve the state-constrained discrete problem, the convergence proof uses the fact that the state vector at each discrete time belongs to a convex compact set. In this sense then, the liability of the state-constrained problem has now become an asset. The existence of state constraints also introduces a symmetry into the problem, so that the usual sharp distinction between the

(independent) control vectors and (dependent) state vectors largely disappears.

The method described here applies to a recursion relation in the form of a system of inequalities, and might represent a finite difference approximation to a system of differential *inequalities*. By the use of a modified objective function, the problem usually considered corresponding to a system of differential equations can be handled. The "classical" two-point boundary value problem can also be solved in this fashion by allowing the control vector to represent the error in the difference equations and minimizing this error.

It should be emphasized that while the convexity assumption is needed in order to prove convergence, the computational method can be applied even when this assumption is not satisfied. In many such cases the iterative method will still converge, and if so, the trajectory obtained will satisfy the necessary conditions for an optimal trajectory. Furthermore, at each iteration a linear constraint minimization problem with either a convex or linear function is solved. Because of this, the method will almost always converge to a trajectory, which is at least a local minimum of the objective function, rather than an arbitrary stationary trajectory. It should also be mentioned that the method considered here requires only the Jacobian matrix (first partial derivatives) of the system equations, and does not need the Hessian matrix (second partial derivatives) as required by some other computational schemes [6], [10], [12]. For many nonlinear problems this may permit a great reduction in the computation required.

While the iterative method described was developed for problems arising in control theory, it may also be used to solve any finite-dimensional constrained minimization problem of the general type considered. In this respect the method is also a contribution to the solution of nonconvex mathematical programming problems.

2. Problem formulation. The discrete optimal control problem we shall consider here is to determine $m + 1$ state vectors $x_i^* \in E^n$ and m control vectors $u_i^* \in E^r$ which satisfy (2.2), (2.3) and (2.4) and such that

$$\sum_{i=0}^{m-1} \sigma(x_i^*, u_i^*) = \min \sum_{i=1}^{m-1} \sigma(x_i, u_i) \tag{2.1}$$

for all vectors x_i and u_i that satisfy the recursion relation

$$x_{i+1} - x_i = f(x_i, u_i), \qquad i = 0, 1, \cdots, m-1, \tag{2.2}$$

with

$$u_i \in U_i \subset E^r, \qquad i = 0, 1, \cdots, m-1, \tag{2.3}$$

and

$$x_i \in X_i \subset E^n, \qquad i = 0, 1, \cdots, m. \tag{2.4}$$

The subsets X_i and U_i are assumed to be compact and convex. We assume that σ is a convex function from each direct product $X_i \times U_i$ to E^1. We also assume that f is a function from each $X_i \times U_i$ to E^n. An additional assumption on the differentiability and convexity of the components of f will be needed later. It should be mentioned that the results obtained actually hold (with obvious modification) for the more general case where σ and f may depend explicitly on the index i. When the discrete problem is obtained from a continuous problem, this corresponds to the explicit dependence of σ and f on time. However, in order to avoid the complication of additional subscripts we will limit consideration to the simpler case.

A discrete problem of this type may arise directly, or it may arise as a finite difference approximation to a continuous system. For example, suppose that in the original continuous system we wish to determine a control $u(t)$ with range $U(t)$ for each $t \in [0, T]$, and a trajectory $x(t)$ with range $X(t)$ for each $t \in [0, T]$, such that the functional

$$\int_0^T \bar{\sigma}(x(t), u(t))\, dt \tag{2.5}$$

is minimized, and $x(t)$ and $u(t)$ satisfy the system of differential equations

$$\dot{x} = \bar{f}(x, u), \qquad t \in [0, T]. \tag{2.6}$$

The sum (2.1) then represents the simplest approximation to the integral (2.5), and the recursion relation (2.2) the simplest finite difference approximation to the system (2.6), if we let $\Delta t = T/m$, $\sigma = \Delta t\, \bar{\sigma}$, and $f = \Delta t\, \bar{f}$. The form of (2.2) may be retained even when more sophisticated finite difference schemes are used to approximate (2.6), but the relationship between f and $\bar{f}$ will become more complicated. The use of a more accurate implicit finite difference scheme when f is linear has been considered in [14]. It should be emphasized that in this paper we solve the discrete problem for a fixed value of m, and that we are interested in convergence (for fixed m) to an exact solution of the nonlinear discrete problem. The convergence to the solution of the continuous problem as $m \to \infty$ will not be considered here.

In order to show convergence of the iterative procedure we will consider the discrete system (2.1), (2.3) and (2.4), with (2.2) replaced by the system of inequalities

$$x_{i+1} - x_i \leqq f(x_i, u_i), \qquad i = 0, 1, \cdots, m - 1. \tag{2.7}$$

Such a system of inequalities may arise as a discrete approximation to a

system of differential inequalities of the form $\dot{x} \leqq \bar{f}(x, u)$. On the other hand, if one really wants to solve (2.2), this is accomplished by obtaining an optimum solution to (2.7) with an appropriately modified objective function, as discussed at the end of this section.

In order to simplify notation we proceed as in [14], and denote a specific control $(u_0', u_1', \cdots, u_{m-1}')$ and corresponding trajectory $(x_0', x_1', \cdots, x_m')$ by a single vector $z \in E^s$, where $s = m(r + n) + n$. Thus, a solution to the discrete system is specified by the vector

$$z' = (x_0', x_1', \cdots, x_m', u_0', u_1', \cdots, u_{m-1}'). \tag{2.8}$$

We will also denote by $Z \subset E^s$ the direct product of the sets X_i and U_i, so that

$$Z = \prod_{i=0}^{m} X_i \times \prod_{i=0}^{m-1} U_i . \tag{2.9}$$

Since the sets X_i and U_i are convex and compact, Z is also convex and compact. We can now represent the objective by means of the function

$$\phi(z) = \sum_{i=0}^{m-1} \sigma(x_i, u_i). \tag{2.10}$$

It follows from our assumption concerning σ that $\phi(z)$ is convex on Z. Finally we represent the $l = mn$ equations (2.2) or inequalities (2.7) by means of a function $v(z)$ from E^s to E^l. We let

$$v_{i,j} = f_j(x_i, u_i) + x_{i,j} - x_{i+1,j}, \qquad i = 0, 1, \cdots, m-1, \quad j = 1, \cdots, n. \tag{2.11}$$

The equations (2.2) are then given by $v(z) = 0$, and the inequalities (2.7) by $v(z) \geqq 0$. In this notation we can restate our problem (2.1), (2.3), (2.4) and (2.7) as follows:

$$\phi(z^*) = \min_z \{\phi(z) \mid z \in Z, v(z) \geqq 0\}. \tag{2.12}$$

Some remarks on the nature of the admissible set

$$S = \{z \mid z \in Z, v(z) \geqq 0\}$$

are in order here. The set Z is by assumption convex and compact, and in fact will usually be a polyhedral set in E^s. The admissible set corresponding to the original discrete problem (2.2), (2.3) and (2.4) is given by

$$S_1 = \{z \mid z \in Z, v(z) = 0\}.$$

The set S_1 is convex only if $v(z)$ is linear in z, that is, $f(x, u)$ is linear in x and u. If one or more components of f are nonlinear in x or u, the set S_1 is

nonconvex. For a general nonlinear function $f(x, u)$, the set S is also nonconvex. The iterative procedure of the following sections can be applied to such problems and will, in fact, often converge. However, there is no guarantee in the case of a general nonlinear f that the procedure will always converge. In order to prove convergence we require that each component of $v(z)$ be a *convex* function. It should be emphasized that this is *not* the requirement which makes S a convex set (except in the limiting case where $v(z)$ is linear). The set S is convex if each component of $v(z)$ is a *concave* function. Thus the convergence argument holds for the minimization of a convex function over a certain kind of nonconvex region.

If we actually want to satisfy (2.2) we must obtain a solution to the problem $\phi(z^*) = \min_{z \in S_1} \phi(z)$; that is, we require $v(z^*) = 0$. In order to achieve this and still solve a problem in the form of (2.12) we let

$$\varphi(z) = \phi(z) + \alpha \sum_{i,j} v_{i,j}, \tag{2.13}$$

where α is a sufficiently large positive constant. Since each component $v_{i,j}$ is a convex function, $\bar{\phi}(z)$ is a convex function. We then solve $\min_{z \in S} \bar{\phi}(z)$, which is in the form of (2.12). It is shown in the Appendix that provided the constraint set S satisfies a certain condition (essentially the same condition which insures convergence) there will always exist a value of α such that any local minimum of $\bar{\phi}(z)$ for $z \in S$ is also a local minimum of $\phi(z)$ for $z \in S_1$.

We are now able to describe the iterative method for solving the discrete optimal control problem in terms of the (in general, nonconvex) mathematical programming problem (2.12).

3. Linearized problem. Let Z be a compact convex subset of E^s, and $v(z)$ be a function from Z to E^l with $v \in C^2(Z)$. We assume that for some $z^0 \in Z$ we have $v(z^0) > 0$ and define a subset of E^s by

$$S = \{z \mid z \in Z, v(z) \geqq 0\}. \tag{3.1}$$

Since $z^0 \in S$, the set S is not empty. Also since S is a closed subset of Z it is compact but, in general, not convex (see Fig. 1).

If we let $v_z(y)$ be the $l \times s$ Jacobian matrix of v evaluated at $z = y$, we can define for each fixed $y \in Z$ the linear function on Z,

$$w(z, y) = v(y) + v_z(y)[z - y]. \tag{3.2}$$

For each $y \in Z$ we obtain a subset of E^s given by

$$W(y) = \{z \mid w(z, y) \geqq 0\}. \tag{3.3}$$

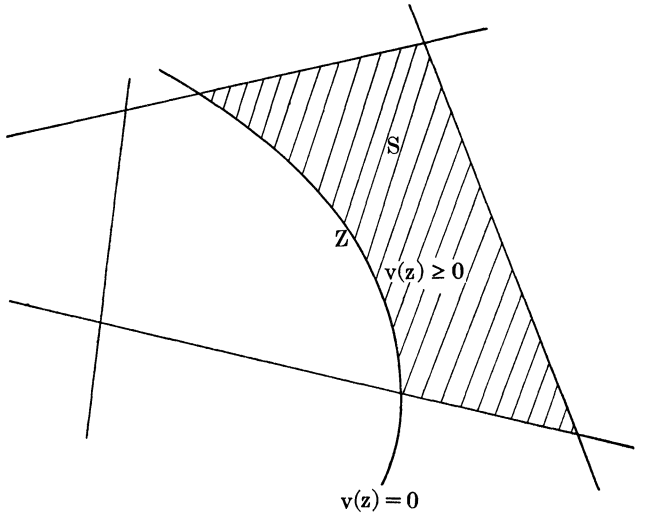

FIG. 1. *The convex set Z and subset S*

Now we consider the point-to-set mapping

$$\Gamma : Z \to Z, \tag{3.4}$$

given by

$$\Gamma y = W(y) \cap Z. \tag{3.5}$$

This is illustrated in Fig. 2.

THEOREM 1. *The set* Γy *is compact and convex. Furthermore, if each component of* $v(z)$ *is convex on* Z, *then for each* $y \in S$,

$$y \in \Gamma y \subset S. \tag{3.6}$$

Proof. For each y, the set $W(y)$ is the intersection of l halfspaces, a closed convex set. Therefore the intersection of $W(y)$ and the compact convex set Z is compact and convex. Next we note that since $y \in S$,

$$w(y, y) = v(y) \geqq 0, \tag{3.7}$$

so that $y \in W(y)$. Then since $y \in Z$, we have $y \in \Gamma y$.

Furthermore, by the convexity of $v(z)$, we have for any $(y, z) \in S \times S$,

$$v(z) \geqq v(y) + v_z(y)[z - y] = w(z, y). \tag{3.8}$$

Then for each $z \in W(y)$,

$$v(z) \geqq w(z, y) \geqq 0, \tag{3.9}$$

so that for every $z \in W(y) \cap Z$ we have $z \in S$, or $\Gamma y \subset S$.

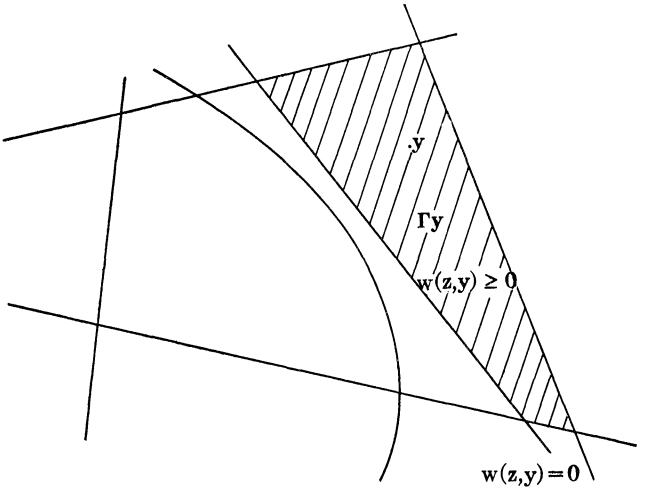

FIG. 2. *The convex subset* $\Gamma y \subset S$ *for* $y \in S$

Directly from (3.6) we get the following.

COROLLARY. Γy *maps* S *onto* S.

The constraints for the problem have now been defined in terms of the convex subset Z and the function $v(z)$. The objective function is given by a function $\phi(z)$ from Z to E^1 which is continuous and convex on Z. The iterative procedure, starting with an initial point $y^0 \in S$ can now be stated in a concise form. A sequence $\{y^j\}$ is obtained which satisfies

$$\phi(y^{j+1}) = \min_{z \in \Gamma y^j} \phi(z), \qquad j = 0, 1, \cdots . \tag{3.10}$$

Such a sequence is obtained by solving a well behaved convex constrained minimization problem with $z \in \Gamma y^j$, to get the minimum $\phi(y^{j+1})$ at a point $y^{j+1} \in \Gamma y^j$. The convexity of the subset Γy^j and the function $\phi(z)$ insure that a global minimum of $\phi(z)$ for $z \in \Gamma y^j$ is attained at $z = y^{j+1}$.

Suppose that the sequence $\{y^j\}$ converges to a limit point y^*. We would like to be able to state that the point y^* is the optimum solution to the partially linearized problem obtained by linearizing the constraints $v(z) \geqq 0$, about $z = y^*$. That is, we want

$$\phi(y^*) = \min_{z \in \Gamma y^*} \phi(z). \tag{3.11}$$

In terms of the original discrete optimal control problem (2.1), (2.3), (2.4) and (2.7), this is equivalent to the statement that the control $u_i^*, i = 0, 1, \cdots, m-1$, and trajectory $x_i^*, i = 0, 1, \cdots, m$, give an

optimal solution to the problem obtained by linearizing (2.7) about u_i^* and x_i^*.

However, without some further assumption, the relationship (3.11) may not hold. This is shown by the following simple two-dimensional example. Let

$$Z = \{z \mid 0 \leqq z_1 \leqq 1, 0 \leqq z_2 \leqq 1\} \tag{3.12}$$

and

$$v(z) = 4(z_1 - \tfrac{1}{2})^2 - z_2\,, \tag{3.13}$$

so that the feasible set S is given by

$$S = \{z \mid 4(z_1 - \tfrac{1}{2})^2 - z_2 \geqq 0, 0 \leqq z_1 \leqq 1, 0 \leqq z_2 \leqq 1\}. \tag{3.14}$$

This is illustrated in Fig. 3. Also let $\phi(z) = z_1$. We have

$$w(z, y) = v(y) + 8(y_1 - \tfrac{1}{2})(z_1 - y_1) - (z_2 - y_2), \tag{3.15}$$

so that for $y^0 = (1, 0)$ we get

$$\Gamma y^0 = \{z \mid 4z_1 - z_2 - 3 \geqq 0, 0 \leqq z_1 \leqq 1, 0 \leqq z_2 \leqq 1\}. \tag{3.16}$$

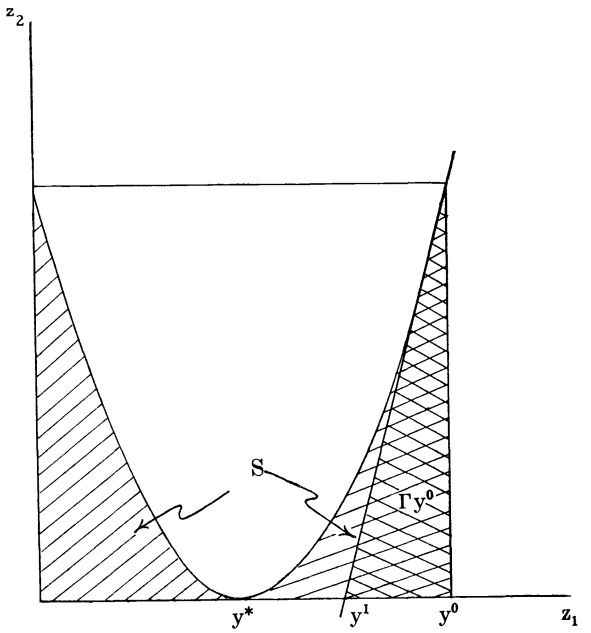

FIG. 3. *Two-dimensional example*

The solution to (3.10) for $j = 0$ is easily seen (from Fig. 3) to be $y^1 = \frac{3}{4}$. The sequence $\{y^j\}$ obtained in this way converges to $y^* = (\frac{1}{2}, 0)$, with $\phi(y^*) = \frac{1}{2}$. But Γy^* is the interval [0, 1] on the z_1 axis, so that $\min_{z \in \Gamma y^*} \phi(z) = 0$, and is attained at $z = (0, 0) \neq y^*$.

In order that the limit point y^* always satisfy (3.11) it is sufficient that the mapping Γy be continuous. The mapping Γy is continuous (both upper and lower semicontinuous) if for any point $y^1 \in S$ and any point $y^2 \in S$ in the neighborhood of y^1, there is *some* point of Γy^1 close to *each* point of Γy^2. The continuity of Γy follows from two assumptions we make concerning the set S.

(1) For each $y \in S$, the Jacobian matrix $v_z(y)$ has full row rank, that is, rank $= l \leqq s$.

(2) For each $y \in S$, the convex set Γy contains interior points.

These two assumptions are essentially the Kuhn-Tucker constraint qualification for the set S (see, for example, [2]). The proof that (1) and (2) imply the continuity of Γy is given in the Appendix. A slightly stronger assumption than (2), which however involves only the rank of an augmented Jacobian matrix, is also given there.

The difficulty in the previous two-dimensional example occurs because the assumption (2) above is not satisfied. In particular, for $y^* = (\frac{1}{2}, 0)$, Γy^* is just the interval [0, 1]. As a result the mapping Γy is not continuous in the neighborhood of y^*.

The first assumption above is always satisfied when the function $v(z)$ is defined by (2.11), as shown in the following.

LEMMA. *If $v(z)$ corresponds to the discrete recursion relation, as given by* (2.11), *then assumption* (1) *is satisfied.*

Proof. Directly from (2.11) we have that

$$\frac{\partial v_{i,j}}{\partial x_{i+1,j}} = -1,$$

$$\frac{\partial v_{i,j}}{\partial x_{i+1,p}} = 0, \quad p \neq j, \tag{3.17}$$

$$\frac{\partial v_{i,j}}{\partial x_{q,p}} = 0, \quad q > i + 1, \quad p = 1, \cdots, n,$$

for $i = 0, 1, \cdots, m - 1; j = 1, \cdots, n$. Therefore the Jacobian matrix v_z contains a square $(mn \times mn)$ lower triangular matrix with elements -1 along its diagonal. Since such a matrix is nonsingular and since v_z has mn rows, v_z has full row rank.

4. Convergence of iterative procedure. The iterative procedure will now be considered in more detail. We again consider the convex function ϕ from

Z to E^1, with $\phi \in C^1(Z)$. Since S is compact, $\phi(z)$ is bounded and attains its minimum for $z \in S$. In particular, let

$$\mu = \min_{z \in S} \phi(z). \tag{4.1}$$

For each $y \in S$, the set Γy is compact so that the minimum of $\phi(z)$ for $z \in \Gamma y$ is attained. We let

$$\Psi(y) = \min_{z \in \Gamma y} \phi(z). \tag{4.2}$$

We now show that because of the continuity of Γy, the function $\Psi(y)$ is continuous for $y \in S$.

LEMMA. *$\Psi(y)$ is continuous for $y \in S$.*

Proof. For $y^1 \in S$, let $\Psi(y^1)$ be attained at $z^1 \in \Gamma y^1$, that is $\Psi(y^1) = \phi(z^1)$. Now choose $y^2 \in S$ close to y^1, and let $\Psi(y^2)$ be attained at z^2, so that $\Psi(y^2) = \phi(z^2)$. Suppose $\phi(z^2) \leqq \phi(z^1)$. Now by the continuity of Γy we can choose $\bar{z}^1 \in \Gamma y^1$ close to z^2. Then by the continuity of $\phi(z)$ we have $\phi(\bar{z}^1)$ close to $\phi(z^2)$. But since $\phi(z^1) \leqq \phi(z)$ for every $z \in \Gamma y^1$, we have

$$\phi(z^2) \leqq \phi(z^1) \leqq \phi(\bar{z}^1), \tag{4.3}$$

so that $\phi(z^1)$ is close to $\phi(z^2)$.

A similar argument holds for $\phi(z^1) \leqq \phi(z^2)$.

Starting with $y^0 \in S$ we generate a sequence of vectors $\{y^j\}$ as follows:

$$\phi(y^{j+1}) = \min_{z \in \Gamma y^j} \phi(z), \qquad j = 0, 1, \cdots. \tag{4.4}$$

Note that if Z is a polyhedral set then Γy^j is a polyhedral set determined by specified linear inequalities. Furthermore, $\phi(z)$ is a convex function, so that for each y^j we solve a straightforward convex programming problem with linear constraints.

THEOREM 2. *Every vector of the sequence $\{y^j\}$ is in S. The corresponding sequence of values $\{\phi(y^j)\}$ is monotonically decreasing. The sequence $\{y^j\}$ contains a convergent subsequence converging to a point $y^* \in S$ such that*

$$\mu \leqq \phi(y^*) \leqq \phi(y^j), \qquad j = 0, 1, \cdots, \tag{4.5}$$

and

$$\phi(y^*) = \min_{z \in \Gamma y^*} \phi(z). \tag{4.6}$$

Proof. By Theorem 1, we have $y^j \in \Gamma y^j \subset S$, so that each y^j is in S. Also since $y^j \in \Gamma y^j$ we must have

$$\phi(y^{j+1}) = \min_{z \in \Gamma y^j} \phi(z) \leqq \phi(y^j), \tag{4.7}$$

so that $\{\phi(y^j)\}$ is monotonically decreasing.

Since S is bounded the sequence $\{y^j\}$ contains a convergent subsequence. Let y^* be the limit point of such a convergent subsequence. Since S is compact, $y^* \in S$, and $\phi(y^*) \geqq \mu$. Furthermore, from the monotonicity of the sequence $\{\phi(y^j)\}$ the relation (4.5) must hold.

To demonstrate (4.6), we observe that since $y^* \in S$, we have $y^* \in \Gamma y^*$, so that

$$\Psi(y^*) = \min_{z \in \Gamma y^*} \phi(z) \leqq \phi(y^*). \tag{4.8}$$

Now suppose that $\Psi(y^*) < \phi(y^*)$. Then by the continuity of $\Psi(y)$ we can pick k sufficiently large so that $\Psi(y^k) < \phi(y^*)$. But from (4.2) and (4.4) we have $\phi(y^{k+1}) = \Psi(y^k)$, so that $\phi(y^{k+1}) < \phi(y^*)$, which contradicts (4.5). Therefore we must have $\Psi(y^*) = \phi(y^*)$.

THEOREM 3. *Let y^* be a limit point of $\{y^j\}$. Then y^* is the global minimum of the partially linearized problem about the point y^*. Furthermore, the optimality conditions (the Kuhn-Tucker necessary conditions) which must be satisfied at a global minimum of the problem* (2.12) *are, in fact, satisfied at y^*.*

Proof. The set Γy^* is the intersection of Z and the convex set $W(y^*)$ obtained by linearizing the constraints $v(z) \geqq 0$, about $z = y^*$. It follows immediately from (4.6) that y^* is a global optimum solution to this partially linearized problem.

As mentioned in the previous section, the assumptions (1) and (2) on the set S are equivalent to the Kuhn-Tucker constraint qualification. It is shown in their original paper [11] that with this qualification the optimum solution z^* to a general nonlinear problem has the property that the gradient $\nabla\phi(z^*)$ must belong to the convex cone of inward normals to the active constraints at z^*. The solution y^* to the partially linearized problem about y^* will, of course, also have this property. Therefore, $\nabla\phi(y^*)$ belongs to the convex cone of inward normals to the active constraints at y^*, i.e., the Kuhn-Tucker necessary conditions for a global minimum are satisfied at y^*.

5. Computational solution. The computational solution of the nonlinear discrete optimal control problem (2.1)–(2.4) is considered in this section. We will assume that the convex compact sets U_i and X_i are convex polytopes defined by specified linear inequaiities (see Appendix). In order to apply the computational method we need only make the additional assumption that the functions $\sigma(x, u)$ and $f(x, u)$ are of class C^1 on each $X_i \times U_i$. However, in order to insure the validity of the convergence proof (Theorem 2) we must make an additional assumption concerning f and an assumption about the linear inequalities defining the X_i and U_i. We assume that each component f_j of f, $j = 1, \cdots, n$, is either convex or concave on $X_i \times U_i$.

For $i = 0, 1, \cdots, m-1$ and $j = 1, \cdots, n$ we let

$$\bar{v}_{i,j} = f_j(x_i, u_i) + x_{i,j} - x_{i+1,j},$$

$$v_{i,j} = \begin{cases} \bar{v}_{i,j} \text{ for } f_j \text{ convex on } X_i \times U_i, \\ -\bar{v}_{i,j} \text{ for } f_j \text{ concave on } X_i \times U_i. \end{cases} \tag{5.1}$$

The function $v(z)$, with components $v_{i,j}$, is thus a convex function on Z. Furthermore, the equations (2.2) are now equivalent to $v(z) = 0$.

As discussed in the Appendix the linear inequalities which define the X_i and U_i are specified in terms of the vector z by $a_i'z - b_i \geqq 0$, $i = 1, \cdots, k$, giving the polyhedral set Z. We make the following assumption about these linear inequalities. Let $\bar{y} \in S$ be a boundary point of Z, i.e., $v(\bar{y}) = 0$, and $a_i'\bar{y} - b_i = 0$, $i = 1, \cdots, \bar{k}$. Then the $(l + \bar{k}) \times s$ matrix consisting of $v_z(\bar{y})$ augmented by the rows a_i', $i = 1, \cdots, \bar{k}$, is of full row rank $(= l + \bar{k})$. According to the Lemma at the end of §3, $v_z(y)$ is always of full row rank, so this assumption is essentially a condition on the vectors a_i. As shown in the Appendix it follows from the full rank condition that Γy is a continuous mapping. The convergence proof of Theorem 2 is applicable because $v(z)$ is convex and Γy is continuous.

At each iteration we wish to solve a mathematical programming problem of the form,

$$\min_z \{\phi(z) \mid a_i'z - b_i \geqq 0, i = 1, \cdots, k; w(z, y) \geqq 0\}. \tag{5.2}$$

This is a linear constraint problem with $m(r + n) + n$ variables and $k + l$ constraints. For small problems a direct computational solution of (5.2) causes no difficulty. In many practical cases however, the number of state variables is greater than the number of control variables, i.e., $r < n$. In such a case there is a considerable computational advantage in treating the linearized problem (5.2) as the linear problem was treated in [14]. In effect, the linear relations $w(z, y) = 0$ are used to solve explicitly for the vectors x_i, $i = 1, \cdots, m$, in terms of x_0 and the u_i, $i = 0, 1, \cdots, m-1$. Substitution for the vectors x_i in $\phi(z)$ and the inequalities $a_i'z - b_i \geqq 0$ reduces the original problem (5.2) to one in only $mr + n$ variables. This reduced problem may then be solved by an appropriate linear constraint method which takes advantage of the particular form of ϕ. For example, if ϕ is quadratic, a quadratic programming method may be used.

In the important case where ϕ is linear, a further efficiency is made possible by treating the reduced problem as the dual problem, and solving the corresponding primal linear programming problem. This permits us to take advantage of the fact that the variables of the reduced problem (the control variables) are not required to be nonnegative, and that there are

more inequality constraints than variables. The corresponding primal problem consists of $mr + n$ equations in $mn + k$ nonnegative variables. The numerical examples discussed below are of this type.

The use of the linear equality relations $w(z, y) = 0$ has the additional computational advantage that no modification of the true objective function is required. On the other hand a possible theoretical difficulty may arise since even with $v(z)$ convex it is usually not true that $y^j \in \Gamma y^j$ when Γy is determined by $w(z, y) = 0$. Thus the monotone behavior of $\phi(y^j)$ is not guaranteed. However, no such difficulty has been observed in the actual numerical calculations.

In order to illustrate the application of the iterative method we will discuss two numerical solutions to a nonlinear problem. The problem considered is a discrete approximation to the following continuous scalar $(n = 1)$ problem:

$$\min \int_0^1 u(t)\, dt,$$

subject to $\dot{x} = f(x, u)$, $|u(t)| \leqq 1$, for $t \in [0, 1]$, and $x(0) = 1$, $x(1) = \frac{1}{2}$, where $f(x, u) = -\frac{3}{2}x + x^2 + u(t)$. An additional state constraint is imposed in the second example. The initial trajectory used to start the iteration was $x^0(t) = 1$, for $t \in [0, 1]$.

For these examples the simplest (forward) finite difference scheme was used, namely,

$$x_{i+1} - x_i = \Delta t f(x_i, u_i), \qquad i = 0, 1, \cdots, m - 1, \tag{5.3}$$

so that

$$\begin{aligned} v_i = \Delta t f(x_i, u_i) + x_i - x_{i+1} &= (1 - \tfrac{3}{2}\Delta t)x_i + \Delta t(x_i)^2 \\ &\quad + \Delta t u_i - x_{i+1}, \qquad i = 0, 1, \cdots, m - 1. \end{aligned} \tag{5.4}$$

For x_i^j known, the linearized system which must be satisfied by x_i^{j+1} and u_i^{j+1} is

$$\begin{aligned} w_i &= -x_{i+1}^{j+1} + [1 + \Delta t(2x_i^j - \tfrac{3}{2})]x_i^{j+1} + \Delta t u_i^{j+1} - \Delta t(x_i^j)^2 \\ &= 0, \qquad i = 0, 1, \cdots, m - 1. \end{aligned} \tag{5.5}$$

This system is solved using the specified initial value for $x(0)$ to give the x_i^{j+1} explicitly as linear functions of the u_i^{j+1},

$$x_i^{j+1} = d_i^{j+1}(u_{i-1}^{j+1}, \cdots, u_0^{j+1}), \qquad i = 1, \cdots, m. \tag{5.6}$$

The following linear programming problem (in the dual form) is then solved at each iteration to give the new optimal control u_i^{j+1},

$i = 0, 1, \cdots, m-1$:

$$\min_{u_i} \left\{ \sum_{i=0}^{m-1} u_i \mid -1 \leqq u_i \leqq 1, i = 0, 1, \cdots, m-1; \right.$$
$$\left. \tfrac{1}{2} \leqq d_m{}^{j+1}(u_{m-1}, u_{m-2}, \cdots, u_0) \leqq \tfrac{1}{2} \right\}. \tag{5.7}$$

The corresponding state trajectory $x_i{}^{j+1}$, $i = 1, \cdots, m$, is then given by (5.6).

The iteration was started with $x_i{}^0 = 1$, $i = 0, 1, \cdots, m$, and a value of $m = 20$ ($\Delta t = 0.05$) was used. The results for the first numerical example are shown in Figs. 4 and 5. Convergence was achieved (within the desired accuracy) in three iterations. However, the difference between x^2 and $x^* = x^3$ is too small to be shown graphically (Fig. 4). Note the rapid convergence even though the initial guess, $x_i{}^0$, for the trajectory was very poor and did not even satisfy the terminal boundary condition. The corresponding optimal control $u_i{}^*$ is shown in Fig. 5. The monotone behavior of the function value is verified by the successive values of $\phi^j = \sum_{i=0}^{m-1} u_i{}^j$. These were $\phi^1 = -0.286$, $\phi^2 = -0.946$, and $\phi^3 = -0.950$.

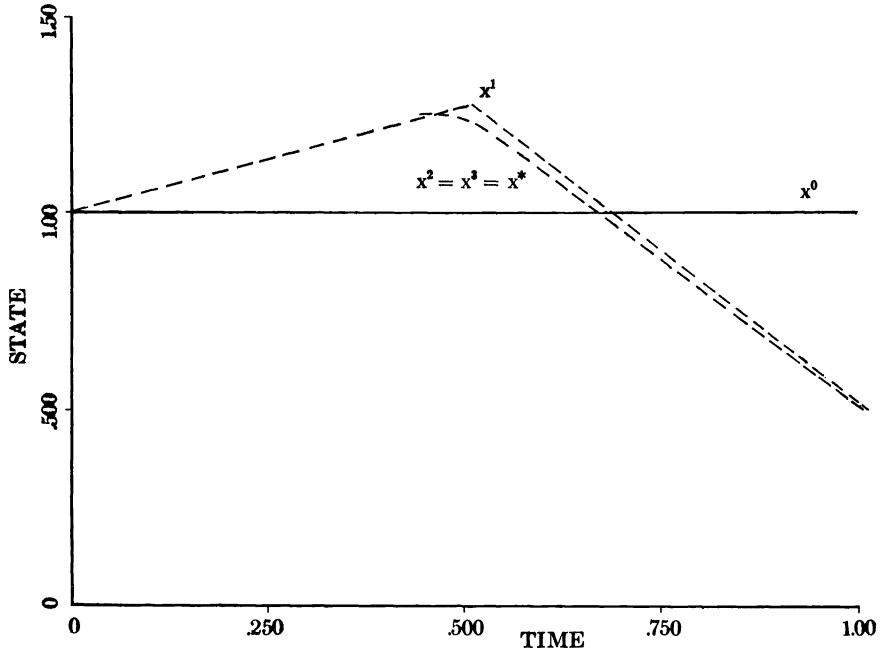

FIG. 4. *Initial and optimal state trajectories for nonlinear numerical example*

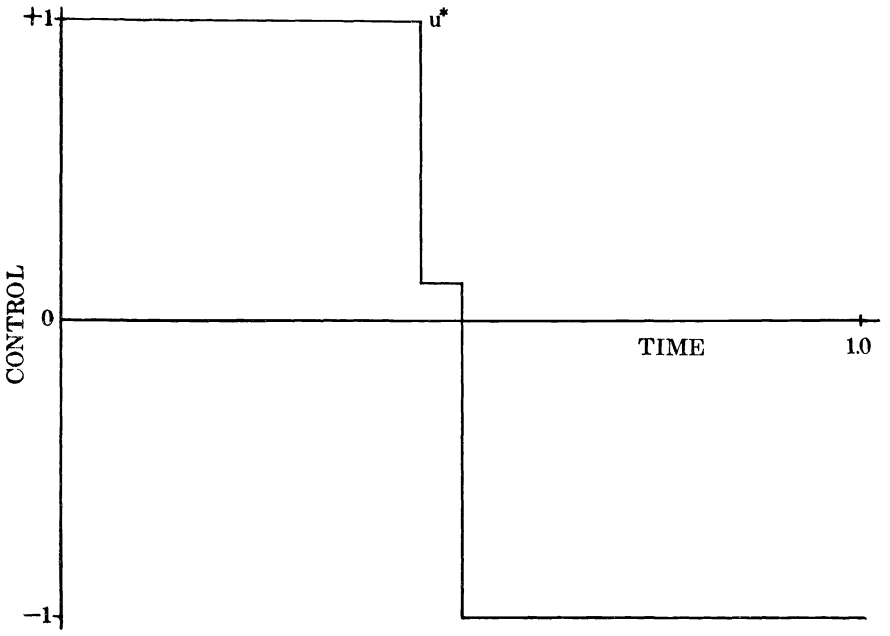

FIG. 5. *Optimal control for nonlinear numerical example*

For the second example the state constraint, $x(\frac{1}{2}) \leqq -\frac{1}{2}$, was imposed. This of course eliminates the solution shown in Fig. 4. The sequence of 5 state trajectories obtained is shown in Fig. 6. The corresponding function values were $\phi^1 = 2.792$, $\phi^2 = -0.144$, $\phi^3 = -0.656$, $\phi^4 = -0.972$, and $\phi^5 = -1.008$. The control from the first iteration u_i^1 and the optimal control u_i^* are shown in Fig. 7. All of the state trajectories (except for the initial guess) are seen to satisfy the state constraints. It is interesting to observe that the method not only converges to a different trajectory x_i^* but that the added state constraint is not active for this limit trajectory. Thus the state constraint forces the solution away from its previous sequence and allows it to converge to a different local minimum. On the other hand, in some other nonlinear state-constrained cases which have been computed by this method, a state inequality constraint of the type imposed here has remained active for the limiting trajectory. Finally, it should be noted that for both cases the limiting control has the properties of an optimal control for a discrete *linear* problem, that is, $n(=1)$ switchings and $m - n(=19)$ values of $u_i^* = \pm 1$.

Appendix. In this Appendix we prove that the assumptions (1) and (2) of §3 imply the continuity of Γy. We also show the validity of the modified objective function (2.13).

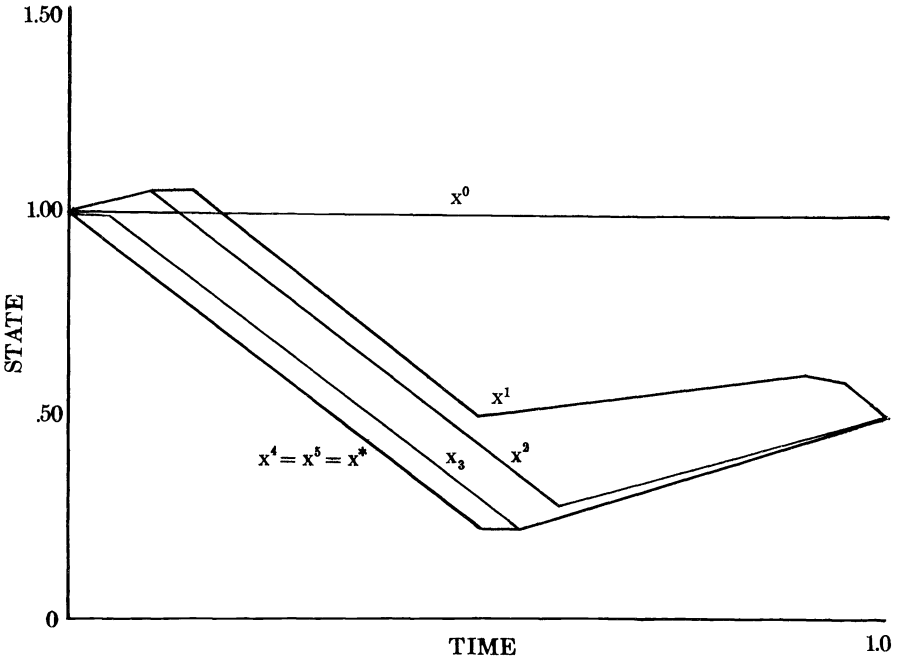

FIG. 6. *State trajectories for nonlinear example with added state constraint*

We first state a condition on the rank of an augmented Jacobian matrix which insures the satisfaction of the assumption (2) of §3. In order to state this condition we must have an explicit statement of the constraints which define the compact set Z.

We will assume that Z is the polyhedral set determined by the system of k linear inequalities

$$a_i'z - b_i \geqq 0, \qquad i = 1, \cdots, k, \tag{A.1}$$

or

$$Z = \{z \mid A'z - b \geqq 0\}, \tag{A.2}$$

where A is an $s \times k$ matrix with specified columns a_i, and $b \in E^k$ is specified. Let $\bar{z}$ denote a boundary point of Z. Then we must have at least one active constraint at $\bar{z}$, that is, $a_i'\bar{z} - b_i = 0$ for at least one value of i. We will denote by $\bar{A}(z)$ the matrix whose columns represent the active constraints at z. Similarly, let $\bar{V}'(z)$ represent the Jacobian matrix of the vector $\bar{v}(z)$ which contains all components of $v(z)$ for which $v_i(z) = 0$. That is, $\bar{v}(z) = 0$, and $\bar{V}'(z) = \bar{v}_z(z)$.

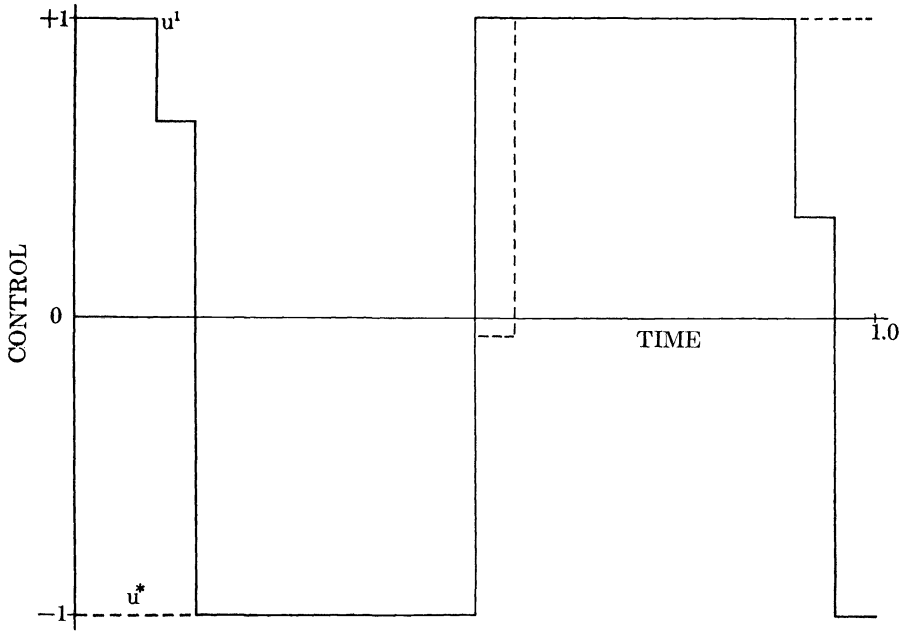

FIG. 7. *Controls for nonlinear example with added state constraint*

We will denote the boundary points of S by ∂S. It follows that for every $y \in \partial S$, the matrix

$$\bar{B}(y) = [\bar{V}(y) \quad \bar{A}(y)] \tag{A.3}$$

is defined and has at least one column. We will say that $\bar{B}(y)$ satisfies the *full rank condition* at $y \in \partial S$ if the columns of $\bar{B}(y)$ are linearly independent.

Assumption (1) implies that $\bar{B}(y)$ satisfies the full rank condition at every $y \in \partial S$ which is also interior to Z. This is true because for such a point $\bar{A}(y) = 0$, and $\bar{V}(y)$ certainly has full column rank since it consists of selected columns of v_z'. Furthermore, assumption (2) is implied by the full rank condition on $\bar{B}$, as shown by the following.

LEMMA. *Let $\bar{B}(y)$ satisfy the full rank condition for every $y \in \partial S$. Then for each $y \in S$, the convex set Γy contains interior points.*

Proof. First suppose $\bar{y} \in S$ is an interior point of S. Since $S \subset Z$, $\bar{y}$ is an interior point of Z. Furthermore, $w(\bar{y}, \bar{y}) = v(\bar{y}) > 0$, so that $\bar{y}$ is an interior point of $W(\bar{y})$. Therefore $\bar{y}$ is an interior point of $\Gamma\bar{y}$.

Now suppose $\bar{y} \in \partial S$. The set $\Gamma\bar{y}$ is the polyhedral set determined by the $k + l$ linear inequalities

$$\Gamma\bar{y} = \{z \mid w(z, \bar{y}) \geqq 0, A'z - b \geqq 0\}. \tag{A.4}$$

Now consider the point $z = \bar{y}$. We may assume without loss of generality that

$$w_i(\bar{y}, \bar{y}) = v_i(\bar{y}) \begin{cases} = 0, & i = 1, \cdots, \bar{l} \leqq l, \\ \geqq \epsilon, & i = \bar{l} + 1, \cdots, l, \end{cases} \tag{A.5}$$

and

$$a_i'\bar{y} - b_i \begin{cases} = 0, & i = 1, \cdots, \bar{k} \leqq k, \\ \geqq \epsilon, & i = \bar{k} + 1, \cdots, k, \end{cases} \tag{A.6}$$

for some $\epsilon > 0$. Then the columns of $\bar{V}(\bar{y})$ are the gradient vectors $\nabla v_i(\bar{y})$, $i = 1, \cdots, \bar{l}$, and the columns of $\bar{A}(\bar{y})$ are the vectors a_i, $i = 1, \cdots, \bar{k}$. Since $\bar{B}(\bar{y})$ satisfies the full rank condition, its columns are linearly independent and there exists no vector $r \in E^{\bar{k}+\bar{l}}$, except $r = 0$, such that $\bar{B}(\bar{y})r = 0$. Then by a variation on the Farkas lemma (see [8, Theorem 2.9, p. 48]), there exists a vector $\bar{z} \in E^s$ such that

$$\bar{z}'\bar{B}(\bar{y}) > 0. \tag{A.7}$$

Now consider the point

$$\bar{\bar{y}} = \bar{y} + \bar{\epsilon}\bar{z}, \tag{A.8}$$

where $\bar{\epsilon} > 0$ is chosen sufficiently small so that

$$\begin{aligned} \bar{\epsilon}\bar{z}'\nabla v_i(\bar{y}) &< \epsilon, \qquad i = \bar{l} + 1, \cdots, l, \\ \bar{\epsilon}\bar{z}'a_i &< \epsilon, \qquad i = \bar{k} + 1, \cdots, k. \end{aligned} \tag{A.9}$$

Now consider $w_i(\bar{\bar{y}}, \bar{y})$, $i = 1, \cdots, l$, and $a_i'\bar{\bar{y}} - b_i$, $i = 1, \cdots, k$. From (A.5), (A.6), (A.7) and (A.8) we have $w_i(\bar{\bar{y}}, \bar{y}) > 0$, $i = 1, \cdots, \bar{l}$, and $a_i\bar{\bar{y}} - b_i > 0$, $i = 1, \cdots, \bar{k}$. From (A.5), (A.6), (A.8) and (A.9) we have $w_i(\bar{\bar{y}}, \bar{y}) > 0$, $i = \bar{l} + 1, \cdots, l$, and $a_i'\bar{\bar{y}} - b_i > 0$, $i = \bar{k} + 1, \cdots, k$. Therefore, $\bar{\bar{y}}$ is interior to every constraint of $\Gamma\bar{y}$ and is an interior point of $\Gamma\bar{y}$.

THEOREM 4. *The mapping Γy is continuous for $y \in S$.*

Proof. Because $v(z) \in C^2$ on the compact set Z a uniform bound γ exists such that for any $(z, y^1, y^2) \in S \times S \times S$,

$$\|w(z, y^1) - w(z, y^2)\| \leqq \gamma\|y^1 - y^2\|. \tag{A.10}$$

Also since $v_z(y)$ is of rank l for $y \in S$, the symmetric matrix $v_z v_z'$ is positive definite at every point of S. Therefore a uniform bound β exists such that

$$\|(v_z v_z')^{-1}\| \leqq \beta^2 \tag{A.11}$$

for every $y \in S$.

Suppose we are given $y^1 \in S$ and $z^1 \in \Gamma y^1$. Then given any $\epsilon > 0$, we

now show that we can choose $\delta > 0$ so that, for each $y^2 \in S$ with $\|y^1 - y^2\| \leqq \delta$, we can find $z^2 \in \Gamma y^2$ such that $\|z^1 - z^2\| \leqq \epsilon$.

If $z^1 \in \Gamma y^2$, the theorem is true with $z^2 = z^1$. Now suppose $z^1 \notin \Gamma y^2$, that is, at least one component of $w(z^1, y^2)$ is negative. Without loss of generality we assume that $w_i(z^1, y^2) < 0$, for $i = 1, \cdots, k \leqq l$, and $w_i(z^1, y^2) \geqq 0$, for $i = k+1, \cdots, l$. Since $z^1 \in \Gamma y^1$, we have $w_i(z^1, y^1) \geqq 0$, for $i = 1, \cdots, l$. Let $\bar{w} \in E^l$ be the vector with $\bar{w}_i = w_i(z^1, y^2) < 0$, $i = 1, \cdots, k$, and $\bar{w}_i = 0$, $i = k+1, \cdots, l$. Then

$$|\bar{w}_i| \leqq |w_i(z^1, y^1) - w_i(z^1, y^2)|, \qquad i = 1, \cdots, l, \tag{A.12}$$

so that

$$\|\bar{w}\| \leqq \|w(z^1, y^1) - w(z^1, y^2)\| \leqq \gamma \|y^1 - y^2\|, \tag{A.13}$$

where the last inequality follows from (A.10).

We first assume z^1 is an interior point of Z. Then there is an ϵ_1 with $0 < \epsilon_1 \leqq \epsilon$, such that $z \in Z$ for $\|z - z^1\| \leqq \epsilon_1$. Choose $\delta = \epsilon_1/\beta\gamma$, and let y^2 be any point in S with $\|y^1 - y^2\| \leqq \delta$. Now choose $\bar{w}$ as above, and let

$$z^2 = z^1 - v_z'(y^2)[v_z(y^2)v_z'(y^2)]^{-1}\bar{w}. \tag{A.14}$$

From (3.2) we have

$$\begin{aligned} w(z^2, y^2) &= v(y^2) + v_z(y^2)[z^1 - y^2] - v_z(y^2)v_z'(y^2)[v_z(y^2)v_z'(y^2)]^{-1}\bar{w} \\ &= w(z^1, y^2) - \bar{w} \geqq 0, \end{aligned} \tag{A.15}$$

so that $z^2 \in W(y^2)$. Furthermore from (A.14) and (A.11) we have

$$\|z^2 - z^1\|^2 = \bar{w}'[v_z(y^2)v_z'(y^2)]^{-1}\bar{w} \leqq \beta^2\|\bar{w}\|^2. \tag{A.16}$$

Since $\|y^1 - y^2\| \leqq \epsilon_1/\beta\gamma$, we get from (A.16) and (A.13) that

$$\|z^2 - z^1\| \leqq \beta\|\bar{w}\| \leqq \beta\gamma\|y^1 - y^2\| \leqq \epsilon_1. \tag{A.17}$$

But this shows that $z^2 \in Z$, and therefore $z^2 \in \Gamma y^2$. Finally since $\epsilon_1 \leqq \epsilon$, we have $\|z^2 - z^1\| \leqq \epsilon$, as was to be shown.

The other possibility we must consider is that $z^1 \in \Gamma y^1$ is a boundary point of Z. Since Γy^1 has interior points and is a convex set there are interior points in the neighborhood of every point of Γy^1 (see, for example, [7]). In particular there exist ϵ_3, $0 < \epsilon_3 \leqq \epsilon/2$, and $z^3 \in \Gamma y^1$, such that $\|z^3 - z^1\| \leqq \epsilon/2$ and $\|z - z^3\| \leqq \epsilon_3$ implies that z is interior to Z. Now choose $\delta = \epsilon_3/\beta\gamma$, and replace z^1 by z^3 in the previous argument. This gives a point $z^2 \in \Gamma y^2$ with $\|z^2 - z^3\| \leqq \epsilon/2$. It follows that $\|z^2 - z^1\| \leqq \epsilon$.

We now prove the statement about the modified objective function (2.13) made at the end of §2. We define the $s \times (l + \bar{k})$ augmented Jacobian matrix $B(y) = [v_z'(y) \quad \bar{A}(y)]$. Let $\bar{\phi}(z)$ be as in (2.13).

THEOREM 5. *Let $B(y)$ have full column rank for every $y \in S$. Then a value of α exists such that every local solution of*

$$\min_z \{\bar{\phi}(z) \mid z \in Z, v(z) \geqq 0\} \tag{A.17}$$

is also a local solution of

$$\min_z \{\phi(z) \mid z \in Z, v(z) = 0\}. \tag{A.18}$$

Proof. Since $\phi \in C^1$ and $B(y)$ has full column rank on the compact set S, there are constants α_1 and ϵ_1 such that for any $y \in S$,

$$\|\nabla\phi(y)\| \leqq \alpha_1, \tag{A.19}$$

and

$$\|B(y)r\| \geqq \epsilon_1\|r\|, \quad r \in E^{l+\tilde{k}}. \tag{A.20}$$

We choose $\alpha > \alpha_1/\epsilon_1$. Let y^* be a local minimum of (A.17). Because of the rank condition on $B(y)$, the necessary Kuhn-Tucker conditions are satisfied at y^*. The relevant conditions are that there exist vectors $p \geqq 0$ and $q \geqq 0$ such that

$$v_z'(y^*)p + \bar{A}(y^*)q = \nabla\bar{\phi}(y^*) = \nabla\phi(y^*) + \alpha\sum_{i=1}^{l} \nabla v_i(y^*) \tag{A.21}$$

and

$$v_i(y^*)p_i = 0, \qquad i = 1, \cdots, l. \tag{A.22}$$

We let $r' = (p_1 - \alpha, \cdots, p_l - \alpha, q_1, \cdots, q_{\tilde{k}})$, and write (A.21) as

$$B(y^*)r = \nabla\phi(y^*). \tag{A.23}$$

From (A.19) and (A.20) it follows that

$$\epsilon_1\|r\| \leqq \|B(y^*)r\| = \|\nabla\phi(y^*)\| \leqq \alpha_1, \tag{A.24}$$

or $\|r\| \leqq \alpha_1/\epsilon_1$. But this requires $|\alpha - p_i| \leqq \alpha_1/\epsilon_1 < \alpha$, $i = 1, \cdots, l$, or $p_i > 0$, $i = 1, \cdots, l$. Then from (A.22) we must have $v_i(y^*) = 0$, $i = 1, \cdots, l$, so that y^* is a feasible solution of (A.18).

Now suppose y^* is not a local minimum of (A.18). Then for some point $y^1 \in Z$, arbitrarily close to y^*, we have $v(y^1) = 0$ and $\phi(y^1) < \phi(y^*)$. But then $\bar{\phi}(y^1) < \bar{\phi}(y^*)$, so that y^* is not a local minimum of (A.17).

REFERENCES

[1] R. ARIS, *Discrete Dynamic Programming*, Blaisdell, New York, 1964.

[2] K. J. ARROW, L. HURWICZ, AND H. UZAWA, *Constraint qualifications in maximization problems*, Naval Res. Logist. Quart., 8 (1961), pp. 175–191.

[3] R. E. BELLMAN, I. GLICKSBERG, AND O. A. GROSS, *Some aspects of the mathe-*

matical theory of control processes, R-313, The RAND Corporation, Santa Monica, 1958.

[4] R. E. Bellman and R. Kalaba, *Dynamic programming, invariant imbedding and quasilinearization: comparisons and interconnections*, Computing Methods in Optimization Problems, Balakrishnan and Neustadt, eds., Academic Press, New York, 1964, pp. 135–145.

[5] L. D. Berkovitz, *On control problems with bounded state variables*, J. Math. Anal. Appl., 5 (1962), pp. 488–498.

[6] A. E. Bryson, W. F. Denham, F. J. Carroll, and K. Mikami, *Lift or drag programs that minimize re-entry heating*, J. Aerospace Sci., 29 (1962), pp. 420–430.

[7] H. G. Eggleston, *Convexity*, Cambridge University Press, Cambridge, 1958, p. 9.

[8] D. Gale, *The Theory of Linear Economic Models*, McGraw-Hill, New York, 1960.

[9] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962, pp. 366–374.

[10] R. E. Kopp and R. McGill, *Several trajectory optimization techniques*, Computing Methods in Optimization Problems, Balakrishnan and Neustadt, eds., Academic Press, New York, 1964, pp. 65–89.

[11] H. W. Kuhn and A. W. Tucker, *Nonlinear programming*, Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1951, pp. 481–492.

[12] C. W. Merriam, *An algorithm for the iterative solution of a class of two-point boundary value problems*, this Journal, 2 (1964), pp. 1–10.

[13] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[14] J. B. Rosen, *Optimal control and convex programming*, Nonlinear Programming—A Course, J. Abadie, ed., North-Holland, Amsterdam, to appear.

J. SIAM CONTROL
Vol. 4, No. 1, 1966
Printed in U.S.A.

STEEPEST DESCENT WITH INEQUALITY CONSTRAINTS ON THE CONTROL VARIABLES*

RINALDO F. VACHINO†

A number of algorithms have been developed in the last few years for the iterative solution of variational problems. The methods of Bryson and Denham [2] and of Kelley et al. [12] have been applied by them and other authors to problems whose control and state variables are subject to inequality contraints.

The present study deals with a particular class of variational problems, those problems characterized by closed control function space and furthermore those problems whose control variables that are subject to inequality constraints appear linearly. This study presents an adaptation of Bryson and Denham's method of steepest descent to solve this class of problems. Other authors have proposed similar schemes for coping with this class of problems; see [5], [9], and [10].

Introduction. Consider a class of problems whose control vector variable,

$$u(t) = [v(t) \quad z(t)],$$

is an m-dimensional vector; $v(t)$ is a k-dimensional vector of continuous functions, such that $v(t) \in V$, where V is an open set of a Euclidean space E^k; and $z(t)$ is an $(m - k)$-dimensional vector, each of whose components is subject to a two-sided constraint of the type

$$|z^i(t)| \leqq 1, \qquad i = 1, \cdots, m - k, \tag{1}$$

that is, $z(t) \in Z$, where Z is a closed, bounded set of an $(m - k)$-dimensional Euclidean space, the unit hypercube. Thus $u(t) \in U = V \times Z$ for all $t \in [t_0, t_f]$. Consider, furthermore, the class of problems whose system equation has the form

$$\dot{x} = f(x, u) = k(x, v) + g(x, v)z, \tag{2}$$

where the vector-valued function k and the matrix g are assumed to be continuous and sufficiently differentiable with respect to all their arguments. The vector-valued function f is assumed to have finite discontinuities at those points where $u(t)$ is discontinuous, and to possess left- and right-hand derivatives with respect to x and u at each discontinuity.

* Received by the editors August 6, 1965. Presented at the First International Conference on Programming and Control, held at the United States Air Force Academy, Colorado, April 15, 1965.

† The Frank J. Seiler Research Laboratory, Office of Aerospace Research, United States Air Force Academy, Colorado.

The preceding equation has a wide variety of linear and nonlinear problems as special cases; it differs from the more general problem formulation by the fact that the control variables subject to inequality constraints appear linearly.

The Mayer problem. The class of problems discussed in the preceding section can be formulated as follows.
Choose the m-dimensional control vector function $u(t)$ from a class of piecewise continuous functions of time such that $|u^j(t)| \leqq 1$, for $j = k + 1, \cdots, m$, which takes the system described by the vector-valued differential equation

$$f(x, u) - \dot{x} = 0 \tag{3}$$

from its initial state x_0 at time t_0, to its intended final state, such that it satisfies the vector-valued terminal condition

$$\Psi[x(t_f), t_f] = 0 \tag{4}$$

and minimizes the cost index

$$\Phi[x(t_f), t_f]. \tag{5}$$

The time t_f is chosen as the first time that one of the terminal conditions, hereafter referred to as the *stopping condition*,

$$\Omega[x(t_f), t_f] = 0, \tag{6}$$

is satisfied.

Using the theory of the maximum principle, one can formulate a Hamiltonian function,

$$H[p(t), x(t), u(t)] = \sum_{i=1}^{n} p_i f^i(x, u), \tag{7}$$

which for autonomous systems can be shown to be stationary if its total derivative vanishes,

$$\frac{dH}{dt} = \sum_{i=1}^{n} \frac{\partial H}{\partial p_i} \frac{dp_i}{dt} + \sum_{i=1}^{n} \frac{\partial H}{\partial x^i} \frac{dx^i}{dt} + \sum_{j=1}^{m} \frac{\partial H}{\partial u^j} \frac{du^j}{dt} = 0. \tag{8}$$

The first two terms can be made to vanish by choosing

$$\frac{dx^i}{dt} = \frac{\partial H}{\partial p_i}, \qquad i = 1, \cdots, n, \tag{9}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial x^i}, \qquad i = 1, \cdots, n, \tag{10}$$

which are the classical canonical equations. The remaining terms of (8)

can be rewritten as

$$\frac{dH}{dt} = \left\langle \nabla_u H, \frac{du}{dt} \right\rangle = 0. \tag{11}$$

The required differentiability of H is assured by the choice of the functions p according to (10), whose solutions possess continuous time derivatives everywhere, except at the points of discontinuity of $u(t)$, and by the definition of system (3).

Thus if H is to be an extremal with respect to $u(t)$, its variation must vanish with respect to all possible variations in $u(t)$ from the optimum function $u^0(t)$. One can write the condition

$$H(p, x, u^0) \geqq H(p, x, u), \tag{12}$$

and the optimum control can be defined as

$$u^0(t) = \arg \max_{u \in U} H(p, x, u), \tag{13}$$

if and only if

$$H(p, x, \arg \max_{u \in U} H(p, x, u)) = \max_{u \in U} H(p, x, u).$$

Returning to the specific formulation appearing on the right-hand side of (7) one can show that the components of $z(t)$ take on only their extreme values; that is, they exhibit a bang-bang behavior and have a finite number of finite discontinuities where these control variables switch from one extreme value to the other one. Forming a Hamiltonian with the differential condition (2),

$$H(p, x, v, z) = p^T k(x, v) + p^T g(x, v) z, \tag{14}$$

it can be verified that the optimum choice of z stems from

$$H(p, x, v^0, z^0) \geqq H(p, x, v, z), \tag{15}$$

and that this condition is fulfilled if one maximizes

$$\langle p, g(x, v)z \rangle = \langle g^T(x, v)p, z \rangle, \tag{16}$$

for all $t \in [t_0, t_f]$ by choosing

$$z = \text{signum}\,[g^T(x, v)p]. \tag{17}$$

Thus the components of z take on only their extreme values; each component is piecewise constant and changes value when the corosponding component of the argument of (17) changes sign. This condition guarantees that systems describable by differential equations linear in the control variables that are subject to ineqality constraints can be described during succeeding

subintervals of time by a different system of differential equations. This condition fails, however, in those cases when the argument of (17) vanishes for finite intervals of time, thus giving rise to singular subarcs.

Reformulated problem. Let N be the total number of distinct discontinuities in all of the components of $z(t)$ in (t_0, t_f), and let $t = t_s$, $s = 1, \cdots, N$, be the ordered times at which the N discontinuities occur. For each subinterval $[t_{s-1}, t_s]$, $s = 1, \cdots, N+1$, where $t_f = t_{N+1}$, the vector function $z(t)$ is constant, as has been determined from an application of Pontryagin's maximum principle. Thus one can consider the system to be described by different sets of differential equations, each set corresponding to a successive interval of time; for $s = 1, \cdots, N+1$,

$$\dot{x}_s = f_s(x, u), \qquad t_{s-1} \leqq t \leqq t_s . \tag{18}$$

Because of the conditions stipulated on $f(x, u)$, one concludes that during each subinterval the right-hand side of (18) is a continuous and continuously differentiable function of the state and the control. Given a control vector $u(t)$, one can prove the existence and uniqueness of a solution $x(t)$ of (18) which is absolutely continuous and such that $\dot{x}_s = f_s(x)$ is valid everywhere in the interval $[t_{s-1}, t_s]$ and is subject to the initial conditions $x(t_{s-1})$ at time t_{s-1}, if the function $f_s(x)$ is continuous and Lipschitzian. The solution for the larger interval $[t_0, t_f]$ can be obtained by piecing together the solution for each subarc, with the initial state for each subarc corresponding to the terminal state of the preceding subarc.

The solution $x(t)$ for each subarc can also be shown to be a continuous and a continuously differentiable function of the initial state and the initial time at t_{s-1}, if the right-hand side of (18) is continuous and Lipschitzian. The state vector at t_{s-1} is similarly continuously dependent and continuously differentiable with respect to the preceding corner time t_{s-2}. By an inductive application of the results of basic theorems of differential equations, as given in [4], one can show that the state of the system at any time t is a continuous and continuously differentiable function of every corner time that precedes it, as well as being a continuous function of time for all $t \in [t_0, t)$.

The steepest descent algorithm. The generalization of Bryson and Denham's method of steepest descent to problems with discontinuous optimum solutions consists of treating the differential equations of the system as a succession of different systems of equations, one for each succeeding interval of time.

The succession of system equations (18) of the preceding section can be expressed by using Heaviside step functions as

$$\dot{x} = \sum_{s=1}^{N+1} f_s(x, u)[h(t - t_{s-1}) - h(t - t_s)]. \tag{19}$$

The variation of this differential equation relates the variation of the rate of change of the state of the system with the variation of the state, the control, and the switching times of the system.

The switching times are assumed to be variable functions of time: $t_s = t_s(t)$, $s = 1, \cdots, N$. The effect of the variation of the corner time on the variation of the step function $h(t - t_s)$ can be evaluated by considering the definition of the derivative of a step function and by introducing the variation of the corner times. For each corner time one obtains the variational relation

$$\delta h(t - t_s) = -\Delta(t - t_s)\delta t_s, \qquad s = 1, \cdots, N, \tag{20}$$

where $\Delta(t - t_s)$ is a delta function occurring at $t = t_s$.

Given a nominal choice of the control function $v(t)$ and a choice of the switching times, one obtains a trajectory in state space corresponding to it. If the initial control vector and the initial switching times are allowed to vary by a small amount, one obtains a corresponding variation in the state trajectory. The equation of variation can be obtained as the principal part of the Taylor series expansion of (19),

$$\begin{aligned} \delta\dot{x}(t) = &\sum_{s=1}^{N+1} [F_s(t)\delta x(t) + G_s(t)\delta v(t)][h(t - t_{s-1}) - h(t - t_s)] \\ &+ \sum_{s=1}^{N+1} f_s(x, u)[\Delta(t - t_s)\delta t_s - \Delta(t - t_{s-1})\delta t_{s-1}], \end{aligned} \tag{21}$$

where the matrices

$$F_s = \left[\frac{\partial f_s{}^i}{\partial x^j}\right], \quad G_s = \left[\frac{\partial f_s{}^i}{\partial v^r}\right], \tag{22}$$

for $i, j = 1, \cdots, n$, and $r = 1, \cdots, k$.

Equation (21) can be solved for the variation $\delta x(t)$ in terms of the variations $\delta v(t)$ and $\delta t_1, \cdots, \delta t_N$; this solution is readily possible by using the system of equations adjoint to (21),

$$\dot{\lambda}(t) = -\sum_{s=1}^{N+1} F_s{}^T(t)\lambda(t)[h(t - t_{s-1}) - h(t - t_s)], \tag{23}$$

where $\lambda(t)$ is an n-dimensional vector of adjoint variables. If (21) is premultiplied by $\lambda^T(t)$ and (23) is premultiplied by $\delta x^T(t)$ and then transposed, the sum of these two products can be integrated to give

$$\lambda^T(t_f)\,\delta x(t_f) = \sum_{s=1}^{N+1} \int_{t_{s-1}}^{t_s} \lambda^T(\tau) G_s(\tau) \delta v(\tau)\, d\tau$$

(24)

$$+ \sum_{s=1}^{N+1} \int_{t_{s-1}}^{t_s} \lambda^T(\tau) f_s(x, u)[\Delta(\tau - t_s)\delta t_s - \Delta(\tau - t_{s-1})\delta t_{s-1}]\, d\tau.$$

The inner product of the adjoint vector and the variation of the state vector at t_f can be rearranged to show the explicit dependence on the variation of the final time,

$$\lambda^T(t_f)\delta x(t_f) = \sum_{s=1}^{N+1} \int_{t_{s-1}}^{t_s} \lambda^T L_s\, \delta w\, d\tau + \lambda^T(t_f) f_{N+1}(t_f)\delta t_f\,, \tag{25}$$

where

$$L_s = [G_s \,\vdots\, C_s], \tag{26}$$

$$G_s = \left[\frac{\partial f_s{}^i}{\partial v^j}\right], \qquad i = 1, \cdots, n, \quad j = 1, \cdots, k, \tag{27}$$

$$C_s = [(f_s{}^i - f_{s+1}^i)\Delta(t - t_s)], \qquad s = 1, \cdots, N, \quad i = 1, \cdots, n, \tag{28}$$

$$\delta w = [\delta v^1 \cdots \delta v^k\, \delta t_1 \cdots \delta t_N]. \tag{29}$$

The matrix C_s accounts for the influence of the variation of the corner times. If there were no discontinuities in the function $f(x, u)$, then C_s would vanish and the total variation of the state vector at the terminal time would depend only on the variation $\delta v(t)$ and the variation of the terminal time δt_f, as obtained in [2]. The resulting variation $\delta w(t)$ which arises from the last equation involves the variation of time-dependent functions $v^1(t), \cdots, v^k(t)$, as well as variations of time-invariant quantities $t_1, \cdots, t_N$. The latter can be considered to be constant functions of time, $t_1(t), \cdots, t_N(t)$, and their variation can be obtained iteratively in the same manner as the variation of the continuous control vector.

The recursive variation. The Mayer problem that has been formulated deals with cost index and terminal conditions that are continuous point functions of the state and time of the system; hence these point functions are continuous functions of all switching times. Any function of the terminal state and of the terminal time will have a variation which, as a first approximation, can be represented as a linear combination of the variation of the state vector and of the terminal time. Choosing the terminal conditions on the adjoint vector

$$\lambda_\Phi(t_f) = \frac{\partial \Phi}{\partial x}\bigg|_{t_f}, \tag{30}$$

$$\lambda_{\Psi}(t_f) = \left.\frac{\partial \Psi}{\partial x}\right|_{t_f}, \tag{31}$$

$$\lambda_{\Omega}(t_f) = \left.\frac{\partial \Omega}{\partial x}\right|_{t_f}, \tag{32}$$

one can relate the total variations $d\Phi$, $d\Psi$, and $d\Omega$ to (25), where in each case the adjoint vector is a solution of (23) subject to one of the conditions (30), (31), or (32) with the corresponding subscript.

From this point on, the derivation of the recursive variation δw follows that outlined in the original work of Bryson and Denham [2]. In its final form the variation of the control w is obtained as

$$\begin{aligned}\delta w(t) = \pm \sum_{s=1}^{N+1} \Lambda^{-1} L_s{}^T(t)[\lambda_{\Phi\Omega}(t) - \lambda_{\Psi\Omega}(t) I_{\Psi\Psi}^{-1} I_{\Psi\Phi}] \\ \cdot \sqrt{\frac{(dP)^2 - d\Psi^T I_{\Psi\Psi}^{-1} d\Psi}{I_{\Phi\Phi} - I_{\Psi\Phi}^T I_{\Psi\Psi}^{-1} I_{\Psi\Phi}}} + \sum_{s=1}^{N+1} \Lambda^{-1} L_s{}^T(t) \lambda_{\Psi\Omega}(t) I_{\Psi\Psi}^{-1} \, d\Psi,\end{aligned} \tag{33}$$

where

$$\lambda_{\Phi\Omega}^T(t) = \lambda_{\Phi}{}^T(t) - \frac{\dot{\Phi}(t_f)}{\dot{\Omega}(t_f)} \lambda_{\Omega}{}^T(t), \tag{34}$$

$$\lambda_{\Psi\Omega}^T(t) = \lambda_{\Psi}{}^T(t) - \frac{\dot{\Psi}(t_f)}{\dot{\Omega}(t_f)} \lambda_{\Omega}{}^T(t), \tag{35}$$

$$(dP)^2 = \int_{t_0}^{t_f} \delta w(\tau)^T \Lambda \delta w(\tau) \, d\tau, \tag{36}$$

$$I_{\Phi\Phi} = \sum_{s=1}^{N+1} \int_{t_{s-1}}^{t_s} \lambda_{\Phi\Omega}^T L_s \Lambda^{-1} L_s{}^T \lambda_{\Phi\Omega} \, d\tau, \tag{37}$$

$$I_{\Psi\Phi} = \sum_{s=1}^{N+1} \int_{t_{s-1}}^{t_s} \lambda_{\Psi\Omega}^T L_s \Lambda^{-1} L_s{}^T \lambda_{\Phi\Omega} \, d\tau, \tag{38}$$

$$I_{\Psi\Psi} = \sum_{s=1}^{N+1} \int_{t_{s-1}}^{t_s} \lambda_{\Psi\Omega}^T L_s \Lambda^{-1} L_s{}^T \lambda_{\Psi\Omega} \, d\tau. \tag{39}$$

Equation (33) gives the recursive variation of the generalized control that produces the desired variation in the cost index and the terminal conditions; this expression is identical in form to the variation developed by Bryson and Denham. Hence in the same formalism developed by these authors, one can solve for the optimum control of a different class of control problems, and obtain iterative corrections to the choice of initial control vectors and an iterative correction to the initial choice of switching times.

Numerical example. In this section the method of steepest descent outlined in the preceding sections is applied to solve a problem with multiple subarcs. The specific problem solved is one for which it is possible to determine *a priori* the number of subarcs that compose the optimum solution. The problem that has been chosen for application of this method is one that has appeared in the literature numerous times; perhaps the most complete theoretical treatment of this problem is that of Leitmann [13], who analyzed it by applying classical techniques of the calculus of variations and thus deduced the nature of the optimal control.

The problem. The problem can be formulated as that of transferring a point mass a given distance over a flat surface, under the influence of a constant gravitational field, and in the absence of atmospheric and other disturbances. The object is to seek the minimum fuel trajectory traversed by the point mass, from all the possible thrust profiles that will cause the point mass to execute the desired change of state.

The system is the point mass, whose state is described by the system of simultaneous differential equations

$$\dot{x} = f(x, u) = \begin{bmatrix} \dot{x}^1 \\ \dot{x}^2 \\ \dot{x}^3 \\ \dot{x}^4 \\ \dot{x}^5 \end{bmatrix} = \begin{bmatrix} x^3 \\ x^4 \\ (c\beta u^1/x^5) \cos u^2 \\ (x\beta u^1/x^5) \sin u^2 - g \\ -\beta u^1 \end{bmatrix}, \tag{40}$$

where the state vector $x(t)$ is

$$x(t) = \begin{bmatrix} x^1(t) \\ x^2(t) \\ x^3(t) \\ x^4(t) \\ x^5(t) \end{bmatrix} = \begin{bmatrix} \text{horizontal displacement: ft} \\ \text{verticle displacement: ft} \\ \text{horizontal velocity: fps} \\ \text{vertical velocity: fps} \\ \text{mass: slugs} \end{bmatrix},$$

the control vector $u(t)$ is

$$u(t) = \begin{bmatrix} u^1(t) \\ u^2(t) \end{bmatrix} = \begin{bmatrix} \text{on-off thrust magnitude control} \\ \text{direction of thrust from the horizontal} \end{bmatrix},$$

c = effective exhaust velocity : 10,000 fps,

β = fuel rate : 0.0622 slugs/sec,

g = gravitational acceleration : 5.27 ft/sec^2.

The thrust acting on the point mass is defined as the product $c\beta$. In the presence of an inequality constraint on the fuel flow rate of the type $0 \leqq \beta \leqq \beta_{\max}$, one can redefine the fuel flow rate as $\beta = \beta_{\max} u^1(t)$, where

$0 \leqq u^1(t) \leqq 1$. The case where $u^1(t) \equiv 1$ corresponds to the case of continuous constant thrust; for the present case $u^1(t) \not\equiv 1$.

The initial state and the desired terminal state of the system are, respectively,

$$x(t_0) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 47.2 \end{bmatrix} \quad \text{and} \quad x(t_f) = \begin{bmatrix} 5000 \\ 0 \\ 0 \\ 0 \\ x^5(t_f) \end{bmatrix}. \tag{41}$$

The final mass is unspecified and is the object of the maximization process. Thus the cost index to be minimized is

$$\Phi = 47.2 - x^5(t_f); \tag{42}$$

the terminal constraints are

$$\begin{aligned} \Psi^1 &= [5000 - x^1(t_f)] = 0, \\ \Psi^2 &= [-x^3(t_f)] = 0, \\ \Psi^3 &= [-x^4(t_f)] = 0; \end{aligned} \tag{43}$$

and the stopping condition is

$$\Omega = [-x^2(t_f)] = 0. \tag{44}$$

For this problem one can show that the optimum thrust profile is composed of three subarcs; hence the influence of two discontinuities in the thrust must be calculated. The times when each of these discontinuities occurs in the initial thrust profile are to be assumed, as well as the direction of the thrust with respect to the horizontal as a function of time.

Application of the method of steepest descent. The times t_1 and t_2 are now assumed for the times at which these discontinuities occur in the component $u^1(t)$ of the control vector,

$$u^1(t) = \begin{cases} 1 & \text{if} \quad t_0 \leqq t \leqq t_1, \\ 0 & \text{if} \quad t_1 < t \leqq t_2, \\ 1 & \text{if} \quad t_2 < t \leqq t_3. \end{cases} \tag{45}$$

At the times t_1 and t_2 the system equations change because of the change in the control function $u^1(t)$. The system equation given in (40) can be rewritten using Heaviside step functions, for each of the subintervals indicated in (45). See Fig. 1. The rewritten system equation is

$$\begin{aligned} \dot{x} = f_1[1 - h(t - t_1)] + f_2[h(t - t_1) - h(t - t_2)] \\ + f_3[h(t - t_2) - h(t - t_3)], \end{aligned} \tag{46}$$

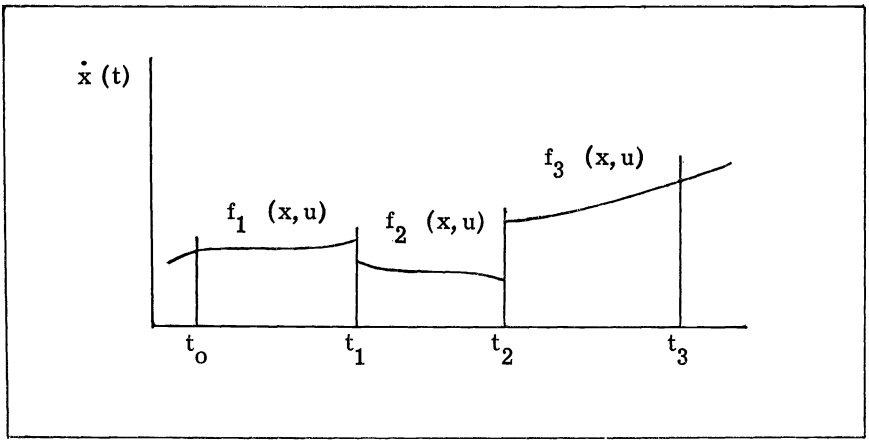

FIG. 1. *System equations*

which yields the equation of variation,

$$\begin{aligned}\delta\dot{x} = {} & \delta f_1[1 - h(t - t_1)] + \delta f_2[h(t - t_1) - h(t - t_2)] \\ & + \delta f_3[h(t - t_2) - h(t - t_3)] + (f_1 - f_2)\Delta(t - t_1)\delta t_1 \\ & + (f_2 - f_3)\Delta(t - t_2)\delta t_2 + f_3\Delta(t - t_3)\delta t_3 .\end{aligned} \tag{47}$$

This vector equation can be rewritten in component form:

$$\begin{aligned}\delta\dot{x}^1 &= \delta x^3, \\ \delta\dot{x}^2 &= \delta x^4, \\ \delta\dot{x}^3 &= -\left[\frac{c\beta u^1}{(x^5)^2}(\cos u^2)\delta x^5 + \frac{c\beta u^1}{x^5}(\sin u^2)\delta u^2\right] \\ &\quad + \frac{c\beta}{x^5}(\cos u^2)\Delta(t - t_1)\delta t_1 - \frac{c\beta}{x^5}(\cos u^2)\Delta(t - t_2)\delta t_2 \\ &\quad + \frac{c\beta}{x^5}(\cos u^2)\Delta(t - t_3)\delta t_3 , \\ \delta\dot{x}^4 &= -\left[\frac{c\beta u^1}{(x^5)^2}(\sin u^2)\delta x^5 - \frac{c\beta u^1}{x^5}(\cos u^2)\delta u^2\right] \\ &\quad + \frac{c\beta}{x^5}(\sin u^2)\Delta(t - t_1)\delta t_1 - \frac{c\beta}{x^5}(\sin u^2)\Delta(t - t_2)\delta t_2 \\ &\quad + \frac{c\beta}{x^5}(\sin u^2)\Delta(t - t_3)\delta t_3 , \\ \delta\dot{x}^5 &= -\beta\Delta(t - t_1)\delta t_1 + \beta\Delta(t - t_2)\delta t_2 - \beta\Delta(t - t_3)\delta t_3 .\end{aligned} \tag{48}$$

Written in matrix form these equations of variation and their adjoints become

$$\delta\dot{x} = F\delta x + L\delta w + f_3\Delta(t - t_3)\delta t_3\,, \tag{49}$$

$$\dot{\lambda} = -F^T\lambda, \tag{50}$$

where $\delta w^T = [\delta u^2\ \delta t_1\ \delta t_2]$,

$$F = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dfrac{-c\beta u^1}{(x^5)^2}\cos u^2 \\ 0 & 0 & 0 & 0 & \dfrac{-c\beta u^1}{(x^5)^2}\sin u^2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$L = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \dfrac{-c\beta u^1}{x^5}(\sin u^2) & \dfrac{c\beta}{x^5}(\cos u^2)\Delta(t - t_1) & \dfrac{-c\beta}{x^5}(\cos u^2)\Delta(t - t_2) \\ \dfrac{c\beta u^1}{x^5}(\cos u^2) & \dfrac{c\beta}{x^5}(\sin u^2)\Delta(t - t_1) & \dfrac{-c\beta}{x^5}(\sin u^2)\Delta(t - t_2) \\ 0 & -\beta\Delta(t - t_1) & \beta\Delta(t - t_2) \end{bmatrix}.$$

The equations of variation and their adjoints have the property that

$$\lambda^T(t_3)\delta x(t_3) = \int_0^{t_3} \lambda^T(\tau)L(\tau)\delta w(\tau)\,d\tau + \lambda^T(t_3)\dot{x}(t_3)\delta t_3\,, \tag{51}$$

from which one can obtain the variation in the cost index, the terminal and the stopping condition by the appropriate choice of the terminal conditions on the adjoint variables at t_3 as

$$\begin{bmatrix} \lambda_\Phi \\ \lambda_{\Psi^1} \\ \lambda_{\Psi^2} \\ \lambda_{\Psi^3} \\ \lambda_\Omega \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 \end{bmatrix}.$$

The terminal conditions at t_3 on the adjoint variables and the initial conditions on the state variables at t_0 comprise the total number of conditions that are necessary in order to apply the method of steepest descent.

Digital computer results. The initial control function $u^2(t)$ that was chosen to initiate the iteration process is shown in Fig. 2. The times t_1 and t_2 associated with this curve were chosen arbitrarily; the curve itself

was chosen to be a ramp function

$$u^2(t) = 1 + 0.005t, \tag{52}$$

where the units of u^2 are in radians. The positive zero intercept was chosen to insure that every point in the trajectory prior to the satisfaction of the stopping condition has a positive altitude. This was a necessary precaution motivated by the fact that the stopping condition was chosen as the first instant in time when the vertical displacement of the trajectory becomes negative, that is, when the point mass reaches the ground again, after takeoff.

Fig. 2 and Fig. 3 show the trajectories in control and state space, respectively, at two stages of the iteration process. Table 1 gives the numerical value of the significant variables of the problem as a function of the number of iterations. It can be seen that the initial choice of $u^2(t)$ caused the point mass to translate horizontally by -4141.3 ft, with a resultant error along the ground of 9141.3 ft and with terminal horizontal and vertical velocities of -330.4 fps and -167.7 fps, respectively. The data tabulated in Table 1 indicate the errors in the terminal conditions; as (43) indicates, the terminal velocities are the negatives of the terminal velocity errors, and the distance error equals the desired impact distance, 5000 ft, minus the actual impact distance.

Variation in the number of discontinuities. In the problem that has been solved in the preceding sections one could determine a priori the exact number of discontinuities present in the optimum control. Hence in the application of the steepest descent algorithm it is only necessary to determine their optimum location. Because in general it is not possible to apply analytical methods to determine the number of discontinuities for all problems of this type, one is forced to assume a certain number of them. The recursive variations must then correct this number of corner times, as well as to vary the time of their occurrence. Even though the present problem is sufficiently simple so that the composition of its optimum trajectory can be determined a priori, it is interesting to attempt to solve it by assuming an initial angle of thrust with too many or too few subarcs.

Too few subarcs. Fig. 4 illustrates the nominal control function used to start the iterative process with too few subarcs. The top curve illustrates the portion of the initial ramp which is also shown in Fig. 2; this portion of the initial ramp served as a first thrusting subarc. The equations of motion of the system were integrated using this first thrusting subarc; after time t_A the system was allowed to coast until the stopping condition was met at some unspecified later time t_B. On the succeeding iteration the variations of the thrusting subarc and of the time t_A were added to the initial choices of these functions; the resulting nominal initial subarc and cutoff time $t_A + \delta t_A$ were again used to integrate the differential equations of the system.

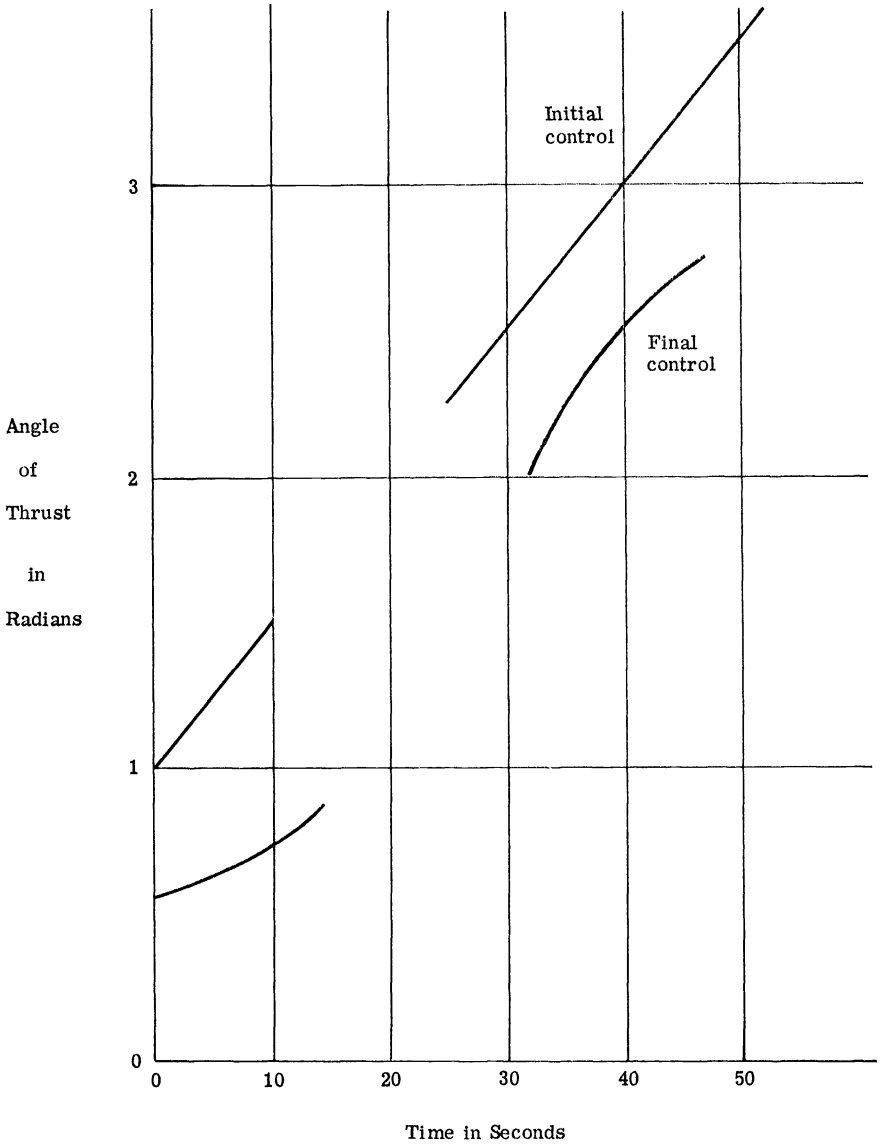

FIG. 2. *Trajectories in control space*

At time $t_B' = t_B - \Delta$, where Δ is the Runge-Kutta integration step size, a second thrusting subarc was introduced. Because of this additional thrusting subarc, of time duration Δ, the stopping condition was satisfied at some other time t_B''. In subsequent iterations the corrections to the times t_B' and t_B'' tended monotonically to lengthen the duration of the third

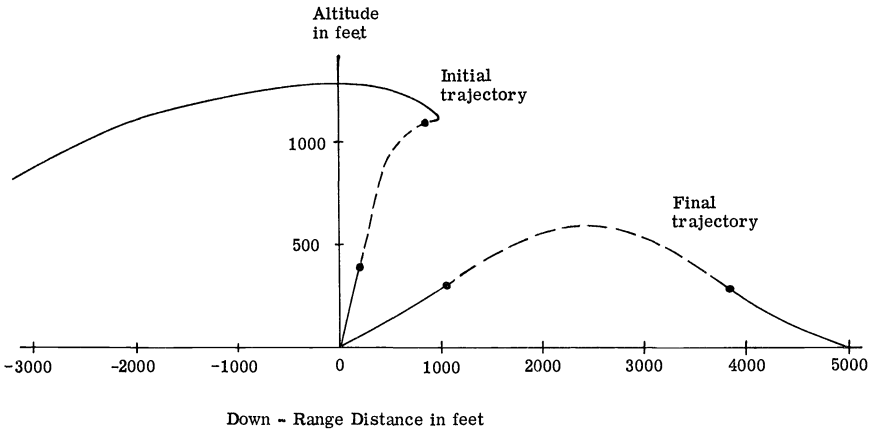

FIG. 3. *Trajectories in state space*

TABLE 1

cle	Fuel used	Distance error	Horizontal velocity error	Vertical velocity error	Gradient	Time of cut-off	Time of reignition
	slugs	*ft*	*fps*	*fps*		*sec*	*sec*
1	2.5977	9141.3	330.4	167.7	0.03014391	10.0000	25.0000
30	2.0895	5047.3	195.6	120.6	0.01728153	10.3500	28.2500
60	1.9985	3207.1	120.2	97.5	0.01411662	12.3500	29.9700
90	1.9689	2043.7	68.8	63.6	0.00896071	13.7181	31.5301
120	1.8730	1230.9	35.1	33.7	0.00317125	13.7965	31.4752
150	1.8121	782.8	9.7	12.2	0.00097550	14.1077	31.1275
180	1.7926	27.5	1.0	1.8	0.00010507	14.4682	31.1031

subarc, by decreasing t_B' and increasing t_B''. Simultaneously the variations in the magnitude of the third subarc with time modified the third subarc toward its optimum form. During this process the first thrusting subarc was also lengthened in duration and altered in shape towards its optimum form.

The lower curve in Fig. 4 shows the optimum first subarc; this first subarc was also used as a starting control function and was followed by a coasting subarc until some later time t_C when the stopping condition was met. At $t_C - \Delta$ the same impulsive thrust was introduced as was described in the preceding paragraph. Again it was noted that the leading and trailing edges of this very short thrusting subarc moved to the left and to the right, respectively, for a number of succeeding iterations, thus developing a full third subarc.

Too many subarcs. The two runs made with too many subarcs started with the two curves of Fig. 5. The top curve shows the initial ramp with

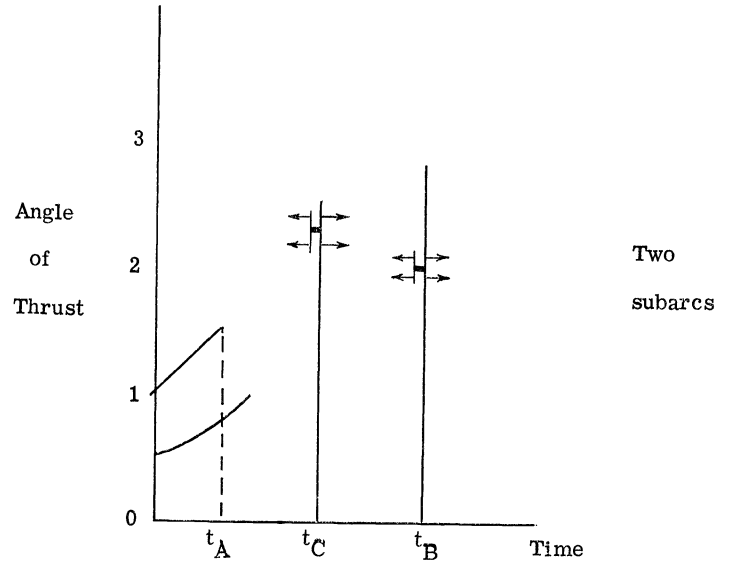

FIG. 4. *Two subarcs*

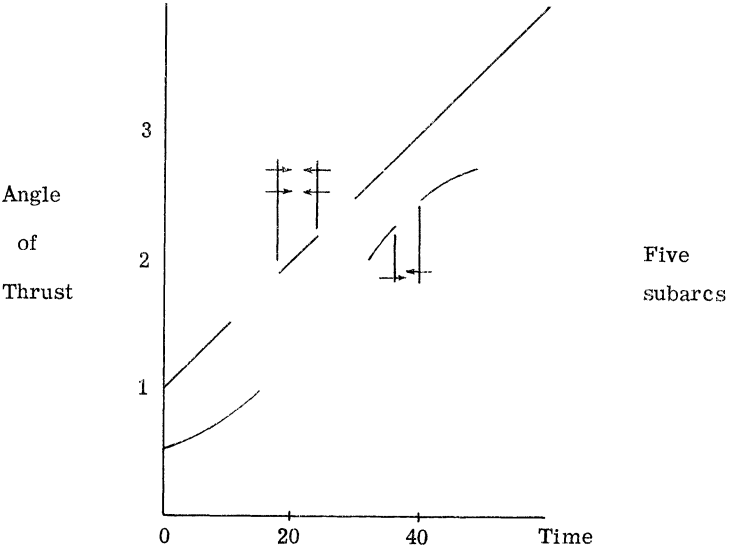

FIG. 5. *Five subarcs*

three thrusting subarcs. The additional thrusting occurs somewhere in the middle of what was the coasting subarc in the ramp illustrated in Fig. 2. As indicated by the arrows at the extremities of this additional thrusting subarc, the variation of these corner times caused this subarc to decrease in

duration until the two times overlapped; at this time the additional center thrusting subarc vanished.

A similar behavior was noticed when the iterative process was begun with the lower curve of Fig. 5. This curve illustrates the "optimum" control obtained previously, and shown in Fig. 2, but with an additional coasting subarc introduced during the third, thrusting subarc. The arrows indicate that here also the variations in the extremities of this extra coasting subarc during succeeding iterations tended toward each other.

In conclusion, it can be said that in the present problem there is a clear tendency to correct the number as well as the location of the times at which the discontinuities in the optimum control occur. Only in one case was the progression toward the "optimum" control slow. This was the case involving the use of the "optimum" control with the additional coasting subarc, as the initial control function. The slower convergence can be explained by the fact that the step size $(dP)^2$ was purposely related to the magnitude of the gradient; consequently as the gradient tended toward smaller values, the step size also tended toward zero. Thus the use of a "quasioptimal" control to start the iteration process produced a small gradient, with correspondingly small step sizes, which limited the progress in closing the superfluous coasting subarc.

Conclusions. The adaptation of Bryson and Denham's method discussed in this study gives hope of being applicable to other problems, notably to problems that one cannot analyze in detail a priori, using some of the classical tools of the calculus of variations, and for which one cannot determine the composition of the optimum trajectory.

Other problems, for which one knows how many discontinuities to expect in the state variable, can also be formulated and solved as the class of problems that has been illustrated in this study.

Finally it must be pointed out that the extremal properties of the control function that is generated in the present algorithm have not been determined. Thus one can only be certain of having identified a possible minimizing curve; the further identification of the properties of this curve can only be done by means of further tests.

Acknowledgments. The author is indebted to his colleagues at the Seiler Research Laboratory for many stimulating discussions. In particular he owes much to the encouragement and comments of Lt. Colonel O. J. Manci, Jr.

REFERENCES

[1] L. D. Berkovitz, *On control problems with bounded state variables*, J. Math. Anal. Appl., 5 (1962), pp. 448–489.

[2] A. E. BRYSON AND W. F. DENHAM, *A steepest ascent method for solving optimum programming problems*, Trans. ASME Ser. E. J. Appl. Mech., 29 (1962), pp. 247–257.

[3] A. E. BRYSON ET AL., *Optimal programming problems with inequality constraints I: Necessary conditions for extremal solutions*, AIAA J., 1 (1963), pp. 2544–2550.

[4] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[5] W. F. DENHAM AND A. E. BRYSON, *Optimal programming problems with inequality constraints II: Solution by steepest descent*, AIAA J., 2 (1964), pp. 25–34.

[6] S. DREYFUS, *Variational problems with state variable inequality constraints*, RAND Corporation, P-2605, 1956.

[7] R. V. GAMKRELIDZE, *Optimal processes with restricted phase coordinates*, Izv. Akad. Nauk SSSR Ser. Mat., 24 (1960), pp. 315–356.

[8] ———, *Optimum rate processes with bounded phase coordinates*, Dokl. Akad. Nauk SSSR, 125 (1959), pp. 475–478.

[9] R. G. GRAHAM, *The effect of state variable discontinuities in the solution of variational problems*, Aerospace Corporation, SSD-TDR 64-142, 1964.

[10] R. H. HILLSLEY AND H. M. ROBBINS, *Steepest ascent trajectory optimization method which reduces memory requirements*, Computing Methods in Optimization Problems, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1964, pp. 107–133.

[11] Y. C. Ho, *A successive approximation technique for optimal control systems subject to input saturation*, Trans. ASME Ser. D. J. Basic Engrg., 84 (1962), pp. 33–40.

[12] H. J. KELLEY ET AL., *Successive approximation techniques for trajectory optimization*, IAS Vehicle Systems Optimization Symposium, Garden City, New York, 1961.

[13] G. LEITMANN, *On a class of variational problems in rocket flight*, J. Aerospace Sci., 26 (1959), pp. 586–591.

[14] B. PAIEWONSKI, *Time optimal control of linear systems with bounded controls*, Nonlinear Differential Equations and Nonlinear Mechanics, J. P. LaSalle and S. Lefschetz, eds., Academic Press, New York, 1963, pp. 333–365.

[15] L. S. PONTRYAGIN ET AL., *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[16] V. A. TROITZKIY, *Variational problems in the optimization of control processes for equations with discontinuous right sides*, Prikl. Mat. Meh., 26 (1962), pp. 233–246.

[17] ———, *Variational problems in the optimization of control processes for systems with bounded coordinates*, Ibid., 26 (1962), pp. 431–443.

[18] R. F. VACHINO, *Steepest descent solution of a class of variational problems subject to inequality constraints on the control variable*, Ph.D. thesis, University of Michigan, Ann Arbor.

[19] F. A. VALENTINE, *The problem of Lagrange with differential inequalities as added side conditions*, Contributions to the Theory of the Calculus of Variations, University of Chicago Press, Chicago, 1937, pp. 407–448.

J. SIAM CONTROL
Vol. 4, No. 2, 1966
Printed in U.S.A.

DIRECTIONAL CONVEXITY AND THE MAXIMUM PRINCIPLE FOR DISCRETE SYSTEMS*

J. M. HOLTZMAN† AND H. HALKIN‡

Abstract. Directional convexity is a property of sets closely related to, but weaker than, convexity. It is the existence of supporting hyperplanes at all boundary points of a convex set that makes convexity important in optimal control theory. However, there need be supporting hyperplanes on only one side of the sets for much of the development. Directionally convex sets have this property. The concept of directional convexity, a property of sets, is a generalization, in the following sense, of the concept of convexity, a property of functions. The graph of every convex function is a directionally convex set. However, not every directionally convex set is a graph of a function. Since directional convexity is more general than convexity (for both sets and functions), it may unify a method of investigation which uses convex (or concave) functions with another which uses convex sets. Properties of directional convexity and of matrices that preserve directional convexity are given.

Directional convexity was introduced recently to extend the applicability of results on the optimal control of discrete-time systems. It is shown here that the results may be further generalized.

1. Introduction. A derivation of the maximum principle for a class of discrete systems was given in [1]. Actually, two derivations were presented in [1]. The first approach assumed the existence of tangent hyperplanes at points on the sets of reachable events. The second approach did not require the assumption of the existence of tangent hyperplanes. In both derivations, a convexity requirement was placed on the difference equations (including the performance state variable). It was shown in [2] that this convexity requirement is restrictive for practical systems. It is almost as restrictive as requiring linearity in the control. It was also shown in [2] that the first approach is valid with a requirement weaker than convexity. This extends its applicability to much wider classes of practical systems (an example of which is given in the present paper). A further discussion of this new requirement, "directional convexity," introduced in [2], will be given here. It will be shown here that the second approach of [1] is also valid with the new requirement of directional convexity. Other references on the discrete maximum principle are given in [1] and [2].

2. Directional convexity.

DEFINITION.[1] If z is a nonzero vector and A is a set we shall say that A

* Received by the editors June 7, 1965, and in revised form September 13, 1965.

† Bell Telephone Laboratories, Whippany, New Jersey.

‡ Bell Telephone Laboratories, Whippany, New Jersey. Now at Department of Mathematics, University of California at San Diego, La Jolla, California.

[1] In this paper, it is assumed that all sets are subsets of some finite-dimensional real Euclidean space and that all vectors are real.

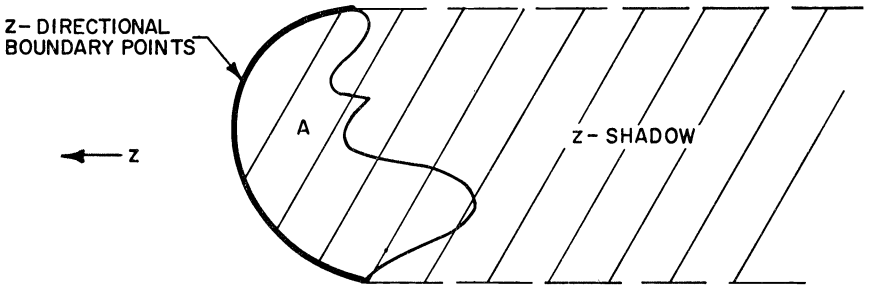

FIG. 1. *Illustration of definitions*

is *z-directionally convex* if for each $a, b \in A$, each $\mu \in [0, 1]$, there exists a $\beta \geqq 0$ such that

$$\mu a + (1 - \mu)b + \beta z \in A. \tag{2.1}$$

It is seen that directional convexity is weaker than convexity; all convex sets are z-directionally convex for any vector z since in the case of a convex set the relation (2.1) is always satisfied with $\beta = 0$. Some terminology is now introduced.

DEFINITION. A point a is said to be a *z-directional* boundary point of a set A if

(i) for every $\epsilon > 0$, there exists $b \in A$ such that $\| a - b \| < \epsilon$,

and

(ii) for every $\beta > 0$, $a + \beta z \notin A$.

DEFINITION. The *z-shadow* of a set A is the set

$$\{a - \lambda z : a \in A, \lambda \geqq 0\}.$$

The preceding definitions are illustrated in Fig. 1. It is easily shown that a z-directional boundary point of a z-directionally convex set A is a boundary point of the z-shadow of A.

THEOREM 2.1. *The z-shadow of a z-directionally convex set A is convex.*

Proof. Let S be the z-shadow of a z-directionally convex set A.

If $a, b \in S$, there exist $\lambda_1 \geqq 0$ and $\lambda_2 \geqq 0$ such that $a + \lambda_1 z \in A$ and $b + \lambda_2 z \in A$.

Then, from the z-directional convexity of A, for all $\mu \in [0, 1]$ there is a $\beta \geqq 0$ such that $\mu(a + \lambda_1 z) + (1 - \mu)(b + \lambda_2 z) + \beta z \in A$ or $c = \mu a + (1 - \mu)b + \beta^* z \in A$, where

$$\beta^* = \mu\lambda_1 + (1 - \mu)\lambda_2 + \beta \geqq 0.$$

Since S is the z-shadow of A,

$$d = c - \beta^* z \in S.$$

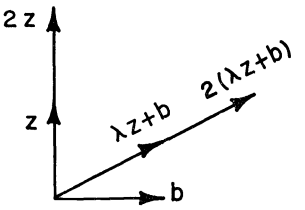

FIG. 2. *Construction for Theorem 2.2*

Since

$$d = \mu a + (1 - \mu)b \in S,$$

we have shown that S is convex.

DEFINITION. The matrix M is said to be a *z-directional matrix* if for every z-directionally convex set A the set

$$B = \{Mx: x \in A\}$$

is z-directionally convex.

The z-directional matrices are thus those that preserve z-directional convexity. Some properties of them are derived here. It will be seen that for M to be a z-directional matrix it is necessary that z be an eigenvector of M. For a large class of matrices (but not all), the associated eigenvalue must be nonnegative.[2]

THEOREM 2.2. *M is a z-directional matrix implies $Mz = \lambda z$ for some real λ.*

Proof. Assume $Mz = \lambda z + b$, $b \neq 0$, $\langle b \mid z \rangle = 0$.[3] Let $A = \{z, 2z\}$. A is z-directionally convex but $B = \{\lambda z + b, 2(\lambda z + b)\}$ is not z-directionally convex (see Fig. 2). Thus M is not a z-directional matrix.

THEOREM 2.3. *$Mz = \lambda z$, $\lambda \geqq 0$, implies M is a z-directional matrix.*

Proof. Let A be a z-directionally convex set and $B = \{Mx: x \in A\}$. We must prove that B is a z-directionally convex set. Let a and $b \in B$ and $\mu \in [0, 1]$. We must show that there exists a $\beta \geqq 0$ such that $\mu a + (1 - \mu)b + \beta z \in B$. There exist some a^* and $b^* \in A$ such that $a = Ma^*$ and $b = Mb^*$. Since A is a z-directionally convex set there exists a β^* such that $\mu a^* + (1 - \mu)b^* + \beta^* z \in A$. Then $\mu a + (1 - \mu)b + \beta^* \lambda z \in B$. We obtain the desired result by letting $\beta = \beta^* \lambda$.

Note that it is not true, in general, that the implication of Theorem 2.3 may be reversed. For example, the matrix

$$M = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$$

[2] Note that we are concerned only with real vector spaces over a real field so that it is possible to have only real eigenvalues.

[3] $\langle a \mid b \rangle$ is the notation used to denote the scalar product of a and b.

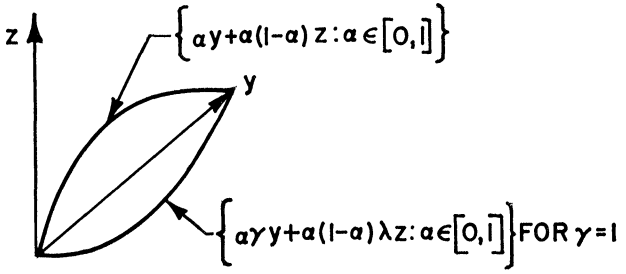

FIG. 3. *Construction for Theorem 2.4*

is a z-directional matrix with

$$z = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and $Mz = \lambda z$, $\lambda = -1$. There are, however, important cases in which the implication of Theorem 2.3 can be reversed.

THEOREM 2.4. *M has a nonzero eigenvalue γ with eigenvector y: $My = \gamma y$, and z is linearly independent of y. Then $Mz = \lambda z$, $\lambda \geqq 0$ if and only if M is a z-directional matrix.*

Proof. (Necessity.) From Theorem 2.3.

(Sufficiency.) Assume $Mz = \lambda z$, $\lambda < 0$. Let $A = \{\alpha y + \alpha(1 - \alpha)z: \alpha \in [0, 1]\}$. A is a z-directionally convex set. But $B = \{\alpha\gamma y + \alpha(1 - \alpha)\lambda z: \alpha \in [0, 1]\}$ is not z-directionally convex (see Fig. 3). Thus M is not a z-directional matrix.

THEOREM 2.5. *If M is nonsingular, then $Mz = \lambda z$, $\lambda > 0$, if and only if M is a z-directional matrix.*

Proof. (Necessity.) From Theorem 2.3.

(Sufficiency.) Two cases can exist:

(i) M has an eigenvector linearly independent of z (the corresponding eigenvalue must be nonzero because M is nonsingular); then Theorem 2.4 may be used.

(ii) M has no eigenvectors linearly independent of z. Let y be a nonzero vector with $\langle y \mid z \rangle = 0$. Then $My = cy + b$, $\langle b \mid y \rangle = 0$, $b \neq 0$, c a real scalar. Assume that

$$Mz = \lambda z, \qquad \lambda < 0.$$

Let

$$A = \{\alpha y + \alpha(1 - \alpha)z: \alpha \in [0, 1]\}.$$

A is a z-directionally convex set. But

$$B = \{\alpha cy + \alpha b + \alpha(1 - \alpha)\lambda z : \alpha \in [0, 1]\}$$

is not a z-directionally convex set. Thus M is not a z-directional matrix.

Next we shall consider additive and multiplicative properties of z-directional matrices. The sum of two z-directional matrices is not, in general, a z-directional matrix as can be seen by the following example.

$$M_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad M_2 = \begin{bmatrix} 0 & 0 \\ 0 & -2 \end{bmatrix}, \qquad M_1 + M_2 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

M_1 and M_2 are both z-directional matrices with $z = (0, 1)^T$. However, $(M_1 + M_2)$ is not (see Theorem 2.5). A sufficient condition for $(M_1 + M_2)$ to be a z-directional matrix is that M_1 and M_2 both be nonsingular z-directional matrices. This is proved in:

THEOREM 2.6. *If M_1 and M_2 are nonsingular z-directional matrices, α_1 and $\alpha_2 \geqq 0$, then $(\alpha_1 M_1 + \alpha_2 M_2)$ is a z-directional matrix.*

Proof. From Theorem 2.5, $M_1 z = \lambda_1 z$, $\lambda_1 \geqq 0$, and $M_2 z = \lambda_2 z$, $\lambda_2 \geqq 0$. Then

$$(\alpha_1 M_1 + \alpha_2 M_2)z = (\alpha_1\lambda_1 + \alpha_2\lambda_2)z, \qquad (\alpha_1\lambda_1 + \alpha_2\lambda_2) \geqq 0.$$

By Theorem 2.3, $\alpha_1 M_1 + \alpha_2 M_2$ is a z-directional matrix.

THEOREM 2.7. *If $\| M \| < 1$ and M is a z-directional matrix, then $(I + M)$ is a z-directional matrix.*

Proof. Since M is a z-directional matrix we have, from Theorem 2.2, $Mz = \lambda z$, λ real. Then $(I + M)z = (1 + \lambda)z$ and since $\| M \| < 1$, $| \lambda | < 1$. Thus $1 + \lambda > 0$. From Theorem 2.3, $I + M$ is a z-directional matrix.

The product of two z-directional matrices is always a z-directional matrix. This is proved in:

THEOREM 2.8. *If M_1 and M_2 are z-directional matrices, then $M_1 M_2$ is a z-directional matrix.*

Proof. Let A be a z-directionally convex set. Then we have to show that $B = \{M_1 M_2 x : x \in A\}$ is a z-directionally convex set. We can also write B as $B = \{M_1 y : y \in C\}$, where $C = \{M_2 x : x \in A\}$ is a z-directionally convex set. Since M_2 is a z-directional matrix, B is a z-directionally convex set.

THEOREM 2.9. *If A and B are z-directionally convex sets then $A + B$ is a z-directionally convex set.*[4]

Proof. Let x_1 and $x_2 \in A + B$. Then $x_1 = a_1 + b_1$ and $x_2 = a_2 + b_2$ with $a_1, a_2 \in A$ and $b_1, b_2 \in B_1$. We have

$$\mu x_1 + (1 - \mu)x_2 = \mu a_1 + (1 - \mu)a_2 + \mu b_1 + (1 - \mu)b_2 .$$

Since A and B are z-directionally convex sets, for each $\mu \in [0, 1]$, there are $\beta_1 \geqq 0$ and $\beta_2 \geqq 0$ such that

[4] $A + B$ is defined as the set $\{x : x = x_1 + x_2, x_1 \in A, x_2 \in B\}$.

$$\mu a_1 + (1 - \mu)a_2 + \beta_1 z \in A, \qquad \mu b_1 + (1 - \mu)b_2 + \beta_2 z \in B.$$

Therefore,

$$\mu x_1 + (1 - \mu)x_2 + (\beta_1 + \beta_2)z \in A + B.$$

Two convex sets A and B are separable if there exist a nonzero vector p and a scalar α such that

$$\langle x \mid p \rangle \leqq \alpha \quad \text{for all} \quad x \in A,$$

$$\langle x \mid p \rangle \geqq \alpha \quad \text{for all} \quad x \in B.$$

In other words, the two convex sets A and B are separable if there exists a hyperplane P such that the set A is on one side of P and the set B is on the other side. The hyperplane P is the set of all vectors x such that

$$\langle x \mid p \rangle = \alpha.$$

An important property of a convex set A is that there exists a supporting hyperplane passing through every boundary point of A (a hyperplane which separates the boundary point and the set A, see [3]). For z-directionally convex sets we have the following fundamental separation theorem.

THEOREM 2.10. *If the set A is z-directionally convex and a is a z-directional boundary point of A, then there exists a nonzero vector p such that*

$$\langle p \mid x \rangle \leqq \langle p \mid a \rangle \quad \textit{for all} \quad x \in A.$$

Proof. If a is a z-directional boundary point of A then a is a boundary point of the z-shadow of A which is convex (Theorem 2.1). Hence there exists a hyperplane separating a from the z-shadow of A and thus from the set A itself.

3. The discrete optimization problem. We are concerned with the system described by the difference equation

$$x(i + 1) - x(i) = A(i)x(i) + g(i, u(i)), \tag{3.1}$$

where x is an n-vector (state variable), A is an $n \times n$ matrix defined for every $i = 0, 1, \cdots, k - 1$, and g is an n-vector defined for every $i = 0, 1, \cdots, k - 1$ and every control u in Ω, a given set of admissible controls. We are given an initial condition vector x_0. We are also given a nonzero vector z and x_1. The set S_1 is defined by

$$S_1 = \{x_1 + \lambda z : \lambda \text{ real}\}. \tag{3.2}$$

We make the following assumptions:

(i) The sets $w(i) = \{g(i, u) : u \in \Omega\}$ are closed, bounded, and z-directionally convex for all $i = 0, 1, \cdots, k - 1$.

(ii) The matrices $I + A(i)$ are nonsingular for all $i = 0, 1, \cdots, k - 1$.

(iii) The matrices $I + A(i)$ are z-directional matrices for all $i = 0, 1, \cdots, k-1$.[5]

Before stating the optimization problem, some additional notation is is given. The letter u will represent a control strategy:

$$u = \{(i, u(i)) : i = 0, 1, \cdots, k-1\}. \tag{3.3}$$

The strategy u will be called "admissible" if

$$u(i) \in \Omega \quad \text{for all} \quad i = 0, 1, \cdots, k-1.$$

The letter F will represent the set of all admissible strategies.

We shall denote by $x(j; u)$ the value of the state variable at step j corresponding to the solution of the difference equation (3.1) satisfying

$$x(0; u) = x_0 \tag{3.4}$$

and with the strategy u. The optimization problem is to find a strategy $u \in F$ such that[6]

$$x(k; u) \in S_1 \tag{3.5}$$

and

$$\langle z \mid x(k; u) \rangle \quad \text{is maximum.} \tag{3.6}$$

A strategy satisfying the above will be denoted s.

There are two differences between the optimization problem of [1] and the one considered here. One is that, in [1], the objective is to maximize the nth element of $x(k)$. In our treatment, we are being slightly more general with maximizing in an arbitrary direction (the vector z). The second difference is that in [1] the sets

$$w(i) = \{g(i, u) : u \in \Omega\}, \quad i = 0, 1, \cdots, k-1, \tag{3.7}$$

are assumed to be convex (in addition to being closed and bounded). Here we relax the convexity requirement on $w(i)$ to z-directional convexity. This relaxation of the convexity requirement considerably extends the practical applicability of the results (see [2]).[7] The set defined at the end of §5 of the present paper could represent a $w(i)$ and is z-directionally convex but not convex.

The following are the results we wish to prove:

MAXIMUM PRINCIPLE. *If s is an optimal strategy, it is necessary that there*

[5] Conditions (ii) and (iii) will be satisfied if $A(i)$ is a z-directional matrix and $\| A(i) \| < 1$. $I + A(i)$ has an inverse; see, e.g., [4, p. 92]. $I + A(i)$ is a z-directional matrix by Theorem 2.7.

[6] The treatment of the case of free end conditions is straightforward.

[7] In [2], G_i denotes what is here called $w(i)$.

exists a nonzero vector $p(i, \nu)$ *satisfying the difference equation*

$$p(i, \nu) - p(i + 1, \nu) = A^T(i)p(i + 1, \nu) \tag{3.8}$$

and condition

$$\langle p(k, \nu) | z \rangle \geqq 0 \tag{3.9}$$

and such that

$$H(i, x(i, \nu), v(i), p(i + 1, \nu)) \geqq H(i, x(i, \nu), u, p(i + 1, \nu)) \tag{3.10}$$

for all $i = 0, 1, \cdots, k - 1$ *and all* $u \in \Omega$, *where* H *is defined by*

$$H(i, x, u, p) = \langle A(i)x + g(i, u) | p \rangle. \tag{3.11}$$

Most of the steps in proving the maximum principle are the same as those in [1]. However, for some steps, especially where convexity is used in [1], new proofs must be given here.

The "comoving space along a trajectory" is introduced in [1]. We shall summarize some of the notions of [1] to which the reader is referred for more detail. Let $G(i)$ be an $n \times n$ matrix for $i = 0, 1, \cdots, k$ and defined by

$$G(k) = I, \tag{3.12}$$

$$G(i) - G(i + 1) = G(i + 1)A(i), \quad i = 0, 1, \cdots, k - 1. \tag{3.13}$$[8]

$G^{-1}(i)$ exists for all $i = 0, 1, \cdots, k - 1$ ([1, §6]). Since

$$G(i) = (I + A(k - 1))(I + A(k - 2)) \cdots (I + A(i))$$

and $I + A(j)$ is a z-directional matrix for $j = 0, 1, \cdots, k - 1$, then $G(i)$ is a z-directional matrix.

Let Y be an n-dimensional Euclidean space with elements y. We shall consider the mapping from $X \times T$ into $Y \times T$ defined by the relation

$$y = G(i)(x - x(i, \nu)); \tag{3.14}$$

$y(i, \mu, \nu)$ is defined by

$$y(i, \mu, \nu) = G(i)(x(i, \mu) - x(i, \nu)). \tag{3.15}$$

The set of reachable events for the space X is defined by

$$W(i) = \{x(i, \mu) : \mu \in F\}. \tag{3.16}$$

A set of reachable events for the space Y is

$$W(i, \nu) = \{y(i, \mu, \nu) : \mu \in F\}. \tag{3.17}$$

[8] $G(i)$ is the discrete-time analog of the fundamental solution matrix or transition matrix for linear differential equations.

The space Y is called the comoving space along the optimal trajectory because the optimal trajectory in the space X is transformed into the trajectory $y = 0$ of the space Y (see (3.14)).

THEOREM 3.1. *The set $W(k, \nu)$ is z-directionally convex.*

Proof.

$$W(k, \nu) = \left\{\sum_{j=0}^{k-1} G(j+1)\big(g(j, u(j)) - g(j, v(j))\big) : u \in F\right\}$$

$$= \left\{\sum_{j=0}^{k-1} G(j+1)\big(g_j - g(j, v(j))\big) : g_j \in w(j)\right\},$$

where

$$w(j) = \{g(j, u) : u \in \Omega\}.$$

Then

$$W(k, \nu) = \{-h\} + \{G(1)g_0 : g_0 \in w(0)\} + \{G(2)g_1 : g_1 \in w(1)\}$$
$$+ \cdots + \{G(k)g_{k-1} : g_{k-1} \in w(k-1)\},$$

where

$$h = \sum_{j=0}^{k-1} G(j+1)g(j, v(j)).$$

Since $w(j)$ is z-directionally convex for $j = 0, 1, \cdots k-1$, and $G(j+1)$ is a z-directional matrix, then $W(k, \nu)$ is z-directionally convex by Theorem 2.9.

4. Proof of the maximum principle.

THEOREM 4.1. *If ν is an optimal strategy then $x(k, \nu)$ is a z-directional boundary point*[9] *of $W(k)$.*

Proof. If $x(k, \nu)$ is not a z-directional boundary point of $W(k)$ then there must exist a strategy u such that

$$x(k, u) = x(k, \nu) + \beta z \in W(k) \tag{4.1}$$

for some $\beta > 0$. Then

$$\langle z \mid x(k, u)\rangle > \langle z \mid x(k, \nu)\rangle, \tag{4.2}$$

which contradicts the optimality of ν.

THEOREM 4.2. *If $x(k, \nu)$ is a z-directional boundary point of $W(k)$ then $y = 0$ is a z-directional boundary point of $W(k, \nu)$.*

Proof. From the definitions of $W(i, \nu)$ and $y(i, u, \nu)$ we have

[9] It is also easily shown that $x(i, \nu)$ is a z-directional boundary point of $W(i)$ for all $i = 0, 1, \cdots, k-1$. This result makes Halkin's "principle of optimal evolution" [1] even more precise and indicates why only z-directional convexity is required.

$$W(k, \varsigma) = \{x - x(k, \varsigma) : x \in W(k)\} \tag{4.3}$$

and Theorem 4.2 follows immediately.

THEOREM 4.3. *If $y = 0$ is a z-directional boundary point of $W(k, \varsigma)$ then there is a nonzero vector $\phi(\varsigma)$ such that*

$$\langle G(i+1)(g(i, u) - g(i, v(i)))\,|\,\phi(\varsigma)\rangle \leqq 0 \tag{4.4}$$

for all $i = 0, 1, \cdots, k-1$ and all $u \in \Omega$.

Proof. The set $W(k, \varsigma)$ is z-directionally convex (Theorem 3.1) so that there exists a supporting hyperplane passing through $y = 0$ (Theorem 2.10). In other words there exists a nonzero vector $\phi(\varsigma)$, normal to the supporting hyperplane, such that

$$\langle y \,|\, \phi(\varsigma)\rangle \leqq 0 \quad \text{for all} \quad y \in W(k, \varsigma). \tag{4.5}$$

If $\phi(\varsigma)$ does not satisfy the condition (4.4), then there exist $j \in \{0, 1, \cdots, k-1\}$ and $u \in \Omega$ such that

$$\langle G(j+1)(g(j, u) - g(j, v(j)))\,|\,\phi(\varsigma)\rangle > 0. \tag{4.6}$$

Let $\bar{\varsigma} \in F$ be constructed by the following two relations

$$\bar{v}(i) = v(i), \qquad i \neq j, \tag{4.7}$$

$$\bar{v}(j) = u. \tag{4.8}$$

It is a trivial matter to verify that

$$\langle y(k, \bar{\varsigma}, \varsigma)\,|\,\phi(\varsigma)\rangle > 0, \tag{4.9}$$

which contradicts (4.5).

THEOREM 4.4. *If there exists a nonzero vector $\phi(\varsigma)$ satisfying* (4.4), *then there exists a nonzero vector $p(i, \varsigma)$ defined for $i = 0, 1, \cdots, k$ such that*

$$p(i, \varsigma) = G^T(i)\phi(\varsigma), \tag{4.10}$$

$$H(i, x(i, \varsigma), u, p(i+1, \varsigma)) \leqq H(i, x(i, \varsigma), v(i), p(i+1, \varsigma)), \tag{4.11}$$

for all $i = 0, 1, \cdots, k-1$ and all $u \in \Omega$,

$$p(i, \varsigma) - p(i+1, \varsigma) = A^T(i)p(i+1, \varsigma), \tag{4.12}$$

for all $i = 0, 1, 2, \cdots, k-1$.

Proof. Identical to the proof of [1, Theorem 8.4].

THEOREM 4.5. *If $x = x(k, \varsigma)$ is a z-directional boundary point of the set $W(k)$ then there exists a vector $p(i, \varsigma)$ defined for $i = 0, 1, \cdots, k$ and satisfying the relations* (4.10), (4.11), *and* (4.12).

Proof. Theorem 4.5 is a direct consequence of Theorems 4.2, 4.3, and 4.4. Indeed, if $x = x(k, \varsigma)$ is a z-directional boundary point of the set $W(k)$

then, by Theorem 4.2, $y = 0$ is a z-directional boundary point of $W(k, \varsigma)$. Then we may use Theorems 4.3 and 4.4.

THEOREM 4.6. *If ς is an optimal strategy then there exists a vector $p(i, \varsigma)$ defined for $i = 0, 1, \cdots, k$ and satisfying the relations* (4.10), (4.11), *and* (4.12).

Proof. By Theorem 4.1, if ς is an optimal strategy, then $x(k, \varsigma)$ is a z-directional boundary point of $W(k)$. Then we may use Theorem 4.5.

THEOREM 4.7. *If ς is an optimal strategy then*

$$\langle p(k, \varsigma) | z \rangle \geqq 0.$$

Proof. From (4.10) and (3.12) we have

$$p(k, \varsigma) = \phi(\varsigma).$$

From (4.5) and (3.14),

$$\langle x - x(k, \varsigma) | \phi(\varsigma) \rangle \leqq 0 \quad \text{for all} \quad x \in W(k).$$

Also,

$$\langle r - x(k, \varsigma) | \phi(\varsigma) \rangle \leqq 0 \quad \text{for all} \quad r \in z\text{-shadow of} \quad W(k).$$

Since

$$x(k, \varsigma) - z \in z\text{-shadow of} \quad W(k),$$

then

$$\langle z \mid \phi(\varsigma) \rangle \geqq 0.$$

5. Connection between directionally convex sets and convex functions. It may be observed that there is similarity between directional convexity, which is a property of sets, and convexity, a property of functions. Directional convexity is, in the following sense, a more general concept than convexity of functions. The graph of a convex function is always directionally convex (this will be made more precise in the theorem proved below). However, not all directionally convex sets may be obtained from the graph of a function (for example, a directionally convex set which "looks like" a quarter moon).

THEOREM 5.1. *If $x \in X$, a convex set, and $f(x)$ is a convex function of x, then the graph*

$$G = \{(x, f(x)) : x \in X\}$$

of that function is z-directionally convex with $z = (0, -1)$.

Proof. We have to show that for each $\mu \in [0, 1]$, each $x_1, x_2 \in X$, there is a $\beta \geqq 0$ such that

$$y = \mu(x_1, f(x_1)) + (1 - \mu)(x_2, f(x_2)) + \beta(0, -1) \in G.$$

We have

$$y = (\mu x_1 + (1 - \mu)x_2 , \mu f(x_1) + (1 - \mu)f(x_2)) + (0, -\beta).$$

From the convexity of $f(x)$,

$$y = (\mu x_1 + (1 - \mu)x_2 , f(\mu x_1 + (1 - \mu)x_2) - \gamma) + (0, -\beta)$$

for some $\gamma \geqq 0$. Thus by setting $\beta = \gamma$ we have

$$y = (\mu x_1 + (1 - \mu)x_2 , f(\mu x_1 + (1 - \mu)x_2)) \in G.$$

In a similar manner, it may easily be shown that, if $g_i(u)$ is linear in u for $i = 1, 2, \cdots, j$; Ω is a convex set; b_i are real scalars; and $f(u)$ is a concave function of u, then the set of vectors

$$\left\{ \begin{bmatrix} g_1(u) + b_1 \\ \vdots \\ g_j(u) + b_j \\ f(u) + b_{j+1} \end{bmatrix} : u \in \Omega \right\}$$

is z-directionally convex with $z = (0, 0, \cdots, 1)^T$. The above set is not convex.

6. Another application of directional convexity. Let us note here an interesting parallelism between two apparently unrelated problems: convexity is the fundamental tool of the known existence theorems [5], [6], [7], [8], for the optimal solution of systems described by differential equations and convexity is also the fundamental tool of the necessary condition derived in Halkin [1] for the optimal solution of systems described by difference equations. In the present paper we show that for the second of these problems the concept of convexity can be replaced by the more general concept of directional convexity. Similarly in [9] it is shown that for the first of these two problems the concept of convexity can also be replaced by directional convexity.

REFERENCES

[1] H. Halkin, *Optimal control for systems described by difference equations*, Advances in Control Systems: Theory and Applications, C. T. Leondes, ed., Academic Press, New York, pp. 173–196.

[2] J. M. Holtzman, *Convexity and the maximum principle for discrete systems*, IEEE Trans. Automatic Control, AC-11(1966), pp. 30–35.

[3] H. G. Eggleston, *Convexity*, Cambridge University Press, Cambridge, 1958.

[4] L. A. Liusternik and V. J. Sobolev, *Elements of Functional Analysis*, Ungar, New York, 1961.

[5] A. F. Filippov, *On certain questions in the theory of optional control*, Vestnik Moskov. Univ. Ser. Mat. Mekh. Astr. Fiz. Khim., 2(1959), pp. 25–32; English transl., this Journal, 1(1962), pp. 76–84.

[6] E. Roxin, *The existence of optimal controls*, Michigan Math. J., 9(1962), pp. 109–119.

[7] J. Warga, *Relaxed variational problems*, J. Math. Anal. Appl., 4(1962), pp. 111–128.

[8] R. V. Gamkrelidze, *Optimal sliding states*, Dokl. Akad. Nauk SSSR, 143(1962), pp. 1243–1245.

[9] J. M. Holtzman and H. Halkin, *Directional convexity and the existence of optimal controls*, Bell Telephone Laboratories, Technical Memorandum, 1965.

J. SIAM CONTROL
Vol. 4, No. 2, 1966
Printed in U.S.A.

OPTIMAL REGULATION OF LINEAR SYMMETRIC HYPERBOLIC SYSTEMS WITH FINITE DIMENSIONAL CONTROLS*

DAVID L. RUSSELL†

0. Introduction. A number of mathematicians have recently studied problems of optimal control for systems whose motion is described by partial differential equations. A short, and by no means complete, bibliography is given at the end of this paper. In most of these papers the control set is a subset of a function space. This corresponds to the assumption that control can be exercised by a force arbitrarily distributed throughout the body under consideration. Exceptions to this rule are to be found in [5], [8] and [11] where boundary value control of the heat equation is considered.

It is clear that in most applications the control set will be finite-dimensional. For example, the flexing of a large rocket booster is controlled by the essentially two-dimensional deflection of the rocket exhaust.

In this paper we study finite-dimensional control of linear symmetric hyperbolic systems of partial differential equations. The control criterion is minimization of the total energy of the system at a given time T after the exercise of control is begun. We show that the optimal control exists and satisfies a certain maximum principle. The maximum principle may, or may not, characterize the optimal control as a "bang-bang" control. A question akin to that of normality in ordinary differential equations arises and is studied.

We show that a class of boundary value control problems can be included in the problems to which our theory applies. As an example, we study the question of optimal energy dissipation in a vibrating string with control exercised at one endpoint.

1. The system. We will treat systems

$$E(x)\frac{\partial u}{\partial t} = A(x)\frac{\partial u}{\partial x} + C(x)u + B(x)f(t), \tag{L}$$

where E, A and C are n by n matrices while B is an n by m matrix, $m \leqq n$. We assume that these matrices are in class C^2 on the closed interval

$$0 \leqq x \leqq 1. \tag{1.1}$$

* Received by the editors April 19, 1965, and in final revised form October 15, 1965.

† Mathematics Research Center, United States Army, University of Wisconsin, Madison, Wisconsin. Contract No.: DA-11-022-ORD-2059. Now at the Department of Mathematics, University of Wisconsin. The author wishes to acknowledge a number of helpful suggestions by the referee which resulted in simplification of the proofs of Theorems 2.1 and 4.1.

The vector u is n-dimensional while f is an m-dimensional vector function measurable on the interval

$$0 \leqq t \leqq \infty \tag{1.2}$$

and such that for some positive number K,

$$\|f(t)\| \leqq K, \qquad t \in [0, \infty). \tag{1.3}$$

Here $\|\ \|$ denotes the Euclidean norm in E^m.

The matrices E and A are both symmetric and E is strictly positive definite on the interval (1.1). Thus the roots $\lambda_1(x), \lambda_2(x), \cdots, \lambda_n(x)$ of the equation

$$\det\,(E(x)\lambda - A(x)) = 0 \tag{1.4}$$

are real valued functions of x which are in $C^2[0, 1]$. The characteristics are the solution curves of

$$\frac{dx}{dt} = \lambda_i(x), \qquad i = 1, 2, \cdots, n. \tag{1.5}$$

For this system L we shall pose the following initial-boundary value problem (IBVP). As initial conditions we require that

$$u(x, 0) = u_0(x), \qquad x \in [0, 1], \tag{1.6}$$

where $u_0(x)$ is absolutely continuous on [0, 1], and that there is an $M > 0$ such that

$$\left\| \frac{du_0}{dx} \right\| \leqq M \tag{1.7}$$

wherever this derivative exists, i.e., almost everywhere. For boundary conditions we stipulate that there are n by n matrices A_0, A_1 such that

$$A_0u(0, t) \equiv 0, \qquad A_1u(1, t) \equiv 0, \qquad 0 \leqq t < \infty, \tag{1.8}$$

and we assume A_0 and A_1 are such that (1.8) implies

$$u(0, t)\cdot A(0)u(0, t) \equiv 0, \qquad u(1, t)\cdot A(1)u(1, t) \equiv 0, \qquad 0 \leqq t < \infty. \tag{1.9}$$

Finally we impose the consistency conditions

$$A_0u_0(0) = 0, \qquad A_1u_0(1) = 0. \tag{1.10}$$

We remark that the boundary conditions (1.8) cannot be arbitrarily imposed. They must be related to the sets of positive and negative eigenvalues $\lambda_i(x)$ of the system. For more on this, see §7.

The homogeneous system corresponding to (L) is the system

$$E(x)\frac{\partial u}{\partial t} = A(x)\frac{\partial u}{\partial x} + C(x)u. \tag{LH}$$

Under appropriate hypotheses the following theorems are true.

THEOREM 1.1. *For a given $f(t)$ obeying* (1.3) *the above described IBVP for* (L) *has a unique solution $u(x, t)$ in*

$$D = \{(x, t) | \ x \in [0, 1], t \in [0, \infty)\} \tag{1.11}$$

with the properties

(i) *u is continuous in D;*

(ii) *$\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial t}$ exist almost everywhere in D;*

(iii) *u, $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial t}$ and f satisfy the system equation* (L) *almost everywhere in D;*

(iv) *u, $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial t}$ are uniformly bounded in compact subsets of D and this bound is independent of f so long as f obeys* (1.3);

(v) *$u(x, t)$ is absolutely continuous on any line $x = c$, $0 \leqq c \leqq 1$, and on any line $t = c$, $0 \leqq c < \infty$, insofar as that line lies in D.*

THEOREM 1.2. *For a given $f(t)$ obeying* (1.3) *let $u(x, t)$ be the unique solution to the above IBVP and let $u_H(x, t)$ be the solution to the same IBVP for the system* (LH). *There is an n by m matrix function $V(x, t)$ which is piecewise of class C^1 in D and which depends only upon the system* (LH) *(i.e, not on f) such that*

$$u(x, t) = u_H(x, t) + \int_0^t V(x, t - \tau)f(\tau)\, d\tau. \tag{1.12}$$

Theorem 1.2 is known as Duhamel's principle. The columns of $V(x, t)$ are derivatives of solutions of the IBVP for (L) with zero initial conditions and special choices of f. Thus, at least in principle, $V(x, t)$ in computable.

Theorem 1.1 is proved in [14] under the assumption that $\lambda_1(x) > \lambda_2(x) > \cdots > \lambda_r(x) > 0 > \lambda_{r+1}(x) > \cdots > \lambda_n(x)$. Related material may be found in [15], [16], and [17].

2. The problem. Let $u(x, t)$ be the solution of the IBVP for (L) for some measurable $f(t)$ satisfying (1.3). The *energy distribution* associated with u is the quadratic form $u(x, t) \cdot E(x)u(x, t)$. By the *total energy* associated with u at a time t we shall understand the quantity $\mathcal{E}(u, t)$ given by

$$\mathcal{E}(u, t) = \frac{1}{2}\int_0^1 u(x, t) \cdot E(x)u(x, t)\, dx. \tag{2.1}$$

Now let $v(x)$ and $w(x)$ be bounded measurable n-vector functions de-

fined on the interval [0, 1]. The *energy inner product* $[v, w]$ is defined by

$$[v, w] = \int_0^1 v(x) \cdot E(x) w(x)\, dx. \tag{2.2}$$

It is easily verified that the set of all measurable n-vector functions $v(x)$ defined on [0, 1] for which

$$\int_0^1 v(x) \cdot E(x)\, v(x)\, dx < \infty \tag{2.3}$$

is a real Hilbert space with this inner product. Thus

$$\mathcal{E}(u, t) = \tfrac{1}{2}[u(\cdot, t), u(\cdot, t)] = \tfrac{1}{2} \| u \|^2, \tag{2.4}$$

where $u(\cdot, t)$ is the function of $x \in [0, 1]$ which we obtain from $u(x, t)$ by fixing t.

Consider the system

$$E(x) \frac{\partial v}{\partial t} = A(x) \frac{\partial v}{\partial x} + \left[\frac{dA(x)}{dx} - C^T(x)\right] v, \tag{L*}$$

where E, A and C are the matrices in (L) and C^T denotes the transpose of C. This system (L*) will be called the adjoint system for (L).

Let $v(x, t)$ be a solution of (L*) which satisfies the boundary conditions (1.8). It has properties similar to those of $u(x, t)$ if it satisfies appropriate initial (or terminal) conditions.

THEOREM 2.1. *The energy inner product* $[v(\cdot\ , t), u(\cdot\ , t)]$ *satisfies*

$$\frac{d}{dt}[v(\cdot, t), u(\cdot, t)] = \int_0^1 v(x, t) B(x)\, dx \cdot f(t) \tag{2.5}$$

for almost all $t \in [0, \infty)$.

Proof. Let t_1 be a positive real number. Since v and u are absolutely continuous on lines $x = c$ in D it is readily seen that

$$\begin{aligned}&[v(\cdot, t_1), u(\cdot, t_1)] - [v(\cdot, 0), u(\cdot, 0)] \\ &= \int_0^1 \left[\int_0^{t_1} \left(\frac{\partial v(x, t)}{\partial t} \cdot E(x) u(x, t) + v(x, t) \cdot E(x) \frac{\partial u(x, t)}{\partial t}\right) dt\right] dx.\end{aligned} \tag{2.6}$$

The properties of v and u (see Theorem 1.1) allow us to use Fubini's theorem and the fact that u and v satisfy (L) and (L*), respectively, to show that

$$\begin{aligned}&[v(\cdot, t_1), u(\cdot, t_1)] - [v(\cdot, 0), u(\cdot, 0)] \\ &= \int_0^{t_1} \left[\int_0^1 \left(u(x, t) \cdot A(x) \frac{\partial v(x, t)}{\partial x} + u(x, t) \cdot \frac{dA(x)}{dx} v(x, t)\right.\right. \\ &\qquad \left.\left. + v(x, t) \cdot A(x) \frac{\partial u(x, t)}{\partial x} + v(x, t) \cdot B(x) f(t)\right) dx\right] dt.\end{aligned} \tag{2.7}$$

The first three terms of the integrand in (2.7) together add to $\frac{\partial}{\partial x}(v(x, t)\cdot A(x)u(x, t))$. Then since v and u obey (1.8) (and hence (1.9)) it is easy to see that (2.7) implies

$$\begin{aligned}[v(\cdot, t_1), u(\cdot, t_1)] - [v(\cdot, 0), u(\cdot, 0)] &= \int_0^{t_1} [v(x, t)\cdot A(x)u(x, t)]_{x=0}^{x=1}\, dt \\ &\quad + \int_0^{t_1} \left(\int_0^1 v(x, t)B(x)\, dx\cdot f(t)\right) dt \\ &= \int_0^{t_1} \left(\int_0^1 v(x, t)B(x)\, dx\cdot f(t)\right) dt,\end{aligned} \tag{2.8}$$

and from this the result (2.5) follows immediately.

When

$$C(x) + C^T(x) \equiv \frac{dA(x)}{dx} \tag{2.9}$$

the systems (LH) and (L*) coincide. In this case a solution $u(x, t)$ of the IBVP for (LH) satisfies

$$\frac{d}{dt}[u(\cdot, t), u(\cdot, t)] = 0, \text{ a.e.}, \tag{2.10}$$

and hence, for all $t > 0$,

$$\mathcal{E}(u, t) = \mathcal{E}(u, 0), \tag{2.11}$$

i.e., the energy is conserved. Thus systems (LH) which are self-adjoint ((LH) is the same as (L*)) correspond to conservative systems in the language of physics. We shall not confine ourselves to conservative systems —the energy may increase or we may, as is usually the case, have a dissipative system.

Let Ω denote a compact convex subset of E^m. A measurable m-vector function $f(t)$ will be called an admissible control on $[0, T]$ if $f(t) \in \Omega$ a.e. in $[0, T]$.

THE OPTIMAL CONTROL PROBLEM. *Let an initial state $u(x, t) = u_0(x)$ be given. Corresponding to each bounded measurable m-vector function $f(t)$ let $u^f(x, t)$ be the solution of the IBVP. It is required to find an admissible control $f_0(t)$ such that*

$$\mathcal{E}(u^{f_0}, T) = \min_{f \text{ admissible}} \mathcal{E}(u^f, T),$$

Consider the following time optimal control problem: let E be a non-

negative real number less than $\mathcal{E}(u, 0)$. Find $f_0(t)$ which most quickly reduces $\mathcal{E}(u, t)$ to the value E. A solution of this time optimal problem is readily seen to be a solution of the above-posed optimal control problem where T is the first time at which $\mathcal{E}(u^{f^0}, t)$ is equal to E. The converse need not be true, however. A solution of the optimal control problem we have posed need not be time optimal in any sense.

3. Existence and uniqueness. Let a time $T > 0$ be fixed. Let $\mathfrak{H}$ denote the Hilbert space of n-vector functions defined on $[0, 1]$ for which the inequality (2.3) is satisfied and let the inner product in this Hilbert space be the energy inner product given in (2.2). By $\mathfrak{A}(u_0, T)$ we shall denote the set of all functions $u^f(\cdot, T)$, where $u^f(x, t)$ is the solution of the IBVP for (L) with initial state $u^f(x, 0) = u_0(x)$ for some admissible control $f(t)$. $\mathfrak{A}(u_0, T)$ is the set of attainability from u_0 at time T. It has been shown in [13] that $\mathfrak{A}(u_0, T)$ is closed. In the present instance we can prove in addition that $\mathfrak{A}(u_0, T)$ is compact.

According to Theorem 1.2, for a given admissible control $f(t)$, we have

$$u(x, T) = u_H(x, T) + \int_0^T V(x, T - \tau)f(\tau)\, d\tau, \qquad x \in [0, 1]. \tag{3.1}$$

Let $\{u_k(\cdot, T)\}$ be an arbitrary sequence of points in $\mathfrak{A}(u_0, T)$. Then for each positive integer k we have

$$u_k(x, T) = u_H(x, T) + \int_0^T V(x, T - \tau)f_k(\tau)\, d\tau, \qquad x \in [0, 1]. \tag{3.2}$$

Consider now the Hilbert space $\mathcal{G}$ consisting of all m-vector functions $f(t)$ defined on $[0, T]$ whose norms (in the usual Euclidean sense) are square integrable over that interval. We employ the usual inner product. Since Ω is compact and convex, the set of all admissible controls $f(t)$ on $[0, T]$ is a closed, bounded, and convex subset of $\mathcal{G}$. A familiar theorem from functional analysis states that such a subset of $\mathcal{G}$ is compact in the weak topology of $\mathcal{G}$.

Let $\{f_{k_l}(t)\}$ be a subsequence of $\{f_k(t)\}$ which converges to some $f(t) \in \mathcal{G}$ in the weak topology of $\mathcal{G}$. From what we have written above it is clear that $f(t)$ is also an admissible control. Then for each x in $[0, 1]$,

$$\lim_{l\to\infty} \int_0^T V(x, T - \tau)f_{k_l}(\tau)\, d\tau = \int_0^T V(x, T - \tau)f(\tau)\, d\tau, \tag{3.3}$$

and it follows that the subsequence $\{u_{k_l}(\cdot, T)\}$ of $\{u_k(\cdot, T)\}$ converges pointwise on $[0, 1]$ to $u(\cdot, T)$, where $u(x, t)$ is the response to $f(t)$, given for each $(x, t) \in [0, 1] \times [0, T]$ by the formula (1.12), But, using Theorem 1.1(iv), the $u_{k_l}(\cdot, T)$ and $u(\cdot, T)$ satisfy a common uniform Lipschitz condi-

tion (i.e., are equicontinuous) for $x \in [0, 1]$. Hence $\{u_{k_l}(\cdot\,, T)\}$ converges uniformly to $\{u(\cdot\,, T)\}$ for $x \in [0, 1]$. Then certainly $\{u_{k_l}(\cdot\,, T)\}$ converges to $u(\cdot\,, T)$ in the energy inner product topology of $\mathfrak{H}$. Since $\{u_k(\cdot\,, T)\}$ may be any sequence in $\mathfrak{A}(u_0\,, T)$ and $\mathfrak{H}$ is a metric space we see that $\mathfrak{A}(u_0\,, T)$ is compact.

Since the function $\mathcal{E}(u, T)$ is continuous for $u \in \mathfrak{A}(u_0\,, T)$ it assumes an absolute minimum on that set. Hence:

THEOREM 3.1. *Let the IBVP be posed for the system* (L) *and consider the optimization problem on the interval* $[0, T]$. *There is at least one solution* $f_0(t)$ *for this optimization problem.*

Our next task will be to discover in what sense, if any, the optimal control $f_0(t)$ is unique.

Let E be a positive real number. Consider the subset of $\mathfrak{H}$ given by

$$N_E = \{v \in \mathfrak{H} \mid \tfrac{1}{2} \| v \|^2 \leqq E\}. \tag{3.4}$$

Given an initial state $u_0(x)$, it is clear that there is an admissible control f^* such that $\mathcal{E}(u^{f^*}, T) = E$ if and only if $\mathfrak{A}(u_0\,, T) \cap N_E \neq \emptyset$, where $\emptyset$ denotes the empty set.

Let $f_0(t)$ be any solution of the optimization problem. Then it is clear that $E = \mathcal{E}(u^{f_0}, T)$ is the smallest number such that $\mathfrak{A}(u_0\,, T) \cap N_E \neq \emptyset$.

THEOREM 3.2. *Let the initial state* $u_0(x)$ *and the terminal time* T *be given and let* $f_1(t)$ *and* $f_2(t)$ *be any two solutions of the optimization problem. Then* $u^{f_1}(x, T) \equiv u^{f_2}(x, T)$. *Hence there is a unique terminal state* $u_T(x)$ *for the given optimization problem which is independent of the optimal control.*

Proof. Let $E = \mathcal{E}(u^{f_1}, T) = \mathcal{E}(u^{f_2}, T)$. Then it is clear that

$$u^{f_1}(\cdot\,, T) \in \mathfrak{A}(u_0\,, T) \cap N_E\,, \qquad u^{f_2}(\cdot\,, T) \in \mathfrak{A}(u_0\,, T) \cap N_E\,. \tag{3.5}$$

Since $\mathfrak{A}(u_0\,, T)$ and N_E are convex, so is their intersection. It follows that

$$\tfrac{1}{2}u^{f_1}(\cdot\,, T) + \tfrac{1}{2}u^{f_2}(\cdot\,, T) = u^{(\frac{1}{2}f_1 + \frac{1}{2}f_2)}(\cdot\,, T) \in \mathfrak{A}(u_0\,, T) \cap N_E\,. \tag{3.6}$$

If $u^{f_1}(x, T) \not\equiv u^{f_2}(x, T)$, it is easy to show that $\mathcal{E}(\frac{1}{2}u^{f_1} + \frac{1}{2}u^{f_2}, T) < E$ and hence $\mathfrak{A}(u_0\,, T) \cap N_{E'} \neq \emptyset$ for some $E' < E$. But this would violate the assumption that f_1 and f_2 are optimal. Hence

$$u^{f_1}(\cdot, T) = u^{f_2}(\cdot, T), \tag{3.7}$$

and the proof of the theorem is complete.

We are still far from showing the uniqueness of an optimal control function. But the uniqueness of the terminal state $u_T(x)$ will enable us to show that an optimal control must satisfy a certain maximum principle. From the maximum principle a certain "normality" condition can be developed which will show that an optimal control is unique under certain circumstances.

4. The maximum principle. We have shown in the preceding section that the unique terminal state $u_T(x)$ corresponding to the initial state $u_0(x)$ and an optimal control $f_0(t)$ is the unique point of intersection of $\mathfrak{A}(u_0, T)$ and N_E, where E is the minimum attainable energy.

LEMMA 4.1. *For each* $u^f(\cdot, T) \in \mathfrak{A}(u_0, T)$, *we have*

$$[u_T, u_T] \leqq [u_T, u^f(\cdot, T)]. \tag{4.1}$$

Proof. If $u_T(x) \equiv 0$, there is nothing to prove. Let us assume $u_T(x) \not\equiv 0$ and let $w \in \mathfrak{IC}$ be such that

$$[u_T, u_T] > [u_T, w]. \tag{4.2}$$

For $0 \leqq \lambda \leqq 1$ consider

$$w_\lambda(x) = (1 - \lambda)u_T(x) + \lambda w(x). \tag{4.3}$$

Then letting $[u_T, w] = [u_T, u_T] - \alpha$, $\alpha > 0$, we have

$$\begin{aligned}[w_\lambda, w_\lambda] &= (1 - \lambda)^2[u_T, u_T] + \lambda^2[w, w] + 2\lambda(1 - \lambda)[u_T, w] \\ &= \{(1 - \lambda)^2 + 2\lambda(1 - \lambda)\}[u_T, u_T] \\ &\quad - 2\lambda\alpha + \lambda^2(2\alpha + [w, w]) \\ &= (1 - \lambda^2)[u_T, u_T] - 2\lambda\alpha + \lambda^2(2\alpha + [w, w]).\end{aligned} \tag{4.4}$$

It is clear then that for sufficiently small $\lambda > 0$,

$$[w_\lambda, w_\lambda] < [u_T, u_T] = E. \tag{4.5}$$

Hence for such λ, $w_\lambda \notin \mathfrak{A}(u_0, T)$ and this implies $w \notin \mathfrak{A}(u_0, T)$. This completes the proof.

Now let $v(x, t)$ be the solution of the boundary value problem for (L*) which also satisfies the terminal condition

$$v(x, T) = u_T(x), \qquad x \in [0, 1]. \tag{4.6}$$

From Theorem 2.1 we have, for each admissible control $f(t)$,

$$\begin{aligned}[u_T, u^f(\cdot, T)] &= [v(\cdot, T), u^f(\cdot, T)] \\ &= [v(\cdot, 0), u_0] + \int_0^T \left[\int_0^1 v(x, t)B(x)\, dx\right] \cdot f(t)\, dt.\end{aligned} \tag{4.7}$$

Using the result of Lemma 4.1 together with (4.7) immediately yields:

THEOREM 4.1. (Maximum principle) *Let the IBVP for* (L) *be given with initial state* $u_0(x)$. *Let* $f_0(t)$ *be a solution of the optimization problem for* $T > 0$. *Let* $u_T(x)$ *be the unique terminal state for this problem and let* $v(x, t)$ *be the solution of the boundary value problem for* (L*) *with terminal condition*

(4.6). *Then for almost all* $t \in [0, T]$,

$$(4.8) \quad \left[-\int_0^1 v(x,t)B(x)\,dx\right]\cdot f_0(t) = \max_{f(t)\in\Omega}\left[-\int_0^1 v(x,t)B(x)\,dx\right]\cdot f(t).$$

5. Nature of the optimal control. In §4 we developed a necessary condition for optimality of a control function $f_0(t)$ in the form of a maximum principle. The usefulness of this maximum principle is not yet clear. For example, suppose it is possible to reduce the system to the zero state at time T, i.e. $u_T(x) \equiv 0$. Then $v(x, t) \equiv 0$ and the maximum principle is vacuous. We will later show by example that in such a case the optimal control function $f_0(t)$ may be a smooth, even analytic, function of time. Thus we cannot state any equivalent of the "bang-bang" principle for the system as a whole which will be strictly comparable to the corresponding principle for ordinary differential equations.

We will concern ourselves now with those situations wherein the terminal state $u_T(x) \equiv 0$ is not attainable and hence $v(x, t) \not\equiv 0$. Here we may expect that the maximum principle will yield some information on the properties of $f_0(t)$, an optimal control. For example, on any subinterval $[t_0, t_1] \subseteq [0, T]$ where $\int_0^1 v(x, t)B(x)\,dx \not\equiv 0$, the function $f_0(t)$ must almost everywhere lie on the boundary of Ω.

In order to develop an analog of the normality condition already known for linear ordinary differential equations, let us suppose now that our system is

$$(\hat{\mathrm{L}}) \qquad E\frac{\partial u}{\partial t} = A\frac{\partial u}{\partial x} + \left\{\sum_{l=0}^{q} B_l x^l\right\} f(t) = A\frac{\partial u}{\partial x} + B(x)f(t),$$

where E, A and B_l, $l = 0, \cdots, q$, are constant matrices. Moreover, let us assume Ω to be a polyhedron in E^m whose one-dimensional faces (i.e., edges) are parallel to unit vectors $w_1, w_2, \cdots, w_r$ in E^m.

The optimal control $f_0(t)$ can fail to be a uniquely determined piecewise constant function only if there is an interval $[t_0, t_1] \subseteq [0, T]$ of nonzero length and a w_i such that

$$(5.1) \qquad \int_0^1 v(x,t)B(x)\,dx\cdot w_i \equiv 0, \qquad t \in [t_0, t_1].$$

Then, clearly,

$$(5.2) \qquad \frac{d}{dt}\left[\int_0^1 v(x,t)B(x)\,dx\cdot w_i\right] \equiv 0, \qquad t \in (t_0, t_1).$$

Inasmuch as $v(x, t)$ has the same differentiability properties as stated for $u(x, t)$ in Theorem 1.1 we can write

$$\int_0^1 \frac{\partial v(x,t)}{\partial t} B(x)\, dx \cdot w_i \equiv 0, \qquad t \in (t_0, t_1). \tag{5.3}$$

Using the fact that $v(x, t)$ obeys $(\hat{\mathrm{L}}^*)$ and integrating by parts we see that

$$\left\{ [(E^{-1}Av(x,t))B(x)]_{x=0}^{x=1} - \int_0^1 (E^{-1}Av(x,t)) \frac{dB(x)}{dx}\, dx \right\} \cdot w_i \equiv 0, \qquad t \in (t_0, t_1). \tag{5.4}$$

If this process of

(i) differentiation with respect to t,
(ii) use of the equation $(\hat{\mathrm{L}}^*)$,
(iii) integration by parts

is repeated $q + 1$ times we find that

$$\left\{ \sum_{k=0}^{q} \left[\left((E^{-1}A)^{k+1} (-1)^k \frac{\partial^{q-k} v(x,t)}{\partial t^{q-k}} \right) \frac{d^k B(x)}{dx^k} \right]_{x=0}^{x=1} + (-1)^{q+1} \cdot \int_0^1 ((E^{-1}A)^{q+1} v(x,t)) \frac{d^{q+1} B(x)}{dx^{q+1}}\, dx \right\} \cdot w_i \equiv 0, \qquad t \in (t_0, t_1). \tag{5.5}$$

But, since $B(x)$ is a polynomial of degree q in x,

$$\frac{d^{q+1} B(x)}{dx} \equiv 0, \qquad x \in [0, 1], \tag{5.6}$$

and

$$\frac{d^k}{dx^k} \left(\sum_{l=0}^{q} B_l x^l \right) = \sum_{l=k}^{q} \frac{l!}{(l-k)!} B_l x^{l-k}, \qquad k \leqq q. \tag{5.7}$$

Hence

$$\left\{ \sum_{k=0}^{q} \left[\left((E^{-1}A)^{k+1} (-1)^k \frac{\partial^{q-k} v(1,t)}{\partial t^{q-k}} \right) \sum_{l=k}^{q} \frac{l!}{(l-k)!} B_l \right] - \sum_{k=0}^{q} \left[\left((E^{-1}A)^{k+1} (-1)^k \frac{\partial^{q-k} v(0,t)}{\partial t^{q-k}} \right) k!\, B_k \right] \right\} \cdot w_i \equiv 0, \qquad t \in (t_0, t_1). \tag{5.8}$$

The condition (5.8) involves only the boundary values of $v(x, t)$ and its derivatives with respect to t.

If we denote the left-hand side of (5.8) by $h(t) \cdot w_i$, then the condition

$$h(t) \cdot w_i \neq 0 \text{ on any subinterval of } [0, T] \tag{5.9}$$

may be regarded as a sort of *normality condition*. We must strongly emphasize, however, that normality is now a property of the initial state $u_0(x)$

and T via $u_T(x)$ and $v(x, t)$ and not, in general, a property associated with the system $(\hat{\mathrm{L}})$ as such. In addition, precautions must be taken to assure that the differentiations performed above are legitimate. An example will help to clarify these questions.

6. Example. The vibrating string. Let a string with uniform linear density ρ be stretched over a unit distance with tension τ. Let us measure distance along the string by means of the variable x. We shall study vibrations of the string in a plane. Hence we denote the displacement of the string from its equilibrium position by a scalar function $w(x)$. Under suitable conditions the equation of motion can be taken to be

$$\rho \frac{\partial^2 w}{\partial t^2} - \tau \frac{\partial^2 w}{\partial x^2} = 0. \tag{6.1}$$

Let us impose at $x = 0$ the boundary condition

$$w(0, t) \equiv 0, \tag{6.2}$$

and let control be exercised at $x = 1$ by setting

$$w(1, t) = \tilde{f}(t). \tag{6.3}$$

We shall require that $d\tilde{f}(t)/dt$ be an absolutely continuous function whose derivative, $d^2\tilde{f}(t)/dt^2$, which exists almost everywhere, satisfies

$$-1 \leqq \frac{d^2\tilde{f}(t)}{dt^2} \leqq 1. \tag{6.4}$$

Consider the change of variable

$$w(x, t) = y(x, t) + x\tilde{f}(t). \tag{6.5}$$

It is readily seen that $y(x, t)$ obeys the equation

$$\frac{\partial^2 y}{\partial t^2} - c^2 \frac{\partial^2 y}{\partial x^2} = xf(t) \tag{6.6}$$

along with the boundary conditions

$$y(0, t) \equiv 0, \qquad y(1, t) \equiv 0. \tag{6.7}$$

In (6.6) we have taken $c^2 = \tau/\rho$ and $f(t) = -d^2\tilde{f}(t)/dt^2$. Hence

$$-1 \leqq f(t) \leqq 1. \tag{6.8}$$

At a given moment t the vibrational energy (as distinguished from energy due to translational or rotational motion of the entire string) is given by

$$\begin{aligned} E_v(y, t) &= \frac{1}{2}\left[\rho\left(\frac{\partial y}{\partial t}\right)^2 + \tau\left(\frac{\partial y}{\partial x}\right)^2\right] dx \\ &= \frac{\rho}{2}\int_0^1 \left[\left(\frac{\partial y}{\partial t}\right)^2 + c^2\left(\frac{\partial y}{\partial x}\right)^2\right] dx. \end{aligned} \tag{6.9}$$

For convenience we set

$$\mathcal{E}(y,t) = \frac{1}{\rho} E_v(y,t). \tag{6.10}$$

To obtain a system of first order partial differential equations we let

$$u_1 = \frac{\partial y}{\partial t}, \qquad u_2 = \frac{\partial y}{\partial x}, \qquad u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{6.11}$$

Then we have

$$\frac{\partial}{\partial t}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 & c^2 \\ 1 & 0 \end{pmatrix} \frac{\partial}{\partial x}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + \begin{pmatrix} x \\ 0 \end{pmatrix} f(t). \tag{6.12}$$

Multiplying on the left by the matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & c^2 \end{pmatrix}$$

we obtain

$$E\frac{\partial u}{\partial t} = A\frac{\partial u}{\partial x} + B(x)f(t), \tag{6.13}$$

where

$$E = \begin{pmatrix} 1 & 0 \\ 0 & c^2 \end{pmatrix} \tag{6.14}$$

and hence is symmetric and positive definite,

$$A = \begin{pmatrix} 0 & c^2 \\ c^2 & 0 \end{pmatrix} \tag{6.15}$$

and is thus symmetric, while

$$B(x) = \begin{pmatrix} x \\ 0 \end{pmatrix}. \tag{6.16}$$

The boundary conditions (6.7) clearly imply

$$u_1(0,t) \equiv 0, \qquad u_1(1,t) \equiv 0, \tag{6.17}$$

and this in turn implies

$$u(0,t)\cdot Au(0,t) \equiv 0, \qquad u(1,t)\cdot Au(1,t) \equiv 0. \tag{6.18}$$

Hence the system (8.13) is of the type considered in this paper.

The total energy

$$\mathcal{E}(u,t) = \frac{1}{2}\int_0^1 u(x,t)\cdot Eu(x,t)\,dx \tag{6.19}$$

is clearly such that $\mathcal{E}(u, t) \equiv \mathcal{E}(y, t)$. Given an initial condition $u(x, 0) \equiv u_0(x)$, where $u_0(x)$ is absolutely continuous with uniformly bounded derivative on [0, 1] and satisfies

$$u_{01}(0) = 0, \qquad u_{01}(1) = 0, \tag{6.20}$$

let us consider the problem of determining a control function $f_0(t)$ such that $\mathcal{E}(u, T)$ is as small as possible for some $T > 0$.

It is clear that in this case the systems (LH) and (L*) coincide:

$$E\frac{\partial u}{\partial t} = A\frac{\partial u}{\partial x}. \tag{LH}$$

Thus (LH) is a conservative system. Hence let $v(x, t)$ be the solution of (LH) which satisfies

$$v_1(0, t) \equiv 0, \qquad v_1(1, t) \equiv 0, \tag{6.21}$$

and the terminal condition

$$v(x, T) \equiv u_0(x, T). \tag{6.22}$$

Then the optimal control $f_0(t)$ satisfies

$$\left(-\int_0^1 v(x, t)B(x)\, dx\right) f_0(t) = \max_{-1\leq f\leq 1}\left\{\left(-\int_0^1 v(x, t)B(x)\, dx\right) f\right\} \tag{6.23}$$

or

$$\begin{aligned}\left(-\int_0^1 xv_1(x, t)\, dx\right) f_0 &= \max_{-1\leq f\leq 1}\left\{\left(-\int_0^1 xv_1(x, t)\, dx\right) f\right\} \\ &= \left|\int_0^1 xv_1(x, t)\, dx\right|.\end{aligned} \tag{6.24}$$

With the identifications (see (6.11))

$$v_1 = \frac{\partial z}{\partial t}, \qquad v_2 = \frac{\partial z}{\partial x}, \tag{6.25}$$

we see that

$$\left(-\int_0^1 x\frac{\partial z(x, t)}{\partial t}\, dx\right) f_0(t) = \max_{-1\leq f\leq 1}\left\{\left(-\int_0^1 x\frac{\partial z(x, t)}{\partial t}\, dx\right) f\right\}. \tag{6.26}$$

Thus whenever

$$\int_0^1 x\frac{\partial z(x, t)}{\partial t}\, dx \neq 0,$$

we have

$$f_0(t) = -\operatorname{sgn}\left(\int_0^1 x \frac{\partial z(x,t)}{\partial t}\, dx\right). \tag{6.27}$$

In order to explore this situation further, let us suppose there is a subinterval $[t_0, t_1] \subseteq [0, T]$ whereon

$$\int_0^1 x \frac{\partial z(x,t)}{\partial t}\, dx \equiv 0. \tag{6.28}$$

Then for $t \in (t_0, t_1)$ it can be shown that

$$\int_0^1 x \frac{\partial^2 z(x,t)}{\partial t^2}\, dx = \int_0^1 xc^2 \frac{\partial^2 z(x,t)}{\partial x^2}\, dx \equiv 0. \tag{6.29}$$

But, integrating by parts and noting that $z(0,t) \equiv z(1,t) \equiv 0$, we see that (6.29) implies

$$\frac{\partial z(x,t)}{\partial x}\bigg|_{x=1} \equiv 0, \qquad t \in (t_0, t_1). \tag{6.30}$$

In order to progress a little further, let us consider the very special case wherein

$$\frac{\partial z(x,t)}{\partial t}\bigg|_{t=T} \equiv V_1(x,T) \equiv 0.$$

Since $z(x,t)$ is a solution of

$$\frac{\partial^2 z}{\partial t^2} - c^2 \frac{\partial^2 z}{\partial x^2} = 0, \quad z(0,t) \equiv z(1,t) \equiv 0, \quad z(x,T) = y(x,T), \tag{6.31}$$

it is a well-known fact that

$$z(x,t) = \tfrac{1}{2}[y(x + c(T-t)) + y(x - c(T-t))], \tag{6.32}$$

where we have extended the definition of $y(\cdot, T)$ to $(-\infty, \infty)$ by requiring

$$y(x) = -y(-x), \tag{6.33}$$

$$y(x+2) = y(x). \tag{6.34}$$

Then (6.30) implies

$$\begin{aligned} &\frac{1}{2}\left[\frac{\partial y(x + c(T-t))}{\partial x} + \frac{\partial y(x - c(T-t))}{\partial x}\right]_{x=1} \\ &\qquad = \frac{\partial y(x + c(T-t))}{\partial x}\bigg|_{x=1} \equiv 0, \qquad t \in (t_0, t_1). \end{aligned} \tag{6.35}$$

Hence we conclude that the optimal control function $f_0(t)$ always assumes unique values ± 1 on $[0, T]$ unless $y(\cdot, T)$ contains a segment of the form

$y = \text{const.}$ In particular, if $y(x, T) \not\equiv 0$, $x \in [0, 1]$, then $f_0(t)$ assumes unique extremal values on some subintervals of $[0, T]$ if $T \geqq 1/c$.

The calculations in (6.30) are a special case of the work done for a general system in §5. The fact that (6.13) comes from the second order scalar equation (6.6) simplifies things in that the steps (i), (ii), (iii) following (5.4) need only be performed once. For the general first order system where $B(x)$ is a polynomial of degree 1 these steps must be performed twice.

If $y(\cdot, 0) = w(\cdot, 0)$ is such that $\partial^2 y(x, 0)/\partial x^2$ is uniformly small on $[0, 1]$ and $y(\cdot, 0)$ is a smooth function, say analytic on $[0, 1]$, then the results given in [15, pp. 508–511] show that there is one and only one analytic function $f_0(t)$ defined on $[0, 1/c]$ such that $y(\cdot, 1/c) = w(\cdot, 1/c) = 0$. Such a control is obviously optimal. In addition, if $\partial y(x, 0)/\partial x \not\equiv 0$ in all neighborhoods of $x = 0$, then it is also time optimal.

We summarize the results for the vibrating string as follows: (i) if $y(\cdot, T)$ contains no segment $y = \text{const.}$, the optimal control is a unique "bang-bang", i.e., extremal, control; (ii) if $y(\cdot, T)$ is not identically zero and $T \geqq 1/c$, then $f_0(t)$ is uniquely determined and extremal on at least one subinterval of $[0, T]$; (iii) in certain cases where $y(\cdot, T)$ can be made equal to zero, the control may never be extremal—it may be a smoothly varying function which never reaches the boundary of the control set. If $T > 1/c$ one can construct examples wherein the optimal control is not unique.

It would be preferable, of course, to characterize the optimal control in terms of $y(\cdot, 0)$ but this appears to be very difficult.

7. Boundary value control. In the example of §6 it was very easy to transform the problem into one for which our theory is applicable. In general, the transformations which carry a boundary value control problem into a problem of the type described in §1 are more complicated.

In this section we will discuss the control of a system

$$(L_1) \qquad E(x)\frac{\partial w}{\partial t} = A(x)\frac{\partial w}{\partial x} + C(x)w$$

by means of varying boundary conditions

$$(7.1) \qquad A_0 w(0, t) = B_0 f(t), \qquad A_1 w(1, t) = B_1 g(t).$$

We assume that $f(t)$ and $g(t)$ are absolutely continuous vector functions of dimensions r and s, respectively, such that, where defined,

$$(7.2) \qquad \frac{df}{dt} \in \Omega_r, \qquad \frac{dg}{dt} \in \Omega_s,$$

where Ω_r and Ω_s are compact convex subsets of E^r and E^s, respectively,

whose interiors we assume to be nonempty. A_0 and A_1 are the matrices described in §1.

If the boundary conditions (7.1) are to make sense a number of conditions must be fulfilled. Our discussion of these conditions is very brief and the reader interested in greater detail is referred to [15, pp. 471–475].

We will assume that the equation

$$\det\,(E(x)\lambda - A(x)) = 0 \tag{7.3}$$

has distinct roots $\lambda_1(x), \lambda_2(x), \cdots, \lambda_n(x)$ such that

$$\lambda_i(x) > \lambda_{i+1}(x), \qquad x \in [0, 1], \qquad i = 1, 2, \cdots, n-1, \tag{7.4}$$

and for some m, $1 \leqq m < n$,

$$\lambda_m(x) > 0 > \lambda_{m+1}(x), \qquad x \in [0, 1]. \tag{7.5}$$

Corresponding to each of the eigenvalues $\lambda_i(x)$ of the matrix $E^{-1}(x)A(x)$ there is a left (i.e., covariant) eigenvector $l_i(x)$ which is a unit vector with the property that

$$l_i(x)E^{-1}(x)A(x) = \lambda_i(x)l_i(x). \tag{7.6}$$

For each $x \in [0, 1]$ the set $\{l_i(x) \mid i = 1, 2, \cdots, n\}$ forms a basis for E^n.

The scalar quantities $l_i(x)\cdot w(x, t)$ satisfy certain ordinary differential equations along the corresponding characteristic curves $dx/dt = \lambda_i(x)$. Consider some boundary point $(0, t)$, $t > 0$. The "incoming" characteristics are those for which $\lambda_i(x) < 0$, i.e., those for which $m < i \leqq n$. Hence the quantities $l_i(x)\cdot w(0, t)$, $m < i \leqq n$, are fixed as terminal values for certain solutions of ordinary differential equations. Thus for some real numbers μ_i, $m < i \leqq n$, we already have

$$l_i(x)\cdot w(0, t) < \mu_i, \qquad m < i \leqq n. \tag{7.7}$$

The equations (7.7) together with

$$A_0 w(0, t) = B_0 f(t) \tag{7.8}$$

must determine $w(0, t)$ uniquely. Hence the matrix $\begin{pmatrix} A_0 \\ L_0 \end{pmatrix}$, where L_0 is the $n - m$ by n matrix whose rows are the $l_i(x)$, $m < i \leqq n$, must have rank n and in addition, (7.7) and (7.8) must be consistent. A similar condition holds on the matrix $\begin{pmatrix} A_1 \\ L_1 \end{pmatrix}$, where the rows of L_1 are the vectors $l_i(x)$, $1 \leqq i \leqq m$.

Since Ω_r and Ω_s have nonempty interiors, (7.1) imply that

$$\text{range } B_0 \subseteq \text{range } A_0, \qquad \text{range } B_1 \subseteq \text{range } A_1. \tag{7.9}$$

Let F_0 and F_1 be complementary subspaces for the null spaces of A_0 and A_1, respectively. Then the restrictions $\hat{A}_0$ and $\hat{A}_1$ of A_0 to F_0 and A_1 to F_1, respectively, are invertible. Setting

$$\tilde{B}_0 = \hat{A}_0^{-1}B_0, \qquad \tilde{B}_1 = \hat{A}_1^{-1}B_1, \tag{7.10}$$

(7.1) becomes

$$A_0(w(0, t) - \tilde{B}_0 f(t)) = 0, \qquad A_1(w(1, t) - \tilde{B}_1 g(t)) = 0. \tag{7.11}$$

We define a new dependent variable $v(x, t)$ by

$$v(x, t) = w(x, t) - G(x)h(t). \tag{7.12}$$

Here $h(t)$ is the $(r + s)$-dimensional vector whose first r components are those of $f(t)$ and whose last s components are those of $g(t)$, while

$$G(x)h(t) = (1 - x)\tilde{B}_0 f(t) + x\tilde{B}_1 g(t). \tag{7.13}$$

Then $v(x, t)$ satisfies

$$\begin{aligned} E(x)\frac{\partial v}{\partial t} &= A(x)\frac{\partial v}{\partial x} + C(x)v \\ &\quad + \left[A(x)\frac{dG(x)}{dx} + C(x)G(x)\right]h(t) - E(x)G(x)\frac{dh(t)}{dt}. \end{aligned} \tag{7.14}$$

Let

$$D_1(x) = A(x)\frac{dG(x)}{dx} + C(x)G(x) \tag{7.15}$$

and

$$D_2(x) = -E(x)G(x). \tag{7.16}$$

Let $u(x, t)$ be the vector of dimension $n + r + s$ whose first n components are those of $v(x, t)$ and whose last $r + s$ components are those of $h(t)$. Let

$$\varphi(t) = \frac{dh(t)}{dt}. \tag{7.17}$$

Then $\varphi(t)$ is a measurable vector function such that $\varphi(t) \in \Omega = \Omega_r \times \Omega_s$, $t \geqq 0$. We see then that $u(x, t)$ satisfies the system

$$\begin{aligned} \begin{pmatrix} E(x) & 0 \\ 0 & I_{r+s} \end{pmatrix}\frac{\partial u}{\partial t} &= \begin{pmatrix} A(x) & 0 \\ 0 & 0 \end{pmatrix}\frac{\partial u}{\partial x} \\ &\quad + \begin{pmatrix} C(x) & D_1(x) \\ 0 & 0 \end{pmatrix}u + \begin{pmatrix} D_2(x) \\ I_{r+s} \end{pmatrix}\varphi(t), \end{aligned} \tag{L_2}$$

where I_{r+s} is the $r + s$ by $r + s$ identity matrix. The boundary conditions

now are

$$\begin{pmatrix} A_0 & 0 \\ 0 & 0 \end{pmatrix} u(0,t) \equiv 0, \qquad \begin{pmatrix} A_1 & 0 \\ 0 & 0 \end{pmatrix} u(1,t) \equiv 0, \tag{7.18}$$

and these clearly imply, because of (1.9), that

$$u(0,t)\cdot\begin{pmatrix} A_0 & 0 \\ 0 & 0 \end{pmatrix} u(0,t) \equiv 0, \qquad u(1,t)\cdot\begin{pmatrix} A_1 & 0 \\ 0 & 0 \end{pmatrix} u(1,t) \equiv 0. \tag{7.19}$$

Hence the system (L_2) satisfies the conditions laid down in §1 and our theory is applicable to it.

Minimization of

$$\begin{aligned} \hat{\mathcal{E}}(u,T) &= \frac{1}{2}\int_0^1 u(x,T)\cdot\begin{pmatrix} E(x) & 0 \\ 0 & I_{r+s} \end{pmatrix} u(x,T)\,dx \\ &= \frac{1}{2}\int_0^1 v(x,T)\cdot E(x)v(x,T)\,dx + \frac{1}{2}\|\,h(t)\,\|^2 \end{aligned} \tag{7.20}$$

constitutes minimization of the energy in the v coordinates plus a term which serves to measure the displacement of the v coordinates from the w coordinates. (See (7.12).) One may introduce a weighting scheme here, if so desired, by multiplying both sides of (L_2) by any matrix of the form

$$\begin{pmatrix} I_n & 0 \\ 0 & P \end{pmatrix} \tag{7.21}$$

where P is any symmetric and positive definite $r + s$ by $r + s$ matrix.

Finally we remark that Theorem 1.1 is true for solutions of L_2 because it is true for solutions of L_1 and $\varphi(t)$ satisfies (7.17).

8. Conclusion. We have presented here a rather brief introduction to finite-dimensional control of partial differential equations of a certain type. While a maximum principle has been obtained it is clear that it does not tell as much about the optimal control as the corresponding principle does for ordinary differential equations. We hope this paper will stimulate further research in this area. Results concerning controllability would be particularly interesting.

Finally we remark that the result of Theorem 1.2 lends itself to approximate numerical solution of the optimization problem by means of quadratic programming techniques. For partial differential equations to which the method of separation of variables is applicable one can approximate the solutions using certain systems of ordinary differential equations. More detail on these approximate methods may be found in [14] which is an earlier version of this paper.

REFERENCES

[1] A. G. Butkovskii and A. Ya. Lerner, *Optimal control of systems with distributed parameters*, Avtomat. i Telemek., 21(1960), pp. 682–691.

[2] ———, *Optimal control systems with distributed parameters*, Dokl. Akad. Nauk SSSR, 134(1960), pp. 778–781.

[3] A. G. Butkovskii, *Optimum processes in systems with distributed parameters*, Avtomat. i Telemek., 22(1961), pp. 17–26.

[4] ———, *The maximum principle for optimum systems with distributed parameters*, Ibid., 22(1961), pp. 1288–1301.

[5] Ju. V. Egorov, *Certain problems in the theory of optimal control*, Dokl. Akad. Nauk SSSR, 145(1962), pp. 720–723.

[6] ———, *On optimal control of processes in distributed objects*, Prikl. Mat. Meh., 27(1963), pp. 688–696.

[7] ———, *Optimal control in Banach space*, Soviet Math. Dokl., 4(1963), pp. 630–633.

[8] P. K. C. Wang and F. Tung, *Optimum control of distributed parameter systems*, Proceedings of the Joint Automatic Control Conference, 1963, pp. 16–31.

[9] P. K. C. Wang, *Asymptotic stability of a time delayed diffusion system*, Trans. ASME Ser. E. J. Appl. Mech., 30E(1963), pp. 500–504.

[10] A. V. Balakrishnan, *Optimal control problems in Banach spaces*, this Journal, 3(1965), pp. 152–180.

[11] ———, *Semigroup theory and control theory*, Proceedings, IFIP Congress, 1965.

[12] H. O. Fattorini, *Time-optimal control of solutions of operational differential equations*, this Journal, 2(1964), pp. 54–59.

[13] Peter L. Falb, *Infinite dimensional control problems I: On the closure of the set of attainable states for linear systems*, J. Math. Anal. Appl., 9(1964), pp. 12–22.

[14] D. L. Russell, *Optimal regulation of linear symmetric hyperbolic systems with finite dimensional controls*, Tech. Summary Rpt. 556, Mathematics Research Center, United States Army, University of Wisconsin, Madison, 1965.

[15] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, vol. II, Partial Differential Equations, Interscience, New York, 1962.

[16] K. O. Friedrichs, *Symmetric hyperbolic linear differential equations*, Comm. Pure Appl. Math., 7(1954), pp. 345–392.

[17] Peter D. Lax, *On Cauchy's problem for hyperbolic equations and the differentiability of solutions of elliptic equations*, Ibid., 8(1955), pp. 615–633.

J. SIAM CONTROL
Vol. 4, No. 2, 1966
Printed in U.S.A.

STABILITY OF A CLASS OF DIFFERENTIAL EQUATIONS WITH A SINGLE MONOTONE NONLINEARITY*

KUMPATI S. NARENDRA† AND CHARLES P. NEUMAN‡

1. Introduction. In recent years considerable interest has been shown in the study of the stability of dynamical systems using the second method of Lyapunov. By this approach, the stability of a dynamical system is guaranteed by the determination of a positive definite function whose total time derivative along a motion of the system is negative definite. The class of nonlinear dynamical systems discussed in this paper is composed of a linear plant characterized by the transfer function $G(s)$ and a nonlinear feedback element $f(\cdot)$ whose argument is a linear combination of the system state variables.

The purpose of this paper is to examine in detail the absolute stability (global asymptotic stability) of a class of dynamical systems which satisfy neither the Popov theorem [1] nor the extended Popov theorem [2], [3], [4]. Specifically, by introducing a new Lyapunov function and utilizing the Meyer [5], Kalman [6], and Yakubovich [7] Lemma, frequency domain stability criteria are obtained for the linear plant $G(s)$ in the case of both monotone increasing and odd monotone increasing nonlinear feedback functions. The study of this class of dynamical systems was initiated recently by Zames [8] and Brockett and Willems [9]. For infinite sector problems these frequency domain stability criteria (sufficient conditions for absolute stability) are applicable to linear plants whose transfer functions have some real nonzero zeros; for finite sector problems these criteria are applicable to a system whose characteristic equation evaluated at the maximum stable feedback gain has some real nonzero zeros or real nonzero poles.

The results presented in this paper demonstrate that the assumptions of monotone increasing and odd monotone increasing feedback functions lead successively to less and less restrictive conditions on the linear part of the system, the plant. Moreover, the frequency domain stability criteria derived in this paper provide a direct, systematic method for the generation

* Received by the editors June 3, 1965, and in final revised form October 27, 1965. This work was supported in part by the Joint Services Electronics Program (United States Army, United States Navy, and United States Air Force) under Contract Nonr-1866(16), and by the Division of Engineering and Applied Physics, Harvard University.

† Department of Engineering and Applied Science, Yale University, New Haven, Connecticut.

‡ Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts.

of explicit Lyapunov functions. Finally, examples are included in order to illustrate the ideas developed in this paper.

2. Problem statement. The completely controllable, completely observable, single-input, single-output continuous-time dynamical system [10] treated in this paper is described by the vector matrix equations

$$\begin{aligned} \dot{x} &= Ax + bu, \\ u &= -f(\sigma), \\ \sigma &= h^T x, \end{aligned} \tag{2.1}$$

for all $t \geqq t_0$, where x, the state of the system, is a real n-vector; A is a real constant $n \times n$ stable matrix; $u(\cdot)$, the system input or control function, and $\sigma(\cdot)$, the system output, are real scalar time functions; and b, the input transformation, and h, the output transformation, are real constant n-vectors.

Since the triple $[A, b, h]$ is completely controllable and completely observable, a basis in the state space X may be chosen so that this triple has the canonical (phase-variable) form [11]

$$A = \begin{bmatrix} 0 & 1 & & & 0 \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \\ 0 & & & 0 & 1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} \quad -a_n \end{bmatrix}, \tag{2.2}$$

$$b = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \qquad h = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{n-1} \\ h_n \end{bmatrix}.$$

The transfer function of the linear part of the system, the plant, is computed by taking the formal Laplace transforms of (2.1a) and (2.1c) and utilizing the canonical representation (2.2). This transfer function is

$$G(s) = h^T(sI - A)^{-1}b = \frac{h_n s^{n-1} + h_{n-1} s^{n-2} + \cdots + h_2 s + h_1}{s^n + a_n s^{n-1} + a_{n-1} s^{n-2} + \cdots + a_2 s + a_1}. \tag{2.3}$$

Moreover, since the dynamical system is completely controllable and completely observable, the transfer function $G(s) = p(s)/q(s)$, where $p(s)$ and $q(s)$ are relatively prime polynomials with real coefficients and $q(s)$ is the characteristic polynomial of A. The degree of the numerator polynomial is less than the degree of the denominator polynomial so that $G(\infty) = 0$.

In addition, the control $u = -f(\sigma)$, where $\sigma = h^T x$ is a linear combination of the system state variables. The continuous nonlinear function $f(\sigma)$ satisfies

$$f(0) = 0, \qquad \sigma f(\sigma) > 0, \quad \sigma \neq 0, \tag{2.4}$$

and is either (a) a monotone increasing or (b) an odd monotone increasing feedback function. These two classes of nonlinearities have the additional geometrical properties which are expressed by:

LEMMA 2.1. *If $f(\cdot)$ is a monotone increasing function, then for any σ_1 and σ_2,*

$$\{f(\sigma_1) - f(\sigma_2)\}(\sigma_1 - \sigma_2) \geqq 0.$$

LEMMA 2.2. *If $f(\cdot)$ is an odd monotone increasing function, then for any σ_1 and σ_2, and any $f(\cdot)$ such that $0 < \sigma f(\sigma) < \bar{K}\sigma^2$, $0 \leqq \epsilon < \bar{K} \leqq +\infty$,*

$$\left[\sigma_1 f(\sigma_2) - \sigma_2 f(\sigma_1) - \frac{1}{\bar{K}} f(\sigma_1) f(\sigma_2)\right] \leqq [\sigma_1 f(\sigma_1) + \sigma_2 f(\sigma_2)]. \tag{2.5}$$

Proof.

(i) If σ_1 or σ_2 is zero, then the left-hand side of (2.5) is identically zero and the right side is nonnegative.

(ii) If $\operatorname{sgn} \sigma_1 = \operatorname{sgn} \sigma_2$, then $\sigma_1 f(\sigma_2) \leqq \sigma_1 f(\sigma_1) + \sigma_2 f(\sigma_2)$.

(iii) If $\operatorname{sgn} \sigma_1 = -\operatorname{sgn} \sigma_2$, then the left-hand side is less than $-\sigma_2 f(\sigma_1)$ and $-\sigma_2 f(\sigma_1) \leqq \sigma_1 f(\sigma_1) + \sigma_2 f(\sigma_2)$.

These lemmas will be employed to derive sufficient conditions for the absolute stability of the class of dynamical systems with monotone increasing feedback functions (§4) and of the class of dynamical systems with odd monotone increasing feedback functions (§5).

Within this framework, therefore, the purpose of this paper is to investigate in detail the following problem:

Given. The completely controllable, completely observable, single-input, single-output continuous-time dynamical system (2.1) whose linearized system $f(\sigma) = K\sigma$ is asymptotically stable for all feedback gains K lying in the open sector $(0, \bar{K})$.

Find. Sufficient conditions for the absolute stability of this dynamical system when the open-loop system is asymptotically stable and the nonlinearity is (a) a monotone increasing feedback function and (b) an odd monotone increasing feedback function.

3. Mathematical preliminaries. The ideas required for the development of this paper are now presented through a series of three lemmas [12].

LEMMA 3.1. *If $-\eta$ and s are not eigenvalues of A then*

$$(s + \eta)(\eta I + A)^{-1}(sI - A)^{-1} = (sI - A)^{-1} + (\eta I + A)^{-1}. \tag{3.1}$$

Proof.

$$(s+\eta)(\eta I + A)^{-1}(sI - A)^{-1}$$
$$= (\eta I + A)^{-1}\{(\eta I + A) + (sI - A)\}\,(sI - A)^{-1}$$
$$= (sI - A)^{-1} + (\eta I + A)^{-1}.$$

If $G(s)$ has a zero at $s = -\eta$, then

$$G(s) = \frac{(s+\eta)(r_{n-1}\,s^{n-2} + r_{n-2}\,s^{n-3} + \cdots + r_2\,s + r_1)}{s^n + a_n\,s^{n-1} + \cdots + a_3\,s^2 + a_2\,s + a_1} = (s+\eta)R(s),$$

where $R(s) = r^T(sI - A)^{-1}b$ and $r = \text{col}\,(r_1\ r_2\ r_3\ \cdots\ r_{n-2}\ r_{n-1}\ 0)$.

The important result for transfer functions with nonzero zeros is stated by the following:

LEMMA 3.2. *If $G(s)$ has a zero at $s = -\eta$, then $h^T(\eta I + A)^{-1}b = 0$, and*

$$h^T(\eta I + A)^{-1}(sI - A)^{-1}b = \frac{G(s)}{s+\eta} \equiv R(s).$$

Proof. Premultiplying (3.1) by h^T and postmultiplying (3.1) by b yields

$$(s+\eta)h^T(\eta I + A)^{-1}(sI - A)^{-1}b = h^T(sI - A)^{-1}b + h^T(\eta I + A)^{-1}b,$$

so that

$$h^T(\eta I + A)^{-1}(sI - A)^{-1}b = \frac{G(s)}{s+\eta}.$$

Finally, Lemma 3.1 is applied to yield the analogous lemma for finite sector problems.

LEMMA 3.3. *If $h^T(\eta I + A)^{-1}b = 1/\bar{K}$, then*

$$h^T(\eta I + A)^{-1}(sI - A)^{-1}b = \frac{G(s) + 1/\bar{K}}{s+\eta},$$

where $s = -\eta$ is a zero of $G(s) + 1/\bar{K}$.

The proof of Lemma 3.3 is identical to that of Lemma 3.2; the algebraic details are omitted here.

4. Monotone increasing functions. The second method of Lyapunov is now employed to derive sufficient conditions for the absolute stability of the dynamical system (2.1) with monotone increasing feedback functions. For this purpose introduce as a candidate for the Lyapunov function

$$V(x) = \frac{1}{2}x^TPx + \beta_0\int_0^{h^Tx} f(\zeta)\,d\zeta + \sum_{i=1}^{v}\beta_i\int_0^{r_i^Tx} f(\zeta)\,d\zeta, \qquad (P = P^T \geqq 0), \tag{4.1}$$

where all scalar multipliers (Greek letters) are nonnegative and the sequence of n-vectors $\{r_i\}$ is to be determined. The infinite and finite sector problems are treated separately.

The following statement of the Meyer-Kalman-Yakubovich (MKY) Lemma [5] is employed throughout the development of this paper.

LEMMA. *Let A be a real $n \times n$ matrix all of whose characteristic roots have negative real parts; let γ be a real nonnegative number, and let b and k be two real n-vectors. If*

$$\gamma + \operatorname{Re}\{k^T(i\omega I - A)^{-1}b\} \geqq 0 \tag{4.2}$$

for all real ω, then there exist two real $n \times n$ symmetric matrices P and D and a real n-vector q such that

(i) $PA + A^TP = -2qq^T - 2D$,

(ii) $Pb - k = 2\sqrt{\gamma}q$,

(iii) *D is positive semidefinite and P is positive definite,*

(iv) $\{x \in R^n: x^TDx = 0\} \cap \{x \in R^n: q^Te^{At}x \equiv 0\} = \{0\}$.

4.1. The infinite sector problem for monotone increasing functions. First, the following sufficient conditions are derived for the absolute stability of the system (2.1) in the case where the linearized system $f(\sigma) = K\sigma$ is asymptotically stable for all positive feedback gains K.

THEOREM 4.1. *Consider the single-input, single-output, completely controllable, completely observable dynamical system (2.1) where A is a real $n \times n$ stable matrix, the linearized system is asymptotically stable for all $K > 0$, and $f(\cdot)$ is a monotone increasing function such that*

$$f(0) = 0, \qquad \sigma f(\sigma) > 0 \quad \text{for} \quad \sigma \neq 0.$$

Let $\eta_1, \eta_2, \cdots, \eta_v$ be real numbers such that $h^T(\eta_i I + A)^{-1}b = 0$ for $i = 1, 2, \cdots, v$. Then the dynamical system (2.1) is absolutely stable if there exist nonnegative constants $\alpha, \beta_0, \beta_1, \cdots, \beta_v, \gamma_1, \cdots, \gamma_v$, such that

(a) $\beta_i\eta_i - \gamma_i = \epsilon_i \geqq 0$; $\beta_i = 0$ *if and only if* $\gamma_i = 0$, $i = 1, 2, \cdots, v$; $\alpha \neq 0$;

$$\text{(b)} \qquad \operatorname{Re}\left\{\alpha + \beta_0 s + \sum_{i=1}^{v} \gamma_i \left(1 - \frac{\gamma_i}{\beta_i} \frac{1}{(s + \eta_i)}\right)\right\} G(s) \geqq 0.$$

Proof. Since $h^TA(sI - A)^{-1}b = sh^T(sI - A)^{-1}b - h^Tb$, Lemma 3.2 shows that hypothesis (b) is equivalent to

$$\beta_0 h^Tb + \operatorname{Re}\left\{\alpha h^T + \beta_0 h^TA + \sum_{i=1}^{v} \gamma_i \left[h^T - \frac{\gamma_i}{\beta_i} h^T(\eta_i I + A)^{-1}\right]\right\} (sI - A)^{-1}b \geqq 0.$$

Let $r_i^T = (\gamma_i/\beta_i)h^T(\eta_i I + A)^{-1}$, where $s = -\eta_i$ is a zero of $G(s)$; so that $r_i^T b = 0$, $i = 1, 2, \cdots, v$.

The MKY Lemma shows that there exists a positive definite matrix P and an n-vector q such that

$$PA + A^T P = -2qq^T - 2D, \tag{4.3}$$

$$\left[Pb - \left\{\alpha h + \beta_0 A^T h + \sum_{i=1}^{v} \gamma_i(h - r_i)\right\}\right] = 2\sqrt{\gamma}q, \tag{4.4}$$

where $\gamma = \beta_0 h^T b$. Since $G(s)$ is a minimum phase transfer function, $h^T b \geqq 0$.

Hypothesis (b) and the MKY Lemma demonstrate that the function (4.1) is positive definite.

The time derivative of (4.1) along any motion of system (2.1) is

$$\begin{aligned}\dot{V}(x) = {} & \tfrac{1}{2}x^T(PA + A^T P)x - f(h^T x)x^T\{Pb - \beta_0 A^T h\} - \beta_0 h^T b\{f(h^T x)\}^2 \\ & + \sum_{i=1}^{v} \beta_i r_i^T A x f(r_i^T x) - \sum_{i=1}^{v} \beta_i r_i^T b f(h^T x) f(r_i^T x).\end{aligned} \tag{4.5}$$

By adding and subtracting $\alpha\sigma f(\sigma)$, $\alpha > 0$, and the semidefinite quantity

$$\sum_{i=1}^{v} \gamma_i\{f(h^T x) - f(r_i^T x)\}(h - r_i)^T x \geqq 0$$

to (4.5), this time-derivative is rewritten

$$\begin{aligned}\dot{V}(x) = {} & \frac{1}{2}x^T(PA + A^T P)x \\ & - \left[Pb - (\alpha h + \beta_0 A^T h + \sum_{i=1}^{v} \gamma_i(h - r_i)\right]^T x f(h^T x) \\ & - \beta_0 h^T b f^2(h^T x) - \alpha\sigma f(\sigma) - \sum_{i=1}^{v} \gamma_i(h - r_i)^T x\{f(h^T x) - f(r^T x)\} \\ & - \sum_{i=1}^{v} \beta_i r_i^T b f(h^T x) f(r_i^T x) + \sum_{i=1}^{v} [\beta_i r^T A - \gamma_i(h - r_i)^T] x f(r_i^T x).\end{aligned} \tag{4.6}$$

The definition of r_i implies

$$\beta_i r_i^T A + \gamma_i(r_i - h)^T = -\epsilon_i r_i^T, \quad \beta_i \eta_i - \gamma_i = \epsilon_i, \quad i = 1, 2, \cdots, v. \tag{4.7}$$

By substituting (4.3), (4.4), and (4.7) into (4.6), this time-derivative of (4.1) becomes

$$\begin{aligned}\dot{V}(x) = {} & -x^T qq^T x - 2\sqrt{\gamma}q^T x f(h^T x) - \gamma f^2(h^T x) - \alpha\sigma f(\sigma) \\ & - x^T D x - \sum_{i=1}^{v} \gamma_i(h - r_i)^T x\{f(h^T x) - f(r^T x)\} - \sum_{i=1}^{v} \epsilon_i r_i^T x f(r_i^T x).\end{aligned} \tag{4.8}$$

Upon completing the square, this time-derivative (4.8) of the function (4.1) is written

$$\dot{V}(x) = -(q^T x + \sqrt{\gamma} f(\sigma))^2 - \alpha\sigma f(\sigma) - x^T D x - \sum_{i=1}^{v} \gamma_i (h - r_i)^T x \{f(h^T x) - f(r^T x)\} - \sum_{i=1}^{v} \epsilon_i r_i{}^T x f(r_i{}^T x). \tag{4.9}$$

Lemma 2.1 shows that this time-derivative is nonpositive definite. It therefore remains to demonstrate that the only solution of (2.1) that remains in the set where $\dot{V}(x) = 0$ is the null solution $x(t) \equiv 0$. Assume $x(t)$ is a solution of (2.1) that remains in the set where $\dot{V}(x) = 0$ for all t and $x(0) = x_0$. From the second term $\sigma(t) = h^T x = 0$. Thus, $x(t)$ is a solution of $\dot{x} = Ax$ so that $x(t) = e^{At} x_0$. From the first term, $q^T e^{At} x_0 = 0$. From the third term, $x' D x = 0$ and so by part (iv) of the MKY lemma $x_0 = 0$. Thus the only solution of (2.1) that remains in the set where $\dot{V} \equiv 0$ is the trivial solution $x(t) \equiv 0$.

Hence, $V(x)$ defined by (4.1) is a Lyapunov function which demonstrates the absolute stability of the dynamical system (2.1) for all monotone increasing feedback functions (2.4).

Comments. (i) Physically, hypothesis (b) of Theorem 4.1 requires that the product of the plant transfer function $G(s)$ and the $\{\ \}$ multiplier be a positive real function and therefore be realizable as the driving-point impedance of a passive electrical network. Moreover, since

$$1 - \frac{\gamma_i}{\beta_i \eta_i} = \frac{\epsilon_i}{\beta_i \eta_i} \geqq 0,$$

this multiplier is recognized to be the partial-fraction expansion of the driving point impedance of an RL electrical network and may be rewritten

$$\left\{\alpha + \beta_0 s + \sum_{i=1}^{v} \gamma_i \left(\frac{s + c_i \eta_i}{s + \eta_i}\right)\right\},$$

where $0 \leqq c_i \leqq 1$, $i = 1, 2, \cdots, v$.

(ii) It is interesting to note that the Popov multiplier $\alpha + \beta s$ is the driving point impedance of the RL network composed of a series resistor-inductor combination. The phase angle of this multiplier starts at zero and increases monotonically to 90°. The phase angle of the multiplier used in Theorem 4.1 also lies between 0 and 90° but is not a monotone increasing function of frequency. Thus, a necessary condition for the application of the Popov theorem and Theorem 4.1 is that the Nyquist plot of $G(s)$ be restricted to the first, third, and fourth quadrants.

(iii) In Theorem 4.1 there is an n-vector r_i and there are scalars β_i, γ_i, and ϵ_i for each real nonzero zero of $G(s)$. Thus, these results are applicable to systems whose plant transfer functions have some real nonzero zeros.

The application of the systematic procedure of Theorem 4.1 for the construction of the explicit Lyapunov function (4.1) is now illustrated by the following example. Only the main results of this example are presented here; the tedious algebraic calculations have been omitted.

Example 4.1. *Fourth order system with two real zeros.* Consider the single degree of freedom dynamical system described by

$$\dddot{x} + (a + b)\ddot{x} + (ab + c)\dot{x} + c(a + b)\dot{x} + abcx + f[h_3(\alpha\beta x + (\alpha + \beta)\dot{x} + \ddot{x}] = 0, \tag{4.10}$$

where

$$b > \beta > a > \alpha > 0.$$

When linear feedback is employed, the Routh-Hurwitz conditions indicate that the linearized system is asymptotically stable for all positive feedback gains. The purpose of this example is to determine whether the nonlinear system (4.10) is absolutely stable for all monotone increasing feedback functions.

For this problem

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -abc & -c(a+b) & -(ab+c) & -(a+b) \end{bmatrix},$$

$$b = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad h = h_3 \begin{bmatrix} \alpha\beta \\ \alpha + \beta \\ 1 \\ 0 \end{bmatrix},$$

so that

$$G(s) = \frac{h_3(s + \alpha)(s + \beta)}{(s^2 + c)(s + a)(s + b)}.$$

Since $b > \beta > a > \alpha > 0$, a suitable RL driving point impedance multiplier is

$$Z_{RL}^{(s)} = \frac{s(s + a)(s + b)}{(s + \alpha)(s + \beta)},$$

and since $G(s)Z_{RL}^{(s)} = h_3 s/(s^2 + c)$ is an LC driving point impedance (a positive real function), the hypotheses of Theorem 4.1 are satisfied and (4.10) is absolutely stable for all monotone increasing feedback nonlinearities. Moreover, the Lyapunov function which demonstrates the absolute stability of (4.10) for all monotone increasing feedback nonlinearities is

$$
\begin{aligned}
V(x) &= \frac{h_3}{2}\, x' \\
&\cdot \begin{bmatrix} a^2b^2c & abc(a+b) & abc & 0 \\ abc(a+b) & a^2b^2 + c(a+b)^2 & (a+b)(ab+c) & ab \\ abc & (a+b)(ab+c) & (a+b)^2 + c & (a+b) \\ 0 & ab & (a+b) & 1 \end{bmatrix} x \\
&+ \int_0^{h_3[\alpha\beta x_1 + (\alpha+\beta)x_2 + x_3]} f(\zeta)\, d\zeta \\
&+ \frac{(a-\alpha)(b-\alpha)}{\alpha(\beta-\alpha)} \int_0^{\alpha h_3(\beta x_1 + x_2)} f(\zeta)\, d\zeta \\
&+ \frac{(b-\beta)(\beta-a)}{\beta(\beta-\alpha)} \int_0^{\beta h_3(\alpha x_1 + x_2)} f(\zeta)\, d\zeta.
\end{aligned} \tag{4.11}
$$

The time-derivative of this Lyapunov function along a motion of system (4.10) is

$$
\begin{aligned}
\dot{V}(x) = &-\frac{(a-\alpha)(b-\alpha)}{(\beta-\alpha)} \{f[h_3(\alpha\beta x_1 + (\alpha+\beta)x_2 + x_3)] \\
&- f(\alpha h_3(\beta x_1 + x_2))]\} h_3(\beta x_2 + x_3) \\
&- \frac{(b-\beta)(\beta-a)}{(\beta-\alpha)} \{f[h_3(\alpha\beta x_1 + (\alpha+\beta)x_2 + x_3)] \\
&- f[\beta h_3(\alpha x_1 + x_2)]\} h_3(\alpha x_2 + x_3).
\end{aligned} \tag{4.12}
$$

4.2. The finite sector problem for monotone increasing functions. Next, these ideas are extended to the case where the linearized system is asymptotically stable for all feedback gains K lying in the finite open sector $(0, \bar{K})$.

THEOREM 4.2. *Consider the single-input, single-output, completely controllable, completely observable dynamical system* (2.1) *whose linearized system is asymptotically stable for all feedback gains K lying in the finite open sector $(0, \bar{K})$. A is a real $n \times n$ stable matrix and $f(\cdot)$ is a monotone increasing function satisfying*

$$
f(0) = 0, \qquad 0 < \sigma f(\sigma) < \bar{K}\sigma^2 \quad \text{for} \quad \sigma \neq 0. \tag{4.13}
$$

Let $\eta_1, \eta_2, \cdots, \eta_v$ be real numbers such that $\bar{K}h^T(\eta_i I + A)^{-1}b = 1$ for $i = 1, 2, \cdots, v$. Then the dynamical system (2.1) *is absolutely stable for all monotone increasing feedback functions* (4.13) *if there exist nonnegative constants $\alpha, \beta_0, \beta_1, \cdots, \beta_v, \gamma_1, \cdots, \gamma_v$ such that*
(a) $\beta_i\eta_i - \gamma_i = \epsilon_i \geqq 0$; $\beta_i = 0$ *if and only if* $\gamma_i = 0$, $i = 1, 2, \cdots, v$; $\alpha \neq 0$;

(b) $$\operatorname{Re}\left\{\left(G(s)+\frac{1}{\bar{K}}\right)\left[\alpha+\beta_0 s+\sum_{i=1}^{v}\gamma_i\left(1-\frac{\gamma_i}{\beta_i}\frac{1}{s+\eta_i}\right)\right]\right\}-\frac{1}{\bar{K}}\sum_{i=1}^{v}\gamma_i \geqq 0.$$

Proof. Lemma 3.3 demonstrates that hypothesis (b) is equivalent to

$$\frac{\alpha}{\bar{K}}+\beta_0 h^T b+\operatorname{Re}\left\{\alpha h^T+\beta_0 h^T A+\sum_{i=1}^{v}\gamma_i\left[h^T-\frac{\gamma_i}{\beta_i}h^T(\eta_i I+A)^{-1}\right]\right\}(sI-A)^{-1}b \geqq 0.$$

Define $r_i{}^T = (\gamma_i/\beta_i)h^T(\eta_i I + A)^{-1}$, where $s = -\eta_i$ is a zero of $G(s) + 1/\bar{K}$; so that $r_i{}^T b = (\gamma_i/\beta_i)(1/\bar{K})$; $i = 1, 2, \cdots, v$.

The MKY Lemma shows that there exist a positive definite matrix P and an n-vector q such that

$$PA + A^T P = -2qq^T - 2D \tag{4.14}$$

and

$$Pb - \{\alpha h + \beta_0 A^T h + \sum_{i=1}^{v}\gamma_i(h - r_i)\} = 2\sqrt{\gamma}\, q, \tag{4.15}$$

where $\gamma = \beta_0 h^T b + \alpha/\bar{K} \geqq 0$.

By the development of this section, the time-derivative of the function (4.1) may be written

$$\begin{aligned}\dot{V}(x) = {} & \frac{1}{2}x^T(PA + A^T P)x - [Pb - (\alpha h + \beta_0 A^T h \\ & + \sum_{i=1}^{v}\gamma_i(h - r_i))]^T x f(h^T x) \\ & - \left[\frac{\alpha}{\bar{K}} + \beta_0 h^T b\right] f^2(h^T x) - \alpha f(\sigma)\left[\sigma - \frac{f(\sigma)}{\bar{K}}\right] - \sum_{i=1}^{v}\epsilon_i r_i{}^T x f(r_i{}^T x) \\ & - \sum_{i=1}^{v}\gamma_i\left[(h - r_i)^T x\{f(h^T x) - f(r_i{}^T x)\} + \frac{1}{\bar{K}}f(h^T x)f(r_i{}^T x)\right].\end{aligned} \tag{4.16}$$

Substitution of (4.14) and (4.15) into (4.16) and completion of the square yields

$$\begin{aligned}\dot{V}(x) = {} & -[q^T x + \sqrt{\gamma}\, f(\sigma)]^2 - \alpha f(\sigma)\left[\sigma - \frac{f(\sigma)}{\bar{K}}\right] \\ & - x^T D x - \sum_{i=1}^{v}\epsilon_i r_i{}^T x f(r_i{}^T x) \\ & - \sum_{i=1}^{v}\gamma_i\left[(h - r_i)^T x\{f(h^T x) - f(r_i{}^T x)\} + \frac{1}{\bar{K}}f(h^T x)f(r_i{}^T x)\right].\end{aligned} \tag{4.17}$$

Since the monotone increasing nonlinearity is restricted to lie in the finite sector $(0, \bar{K})$, the last term

$$-\left[(h - r_i)^T x\{f(h^T x) - f(r_i{}^T x)\} + \frac{1}{\bar{K}} f(h^T x) f(r_i{}^T x)\right]$$

is nonpositive definite for all $h^T x$ and $r_i{}^T x$, $i = 1, 2, \cdots, v$. Moreover, the MKY Lemma demonstrates that the only solution of (2.1) that remains in the set where this time-derivative of (4.1) vanishes is the null solution $x(t) \equiv 0$. Thus the system (2.1) is absolutely stable for all monotone increasing nonlinearities lying in the finite open sector $(0, \bar{K})$.

Comments. (i) Hypothesis (a) shows that the multiplier used in Theorem 4.2 is the partial-fraction expansion of the driving point impedance of an RL network.

(ii) Since the stability properties of a system with the transfer function $G(s)$ in the forward path and a monotone increasing feedback function (4.13) are completely equivalent to those of a system with the plant transfer function $1/[G(s) + 1/\bar{K}]$ and any monotone increasing feedback function, Theorem 4.1 and Theorem 4.2 demonstrate that a theorem, completely analogous to Theorem 4.2, can be stated for a multiplier which is the driving point impedance of an RC electrical network whose zeros are the real poles of $G(s) + 1/\bar{K}$.

5. Odd monotone increasing functions. The additional assumption of an odd monotone increasing feedback function leads to even less restrictive conditions on the linear part of the system than the conditions derived in §4. In order to derive frequency domain stability criteria for the absolute stability of the dynamical system (2.1) with odd monotone increasing feedback functions, employ the Lyapunov function

$$V(x) = \frac{1}{2} x^T P x + \beta_0 \int_0^{h^T x} f(\zeta)\, d\zeta + \sum_{i=1}^{v_1} \beta_i \int_0^{r_i{}^T x} f(\zeta)\, d\zeta + \sum_{j=1}^{v_2} \beta_j{}' \int_0^{r_j{}'{}^T x} f(\zeta)\, d\zeta, \tag{5.1}$$

where all scalar multipliers (Greek letters) are nonnegative and the sequences of n-vectors $\{r_i\}$ and $\{r_j'\}$ are specified by (4.7).

Theorems applicable to infinite and finite sector problems are presented separately. Since the proofs of these theorems are completely analogous to those of Theorem 4.1 and Theorem 4.2, the theorems presented in this section are stated without proof.

Lemma 2.2 and Lemma 2.3 demonstrate that in the case of odd monotone nonlinearities, the sequence of scalars $\{\gamma_i\}$ employed in Theorem 4.1 and Theorem 4.2 may be either positive or negative.

THEOREM 5.1 (*Infinite sector*). *Consider the single-input, single-output,*

completely controllable, completely observable dynamical system (2.1) *whose linearized system is asymptotically stable for all positive feedback gains K. A is a real* $n \times n$ *stable matrix and* $f(\cdot)$ *is an odd monotone increasing function such that*

$$f(0) = 0, \qquad \sigma f(\sigma) > 0 \quad \text{for} \quad \sigma \neq 0.$$

Let $\{\eta_i\}$ *and* $\{\eta_j'\}$ *be finite sequences of real numbers such that* $h^T(\eta_i I + A)^{-1}b = 0$ *for* $i = 1, 2, \cdots, v_1$ *and* $h^T(\eta_j' I + A)^{-1}b = 0$ *for* $j = 1, 2, \cdots, v_2$. *Then the dynamical system* (2.1) *is absolutely stable if there exist nonnegative constants* α *and* β_0 *and finite sequences of nonnegative constants* $\{\beta_i\}$, $\{\beta_j'\}$, $\{\gamma_i\}$, $\{\gamma_j'\}$ *such that*

(a) $\beta_i\eta_i - \gamma_i = \epsilon_i \geqq 0$; $\beta_i = 0$ *if and only if* $\gamma_i = 0$, $i = 1, 2, \cdots, v_1$; $\alpha \neq 0$; $\beta_j'\eta_j' - \gamma_j' = \epsilon_j' \geqq 0$; $\beta_j' = 0$ *if and only if* $\gamma_j' = 0$, $j = 1, 2, \cdots, v_2$;

(b)
$$\operatorname{Re}\left\{G(s)\left[\alpha + \beta_0 s + \sum_{i=1}^{v_1}\gamma_i\left(1 - \frac{\gamma_i}{\beta_i}\frac{1}{s+\eta_i}\right) + \sum_{j=1}^{v_2}\gamma_j'\left(1 + \frac{\gamma_j'}{\beta_j'}\frac{1}{s+\eta_j'}\right)\right]\right\} \geqq 0.$$

By the hypothesis of Theorem 5.1 and application of the MKY Lemma, the negative semidefinite upper bound of the time-derivative of the Lyapunov function (5.1) is

$$\dot{V}(x) \leqq -[q^Tx + x\sqrt{\gamma}f(\sigma)]^2 - \sum_{i=1}^{v_1}\gamma_i(h - r_i)^Tx[f(h^Tx) - f(r_i^Tx)]$$
$$- x^TDx - \alpha\sigma f(\sigma) - \sum_{i=1}^{v_1}\epsilon_i r_i^T x f(r_i^Tx) - \sum_{j=1}^{v_2}\epsilon_j' r_j'^T x f(r_j'^Tx),$$

where $\gamma = \beta_0 h^Tb \geqq 0$.

The MKY Lemma shows that the only solution of (2.1) that remains in the set where this upper bound of the time-derivative of the Lyapunov function (5.1) vanishes is the null solution $x(t) \equiv 0$ so that the system is absolutely stable.

Comments. The multiplier of Theorem 5.1 is the partial-fraction expansion of the driving point impedance of a special class of RLC electrical networks. This class of driving point impedance has the representation

$$\hat{Z}_{RLC}^{(s)} = R_\infty + sL + Z_{RL}^{(s)} + Z_{RC}^{(s)} \tag{5.2}$$

and has poles which are the real nonzero zeros of the plant transfer function $G(s)$. Thus, Theorem 5.1 may be applied to the class of systems (2.1) whose linear plants have transfer functions with Nyquist plots lying in all four quadrants.

Moreover, hypothesis (a) shows that the multiplier used in Theorem 5.1 may be rewritten

$$\alpha + \beta_0 s + \sum_{i=1}^{v_1} \gamma_i \left(\frac{s + c_i \eta_i}{s + \eta_i} \right) + \sum_{j=1}^{v_2} \gamma_j' \left(\frac{s + c_j' \eta'}{s + \eta_j'} \right),$$

where $0 \leqq c_i \leqq 1$, $i = 1, 2, \cdots, v_1$, and $1 \leqq c_j' \leqq 2$, $j = 1, 2, \cdots, v_2$.

The following example is presented in order to illustrate the application of Theorem 5.1.

Example 5.1. Let

$$G(s) = \frac{s(s + \alpha a)}{(s^2 + b)(s + a)}$$

so that a suitable multiplier is

$$Z(s) = \frac{s + a}{s + \alpha a} = 1 + \frac{a(1 - \alpha)}{s + \alpha a}.$$

If $\alpha < 1$, Theorem 5.2 shows that this system is absolutely stable for all odd monotone increasing feedback functions. If, however, $\alpha > 1$, Theorem 4.1 demonstrates that this system is absolutely stable for all monotone increasing feedback functions.

THEOREM 5.1 (*Finite sector*). *Consider the single-input, single-output, completely controllable, completely observable dynamical system* (2.1) *whose linearized system is asymptotically stable for all feedback gains* K *lying in the finite open sector* $(0, \bar{K})$. *A is a real $n \times n$ stable matrix and $f(\cdot)$ is an odd monotone increasing function satisfying*

$$f(0) = 0, \qquad 0 < \sigma f(\sigma) < \bar{K}\sigma^2 \quad \text{for} \quad \sigma \neq 0. \tag{5.3}$$

Let $\{\eta_i\}$ and $\{\eta_j'\}$ be finite sequences of real numbers such that $\bar{K}h^T(\eta_i I + A)^{-1}b = 1$ for $i = 1, 2, \cdots, v_1$ and $\bar{K}h^T(\eta_j' I + A)^{-1}b = 1$ for $j = 1, 2, \cdots, v_2$. Then the dynamical system (2.1) *is absolutely stable for all odd monotone increasing functions* (5.3) *if there exist nonnegative constants α and β_0 and finite sequences of nonnegative constants $\{\beta_i\}$, $\{\beta_j'\}$, $\{\gamma_i\}$ and $\{\gamma_j'\}$ such that*

(a) $\beta_i\eta_i - \gamma_i = \epsilon_i \geqq 0$; $\beta_i = 0$ *if and only if* $\gamma_i = 0$, $i = 1, 2, \cdots, v_1$; $\beta_j'\eta_j' - \gamma_j' = \epsilon_j' \geqq 0$; $\beta_j' = 0$ *if and only if* $\gamma_j' = 0$, $j = 1, 2, \cdots, v_2$; $\alpha \neq 0$;

(b)
$$\operatorname{Re}\left\{\left(G(s) + \frac{1}{\bar{K}}\right)\left[\alpha + \beta_0 s + \sum_{i=1}^{v_1} \gamma_i \left(1 - \frac{\gamma_i}{\beta_i} \frac{1}{s + \eta_i}\right) + \sum_{j=1}^{v_2} \gamma_j' \left(1 + \frac{\gamma_j'}{\beta_j'} \frac{1}{s + \eta_j'}\right)\right]\right\} - \frac{1}{\bar{K}}\left(\sum_{i=1}^{v_1} \gamma_i + \sum_{i=1}^{v_2} \gamma_j'\right) \geqq 0.$$

By the hypotheses of Theorem 5.2 and application of the MKY Lemma, the negative semidefinite upper bound of the time-derivative of the Lyapunov function (5.1) is

$$\dot{V}(x) \leqq -[q^T x + \sqrt{\gamma}\, f(\sigma)]^2 - \sum_{i=1}^{v_1} \epsilon_i\, r_i{}^T x f(r_i{}^T x) - \alpha f(\sigma)\left[\sigma - \frac{f(\sigma)}{\bar{K}}\right]$$

$$- x^T D x - \sum_{i=1}^{v_1} \gamma_i \left[(h - r_i)^T x\{f(h^T x) - f(r_i{}^T x)\} + \frac{1}{\bar{K}} f(h^T x) f(r_i{}^T x)\right]$$

$$- \sum_{j=1}^{v_2} \epsilon_j'\, r_j'{}^T x f(r_j'{}^T x),$$

where $\gamma = \beta_0 h^T b + \alpha/\bar{K}$.

Moreover, the MKY Lemma shows that the only solution of (2.1) that remains in the set where this upper bound vanishes is the null solution $x(t) \equiv 0$. Thus, the system (2.1) is absolutely stable for all odd monotone increasing feedback functions lying in the finite open sector $(0, \bar{K})$.

Comments. Hypothesis (a) demonstrates that the multiplier employed in Theorem 5.2 is the partial-fraction expansion of the special class of RLC networks having the driving point impedance $\hat{Z}_{\text{RLC}}^{(s)}$ represented by (5.2).

6. Acknowledgment. The authors are indebted to the referee who is responsible for many corrections and improvements in this paper.

REFERENCES

[1] V. M. Popov, *Absolute stability of nonlinear systems of automatic control*, Automat. Remote Control, 22 (1962), pp. 857–875.

[2] K. S. Narendra and R. M. Goldwyn, *Existence of quadratic type Liapunov functions for a class of nonlinear systems*, Internat. J. Engrg. Sci., 2 (1964), pp. 367–377.

[3] V. M. Popov and A. Halaney, *On the stability of nonlinear automatic control systems with lagging arguments*, Automat. Remote Control, 23 (1963), pp. 783–786.

[4] Z. V. Rekasius, *A stability criterion for feedback systems with one nonlinear element*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 46–50.

[5] K. R. Meyer, *Liapunov functions for the problem of Lur'e*, Proc. Nat. Acad. Sci. U.S.A., 53 (1965), pp. 501–503.

[6] R. E. Kalman, *Lyapunov functions for the problem of Lur'e in automatic control*, Ibid., 49 (1963), pp. 201–205.

[7] V. A. Yakubovich, *The solution of certain matrix inequalities in automatic control theory*, Dokl. Akad. Nauk SSSR, 143 (1962), pp. 1304–1307.

[8] G. Zames, *Nonlinear, time-varying systems—contracting transformations for iteration and stability*, to appear.

[9] R. W. Brockett and J. L. Willems, *Frequency domain stability criteria*, Joint Automatic Control Preprints (1965).

[10] R. E. Kalman, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

[11] ———, *When is a linear control system optimal?*, Trans. ASME Ser. D. J. Basic Engrg., 86D (1964), pp. 51–60.

[12] K. S. Narendra and C. P. Neuman, *Stability of a class of discrete time systems with a single feedback nonlinearity*, to appear.

J. SIAM CONTROL
Vol. 4, No. 2, 1966
Printed in U.S.A.

THE SECOND VARIATION FOR THE SINGULAR BOLZA PROBLEM*

B. S. GOH†

1. Introduction. The Bolza problem of the calculus of variations may be formulated thus: $y_i(x)$, $i = 1, 2, \cdots, n$, is a set of n functions defined over an interval $a < x < b$ and satisfying differential constraints

$$\phi_\beta(x, y, y') = 0, \qquad \beta = 1, 2, \cdots, m < n, \tag{1}$$

where y, y' denote the whole set of functions and their derivatives. The derivatives $y_i'(x)$ are assumed continuous except for a finite number of finite discontinuities. x_1, x_2 are endpoints satisfying $a < x_1 < x_2 < b$. The values of the functions at these endpoints are required to satisfy end conditions

$$\psi_\mu[x_1, y(x_1), x_2, y(x_2)] = 0, \qquad \mu = 1, 2, \cdots, p \leqq 2n + 2. \tag{2}$$

The problem is to determine the set of functions (if it exists) and endpoints satisfying these end conditions and the differential constraints, which minimizes a quantity J given by

$$J = g[x_1, y(x_1), x_2, y(x_2)] + \int_{x_1}^{x_2} f(x, y, y')\, dx, \tag{3}$$

where g and f are known functions.

Bliss [1] shows that, assuming the functions f, g, ϕ_β, ψ_μ satisfy certain conditions of a general nature, a necessary condition to be satisfied by functions $y_i(x)$ and endpoints x_1, x_2 minimizing J, is the so-called multiplier rule. He also derives further necessary conditions associated with the names of Weierstrass and Clebsch. By consideration of the second variation of J, a fourth necessary condition is obtained. In this paper, we shall examine this last condition.

Given functions $y_i(x)$ define an arc E in an $(n + 1)$-dimensional space in which $(x, y_1, y_2, \cdots, y_n)$ are coordinates. A set of admissible variations along E is a set of two quantities ξ_1, ξ_2 and n functions $\eta_i(x)$ satisfying the equations of variation

$$\frac{\partial \phi_\beta}{\partial y_i'} \eta_i' + \frac{\partial \phi_\beta}{\partial y_i} \eta_i = 0, \tag{4}$$

* Received by the editors September 7, 1965.

† Department of Mathematics, University of Canterbury, Christchurch, New Zealand.

$$(5)\quad \left(\frac{\partial \psi_\mu}{\partial x_1} + y'_{i1}\frac{\partial \psi_\mu}{\partial y_{i1}}\right)\xi_1 + \frac{\partial \psi_\mu}{\partial y_{i1}}\eta_{i1} + \left(\frac{\partial \psi_\mu}{\partial x_2} + y'_{i2}\frac{\partial \psi_\mu}{\partial y_{i2}}\right)\xi_2 + \frac{\partial \psi_\mu}{\partial y_{i2}}\eta_{i2} = 0,$$

where the usual repeated subscript summation convention has been employed and $y_{i1} = y_i(x_1)$, $y_{i2} = y_i(x_2)$. The partial derivatives in (4) and (5) are to be calculated along E. It is demonstrated by Bliss that, provided the arc E satisfies a normality condition (i.e., the multiplier l_0 occurring in his multiplier rule does not vanish) variations satisfying the equations of variation (4), (5) certainly exist. Given such a set of variations along a minimizing arc E for which the second derivatives $y_i''(x)$ are continuous (i.e., E is an extremal) the second variation of J with respect to the variations is given by

$$(6)\quad J_2 = 2\gamma[\xi_1, \eta(x_1), \xi_2, \eta(x_2)] + \int_{x_1}^{x_2} 2\omega(x, \eta, \eta')\, dx,$$

where 2γ is a certain homogenous quadratic form in its arguments and

$$(7)\quad 2\omega = R_{ij}\eta_i'\eta_j' + 2Q_{ij}\eta_i'\eta_j + P_{ij}\eta_i\eta_j .$$

The coefficients R_{ij}, Q_{ij}, P_{ij} are evaluated along the arc E from the equations

$$(8)\quad R_{ij} = \frac{\partial^2 F}{\partial y_i' \partial y_j'}, \qquad Q_{ij} = \frac{\partial^2 F}{\partial y_i' \partial y_j}, \qquad P_{ij} = \frac{\partial^2 F}{\partial y_i \partial y_j},$$

where

$$(9)\quad F = f + l_\beta \phi_\beta ,$$

the l_β being multipliers occurring in the multiplier rule. The fourth necessary condition to be satisfied by E, if it is to minimize J, is that

$$(10)\quad J_2(\xi, \eta) \geqq 0$$

for all admissible variations ξ_1, ξ_2, $\eta_i(x)$.

A particular set of admissible variations satisfying the conditions (4), (5) is given by

$$(11)\quad \xi_1 = 0,\ \xi_2 = 0,\ \eta_i(x) \equiv 0.$$

For this set of variations J_2 vanishes. Hence, if E satisfies the condition (10), this set of variations must minimize J_2. We are accordingly led to consider the *accessory minimizing problem*, viz., to minimize J_2 with respect to the set of functions $\eta_i(x)$ satisfying the differential and end constraints (4), (5). The minimizing set (11) must satisfy the Clebsch condition for this accessory problem. The function F occurring in the multiplier rule (see (9)) takes the following form for the accessory problem:

$$F = 2\omega + \lambda_\beta \left(\frac{\partial \phi_\beta}{\partial y_i'} \eta_i' + \frac{\partial \phi_\beta}{\partial y_i} \eta_i \right). \tag{12}$$

Hence,

$$\frac{\partial^2 F}{\partial \eta_i' \partial \eta_j'} = R_{ij}, \tag{13}$$

and the Clebsch condition requires that

$$R_{ij}\pi_i\pi_j \geqq 0 \tag{14}$$

at each point of E for every set $(\pi_1, \pi_2, \cdots, \pi_n)$ satisfying the constraints

$$\frac{\partial \phi_\beta}{\partial y_i'} \pi_i = 0. \tag{15}$$

This condition is identical with the Clebsch condition for the original Bolza problem, so that this approach to the accessory problem provides no fresh information. However, if the Bolza problem happens to be *singular*, it will be shown that the second variation may be transformed into a new form for which this approach yields an independent necessary condition.

A minimizing arc E is said to be singular in the following circumstances: Let R be the $n \times n$ matrix with elements R_{ij} and let ϕ denote the $m \times n$ matrix with elements $\phi_{\beta i} = \partial \phi_\beta / \partial y_i'$. Then, if ϕ^T denotes the transpose of ϕ and we form the $(m + n) \times (m + n)$ determinant

$$\Delta = \begin{vmatrix} R & \phi^T \\ \phi & 0 \end{vmatrix}, \tag{16}$$

0 being the zero $m \times m$ matrix, then E is singular if Δ vanishes at any point on it. We shall study the case where Δ vanishes at every point of E.

2. Elementary illustration. Consider the problem of minimizing the integral

$$J = \int_a^b (P + Qy')\, dx \tag{17}$$

with respect to the function $y(x)$, where

$$P = P(x, y), \qquad Q = Q(x, y), \tag{18}$$

and y is to satisfy fixed end conditions

$$y(a) = A, \qquad y(b) = B. \tag{19}$$

There are no differential constraints so that

$$\Delta = R = 0 \tag{20}$$

identically; the problem is accordingly singular.

The characteristic equation for the problem proves to be

$$\alpha = \frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x} = 0, \tag{21}$$

defining a single extremal in the xy-plane. Assuming that the end conditions are satisfied by this extremal, it remains to decide whether the extremal does, in fact, minimize J.

The Clebsch (Legendre) condition requires that

$$\frac{\partial^2}{\partial y'^2}(P + Qy') \geqq 0, \tag{22}$$

which is trivially satisfied.

The second variation for an admissible variation $\eta(x)$ with $\eta(a) = \eta(b) = 0$ is given by

$$J_2 = \int_a^b \left[2\frac{\partial Q}{\partial y}\eta\eta' + \left(\frac{\partial^2 P}{\partial y^2} + y'\frac{\partial^2 Q}{\partial y^2}\right)\eta^2 \right] dx. \tag{23}$$

Applying the Clebsch condition to the accessory minimizing problem yields no further information, as we expect. However, integrating the first term in the integrand of (23) by parts and employing the conditions $\eta(a) = 0 = \eta(b)$, we can put J_2 in the form

$$J_2 = \int_a^b \frac{\partial \alpha}{\partial y}\eta^2\, dx, \tag{24}$$

where α is given by (21). Application of the Clebsch condition to this new form of J_2 still yields no new condition. However, if we put

$$\eta(x) = \zeta'(x), \tag{25}$$

so that J_2 takes the form

$$J_2 = \int_a^b \frac{\partial \alpha}{\partial y}\zeta'^2\, dx \tag{26}$$

and now consider the accessory problem with regard to the unknown function $\zeta(x)$, it is easily seen that the Clebsch-Legendre condition requires that

$$\frac{\partial \alpha}{\partial y} \geqq 0. \tag{27}$$

This is a new condition which has already been obtained for this problem by Miele [2].

It should be noted that the end conditions $\eta(a) = \eta(b) = 0$ can be disregarded after J_2 has been expressed in the form (24), since any func-

tion $\eta(x)$ not satisfying these conditions can be replaced by a modified function $\bar{\eta}(x)$, differing from $\eta(x)$ only in arbitrary small neighborhoods of the endpoints and such that $\bar{\eta}(a) = \bar{\eta}(b) = 0$; such a replacement will result in a change in J_2 which is also arbitrarily small and hence can be disregarded for the purpose of the condition $J_2 \geqq 0$. It follows that $\zeta(x)$ is not required to satisfy end conditions on its derivative.

In the next section, we shall generalize the method illustrated above to be applicable to the general singular Bolza problem.

3. Transformation of the second variation. We now consider the second variation for the Bolza problem in the form given by (6). The problem will be supposed singular, such that Δ as given by (16) vanishes identically along the hypothetical minimizing arc.

Employing matrix notation, we can write

$$2\omega = \eta'^T R\eta' + 2\eta'^T Q\eta + \eta^T P\eta, \tag{28}$$

where η is an $n \times 1$ matrix with elements η_i and R, Q, P are $n \times n$ matrices with elements R_{ij}, Q_{ij}, P_{ij}, respectively. η^T denotes the transpose of η. The matrices R, P are clearly symmetric.

The differential constraints (4) can be written

$$\phi\eta' + \theta\eta = 0, \tag{29}$$

where ϕ, θ are $m \times n$ matrices with elements $\partial\phi_\beta/\partial y_i'$, $\partial\phi_\beta/\partial y_i$, respectively.

Suppose R is partitioned by lines running between its mth and $(m+1)$th rows and between its mth and $(m+1)$th columns thus:

$$R = \begin{pmatrix} R_2 & R_3^T \\ R_3 & R_1 \end{pmatrix}, \tag{30}$$

so that R_1 is an $(n-m) \times (n-m)$ matrix, R_3 is a $(n-m) \times m$ matrix and R_2 is an $m \times m$ matrix. ϕ is similarly partitioned thus:

$$\phi = (N \mid M), \tag{31}$$

where N is an $m \times m$ matrix and M is an $m \times (n-m)$ matrix. According to the hypothesis applying to the Bolza problem as formulated by Bliss, ϕ must be of rank m and hence we can arrange the nomenclature so that N is nonsingular. Thus N^{-1} exists and premultiplying (29) by this matrix, we obtain

$$(I_m \mid A)\eta' + B\eta = 0, \tag{32}$$

where I_m is the $m \times m$ unit matrix and

$$A = N^{-1}M, \tag{33}$$

$$B = N^{-1}\theta. \tag{34}$$

Partitioning η thus:

$$\eta = \begin{pmatrix} \rho \\ \pi \end{pmatrix}, \tag{35}$$

where ρ is an $m \times 1$ matrix and π is an $(n - m) \times 1$ matrix, we have

$$\begin{aligned} \eta'^T R \eta' &= (\rho'^T \mid \pi'^T) \begin{pmatrix} R_2 & R_3^T \\ R_3 & R_1 \end{pmatrix} \begin{pmatrix} \rho' \\ \pi' \end{pmatrix} \\ &= \rho'^T R_2 \rho' + \pi'^T R_3 \rho' + \rho'^T R_3^T \pi' + \pi'^T R_1 \pi'. \end{aligned} \tag{36}$$

Equation (32) can be written

$$(I_m \mid A) \begin{pmatrix} \rho' \\ \pi' \end{pmatrix} + B\eta = 0, \tag{37}$$

or

$$\rho' + A\pi' + B\eta = 0. \tag{38}$$

Eliminating ρ' between (36) and (38), we find that

$$\eta'^T R \eta' = \pi'^T R_1^* \pi' + \eta^T H \pi' + \pi'^T K \eta + \eta^T L \eta, \tag{39}$$

where

$$R_1^* = A^T R_2 A - R_3 A - A^T R_3^T + R_1, \tag{40}$$

$$H = B^T (R_2 A - R_3^T), \tag{41}$$

$$K = (A^T R_2 - R_3) B, \tag{42}$$

$$L = B^T R_2 B. \tag{43}$$

Substituting from (39) into (28), we see that it is possible to express 2ω in the form

$$2\omega = \eta'^T R^* \eta' + 2\eta'^T Q^* \eta + \eta^T P^* \eta, \tag{44}$$

where

$$R^* = \begin{pmatrix} 0 & 0 \\ 0 & R_1^* \end{pmatrix}, \tag{45}$$

$$P^* = P + L. \tag{46}$$

Clearly, R^* and P^* are symmetric.

We have now expressed the second variation J_2 in the form given in (6), but with the integrand 2ω in the form given by (44). The variations η_i are subjected to the differential constraints (38) and the end conditions

(5). We shall now prove that the new form of J_2 is singular if, and only if, the original form is singular.

We define an $(n + m) \times (n + m)$ matrix C in the partitioned form

$$C = \begin{pmatrix} D & -A & 0 \\ 0 & I_{n-m} & 0 \\ 0 & 0 & I_m \end{pmatrix}, \tag{47}$$

where $D = N^{-1}$. Taking determinants and expanding by minors in the lower n rows, we find that

$$|C| = |D| = |N|^{-1} \neq 0. \tag{48}$$

Thus, C is regular. Also

$$\begin{aligned} C^T \begin{pmatrix} R & \phi^T \\ \phi & 0 \end{pmatrix} C &= C^T \begin{pmatrix} R_2 & R_3{}^T & N^T \\ R_3 & R_1 & M^T \\ N & M & 0 \end{pmatrix} C, \\ &= \begin{pmatrix} E_1 & E_2 & E_3 \\ E_4 & E_5 & E_6 \\ I_m & E_7 & 0 \end{pmatrix}, \end{aligned} \tag{49}$$

where

$$E_1 = D^T R_2 D, \tag{50}$$

$$E_2 = D^T(-R_2 A + R_3{}^T), \tag{51}$$

$$E_3 = D^T N^T = (ND)^T = I_m, \tag{52}$$

$$E_4 = (-A^T R_2 + R_3)D = E_2{}^T \tag{53}$$

as R_2 is symmetric,

$$E_5 = A^T(R_2 A - R_3{}^T) - R_3 A + R_1 = R_1{}^* \tag{54}$$

by (40),

$$E_6 = -A^T N^T + M^T = (-NA + M)^T = 0, \tag{55}$$

$$E_7 = -NA + M = 0 \tag{56}$$

by (33). Thus, (49) can be written

$$C^T \begin{pmatrix} R & \phi^T \\ \phi & 0 \end{pmatrix} C = \begin{pmatrix} E_1 & E_4{}^T & I_m \\ E_4 & R_1{}^* & 0 \\ I_m & 0 & 0 \end{pmatrix}. \tag{57}$$

Taking determinants of both members of this equation and expanding the right-hand determinant by minors of the last m columns and the last m rows, we find that

$$|N|^{-2}\Delta = (-1)^m |R_1^*|. \tag{58}$$

It now follows that $|R_1^*|$ vanishes if, and only if, Δ vanishes, i.e., if the minimizing arc is singular.

But the transformed second variation is singular if the determinant

$$\begin{vmatrix} 0 & 0 & I_m \\ 0 & R_1^* & A^T \\ I_m & A & 0 \end{vmatrix} = (-1)^m |R_1^*| \tag{59}$$

vanishes and we have accordingly proved that the transformed J_2 is singular if, and only if, J_2 in its original form is singular.

4. Further transformation of second variation. We will now examine the effect upon J_2 as given by (6) and the constraining equations (4), (5), of a nonsingular linear transformation from variations η_i to functions ζ_i determined by the transformation equation

$$\eta = V\zeta. \tag{60}$$

In this equation, V represents an $n \times n$ matrix with elements V_{ij} which are functions of x and such that $|V|$ does not vanish for any x in the interval $[x_1, x_2]$.

Now,

$$\eta' = V\zeta' + V'\zeta \tag{61}$$

and hence, substituting from the last two equations into (28) and (29), we obtain

$$2\omega = \zeta'^T R_1 \zeta' + 2\zeta'^T Q_1 \zeta + \zeta^T P_1 \zeta, \tag{62}$$

where

$$R_1 = V^T R V, \tag{63}$$

$$Q_1 = V^T Q V + V^T R V', \tag{64}$$

$$P_1 = V^T P V + V'^T Q V + V^T Q^T V' + V'^T R V', \tag{65}$$

and

$$\phi_1 \zeta' + \theta_1 \zeta = 0, \tag{66}$$

where

$$\phi_1 = \phi V, \tag{67}$$

$$\theta_1 = \theta V + \phi V'. \tag{68}$$

R_1, P_1 are clearly symmetric matrices.

We shall further suppose that the quadratic form γ in (6) is transformed into a quadratic form γ_1 in the arguments ξ_1, ξ_2, ζ_{i1}, ζ_{i2} and that the end constraints (5) are transformed into

$$W_{\mu}\xi_1 + X_u\xi_2 + Y_{\mu i}\zeta_{i1} + Z_{\mu i}\zeta_{i2} = 0. \tag{69}$$

Suppose that 2ω has already been transformed into the form given in (44) and the differential constraints into the form of (38). Since $R_1{}^*$ is a singular symmetric matrix, a nonsingular $(n - m) \times (n - m)$ matrix U can be found such that

$$U^T R_1{}^* U = \begin{pmatrix} R_4 & 0 \\ 0 & 0 \end{pmatrix}, \tag{70}$$

where R_4 is an $r \times r$ diagonal matrix and r is the rank of $R_1{}^*$. Choosing

$$V = \begin{pmatrix} I_m & 0 \\ 0 & U \end{pmatrix}, \tag{71}$$

$|V| = |U| \neq 0$ and the transformation is regular, and it follows that 2ω is transformed from the form of (44) to that of (62) with

$$\begin{aligned} R_1 = V^T R^* V &= \begin{pmatrix} I_m & 0 \\ 0 & U^T \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & R_1{}^* \end{pmatrix} \begin{pmatrix} I_m & 0 \\ 0 & U \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 \\ 0 & U^T R_1{}^* U \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & R_4 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \end{aligned} \tag{72}$$

Further, this transformation replaces the constraining equation (37) by (66) with

$$\phi_1 = (I_m \,|\, A) \begin{pmatrix} I_m & 0 \\ 0 & U \end{pmatrix} = (I_m \,|\, AU). \tag{73}$$

We now further partition ϕ_1 thus:

$$\phi_1 = (I_m \mid F \mid G), \tag{74}$$

F being an $m \times r$ matrix and G an $m \times (n - m - r)$ matrix. Performing a further regular transformation having matrix V_1, where

$$V_1 = \begin{pmatrix} I_m & 0 & -G \\ 0 & I_r & 0 \\ 0 & 0 & I_q \end{pmatrix}, \tag{75}$$

where $m + r + q = n$, we calculate that R_1, ϕ_1 are transformed into R_2, ϕ_2, respectively, where

$$R_2 = \begin{pmatrix} I_m & 0 & 0 \\ 0 & I_r & 0 \\ -G^T & 0 & I_q \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & R_4 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} I_m & 0 & -G \\ 0 & I_r & 0 \\ 0 & 0 & I_q \end{pmatrix} \tag{76}$$

$$= \begin{pmatrix} 0 & 0 & 0 \\ 0 & R_4 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{77}$$

$$\phi_2 = (I_m \mid F \mid G) \begin{pmatrix} I_m & 0 & -G \\ 0 & I_r & 0 \\ 0 & 0 & I_q \end{pmatrix} = (I_m \mid F \mid 0). \tag{78}$$

It is now clear that we have succeeded in transforming 2ω into the form

$$2\omega = \zeta'^T R_2 \zeta' + 2\zeta'^T Q_2 \zeta + \zeta^T P_2 \zeta, \tag{79}$$

in which the derivatives ζ_i', $i > m + r$, only occur in the middle term, and the differential constraints into the form

$$\phi_2 \zeta' + \theta_2 \zeta = 0, \tag{80}$$

in which the derivatives ζ_i', $i > m + r$, do not occur at all. We have accordingly shown that the assumption that J_2 and its associated constraining equations are of this special form involves no loss of generality.

5. A necessary condition for minimization. Suppose that J_2 as given by (6) and the constraining equations (4), (5) is of the special reduced form described at the end of the previous section. Then, if the problem is singular, $m + r < n$. It follows that the derivatives η_{m+r+1}', η_{m+r+2}', $\cdots$, η_n' only occur in 2ω in the expression

$$2Q_{ij}\eta_i'\eta_j = 2\eta'^T Q \eta \tag{81}$$

and do not occur at all in the constraining equations.

Let $n - m > s \geqq r$ and $t = m + s + 1, m + s + 2, \cdots, n$. Suppose the matrix Q is partitioned thus:

$$Q = \begin{pmatrix} Q_2 & Q_4 \\ Q_3 & Q_1 \end{pmatrix}, \tag{82}$$

so that Q_1 is an $(n - m - s) \times (n - m - s)$ *symmetric* matrix, Q_2 is an $(m + s) \times (m + s)$ matrix, Q_3 is an $(n - m - s) \times (m + s)$ matrix, and Q_4 is an $(m + s) \times (n - m - s)$ matrix. This can always be done with Q_1 equal to the 1×1 symmetric matrix (Q_{nn}) and, in general, this is the only such partition. η is then partitioned as

$$\eta = \begin{pmatrix} h \\ k \end{pmatrix}, \tag{83}$$

where h is an $(m + s) \times 1$ matrix and k is an $(n - m - s) \times 1$ matrix. Therefore

$$2\eta'^T Q\eta = 2h'^T Q_2 h + 2h'^T Q_4 k + 2k'^T Q_3 h + 2k'^T Q_1 k. \tag{84}$$

Now, we have

$$\int_{x_1}^{x_2} k'^T Q_3 h \, dx = k^T Q_3 h \Big|_{x_1}^{x_2} - \int_{x_1}^{x_2} k^T \frac{d}{dx}(Q_3 h) k \, dx, \tag{85}$$

$$\int_{x_1}^{x_2} k'^T Q_1 k \, dx = \frac{1}{2} k^T Q_1 k \Big|_{x_1}^{x_2} - \frac{1}{2} \int_{x_1}^{x_2} k^T \frac{d}{dx}(Q_1) k \, dx, \tag{86}$$

as Q_1 is assumed to be symmetric. It is clear that the η_t' can now be completely eliminated from the integrand 2ω.

Assuming that this elimination has been effected, we make the final transformation

$$\begin{aligned} \eta_i &= \zeta_i, \qquad i = 1, 2, \cdots, m + s, \\ \eta_i &= \zeta_i', \qquad i = t = m + s + 1, m + s + 2, \cdots, n. \end{aligned} \tag{87}$$

Because of the absence of the derivatives η_t', after transformation J_2 and the constraining equations remain in the form of (4)–(6), except that the end values of the derivatives ζ_t', viz., ζ_{t1}', ζ_{t2}' occur in the end constraints (5) and in the quadratic form 2γ. However, any restrictions placed upon these two sets of end values do not reduce the class of sets of admissible functions ζ_i satisfying the differential constraints, for any such set can be made to satisfy such restrictions by infinitesimal adjustments over small neighborhoods of the endpoints and these adjustments will only affect the integral in J_2 infinitesimally. As a consequence, the quantities ζ_{t1}', ζ_{t2}' can be treated as parameters playing roles similar to those of ξ_1 and ξ_2.

Having performed the final transformation (87), the Clebsch condition is applied to the accessory minimizing problem and can be expected to yield a new necessary condition. We will illustrate this remark by applying the method to a rocket optimization problem in the next section.

6. Application to the intermediate-thrust arcs of rocket trajectories. We will now apply the results of §§4, 5 to the problem of deciding the status of the intermediate thrust arcs which arise in optimal rocket trajectory problems (see [3], [4]). The conclusions to which we shall be led are in agreement with those obtained by Kopp and Moyer [5], Robbins [6], and Kelley [7], [8] employing different methods. In this section we will use the following set of indices: $i, j, k = 1, 2, 3$; $r = 1, 2, \cdots, 10$.

The optimal rocket trajectory problem may be formulated thus: $Ox_1x_2x_3$ is an inertial frame. At time t a rocket has coordinates x_i and velocity components v_i : its motor thrust acts in a direction having direction cosines l_i and the mass rate of propellant consumption is m. Then if $g_i(x_1, x_2, x_3, t)$ are the components of the gravitational field in the frame, M is the rocket mass, and c the exhaust velocity, the equations of motion are

$$\dot{v}_i = (cml_i)/M + g_i, \tag{88}$$

$$\dot{x}_i = v_i, \tag{89}$$

$$\dot{M} = -m. \tag{90}$$

Employing the usual spherical polar angles θ, ϕ, the direction cosines can be expressed thus:

$$l_1 = \sin\theta\cos\phi, \qquad l_2 = \sin\theta\sin\phi, \qquad l_3 = \cos\theta. \tag{91}$$

The rocket is to be transferred from a given point in a gravitational field at which it has a specified velocity and mass to another given point at which it is to have another specified velocity. This leads to end conditions

$$v_i = v_{i0}, \qquad x_i = x_{i0}, \qquad M = M_0 \quad \text{at} \quad t = t_0, \tag{92}$$

$$v_i = v_{i1}, \qquad x_i = x_{i1}, \quad \text{at} \quad t = t_1. \tag{93}$$

The problem is to choose the functions $v_i(t)$, $x_i(t)$, $M(t)$, $\theta(t)$, $\phi(t)$, $m(t)$ subject to the constraints (88)–(90) and end conditions (92), (93) so that M_1 is maximized, i.e., the propellant expenditure is a minimum. This is equivalent to requiring that

$$J = -M_1 \tag{94}$$

should be minimized. The times of departure and arrival, t_0, t_1, may, or may not, be open to choice.

Treating θ, ϕ, m as derivatives of functions $\alpha(t)$, $\beta(t)$, $n(t)$, the above problem is set in the form of a Mayer problem as formulated by Bliss [1] and the argument of the previous sections is immediately applicable. Thus the equations

$$\theta = \dot{\alpha}, \qquad \phi = \dot{\beta}, \qquad m = \dot{n}, \tag{95}$$

are employed to eliminate θ, ϕ, m from the problem.

Defining a Lagrange function F by the equation

$$F = \lambda_i \left(\dot{v}_i - \frac{1}{M} cml_i - g_i \right) + \lambda_{i+3}(\dot{x}_i - v_i) + \lambda_7(\dot{M} + m), \tag{96}$$

it has been shown by Lawden [3], [4] that the multiplier rule and the necessary condition of Weierstrass can be satisfied by trajectories, called intermediate thrust arcs, along which

$$\lambda_i = pl_i , \tag{97}$$

where p is a constant. It may be verified that the determinant Δ (see (16)) for the problem vanishes identically, so that the trajectories under consideration are singular.

Denoting the variations of v_i , x_i , M, α, β, n by η_i , η_{i+3} , η_7 , η_8 , η_9 , η_{10} , respectively, the equations of variation prove to be

$$\dot{\eta}_i = \frac{1}{M} cm \left(\frac{\partial l_i}{\partial \theta} \dot{\eta}_8 + \frac{\partial l_i}{\partial \phi} \dot{\eta}_9 \right) + \frac{1}{M} c l_i \dot{\eta}_{10} + \frac{\partial g_i}{\partial x_j} \eta_{j+3} - \frac{1}{M^2} cml_i \eta_7 , \tag{98}$$

$$\dot{\eta}_{i+3} = \eta_i , \tag{99}$$

$$\dot{\eta}_7 = -\dot{\eta}_{10} . \tag{100}$$

Substituting from (96) into (8), it will be found from (7) that for this problem

$$\begin{aligned} 2\omega = {} & \frac{1}{M} pcm(\dot{\eta}_8^2 + \dot{\eta}_9^2 \sin^2 \theta) \\ & + \frac{2pc}{M^2} \dot{\eta}_{10} \eta_7 - \lambda_k \frac{\partial^2 g_k}{\partial x_i \partial x_j} \eta_{i+3} \eta_{j+3} - \frac{2mpc}{M^3} \eta_7^2 . \end{aligned} \tag{101}$$

By employing (100) this can be simplified thus:

$$\begin{aligned} \int_{t_0}^{t_1} \frac{2pc}{M^2} \dot{\eta}_{10} \eta_7 \, dt &= -\int_{t_0}^{t_1} \frac{pc}{M^2} \frac{d}{dt} (\eta_7^2) \, dt \\ &= -\left[\frac{pc}{M^2} \eta_7^2 \right]_{t_0}^{t_1} + \int_{t_0}^{t_1} \frac{2mpc}{M^3} \eta_7^2 \, dt, \end{aligned} \tag{102}$$

where use has been made of (90) after integration by parts. Absorbing the terms involving end values in the form 2γ, 2ω reduces to the form

$$2\omega = \frac{1}{M} pcm(\dot{\eta}_8^2 + \dot{\eta}_9^2 \sin^2 \theta) - \lambda_k \frac{\partial^2 g_k}{\partial x_i \partial x_j} \eta_{i+3} \eta_{j+3} . \tag{103}$$

It is clear that 2ω and the equations of variation are already in forms corresponding to (72) and (73). The next step is accordingly to perform the transformation whose matrix is (75) on the variations η_r . This is the transformation to variables ξ_r where

$$\eta_i = \xi_i + \frac{1}{M} cl_i \xi_{10}, \tag{104}$$

$$\eta_{i+3} = \xi_{i+3}, \tag{105}$$

$$\eta_7 = \xi_7 - \xi_{10}, \tag{106}$$

$$\eta_{7+i} = \xi_{7+i}. \tag{107}$$

In terms of the ξ's, the equations of variation now take the form

$$\dot{\xi}_i = \frac{1}{M} cm\left(\frac{\partial l_i}{\partial \theta}\dot{\xi}_8 + \frac{\partial l_i}{\partial \phi}\dot{\xi}_9\right) + \frac{\partial g_i}{\partial x_j}\xi_{j+3} - \frac{1}{M^2} cml_i \xi_7 - \frac{1}{M} c\dot{l}_i \xi_{10}, \tag{108}$$

$$\dot{\xi}_{i+3} = \xi_i + \frac{1}{M} cl_i \xi_{10}, \tag{109}$$

$$\dot{\xi}_7 = 0, \tag{110}$$

and

$$2\omega = \frac{1}{M} pcm(\dot{\xi}_8^2 + \dot{\xi}_9^2 \sin^2\theta) - \lambda_k \frac{\partial^2 g_k}{\partial x_i \partial x_j}\xi_{i+3}\xi_{j+3}. \tag{111}$$

$\dot{\xi}_{10}$ does not appear in the equations of variation or in 2ω. We can therefore replace ξ_{10} by $\dot{\xi}_{10}$ everywhere, as explained in §5. When this has been done, it will be found that the accessory problem remains singular and hence that the above process can be repeated.

We now transform to variations ζ_i, again employing a transformation matrix in the form (75). The equations of transformation prove to be

$$\xi_i = \zeta_i - \frac{1}{M} c\dot{l}_i \zeta_{10}, \tag{112}$$

$$\xi_{i+3} = \zeta_{i+3} + \frac{1}{M} cl_i \zeta_{10}, \tag{113}$$

$$\xi_7 = \zeta_7, \tag{114}$$

$$\xi_{i+7} = \zeta_{i+7}. \tag{115}$$

The equations of variation now take the form

$$\dot{\zeta}_i = \frac{1}{M} cm\left(\frac{\partial l_i}{\partial \theta}\dot{\zeta}_8 + \frac{\partial l_i}{\partial \phi}\dot{\zeta}_9\right) + \frac{\partial g_i}{\partial x_j}\left(\zeta_{j+3} + \frac{1}{M} cl_j \zeta_{10}\right) - \frac{1}{M^2} cml_i \zeta_7 + \zeta_{10}\frac{d}{dt}\left(\frac{1}{M} c\dot{l}_i\right), \tag{116}$$

$$\dot{\zeta}_{i+3} = \zeta_i - \left(\frac{2}{M} c\dot{l}_i + \frac{1}{M^2} mcl_i\right)\zeta_{10}, \tag{117}$$

$$\dot{\zeta}_7 = 0, \tag{118}$$

and

$$2\omega = \frac{1}{M} pcm(\dot{\zeta}_8^{\,2} + \dot{\zeta}_9^{\,2} \sin^2\theta) - \lambda_k \frac{\partial^2 g_k}{\partial x_i \partial x_j} \cdot \left(\zeta_{i+3} + \frac{1}{M} cl_i \zeta_{10}\right)\left(\zeta_{j+3} + \frac{1}{M} cl_j \zeta_{10}\right). \tag{119}$$

Since $\dot{\zeta}_{10}$ does not occur in either the equations of variation or 2ω, we can now replace ζ_{10} by $\dot{\zeta}_{10}$ everywhere. The resulting form of the accessory problem is no longer singular and the condition (14) subject to the constraints (15) takes the form

$$\frac{1}{M} pcm(\pi_8^{\,2} + \pi_9^{\,2} \sin^2\theta) - c^2\lambda_k \frac{\partial^2 g_k}{\partial x_i \partial x_j} l_i l_j \pi_{10}^2 \geqq 0 \tag{120}$$

subject to the constraints

$$\pi_i = \frac{1}{M} cm\left(\frac{\partial l_i}{\partial\theta}\pi_s + \frac{\partial l_i}{\partial\phi}\pi_9\right) + \left[\frac{1}{M} cl_j \frac{\partial g_i}{\partial x_j} + \frac{d}{dt}\left(\frac{1}{M} c\dot{l}_i\right)\right]\pi_{10}, \tag{121}$$

$$\pi_{i+3} = -\left(\frac{2}{M} c\dot{l}_i + \frac{1}{M^2} mcl_i\right)\pi_{10}, \tag{122}$$

$$\pi_7 = 0. \tag{123}$$

It is evident that the constraints leave π_8, π_9, π_{10} arbitrary and hence (120) implies that

$$\frac{1}{M} pcm \geqq 0, \qquad \frac{1}{M} pcm \sin^2\theta \geqq 0, \qquad \frac{c^2}{M^2}\lambda_k \frac{\partial^2 g_k}{\partial x_i \partial x_j} l_i l_j \leqq 0. \tag{124}$$

It follows that p must be positive and, by (97), that

$$\lambda_i \lambda_j \lambda_k \frac{\partial^2 g_k}{\partial x_i \partial x_j} \leqq 0. \tag{125}$$

In the special case when the gravitational field is an inverse square law of attraction towards the origin, we have

$$g_k = -\frac{\mu x_k}{r^3}, \tag{126}$$

where $r^2 = x_i x_i$, and it follows that

$$\frac{\partial^2 g_k}{\partial x_i \partial x_j} = \frac{3\mu}{r^5}(\delta_{ij} x_k + \delta_{jk} x_i + \delta_{ki} x_j) - \frac{15\mu}{r^7} x_i x_j x_k. \tag{127}$$

Hence,

$$\lambda_i\lambda_j\lambda_k \frac{\partial^2 g_k}{\partial x_i \partial x_j} = \frac{9\mu}{r^5}(\lambda_i\lambda_j\lambda_k x_j) - \frac{15\mu}{r^7}(\lambda_i x_i)^3. \tag{128}$$

But $\lambda_i\lambda_i = p^2$ and $\lambda_i x_i = pr\cos\psi$, where ψ is the angle between the radius vector from O to the vehicle and the direction of thrust. Condition (125) is accordingly equivalent to

$$\frac{3\mu p^3 s}{r^4}(3 - 5s^2) \leqq 0, \tag{129}$$

where $s = \cos\psi$ and hence,

$$\text{either} \quad s \geqq (\tfrac{3}{5})^{1/2} \quad \text{or} \quad 0 \geqq s \geqq -(\tfrac{3}{5})^{1/2}. \tag{130}$$

In Lawden [3], [4] it is shown that for the case of the two-dimensional IT-arcs where the transit time is not predetermined, s satisfies the condition

$$0 \leqq s \leqq (\tfrac{1}{3})^{1/2}. \tag{131}$$

The conditions (130) are accordingly violated and these IT-arcs cannot form part of an optimal trajectory.

Leitmann [9, p. 182] has proved all two-dimensional IT-arcs satisfy the condition

$$s^2 \leqq (\tfrac{1}{3})^{1/2}. \tag{132}$$

In exactly the same way we can prove all three-dimensional IT-arcs satisfy (132). It follows that IT-arcs can be optimal only when s is negative, i.e., the thrust possesses a component which is inwardly directed to the centre of attraction.

7. Notes. In the investigations [5], [6] special impulsive variations are used. These variations are generalizations of the impulsive variations used by Kelley [10] to deduce a generalized Clebsch condition which is, however, ineffective when applied to the IT-arcs of the rocket problem studied in §6.

In another approach Kelley [7] first transforms the state and control variables according to a rule fully discussed in Kelley [11]. The main advantage of this approach is that it provides a method of deriving the singular extremals. However, as singular extremals are transformed into singular extremals there is no guarantee that the classical Clebsch condition, applied in the new system of variables, will always impose a new and effective optimality condition, as it does in the case of the singular extremals of Lawden's problem in optimal rocket flight [8].

In this paper a procedure is put forward by which a generalized Clebsch necessary condition for singular extremals may be deduced. One advantage of this approach is that in some cases the final form of the accessory minimum problem is a nonsingular Bolza problem which would permit us to

define a generalized Jacobi's condition. Moreover this approach allows us to derive, without much more labor, the full generalized Clebsch condition for problems involving more than one control variable. Thus, although condition (125) could have been obtained by studying the accessory minimum problem in the subclass of variations $\eta_8 = \eta_9 \equiv 0$, we have derived the full generalized Clebsch condition because terms involving $\pi_8\pi_{10}$ and $\pi_9\pi_{10}$ could have occurred in condition (120).

Acknowledgment. The author wishes to express his indebtedness to Professor D. F. Lawden for his many valuable suggestions, including the original problem, and for his help in the preparation of this paper.

REFERENCES

[1] G. A. Bliss, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, 1946.

[2] A. Miele, *Extremization of linear integrals by Green's theorem*, Optimization Techniques, G. Leitmann, ed., Academic Press, New York, 1962, Chap. 3.

[3] D. F. Lawden, *Optimal intermediate-thrust arcs in a gravitational field*, Astronaut. Acta, 8(1962), pp. 106–123.

[4] ———, *Optimal Trajectories for Space Navigation*, Butterworths, Washington, 1963, Chaps. 3, 5.

[5] R. E. Kopp and H. G. Moyer, *Necessary conditions for singular extremals*, AIAA J., 3(1965), pp. 1439–1444.

[6] H. M. Robbins, *Optimality of intermediate-thrust arcs of rocket trajectories*, Ibid., 3(1965), pp. 1094–1098.

[7] H. J. Kelley, *Singular extremals in Lawden's problem of optimal rocket flight*, Ibid., 1(1963), pp. 1578–1580.

[8] H. J. Kelley et al., *Singular extremals in optimal control*, Optimization Techniques, vol. 2, G. Leitmann, ed., Academic Press, New York, to appear.

[9] G. Leitmann, *Variational problems with bounded control variables*, Optimization Techniques, G. Leitmann, ed., Academic Press, New York, 1962, Chap. 5.

[10] H. J. Kelley, *A second variation test for singular extremals*, AIAA J., 2(1964), pp. 1380–1382.

[11] ———, *A transformation approach to singular subarcs in optimal trajectory and control problems*, this Journal, 2(1965), pp. 234–240.

J. SIAM CONTROL
Vol. 4, No. 2, 1966
Printed in U.S.A.

SUFFICIENT CONDITIONS FOR OPTIMALITY AND THE JUSTIFICATION OF THE DYNAMIC PROGRAMMING METHOD*

V. G. BOLTYANSKII†

Abstract. The paper contains a detailed presentation of results which were published earlier in brief in [3]. The problem of the optimal control of a plant described by ordinary differential equations is considered. Sufficient optimality conditions are derived, one of which essentially gives a correct foundation to the dynamic programming method (for the class of problems being studied), while the other shows that under the condition of existence of regular synthesis the maximum principle is not only a necessary but also a sufficient optimality condition. Examples of the synthesis of nonlinear second-order systems are given.

1. Introduction. We shall study the motion of a controlled plant which is described by the system of differential equations

$$\frac{dx^i}{dt} = f^i(x^1, \cdots, x^n, u), \qquad i = 1, \cdots, n, \tag{1}$$

or, in vector form, by the equation

$$\frac{dx}{dt} = f(x, u). \tag{2}$$

The control parameter u which occurs on the right-hand side of system (1) can vary within the limits of a certain control region U. We pose the following optimal problem for (1) (see [1, p. 13]): In the phase space X of the variables $x^1, \cdots, x^n$, the two points x_0 and x_1 are given; *from among all the piecewise-continuous controls $u(t)$ which transfer the phase point moving in accordance with* (1) *from the position x_0 to the position x_1, find the one for which the functional*

$$J = \int_{t_0}^{t_1} f^0(x(t), u(t))\, dt$$

takes the smallest possible value. The control $u(t)$ which solves the stated problem and the trajectory corresponding to this control will, as usual, be called optimal.

Theorem 3 (or 4), proved below, is essentially the well-known dynamic programming principle of Bellman [2] as applied to the optimal control

* Originally published in Izv. Akad. Nauk SSSR Ser. Mat., 28 (1964), pp. 481–514. Submitted on January 8, 1963, for publication. This translation into English has been prepared by N. H. Choksy.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† V. A. Steklov Mathematical Institute, Academy of Sciences, U.S.S.R.

problem considered above. As is well-known, the dynamic programming principle, which has been completely substantiated in the case of difference equations does not as yet have a firm foundation for the case of differential equations. The reasoning which is usually used to substantiate this principle (see [1, pp. 69–72]) requires the continuous differentiability of the so-called Bellman function (see below), but this condition is not satisfied in even the simplest examples. Theorem 3 (or 4) gives an absolutely correct foundation of a somewhat refined dynamic programming principle (in the case of the optimal problem being studied). The dynamic programming principle appears in Theorems 3 and 4 as a sufficient (and not, as is usually done, as a necessary) optimality condition; moreover, it is proved to such a degree of generality that it allows us to include all the known examples and a number of new ones.

As is well-known, the maximum principle [1, p. 19] is the correctly-substantiated necessary optimality condition. We derive below the sufficient optimality condition in the form of a maximum principle (Theorem 5) in addition to the sufficient optimality condition in the form of the dynamic programming principle. In the former form the sufficient condition is convenient for practical application (apparently, in general, the dynamic programming principle has the advantage of greater generality but yields to the maximum principle in rigor of foundation and in convenience of use). The proof of Theorem 5 is based on Theorem 3 and allows us to establish the connection between the maximum principle and the dynamic programming principle.

The sufficient optimality conditions which are obtained play an important role for the following reason. The maximum principle allows us in a number of cases to pick out uniquely the trajectories which may be optimal. Are these trajectories actually optimal? To answer this question in the case where system (1) is linear we use the existence theorem for optimal controls [1, p. 127]: since the optimal control exists and since the maximum principle uniquely determines a trajectory which may be optimal, then it is the (unique) optimal trajectory joining the two given points. However, the existence theorem has been proved only for linear systems (1) and, moreover, only for the time-optimal case; fundamental difficulties are encountered when we attempt to prove it for nonlinear systems. Therefore, for even the simplest nonlinear systems there is no certainty that the trajectories found by synthesizing on the basis of the maximum principle are actually optimal. Theorem 5 indicates an escape from this situation since, as a rule, it allows us to assert that a synthesis effected on the basis of the maximum principle does indeed lead to optimal trajectories.

As examples we consider certain nonlinear equations of form (1). For these examples Theorem 5 is the only means of establishing optimality

since the existence theorem is inapplicable to the equations being considered.

The results of this article were presented earlier in [3] in an abbreviated form.

2. Fundamental lemmas. In what follows we shall assume that the functions $f^i(x, u)$, $i = 0, 1, \cdots, n$, are defined, continuous, and continuously differentiable with respect to $x^1, x^2, \cdots, x^n$, where the point $x = (x^1, x^2, \cdots, x^n)$ is located in some open set V of the space X. In other words, the functions

$$f^i(x, u), \qquad \frac{\partial f^i(x, u)}{\partial x^j}, \qquad i, j = 0, 1, \cdots, n,$$

are defined and are continuous on the direct product $V \times U$. The control region U can be, for example, a certain set in the r-dimensional vector space of the variable $u = (u^1, \cdots, u^r)$. We shall say that a piecewise-continuous control $u(t)$, specified on the interval $t_0 \leqq t \leqq t_1$, is admissible relative to the point $x_0 \in V$ if the solution of the equation

$$\frac{dx}{dt} = f(x, u(t)) \tag{3}$$

with initial condition $x(t_0) = x_0$ is defined and remains in the region V for $t_0 \leqq t \leqq t_1$. Further, we shall say that the control $u(t)$, $t_0 \leqq t \leqq t_1$, which is admissible relative to the point x_0, transfers the phase point from the position x_0 to the position x_1 if the solution of (3) with initial condition $x(t_0) = x_0$ satisfies the relation $x(t_1) = x_1$. The basic problem which we shall consider in this article is the following: *two points, x_0 and x_1, are given in the region V; from all the controls admissible relative to the point x_0, which transfer the phase point from the position x_0 to the position x_1, choose the one which effects the transfer from position x_0 to position x_1 with a minimal value of the functional J.*

The control $u(t)$ which solves this problem will be called *optimal in region V*. The corresponding trajectory $x(t)$ will also be called *optimal in region V*. Note that the control and the trajectory which are optimal in region V may cease to be optimal if region V increases. If $V = X$ we shall omit the phrase "in region V" and speak simply of optimal controls and trajectories.

We shall now introduce the concept of a piecewise-smooth set, which is important for what follows. Let K be some bounded, s-dimensional, convex polyhedron, $s \leqq n$, located in a vector space Ξ of the variables $\xi^1, \xi^2, \cdots, \xi^s$, and let us consider it together with its boundary (i.e., as a closed set). Let us assume that in a certain open set of space Ξ containing the

polyhedron K there are given n continuously-differentiable functions

$$\varphi^i(\xi^1, \xi^2, \cdots, \xi^s), \qquad i = 1, 2, \cdots, n, \tag{4}$$

possessing the property that the functional matrix

$$\left(\frac{\partial \varphi^i}{\partial \xi^j}\right), \qquad i = 1, \cdots, n, \quad j = 1, \cdots, s,$$

has the rank s at every point $\xi \in K$. The functions (4) effect a smooth mapping φ of the polyhedron K into the space X by the formulas:

$$x^i = \varphi^i(\xi^1, \cdots, \xi^s), \qquad i = 1, \cdots, n. \tag{5}$$

If the mapping is one-to-one (i.e., different points of the polyhedron K lead to different points of the space X), then the image $L = \varphi(K)$ of the polyhedron K will be called a curvilinear s-dimensional polyhedron in the space X. It is obvious that a curvilinear polyhedron is a closed and bounded (i.e., compact) set in the space X.

Any set $M \subset V$ that is the union of a finite or denumerable number of curvilinear polyhedra arranged in such a way that only a finite number of these polyhedra intersect every closed bounded set lying in V will be called a *piecewise-smooth set* in V. (The polyhdera may "cluster" at the boundary of the set V.) If among the curvilinear polyhedra whose union is a piecewise-smooth set M there is even one polyhedron of dimension k while all the remaining polyhedra have dimensions $\leqq k$, then the piecewise-smooth set M is said to be *k-dimensional*. In particular, any smooth surface of dimension less than n which is closed in V is a piecewise-smooth set in V because, as was proved in [4], it can be decomposed into curvilinear polyhedra. It is obvious that any set of dimension less than n which is piecewise-smooth in V does not contain interior points.

Let $L = \varphi(K)$ be a curvilinear polyhedron (see [5]) located in the region V, and let $x(t)$, $t_0 \leqq t \leqq t_1$, be a phase trajectory in V corresponding to an admissible control $u(t)$, $t_0 \leqq t \leqq t_1$, i.e., it is a solution of (3). We shall say that the phase trajectory $x(t)$ has a common position with the polyhedron L of dimension less than $n - 1$ if it does not intersect it; it has a common position with the polyhedron L of dimension $n - 1$ if the following conditions are satisfied:

1. The trajectory $x(t)$ does not intersect the boundary of the polyhedron L;

2. If $\tau_1, \tau_2, \cdots, \tau_{k-1}$ are all the points of discontinuity of the control $u(t)$, then none of the points

$$x(t_0), x(\tau_1), x(\tau_2), \cdots, x(\tau_{k-1}), x(t_1)$$

belong to the polyhedron L;

3. The trajectory $x(t)$ is not tangent to the polyhedron L at any of its points, i.e., if $x(t') \in L$, $t_0 < t' < t_1$, then the vector $f(x(t'), u(t'))$ does not lie in the tangent plane of the polyhedron L drawn at the point $x(t')$. In particular, this vector is not zero.

If the phase trajectory $x(t)$ has a common position with the polyhedron L (of dimension $n - 1$), then each of their common points is an isolated point on the trajectory $x(t)$ and, therefore, because the trajectory $x(t)$, $t_0 \leqq t \leqq t_1$, is compact, there exist only a finite number of points of intersection of the trajectory $x(t)$ with the polyhedron L.

LEMMA 1. *Let $u(t)$, $t_0 \leqq t \leqq t_1$, be an admissible control relative to the point $x_0 \in V$ and let L be a curvilinear polyhedron of dimension $\leqq n - 1$ located in V. Then, in any neighborhood W_0 of the point x_0 there exists an open set $G \subset W_0$ such that for any point $y_0 \in G$ the solution $y(t)$ of (3) with the initial condition $y(t_0) = y_0$ is defined on the whole interval $t_0 \leqq t \leqq t_1$ and has a common position with the polyhedron L.*

Proof. First of all we can assume, by reducing the neighborhood W_0 if necessary, that for any point $y_0 \in W_0$ the solution $y(t)$ of (3) with initial condition $y(t_0) = y_0$ is defined on the whole interval $t_0 \leqq t \leqq t_1$ (see [5, Theorem 16]). At first let the polyhedron L have dimension $\leqq (n - 2)$. In V we choose an open manifold N of the same dimension as L containing the polyhedron L. Such a manifold exists because the mapping (5) which gives the curvilinear polyhedron L is defined not only on the polyhedron K but also in some neighborhood of it.

The direct product $N \times [t_0, t_1]$ of the manifold N and the interval $[t_0, t_1]$ (see [6, p. 14]) is a manifold with an edge, having dimension $\leqq (n - 1)$. By P we denote the set of all points $(x', t') \in N \times [t_0, t_1]$ such that the solution $x(t; x', t')$ of (3) with the initial condition $x(t') = x'$ is defined on the interval $t_0 \leqq t \leqq t'$. The set P is an open subset, i.e., a submanifold, of the manifold $N \times [t_0, t_1]$. For every point $(x', t') \in P$ we consider the trajectory $x(t_0\, x', t')$, i.e., the solution of (3) with the initial condition $x(t') = x'$, and set

$$x(t_0\,; x', t') = \psi(x', t').$$

We obtain a mapping ψ (obviously continuous) of the manifold P into the region V.

Let us show that the image $\psi(P)$ of this mapping is a first-category set in V, i.e., the union of not more than a denumerable number of nowhere dense sets. To do this we set $\tau_0 = t_0$, $\tau_k = t_1$. Recall that $\tau_1, \tau_2, \cdots, \tau_{k-1}$ are all the points of discontinuity of the control $u(t)$. By P_i, $i = 1, 2, \cdots, k$, we denote the set of all points $(x', t') \in P$ for which $\tau_{i-1} \leqq t' \leqq \tau_i$. It is obvious that

$$P = P_1 \cup P_2 \cup \cdots \cup P_k,$$

and, therefore,

$$\psi(P) = \psi(P_1) \cup \psi(P_2) \cup \cdots \cup \psi(P_k).$$

Thus, it is sufficient to prove that each of the sets

$$\psi(P_1), \psi(P_2), \cdots, \psi(P_k)$$

is of first category in V. This proof is carried out in the same way for all the sets $\psi(P_1), \cdots, \psi(P_k)$. Let us carry it out for $\psi(P_k)$.

Since for $\tau_{k-1} \leqq t \leqq \tau_k$ the right-hand side of (3) depends continuously on $x^1, \cdots, x^n$, t and is continuously differentiable with respect to $x^1, \cdots, x^n$, the point $x(t; x', t')$ is continuously differentiable with respect to x', t' when $(x', t') \in P_k$, $\tau_{k-1} \leqq t \leqq t'$, by virtue of the theorem on the differentiability of solutions with respect to initial values (see [5, Theorem 18]). Therefore, the mapping $(x', t') \to x(\tau_{k-1}; x', t')$ is a smooth (of Class 1) mapping of the manifold P_k into V. By virtue of the same theorem on differentiability with respect to initial values, applied to (3) for $\tau_{k-2} \leqq t \leqq \tau_{k-1}$, the point

$$x(\tau_{k-2}; x', t') = x(\tau_{k-2}; x(\tau_{k-1}; x', t'), \tau_{k-1})$$

smoothly depends on $x(\tau_{k-1}; x', t')$, i.e., in view of what has already been proved, smoothly depends on $(x', t') \in P_k$. By next considering the intervals

$$\tau_{k-3} \leqq t \leqq \tau_{k-2}, \quad \cdots, \quad \tau_0 \leqq t \leqq \tau_1,$$

we finally get that the point $x(t_0; x', t') = \psi(x', t')$ smoothly depends on x', t' when $(x', t') \in P_k$. In other words, the mapping ψ, considered on P_k, is a smooth (of Class 1) mapping. Consequently, the set $\psi(P_k)$ is of first category in V (see [6, Theorem 1, p. 15]).

Thus, the set $\psi(P)$ is of first category in V. Therefore, in W_0 there exists a point $y_0 \notin \psi(P)$. Let us consider the solution $y(t)$ with the initial condition $y(t_0) = y_0$. By virtue of the choice of the neighborhood W_0 this solution is defined on the whole interval $t_0 \leqq t \leqq t_1$. Further, the solution $y(t)$ does not intersect the manifold N. Indeed, if a point t', $t_0 \leqq t' \leqq t_1$, were to exist such that $y(t') \in N$, then we would have $(y(t'), t') \in P$ because the solution $y(t)$ is defined on the interval $t_0 \leqq t \leqq t'$. By definition the solution $x(t; y(t'), t')$ satisfies the initial condition $x(t') = y(t')$ and therefore, by virtue of the uniqueness theorem, coincides with the solution $y(t)$. But then

$$y_0 = y(t_0) = x(t_0; y(t'), t') = \psi(y(t'), t') \in \psi(P),$$

which contradicts the choice of the point y_0.

Thus, the trajectory $y(t)$ with initial condition $y(t_0) = y_0$ is defined on

the whole interval $t_0 \leqq t \leqq t_1$ and does not intersect the manifold N and, consequently, not the polyhedron L. From the theorem on the continuous dependence of a solution on the initial values it follows, because the polyhedron L is compact, that there exists a neighborhood $W_0' \subset W_0$ of the point y_0 such that any solution $x(t)$ of (3), satisfying the condition $x(t_0) \in W_0'$, does not intersect L. This solution is defined on the whole interval $t_0 \leqq t \leqq t_1$ by virtue of the inclusion $W_0' \subset W_0$. The case where the polyhedron has dimension less than $n - 1$ has now been completely studied.

Now let the dimension of the polyhedron L equal $n - 1$. Then, the reasoning above is applicable to any face of the polyhedron L. Since the polyhedron L has only a finite number of faces, there exists an open set $W_0'' \subset W_0$ such that the solutions $x(t)$, satisfying the condition $x(t_0) \in W_0''$, are defined on the whole interval $t_0 \leqq t \leqq t_1$ and do not intersect the boundary of the polyhedron L. Thus, for any open set $G \subset W_0''$ condition 1 stated in the definition of a common position is satisfied.

Let us proceed to consider condition 2. In V we choose an open manifold N of dimension $n - 1$ containing the polyhedron L. By N_i, $i = 0, 1, \cdots, k$, we denote the set of points $x' \in N$ such that the solution $x_i(t; x')$ of (3) with the initial condition $x(\tau_i) = x'$ is defined on the interval $\tau_0 \leqq t \leqq \tau_i$. Then N_i is an open subset, i.e., submanifold, of the manifold N. For every point $x' \in N_i$ we consider the solution $x_i(t; x')$, i.e., the solution of (3) with initial condition $x(\tau_i) = x'$, and we set $\psi_i(x') = x_i(\tau_0 ; x')$. We obtain the mapping ψ_i (obviously continuous) of the manifold N_i into the region V.

As above, it can be established that ψ_i is a smooth (of Class 1) mapping of the manifold N_i into the region V. Since the manifold N (and hence also N_i) has dimension $n - 1$, the image $\psi_i(N_i)$ is a first-category set in V. Consequently,

$$\psi_0(N_0) \cup \psi_1(N_1) \cup \cdots \cup \psi_k(N_k)$$

is a first-category set in V. Therefore, there exists in W_0'' a point

$$y_0 \notin \psi_0(N_0) \cup \psi_1(N_1) \cup \cdots \cup \psi_k(N_k).$$

Let us consider the solution $y(t)$ with the initial condition $y(t_0) = y_0$. This solution is defined on the whole interval $t_0 \leqq t \leqq t_1$ and, as is easily seen, satisfies the condition $y(\tau_i) \notin N$, $i = 0, 1, \cdots, k$, and, consequently, also the condition $y(\tau_i) \notin L$, $i = 0, 1, \cdots, k$. From the theorem on the continuous dependence of a solution on the initial values follows the existence of a neighborhood $W_0''' \subset W_0''$ of the point y_0 such that for any solution $x(t)$ of (3) satisfying the condition $x(t_0) \in W_0'''$, the relations $x(\tau_i) \notin L$, $i = 0, 1, \cdots, k$, are satisfied. This solution is defined on the whole interval $t_0 \leqq t \leqq t_1$.

Thus, conditions 1 and 2 stated in the definition of a common position are satisfied for any open set $G \subset W_0'''$.

Finally, let us turn to the consideration of condition 3. The direct product $N \times [t_0, t_1]$ is a manifold with an edge of dimension n. By Q we denote the set of all points $(x', t') \in N \times [t_0, t_1]$ such that the solution $x(t; x', t')$ of (3) with initial condition $x(t') = x'$ is defined on the interval $t_0 \leqq t \leqq t'$. The set Q is a submanifold of the manifold $N \times [t_0, t_1]$. We subdivide the manifold Q into the parts $Q_1, Q_2, \cdots, Q_k$ by referring to Q_i as the set of all those points $(x', t') \in Q$ for which $\tau_{i-1} \leqq t' \leqq \tau_i$. Further, we define the mapping ψ of the manifold Q into the region V by setting

$$\psi(x', t') = x(t_0; x', t').$$

As before, the mapping ψ, considered on Q_i, is a smooth (of Class 1) mapping.

Let us assume that the trajectory $y(t)$ with initial condition $y(t_0) = y_0 \in W_0'''$ is, at the instant t', $t_0 \leqq t' \leqq t_1$, tangent to the manifold N at the point x'. The instant t' differs from $\tau_0, \tau_1, \cdots, \tau_k$ by virtue of condition 2 which is already satisfied when $y_0 \in W_0'''$. Then

$$x(t' + dt; x', t') = x' + dx,$$

where dx is some vector tangent to the manifold N and $dt \neq 0$. In other words,

$$x(t' + dt; x', t') = x(t' + dt; x' + dx, t' + dt),$$

whence, by virtue of the uniqueness theorem, it follows that

$$x(t; x', t') = x(t; x' + dx, t' + dt).$$

In particular, when $t = t_0$ we get

$$\psi(x', t') = \psi(x' + dx, t' + dt).$$

This signifies that a nonzero tangent vector (dx, dt) of the manifold $N \times [t_0, t_1]$ at the point (x', t') goes to zero when the mapping which is tangent to ψ, i.e., the tangent mapping at the point (x', t'), degenerates. In other words, the point (x', t') is not a regular point of the mapping ψ, and therefore the point $\psi(x', t') = y_0$ belongs to the image of a set of irregular points. (For definitions of regular and irregular points see [6, p. 10].)

Thus, if a trajectory starting from the point $y_0 \in W_0'''$ is tangent to the manifold N, then y_0 belongs to the image of a set of irregular points under the mapping ψ. But, according to [7], under a smooth (of Class 1) mapping of an n-dimensional manifold into an n-dimensional manifold, the image of a set of irregular points is of the first category[1]. Therefore, in W_0''' there

[1] Sard's theorem, proved by him in [7], was published considerably later by Dubovickii [8]; thus, for the Russian reader it is more convenient to find the formulation and the proof of this interesting theorem in [8]. Pontryagin [6] also cites this theorem (Dubovickii's Theorem, p. 25), but only a weakened estimate of the smoothness class is given, which is not useful to us.

exists a point y_0 not belonging to an image of irregular points. The trajectory $y(t)$ starting from this point y_0 is not tangent to the manifold N, i.e., is not tangent to the polyhedron L, and therefore has a common position with L. In particular, the trajectory $y(t)$ intersects the polyhedron L at only a finite number of points without being tangent to it. From this it is not difficult to deduce the existence of a neighborhood $G \subset W_0'''$ of the point y_0 such that when $x(t_0) \in G$ the trajectory $x(t)$, defined on the interval $t_0 \leqq t \leqq t_1$ and satisfying conditions 1 and 2, also intersects L at a finite number of points without being tangent to it. In other words, when $x(t_0) \in G$ the trajectory $x(t)$ satisfies all the conditions 1, 2, 3.

Thus, Lemma 1 is completely proved.

LEMMA 2. *Let $u(t)$, $t_0 \leqq t \leqq t_1$, be an admissible control relative to the point $x_0 \in V$ and let M be a set of dimension $\leqq n-1$ which is piecewise smooth in V. Then in any neighborhood W_0 of the point x_0 we can find a point y_0 such that the control $u(t)$, $t_0 \leqq t \leqq t_1$, is admissible relative to the point y_0, while the trajectory $y(t)$, $t_0 \leqq t \leqq t_1$, starting from the point y_0 and corresponding to the control $u(t)$, intersects M at only a finite number of points (i.e., there exist only a finite number of instants t, $t_0 \leqq t \leqq t_1$, for which $y(t) \in M$).*

Proof. Let $x(t)$, $t_0 \leqq t \leqq t_1$, denote the trajectory corresponding to the control $u(t)$ and starting from the point x_0. This trajectory lies wholly inside the region V and is compact. Therefore, there exists a neighborhood $W \subset V$ of this trajectory which intersects only a finite number of the curvilinear polyhedra comprising set M. Let us enumerate these polyhedra as $L_1, L_2, \cdots, L_\nu$.

We shall study (3), which is obtained by substituting in place of u in the right-hand side of (2) the control $u(t)$ mentioned in the lemma. Then $x(t)$ is the solution of (3) defined for $t_0 \leqq t \leqq t_1$ and satisfying the initial condition $x(t_0) = x_0$. By virtue of the theorem on the continuous dependence of a solution on the initial values, there exists a neighborhood $W_0' \subset W_0$ of the point x_0 such that any solution $y(t)$ of (3) for which $y(t_0) \in W_0'$ is defined on the whole interval $t_0 \leqq t \leqq t_1$ and is situated wholly in the region W.

By virtue of Lemma 1 there exists a neighborhood $W_0^{(1)} \subset W_0'$ such that any solution $y(t)$ for which $y(t_0) \in W_0^{(1)}$ has a common position with the polyhedron L_1. By virtue of Lemma 1 there exists an open set $W_0^{(2)} \subset W_0^{(1)}$ such that when $y(t_0) \in W_0^{(2)}$ the solution $y(t)$ has a common position with the polyhedron L_2. Continuing thus, we obtain the open sets

$$W_0^{(\nu)} \subset \cdots \subset W_0^{(2)} \subset W_0^{(1)} \subset W_0' \subset W_0 .$$

When $y(t_0) \in W_0^{(\nu)}$ the solution $y(t)$ has a common position with all the polyhedra $L_1, L_2, \cdots, L_\nu$. Moreover, it is situated wholly in W because

$y(t_0) \in W_0'$, and therefore does not intersect any of the other polyhedra. Consequently, when $y_0 = y(t_0) \in W_0^{(\nu)} \subset W_0$, the solution $y(t)$ intersects M at only a finite number of points.

Lemma 2 is proved.

3. Estimates of transient response performance.

LEMMA 3. *In the region V let there be given a piecewise-smooth set M of dimension $\leqq n - 1$ and the function $\omega(x) = \omega(x^1, \cdots, x^n)$, continuous in V, which in the set $V - M$ is continuously differentiable with respect to $x^1, \cdots, x^n$ and satisfies the condition*

$$\sum_{\alpha=1}^{n} \frac{\partial \omega(x)}{\partial x^\alpha} f^\alpha(x, u) \leqq f^0(x, u), \qquad x \in V - M, \quad u \in U. \tag{6}$$

Then if $u(t)$, $t_0 \leqq t \leqq t_1$, is an admissible control relative to the point x_0 which transfers the phase point from position x_0 to the position x_1, and if $x(t)$ is the corresponding trajectory, then

$$\int_{t_0}^{t_1} f^0(x(t), u(t))\, dt \geqq \omega(x_1) - \omega(x_0). \tag{7}$$

Proof. Let us select an arbitrary number $\epsilon > 0$ and let W_0 and W_1 be neighborhoods of the points x_0 and x_1, respectively, such that

$$|\omega(x) - \omega(x_0)| < \epsilon \quad \text{if} \quad x \in W_0,$$

$$|\omega(x) - \omega(x_1)| < \epsilon \quad \text{if} \quad x \in W_1.$$

The trajectory corresponding to the control $u(t)$ and starting from the point x_0 is denoted by $x(t)$. Then $x(t)$ is a solution of (3) defined for $t_0 \leqq t \leqq t_1$ and satisfying the conditions $x(t_0) = x_0$, $x(t_1) = x_1$. By virtue of the theorem on the continuous dependence of a solution of a system of differential equations on the initial conditions, there exists a neighborhood $W_0' \subset W_0$ of the point x_0 such that any solution $y(t)$ of (3) (with the same control $u(t)$, $t_0 \leqq t \leqq t_1$), for which $y(t_0) \in W_0'$, is defined on the whole interval $t_0 \leqq t \leqq t_1$ and satisfies the relations

$$y(t_1) \in W_1, \quad |f^0(x(t), u(t)) - f^0(y(t), u(t))| < \epsilon, \qquad t_0 \leqq t \leqq t_1.$$

By virtue of Lemma 2 there exists a solution $y(t)$, $t_0 \leqq t \leqq t_1$, of (3) which satisfies the condition $y(t_0) \in W_0'$ and intersects M at only a finite number of points. Let $\theta_1, \theta_2, \cdots, \theta_{m-1}$ be the instants at which the trajectory $y(t)$ intersects M and, moreover, let

$$\theta_1 < \theta_2 < \cdots < \theta_{m-1}.$$

Further, let us set

$$\theta_0 = t_0, \quad \theta_m = t_1.$$

When $\theta_{i-1} < t < \theta_i$, $i = 1, 2, \cdots, m$, the point $y(t)$ is located in the set $V - M$ and, therefore, the function ω is continuously differentiable at the point $y(t)$ and satisfies (6). Therefore, when $\theta_{i-1} < t < \theta_i$ (if t differs from the values $\tau_1, \cdots, \tau_{k-1}$ at which the control $u(t)$ is discontinuous and (2), therefore, is not satisfied) we have

$$\frac{d\omega(y(t))}{dt} = \sum_{\alpha=1}^{n} \frac{\partial\omega(y(t))}{\partial x^\alpha} \cdot \frac{dy^\alpha(t)}{dt}$$

$$= \sum_{\alpha=1}^{n} \frac{\partial\omega(y(t))}{\partial x^\alpha} f^\alpha(y(t), u(t)) \leqq f^0(y(t), u(t)).$$

Thus, everywhere on the interval $t_0 \leqq t \leqq t_1$ except at the points $\theta_0, \theta_1, \cdots, \theta_m$ and $\tau_1, \cdots, \tau_{k-1}$, i.e., everywhere except at a *finite* number of points, the function $\omega(y(t))$ has a continuous derivative and satisfies the relation

$$\frac{d\omega(y(t))}{dt} \leqq f^0(y(t), u(t)).$$

From this, because the function $\omega(y(t))$ is continuous, it follows that the inequality

$$\omega(y(t_1)) - \omega(y(t_0)) \leqq \int_{t_0}^{t_1} f^0(y(t), u(t))\, dt \tag{8}$$

holds

Further, since $y(t_0) \in W_0' \subset W_0$, then $y(t_1) \in W_1$. Consequently, by virtue of the choice of the neighborhoods W_0 and W_1 we have:

$$|\omega(y(t_0)) - \omega(x_0)| < \epsilon, \quad |\omega(y(t_1)) - \omega(x_1)| < \epsilon,$$

$$|f^0(x(t), u(t)) - f^0(y(t), u(t))| < \epsilon, \qquad t_0 \leqq t \leqq t_1.$$

In particular,

$$\omega(y(t_0)) - \omega(x_0) < \epsilon, \tag{9}$$

$$-\omega(y(t_1)) + \omega(x_1) < \epsilon, \tag{10}$$

and, further,

$$\int_{t_0}^{t_1} (f^0(y(t), u(t)) - f^0(x(t), u(t)))\, dt < \int_{t_0}^{t_1} \epsilon\, dt = \epsilon(t_1 - t_0),$$

i.e.,

$$\int_{t_0}^{t_1} f^0(y(t), u(t))\, dt < \int_{t_0}^{t_1} f^0(x(t), u(t))\, dt + \epsilon(t_1 - t_0). \tag{11}$$

Adding inequalities (8)–(11) we find

$$\omega(x_1) - \omega(x_0) < \int_{t_0}^{t_1} f^0(x(t), u(t))\, dt + 2\epsilon + \epsilon(t_1 - t_0).$$

In view of the arbitrariness of ϵ, from here we also obtain the required relation (7).

Lemma 3 is proved.

LEMMA 4.[2] *In the region V let there be given a closed set M of measure zero in V and a continuous function $\omega(x)$ which satisfies a Lipschitz condition locally in V and which on the set $V-M$ is continuously differentiable with respect to $x^1, \cdots, x^n$ and satisfies* (6). *Then for any admissible control $t_0 \leqq t \leqq t_1$, relative to the point x_0, which transfers the phase point from the position x_0 to the position x_1,* (7) *is satisfied.*

Proof. Let us select a neighborhood W_0 of the point x_0 possessing the property that the control $u(t)$, $t_0 \leqq t \leqq t_1$, is admissible relative to any point $y_0 \in W_0$. Such a neighborhood exists by virtue of the continuous dependence of a solution on the initial values. From the neighborhood W_0 and a number $\epsilon > 0$ let us choose the neighborhood W_0' in the same way as in the proof of Lemma 3. We denote by $x(t; y_0)$ a solution of (3) satisfying the initial condition $x(t_0) = y_0 \in W_0'$. The formula

$$\psi(y_0, t) = (x(t; y_0), t)$$

defines a continuous mapping of the direct product $W_0' \times [t_0, t_1]$ into the direct product $V \times [t_0, t_1]$. We choose the points $\tau_0, \tau_1, \cdots, \tau_k$ as before. Then, on the product

$$W_0' \times (\tau_{i-1}, \tau_i), \quad i = 1, 2, \cdots, k,$$

the mapping ψ is a smooth (of Class 1) mapping by virtue of the theorem on the differentiability of a solution with respect to the initial conditions. Furthermore, this mapping is regular at every point of the product $W_0' \times (\tau_{i-1}, \tau_i)$, i.e., has a nonzero Jacobian. Indeed, by using the variational equations it is easy to show that the vectors

$$\frac{\partial \psi(y_0, t)}{\partial y_0{}^i}, \quad i = 1, \cdots, n,$$

where $y_0{}^1, \cdots, y_0{}^n$ are the coordinates of the point y_0, are linearly independent at any instant t. These vectors lie in the "layer" $t = \text{const.}$ of the direct product $V \times [t_0, t_1]$. However, the vector $\partial\psi(y_0, t)/\partial t$, defined for $t \neq \tau_0, \tau_1, \cdots, \tau_k$, is different from zero and does not lie in this layer. Thus, the vectors

[2] This lemma and the proofs of Theorems 2 and 4 which are based on it form an entity in themselves and are not connected with the rest of the paper. Therefore, Lemma 4 and Theorems 2 and 4 can be omitted without affecting the understanding of the remaining part of the paper.

$$\frac{\partial\psi(y_0, t)}{\partial y_0^1}, \cdots, \frac{\partial\psi(y_0, t)}{\partial y_0^n}, \frac{\partial\psi(y_0, t)}{\partial t}$$

are linearly independent and, therefore, the mapping ψ is regular. Moreover, the mapping ψ is homeomorphic, by virtue of the uniqueness theorem.

Since M has measure zero in V, the set $M \times [t_0, t_1]$ has measure zero in the manifold $V \times [t_0, t_1]$. From this it easily follows that the set $\psi^{-1}(M \times [t_0, t_1])$ has measure zero in the manifold $W_0' \times [t_0, t_1]$. Indeed, the part of the set $\psi^{-1}(M \times [t_0, t_1])$ which is located in $W_0' \times (\tau_{i-1}, \tau_i)$ has measure zero because ψ is a smooth homeomorphism. Further, the part of the set $\psi^{-1}(M \times [t_0, t_1])$ which is located in $W_0' \times \tau_i$ has measure zero since the set $W_0' \times \tau_i$ itself has measure zero in $W_0' \times [t_0, t_1]$.

Thus, $\psi^{-1}(M \times [t_0, t_1])$ is a set of measure zero in $W_0' \times [t_0, t_1]$. From this, according to the general theorems of measure theory (for example, see [9, pp. 367–371]), it follows that for almost all points $y_0 \in W_0'$ the segment $I(y_0)$, consisting of points of the form (y_0, t), $t \in [t_0, t_1]$, intersects $\psi^{-1}(M \times [t_0, t_1])$ in a set of measure zero (in the sense of measure on this interval). Let us select such a point y_0 and consider the solution $y(t)$ of (3) with the initial condition $y(t_0) = y_0$.

By virtue of the definition of mapping ψ we have

$$(y(t), t) = \psi(y_0, t), \qquad t_0 \leqq t \leqq t_1.$$

It is easy to understand that the inclusions

$$(y_0, t) \in \psi^{-1}(M \times [t_0, t_1]), \qquad (y(t), t) \in M \times [t_0, t_1], \quad \text{and} \quad y(t) \in M$$

are equivalent. Therefore, because of the way the point y_0 was chosen, the set of those points t for which $y(t) \in M$ has measure zero on the interval $t_0 \leqq t \leqq t_1$. Let us denote this set by R. Then the set $[t_0, t_1] - R$ is of complete measure on the interval $[t_0, t_1]$. When $t \in [t_0, t_1] - R$, the point $y(t)$ lies in $V - M$ and, therefore, the function $\omega(y(t))$ has at this point a derivative which satisfies the condition

$$\frac{d\omega(y(t))}{dt} \leqq f^0(y(t), u(t)). \tag{12}$$

(Compare the proof of Lemma 3.) Thus, almost everywhere on the interval $[t_0, t_1]$ the function $\omega(y(t))$ has a derivative with respect to t and satisfies (12).

Furthermore, the function $\omega(y(t))$ is absolutely continuous. Indeed, since the trajectory $y(t)$, $t_0 \leqq t \leqq t_1$, is a compact set, the function $|f(y(t), u(t))|$ (see (2)) is bounded on this trajectory. Let K_1 be the upper bound of this function. Then for any two values t', t'' lying in the interval $[t_0, t_1]$ we have

$$\rho(y(t'), u(t'')) \leq K_1 |t' - t''|,$$

where ρ denotes distance in the space X. Since, further, the function $\omega(x)$ locally satisfies a Lipschitz condition, then for any $t \in [t_0, t_1]$ there exists a constant K_2 such that

$$|\omega(x') - \omega(x'')| \leqq K_2\rho(x', x''),$$

provided only that x' and x'' are contained in a sufficiently small neighborhood of the point $y(t)$. Thus, if t' and t'' are sufficiently close to t, then

$$|\omega(y(t')) - \omega(y(t''))| \leqq K_2\rho(y(t'), y(t'')) \leqq K_2K_1 \cdot |t' - t''|.$$

In other words, the function $\omega(y(t))$ locally satisfies a Lipschitz condition and, therefore, is absolutely continuous.

Thus, the function $\omega(y(t))$ is absolutely continuous and satisfies (12) almost everywhere on the interval $[t_0, t_1]$. From this it follows that (8) is satisfied for the function $\omega(y(t))$. Moreover, (9)–(11) are satisfied by virtue of the inclusion $y_0 \in W_0'$ and the choice of the neighborhood W_0'. The validity of Lemma 4 ensues from (8)–(11).

4. Sufficient optimality conditions for an individual trajectory.

THEOREM 1. *Let $u^*(t)$, $t_0^* \leqq t \leqq t_1^*$, be an admissible control relative to the point x_0 transferring the phase point from the position x_0 to the position x_1, and let $x^*(t)$ be the corresponding trajectory. For the optimality (in V) of the control $u^*(t)$ and the trajectory $x^*(t)$ it is sufficient that there exist a piecewise-smooth set $M \subset V$ of dimension $\leqq n - 1$ and a function $\omega(x^1, \cdots, x^n)$ which is continuous in V, which is continuously differentiable with respect to $x^1, \cdots, x^n$ on the set $V - M$, and which satisfies the conditions*:

$$\sum_{\alpha=1}^{n} \frac{\partial\omega(x)}{\partial x^\alpha} f^0(x, u) \leqq f_0(x, u), \quad x \in V - M, u \in U, \tag{13}$$

$$\int_{t_0^*}^{t_1^*} f^0(x^*(t), u^*(t))\, dt = \omega(x_1) - \omega(x_0). \tag{14}$$

Proof. By virtue of Lemma 3, (7) is satisfied for any control $u(t)$, $t_0 \leqq t \leqq t_1$, that transfers the phase point from the position x_0 to the position x_1; whence, according to (14) it follows also that the control $u^*(t)$ and the trajectory $x^*(t)$, $t_0 \leqq t \leqq t_1$, are optimal.

The following theorem follows analogously from Lemma 4.

THEOREM 2. *Let $u^*(t)$, $t_0^* \leqq t \leqq t_1^*$, be an admissible control relative to the point x_0 transferring the phase point from the position x_0 to the position x_1, and let $x^*(t)$ be the corresponding trajectory. For the optimality (in V) of the control $u^*(t)$ and the trajectory $x^*(t)$ it is sufficient that there exist a closed set $M \subset V$ and a function $\omega(x) = \omega(x^1, \cdots, x^n)$ which is continuous in V, locally satisfies a Lipschitz condition in V, is continuously differentiable with respect to $x^1, \cdots, x^n$ on the set $V - M$ and satisfies* (13) *and* (14).

Note that for the time-optimal case, i.e., the case where $f^0(x, u) \equiv 1$ and the functional $J = t_1 - t_0$ is the transfer time from the point x_0 to the point x_1, (13) and (14) take the following forms:

$$\sum_{\alpha=1}^{n} \frac{\partial \omega(x)}{\partial x^\alpha} f^\alpha(x, u) \leqq 1, \quad x \in V - M, \quad u \in U, \tag{13'}$$

$$t_1^* - t_0^* = \omega(x_1) - \omega(x_0). \tag{14'}$$

5. The dynamic programming principle as a sufficient optimality condition. The continuous function

$$\omega(x) = \omega(x^1, \cdots, x^n),$$

specified in the region V, will be called a Bellman function relative to a point $a \in V$ if it possesses the following properties:

1. $\omega(a) = 0$;

2. there exists a set M (the singular set of Bellman functions $\omega(x)$) which is closed in V and does not contain interior points such that the function $\omega(x)$ is continuously differentiable with respect to $x^1, \cdots, x^n$ on the set $V - M$ and satisfies the condition

$$\sup_{u \in U} \left(\sum_{\alpha=1}^{n} \frac{\partial \omega(x)}{\partial x^\alpha} f^\alpha(x, u) - f^0(x, u) \right) = 0, \quad x \in V - M. \tag{15}$$

Relation (15) is called the Bellman equation.

Note that if $f^0(x, u) \equiv 1$ the functional J takes the value $J = t_1 - t_0$ (the time-optimal problem); in this case the Bellman equation becomes

$$\sup_{u \in U} \sum_{\alpha=1}^{n} \frac{\partial \omega(x)}{\partial x^\alpha} f^\alpha(x, u) = 1, \qquad x \in V - M.$$

It is understood that all the subsequent results remain valid for this special case also.

THEOREM 3. *Let us assume that for* (2) *given in a region* $V \subset X$ *there exists in* V *a Bellman function* $\omega(x)$ *relative to the point* $a \in V$ *with a piecewise-smooth singular set. Let us further assume that for any point* $x_0 \in V$ *there exists a control* $u(t) = u_{x_0}(t)$ *which is admissible relative to the point* x_0 *and which transfers the phase point from the position* x_0 *to the position* a *and satisfies the relation*

$$\int_{t_0}^{t_1} f^0(x(t), u(t))\, dt = -\omega(x_0). \tag{16}$$

Then, all the indicated controls $u_{x_0}(t)$ *are optimal in* V.

Proof. The result follows directly from Theorem 1 if we note that (15) by itself implies the fulfillment of (13) and that (16) coincides with (14) since $\omega(x_1) = \omega(a) = 0$.

The following theorem follows analogously from Theorem 2.

THEOREM 4. *Let us assume that for* (2) *given in a region* $V \subset X$ *there exists in* V *a Bellman function* $\omega(x)$ *relative to the point* $a \in V$ *with a closed singular set* M *of measure zero, which locally* (*close to every point* $x \in V$) *satisfies a Lipschitz condition. Let us further assume that for any point* $x_0 \in V$ *there exists a control* $u(t) = u_{x_0}(t)$ *which is admissible relative to the point* x_0 *and which transfers the phase point from the position* x_0 *to the position* a *and satisfies* (16). *Then, all the indicated controls* $u_{x_0}(t)$ *are optimal in* V.

6. Sufficient optimality condition in the form of the maximum principle. We shall first introduce the notion of regular synthesis for (2), wherein we shall now assume the continuity of the derivatives $\partial f^i/\partial x^j$, $\partial f^i/\partial u^k$ and the validity of the relation $f^0(x, u) > 0$. Let us assume that we are given a piecewise-smooth set N of dimension $\leqq n-1$, the piecewise-smooth sets

$$P^0 \subset P^1 \subset P^2 \subset \cdots \subset P^{n-1} \subset P^n = V, \tag{17}$$

and the function $v(x)$ defined in V and taking values in U. We shall say that the sets (17) and the function $v(x)$ effect a regular synthesis for (2) in the region V if the following conditions are satisfied:

A. The set P^0 consists of only the one point a. Every component of the set $P^i - (P^{i-1} \cup N)$, $i = 1, 2, \cdots, n$, is an i-dimensional smooth manifold in V; we shall call these components i-dimensional cells. The function $v(x)$ is continuous and continuously differentiable on each cell and can be extended into a continuously differentiable function on the neighborhood of the cell.

B. All the cells are grouped into cells of the first and second kinds. All n-dimensional cells are cells of the first kind.

C. If σ is a certain i-dimensional cell of the first kind, $i > 1$, then through every point of this cell there passes a unique trajectory of the equation

$$\frac{dx}{dt} = f(x, v(x)) \tag{18}$$

(passing with respect to the cell σ). There exists an $(i-1)$-dimensional cell $\Pi(\sigma)$ such that every trajectory of (18) which moving around in cell σ leaves the cell σ after a finite time during which it strikes against the cell $\Pi(\sigma)$ at a nonzero angle and approaches it with a nonzero phase velocity. If σ is a one-dimensional cell of the first kind, then it is a segment of the phase trajectory of (18) approaching the point a with nonzero phase velocity.

D. If σ is a certain i-dimensional cell of the second kind, $i \geqq 1$, then there exists an $(i+1)$-dimensional cell $\Sigma(\sigma)$, a cell of the first kind, such that from any point of the cell σ there issues a unique trajectory of (18) moving

around in the cell $\Sigma(\sigma)$; moreover, the function $v(x)$ is continuous and continuously differentiable on $\sigma \cup \Sigma(\sigma)$.

E. The conditions enumerated above ensure the possibility of extending the trajectories of (18) from cell to cell: from the cell σ into the cell $\Pi(\sigma)$ if $\Pi(\sigma)$ is of the first kind, and from the cell σ into the cell $\Sigma(\Pi(\sigma))$ if the cell $\Pi(\sigma)$ is of the second kind. It is required that every such trajectory go through only a finite number of cells (i.e., each trajectory "pierces" cells of the second kind only a finite number of times). In this connection any trajectory terminates at the point a. We shall refer to the indicated trajectories as being marked. Thus, from every point of the set $V - N$ there issues a unique marked trajectory that leads to the point a. It is also required that from every point of the set N there issues a trajectory of (18) leading to the point a, which is not necessarily unique and which is also said to be marked.

F. All the marked trajectories satisfy the maximum principle.

G. The value of the functional J computed along the marked trajectories that terminate at the point a is a continuous function of the initial point x_0. In particular, if several marked trajectories start from a point $x_0 \in N$, then the value of the functional J is the same for each.

All the known examples of linear time-optimal synthesis are special cases of regular synthesis.

THEOREM 5. *If a regular synthesis for* (2) *is effected in the set V under the assumptions that there exist the continuous derivatives $\partial f^i/\partial x^j$ and $\partial f^i/\partial u^k$ and that $f^0(x, u) > 0$, then all the marked trajectories are optimal (in region V). In this sense the maximum principle is a sufficient optimality condition.*

Proof. We shall first prove Theorem 5 under the assumption that $f^0(x, u) \equiv 1$. In this case the functional J is the time it takes the phase point to move from position x_0 to the point a. Let $\omega(x)$ denote the value of the functional J (i.e., the transfer time) for going from a point x to the point a by means of a marked trajectory. The set $P^{n-1} \cup N$ is denoted by M. Let us prove that $\omega(x)$ is a Bellman function with M as its singular set. Then, Theorem 5 will follow immediately from Theorem 3. Thus, it is necessary to prove only that the function $\omega(x)$ is differentiable on the set $V - M$ and that it satisfies (15).

Let x be an arbitrary point belonging to a certain n-dimensional cell σ. Let us select an arbitrary number t_0 and let $t_0 + \theta_1(x)$ be the instant at which a trajectory of (18), which starts from the point x at the instant t_0, hits the cell $\Pi(\sigma)$, i.e., $\theta_1(x)$ is the time of motion from the point x to the cell $\Pi(\sigma)$. The point at which this trajectory "lands" on the set $\Pi(\sigma)$ will be denoted by $\xi_1(x)$. From the general theorems on the differentiability of solutions with respect to parameters it follows that $\xi_1(x)$ and $\theta_1(x)$ are continuously differentiable functions of x. Indeed, let x_0 be an arbitrary

point of the n-dimensional cell we are considering. Let us reverse the direction of the time flow, i.e., we shall consider the system

$$\frac{dy}{dt} = -f(y, v(y)) \tag{19}$$

on the cell σ. The trajectories of this system (in the cell σ) coincide with the trajectories of (18) but run in the opposite direction. For any point $\xi \in \Pi(\sigma)$ close to $\xi_1(x_0)$ we let $y(t, \xi)$ denote the solution of (19) with the initial condition $y(0, \xi) = \xi$. Then, the function $y(t, \xi)$ is continuously differentiable with respect to the set of variables t, ξ, for $t > 0$ and $\xi \in \Pi(\sigma)$. Obviously we have

$$y(\theta_1(x_0), \xi_1(x_0)) = x_0 . \tag{20}$$

It is not difficult to see that the functional determinant[3]

$$\left.\frac{D(y(t, \xi))}{D(t, \xi)}\right|_{t=\theta_1(x_0), \xi=\xi_1(x_0)}$$

is different from zero. Indeed, when $t = 0$ and $\xi = \xi_1(x_0)$, this functional determinant differs from zero since by virtue of condition C the trajectory $x(t)$ of (18), starting from the point x_0, approaches the cell $\Pi(\sigma)$ at a non-zero angle. Consequently, this functional determinant differs from zero also when $t = \theta_1(x_0)$ and $\xi = \xi_1(x_0)$ since the system of variational equations is linear (see [5, p. 198, Theorem 18]).

Therefore, when x is close to x_0 the equation $y(t, \xi) = x$ can be solved uniquely (see (20)):

$$\xi = \xi_1(x), \quad t = \theta_1(x).$$

Moreover, the functions $\xi_1(x)$ and $\theta_1(x)$ are continuously differentiable with respect to x.

Further, from the point $\xi_1(x)$ the trajectory is extended into the cell $\Pi(\sigma)$ or $\Sigma(\Pi(\sigma))$. It can be established analogously that the point $\xi_2(x)$ at which this trajectory leaves the cell $\Pi(\sigma)$ (or $\Sigma(\Pi(\sigma))$) and the time $\theta_2(x)$ of motion within this cell depend differentiably on $\xi_1(x)$, and hence also on x. Continuing in this manner we find that the total time

$$-\omega(x) = \theta_1(x) + \theta_2(x) + \cdots$$

of motion along a marked trajectory from the point x to the point a is (inside the cell σ) a continuously differentiable function of the point x.

Thus, the function $\omega(x)$ is continuously differentiable on $V - M$.

It remains to establish that the function $\omega(x)$ satisfies the Bellman equa-

[3] The point $y(t, \xi)$ has n coordinates (in the cell σ); the point ξ has $n - 1$ coordinates (in the cell $\Pi(\sigma)$).

tion (15) on $V - M$. Let $x_0 \in V - M$; $x(t)$ denotes the marked trajectory starting from the point x_0 at the instant t_0, and t_1 denotes the instant at which it hits the point a. Let us consider the set S consisting of all the points which satisfy the condition

$$\omega(x) = \omega(x_0).$$

Then, close to the point x_0 the set S is a smooth hypersurface in V with a normal vector

$$\operatorname{grad} \omega(x_0) = \left(\frac{\partial\omega(x_0)}{\partial x^1}, \cdots, \frac{\partial\omega(x_0)}{\partial x^n}\right).$$

This vector differs from zero by virtue of the relation

$$\sum_{\alpha=1}^{n} \frac{\partial\omega(x_0)}{\partial x^\alpha} \cdot f^\alpha(x_0, v(x_0)) = \frac{d\omega(x(t))}{dt}\bigg|_{x(t)=x_0} = 1. \tag{21}$$

According to condition F the trajectory $x(t)$ satisfies the maximum principle. Let $\psi(t) = (\psi_1(t), \cdots, \psi_n(t))$ denote the covariant vector-function corresponding to the trajectory $x(t)$ as called for in the maximum principle (see [1, p. 18]). We now show that the vector $\psi(t_0)$ is orthogonal to the surface S at the point x_0, i.e.,

$$\psi(t_0) = \lambda \operatorname{grad} \omega(x_0),$$

or, alternatively,

$$\psi_\alpha(t_0) = \lambda \cdot \frac{\partial\omega(x_0)}{\partial x^\alpha}, \qquad \alpha = 1, 2, \cdots, n. \tag{22}$$

Let us suppose that (22) is established. Then by virtue of the maximum principle we have

$$H = \sum_{\alpha=1}^{n} \psi_\alpha(t_0) f^\alpha(x_0, v(x_0)) = \lambda \sum_{\alpha=1}^{n} \frac{\partial\omega(x_0)}{\partial x^\alpha} f^\alpha(x_0, v(x_0)) = \lambda$$

(see (21)). From the relation $H \geqq 0$ occurring in the maximum principle we conclude that $\lambda \geqq 0$. Moreover, $\lambda \neq 0$ because otherwise $\psi(t_0) = 0$ (see (22)). Thus we have $\lambda > 0$. Further, from the maximum principle we find that

$$H(\psi(t_0), x_0, v(x_0)) \geqq H(\psi(t_0), x_0, u) \quad \text{for every} \quad u \in U,$$

whence, by virtue of (21), (22), and the relation $\lambda > 0$, we obtain

$$\begin{aligned} 1 &= \sum_{\alpha=1}^{n} \frac{\partial\omega(x_0)}{\partial x^\alpha} f^\alpha(x_0, v(x_0)) = \frac{1}{\lambda} \sum_{\alpha=1}^{n} \psi_\alpha(t_0) f^\alpha(x_0, v(x_0)) \\ &= \frac{1}{\lambda} H(\psi(t_0), x(t_0), v(x_0)) \geqq \frac{1}{\lambda} H(\psi(t_0), x(t_0), u) \\ &= \frac{1}{\lambda} \sum_{\alpha=1}^{n} \psi_\alpha(t_0) f^\alpha(x_0, u) = \sum_{\alpha=1}^{n} \frac{\partial\omega(x_0)}{\partial x^\alpha} f^\alpha(x_0, u) \end{aligned}$$

for every $u \in U$. Thus, (15) is fulfilled in $V - M$ because $f^0 \equiv 1$.

It remains to establish the validity of (22). Let $\sigma_1, \sigma_2, \cdots, \sigma_q$ be cells of the first kind through which the trajectory $x(t)$ passes in succession such that $x_0 \in \sigma_1$ and the cell σ_q is one-dimensional and adjoins point a. Let us set $t_0 = \tau_0$, $t_1 = \tau_q$ and let $\tau_1, \cdots, \tau_{q-1}$ denote the "switching instants" (i.e., the instants of transition from cell to cell), so that on the interval $\tau_{i-1} < t < \tau_i$ the trajectory $x(t)$ moves within the cell σ_i, $i = 1, 2, \cdots, q$. For every two adjacent cells σ_i and σ_{i+1} in the sequence $\sigma_1, \sigma_2, \cdots, \sigma_q$ one of the two following cases is possible (see conditions C and D in the definition of regular synthesis):

(a) both the cells σ_i and σ_{i+1} have the same dimension k, and then $\sigma_{i+1} = \Sigma(\Pi(\sigma_i))$. In this case the trajectory $x(t)$ pierces the cell $\Pi(\sigma_i)$ at the switching instant τ_i, this cell being a $(k-1)$-dimensional cell of the second kind;

(b) the cell σ_i has dimension k while the cell σ_{i+1} has dimension $k-1$ and coincides with the cell $\Pi(\sigma_i)$.

In both cases the "switching point" $x(\tau_i)$ is an interior point of the cell $\Pi(\sigma_i)$ and, moreover, from any point of the cell $\Pi(\sigma_i)$ a unique trajectory of (18) starts which moves around in the cell σ_{i+1}. Therefore, the trajectory $x^*(t)$ of (18), which at the instant t_0 starts from any interior point x_0^* of the cell σ_1, will pass through the same sequence of cells $\sigma_1, \sigma_2, \cdots, \sigma_q$ and will arrive at the point a. We shall assume that the point x^* lies on the hypersurface S sufficiently close to x_0 so that the time of motion along the trajectory $x^*(t)$ from the point x_0^* to the point a coincides with the time of motion along the trajectory $x(t)$. In other words, both the trajectories $x(t)$ and $x^*(t)$, which at the instant t_0 start from the points x_0 and x_0^*, arrive at the point a at one and the same instant $t_1 = -\omega(x_0) = -\omega(x_0^*)$.

As we have already seen above, the switching instants

$$\tau_0^* = t_0, \tau_1^*, \tau_2^*, \cdots, \tau_q^* = t_1$$

for the trajectory $x^*(t)$, and the corresponding switching points

$$x^*(\tau_1^*), x^*(\tau_2^*), \cdots, x^*(\tau_{q-1}^*),$$

which are interior points of the cells $\Pi(\sigma_1), \Pi(\sigma_2), \cdots, \Pi(\sigma_{q-1})$, are differentiably dependent on the point $x_0^* \in \sigma_1$.

If the point x_0^* is sufficiently close to the point x_0, the inequalities $\tau_{i-1} < \tau_i^*$ and $\tau_{i-1}^* < \tau_i$ are satisfied since $\tau_{i-1} < \tau_i$. Let δ_i denote the time interval between the instants τ_i and τ_i^*, $i = 1, 2, \cdots, q-1$. Since each of the numbers τ_i, τ_i^* is less than each of the numbers τ_{i+1}, τ_{i+1}^* if the point x_0^* is sufficiently close to x_0, then the entire interval δ_i is located on the real axis to the left of the interval δ_{i+1}. Let Δ_1 denote the interval from the instant t_0 up to the left end of the interval δ_1; let Δ_i, $i = 2, 3, \cdots, q-1$, denote the interval from the right end of the interval δ_{i-1} up to the left end of the interval δ_i; and let Δ_q denote the interval from the

right end of the interval δ_{q-1} up to the point t_1. Thus, the intervals

$$\Delta_1, \delta_1, \Delta_2, \delta_2, \Delta_3, \cdots, \delta_{q-1}, \Delta_q$$

directly abut each other on the real axis. Here, during the time interval Δ_i both the phase points $x(t)$ and $x^*(t)$ are found in the cell σ_i, while during the interval δ_i one of them is found in the cell σ_i and the other in the cell σ_{i+1}.

Let ϵ denote the distance between the points x_0 and x_0^* (we assume that ϵ is sufficiently small). Because the switching instants $\tau_1^*, \tau_2^*, \cdots, \tau_{q-1}^*$ and the corresponding switching points $x^*(\tau_1^*), x^*(\tau_2^*), \cdots, x^*(\tau_{q-1}^*)$ are differentiably dependent on the point $x_0^* \in \sigma_1$, it easily follows that there exists a positive constant C such that the length of each of the intervals δ_i does not exceed $C\epsilon$ and the trajectories $x(t)$ and $x^*(t)$ are at a distance of the order of ϵ from each other:

$$|x(t) - x^*(t)| \leqq C\epsilon, \quad t_0 \leqq t \leqq t_1. \tag{23}$$

As before, we let $\psi(t) = (\psi_1(t), \cdots, \psi_n(t))$ denote the covariant vector-function corresponding to the trajectory $x(t)$ by virtue of the maximum principle. We have by virtue of the relation $x(t_1) = x^*(t_1) = a$:

$$\begin{aligned}
&-\sum_{\alpha=1}^{n} [x^\alpha(t_0) - x^{*\alpha}(t_0)]\psi_\alpha(t_0) \\
&= \sum_{\alpha=1}^{n} [x^\alpha(t_1) - x^{*\alpha}(t_1)]\psi_\alpha(t_1) - \sum_{\alpha=1}^{n} [x^\alpha(t_0) - x^{*\alpha}(t_0)]\psi_\alpha(t_0) \\
&= \int_{t_0}^{t_1} \frac{d}{dt}\left\{\sum_{\alpha=1}^{n} [x^\alpha(t) - x^{*\alpha}(t)]\psi_\alpha(t)\right\} dt \\
&= \int_{t_0}^{t_1} \left\{\sum_{\alpha=1}^{n} \psi_\alpha(t) \frac{d}{dt}[x^\alpha(t) - x^{*\alpha}(t)] + \sum_{\alpha=1}^{n} [x^\alpha(t) - x^{*\alpha}(t)] \frac{d\psi_\alpha(t)}{dt}\right\} dt \\
&= \int_{t_0}^{t_1} \left\{\sum_{\alpha=1}^{n} \psi_\alpha(t)[f^\alpha(x(t), v(x(t))) - f^\alpha(x^*(t), v(x^*(t)))]\right. \\
&\qquad \left. - \sum_{\alpha=1}^{n} [x^\alpha(t) - x^{*\alpha}(t)] \frac{\partial H(\psi(t), x(t), v(x(t)))}{\partial x^\alpha}\right\} dt \\
&= \int_{t_0}^{t_1} \left\{H(\psi(t), x(t), v(x(t))) - H(\psi(t), x^*(t), v(x^*(t)))\right. \\
&\qquad \left. - \sum_{\alpha=1}^{n} [x^\alpha(t) - x^{*\alpha}(t)] \frac{\partial H(\psi(t), x(t), v(x(t)))}{\partial x^\alpha}\right\} dt.
\end{aligned}$$

Thus,

$$-\sum_{\alpha=1}^{n} (x^\alpha(t_0) - x^{*\alpha}(t_0))\psi_\alpha(t_0) = \int_{t_0}^{t_1} F(t)\, dt, \tag{24}$$

where

$$F(t) = H(\psi(t), x(t), v(x(t))) - H(\psi(t), x^*(t), v(x^*(t))) - \sum_{\alpha=1}^{n} [x^\alpha(t) - x^{*\alpha}(t)] \frac{\partial H(\psi(t), x(t), v(x(t)))}{\partial x^\alpha}$$

$$= H(\psi(t), x(t), v(x(t))) - \left\{ H(\psi(t), x(t), v(x^*(t))) + \sum_{\alpha=1}^{n} [x^{*\alpha}(t) - x^\alpha(t)] \frac{\partial H(\psi(t), \xi, v(x^*(t)))}{\partial x^\alpha} \right\} - \sum_{\alpha=1}^{n} [x^\alpha(t) - x^{*\alpha}(t)] \frac{\partial H(\psi(t), x(t), v(x(t)))}{\partial x^\alpha}$$

$$= H(\psi(t), x(t), v(x(t))) - H(\psi(t), x(t), v(x^*(t))) + \sum_{\alpha=1}^{n} [x^\alpha(t) - x^{*\alpha}(t)] \left\{ \frac{\partial H(\psi(t), \xi, v(x^*(t)))}{\partial x^\alpha} - \frac{\partial H(\psi(t), x(t), v(x(t)))}{\partial x^\alpha} \right\}$$

and ξ is some point of the segment connecting $x(t)$ and $x^*(t)$. Hence we get

$$F(t) \geqq \sum_{\alpha=1}^{n} (x^\alpha(t) - x^{*\alpha}(t)) \left\{ \frac{\partial H(\psi(t), \xi, v(x^*(t)))}{\partial x^\alpha} - \frac{\partial H(\psi(t), x(t), v(x(t)))}{\partial x^\alpha} \right\}$$

since $H(\psi(t), x(t), v(x(t))) \geqq H(\psi(t), x(t), u)$ for every $u \in U$.

Thus,

$$F(t) \geqq G(t), \tag{25}$$

where

$$G(t) = \sum_{\alpha=1}^{n} (x^\alpha(t) - x^{*\alpha}(t)) \cdot \left\{ \frac{\partial H(\psi(t), \xi, v(x^*(t)))}{\partial x^\alpha} - \frac{\partial H(\psi(t), x(t), v(x(t)))}{\partial x^\alpha} \right\}. \tag{26}$$

Now if the point t belongs to one of the intervals Δ_i, then the points $x(t)$ and $x^*(t)$ belong to one and the same cell σ_i, on which the function $v(x)$ is continuously differentiable. Moreover, the trajectory $x^*(t)$ lies in a small neighborhood of the trajectory $x(t)$ (see (23)) which is a compact set. Consequently, the following estimate is valid:

$$| v(x^*(t)) - v(x(t)) | \leqq C' \cdot | x^*(t) - x(t) | \leqq CC'\epsilon, \quad t \in \Delta_i .$$

In precisely the same way

$$| \xi - x(t) | \leqq | x^*(t) - x(t) | \leqq C\epsilon.$$

From the continuity of the function $\partial H/\partial x^\alpha$ in its arguments it now fol-

lows that, when $t \in \Delta_i$, the difference

$$\frac{\partial H(\psi(t), \xi, v(x^*(t)))}{\partial x^\alpha} - \frac{\partial H(\psi(t), x(t), v(x(t)))}{\partial x^\alpha} \tag{27}$$

is infinitesimal along with ϵ (i.e., the difference tends to zero as $\epsilon \to 0$; moreover, it converges uniformly with respect to t). Finally, taking (23) into account, by virtue of (26), we obtain

$$\lim_{\epsilon \to 0} \frac{G(t)}{\epsilon} = 0$$

uniformly with respect to $t \in \Delta_i$, whence we find that

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{\Delta_i} G(t)\, dt = 0, \qquad i = 1, 2, \cdots, q. \tag{28}$$

If the point t belongs to one of the intervals δ_i, then we can no longer assert that the difference (27) is infinitesimal along with ϵ, since the points $x(t)$ and $x^*(t)$ belong to different cells and the function $v(x)$ can suffer a discontinuity during the transition from cell to cell. However, the difference (27) remains bounded for all t because the trajectory $x(t)$ is compact. Therefore, by virtue of (23), we have

$$\lim_{\epsilon \to 0} G(t) = 0$$

uniformly in t. Since the length of the interval δ_i does not exceed $C\epsilon$, we obtain

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{\delta_i} G(t)\, dt = 0, \qquad i = 1, 2, \cdots, q - 1. \tag{29}$$

Adding all of the relations (28) and (29) we find that

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{t_0}^{t_1} G(t)\, dt = 0. \tag{30}$$

Now let the point x_0^* approach the point x_0 on the surface S, tangent to a certain vector $p = (p^1, p^2, \cdots, p^n)$. In other words,

$$\lim_{\epsilon \to 0} \frac{x^*(t_0) - x(t_0)}{\epsilon} = \lim_{\epsilon \to 0} \frac{x_0^* - x_0}{\epsilon} = p.$$

Then by virtue of (24), (25), (30) we have

$$\begin{aligned} \sum_{\alpha=1}^{n} p^\alpha \psi_\alpha(t_0) &= \lim_{\epsilon \to 0} \sum_{\alpha=1}^{n} \frac{1}{\epsilon} (x^{*\alpha}(t_0) - x^\alpha(t_0)) \psi_\alpha(t_0) \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{t_0}^{t_1} F(t)\, dt \geqq \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{t_0}^{t_1} G(t)\, dt = 0. \end{aligned}$$

Since the relation

$$\sum_{\alpha=1}^{n} p^{\alpha}\psi_{\alpha}(t_0) \geqq 0$$

is correct for any vector p that is tangent to S, then

$$\sum_{\alpha=1}^{n} p^{\alpha}\psi_{\alpha}(t_0) = 0$$

for any vector p tangent to the hypersurface S; whence follows (22).

Thus, Theorem 5 is proved for the case $f^0 \equiv 1$.

Let us now proceed to prove Theorem 5 in the general case. Let all the conditions A–G indicated in the definition of regular synthesis be satisfied. We select an arbitrary point $x_0 \in V$ and denote by $x(t)$ the marked trajectory starting from the point x_0 at the instant t_0 and by t_1 the instant at which it hits the point a. Since the trajectory $x(t)$ is marked, it satisfies (18) and the relation

$$\int_{t_0}^{t_1} f^0(x(t), v(x(t)))\, dt = -\omega(x_0); \tag{31}$$

furthermore, it satisfies the maximum principle.

Let us introduce the function

$$\tau(t) = \int_{t_0}^{t} f^0(x(t), v(x(t)))\, dt, \qquad t_0 \leqq t \leqq t_1.$$

Since $f^0(x, u) > 0$ for all x, u, the function $\tau(t)$ increases monotonically. Moreover,

$$\tau(t_0) = t_0, \qquad \tau(t_1) = -\omega(x_0). \tag{32}$$

Consequently, on the interval $0 \leqq \tau \leqq -\omega(x_0)$ we can define the function $t(\tau)$ which is the inverse of the function $\tau(t)$. Here

$$t(0) = t_0, \qquad t(-\omega(x_0)) = t_1.$$

Let us now set

$$y(\tau) = x(t(\tau)), \qquad 0 < \tau < -\omega(x_0). \tag{33}$$

Trajectory (33) starts from the point $y(0) = x_0$ and arrives at the point a.

We first establish the fact that the function $v(x)$, the set N, and the piecewise-smooth set (17), constructed for (2) and the functional $J = \int f^0(x, u)\, dt$, effect regular synthesis also for the system of equations

$$\frac{dy^i}{d\tau} = \frac{f^i(y, u)}{f^0(y, u)}, \qquad i = 1, 2, \cdots, n, \tag{34}$$

and time-optimality; moreover, the marked trajectories here are trajectories of the form (33).

Indeed, since the trajectory $x(t)$ satisfies (18), then for trajectory (33) we have:

$$\frac{dy^i(\tau)}{d\tau} = \frac{dx^i(t(\tau))}{d\tau} = \frac{dx^i(t(\tau))}{dt} \cdot \frac{dt(\tau)}{d\tau}$$

$$= f^i(x(t(\tau)), v(x(t(\tau)))) \cdot \frac{1}{f^0(x(t(\tau)), v(x(t(\tau))))}$$

$$= \frac{f^i(y(\tau), v(y(\tau)))}{f^0(y(\tau), v(y(\tau)))}, \qquad i = 1, \cdots, n.$$

Thus, trajectories of the form (33) satisfy the system of equations

$$\frac{dy^i}{d\tau} = \frac{f^i(y, v(y))}{f^0(y, v(y))}, \qquad i = 1, \cdots, n, \tag{35}$$

which plays the same role relative to (34) as (18) does relative to (2).

Let us now verify conditions A–G of regular synthesis. Conditions A and B are not related to any system of equations; since they are fulfilled for (2) they remain valid also when they are applied to (34). Condition C (in which now (35) replaces (18)) is also fulfilled since trajectory (33) satisfies (35). Note that trajectory (33) coincides geometrically with the trajectory $x(t)$ and, therefore, it approaches the cell $\Pi(\sigma)$ at the same angle as the trajectory $x(t)$. The phase velocities of the approaches are connected by the relation

$$\frac{dy(\tau)}{d\tau} = \frac{dx(t(\tau))}{d\tau} = \frac{dx(t(\tau))}{dt} \cdot \frac{dt(\tau)}{d\tau} = \frac{dx(t)}{dt} \cdot \frac{1}{f^0(x(t), v(x(t)))},$$

and, therefore, from the relation $dx/dt \neq 0$ it follows that $dy/d\tau \neq 0$, i.e., trajectory (33), as also the trajectory $x(t)$, approaches the cell $\Pi(\sigma)$ with a nonzero phase velocity.

Condition D is retained without change. Condition E is also satisfied: the marked trajectories now are the trajectories (33). The validity of condition G is also easily verified: the time of motion along trajectory (33) from the point x_0 to the point a equals $-\omega(x_0)$ (see (32)) and therefore it depends continuously on the initial point x_0. It remains only to verify condition F, i.e., to prove that the trajectory (33) of (34) satisfies the maximum principle.

We first write the function H for (34), in which we denote the auxiliary variables by $\varphi_1, \cdots, \varphi_n$:

$$H = \sum_{\alpha=1}^{n} \varphi_\alpha \frac{f^\alpha(y, u)}{f^0(y, u)}; \tag{36}$$

and the system of differential equations for the auxiliary variables:

$$\frac{d\varphi_i}{d\tau} = -\frac{\partial H}{\partial y^i} = -\sum_{\alpha=1}^{n} \frac{\varphi_\alpha}{f^0(y, u)} \cdot \frac{\partial f^\alpha(y, u)}{\partial y^i} + \sum_{\alpha=1}^{n} \frac{\varphi_\alpha f^\alpha(y, u)}{(f^0(y, u))^2} \cdot \frac{\partial f^0(y, u)}{\partial y^i}, \qquad i = 1, \cdots, n. \tag{37}$$

To prove that trajectory (33) satisfies the maximum principle we denote by $\psi(t) = (\psi_0, \psi_1(t), \cdots, \psi_n(t))$ the covariant vector-function corresponding to the trajectory $x(t)$ of (2) by virtue of the maximum principle. The function $\mathfrak{H}$ (see [1, p. 18]) for (2) has the form

$$\mathfrak{H} = \sum_{\alpha=0}^{n} \psi_\alpha f^\alpha(x, u). \tag{38}$$

The auxiliary equations are:

$$\frac{d\psi_i}{dt} = -\frac{\partial \mathfrak{H}}{\partial x^i} = -\sum_{\alpha=0}^{n} \psi_\alpha \frac{\partial f^\alpha(x, u)}{\partial x^i}, \qquad i = 1, \cdots, n. \tag{39}$$

Since $\psi(t)$, $x(t)$, $v(x(t))$ satisfy the maximum principle, these functions satisfy (39) and, furthermore, the relations

$$\mathfrak{H}(\psi(t), x(t), v(x(t))) \equiv 0, \qquad \psi_0 \leqq 0, \tag{40}$$

$$\mathfrak{H}(\psi(t), x(t), v(x(t))) \geqq \mathfrak{H}(\psi(t), x(t), u) \quad \text{for every} \quad u \in U, \tag{41}$$

are satisfied. On the basis of these relations we now show that the maximum principle is satisfied by the following vector-functions:

$$y_i(\tau) = x_i(t(\tau)), \qquad \varphi_i(\tau) = \psi_i(t(\tau)), \quad i = 1, \cdots, n, \quad u(\tau) = v(y(\tau)).$$

First of all we note that the functions $\varphi_1(\tau), \cdots, \varphi_n(\tau)$ do not vanish simultaneously. Indeed, if for some τ the relations $\varphi_1 = \varphi_2 = \cdots = \varphi_n = 0$ were fulfilled, then for the corresponding value of $t = t(\tau)$ we would have

$$\psi_1 = \psi_2 = \cdots = \psi_n = 0,$$

and therefore by virtue of (40) and (38),

$$\psi_0 f^0(x(t), u(t)) = 0,$$

whence $\psi_0 = 0$ since $f^0 \neq 0$. Consequently, the vector function $\psi(t) = (\psi_0, \psi_1(t), \cdots, \psi_n(t))$ would vanish, which is impossible. Thus, the vector-function $\varphi(t) = (\varphi_1(t), \varphi_2(t), \cdots, \varphi_n(t))$ is nontrivial.

Further, from (39), (38), (40) we have

$$\frac{d\varphi_i(\tau)}{d\tau} = \frac{d\psi_i(t(\tau))}{d\tau} = \frac{d\psi_i(t(\tau))}{dt} \cdot \frac{dt(\tau)}{d\tau}$$

$$= \left\{-\sum_{\alpha=0}^{n} \psi_\alpha(t(\tau)) \cdot \frac{\partial f^\alpha(x(t(\tau)), v(x(t(\tau))))}{\partial x^i}\right\} \cdot \frac{1}{f^0(x(t(\tau)), v(x(t(\tau))))}$$

$$= -\sum_{\alpha=1}^{n} \frac{\varphi_\alpha(\tau)}{f^0(y(\tau), v(y(\tau)))} \cdot \frac{\partial f^\alpha(y(\tau), v(y(\tau)))}{\partial y^i}$$

$$- \frac{\psi_0}{f^0(x(t(\tau)), v(x(t(\tau))))} \cdot \frac{\partial f^0(y(\tau), v(y(\tau)))}{\partial y^i}$$

$$= -\sum_{\alpha=1}^{n} \frac{\varphi_\alpha(\tau)}{f^0(y(\tau), v(y(\tau)))} \cdot \frac{\partial f^\alpha(y(\tau), v(y(\tau)))}{\partial y^i}$$

$$- \frac{\psi_0 f^0(x(t(\tau)), v(x(t(\tau))))}{[f^0(x(t(\tau)), v(x(t(\tau))))]^2} \cdot \frac{\partial f^0(y(\tau), v(y(\tau)))}{\partial y^i}$$

$$= -\sum_{\alpha=1}^{n} \frac{\varphi_\alpha(\tau)}{f^0(y(\tau), v(y(\tau)))} \cdot \frac{\partial f^\alpha(y(\tau), v(y(\tau)))}{\partial y^i}$$

$$- \frac{\mathcal{H}(\psi(t(\tau)), x(t(\tau)), v(x(t(\tau)))) - \sum_{\alpha=1}^{n} \psi_\alpha(t(\tau)) f^\alpha(x(t(\tau)), v(x(t(\tau))))}{[f^0(x(t(\tau)), v(x(t(\tau))))]^2}$$

$$\cdot \frac{\partial f^0(y(\tau), v(y(\tau)))}{\partial y^i}$$

$$= -\sum_{\alpha=1}^{n} \frac{\varphi_\alpha(\tau)}{f^0(y(\tau), v(y(\tau)))} \cdot \frac{\partial f^\alpha(y(\tau), v(y(\tau)))}{\partial y^i}$$

$$+ \sum_{\alpha=1}^{n} \frac{\varphi_\alpha(\tau) f^\alpha(y(\tau), v(y(\tau)))}{[f^0(y(\tau), v(y(\tau)))]^2} \cdot \frac{\partial f^0(y(\tau), v(y(\tau)))}{\partial y^i}.$$

Thus, the vector-functions $\varphi(\tau)$, $y(\tau)$, $v(y(\tau))$ satisfy (37). Finally, we have

$$H(\varphi(\tau), y(\tau), u) = \sum_{\alpha=1}^{n} \varphi_\alpha(\tau) \cdot \frac{f^\alpha(y(\tau), u)}{f^0(y(\tau), u)}$$

$$= \frac{1}{f^0(x(t(\tau)), u)} \cdot \sum_{\alpha=1}^{n} \psi_\alpha(t(\tau)) f^\alpha(x(t(\tau)), u)$$

$$= \frac{1}{f^0(x(t(\tau)), u)} \cdot \{\mathcal{H}(\psi(t(\tau)), x(t(\tau)), u) - \psi_0 f^0(x(t(\tau)), u)\}$$

$$= \frac{1}{f^0(x(t(\tau)), u)} \cdot \mathcal{H}(\psi(t(\tau)), x(t(\tau)), u) - \psi_0.$$

Therefore, by virtue of (40) and (41), taking into account that $f^0 > 0$, we obtain

$$H(\varphi(\tau), y(\tau), u) \leqq -\psi_0, \qquad H(\varphi(\tau), y(\tau), v(y(\tau))) = -\psi_0,$$

i.e.,

$$H(\varphi(\tau), y(\tau), u) \leqq H(\varphi(\tau), y(\tau), v(y(\tau))) = -\psi_0 \geqq 0, \quad u \in U.$$

Thus, trajectory (33) satisfies the maximum principle, and condition F also is fulfilled.

Thus, trajectories of the form (33) effect regular synthesis for (34) and time-optimality. Since Theorem 5 has already been proved for the case of time-optimality, it follows from this that all the trajectories (33) are (time-)optimal trajectories of (34).

Now we can prove Theorem 5 in the general case without difficulty. Indeed, let $u_*(t)$, $t_0 \leqq t \leqq t_*$, be an arbitrary admissible control which transfers the phase point moving under law (2) from the position x_0 to the position a; let $x_*(t)$ denote the corresponding trajectory. Thus, the functions $x_*(t)$, $u_*(t)$ satisfy (2), and the relations $x_*(t_0) = x_0$, $x_*(t_*) = a$ are satisfied. The corresponding value of the functional J is denoted by J_* :

$$J_* = \int_{t_0}^{t_*} f^0(x_*(t), u_*(t))\, dt.$$

Let us set

$$\tau_*(t) = \int_{t_0}^{t} f^0(x_*(t), u_*(t))\, dt, \qquad t_0 \leqq t \leqq t_*;$$

then $\tau_*(t)$ is a monotonically increasing function with a defined inverse $t_*(\tau)$, $0 \leqq \tau \leqq J_*$; moreover, we have $t_*(0) = t_0$, $t_*(J_*) = t_*$. It is not difficult to verify that the functions $x_*(t_*(\tau))$, $u_*(t_*(\tau))$ satisfy (34) and that here the time of motion along the trajectory $y_*(\tau) = x_*(t_*(\tau))$ from position x_0 to position a equals J_*. But we already know that trajectory (33) is time-optimal and, therefore, the time of motion J_* along the trajectory $y_*(\tau)$ can be only larger than the time of motion $-\omega(x_0)$ along the trajectory (33):

$$J_* \geqq -\omega(x_0).$$

Taking (41) and (31) into account, we obtain from this the inequality

$$\int_{t_0}^{t_*} f^0(x_*(t), u_*(t))\, dt \geqq \int_{t_0}^{t_1} f^0(x(t), v(x(t)))\, dt.$$

But this signifies that the marked trajectory $x(t)$ leading from the point x_0 to the point a is optimal.

Theorem 5 is completely proved.

7. Examples.

Example 1. Let us consider a controlled plant whose behavior is described by the system

$$\frac{dx^1}{dt} = x^2, \qquad \frac{dx^2}{dt} = f(x^2, u), \tag{42}$$

with a one-dimensional control region U defined by the inequalities

$$-1 \leqq u \leqq 1. \tag{43}$$

We shall assume that the function f is continuously differentiable in both its arguments and satisfies the relations

$$\frac{\partial f(x^2, u)}{\partial u} > 0 \quad \text{for all} \quad x^2, u, \tag{44}$$

$$f(x^2, 1) > c > 0, \qquad f(x^2, -1) < -c, \quad \text{for all} \quad x^2. \tag{45}$$

Any plant described by the equation $\ddot{x} = f(\dot{x}, u)$, "differing slightly" from the equation $\ddot{x} = u$, will satisfy the stated conditions. As an example we can cite the nonlinear equation

$$\ddot{x} = u + \tfrac{1}{2} \sin (\dot{x} + u).$$

For the controlled plant (42)–(43) we investigate the time-optimal problem of hitting the origin from a specified initial phase state. We first write the function

$$H = \psi_1 x^2 + \psi_2 f(x^2, u)$$

and the system of equations for the auxiliary unknowns ψ_1, ψ_2:

$$\text{(a)} \qquad \dot{\psi}_1 = -\frac{\partial H}{\partial x^1} = 0,$$

$$\text{(b)} \qquad \dot{\psi}_2 = -\frac{\partial H}{\partial x^2} = -\psi_1 - \psi_2 \frac{\partial f(x^2, u)}{\partial x^2}. \tag{46}$$

From the condition that the function H should be maximum it follows that the optimal control u is determined by maximizing $f(x^2, u)$ when $\psi_2 > 0$ and by minimizing $f(x^2, u)$ when $\psi_2 < 0$. (The control u is not defined when $\psi_2 = 0$.) But (44) shows that $f(x^2, u)$ is maximum when $u = +1$ (see (43)) and $f(x^2, u)$ is minimum when $u = -1$. In other words, the optimal control satisfies the conditions: $u = +1$ when $\psi_2 > 0$ and $u = -1$ when $\psi_2 < 0$.

It remains to trace out the law whereby the quantity ψ_2 varies. From (46a) we see that $\psi_1 = \text{const}$. Now by solving (46b) as a linear equation in ψ_2 (considering that $u(t)$ and $x^2(t)$ are known) we find that

$$(47) \qquad \psi_2 = \exp\left(-\int_{t_0}^{t} \frac{\partial f}{\partial x^2}\, dt\right) \cdot \left\{\psi_{20} - \psi_1 \int_{t_0}^{t} \exp\left(\int_{t_0}^{\tau} \frac{\partial f}{\partial x^2}\, dt\right) d\tau\right\}.$$

From this formula we see that if the quantity ψ_2 vanishes at some instant t_0 (i.e., $\psi_{20} = 0$), then the function ψ_2 has no other zeros (because the function e^z is positive for all real z) and, moreover, the function ψ_2 changes sign as it passes through zero. Thus, the function ψ_2 does not change sign more than once, i.e., by virtue of the maximum principle the optimal control does not switch more than once. Here, any control $u = \pm 1$ having no more than one switching satisfies the maximum principle, because from (47) we see that by choosing the sign of the constant ψ_1 properly (and assuming $\psi_{20} = 0$), we can make the function $\psi_2(t)$ change sign at any specified instant t_0 either from minus to plus or from plus to minus according to our wish.

Now it is not difficult to find all the trajectories leading to the origin and satisfying the maximum principle. This is done in the same way as in [1, Example 1, pp. 23–27]. Indeed, by virtue of what has been said above, every trajectory which satisfies the maximum principle consists of two segments (one of them may be absent), on the first of which $u = +1$, and on the second of which $u = -1$ (or vice versa). Let us denote by σ_+^1 the semitrajectory of (42) obtained when $u \equiv +1$ and terminating at the origin, and by σ_-^1 the analogous semitrajectory obtained when $u \equiv -1$ (see Fig. 1). Both semitrajectories together form a curve which we denote by P^1. This curve divides the whole plane P^2 of the variables x^1, x^2 into two parts. The part located below the curve P^1 is denoted by σ_+^2, and the part located above, by σ_-^2. The final segment of each trajectory which satisfies the maximum principle is either along σ_+^1 or along σ_-^1, while the initial segment is either in the cell σ_+^2 (when $u = +1$) or in the cell σ_-^2 (when $u = -1$). The indicated trajectories fill the entire plane P^2 and satisfy the maximum principle. By taking all the four cells σ_+^1, σ_-^1, σ_+^2, σ_-^2 to be cells of the first kind, we obtain regular synthesis. Indeed, conditions A–G are verified without difficulty. Let us only note that the trajectories moving in the cell σ_+^2 approach the cell σ_-^1 at a nonzero angle, and the trajectories moving in the cell σ_-^2 approach the cell σ_+^1 at a nonzero angle. This follows from the fact that the phase velocity vector $\{x^2, f(x^2, u)\}$ has two noncollinear directions when $u = +1$ and $u = -1$ (see (45)). From Theorem 5 it now follows that all the trajectories we have constructed are indeed optimal. The corresponding function $v(x)$ which effects the synthesis of the optimal control has the following form:

$$v(x) = \begin{cases} +1 \text{ on the cells } \sigma_+^2 \text{ and } \sigma_+^1, \\ -1 \text{ on the cells } \sigma_-^2 \text{ and } \sigma_-^1. \end{cases}$$

Note that all these discussions remain in force also in the case where

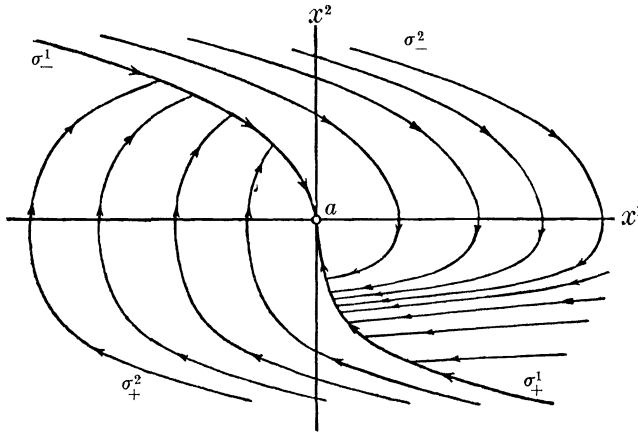

FIG. 1

the constant c turns out to equal zero in (45), i.e., where (45) is replaced by

$$f(x^2, 1) > 0, \qquad f(x^2, -1) < 0, \quad \text{for all} \quad x^2.$$

However, in this case, depending on the form of the function $f(x^2, u)$, the optimal control can be synthesized either on the whole plane, or in some halfplane, or in a strip (whose edges are parallel to the x^1-axis).

Example 2. As a second example consider the plant described by the system

$$\begin{aligned} \frac{dx^1}{dt} &= x^2, \\ \frac{dx^2}{dt} &= -x^1 + g(x^2, u), \end{aligned} \tag{48}$$

with the same control region $-1 \leqq u \leqq 1$. We assume that the function g is continuously differentiable in both arguments and satisfies the conditions

$$\left| \frac{\partial g(x^2, u)}{\partial x^2} \right| < c < 2 \quad \text{for all} \quad x^2, u, \tag{49}$$

$$g(x^2, 1) > c' > 0, \qquad g(x^2, -1) < -c', \quad \text{for all} \quad x^2. \tag{50}$$

Any plant described by the equation $\ddot{x} + x = g(\dot{x}, u)$, "differing slightly" from the equation $\ddot{x} + x = u$, will satisfy the stated conditions. (Compare [1, Example 2, pp. 27–35].) As an example we can cite the nonlinear equation

$$\ddot{x} + x = u + \tfrac{1}{2}\sin(\dot{x} + u).$$

For the controlled plant (48) also we investigate the time-optimal problem of hitting the origin from a specified initial phase state. The function H is

$$H = \psi_1 x^2 + \psi_2(-x^1 + g(x^2, u)),$$

and the system of equations for the auxiliary unknowns is

$$\begin{aligned} \dot{\psi}_1 &= -\frac{\partial H}{\partial x^1} = \psi_2, \\ \dot{\psi}_2 &= -\frac{\partial H}{\partial x^2} = -\psi_1 - \psi_2 \frac{\partial g(x^2, u)}{\partial x^2}. \end{aligned} \tag{51}$$

From the condition that the function H be maximum, as in the preceding example, it follows that $u = \operatorname{sgn} \psi_2$.

Let us now trace out the law whereby the quantity ψ_2 varies. By virtue of (51) we find that

$$\frac{d}{dt}\left(\tan^{-1}\frac{\psi_2}{\psi_1}\right) = \frac{\psi_1\dot{\psi}_2 - \dot{\psi}_1\psi_2}{(\psi_1)^2 + (\psi_2)^2} = \frac{\psi_1\left(-\psi_1 - \psi_2\dfrac{\partial g(x^2, u)}{\partial x^2}\right) - (\psi_2)^2}{(\psi_1)^2 + (\psi_2)^2}$$

$$= -1 - \frac{\psi_1\psi_2}{(\psi_1)^2 + (\psi_2)^2} \cdot \frac{\partial g(x^2, u)}{\partial x^2}.$$

Since

$$\left|\frac{\psi_1\psi_2}{(\psi_1)^2 + (\psi_2)^2}\right| \leqq \frac{1}{2}$$

for all real values of ψ_1 and ψ_2 (not vanishing simultaneously), by virtue of (49) we get

$$-1 - \frac{c}{2} \leqq \frac{d}{dt}\tan^{-1}\frac{\psi_2}{\psi_1} \leqq -1 + \frac{c}{2} < 0.$$

Consequently, the vector (ψ_1, ψ_2) rotates clockwise (and possibly changes its length) with an angular velocity not less than $1 - (c/2) > 0$ and not greater than $1 + (c/2)$. Therefore, the zeros of the function $\psi_2(t)$ (for any $x(t)$, $u(t)$) are distributed not more sparsely than by $2\pi/(2 - c)$ and not more frequently than by $2\pi/(2 + c)$.

Now, let $x_0 = (x_0^1, x_0^2)$ be an arbitrary point on the phase plane P^2. Let us investigate the solution $x(t)$, $\psi(t)$ of (48) and (51), where $u = 1$, satisfying the initial conditions

$$x^1(0) = x_0^1, \quad x^2(0) = x_0^2, \quad \psi_1(0) = 1, \quad \psi_2(0) = 0. \tag{52}$$

We denote by $T_+(x_0)$ the negative root of the function $\psi_2(t)$ closest to zero (occurring in the constructed solution). Then, on the interval $T_+(x_0) < t < 0$ the function $\psi_2(t)$ maintains a constant sign, namely, the sign $+$, since by virtue of (52), from the second equation of (51) we get that $\dot{\psi}_2(0) = -1$. Thus, for the constructed solution we have

$$\psi_2(t) > 0 \quad \text{when} \quad T_+(x_0) < t < 0; \qquad \psi_2(t) = 0 \quad \text{when} \quad t = T_+(x_0).$$

The segment of the phase trajectory $x(t)$ which we have found, corresponding to the time interval $T_+(x_0) \leqq t \leqq 0$, is denoted by $K_+(x_0)$. This segment ends at the point x_0 ; its start is denoted by $\xi_+(x_0)$. Note that ξ_+ is a mapping of the plane P^2 into itself: it puts every point x_0 into correspondence with the point $\xi_+(x_0)$.

Completely analogously, by considering the initial conditions

$$x^1(0) = x_0^1, \qquad x^2(0) = x_0^2, \qquad \psi_1(0) = -1, \qquad \psi_2(0) = 0,$$

we construct the solution $x(t)$, $\psi(t)$ of (48) and (51) where $u = -1$, and denote by $T_-(x_0)$ the negative root of the function $\psi_2(t)$ closest to zero. Then, analogously we obtain

$$\psi_2(t) < 0 \quad \text{when} \quad T_-(x_0) < t < 0; \qquad \psi_2(t) = 0 \quad \text{when} \quad t = T_-(x_0).$$

The segment of the phase trajectory $x(t)$ which we have found, corresponding to the time interval $T_-(x_0) \leqq t \leqq 0$, is denoted by $K_-(x_0)$, and the initial point of this segment by $\xi_-(x_0)$. The segment $K_-(x_0)$ ends at the point x_0 .

We now construct the piecewise-smooth curve P^1 on the phase plane P^2 in the following way. First of all, by choosing the origin a as x_0 , we construct the arcs $\sigma_1^+ = K_+(a)$, $\sigma_1^- = K_-(a)$ (see Fig. 2). Further, we determine the arcs σ_i^+ and σ_i^- by setting

$$\sigma_i^+ = \xi_+(\sigma_{i-1}^-), \qquad \sigma_i^- = \xi_-(\sigma_{i-1}^+), \qquad i = 2, 3, \cdots .$$

Finally, we let P^1 denote the union of all the arcs σ_i^+, σ_i^-, $i = 1, 2, \cdots$.

A simple computation shows that the arcs

$$\cdots, \sigma_3^-, \sigma_2^-, \sigma_1^-, \sigma_1^+, \sigma_2^+, \sigma_3^+, \cdots \tag{53}$$

are arranged on the plane in such a way that every two adjacent arcs in (53) have a common endpoint and these arcs have no other point in common. Therefore, the set P^1 is a simple open curve (Fig. 2); the ends of this curve are at infinity.

Let us now set

$$v(x) = \begin{cases} +1 \text{ below the curve } P^1 \text{ and on the cells } \sigma_i^+, \ i = 1, 2, \cdots, \\ -1 \text{ above the curve } P^1 \text{ and on the cells } \sigma_i^-, \ i = 1, 2, \cdots . \end{cases}$$

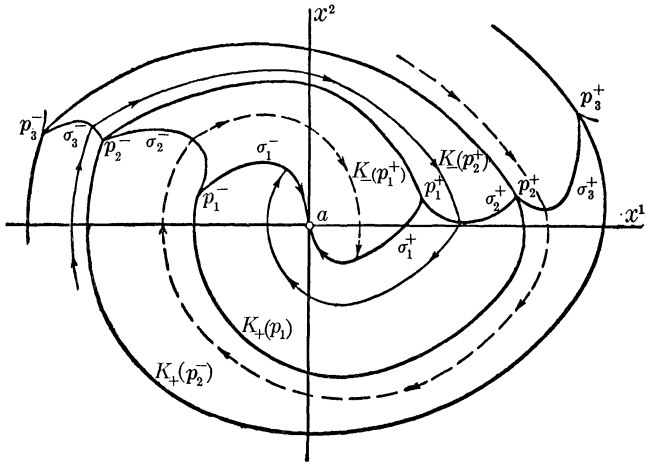

FIG. 2

Finally, we denote by p_i^+ the common endpoint of the arcs σ_i^+ and σ_{i+1}^+ and by p_i^- the common endpoint of the arcs σ_i^- and σ_{i+1}^- for each $i = 1, 2, \cdots$, and we set

$$N = \left[\bigcup_{i=1}^{\infty} K_+(p_i^-)\right] \cup \left[\bigcup_{i=1}^{\infty} K_-(p_i^+)\right].$$

A comparatively simple computation (see [10]) shows that the piecewise-smooth set N, the set $a = P^0 \subset P^1 \subset P^2$, and the function $v(x)$ indicated above effect regular synthesis in the whole phase plane P^2. The corresponding marked trajectories (which by virtue of Theorem 5 are optimal trajectories) are spirals, as shown in Fig. 2. Each such spiral consists either of a certain segment ab of the arc σ_1^+ and the arcs $K_-(b)$, $K_+(\xi_-(b))$, $K_-(\xi_+(\xi_-(b)))$, $\cdots$ (shown in dotted line on Fig. 2), or of a certain segment ac of the arc σ_1^- and the arcs $K_+(c)$, $K_-(\xi_+(c))$, $K_+(\xi_-(\xi_+(c)))$, $\cdots$ (shown in solid line on Fig. 2).

Note that if the constant c' in (50) turns out to equal zero, i.e., if instead of (50) the condition

$$g(x^2, 1) > 0, \qquad g(x^2, -1) < 0, \quad \text{for all} \quad x^2$$

is satisfied, then all the results that have been formulated remain valid except that the synthesis is effected, in general, not in the whole phase plane P^2 but in some region V containing the origin inside it.

Example 3.[4] In conclusion we consider the plant

[4] This example was proposed to the author by A. A. Fel'dbaum (as a simplified electrodrive circuit) and was computed by the student E. Roitenberg while working under the guidance of the author.

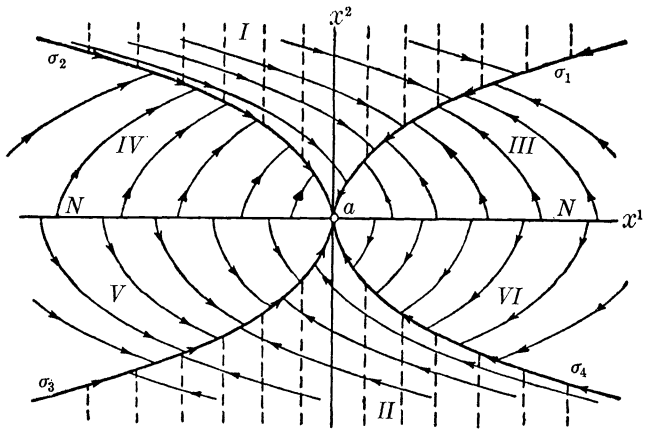

FIG. 3

$$\dot{x}^1 = u^1 x^2, \qquad \dot{x}^2 = u^2, \tag{54}$$

with the control region U defined by the inequalities

$$-1 \leqq u^1 \leqq 1, \qquad -1 \leqq u^2 \leqq 1. \tag{55}$$

For this plant let us consider the time-optimal problem of hitting the origin. Without carrying out the computations (which are not difficult to reproduce) let us mention the final results, i.e., let us describe the regular synthesis for the plant (54)–(55).

The two parabolas $x^1 = \pm\frac{1}{2}(x^2)^2$ together constitute the set P^1; as N we take the axis $x^2 = 0$. Then the set $M = N \cup P^1$ divides the plane P^2 of the variables x^1, x^2 into the six regions I–VI (see Fig. 3), while the point $a = (0, 0)$ divides the set P^1 into four branches (cells) which go off to infinity, σ_1, σ_2, σ_3, σ_4, which are also shown in Fig. 3. Let us now set

$$v^1(x) = \begin{cases} +1 & \text{if } x \in \text{I, II, IV, VI, } \sigma_2, \sigma_4, \\ -1 & \text{if } x \in \text{III, V, } \sigma_1, \sigma_3; \end{cases}$$

$$v^2(x) = \begin{cases} +1 & \text{if } x \in \text{II, III, IV, } \sigma_3, \sigma_4, \\ -1 & \text{if } x \in \text{I, V, VI, } \sigma_1, \sigma_2; \end{cases}$$

$$v(x) = (v^1(x), v^2(x)).$$

It happens that these determine the regular synthesis for the plant (54)–(55). The optimal trajectories consist of segments of the parabolas

$$x^1 = \pm\tfrac{1}{2}(x^2)^2 + \text{const.}$$

They are shown in Fig. 3. Note that two optimal trajectories start from the points of the set N; this does not hinder the application of Theorem 5.

Note also that Theorem 5 does not in any way rule out the existence of other optimal trajectories (besides the marked ones). In Examples 1 and 2 there do not exist other trajectories (besides the marked ones) which satisfy the maximum principle. In Example 3, however, infinitely many optimal trajectories start from each point of cells I, II. Namely, in cell I (or II) we should take $u^2 = -1$ (or $u^2 = +1$), while as u^1 we may take any piecewise-continuous function satisfying (55). At the instant when the phase point hits on the set P^1 a switching occurs and subsequent motion takes place along the set P^1. It is easy to grasp that all the trajectories have one and the same time of motion from the point x_0 to the point a and that all are optimal. For example, in the cells I, II we can take $u^2 = \pm 1$, $u^1 = 0$ (the dotted lines in Fig. 3).

As shown by the cited examples, it is very easy to verify the conditions A–G in Theorem 5. The basic computational difficulty consists of effecting the synthesis on the basis of the maximum principle. If the synthesis has already been effected, then, as a rule, the conditions A–G are automatically satisfied. Thus, the maximum principle (a necessary optimality condition), which, as a rule, permits the synthesis to be effected, is very nearly a sufficient optimality condition.

REFERENCES

[1] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[2] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, 1957.

[3] V. G. Boltyanskii, *Sufficient optimality conditions*, Dokl. Akad. Nauk SSSR, 140 (1961), pp. 994–997; English transl., Soviet Math. Dokl., 2 (1961), pp. 1288–1291.

[4] S. S. Cairns, *On the triangulation of regular loci*, Ann. of Math., 35 (1934), pp. 579–587.

[5] L. S. Pontryagin, *Ordinary Differential Equations*, Addison-Wesley, Reading, Massachusetts, 1962.

[6] ———, *Smooth manifolds and their applications in homotopy theory*, Trudy Mat. Inst. im. Steklov no. 45, Izdat. Akad. Nauk SSSR, Moscow, 1955.

[7] A. Sard, *The measure of the critical values of differentiable maps*, Bull. Amer. Math. Soc., 48 (1942), pp. 883–890.

[8] A. Ya. Dubovickii, *On the differentiable mappings of an n-dimensional cube into a k-dimensional cube*, Mat. Sb., 32(74) (1953), pp. 443–464.

[9] A. F. Timan, *Theory of Approximation of Functions of a Real Variable*, Fizmatgiz, Moscow, 1960.

J. SIAM CONTROL
Vol. 4, No. 2, 1966
Printed in U.S.A.

A NEW REPRESENTATION FOR STOCHASTIC INTEGRALS AND EQUATIONS*

R. L. STRATONOVICH†

Introduction. Stochastic integrals and equations, introduced by Ito [1] (also see [2, chap. VI, §3; chap. IX]), are a convenient means of studying diffusive Markov processes and are being widely used at the present time. In this article we propose another method of representing stochastic integral and differential equations and stochastic integrals. To a significant extent this method is equivalent to Ito's method, but it has a number of advantages in computational techniques. Using this new representation we can work with stochastic integrals in the same way as with the ordinary integrals of smooth functions, for example, we can integrate by parts, etc. In stochastic equations, integral or differential, we can make a change of variables by the usual rules which are suitable in the case of differentiable functions. This is particularly convenient for applications where the investigator has to make actual computations with diffusive processes, just as with the usual (smooth) functions, without having to pay attention to their specific natures, which, in general, requires a more careful treatment.

Stronger reasons for the use of the new method of defining stochastic integrals and equations appear when we have to investigate the questions of the convergence of a non-Markov process to a Markov one by a consideration of the probability functionals, and also the questions on infinitesimal operators which depend on the trajectory of the diffusion process, etc.

The author arrived at the new representation as a result of actual work with smoothed (not completely Markov) processes [3] and with conditional Markov processes [4]. In order to avoid any misunderstanding he takes this opportunity to state explicitly that in articles [3], [4] he used stochastic integrals in the new sense and not in the sense of Ito.

We must remark that in some relations the Ito integral has its own advantages over the new integral, since the former is a martingale and has a mathematical expectation which can be written more concisely. Simple formulas for the transition from one integral to the other allow us at all times to select the representation which is most convenient for any particular purpose.

* Originally published in Vestnik Moskov. Univ. Ser. I Mat. Meh., 1 (1964), pp. 3–12. Submitted on January 15, 1963, for publication. This translation into English has been prepared by N. H. Choksy.

Translated and printed for this journal under a grant-in-aid from the National Science Foundation.

† Department of General Physics, Mechanical-Mathematical Faculty, Moscow State University, Moscow.

1. One-dimensional case. On an interval $T = [a, b]$ let there be given a real diffusive Markov process $\{x(t)\}$ for which

$$(1)\qquad \begin{aligned} &\lim_{h\to 0+0} M\left\{\frac{x(t+h)-x(t)}{h}\,\middle|\, x(t)=\xi\right\} = a(\xi,t),\\ &\lim_{h\to 0+0} M\left\{\frac{[x(t+h)-x(t)]^2}{h}\,\middle|\, x(t)=\xi\right\} = b(\xi,t),\\ &\lim_{h\to 0+0} M\{|\,x(t+h)-x(t)\,|>\delta \mid x(t)=\xi\} = 0, \qquad \delta>0. \end{aligned}$$

We assume that the functions $a(x, t)$ and $b(x, t)$ are continuous in both arguments and, in addition, the second function has the continuous derivative $\partial b(x, t)/\partial x$.

Further, on T let there be given a function $\Phi(x, t)$ continuous in t,having the continuous derivative $\partial\Phi(x, t)/\partial x$ and satisfying the conditions

$$(2)\qquad \begin{aligned} &\int_a^b M\{\Phi(x(t),t)a(x(t),t)\}\,dt < \infty,\\ &\int_a^b M\{|\,\Phi(x(t),t)\,|^2\, b(x(t),t)\}\,dt < \infty. \end{aligned}$$

Let us make a Δ-partitioning:

$$a = t_1^{(\Delta)} < t_2^{(\Delta)} < \cdots < t_N^{(\Delta)} = b, \qquad \Delta = \max\,(t_{j+1} - t_j).$$

DEFINITION. The stochastic integral $\int_a^b \Phi(x(t), t)\,dx(t)$ is defined as the limit-in-the-mean

$$(3)\qquad \begin{aligned} &\int_a^b \Phi(x(t),t)\,dx(t)\\ &\quad = \operatorname*{l.i.m.}_{\Delta\to 0} \sum_{j=1}^{N-1} \Phi\left(\frac{x(t_j)+x(t_{j+1})}{2}, t_j\right)[x(t_{j+1})-x(t_j)]. \end{aligned}$$

On the right-hand side we could also have written

$$\Phi\left(\frac{x(t_j)+x(t_{j+1})}{2}, \frac{t_j+t_{j+1}}{2}\right);$$

however, this does not essentially alter anything.

In what follows we shall prove that the indicated limit exists.

The symmetrized method of defining the integral, mentioned above, differs from the method proposed by Ito. In conformity with the assigned form of the dependency of the function Φ on t and on case (ω) we have

$$(4)\qquad \int_T^* \Phi(x(t),t)\,dx(t) = \operatorname*{l.i.m.}_{\Delta\to 0} \sum_{j-1}^{N-1} \Phi(x(t_j),t_j)[x(t_{j+1})-x(t_j)],$$

where the asterisk denotes the integral in the sense of Ito.

Under the adopted assumptions (the continuity conditions and conditions (2)), the limit in (4) exists and, consequently, the Ito integral exists [1], [2].

Let us prove the existence of the limit in (3) and find the formula relating the two indicated integrals. To do this we select the Δ-partitioning and consider the difference between the limit expressions on the right-hand sides of (3) and (4). Making use of the differentiability with respect to x of the function $\Phi(x, t)$ we get

$$D_\Delta = \sum_{j=1}^{N-1} \left[\Phi\left(\frac{x(t_j) + x(t_{j+1})}{2}, t_j \right) - \Phi(x(t_j), t_j) \right] [x(t_{j+1}) - x(t_j)]$$

$$= \frac{1}{2} \sum_{j=1}^{N-1} \frac{\partial \Phi}{\partial x} [(1 - \theta)x(t_j) + \theta x(t_{j+1}), t_j][x(t_{j+1}) - x(t_j)]^2,$$

$$0 \leqq \theta \leqq \frac{1}{2}, \qquad t_j = t_j^{(\Delta)}.$$

It is not difficult to see that as $\Delta \to 0$ the latter expression tends, with probability 1, to the integral

$$\frac{1}{2} \int \frac{\partial \Phi}{\partial x} (x, t) b(x, t)\, dt.$$

In order to be convinced of this let us make an ϵ-partition of the interval and replace $\partial \Phi(x, t)/\partial x$ by the functions

$$\bar{f}_\epsilon(t) = \sup \left\{ \frac{\partial \Phi}{\partial x} [x(t), t], t \in [t_k^{(\epsilon)}, t_{k+1}^{(\epsilon)}) \right\}, \qquad t \in [t_k^{(\epsilon)}, t_{k+1}^{(\epsilon)}),$$

$$\underline{f}_\epsilon(t) = \inf \left\{ \frac{\partial \Phi}{\partial x} [x(t), t], t \in [t_k^{(\epsilon)}, t_{k+1}^{(\epsilon)}) \right\}, \qquad t \in [t_k^{(\epsilon)}, t_{k+1}^{(\epsilon)}).$$

Then, by denoting

$$\bar{D}_{\Delta\epsilon} = \tfrac{1}{2} \sum \bar{f}_\epsilon(t_j^{(\Delta)})[x(t_{j+1}^{(\Delta)}) - x(t_j^{(\Delta)})]^2,$$

$$\underline{D}_{\Delta\epsilon} = \tfrac{1}{2} \sum \underline{f}_\epsilon(t_j^{(\Delta)})[x(t_{j+1}^{(\Delta)}) - x(t_j^{(\Delta)})]^2,$$

we obviously get

$$\underline{D}_{\Delta\epsilon} < D_\Delta < \bar{D}_{\Delta\epsilon}, \tag{5}$$

where

$$\{t_k^{(\epsilon)}, k = 1, 2, \cdots\} \subset \{t_j^{(\Delta)}, j = 1, 2, \cdots\}.$$

By slightly modified forms of Theorems 2 and 3 from [2, Chap. VIII], the limit,

$$\lim_{\Delta \to 0} \sum_{[x_k^{(\epsilon)}, x_{k+1}^{(\epsilon)})} [x(t_{j+1}^{(\Delta)}) - x(t_j^{(\Delta)})]^2 = \int_{t_k^{(\epsilon)}}^{t_{k+1}^{(\epsilon)}} b(x, t)\, dt,$$

exists with probability 1, so that

$$\lim_{\Delta\to 0} \bar{D}_{\Delta\epsilon} = \frac{1}{2}\int_a^b \bar{f}_\epsilon(x,t)b(x,t)\,dt \equiv \bar{D}_\epsilon,$$
$$\lim_{\Delta\to 0} \underline{D}_{\Delta\epsilon} = \frac{1}{2}\int_a^b \underline{f}_\epsilon(x,t)b(x,t)\,dt \equiv \underline{D}_\epsilon. \tag{6}$$

However, as a consequence of the continuity of the derivative $\partial\Phi(x,t)/\partial x$, the differences $\bar{f}_\epsilon - \underline{f}_\epsilon$ and $\bar{D}_\epsilon - \underline{D}_\epsilon$ may be made as small as desired by decreasing ϵ. Therefore, from (5) and (6) there follows the existence, with probability 1, of the limit,

$$\lim_{\Delta\to 0} D_\Delta = \lim_{\epsilon\to 0} \bar{D}_\epsilon = \lim_{\epsilon\to 0} \underline{D}_\epsilon = \frac{1}{2}\int_a^b \frac{\partial\Phi(x,t)}{\partial x}\, b(x,t)\,dt. \tag{7}$$

Thus, under the stated assumptions, integral (3) exists and is related to the Ito integral by the formula

$$\int_T \Phi[x(t),t]\,dx(t) = \int_T \Phi[x(t),t]\,dx(t) + \frac{1}{2}\int_T \frac{\partial\Phi}{\partial x}[x(t)]b[x(t),t]\,dt \tag{7'}$$

almost certainly.

Example. Let us consider the example cited in [2, p. 392]. Let $x(t)$ be a brownian motion process with diffusion parameter $b(x,t) = 1$. Then, instead of the formula

$$\int_a^{b*} [x(t) - x(a)]\,dx(t) = \frac{1}{2}[x(b) - x(a)]^2 - \frac{1}{2}(b-a),$$

we shall have for integral (3) the simpler formula

$$\int_a^b [x(t) - x(a)]\,dx(t) = \frac{1}{2}[x(b) - x(a)]^2.$$

It can be obtained by a direct integration by parts as for ordinary integrals.

2. Multidimensional generalization. In an analogous manner we define stochastic integrals in the case where we have several diffusion processes $\mathbf{x}(t) = \{x_1(t), \cdots, x_n(t)\}$, described by the drift vector $\mathbf{a}(\mathbf{x},t) = \{a_\alpha(\mathbf{x},t), \alpha = 1, \cdots, n\}$ and by the local diffusion matrix $\{b_{\alpha\beta}(\mathbf{x},t), \alpha,\beta = 1, \cdots, n\}$. Further, let there be given the functions $\{\Phi_\alpha(\mathbf{x},t), \alpha = 1, \cdots, n\}$. We shall assume that the functions

$$a_\alpha(\mathbf{x},t), \quad \frac{\partial b_{\alpha\beta}}{\partial x_\gamma}(\mathbf{x},t), \quad \frac{\partial\Phi_\alpha}{\partial x_\beta}(\mathbf{x},t), \qquad \alpha,\beta,\gamma = 1,\cdots,n, \quad t \in T,$$

are continuous in all their arguments and also that the conditions, which are the multidimensional generalizations of conditions (2), are satisfied. We can then define the stochastic integral

$$(8)\quad \int_a^b \Phi_\alpha(\mathbf{x}, t)\, dx_\alpha = \text{l.i.m.} \sum_{j=1}^{N-1} \Phi_\alpha\left(\frac{x(t_j) + x(t_{j+1})}{2}, t_j\right)[x_\alpha(t_{j+1}) - x_\alpha(t_j)].$$

THEOREM 1. *The limit on the right-hand side of* (8) *exists almost certainly and is related to the integral in the sense of Ito by the relation*

$$(9)\quad \int_a^b \Phi_\alpha(\mathbf{x}, t)\, dx_\alpha = \int_a^{b*} \Phi_\alpha(\mathbf{x}, t)\, dx_\alpha + \frac{1}{2}\int_a^b \frac{\partial \Phi_\alpha}{\partial x_\beta}(\mathbf{x}, t) b_{\alpha\beta}(\mathbf{x}, t)\, dt,$$

which is satisfied with probability 1.

In (8), (9), and in the following, we have assumed a summation over two repeated indices. The proof of Theorem 1 is analogous to the proof in the one-dimensional case.

It is sometimes convenient to treat the stochastic integral as a function of a variable upper limit. Given the function $\Psi(\mathbf{x}, t)$, also continuous, let us consider the sum of the integrals

$$(10)\qquad z(t) = \int_a^t \Psi(\mathbf{x}, t)\, dt + \int_a^t \Phi_\alpha(\mathbf{x}, t)\, dx_\alpha, \qquad t \leqq b.$$

It is interesting to compute limits of type (1) for the indicated integral. In distinction from (1), however, it is advisable to set the condition $\mathbf{x}(t) = \xi$ and not $z(t) = \xi$. It is not difficult to convince ourselves that a continuity condition of the type of the third equation in (1) is satisfied with probability 1. The following theorem is true.

THEOREM 2. *Under the accepted assumptions, integral* (10) *as a function of the upper limit is, almost certainly, characterized by the parameters*

$$\lim_{h\to 0+0} M\left\{\frac{z(t+h) - z(t)}{h} \,\middle|\, \mathbf{x}(t) = \xi\right\} = \Psi(\xi, t) + \Phi_\alpha(\xi, t) a_\alpha(\xi, t) + \frac{1}{2}\frac{\partial \Phi_\alpha}{\partial x_\beta}(\xi, t) b_{\alpha\beta}(\xi, t),$$

$$(11)\quad \lim_{h\to 0+0} M\left\{\frac{[z(t+h) - z(t)]^2}{h} \,\middle|\, \mathbf{x}(t) = \xi\right\} = \Phi_\alpha(\xi, t) b_{\alpha\beta}(\xi, t) \Phi_\beta(\xi, t),$$

$$\lim_{h\to 0+0} M\left\{\frac{[x_\alpha(t+h) - x_\alpha(t)][z(t+h) - z(t)]}{h} \,\middle|\, \mathbf{x}(t) = \xi\right\} = b_{\alpha\beta}(\xi, t)\Phi_\beta(\xi, t).$$

These relations can be proven by utilizing the theory developed for the Ito integral [2, Chap. IX, §5], and also the connecting formula (9). As can be seen from (11), the formula for computing the mean increment $M\{dz/dt \mid \mathbf{x}(t)\}$ is not a trivial one. The term

$$\frac{1}{2}\frac{\partial \Phi_\alpha}{\partial x_\beta} b_{\alpha\beta}$$

which complicates matters, arises because of the presence of correlation between the processes $x_\alpha(t)$, occurring as arguments of Φ, and the increments dx_α.

3. Stochastic equations. In certain special cases the processes $\{x_\alpha(t), \alpha = 1, \cdots, n\}$ and the functions $\Psi, \Phi_\alpha, \alpha = 1, \cdots, n$, are such that the process (10) vanishes identically with probability 1: $z(t) = 0, t \in T$. We shall say that in this case the stochastic equation

$$\int_a^t \Psi[\mathbf{x}(t), t]\, dt + \int_a^t \Phi_\alpha[\mathbf{x}(t), t]\, d\mathbf{x}_\alpha(t) = 0, \qquad t \in T, \tag{12}$$

is satisfied and we shall investigate those connections between the processes $x_1(t), \cdots, x_n(t)$ for which this equation holds.

Special case. Given the two processes $x_1(t) = x(t)$ and $x_2(t) = y(t)$, let the functions Φ_1, Φ_2, Ψ have the following special forms:

$$\Phi_1(\mathbf{x}, t) = -1, \qquad \Phi_2(\mathbf{x}, t) = \sigma(x, t), \qquad \Psi(\mathbf{x}, t) = m(x, t).$$

Then (12) takes the form

$$x(t) = x(a) + \int_a^t m[x(t), t]\, dt + \int_a^t \sigma[x(t), t]\, dy(t). \tag{13}$$

This relation may be called the *stochastic transformation* of process $y(t)$ into $x(t)$. Let us write out the equalities (11) for the given case, taking into account that here the process $z(t)$ and, consequently, also the limits on the left-hand sides, equal zero. This gives

$$m(x, t) - a_1 + \sigma(x, t)a_2 + \frac{1}{2}\frac{\partial\sigma(x, t)}{\partial x} b_{12} = 0,$$

$$b_{11} - 2\sigma(x, t)b_{12} + \sigma^2(x, t)b_{22} = 0,$$

$$-b_{11} + \sigma(x, t)b_{12} = 0, \qquad -b_{12} + \sigma(x, t)b_{22} = 0.$$

Hence, we find that, almost certainly,

$$a_1 = \sigma(x, t)a_2 + m(x, t) + \frac{1}{2}\frac{\partial\sigma(x, t)}{\partial x}\sigma(x, t)b_{22}, \tag{14}$$

$$b_{11} = \sigma^2(x, t)b_{22}, \qquad b_{12} = \sigma(x, t)b_{22}.$$

Going on to a still more special case, let $y(t)$ be a Wiener process, i.e., $a_2 = 0, b_{22} = 1$. Then, from (14) we have

$$a_1 = m(x, t) + \frac{1}{2}\frac{\partial\sigma(x, t)}{\partial x}\sigma(x, t), \qquad b_{11} = \sigma^2(x, t),$$

or, by solving these equalities for $m(x, t)$ and $\sigma(x, t)$, we have

$$m(x, t) = a_1(x, t) - \frac{1}{4}\frac{\partial b_{11}(x, t)}{\partial x}, \qquad \sigma(x, t) = \sqrt{b_{11}(x, t)}. \tag{15}$$

Thus, if the functions $a(x, t)$ and $b(x, t)$ are the drift and the local diffusion of the diffusion process $x(t)$ and if they satisfy the continuity conditions mentioned earlier, then the process $x(t)$ can be described by the stochastic equation

$$dx(t) = \left[a(x, t) - \frac{1}{4}\frac{\partial b(x, t)}{\partial x}\right] dt + \sqrt{b(x, t)}\, dy(t), \tag{16}$$

which is to be understood in the sense of the integral equation (13).

Multidimensional stochastic equation. The latter result has the following multidimensional generalization.

THEOREM 3. *If the multidimensional process* $\mathbf{x}(t) = \{x_1(t), \cdots, x_n(t)\}$ *is described by the equation*

$$dx_\alpha(t) = m_\alpha(\mathbf{x}, t)\, dt + \sigma_{\alpha r}(\mathbf{x}, t)\, dy_r(t), \quad \alpha = 1, \cdots, \quad r = 1, \cdots, l, \tag{17}$$

where $m_\alpha(\mathbf{x}, t)$, $\sigma_{\alpha r}(\mathbf{x}, t)$ *are continuous functions having continuous first derivatives with respect to* $x_1, \cdots, x_n$ *and* $\{y_1(t), \cdots, y_l(t)\}$ *is a system of Wiener processes with a unit local diffusion matrix, then* $\mathbf{x}(t)$ *has the following drift and local diffusion parameters*:

$$a_\alpha(\mathbf{x}, t) = m_\alpha(\mathbf{x}, t) + \frac{1}{2}\frac{\partial \sigma_{\alpha r}}{\partial x_\beta}\sigma_{\beta r}, \qquad b_{\alpha\beta}(\mathbf{x}, t) = \sigma_{\alpha r}(\mathbf{x}, t)\sigma_{\beta r}(\mathbf{x}, t). \tag{18}$$

Equation (17) is to be understood in the sense that the corresponding integral relation is valid by our definition of the integral. As in the one-dimensional case, this result follows from Theorem 2.

As is seen from (17) and (18), it is sometimes convenient to consider, instead of the drift parameter $a_\alpha(\mathbf{x}, t)$ and the local diffusion parameter $b_{\alpha\beta}(\mathbf{x}, t)$, the vectors $m_\alpha(\mathbf{x}, t)$, $\sigma_{\alpha 1}(\mathbf{x}, t)$, $\cdots$, $\sigma_{\alpha l}(\mathbf{x}, t)$ which are defined by (18). The vector

$$m_\alpha(\mathbf{x}, t) = a_\alpha(\mathbf{x}, t) - \frac{1}{2}\frac{\partial \sigma_{\alpha r}}{\partial x_\beta}\sigma_{\beta r} \tag{19}$$

has, in comparison with $a_\alpha(\mathbf{x}, t)$, the advantage that it transforms in a trivial manner under a change of variable. Thus, in the one-dimensional case to the change of variable $x \to \tilde{x} = \int \phi(x)\, dx$ there corresponds the parameter transformations

$$m = \tilde{m} = \phi m, \qquad \sigma \to \phi\sigma, \qquad b \to \phi^2 b$$

($\phi(x)$ is a continuous positive function). The same situation holds in the multidimensional case.

THEOREM 4. *Under the change of variables* $\mathbf{x} = \tilde{\mathbf{x}}(x)$, *the vectors* $\mathfrak{d}_1, \cdots, \mathfrak{d}_l, \mathbf{m}$ *transform covariantly with the vector* $d\mathbf{x}$:

$$\sigma_{\alpha r} = \frac{\partial \tilde{x}_\alpha}{\partial x_\beta}\,\sigma_{\beta r}\,, \qquad \tilde{m}_\alpha = \frac{\partial \tilde{x}_\alpha}{\partial x_\beta}\,m_\beta\,. \tag{20}$$

Consequently, (17) *here transforms as if the processes* $x_1(t), \cdots, x_n(t)$ *were smooth functions of time.*

The statement of the theorem regarding the vectors $\mathfrak{d}_1, \cdots, \mathfrak{d}_l$ follows directly from the tensor nature of the parameters $b_{\alpha\beta}$ and from the definitions in (18) of these vectors. In order to prove the covariancy of the vector m_α let us take into consideration the known formula for the transformation of the drift parameters:

$$\tilde{a}_\alpha = \frac{\partial \tilde{x}_\alpha}{\partial x_\beta}\,a_\beta + \frac{1}{2}\,\frac{\partial^2 \tilde{x}_\alpha}{\partial x_\beta\,\partial x_\gamma}\,b_{\beta\gamma}\,. \tag{21}$$

Further, by the substitution

$$\sigma_{\beta r} = \frac{\partial \tilde{x}_\beta}{\partial x_\gamma}\,\sigma_{\gamma r}\,,$$

we find

$$\frac{\partial \tilde{\sigma}_{\alpha r}}{\partial \tilde{x}_\beta}\,\tilde{\sigma}_{\beta r} = \frac{\partial \tilde{\sigma}_{\alpha r}}{\partial \tilde{x}_\beta}\,\frac{\partial \tilde{x}_\beta}{\partial x_\gamma}\,\sigma_{\gamma r} = \frac{\partial \tilde{\sigma}_{\alpha r}}{\partial x_\gamma}\,\sigma_{\gamma r}\,.$$

Making this substitution for a second time we get

$$\begin{aligned}\frac{\partial \tilde{\sigma}_{\alpha r}}{\partial x_\gamma}\,\sigma_{\gamma r} &= \frac{\partial}{\partial x_\gamma}\left(\frac{\partial \tilde{x}_\alpha}{\partial x_\rho}\,\sigma_{\rho r}\right)\sigma_{\gamma r} \\ &= \frac{\partial \tilde{x}_\alpha}{\partial x_\rho}\cdot\frac{\partial \sigma_{\rho r}}{\partial x_\gamma}\,\sigma_{\gamma r} + \frac{\partial^2 \tilde{x}_\alpha}{\partial x_\gamma\,\partial x_\rho}\,\sigma_{\rho r}\,\sigma_{\gamma r}\,.\end{aligned}$$

Consequently,

$$\frac{\partial \tilde{\sigma}_{\alpha r}}{\partial \tilde{x}_\beta}\,\tilde{\sigma}_{\beta r} = \frac{\partial \tilde{x}_\alpha}{\partial x_\rho}\cdot\frac{\partial \sigma_{\rho r}}{\partial x_\gamma}\,\sigma_{\gamma r} + \frac{\partial^2 \tilde{x}_\alpha}{\partial x_\gamma\,\partial x_\rho}\,b_{\gamma\rho}\,. \tag{22}$$

By subtracting half of (22) from (21) and using (19), we convince ourselves of the validity of the last formula in (20).

Note that as a consequence of the symmetry and nonnegative-definiteness of the local diffusion matrix, there always exists at least one system of real vectors $\mathfrak{d}_1, \cdots, \mathfrak{d}_l$.

Namely, if $U = \| u_{\alpha r} \|$ is an orthogonal transformation reducing this matrix to the diagonal form: $b_{\alpha\beta}u_{\alpha r}u_{\beta s} = b_r^0\delta_{rs}$, then $b_{\alpha\beta} = u_{\alpha r}u_{\beta r}b_r^0$ and, obviously, we can set $\sigma_{\alpha r} = u_{\alpha r}\sqrt{b_r^0}$, where $l = \operatorname{rank} \| b_{\alpha\beta} \|$.

4. Invariant notation for the Kolmogorov equations. An invariant representation of the Kolmogorov equations in arbitrary curvilinear coordinates

was proposed in [5], [6]. In the first of these the consideration was restricted to the case of a nonsingular local diffusion matrix which was chosen as the metric tensor. In the second paper the metric tensor was assumed to be independent, but an essential restriction was introduced along another line, namely, not the whole phase space but only the space corresponding to one-half the variables (the coordinates, not the velocities) was chosen as the metric space. The local diffusion matrix, on the contrary, corresponded only to the velocity space and, moreover, was assumed to be nonsingular.

The vectors $\mathbf{m}$ and $\mathbf{\sigma}_r$ introduced above allow us to obtain an invariant notation for the Kolmogorov equations in the general case of an arbitrary metric phase space. In the special cases mentioned above, this form of notation does not coincide with the forms previously suggested, but is simpler.

Starting here we shall assume that the phase variables are contravariant components of a vector and write them as x^α. According to Theorem 4 the vectors considered therein are also contravariant and therefore we shall write them as m^α and $\sigma^\alpha(r)$ (we write the index r in parentheses since it is not of tensor nature).

Let us consider the Markov probability density $p(x, t; x', t')$ of the transition from the point x to x' during the time from t to t'.

The Kolmogorov equation of the first kind

$$-\frac{\partial p}{dt} = a^\alpha \frac{\partial p}{\partial x^\alpha} + \frac{1}{2} b^{\alpha\beta} \frac{\partial^2 p}{\partial x^\alpha \partial x^\beta},$$

with due regard to (18), transforms to the invariant form

$$-\frac{\partial p}{\partial t} = m^\alpha \frac{\partial p}{\partial x^\alpha} + \frac{1}{2} \sigma^\alpha(r) \frac{\partial}{\partial x^\alpha}\left[\sigma^\beta(r) \frac{\partial p}{\partial x^\beta}\right]. \tag{23}$$

Indeed, as a function of x the transition probability $p(x, t; x', t')$ is a scalar. Therefore, the expressions $m^\alpha(\partial p/\partial x^\alpha)$, $\sigma^\beta(\partial p/\partial x^\beta) = v$ and, consequently, also $\sigma^\alpha(\partial v/\partial x^\alpha)$, are all scalars. Thus, on the right-hand side of (23), as also on the left, we have scalars.

By an analogous substitution of formulas (18), the equation of the second kind

$$\frac{\partial p}{\partial t'} = -\frac{\partial}{\partial x'^\alpha}[a^\alpha p] + \frac{1}{2} \frac{\partial^2}{\partial x'^\alpha \partial x'^\beta}[b^{\alpha\beta} p]$$

transforms to the form

$$\frac{\partial p}{dt'} = -\frac{\partial}{\partial x'^\alpha}[m^\alpha p] + \frac{1}{2} \frac{\partial}{\partial x'^\alpha}\left[\sigma^\alpha(r) \frac{\partial}{\partial x'^\beta}(\sigma^\beta(r) p)\right]. \tag{24}$$

Considered as a function of x', the transition probability $p(x, t; x', t')$ is a scalar density, i.e., it transforms as $\sqrt{g} = \det^{1/2} \| g_{\alpha\beta} \|$. Therefore, the

quantities $m^\alpha p$ and $\sigma^\beta p$ are vector densities. However, if $\mathfrak{A}^\alpha$ is a vector density, then, as is well-known, the divergence $(\partial/\partial x^\alpha)\mathfrak{A}^\alpha$ is again a scalar density. Therefore,

$$\frac{\partial}{\partial x'^\alpha}(m^\alpha p), \qquad \frac{\partial}{\partial x'^\beta}(\sigma^\beta p) = V,$$

and also $\partial(\sigma^\alpha V)/\partial x'^\alpha$ are scalar densities similar to the quantity on the left-hand side of (24).

We can introduce the probability flow

$$\mathfrak{S}^\alpha = m^\alpha p - \frac{1}{2}\sigma^\alpha(r)\frac{\partial[\sigma^\beta(r)p]}{\partial x^\beta},$$

which is a vector density. Then (24) takes the form of the conservation equation

$$\frac{\partial p}{\partial t'} + \frac{\partial \mathfrak{S}^\alpha}{\partial x'^\alpha} = 0.$$

Equations (23) and (24) correspond to one and the same invariant infinitesimal operator

$$dL = \left[m^\alpha \frac{\partial}{\partial x^\alpha} + \frac{1}{2}\sum_r \left(\sigma^\alpha(r)\frac{\partial}{\partial x^\alpha}\right)^2\right] dt.$$

In conclusion we note that the condition, which has been mentioned repeatedly, that the functions $b_{\alpha\beta}$, Φ_α, and their derivatives be continuous, may be weakened. Thus, for example, the results are easily extended to the case of piecewise continuity, etc., but we shall not go into this here.

The author thanks E. B. Dynkin and others for participating in discussions of the author's report on the stated questions in a seminar in the Department of Probability Theory at the Moscow State University.

REFERENCES

[1a] K. Ito, *Stochastic integral*, Proc. Imp. Acad. Tokyo, 20 (1944), pp. 519–524.
[1b] ———, *On a stochastic integral equation*, Proc. Japan Acad., 1–4 (1946), pp. 32–35.
[1c] ———, *On a stochastic differential equation*, Mem. Amer. Math. Soc., 4 (1951), pp. 51–89.
[2] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.
[3] R. L. Stratonovich, *Selected questions on fluctuation theory in radio engineering*, Soviet Radio, Moscow, 1961.
[4a] ———, *On the theory of optimal nonlinear filtering of random functions*, Theor. Probability Appl., 4 (1959), pp. 239–241.
[4b] ———, *Conditional Markov processes*, Ibid., 5 (1960), pp. 172–195.
[5] A. Kolmogoroff, *Zur Umkehrbarkeit der statistischen Naturgesetze*, Math. Ann., 113 (1937), pp. 766–772.
[6] A. M. Yaglom, *On the statistical reversibility of brownian motion*, Mat. Sb., 24(66) (1949), pp. 457–492.

J. SIAM CONTROL
Vol. 4, No. 2, 1966
Printed in U.S.A.

THE EXISTENCE OF OPTIMAL CONTROLS FOR A PERFORMANCE INDEX WITH A POSITIVE INTEGRAND*

M. ANVARI† AND R. F. DATKO‡

1. A control problem of some current interest [1], [2], [3] is to determine a measurable vector valued function

$$u = u(t) = (u_1(t), \cdots, u_m(t))$$

which minimizes the functional

$$J[u] = \int_0^T [\langle x(t), Qx(t)\rangle + c \| u(t) \|^2] \, dt \tag{1.1}$$

and is subject to the following restrictions:

$$\dot{x} = A(t)x + B(t)u(t), \quad \text{where} \quad \dot{x} = \frac{dx}{dt}, \quad x = (x_1, \cdots, x_n), \tag{1.2}$$

$A(t)$ is an $n \times n$ matrix, and $B(t)$ is an $n \times m$ matrix;

$$x(0) = x_0 ; \tag{1.3}$$

$$\lim_{t \to T^-} x(t) = 0; \tag{1.4}$$

$$\| u(t) \| = [u_1^2(t) + \cdots + u_m^2(t)]^{1/2} \leqq 1. \tag{1.5}$$

Here it is assumed that Q is a positive semidefinite $n \times n$ matrix, c is a positive scalar, and $\langle , \rangle$ denotes the usual inner product.

In general a minimizing mapping u does not exist for such a system. However, as will be demonstrated in this note, with suitable restrictions on $A(t)$, $B(t)$, and Q a minimizing mapping does exist.

We consider a generalization of the above problem where $\langle x, Qx\rangle$ is replaced by a positive definite mapping $Q(\| x \|)$ and $c\| u \|^2$ is replaced by a positive convex function $f(\| u \|)$ which satisfies conditions to be described below.

2. Before proceeding to the main result we will establish some facts concerning Orlicz spaces (generalized L^ρ spaces for $1 < \rho < \infty$) which will be needed in the proof of Theorem 2 in §3.

All of the definitions and results in this section with the exception of

* Received by the editors August 6, 1964, and in final revised form August 13, 1965.

† Department of Mathematics, University of British Columbia, Vancouver, British Columbia.

‡ RIAS, Baltimore, Maryland, Now at Department of Mathematics, McGill University, Montreal, Quebec. This research was partially supported by the United States Army Research Office (Durham) under Contract No. DA-36-034-AMC-0221X.

Theorem 1 and property 4 can be found in [4] or [5]. Theorem 1 is established in Appendix 1 and property 4 in Appendix 2.

DEFINITION 1. A function $f: R^+ \to R^+$ is called an *N-function* if it admits the representation

$$f(x) = \int_0^x p(t)\,dt, \tag{2.1}$$

where $p(t)$ is right-continuous for $t \geqq 0$, positive for $t > 0$, nondecreasing, and satisfies the conditions

$$p(0)=0, \qquad p(\infty) = \lim_{t\to\infty} p(t) = \infty.$$

DEFINITION 2. Let f be as in Definition 1. The function g, which is defined by

$$g(y) = \max_{x\geqq 0} (x|\,y\,| - f(x)),$$

is called the *complimentary function* of f.

In [5] it is shown that g is also an N-function. In fact it is the right inverse of f.

DEFINITION 3. An N-function f *satisfies the* Δ_2*-condition* if there exists a constant $K > 1$ such that $f(2x) \leqq Kf(x)$ for all $x \geqq 0$.

Remark. It is easy to show that f is a convex function and from this and the Δ_2-condition that $f(mx) \leqq K(m)f(x)$ for all $x \geqq 0$, where $m \geqq 0$ and K depends only on m.

DEFINITION 4. Let the N-function f and its complimentary function g satisfy the Δ_2-condition and let B be the family of measurable mappings $u: R^+ \to R^m$ such that

$$\int_0^\infty f(\|\,u(t)\,\|)\,dt < \infty, \quad \text{where} \quad \|\,u(t)\,\| = \left[\sum_{i=1}^m u_i^2(t)\right]^{1/2}.$$

For any u in B we define the mapping $\|\cdot\|_B : B \to R^+$ as follows:

$$\|\,u\,\|_B = \inf\left\{k > 0 \,\middle|\, \int_0^\infty f\left(\frac{\|\,u(t)\,\|}{k}\right) dt \leqq 1\right\} \quad \text{if} \quad u(t) \neq 0 \quad \text{a.e. in } R^+;$$

$$\|\,u\,\|_B = 0 \quad \text{if} \quad u(t) = 0 \quad \text{a.e. in } R^+.$$

It is a straightforward exercise to show that the family B is linear and that $\|\cdot\|_B$ is a norm on B (e.g., see [5]).

THEOREM 1. *If f and its complimentary function g satisfy Definitions 1–3, then the family B, with the above norm, is a separable reflexive Banach space.*

The proof is given in Appendix 1.

B has the following properties:

(1) If $\{u_n\}$ is a sequence in B which is bounded in norm, then there exists a subsequence $\{u_{n_1}\}$ such that $\{u_{n_1}\} \to^w u$ in B where $\to^w$ denotes weak convergence.

(2) If $\{u_n\} \to^w u$, then $\underline{\lim} \| u_n \|_B \geqq \| u \|_B$.

(3) If $\| u \|_B \geqq 1$, then $\int_0^\infty f(\| u(t) \|)\, dt \geqq \| u \|_B$.

(4) If $u_n \to^w u$ on the interval $[0, T)$, where $T < \infty$, and where, for each $n = 1, 2, \cdots$ and $t \in [0, T]$, $\| u_n(t) \| \leqq M < +\infty$, then

$$\underline{\lim} \int_0^T f(\| u_n(t) \|)\, dt \geqq \int_0^T f(\| u(t) \|)\, dt.$$

The proof of properties 1 and 2 can be found in any text on functional analysis, property 3 can be found in [5], and property 4 is proved in Appendix 2.

3. Let

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \tag{3.1a}$$

$$\dot{x}^{n+1}(t) = Q(x(t)) + f(\| u(t) \|), \tag{3.1b}$$

where $\| u(t) \| = [\sum_{i=1}^m u_i^2(t)]^{1/2}$ and A and B are respectively continuous $n \times n$ and $n \times m$ matrices which are uniformly bounded for all $t \geqq 0$. We assume u lies in the subset U of B consisting of all mappings $u \in B$ with the property with $u: R^+ \to K$, where K is a compact convex subset of R^m which has a nonempty interior containing the origin, and that $Q: R^n \to R^+$ is a continuous positive definite mapping.

Let $\bar{x}(t) = (x(t), x^{n+1}(t))$ be a solution of (3.1) in the sense of Carathéodory. The coordinate $x^{n+1}(t)$ is what is commonly referred to as a performance index.

Suppose for a given u in U there is a solution $x_u(t)$ of (3.1a) with $x_u(0) = x_0 \neq 0$ and $\lim_{t\to\infty} x_u(t) = 0$ such that $\lim_{t\to\infty} x_u{}^{n+1}(t) < \infty$. Denote by C the set of all u in U with this property and let

$$J[u] = \int_0^\infty [Q(x_u(s) + f(\| u(s) \|)]\, ds.$$

THEOREM 2. *Under the above assumptions there exists an optimal mapping in U, that is, there is a $\bar{u}$ in C such that*

$$J(\bar{u}) = \inf_{u \in C} J[u].$$

Proof. Let $J_0 = \inf_{u\in C} J[u]$, and assume C has an infinite number of members (otherwise there would be nothing to prove). Since $J[u]$ is finite for some u in C, we know $J_0 \geqq 0$. Let $\{J[u_n]\} = \{J_n\}$ be a sequence which converges to J_0 and is bounded above by some finite constant M. Let $\{\epsilon_k\}$

be a null sequence of strictly decreasing positive constants such that $\| x_0 \| > \epsilon_1 > \epsilon_2 > \cdots > \epsilon_k > \cdots$. Let t_{n_1} be the first time the trajectory $x_{u_n}(t) = x_n(t)$ strikes the surface $\partial(\epsilon_1)$ of the ball $\| x \| \leqq \epsilon_1$.

Let $T_1 = \underline{\lim}\, t_{n_1}$. By assumption,

$$M \geqq J_k \geqq x_k^{n+1}(t_{k_1}) \geqq \int_0^{t_{k_1}} Q(x_k(s))\, ds \geqq \eta t_{k_1},$$

where $\eta = \inf_{\| x \| \geqq \epsilon_1} Q(x) > 0$. Thus $M/\eta \geqq t_{k_1}$ for all k and hence $T_1 \leqq M/\eta$.

Consequently we can find a subindex set $\{n_1\} \subset \{n\}$ such that:

(3.2) $\{t_{n_1}\} \to T_1$

(3.3) $\{x_{n_1}(T_1)\} \to x_1 \in \partial(\epsilon_1)$,

(3.4) $\{u_{n_1}\} \to^{w} \bar{u}_1(t)$ on the interval $[0, T_1]$, and $\bar{u}_1 \in U$,

(3.5) $\{x_{n_1}(t)\} \to^{\text{unif}} x_{\bar{u}_1}(t)$ on the interval $[0, T_1]$, where $x_{\bar{u}_1}$ is the solution of (3.1a) with $u = \bar{u}_1$,

(3.6) $x_{\bar{u}_1}(T_1) = x_1$.

Justification. Since $A(t)$, $B(t)$, and $u_n(t)$ are uniformly bounded on $[0, M/\eta]$, the sequence $\{x_n(t)\}$ is equicontinuous; hence we can find a subsequence which satisfies (3.2) and (3.3). Furthermore we can find a subsequence of this subsequence such that $\{u_n\} \to^{w} \bar{u}_1$ in B on the interval $[0, T]$. The mapping $\bar{u}_1$ can be chosen such that $\bar{u}_1(t)$ is in K for each t in $[0, T_1]$. This follows from [7, Exercise 43, p. 439]. Hence (3.4) is true.

Just as in [6] we can show that the solution of (3.1a) with $u = \bar{u}_1$ satisfies (3.5) and hence (3.6).

Let $t_{n_{1,2}}$ be the first time the trajectory $x_{n_1}(t)$ strikes the surface $\partial(\epsilon_2)$ of the unit ball $\| x \| \leqq \epsilon_2$. Let

$$T_2 = \underline{\lim}\, t_{n_{1,2}},$$

which is greater than or equal to T_1 since $\{n_1\} \subset \{n\}$ and $\epsilon_2 < \epsilon_1$. By the same argument used to show the finiteness of T_1 we show that T_2 is finite.

We select a subindex set $\{n_2\} \subset \{n_1\}$ such that (3.2)–(3.6) hold, where T_2 replaces T_1, n_2 replaces n_1, and $\bar{u}_2$ replaces $\bar{u}_1$.

For each ϵ_k we repeat the above process, each time selecting a subindex set $\{n_k\} \subset \{n_{k-1}\}$ and obtaining a T_k which is finite and such that $T_{k-1} \leqq T_k$, $k \geqq 2$.

We apply the Cantor diagonalization process to the index sets $\{n_k\}$ and thus obtain an index set $\{n_\theta\}$ such that:

(1) $\{u_{n_\theta}\} \to^{w} \bar{u}$ on $[0, \infty)$, where $\bar{u}(t) = \bar{u}_k(t)$ for $t \in [0, T_k]$,

(2) the trajectory $x_{\bar{u}}$ corresponding to $\bar{u}$ is such that $\| x_{\bar{u}}(T_k) \| = \epsilon_k$,

(3) $\{x_{n_\theta}(t)\} \to^{\text{unif}} x_{\bar{u}}(t)$ on each interval $[0, T_k]$,

(4) $\bar{u}(t)$ is in K for each t in $[0, \infty)$.

For convenience we relabel n_θ by n in the remainder of the proof.

We claim that $\lim_{t\to\infty} x_{\bar{u}}(t) = 0$ and $J[\bar{u}] = J_0$.

First observe that $x_n(t) \to^{\text{unif}} x(t)$ on any finite interval $[0, T]$. Since Q is continuous this implies that

$$\int_0^t Q(x_n(s))\, ds \to^{\text{unif}} \int_0^t Q(x_{\bar{u}}(s))\, ds$$

on finite intervals.

By property 4 of §2 it follows that

$$\underline{\lim} \int_0^t f(\| u_n(s) \|)\, ds \geqq \int_0^t f(\| u(s) \|)\, ds$$

on finite intervals.

Hence for any $T < \infty$ and $\epsilon > 0$ there is an $n_0(T, \epsilon)$ such that

$$\int_0^T Q(x(s))\, ds + \int_0^T f(\| \bar{u}(s) \|)\, ds$$

$$< \int_0^T Q(x_n(s)\, ds + \int_0^T f(\| u_n(s) \|)\, ds + \epsilon < J_0 + 2\epsilon$$

for all $n \geqq n_0(\epsilon, T)$. Since this is true for all finite T it must be true in the limit, i.e., $J[\bar{u}] \leqq J_0$.

Finally we show that the origin is the only limit point of $x_{\bar{u}}(t)$.

Assume $y_0 \neq 0$ is another limit point of $x_{\bar{u}}(t)$. We construct the shell

$$S = \left\{ x \mid \frac{\| y_0 \|}{4} \leqq \| x \| \leqq \frac{\| y_0 \|}{2} \right\}.$$

If $\{\bar{t}_n\} \to \infty$ and $\{x_{\bar{u}}(\bar{t}_n)\} \to y_0$, then the trajectory $x_{\bar{u}}(t)$ must pass infinitely often through S. By virtue of the boundedness assumptions on $A(t)$, $B(t)$, and $\bar{u}(t)$ it is easy to see that the minimum time of passage through S is bounded away from zero. Let $T_0 > 0$ be a lower bound for the minimum passage time. By the assumptions on Q it follows that there is a $\lambda > 0$ such that $Q(x) \geqq \lambda$ for x in S. Since the passage time for $x_{\bar{u}}(t)$ in S is greater than or equal to T_0, it follows that $J[\bar{u}] = \infty$, which is a contradiction. Hence $x_{\bar{u}}(t)$ has only the origin as a limit.

Remark 1. If in Theorem 2 we replace $f(\| u(t) \|)$ by $f([u^T(t)Lu(t)]^{1/2})$, where ^{T} denotes the transpose of a vector and L is a positive definite symmetric $m \times m$ matrix, then the conclusion of the theorem remains valid. This follows from the fact that $[u^TLu]^{1/2}$ has the property that if $0 < \alpha < 1$ and u_1 and u_2 are given, then $u = (1 - \alpha)u_1 + \alpha u_2$ satisfies

$$[u^TLu]^{1/2} \leqq (1 - \alpha)[u_1{}^TLu_1]^{1/2} + \alpha[u_2{}^TLu_2]^{1/2}.$$

COROLLARY 1. *If* $\dot{x}^{n+1}(t) = x^T(t)Qx(t) + u^T(t)Lu(t)$, *where* Q *and* L *are*

respectively positive definite $n \times n$ and $m \times m$ matrices, then the conclusion of Theorem 2 is valid.

Proof. Let $f(y) = y^2$. By applying Definitions 1–4 we see that f is an N-function whose complimentary function is $g(F) = \frac{1}{4}F^2$ and that both satisfy the Δ_2-condition. By Remark 1, Theorem 2 holds for Q positive definite and $f([u^T(t)Lu(t)]^{1/2}) = u^T(t)Lu(t)$.

Corollary 1 is a result due to Chang [3].

THEOREM 3. *Let $Q(x) = x^TQx$, where Q is a positive definite symmetric matrix. Then the optimal mapping $\bar{u}$ of Theorem 2 is unique up to sets of measure zero.*

Proof. Suppose there is a $\tilde{u}$ in C such that $J(\bar{u}) = J(\tilde{u})$ and $\bar{u}(t) \neq \tilde{u}(t)$ on some set E with Lebesgue measure greater than zero.

Choose $0 < \alpha < 1$ and define the mapping $u_1 = \alpha\bar{u} + (1 - \alpha)\tilde{u}$. Because of the linearity of (3.1a) in x and u it follows that u_1 is in C. Moreover $Q(x)$ is a convex function in x, i.e.,

$$Q(\alpha x_1 + (1 - \alpha)x_2) \leqq \alpha Q(x_1) + (1 - \alpha)Q(x_2) \qquad \text{for } 0 < \alpha < 1.$$

Since f is strictly convex as a function of u and $\bar{u}(t) \neq \tilde{u}(t)$ on a set E with measure greater than zero, it follows that

$$J[u_1] < \alpha J[\bar{u}] + (1 - \alpha)J[\tilde{u}] = J[\bar{u}],$$

which is impossible since $J[\bar{u}]$ is the minimum for all u in C.

Remark 2. Let $f(y) = y^p$, $p > 1$. Then the norm induced by Definition 4 has the property that if $\{u_n\} \to^w u$ and $\{\| u_n \|_B\} \to \| u \|_B$, then $\| u_n - u \|_B \to 0$ as $n \to \infty$, i.e., the convergence is strong.

Proof. The proof of the remark will follow if we can show that the induced norm in the space conjugate to B, i.e., in L_q, is Fréchet differentiable on the unit ball in L_q (see, e.g., [9, §§3, 5, 8, pp. 111–114]).

If $v \neq 0$ is in L_q, then $\| v \|_{L_q}$ satisfies the relationship

$$\int_0^\infty \left[\frac{\| v(\tau) \|}{\| v \|_{L_q}}\right]^q d\tau = 1.$$

Let v and h be mappings in L_q whose norms are equal to one.

Consider the functional

$$\phi(t, k) = \int_0^\infty \frac{\| v(\tau) + th(\tau) \|^q}{k^q} d\tau = 1.$$

For t sufficiently small we can apply Leibniz' rule for differentiation under the integral sign (see, e.g., [10, p. 359]). We obtain

$$\frac{\partial\phi}{\partial k} = \frac{-q}{k},$$

$$\frac{\partial\phi}{\partial t} = \frac{q}{k^q}\int_0^\infty \left[\| v(\tau) + th(\tau) \|^{q-2} \sum_{i=1}^m (v_i(\tau) + th_i(\tau))h_i(\tau)\right] d\tau$$

(where of course the integrand is zero at points where $v(\tau) + th(\tau) = 0$).

By the implicit function theorem we obtain

$$\frac{dk}{dt}(0, k(0)) = \int_0^\infty \left[\| v(\tau) \|^{q-2} \sum_{i=1}^{m} v_i(\tau) h_i(\tau) \right] d\tau,$$

which defines a linear mapping, Γv, from $L_q \to R^1$. This mapping has the representation

$$\Gamma v = \| v(\tau) \|^{q-2} v(\tau)$$

(again the right side is zero if $v(\tau) = 0$) and is called the Gateaux gradient of $\| \cdot \|_{L_q}$ at v.

By direct verification we see that $\| \Gamma v \|^p$ is integrable over R^+ and has norm equal to one. Hence Γv is in B, i.e., in L_p.

Moreover $\Gamma: L_q \to L_p$ is continuous at the point v. To see this, suppose the contrary. Then there exist a sequence $\{v_n\}$ and a constant $\epsilon > 0$ such that $\| v_n - v \|_{L_q} \to 0$, but for all n, $\| \Gamma v - \Gamma v_n \|_B \geqq \epsilon$. However because of the strong convergence of $\{v_n\}$ to v we can select a subsequence $\{v_{n_1}\}$ which converges a.e. to v. From the form of v we see that this implies $\| \Gamma v - \Gamma v_{n_1} \|_B \to 0$ as $n_1 \to \infty$ which is a contradiction. Thus Γ is continuous at v.

We can now apply a result due to Vainberg [11] which states: If the Gateaux gradient of a continuous functional F on a Banach space B exists in a neighborhood of a point v in B and is continuous at v, then it is the Fréchet derivative of F at v.

Thus $\| \cdot \|_{L_q}$ has a Fréchet derivative at every point of the unit ball in L_q, which proves the remark.

THEOREM 4. *Let $f(y) = y^p$, $p > 1$. Then a subsequence u_{n_1} of the original sequence in Theorem 2 tends to $\bar{u}$ in the strong topology, i.e.,*

$$\| u_{n_1} - u \|_B \to 0 \quad as \quad n_1 \to \infty.$$

Proof. Since

$$\int_0^\infty Q(x_{n_1}(s))\, ds \to \int_0^\infty Q(x_{\bar{u}}(s))\, ds$$

for some subsequence, it follows that

$$\int_0^\infty f(\| u_{n_1}(s) \|)\, ds = \int_0^\infty \| u_{n_1}(s) \|^p\, ds \to \int_0^\infty (\| \bar{u}(s) \|^p)\, ds \quad \text{as} \quad n_1 \to \infty.$$

But this implies by Definition 4 that

$$\| u_{n_1} \|_B^p \to \| \bar{u} \|_B^p,$$

which is equivalent to

$$\| u_{n_1} \|_B \to \| \bar{u} \|_B.$$

Hence we can apply Remark 2 to $\{u_{n_1}\}$ and $\bar{u}$ since the norm induced by f is the L^p norm, which has the property of Remark 2.

Appendix 1. Proof of Theorem 1. Let f satisfy Definitions 1–3. Let B_1 denote the family of mappings $\{u\}$ such that $u: R^+ \to R^1$ satisfies

$$\int_0^\infty f(|\, u(t)\,|)\, dt < \infty.$$

In [4] it is shown that B_1 is a separable reflexive Banach space if the norm in B_1 is defined by

$$\|\, u\,\|'_{B_1} = \sup \int_0^\infty |\, u(t)\,|\, v(t)\, dt,$$

where the supremum is taken over all mappings $v: R^+ \to R^1$ such that

$$\int_0^\infty g(|\, v(t)\,|)\, dt \leqq 1.$$

The family B of our theorem is algebraically isomorphic to the direct sum $B_1 \oplus \cdots \oplus B_1$ (taken m times).

From [4, p. 126] we see that the product topology of this direct sum can be given by taking the norm to be

$$\|\, u\,\|'_B = \sup \int_0^\infty \|\, u(t)\,\|\, \|v(t)\,\|\, dt,$$

where the supremum is taken over all mappings $v: R^+ \to R^m$ such that

$$\lambda(v) = \int_0^\infty g(\|\, v(t)\,\|)\, dt \leqq 1.$$

From [8, pp. 1–2] it follows that the conjugate space B^* of B is topologically isomorphic to ${B_1}^* \oplus \cdots \oplus {B_1}^*$ (taken m times) and that the representational form of a continuous linear mapping v from $B \to R^1$ is given by

$$\langle v, u\rangle = \sum_{i=1}^m \int_0^\infty u_i(t) v_i(t)\, dt,$$

where $u_i \in B_1$, $v_i \in {B_1}^*$, for each $i = 1, \cdots, m$.

By [4, Theorem 2, p. 80] we have

$$(1) \qquad \int_0^\infty f\left(\frac{\|\, u(t)\,\|}{\|\, u\,\|'_B}\right) dt \leqq 1.$$

(The proof there is for scalar mappings, but goes through verbatim if we replace the absolute value $|\cdot|$ for scalar functions by the Euclidean length $\|\cdot\|$ for vector valued mappings.)

From Definition 4 it then follows that

$$\| u \|_B \leqq \| u \|'_B \quad \text{for all } u \text{ in } B. \tag{2}$$

By Young's inequality [4, p. 77]

$$\| u(t) \| \| v(t) \| \leqq f(\| u(t) \|) + g(\| v(t) \|). \tag{3}$$

Next we observe that

$$\begin{aligned} \left\| \frac{u}{\| u \|_B} \right\|'_B &\leqq \sup_{\lambda(v) \leqq 1} \int_0^\infty \frac{\| u(t) \| \| v(t) \|}{\| u \|_B} \, dt \\ &\leqq \int_0^\infty f\left(\frac{\| u(t) \|}{\| u \|'_B} \right) + 1, \end{aligned} \tag{4}$$

because of (3).

By (1) we see that the right side of (4) is less than or equal to 2, i.e.,

$$\| u \|'_B \leqq 2 \| u \|_B .$$

Hence $(B, \| \cdot \|'_B)$ and $(B, \| \cdot \|_B)$ are topologically isomorphic under the identity mapping. This proves that $(B, \| \cdot \|_B)$ is a separable reflexive Banach space.

Appendix 2. Proof of property 4. We first establish a lemma needed in the proof of property 4.

LEMMA. *If* $\{u_n\}$ *in* B *is any sequence such that for all* n, $\| u_n(t) \| \leqq M$ *for each* t *in some finite interval* $[0, T]$, *then given any* $\epsilon > 0$ *there is a* $\delta(\epsilon) > 0$ *such that*

$$V_n(\delta) = \left| \int_0^T f\left(\frac{\| u_n(t) \|}{1 - \delta} \right) dt - \int_0^T f(\| u_n(t) \|) \, dt \right| < \epsilon$$

for all n.

Proof. The proof follows from the fact that we can write

$$\begin{aligned} V_n(\delta) &= \int_0^T \left[\int_{\| u_n(t) \|}^{\| u_n(t) \|/(1-\delta)} p(\tau) \, d\tau \right] dt \\ &\leqq \int_0^T \left[\int_M^{M/(1-\delta)} p(\tau) \, d\tau \right] dt = T \int_M^{M/(1-\delta)} p(t) \, dt \\ &= T \left(f\left(\frac{M}{1 - \delta} \right) - f(M) \right). \end{aligned}$$

Proof of property 4. Assume

$$\underline{\lim} \int_0^T f(\| u_n(t) \|) \, dt < \int_0^T f(\| u(t) \|) \, dt = k.$$

Let $\bar{f}(t) = f(t)/k$. Define a new norm $\| \cdot \|'_B$ in B which is determined by $\bar{f}$ and satisfies Definition 4. Because of the Δ_2-condition the norms $\| \cdot \|_B$ and $\| \cdot \|'_B$ are equivalent. In the new norm $\| u \|'_B = 1$ and $\underline{\lim} \| u_n \|'_B \geqq 1$.

Let $k_n = \int_0^T f(\| u_n(t) \|)\, dt$. Then

$$\underline{\lim}\, k_n = k \,\underline{\lim} \int_0^T \bar{f}(\| u_n(t) \|)\, dt < k.$$

Hence, if we let $\bar{k}_n = \int_0^T \bar{f}(\| u_n(t) \|)\, dt$, we see that $\underline{\lim}\, \bar{k}_n = 1 - \alpha$, where α is some positive constant greater than zero and less than 1.

Let $\{n_1\} \subset \{n\}$ be a subindex set such that $\{\bar{k}_{n_1}\} \to 1 - \alpha$ and such that $\bar{k}_{n_1} < 1 - \frac{3}{4}\alpha$ for all n_1. We now apply the above lemma choosing $0 < \delta < 1$ such that

$$\bar{k}_{n_1} \leqq \int_0^T \bar{f}\left(\frac{\| u_{n_1}(t) \|}{1 - \delta}\right) dt < 1 - \frac{\alpha}{2}.$$

Since $\underline{\lim} \| u_n \|'_B \geqq 1$ there exists an $\bar{n}_1$ such that $\| u_{n_1} \|'_B > 1 - \delta$. Hence we obtain the following inequality

$$1 = \int_0^T \bar{f}\left(\frac{\| \bar{u}_{n_1}(t) \|}{\| u_{\bar{n}_1} \|'_B}\right) dt \leqq \int_0^T \bar{f}\left(\frac{\| \bar{u}_{n_1}(t) \|}{1 - \delta}\right) dt < 1 - \frac{\alpha}{2}.$$

This contradiction proves property 4.

REFERENCES

[1] C. D. Johnson and W. M. Wonham, *On a problem of Letov in optimal control*, Proceedings, Joint Automatic Control Conference, Stanford, California, 1964.

[2] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, (2) 5(1960), pp. 102–119.

[3] A. Chang, *An optimal regulator problem*, this Journal, 2(1964), pp. 220–233.

[4] A. C. Zaanen, *Linear Analysis*, North Holland, Amsterdam, The Netherlands, 1956.

[5] M. A. Krasnosel'skii and Ya. B. Rutickii, *Convex Functions and Orlicz Spaces*, Noordhoff, Groningen, The Netherlands, 1961.

[6] E. B. Lee and L. Markus, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8(1961), pp. 36–58.

[7] N. Dunford and J. T. Schwartz, *Linear Operators Part I*, Interscience, New York, 1958.

[8] R. Schatten, *A Theory of Cross-Space*, Princeton University Press, Princeton, 1950.

[9] M. M. Day, *Normed Linear Spaces*, Academic Press, New York, 1962.

[10] E. W. Hobson, *The Theory of Functions of a Real Variable*, Cambridge University Press, Cambridge, 1926.

[11] M. Vainberg, *Concerning differentials and gradients of functionals*, Uspehi Mat. Nauk, 7(1952), pp. 138–146.

J. SIAM CONTROL
Vol. 4, No. 2, 1966
Printed in U.S.A.

ON A SOLUTION OF AN OPTIMIZATION PROBLEM IN LINEAR SYSTEMS WITH QUADRATIC PERFORMANCE INDEX*

YOSHIYUKI SAKAWA†

1. Introduction. We consider a linear control system defined by

$$\frac{dx}{dt} = A(t)x(t) + B(t)u(t), \tag{1}$$

where $x(t)$ is an n-dimensional state vector, $u(t)$ is an r-dimensional control vector, and $A(t)$ and $B(t)$ are $n \times n$ and $n \times r$ matrices which are continuous in the time t. Each component $u_i(t)$ of the control vector is assumed to be constrained as

$$| u_i(t) | \leqq 1, \qquad i = 1, 2, \cdots, r. \tag{2}$$

The control $u(t)$, $0 \leqq t < \infty$, will be called an admissible control if it is measurable and it satisfies the constraints (2).

Optimization of (1), subject to the constraints (2), for a quadratic performance index has been studied by several authors [1]–[4]. Letov [1] discussed the problem using the classical calculus of variations. Wonham, Johnson and Rekasius [2]–[4] used the Hamilton-Jacobi equation for analyzing the problem. Chang [5] showed, under fairly strong conditions, that there exists a unique optimal control for any choice of the initial condition. This paper treats the problem by using a different mathematical procedure from those mentioned above. Since the state variables are expressed, by integrating the linear differential equation (1), in a linear form in the control functions, the quadratic performance index can be expressed as a quadratic functional of the control functions. Thus, we are required to minimize the quadratic functional under the constraints (2). This problem can be considered as an infinite-dimensional nonlinear programming problem. By using the generalized Kuhn-Tucker theorem in nonlinear programming, we derive a system of nonlinear integral equations as a necessary and sufficient condition for the optimal control. The existence and the uniqueness of the solution of the integral equations are studied. Successive approximations for the solution of the integral equations are shown also.

2. Formulation of the optimization problem in the Hilbert space. The solution of (1) with initial value $x(0) = x_0$ is given by

$$x(t) = X(t)x_0 + X(t) \int_0^t X^{-1}(s)B(s)u(s)\, ds, \tag{3}$$

* Received by the editors November 17, 1965.
† Department of Electrical Engineering, Kyoto University, Kyoto, Japan.

where $X(t)$, the fundamental matrix, satisfies

$$\frac{dX(t)}{dt} = A(t)X(t), \qquad X(0) = I \text{ (identity matrix)}. \tag{4}$$

The matrix $X(t)$ is also called the transition matrix. The performance index to be used in this paper is the generalized quadratic error criterion [6]. Let $x_d(t)$ be an n-dimensional desired state vector. Let us also define the error vector to be the difference between the desired state and the actual state, i.e.,

$$e(t) = x_d(t) - x(t).$$

Using (3), we can write

$$e(t) = g(t) - \int_0^t W(t, s)u(s)\, ds, \tag{5}$$

where

$$\begin{aligned} g(t) &= x_d(t) - X(t)x_0\,, \\ W(t, s) &= X(t)X^{-1}(s)B(s). \end{aligned} \tag{6}$$

Clearly, $W(t, s)$ is an $n \times r$ matrix.

The performance index is defined as

$$I(u(t)) = \int_0^T \{e^*(t)Q(t)e(t) + u^*(t)Cu(t)\}\, dt, \tag{7}$$

where $Q(t)$ is an $n \times n$ positive semidefinite symmetric matrix which is continuous in the time t, C is an $r \times r$ positive definite diagonal matrix with positive constant elements, T is a fixed time, and * denotes the transpose of a matrix or a vector. The matrix $Q(t)$ is usually taken to be a diagonal matrix with nonnegative constant elements. The problem is then to choose an appropriate admissible control vector $u(t)$ so that the performance index is minimized.

In this paper, we use notations of functional analysis [6] — [10]. Let H_1 be a real Hilbert space of n-dimensional functions square integrable over $[0, T]$, and H_2 be a real Hilbert space of r-dimensional functions square integrable over $[0, T]$. Then the state vector $x(t)$, $0 \leqq t \leqq T$, will be in H_1 and the control vector $u(t)$ can be taken in H_2. Let us denote the inner product of two n-dimensional vectors x and y in the Hilbert space H_1 by $(x, y)_1$, which is defined by

$$(x, y)_1 = \int_0^T x^*(t)y(t)\, dt = \int_0^T y^*(t)x(t)\, dt.$$

In the same way, let us denote the inner product of two r-dimensional vec-

tors u and v in H_2 by $(u, v)_2$. Then, the performance index (7) can be written as

$$I(u) = (e, Qe)_1 + (u, Cu)_2. \tag{8}$$

We define a linear integral operator L on H_2 by

$$Lu = \int_0^t W(t, s)u(s)\, ds, \qquad 0 \leqq t \leqq T, \tag{9}$$

which maps H_2 into H_1. Since $W(t, s)$ is continuous on the domain $0 \leqq t, s \leqq T$, it is obvious that the linear operator L is bounded and $Lu \in H_1$. The performance index (8) is rewritten as

$$I(u) = (g - Lu, Qg - QLu)_1 + (u, Cu)_2. \tag{10}$$

Equation (10) can be expanded to give

$$I(u) = (g, Qg)_1 - 2(Qg, Lu)_1 + (Lu, QLu)_1 + (u, Cu)_2. \tag{11}$$

Let L^* now be the adjoint operator of L; then L^* maps H_1 into H_2 and satisfies the relation

$$(x, Lu)_1 = (L^*x, u)_2,$$

where $x \in H_1$ and $u \in H_2$. Equation (11) can thus be written as

$$I(u) = (g, Qg)_1 - 2(L^*Qg, u)_2 + (L^*QLu, u)_2 + (Cu, u)_2. \tag{12}$$

It can be proved, as shown in the Appendix, that

$$\begin{aligned} L^*Qg &= \int_s^T W^*(t, s)Q(t)g(t)\, dt, \\ L^*QLu &= \int_0^T Y(s, \tau)u(\tau)\, d\tau, \end{aligned} \tag{13}$$

where

$$Y(s, \tau) = \int_{\max(s,\tau)}^T W^*(t, s)Q(t)W(t, \tau)\, dt. \tag{14}$$

Evidently, $Y(s, \tau)$ is an $r \times r$ continuous matrix and $Y^*(s, \tau) = Y(\tau, s)$. Since

$$(L^*QL)^* = L^*QL,$$

the linear bounded operator L^*QL on H_2 into H_2 is self-adjoint. Moreover, since

$$(L^*QLu, u)_2 = (QLu, Lu)_1 \geqq 0 \tag{15}$$

for arbitrary $u \in H_2$, the operator L^*QL is positive.

Defining such a new operator R by

$$R = L^*QL + C, \tag{16}$$

(12) can be written as

$$I(u) = (Ru, u)_2 - 2(L^*Qg, u)_2 + (g, Qg)_1 . \tag{17}$$

It is clear that the operator R on H_2 into H_2 is self-adjoint and positive definite.

3. Reduction of the optimization problem to a system of integral equations. The constraints (2) can be written as

$$1 - u_i^2(t) \geqq 0, \qquad i = 1, 2, \cdots, r. \tag{18}$$

Thus, the problem is to minimize (17), the quadratic functional of $u(t)$, under the constraints (18). This problem can be considered as an infinite-dimensional nonlinear programming. For this problem, we can apply the generalized Kuhn-Tucker theorem [11] which is an extension of the Kuhn-Tucker theorem on nonlinear programming to more general topological spaces. Defining a mapping ϕ, which maps H_2 into H_2, by

$$\phi(u) = \begin{pmatrix} 1 - u_1^2(t) \\ 1 - u_2^2(t) \\ \vdots \\ 1 - u_r^2(t) \end{pmatrix}, \tag{19}$$

we denote the constraints (18) as

$$\phi(u) \geqq 0. \tag{20}$$

Since the operator R on H_2 is positive definite, it can be easily seen that the functional $I(u)$, as given by (17), is convex. It is clear that $\phi(u)$ defined by (19) is concave. Moreover, it follows that $\phi(0) > 0$, $0 \in H_2$. Therefore, from [11, Theorem V. 3.1], it follows that if u^0 minimizes $I(u)$ subject to $\phi(u) \geqq 0$, then there exists a nonnegative r-dimensional function

$$\lambda^0(t) \geqq 0, \qquad \lambda^0 \in H_2 , \tag{21}$$

such that, for the Lagrangian expression

$$J(u, \lambda) = I(u) - (\lambda, \phi(u))_2 , \tag{22}$$

the saddle-point inequalities

$$J(u, \lambda^0) \geqq J(u^0, \lambda^0) \geqq J(u^0, \lambda) \tag{23}$$

hold for all $u \in H_2$ and all $\lambda \geqq 0$, $\lambda \in H_2$.

Conversely, from [11, Theorem V. 1], it follows that if there exist $u^0 \in H_2$ and $\lambda^0 \geqq 0$, $\lambda^0 \in H_2$, such that the saddle-point inequalities (23) hold for all $u \in H_2$ and all $\lambda \geqq 0$, $\lambda \in H_2$, then

$$\phi(u^0) \geqq 0 \tag{24}$$

and, for all $u \in H_2$ satisfying $\phi(u) \geqq 0$,

$$I(u^0) \leqq I(u). \tag{25}$$

Therefore, the conditions (21) and (23) are necessary and sufficient for u^0 to be an optimal control.

Since $\lambda^0 \geqq 0$ is a fixed vector in H_2, we write

$$J(u, \lambda^0) = J_0(u).$$

Let $\delta J_0(u^0; \xi)$ be the Fréchet differential of J_0 at u^0 with increment ξ, $\xi \in H_2$, which is defined by

$$\delta J_0(u^0; \xi) = \lim_{\epsilon \to 0} \frac{J_0(u^0 + \epsilon\xi) - J_0(u^0)}{\epsilon}, \tag{26}$$

where ϵ is a real number [10]. It can be shown easily that

$$\begin{aligned} J_0(u^0 + \epsilon\xi) - J_0(u^0) &= \epsilon \delta J_0(u^0; \xi) + \epsilon^2 (R\xi, \xi)_2 \\ &\quad + \epsilon^2 \int_0^T \sum_{i=1}^{r} \lambda_i^0(t) \xi_i^2(t)\, dt, \end{aligned} \tag{27}$$

where λ_i^0 and ξ_i, $i = 1, \cdots, r$, are the components of the r-dimensional functions λ^0 and ξ, respectively. Therefore, in order that the first inequality of (23), $J(u, \lambda^0) \geqq J(u^0, \lambda^0)$, be satisfied for all $u \in H_2$, it is necessary and sufficient that

$$\delta J_0(u^0; \xi) = 0 \tag{28}$$

for arbitrary $\xi \in H_2$.

Moreover, the second inequality of (23) implies that

$$(\lambda^0, \phi(u^0))_2 \leqq (\lambda, \phi(u^0))_2$$

for all $\lambda \geqq 0$, $\lambda \in H_2$. Hence, we obtain the inequality $\phi(u^0) \geqq 0$ and

$$(\lambda^0, \phi(u^0))_2 = 0. \tag{29}$$

Thus, the necessary and sufficient conditions for u^0 to be the optimal control are (21), (24), (28), and (29) in all. Henceforth, u^0 and λ^0 are simply written as u and λ, since no confusion may occur.

In view of the definition (26), the Fréchet differential of J_0 at u with in-

crement ξ ($u, \xi \in H_2$) can be evaluated as

$$\delta J_0(u; \xi) = 2(Ru, \xi)_2 - 2(L^*Qg, \xi)_2 - \left(\frac{\partial \phi}{\partial u}\lambda, \xi\right)_2,$$

where $\partial\phi/\partial u$ denotes an $r \times r$ diagonal matrix defined by

$$\frac{\partial \phi}{\partial u} = \left[\frac{\partial \phi_i}{\partial u_j}\right] = -2\begin{pmatrix} u_1 & 0 & \cdots & 0 \\ 0 & u_2 & \cdots & 0 \\ & & \cdot & \\ & & \cdot & \\ & & \cdot & \\ 0 & 0 & \cdots & u_r \end{pmatrix}.$$

Since $\delta J_0(u; \xi)$ vanishes for arbitrary $\xi \in H_2$, it follows that

$$Ru - L^*Qg - \frac{1}{2}\frac{\partial \phi}{\partial u}\lambda = 0. \tag{30}$$

We set $L^*Qg = f$, then from (13) and (16),

$$Ru = \int_0^T Y(s, \tau)u(\tau)\, d\tau + Cu(s),$$

$$L^*Qg = \int_s^T W^*(\tau, s)Q(\tau)g(\tau)\, d\tau = f(s).$$

Clearly, $f(s)$ is an r-dimensional function. Write

$$Y(s, \tau) = \begin{pmatrix} y_{11}(s, \tau) & y_{12}(s, \tau) & \cdots & y_{1r}(s, \tau) \\ & & \cdot & \\ & & \cdot & \\ & & \cdot & \\ y_{r1}(s, \tau) & y_{r2}(s, \tau) & \cdots & y_{rr}(s, \tau) \end{pmatrix}.$$

Then, the relations (21), (29), and (30) can be written for each component as

$$\lambda_i(t) \geqq 0, \qquad i = 1, \cdots, r, \tag{31}$$

$$\lambda_i(t)\{1 - u_i^2(t)\} = 0, \qquad i = 1, \cdots, r, \tag{32}$$

$$\sum_{j=1}^{r} \int_0^T y_{ij}(t, s)u_j(s)\, ds + c_i u_i(t) + \lambda_i(t)u_i(t) = f_i(t), \qquad i = 1, \cdots, r, \tag{33}$$

where the c_i are the elements of the diagonal matrix C and all positive.

From (31) and (32), it follows that

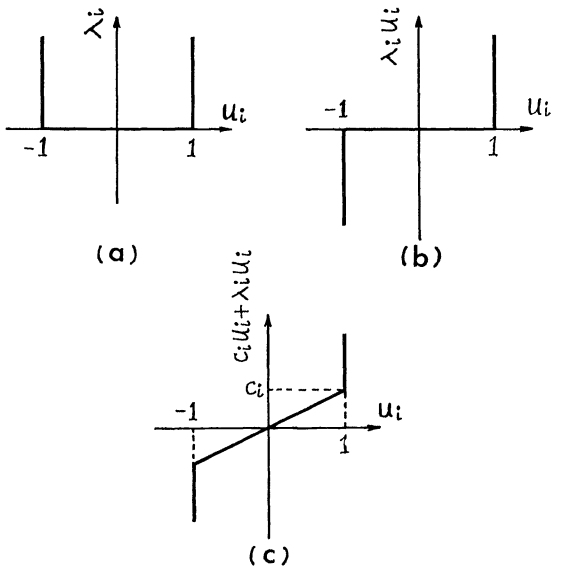

FIG. 1. *Relations between the variables*

$$\lambda_i(t) = 0 \quad \text{if} \quad -1 < u_i(t) < 1,$$

$$\lambda_i(t) \geqq 0 \quad \text{if} \quad u_i(t) = \pm 1.$$

Hence, the relation between $u_i(t)$ and $\lambda_i(t)$ can be shown as Fig. 1a. The relation between $u_i(t)$ and $\lambda_i(t)u_i(t)$ and then the relation between $u_i(t)$ and $c_iu_i(t) + \lambda_i(t)u_i(t)$ can also be obtained successively from Fig. 1a as shown in Figs. 1b and 1c, respectively. By defining the new functions

$$v_i(t) = c_iu_i(t) + \lambda_i(t)u_i(t), \qquad i = 1, \cdots, r,$$

and denoting the relation between $u_i(t)$ and $v_i(t)$ by

$$u_i(t) = \Phi_i(v_i(t)), \qquad i = 1, \cdots, r,$$

(33) can be expressed as

$$v_i(t) + \sum_{j=1}^{r} \int_0^T y_{ij}(t, s)\Phi_j(v_j(s))\, ds = f_i(t), \quad i = 1, 2, \cdots, r, \tag{34}$$

or in vector form, as

$$v(t) + \int_0^T Y(t, s)\Phi(v(s))\, ds = f(t). \tag{35}$$

In (34), the nonlinear function $\Phi_i(v_i)$ is shown in Fig. 2, which can be obtained from Fig. 1c directly. Thus, the optimization problem has been

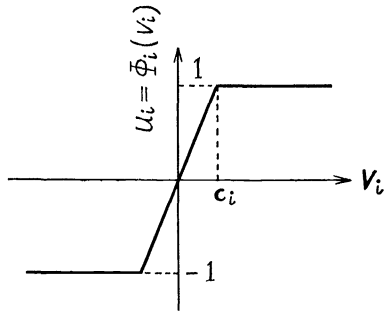

FIG. 2. *Nonlinear characteristic*

reduced to a system of nonlinear integral equations. In other words, (34) is the necessary and sufficient condition for the optimum.

Defining such functions as

$$v_i(t) - f_i(t) = \psi_i(t), \qquad \Phi_i(f_i(t) + \psi_i(t)) = F_i(t, \psi_i(t)), \quad i = 1, \cdots, r,$$

(34) can be written as

$$\psi_i(t) + \sum_{j=1}^{r} \int_0^T y_{ij}(t, s) F_j(s, \psi_j(s))\, ds = 0, \quad i = 1, \cdots, r, \tag{36}$$

or in vector form, as

$$\psi(t) + \int_0^T Y(t, s) F(s, \psi(s))\, ds = 0. \tag{37}$$

Equation (37) is of the vector form of nonlinear integral equations of the Hammerstein type [12], [13].

4. Successive approximations for the solution of the integral equations. Since $c_i > 0$, $i = 1, \cdots, r$, it is clear from Fig. 2 that the functions $F_i(t, \psi_i)$, $i = 1, \cdots, r$, satisfy uniformly a Lipschitz condition of the form

$$| F_i(t, \psi_i^{(1)}) - F_i(t, \psi_i^{(2)}) | \leqq \alpha | \psi_i^{(1)} - \psi_i^{(2)} |, \quad i = 1, \cdots, r, \tag{38}$$

where α is a positive constant such that $\alpha \geqq 1/c_i$, $i = 1, \cdots, r$. Let us define the norm of a vector x in H_2 as

$$\| x \| = (x, x)_2^{1/2} = \left\{ \sum_{i=1}^{r} \int_0^T x_i^2(t)\, dt \right\}^{1/2}.$$

Moreover, let us introduce an r-dimensional function $z(t) = (z_1(t), \cdots, z_r(t))$, where the elements are defined by

$$z_i(t) = \left\{ \sum_{j=1}^{r} \int_0^T y_{ij}^2(t, s)\, ds \right\}^{1/2}, \qquad i = 1, \cdots, r.$$

The functions $y_{ij}(t, s)$, $i, j = 1, \cdots, r$, are continuous on the domain $0 \leqq t,s \leqq T$, hence $z \in H_2$. It can be proved that if

$$\alpha \| z \| < 1, \tag{39}$$

then the successive approximations

$$\psi_i^{(n+1)}(t) = -\sum_{j=1}^{r} \int_0^T y_{ij}(t, s) F_j(s, \psi_j^{(n)}(s))\, ds, \qquad i = 1, \cdots, r, \quad n = 0, 1, 2, \cdots, \tag{40}$$

starting, for instance, with $\psi_i^{(0)}(t) = 0$, converge to a unique solution of (36). It is obvious that the existence of a unique solution of (36) implies the existence of the unique optimal control.

In fact, from (38) it follows that

$$\begin{aligned} &| \psi_i^{(n+1)}(t) - \psi_i^{(n)}(t) | \\ &\qquad \leqq \sum_{j=1}^{r} \int_0^T | y_{ij}(t, s) | \, | F_j(s, \psi_j^{(n)}(s)) - F_j(s, \psi_j^{(n-1)}(s)) | \, ds \\ &\qquad \leqq \alpha \sum_{j=1}^{r} \int_0^T | y_{ij}(t, s) | \, | \psi_j^{(n)}(s) - \psi_j^{(n-1)}(s) | \, ds. \end{aligned}$$

Furthermore, using the Schwarz inequality,

$$\begin{aligned} &| \psi_i^{(n+1)}(t) - \psi_i^{(n)}(t) | \\ &\qquad \leqq \alpha \left\{ \sum_{j=1}^{r} \int_0^T y_{ij}^2(t, s)\, ds \right\}^{1/2} \left\{ \sum_{j=1}^{r} \int_0^T (\psi_j^{(n)}(s) - \psi_j^{(n-1)}(s))^2\, ds \right\}^{1/2} \\ &\qquad = \alpha z_i(t) \, \| \psi^{(n)} - \psi^{(n-1)} \| . \end{aligned}$$

Thus, we obtain

$$\| \psi^{(n+1)} - \psi^{(n)} \| \leqq \alpha \| z \| \| \psi^{(n)} - \psi^{(n-1)} \|. \tag{41}$$

Equation (41) shows that the mapping defined by the right-hand side of (40) is a contraction mapping under the condition (39) [9]. Therefore, under the condition (39), we can show the existence and the uniqueness of the solution of (36).

5. Existence and uniqueness of optimal control. In the case where $c_1 = c_2 = \cdots = c_r$, the nonlinear characteristics Φ_i, $i = 1, \cdots, r$, shown in Fig. 2 coincide with each other. Hence, we express the characteristic as $\hat{\Phi}$. In this case, the system of nonlinear integral equations (34) can be reduced to a single integral equation with a discontinuous kernel in the basic interval $0 \leqq t \leqq rT$:

$$\hat{v}(t) + \int_0^{rT} \hat{y}(t, s) \hat{\Phi}(\hat{v}(s))\, ds = \hat{f}(t), \qquad 0 \leqq t \leqq rT, \tag{42}$$

where

$$\hat{v}(t) = \begin{cases} v_1(t) & \text{if } 0 \leqq t < T, \\ v_2(t - T) & \text{if } T \leqq t < 2T, \\ \vdots & \\ v_r(t - (r-1)T) & \text{if } (r-1)T \leqq t \leqq rT, \end{cases}$$

$$\hat{f}(t) = \begin{cases} f_1(t) & \text{if } 0 \leqq t < T, \\ f_2(t - T) & \text{if } T \leqq t < 2T, \\ \vdots & \\ f_r(t - (r-1)T) & \text{if } (r-1)T \leqq t \leqq rT, \end{cases}$$

and

$$\hat{y}(t, s) = y_{ij}(t - (i-1)T, s - (j-1)T), \tag{43}$$

if $(i-1)T \leqq t < iT$ and $(j-1)T \leqq s < jT$, for $i, j = 1, 2, \cdots, r$. Furthermore, defining such scalar functions as

$$\begin{aligned} \hat{v}(t) - \hat{f}(t) &= \hat{\psi}(t), \\ \hat{\Phi}(\hat{f}(t) + \hat{\psi}(t)) &= \hat{F}(t, \hat{\psi}(t)), \end{aligned} \tag{44}$$

(42) can be written as

$$\hat{\psi}(t) + \int_0^{rT} \hat{y}(t, s)\hat{F}(s, \hat{\psi}(s))\, ds = 0, \qquad 0 \leqq t \leqq rT. \tag{45}$$

Equation (45) is of the standard form of integral equations of the Hammerstein type [12], [13].

Hammerstein [12] proved the existence of the solution of the integral equation of the Hammerstein type, assuming that the iterated kernel

$$\hat{y}_2(t, s) = \int_0^{rT} \hat{y}(t, \tau)\hat{y}(\tau, s)\, d\tau$$

is continuous. However, the kernel function $\hat{y}(t, s)$ defined by (43) is not continuous, hence Hammerstein's existence theorem is not applicable to our problem. In what follows, the existence of the solution of (45) will be shown by using Krasnosel'skii's existence theorem [13], [14].

Let H be a real Hilbert space of functions square integrable over $[0, rT]$. The inner product is defined, as usual, by

$$(x, y) = \int_0^{rT} x(t)y(t)\, dt, \qquad x, y \in H.$$

Let G be an operator on H defined by

$$Gx = \hat{F}(t, x(t)), \qquad x \in H, \tag{46}$$

and K be a linear operator on H defined by

$$Kx = \int_0^{rT} \hat{y}(t, s)x(s)\, ds, \qquad x \in H. \tag{47}$$

Then, the integral equation (45) can be written symbolically as

$$\hat{\psi} + KG\hat{\psi} = 0. \tag{48}$$

Since the matrix $Y(s, \tau)$ defined by (14) is continuous on the closed square domain $0 \leqq s, \tau \leqq T$, it follows that

$$\int_0^{rT}\int_0^{rT} \hat{y}^2(t, s)\, dt\, ds = \sum_{i,j=1}^{r} \int_0^T \int_0^T y_{ij}^2(t, s)\, dt\, ds < \infty.$$

Therefore, the linear operator K is completely continuous [10]. From (15) it follows that

$$\begin{aligned}(Kx, x) &= \int_0^{rT}\int_0^{rT} \hat{y}(t, s)x(t)x(s)\, dt\, ds \\ &= \sum_{i,j=1}^{r} \int_0^T \int_0^T y_{ij}(t, s)x(t - (i-1)T)x(s - (j-1)T)\, dt\, ds \geqq 0,\end{aligned}$$

for an arbitrary function $x \in H$. Hence, the operator K is positive, i.e., all its eigenvalues are positive. Moreover, the operator K is self-adjoint, i.e., $K^* = K$. Therefore, from the spectral theory of operators, the operator K can be decomposed as

$$K = PP^*, \tag{49}$$

where P is a square root of the operator K (i.e., $P = K^{1/2}$) and is a positive self-adjoint completely continuous operator on H into H, and P^* is an adjoint operator of P [14]. Then, the nonlinear integral equation (48) can be written as

$$\hat{\psi} + PP^*G\hat{\psi} = 0. \tag{50}$$

Equation (50) is equivalent to

$$\phi + P^*GP\phi = 0, \qquad \phi \in H, \tag{51}$$

in the sense that to a solution $\phi \in H$ of (51) there corresponds a solution $P\phi \in H$ of (50) and, conversely, to a solution $\hat{\psi} \in H$ of (50) there corresponds a solution $P^*G\hat{\psi} \in H$ of (51). Moreover, Krasnosel'skii [14] shows that the operator $I + P^*GP$, I being an identity operator on H, is a gradient of the functional

$$\Psi(\phi) = \frac{1}{2}(\phi, \phi) + \int_0^{rT} dt \int_0^{P\phi(t)} \hat{F}(t, x)\, dx \tag{52}$$

defined on H, where an operator Γ on H into H is called the gradient of the functional Ψ, if

$$\lim_{\epsilon \to 0} \frac{\Psi(\phi + \epsilon\xi) - \Psi(\phi)}{\epsilon} = (\Gamma\phi, \xi), \qquad \phi, \xi \in H.$$

It is clear that the function $\hat{F}(t, x)$ satisfies the Carathéodory condition [14], i.e., it is continuous with respect to x for almost all $t \in [0, rT]$ and measurable with respect to t for all values of x.

According to [14, Chap. VI, Theorem 1.1], if the functional (52) is increasing, i.e.,

$$\lim_{\|\phi\| \to \infty} \Psi(\phi) = +\infty,$$

then there exists a point ϕ_0 in the Hilbert space H where the functional $\Psi(\phi)$ takes on its minimum value and its gradient vanishes, i.e.,

$$\phi_0 + P^*GP\phi_0 = 0.$$

Thus, if the functional $\Psi(\phi)$ is increasing, then the existence of a solution of (51), and hence the existence of a solution of the fundamental equation (50), can be concluded. Since

$$\int_0^u \hat{F}(t, x)\, dx \leqq \int_0^{|u|} |\, \hat{F}(t, x)\,|\, dx \leqq |\, u\,|,$$

using the Schwarz inequality, it follows that

$$-\int_0^{rT} dt \int_0^{P\phi(t)} \hat{F}(t, x)\, dx \leqq \int_0^{rT} |\, P\phi(t)\,|\, dt \leqq \sqrt{rT}(P\phi, P\phi)^{1/2}$$
$$= \sqrt{rT}(K\phi, \phi)^{1/2} \leqq \sqrt{rT}M(\phi, \phi)^{1/2},$$

where

$$M = \left\{\int_0^{rT} \int_0^{rT} \hat{y}^2(t, s)\, dt\, ds\right\}^{1/2}.$$

Consequently,

$$\Psi(\phi) \geqq \tfrac{1}{2}(\phi, \phi) - \sqrt{rT}M(\phi, \phi)^{1/2}. \tag{53}$$

Equation (53) shows that the functional $\Psi(\phi)$ is increasing. Thus, the existence of the solution of the nonlinear integral equation (45), and hence the existence of the optimal control, has been proved.

If we assume further that the positive operator K defined by (47) is positive definite, i.e.,

$$(\phi, K\phi) > 0 \quad \text{if} \quad \phi \neq 0,$$

then the uniqueness of the solution of (48) can also be proved as follows. Assume that

$$\hat{\psi}_1 + KG\hat{\psi}_1 = 0, \qquad \hat{\psi}_2 + KG\hat{\psi}_2 = 0. \tag{54}$$

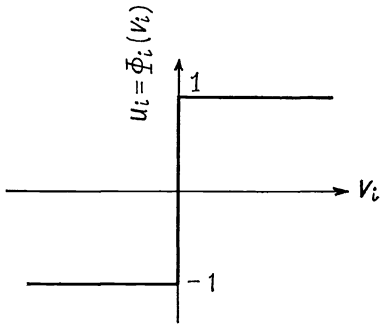

FIG. 3. *Discontinuous nonlinear characteristic*

Subtract these two equations and form the inner product with $(G\hat{\psi}_1 - G\hat{\psi}_2)$:

$$(55)\quad (G\hat{\psi}_1 - G\hat{\psi}_2\,,\, \hat{\psi}_1 - \hat{\psi}_2) + (G\hat{\psi}_1 - G\hat{\psi}_2\,,\, KG\hat{\psi}_1 - KG\hat{\psi}_2) = 0.$$

From the definition of the function $\hat{F}(t, x)$, the first term of (55) is obviously nonnegative. Since the operator K is positive definite, there is a contradiction unless $G\hat{\psi}_1 - G\hat{\psi}_2 = 0$, but in this case from (54) we obtain $\hat{\psi}_1 = \hat{\psi}_2$. Thus, under the assumption that K is positive definite, the uniqueness of the optimal control can be shown.

When $C = 0$ in (7), the nonlinear functions Φ_i, $i = 1, 2, \cdots, r$, become discontinuous as shown in Fig. 3. In Fig. 3 the vertical part of the characteristic corresponds to a singular control which takes on such continuous values as $-1 < u_i(t) < 1$. In this case, however, the existence of the solution can not be claimed since the assumed Carathéodory condition does not hold.

Acknowledgment. The author wishes to express his gratitude to Professor C. Hayashi for his valuable suggestions.

Appendix. Derivation of (13). From the definition (9) of the operator L, it follows that

$$(56)\quad \begin{aligned}(L^*Qg, u)_2 &= (Qg, Lu)_1 \\ &= \int_0^T dt\, g^*(t)Q(t) \int_0^t W(t, s)u(s)\, ds.\end{aligned}$$

The region of integration in (56) is given by $0 \leqq t \leqq T$, $0 \leqq s \leqq t$, which is equivalent to $0 \leqq s \leqq T$, $s \leqq t \leqq T$. Then, changing the order of the integration in (56) yields

$$(57)\qquad (L^*Qg, u)_2 = \int_0^T \int_s^T \{W^*(t,s)Q(t)g(t)\}^*\, dtu(s)\, ds.$$

Equation (57) shows that

$$L^*Qg = \int_s^T W^*(t, s)Q(t)g(t)\, dt. \tag{58}$$

From (58) it follows that

$$L^*QLu = \int_s^T dt\, W^*(t,s)Q(t) \int_0^t W(t, \tau)u(\tau)\, d\tau. \tag{59}$$

The region of integration in (59) is given by $s \leqq t \leqq T, 0 \leqq \tau \leqq t$, which is equivalent to $0 \leqq \tau \leqq T$, $\max(s, \tau) \leqq t \leqq T$. Changing the order of integration in (59) yields

$$\begin{aligned} L^*QLu &= \int_0^T \left\{ \int_{\max(s,\tau)}^T W^*(t, s)Q(t)W(t, \tau)\, dt \right\} u(\tau)\, d\tau \\ &= \int_0^T Y(s, \tau)u(\tau)\, d\tau. \end{aligned} \tag{60}$$

REFERENCES

[1] A. M. Letov, *Analytical design of controllers*, Avtomat. i Telemeh., 21 (1960), pp. 561–568.

[2] W. M. Wonham and C. D. Johnson, *Optimal bang-bang control with quadratic performance index*, Trans. ASME Ser. D. J. Basic Engrg., 86 (1964), pp. 107–115.

[3] C. D. Johnson and W. M. Wonham, *On a problem of Letov in optimal control*, Ibid., 87 (1965), pp. 81–89.

[4] Z. V. Rekasius and T. C. Hsia, *On an inverse problem in optimal control*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 370–375.

[5] A. Chang, *An optimal regulator problem*, this Journal, 2 (1965), pp. 220–233.

[6] H. C. Hsieh, *Synthesis of optimum multivariable control systems by the method of steepest descent*, IEEE Trans. Applications and Industry, 82 (1963), pp. 125–130.

[7] A. V. Balakrishnan, *An operator theoretic formulation of a class of control problems and a steepest descent method of solution*, this Journal, 1 (1963), pp. 109–127.

[8] A. V. Balakrishnan and H. C. Hsieh, *Function space methods in control systems optimization*, presented at the Optimum System Synthesis Conference, Dayton, 1962.

[9] A. N. Kolmogorov and S. V. Formin, *Elements of the Theory of Functions and Functional Analysis*, Graylock Press, Rochester, New York, 1957.

[10] L. V. Kantrovich and G. P. Akirov, *Functional Analysis in Normed Space*, Pergamon Press, London, 1964.

[11] L. Hurwicz, *Programming in linear spaces*, Studies in Linear and Nonlinear Programming, K. J. Arrow, L. Hurwicz and H. Uzawa, eds., Stanford University Press, Stanford, 1958, pp. 38–102.

[12] F. G. Tricomi, *Integral Equations*, Interscience, New York, 1957.

[13] C. L. Dolph and G. J. Minty, *On nonlinear integral equations of the Hammerstein type*, Nonlinear Integral Equations, P. M. Anselone, ed., University of Wisconsin Press, Madison, 1964, pp. 99–154.

[14] M. A. Krasnosel'skii, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, London, 1964.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

NEW RESULTS IN ASYMPTOTIC CONTROL THEORY*

R. S. BUCY†

In a recent paper [1], problems of the asymptotic behavior of optimal control laws were formulated and a class of one-dimensional systems were solved. The solution depended on the explicit construction of a smooth solution to the relevant Hamilton-Jacobi equation for the control problem. Of course this technique for the general n-dimensional problem is ineffective.

Here we consider the problem for n-dimensional systems and resolve it by functional analytic techniques. Our major technical device is to consider the control problem as that of finding the infimum of a functional f on an appropriate Hilbert space of controls. Our existence and uniqueness result for the optimal control, the proof of which is motivated by a rather famous result of Riesz (see [4, p. 25]), has the following corollary: if $f(u_n) \to \inf_{u \in L^2} f(u)$, then u_n tends strongly to $\bar{u}$, the optimal control. This corollary is used to resolve the general problem.

1. Definitions and assumptions. We consider the following linear autonomous n-vector differential equation:

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= F\mathbf{x} + G\mathbf{u}, \\ \mathbf{x}(0) &= \mathbf{c}, \end{aligned} \tag{1.1}$$

where $\mathbf{x}$ and $\mathbf{u}$ are respectively n and m vectors and F and G are respectively $n \times n$ and $n \times m$ matrices. It is well-known that (1.1) has a unique absolutely continuous solution, which is almost everywhere differentiable for every $\mathbf{u}(\cdot)$ in the Hilbert space L_T^2 whose norm is defined by

$$| \mathbf{u} |_T = \left[\int_0^T \| \mathbf{u}(s) \|_R^2 \, ds \right]^{1/2},$$

where R is a positive definite $m \times m$ matrix. We shall denote that solution by $\varphi_{\mathbf{u}}(\cdot, \mathbf{c})$. Now we shall suppose that a function K from R^n to R^+ is given satisfying the assumptions:

(1) $$K(\mathbf{0}) = 0;$$

* Received by the editors September 1, 1965, and in revised form February 28, 1966.

† Department of Aerospace Engineering Sciences, University of Colorado, Boulder, Colorado, and consultant to the RAND Corporation, Santa Monica, California. This research was partially supported by the United States Air Force under Project Rand Contract No. AF 49 (638)-700.

(2) $K \in C^2(R^n)$, and $[K_{x_j x_i}]_{ij}$, $i, j = 1, \cdots, n$, is

positive semidefinite for all $\mathbf{x}$.

We shall associate the following real number with each $\mathbf{c} \in R^n$ and $\mathbf{u}(\cdot) \in L_T^2$:

$$V(T, \mathbf{c}, \mathbf{u}(\cdot)) = \int_0^T (K(\varphi_{\mathbf{u}}(s, \mathbf{c})) + \|\mathbf{u}(s)\|_R^2)\, ds = \beta(\varphi_{\mathbf{u}}(\cdot, \mathbf{c})) + |\mathbf{u}|_T^2. \tag{1.2}$$

Now the control problem consists of the study of the functions

$$V(T, \mathbf{c}) = \inf_{u(\cdot) \in L_T^2} V(T, \mathbf{c}, \mathbf{u}(\cdot)) \tag{1.3}$$

and

$$V^*(\mathbf{c}) = \inf_{\mathbf{u}(\cdot) \in L_\infty^2} V(\infty, \mathbf{c}, \mathbf{u}(\cdot)). \tag{1.4}$$

A function $\mathbf{u}(\cdot) \in L_T^2$ will be called a *control*, and one which achieves the infimum in (1.3) or (1.4) an *optimal control*. Our object here is to resolve the fundamental questions of asymptotic control theory: namely, to prove the existence and uniqueness of the optimal control $u_{T,\mathbf{c}}$ for T fixed, and to prove that

$$V(T, \mathbf{c}) = V(T, \mathbf{c}, \mathbf{u}_{T,\mathbf{c}}(\cdot)) \to V^*(\mathbf{c}) = V(\infty, \mathbf{c}, \mathbf{u}_{\infty,\mathbf{c}}(\cdot))$$

and

$$\mathbf{u}_{T,\mathbf{c}} \to \mathbf{u}_{\infty,\mathbf{c}} \text{ as } T \to \infty.$$

Historically the above problems were resolved in the special case of the filtering problem and the quadratic control problem in [3].

2. Main results. We shall first resolve the existence and uniqueness question.

LEMMA 2.1. *For each fixed T, $0 < T \leqq \infty$, $V(T, \cdot)$ is continuous and there exists an optimal control $\mathbf{u}_{T,\mathbf{c}}(\cdot) \in L_T^2$. Further, if T is finite, the optimal control is unique*[1]; *while if $T = \infty$, $\mathbf{u}_{\infty,\mathbf{c}}(\cdot)$ is unique whenever there exists a control $\mathbf{u}(\cdot) \in L_\infty^2$ such that $V(\infty, \mathbf{c}, \mathbf{u}(\cdot)) < \infty$.*

Proof. For $\mathbf{c}_i \in R^n$ and $\mathbf{u}_i \in L_T^2$ and $0 < \alpha < 1$,

$$\begin{aligned} V(T, \alpha\mathbf{c}_1 + (1 - \alpha)\mathbf{c}_2) \leqq \alpha V(T, \mathbf{c}_1, \mathbf{u}_1(\cdot)) \\ + (1 - \alpha) V(T, \mathbf{c}_2, \mathbf{u}_2(\cdot)), \end{aligned} \tag{2.1}$$

since (1.1) is linear and K is convex. From (2.1) it follows that $V(T, \cdot)$ is convex on an open set and, since by definition it is measurable, $V(T, \cdot)$ is continuous. In order to show the existence of an optimal control, we may

[1] Of course by definition of a control it is an equivalence class of a.e. equal functions.

assume that $V(T, \mathbf{c})$ is finite, for otherwise the conclusion is immediate. By the definition of the infimum it follows that there exists a sequence of controls $\mathbf{u}_n(\cdot)$ such that

$$V(T, \mathbf{c}, \mathbf{u}_n(\cdot)) \to V(T, \mathbf{c}) \quad \text{as} \quad n \to \infty.$$

Now, for ϵ an arbitrary positive real number there exists an n_0 such that for $n > n_0$ and all m,

$$V(T, \mathbf{c}, \mathbf{u}_n(\cdot)) < V(T, \mathbf{c}) + \epsilon$$

and

$$V(T, \mathbf{c}, \mathbf{u}_{n+m}(\cdot)) < V(T, \mathbf{c}) + \epsilon,$$

so that

$$(2.2)\quad \tfrac{1}{2} V(T, \mathbf{c}, \mathbf{u}_n(\cdot)) + \tfrac{1}{2} V(T, \mathbf{c}, \mathbf{u}_{n+m}(\cdot)) - V\left(T, \mathbf{c}, \frac{\mathbf{u}_n(\cdot) + \mathbf{u}_{n+m}(\cdot)}{2}\right) < \epsilon,$$

since L_T^2 is a linear space. Now the Hilbert space identity

$$\left| \frac{\mathbf{u}_n + \mathbf{u}_{n+m}}{2} \right|_T^2 + \left| \frac{\mathbf{u}_n - \mathbf{u}_{n+m}}{2} \right|_T^2 = \frac{1}{2} \left(|\mathbf{u}_n|_T^2 + |\mathbf{u}_{n+m}|_T^2 \right)$$

and (2.2) have the consequence that

$$\left| \frac{\mathbf{u}_n - \mathbf{u}_{n+m}}{2} \right|^2 + \tfrac{1}{2} \beta(\varphi_{\mathbf{u}_n}(\cdot, \mathbf{c})) + \tfrac{1}{2} \beta(\varphi_{\mathbf{u}_{n+m}}(\cdot, \mathbf{c})) - \beta\left(\tfrac{1}{2} \varphi_{\mathbf{u}_n + \mathbf{u}_{n+m}}(\cdot, \mathbf{c})\right) < \epsilon.$$

However, since $\beta(\varphi_{\mathbf{u}}(\cdot, \mathbf{c}))$ is a convex functional of $\mathbf{u}(\cdot)$, it follows that $\{\mathbf{u}_n(\cdot)\}$ is a Cauchy sequence as ϵ was arbitrary; in consequence, there exists a control $\mathbf{u}_{T,\mathbf{c}} \in L_T^2$ such that $\mathbf{u}_n$ converges in norm to $\mathbf{u}_{T,\mathbf{c}}$. By Fatou's theorem,

$$V(T, \mathbf{c}, \mathbf{u}_{T,\mathbf{c}}) \leqq \liminf_n V(T, \mathbf{c}, \mathbf{u}_n(\cdot)) = V(T, \mathbf{c}),$$

so that $\mathbf{u}_{T,\mathbf{c}}$ is optimal. Suppose that $\mathbf{u}_{T,\mathbf{c}}$ and $\mathbf{v}$ are both optimal controls, then by our previous argument the sequence $\{\mathbf{v}_n\}$, where

$$\mathbf{v}_n = \begin{cases} \mathbf{v} & \text{if } n \text{ is even,} \\ \mathbf{u}_{T,\mathbf{c}} & \text{if } n \text{ is odd,} \end{cases}$$

is a Cauchy sequence; and hence $\mathbf{v} = \mathbf{u}_{T,\mathbf{c}}$ a.e. The argument is similar for $T = \infty$ and is omitted.

In actual fact, the preceding proof reveals the following interesting and useful result, which we emphasize as a corollary.

COROLLARY 2.1. *If* $\mathbf{u}_n \in L_T^2$ *and* $V(T, \mathbf{c}, \mathbf{u}_n(\cdot)) \to V(T, \mathbf{c})$ *as* $n \to \infty$, *then* $\mathbf{u}_n$ *converges to* $\mathbf{u}_{T,\mathbf{c}}$ *in norm.*

We may remark that Bellman [2, p. 46] proved the existence and uniqueness of the optimal control for the special case where K is a quadratic form. Using the Alaoglu-Bourbaki theorem for a reflexive space and the Banach result that every weakly convergent sequence possesses a norm convergent sequence in its convex hull, a proof of existence and uniqueness of optimal controls was given in [5] in a slightly more general context than the above lemma, as a reviewer pointed out to the author. The reader should note that this latter proof is not sufficient to establish Corollary 2.1, which, as we shall see, is the key to our later results.

We shall now assume that $V^*(\mathbf{c})$ is finite, for which stability of F or complete controllability of (1.1) is a sufficient condition. The next result gives a priori bounds on $V^*(\mathbf{c})$ and is partly contained in [1].

LEMMA 2.2. *For* $T > 0$,

$$V^*(\mathbf{y}_0) + V(T, \mathbf{c}) \leqq V^*(\mathbf{c}) \leqq V(T, \mathbf{c}) + \min [V^*(\mathbf{y}_1), V^*(\bar{\mathbf{y}}_1)], \tag{2.3}$$

where $\mathbf{y}_0 = \varphi_{\mathbf{u}_{\infty,\mathbf{c}}}(T, \mathbf{c})$, $\mathbf{y}_1 = \varphi_{\mathbf{u}_{T,\mathbf{c}}}(T, \mathbf{c})$, $\bar{\mathbf{y}}_1 = \varphi_{\mathbf{u}_{T,\mathbf{c}}}(T - \epsilon, \mathbf{c})$ *for* $\epsilon > 0$.

Proof. By the optimality of $\mathbf{u}_{T,\mathbf{c}}$ and $\mathbf{u}_{\infty,\mathbf{y}_0}$ it follows that

$$V^*(\mathbf{c}) = V(T, \mathbf{c}, \mathbf{u}_{\infty,\mathbf{c}}(\cdot)) + V(\infty, \mathbf{y}_0, u_{\infty,\mathbf{c}}(\cdot + T)) \geqq V(T, \mathbf{c}) + V^*(\mathbf{y}_0).$$

This establishes the right side of (2.3). Further, by the optimality of $\mathbf{u}_{\infty,\mathbf{c}}$,

$$V^*(\mathbf{c}) \leqq \min \{V(\infty, \mathbf{c}, \mathbf{u}'), V(\infty, \mathbf{c}, \mathbf{u}'')\},$$

where

$$\mathbf{u}'(s) = \begin{cases} \mathbf{u}_{T,\mathbf{c}}(s) & \text{if } 0 \leqq s \leqq T, \\ \mathbf{u}_{\infty,\mathbf{y}_1}(s - T) & \text{if } s > T, \end{cases}$$

and

$$\mathbf{u}''(s) = \begin{cases} \mathbf{u}_{T,\mathbf{c}}(s) & \text{if } 0 \leqq s \leqq T - \epsilon, \\ \mathbf{u}_{\infty,\bar{\mathbf{y}}_1}(s - (T - \epsilon)) & \text{if } s \geqq T - \epsilon. \end{cases}$$

This establishes the left side of (2.3).

It may be remarked that as $V(T, \mathbf{c})$ is monotone nondecreasing in T as T increases, $\lim_{T\to\infty} V(T, \mathbf{c})$ exists. The following theorems provide the major results.

THEOREM 2.1. $V(T, \mathbf{c})$ *converges as* $T \to \infty$ *uniformly on compact subsets of* R^n *to* $V^*(\mathbf{c})$. *Further,* $\mathbf{y}_0(T)$ *and* $\bar{\mathbf{y}}_1(T)$ *tend to zero as* T *tends to infinity.*

Proof. The definition of $V^*(\mathbf{c})$ and the optimality of $\mathbf{u}_{1,\mathbf{y}_0}$ have the consequence that

$$V^*(\mathbf{c}) \geqq V(T, \mathbf{c}, \mathbf{u}_{\infty,\mathbf{c}}(\cdot)) + V(1, \mathbf{y}_0(T)). \tag{2.4}$$

However, $V(T, \mathbf{c}, \mathbf{u}_{\infty,\mathbf{c}}(\cdot)) \uparrow V^*(\mathbf{c})$ as $T \to \infty$ by the monotone convergence theorem, and therefore

$$0 = \lim_{T\to\infty} \sup V(1, \mathbf{y}_0(T)) = \lim_{T\to\infty} V(1, \mathbf{y}_0(T)).$$

Since $V(1, \cdot)$ is continuous and vanishes only at the origin, $\mathbf{y}_0(T) \to 0$ as $T \to \infty$. The argument to show that $\bar{\mathbf{y}}_1(T) \to 0$ as $T \to \infty$ is similar since

$$V(T, \mathbf{c}) \geqq V(T - \epsilon, \mathbf{c}) + V(\epsilon, \bar{\mathbf{y}}_1(T)). \tag{2.5}$$

Now, as $\mathbf{y}_0(T)$ and $\bar{\mathbf{y}}_1(T)$ tend to zero as $T \to \infty$, (2.3) shows that

$$V(T, \mathbf{c}) \to V^*(\mathbf{c}) \quad \text{as} \quad T \to \infty.$$

The convergence is uniform on compact sets by Dini's theorem.

Now consider the sequence of controls $\mathbf{v}_{T,\mathbf{c}} \in L_\infty^2$ defined by

$$\mathbf{v}_{T,\mathbf{c}}(s) = \begin{cases} \mathbf{u}_{T,\mathbf{c}}(s) & \text{if} \quad s \leqq T - \epsilon, \\ \mathbf{u}_{\infty,\bar{\mathbf{y}}_1}(s) & \text{if} \quad s > T - \epsilon, \end{cases}$$

where $\epsilon > 0$. These controls approximate $\mathbf{u}_{\infty,\mathbf{c}}$ as the following theorem indicates.

THEOREM 2.2. *$\mathbf{v}_{T,\mathbf{c}}$ converges in norm in L_∞^2 to $\mathbf{u}_{\infty,\mathbf{c}}$ as $T \to \infty$. Further, for any sequence of real numbers tending to infinity there exists a subsequence T_n such that $\mathbf{v}_{T_n,\mathbf{c}} \to \mathbf{u}_{\infty,\mathbf{c}}$ almost everywhere as n tends to infinity.*

Proof. By the definition of $\mathbf{v}_{T,\mathbf{c}}$ it follows that

$$V(\infty, \mathbf{c}, \mathbf{v}_T) = V(T - \epsilon, \mathbf{c}, \mathbf{u}_{T,\mathbf{c}}(\cdot)) + V^*(\bar{\mathbf{y}}_1),$$

so that

$$\lim_{T\to\infty} V(\infty, \mathbf{c}, \mathbf{v}_T) = \lim_{T\to\infty} V(T - \epsilon, \mathbf{c}, \mathbf{u}_{T,\mathbf{c}}(\cdot))$$

by Theorem 2.1, if $V(T - \epsilon, \mathbf{c}, \mathbf{u}_{T,\mathbf{c}}(\cdot))$ has a limit. Now it is clear that

$$V^*(\mathbf{c}) - V^*(\bar{\mathbf{y}}_1) \leqq V(T - \epsilon, \mathbf{c}, \mathbf{u}_{T,\mathbf{c}}(\cdot)) \leqq V(T, \mathbf{c}),$$

so that $\lim_{T\to\infty} V(T - \epsilon, \mathbf{c}, \mathbf{u}_{T,\mathbf{c}}(\cdot)) = V^*(\mathbf{c})$. Corollary 2.1 implies the first assertion; the second follows since convergence in measure implies the existence of an almost everywhere convergent subsequence.

COROLLARY 2.2. *The optimal control $\varphi_{\mathbf{u}_{T,\mathbf{c}}}(\cdot, \mathbf{c})$ converges a.e. to $\varphi_{\mathbf{u}_{\infty,\mathbf{c}}}(\cdot, \mathbf{c})$ as $T \to \infty$.*

COROLLARY 2.3. *If F is stable, then*

$$\mathbf{v}^*_{T,\mathbf{c}}(s) = \begin{cases} \mathbf{u}_{T,\mathbf{c}}(s) & \text{if} \quad 0 \leqq s \leqq T - \epsilon, \\ 0 & \text{if} \quad s > T - \epsilon, \end{cases}$$

converges in norm to $\mathbf{u}_{\infty,\mathbf{c}}$ as $T \to \infty$. Every sequence of real numbers with a

limit point at $+\infty$ *has a subsequence* T_n *such that*

$$\mathbf{v}_{T_n,c} \longrightarrow \mathbf{u}_{\infty,c} \quad a.e.\ as \quad n \longrightarrow \infty.$$

Acknowledgments. It is a pleasure to acknowledge the help and encouragement and helpful conversations of R. E. Bellman. Many thanks are due to L. Markus and R. B. Lee for the opportunity to examine the notes for their unpublished book [5].

REFERENCES

[1] R. E. Bellman and R. S. Bucy, *Asymptotic control theory*, this Journal, 2 (1964), pp. 11–18.

[2] R. E. Bellman, I. Glicksberg and O. A. Gross, *Some aspects of the mathematical theory of control*, Report R-313, The RAND Corporation, Santa Monica, California, 1958.

[3] R. E. Kalman and R. S. Bucy, *New results in linear filtering and prediction theory*, Trans. ASME Ser. D. J. Basic Engrg., 83D (1961), pp. 95–108.

[4] L. H. Loomis, *Abstract Harmonic Analysis*, Van Nostrand, New York, 1953.

[5] L. Markus and R. B. Lee, *Notes on Control Theory*, University of Minnesota, Minneapolis, 1965, Chap. 3.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

EQUIVALENCE RELATIONS FOR THE CLASSIFICATION AND SOLUTION OF OPTIMAL CONTROL PROBLEMS*

C. D. CULLUM† AND E. POLAK‡

Introduction. This paper is concerned with the use of the concept of equivalence in the study of optimal control problems. The idea of using equivalence relations in the study of problems in system theory is not new, although until recently no apparent attempt had been made to apply this idea to the theory of optimal control. Lately, a number of papers have appeared by Polak [1], [2], [3], Hermes [4], Liu and Leake [5] in which equivalence relations are defined for optimal control problems and used to obtain theoretical or computational results for broad classes of problems. It is the purpose of this paper to formulate the ideas presented in these papers in a more general form. Actually, because of the type of control problem considered by the authors (the so called "open loop" problem), the definitions of equivalence given in [4] and [5] are not subsumed by the structure developed in this paper. However, it should be clear to the reader that a parallel development for closed loop control problems would unite and generalize the equivalence relations defined in [4] and [5].

It is shown in this paper that equivalence relations of the type defined herein lead to problem classification schemes which are both intuitively appealing and computationally useful. To demonstrate the latter, a new computational procedure for solving optimal control problems is presented and illustrated by examples. Finally, it is hoped that this classification scheme will lead to a greatly improved understanding of the invariant properties of optimal control problems.

The idealized physical system. The mathematical structure for constructing relations between optimal control problems will be based on the idealized regulator system shown in Fig. 1. This system consists of the following elements: a plant, describable by differential or difference equations, a controller, a computer, and a switch (sampler). Let X be the state space of the

* Received by the editors July 16, 1965, and in final revised form on April 25, 1966.
This research was conducted partly at the Electronics Research Laboratory of the University of California, Berkeley, and partly at the Electronic Systems Laboratory of the Massachusetts Institute of Technology. It was supported by the National Science Foundation under Grant GK-569 and by the National Aeronautics and Space Administration under Grants NsG-354 (supp 2) and NsG-496 with the Center for Space Research.

† Department of Electrical Engineering, University of California, Berkeley, California.

‡ Department of Electrical Engineering, University of California, Berkeley, California, and Massachusetts Institute of Technology, Cambridge, Massachusetts.

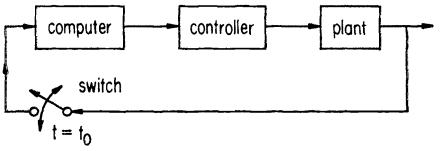

FIG. 1. *Idealized regulator system*

plant, $T = \{-\infty < t < +\infty\}$ the time axis, and V the space of all possible computer outputs, assumed to be such that $V = W \times T$, where W is a set of quantities whose elements determine the "shape" of the forcing functions produced by the contoller. V will be called the control space. When an input $v \in V$ is applied to the controller at time $t = t_0$, it produces a forcing function $u(s; w)$, $0 \leqq s \leqq \tau_v$, $s = (t - t_0)$, $v = (w, \tau_v)$.

The entire regulator system will be assumed to operate as follows. At time $t = t_0$ the switch closes momentarily, enabling the computer to read the plant state $x(t_0) \in X$, while the time t_0 is supplied by a clock. The computer then produces instantaneously a control v, resulting in a forcing function u which takes the plant state from $x(t_0)$ to a point in a given terminal set $X_f \subset X$.

Clearly, since every feedback control law gives rise to a corresponding open loop control law, this definition of the regulator system does not preclude the control laws implemented by the computer from being feedback laws. However, it will be more convenient for the purpose at hand to consider the system as being open loop.

An optimal control problem. An optimal control problem is completely determined by the following seven quantities:

(i) $X^* = X \times T$, the phase space of the system.

(ii) $X_i{}^* \subset X^*$, the set of initial phases.

(iii) $X_f{}^* \subset X^*$, the set of terminal phases.

(iv) V, the control space.

(v) $\mathfrak{A}: X^* \times V \to X^*$, the phase transition law of the system, assumed to have the following properties:

(a) $\mathfrak{A}_v \equiv \mathfrak{A}(\cdot, v): X^* \to X^*$ is one-to-one and onto for all $v \in V$,

(b) $\mathfrak{A}(x_0{}^*, v_0) = (x_1, t_0 + \tau_0)$, where $x_0{}^* = (x_0, t_0)$, $v_0 = (w_0, \tau_0)$, i.e., the last component of the image phase is $t_0 + \tau_0$.

(vi) $F_{\mathfrak{A}}: X^* \times V \to R$, a real valued cost functional depending parametrically on the phase transition law.

(vii) $G = \{g \mid g: X_i{}^* \to V$, and for all $x^* \in X_i{}^*$, $\mathfrak{A}(x^*, g(x^*)) \in X_f{}^*\}$, a set of admissible control laws.

The next step is to impose a partial ordering on the set G.

DEFINITION 1. Let g_1, g_2 be any two elements of G. Then $g_1 \leqq g_2$ if and only if $F_{\mathfrak{A}}(x^*, g_1(x^*)) \leqq F_{\mathfrak{A}}(x^*, g_2(x^*))$ for every $x^* \in X_i{}^*$.

The traditional statement of the optimal control problem can now be enunciated as follows:

Given the seven quantities specified above, find a $g^0 \in G$ such that $g^0 \leqq g$ for every $g \in G$.

Such a g^0 will be called an optimal control law. It is quite clear that we could think of an optimal control problem simply as the septuplet $(X^*, X_i^*, X_f^*, V, \mathfrak{A}, F_{\mathfrak{A}}, G)$, with the task of finding an optimal control $g^0 \in G$ always being implied. However, to make the ensuing discussion less cumbersome, we find it convenient to group the first six quantities in the septuplet together as part of the specification of a feasible solution.

DEFINITION 2. Let $\rho = (X^*, X_i^*, X_f^*, V, \mathfrak{A}, F_{\mathfrak{A}}, g)$, $g \in G$. Then ρ will be called a *feasible solution* to the optimal control problem specified by $X^*, X_i^*, X_f^*, V, \mathfrak{A}, F_{\mathfrak{A}}$, and G.

DEFINITION 3. Let $P = \{\rho \mid \rho = (X^*, X_i^*, X_f^*, V, \mathfrak{A}, F_{\mathfrak{A}}, g), g \in G\}$. Then, for any $\rho_1, \rho_2 \in P$, we define an order relation between ρ_1 and ρ_2 by

$$\rho_1 \leqq \rho_2 \quad \text{if and only if} \quad g_1 \leqq g_2 .$$

This defines a one-to-one ordered correspondence between *feasible solutions*, $\rho \in P$, and *admissible control laws*, $g \in G$.

It is now natural to define an optimal control problem as follows.

DEFINITION 4. An *optimal control problem* is defined to be a set of feasible solutions, differing only in their control laws, partially ordered according to Definition 3.

The next definition is a logical consequence of the preceding definitions.

DEFINITION 5. A feasible solution, $\rho^0 \in P$, is an *optimal solution* to the optimal control problem P if and only if $\rho^0 \leqq \rho$ for every $\rho \in P$.

Remark. For any optimal control problem there is always a question of existence of an optimal solution ρ^0. In what follows we shall always assume that an optimal solution exists.

Properties of control laws. We now establish some properties of control laws which will be required later on.

LEMMA 1. *Consider an optimal control problem P. Let $\rho \in P$ be arbitrary and let g be the corresponding control law. If $X_f^* = \{x_f^*\}$ consists of a single element only, then g is a one-to-one map from X_i^* into V.*

LEMMA 2. *Consider an optimal control problem P. Let $\rho \in P$ be arbitrary and let g be the corresponding control law. If*

$$X_i^* = \{x^* \mid x^* = (x, t_0), x \in X_i, t_0 \text{ fixed}\}$$

and if

$$X_f^* = \{x^* \mid x^* = (x_f, t), x_f \text{ fixed}, \quad t_0 \leqq t < \infty\},$$

then g is a one-to-one map from X_i^* into V.

The proofs of both these lemmas follow immediately from the assumed properties of the phase transition law.

Equivalence relations for optimal control problems. We now investigate possible ways of defining meaningful equivalence relations on a class of optimal control problems. One such definition, which immediately comes to mind, is to say that two optimal control problems, P_1 and P_2, are equivalent if there exists an isomorphism between the partially ordered sets $\{P_1, \leqq\}$ and $\{P_2, \leqq\}$, i.e., if there exists a one-to-one correspondence between the solutions of P_1 and P_2 such that if ρ_1^i, ρ_2^i, $i = 1, 2$, are corresponding solutions in P_1 and P_2 respectively, then $\rho_1^1 \leqq \rho_1^2$ if and only if $\rho_2^1 \leqq \rho_2^2$. This is clearly an equivalence relation. However, so many widely disparate problems are equivalent under this definition that it makes very little sense. Furthermore, this definition results in so little structure that it is doubtful that it could lead to any interesting results. In what follows, the authors propose a definition of equivalence which is more satisfying to one's intuition and which at the same time gives a certain amount of useful mathematical structure. This is accomplished by adding to the definition suggested above the condition that the isomorphism be constructed in a certain manner.

Equivalence. Let $\mathcal{P}$ be a class of optimal control problems and let P_1, P_2 be any two problems in this class, with corresponding subscripts identifying all of the significant quantities of P_1 and P_2. Let $R(G_1) = \bigcup g_1(X_{i1}^*) \subset V_1$, and $R(G_2) = \bigcup g_2(X_{i2}^*) \subset V_2$, with the unions taken over all $g_i \in G_i$, $i = 1, 2$.

DEFINITION 6. We shall say that P_1 is *equivalent* to P_2, written $P_1 \sim P_2$, if and only if there exist two maps φ_{12} and ψ_{12} satisfying

(a) $\varphi_{12}: X_1^* \to X_2^*$, one-to-one and onto, and

(i) $\varphi_{12}(X_{i1}^*) = X_{i2}^*$,

(ii) $\varphi_{12}(X_{f1}^*) = X_{f2}^*$;

(b) $\psi_{12}: R(G_1) \to R(G_2)$, one-to-one and onto;

such that the map π_{12} with domain G_1, which is induced by φ_{12}, ψ_{12} according to the relation

$$\pi_{12}(g_1)(x_2^*) = \psi_{12}(g_1(\varphi_{12}^{-1}(x_2^*))), \quad x_2^* \in X_{i2}^*, \qquad g_1 \in G_1,$$

(c) maps G_1 onto G_2 in a one-to-one manner, and

(d) induces an isomorphism[1] between the partially ordered sets $\{P_1, \leqq\}$ and $\{P_2, \leqq\}$.

[1] The map π_{12} induces a correspondence between solutions of P_1 and P_2 by assigning to every solution $\rho_1 \in P_1$ with control law g_1 the solution $\rho_2 \in P_2$ with control law $\pi_{12}(g_1)$.

Remark. It is trivial to verify that this relation is an equivalence relation.

At first glance, this definition may seem rather complicated and artificial to the reader. However, a little contemplation reveals that it is simply an extension of an intuitive idea of generating an equivalent optimal control problem by making a change of variables on the phase space and/or the control space. This is best illustrated by an example.

Example 1. Consider two problems P_1 and P_2 defined as follows. For P_1 the phase transition law is determined by the linear differential equation of the plant and the characteristics of the controller

$$\dot{x}_1 = Ax_1 + bu, \tag{1}$$

where $x \in E^n$, A is a constant $n \times n$ matrix, b is a constant n-vector, and u is the scalar valued output of the controller, satisfying the condition

$$|u(s; v)| \leqq 1, \qquad 0 \leqq s \leqq \tau, \quad \text{for all} \quad v \in V_1.$$

Hence

$$\mathfrak{A}(x_0^*, v) = \left(e^{A\tau}(x_0 + \int_0^\tau e^{-sA}bu(s; v)\, ds), t_0 + \tau\right). \tag{2}$$

The final and initial sets of phases are defined by

$$X_{f1}^* = \{(0, t_f)\}, \quad \text{a single point},$$

$$X_{i1}^* = \{(x_1, \ t_0), \quad x_1 \in X_{i1}, \quad t_0 \leqq t_f \quad \text{fixed}\}.$$

The set X_{i1} is the set of all states which can be taken to zero by means of admissible forcing functions u in the time $t_f - t_0$.

The cost functional is defined by

$$F_1\, \mathfrak{A}_1(x^*, v) = F(v) = \int_0^\tau |u(s; v)|\, ds, \quad \text{for all} \quad x^* \in X_1^*, v \in V_1.$$

For P_2, the state transition law is determined by the time varying vector differential equation

$$\dot{x}_2 = C(t)x_2 + d(t)u, \tag{3}$$

where $x_2 \in E^n$, $C(t) = L^{-1}(t)AL(t) - L^{-1}(t)\dot{L}(t)$, $d(t) = L^{-1}(t)b$, and $L(t)$ is an $n \times n$ matrix, with bounded components, whose derivative $\dot{L}(t)$ exists and has bounded components. In addition $L(t)$ is such that $L(t_0) = I$, the identity matrix, and $|\det L(t)| \geqq m > 0$ for all $t \in T$. The final and initial phase sets are defined by $X_{i2}^* = X_{i1}^*$, $X_{f2}^* = X_{f1}^*$. The cost functional $F_{2\mathfrak{A}_2} = F$, defined above, $V_2 = V_1$, and $G_2 = G_1$.

Clearly P_2 has been obtained from P_1 by making the change of variables $x_2(t) = L^{-1}(t)x_1(t)$. The reader may verify that these two problems are

equivalent with φ_{12} defined by

$$\varphi_{12}(x_1^*) = \varphi_{12}(x_1, t_1)) = (L^{-1}(t_1)x_1, t_1) \qquad \text{for all} \quad x_1^* \in X_1^*,$$

and the map ψ_{12} taken as the identity map.

The equivalence relation established above partitions the class $\mathcal{P}$ of optimal control problems into equivalence classes in a very desirable manner. It is reasonably clear that if the equivalence maps, φ_{0i}, ψ_{0i}, connecting a problem P_0 with problems P_i in the same equivalence class are known, then, by solving one problem, one has in fact solved the entire class of equivalent problems. Furthermore, one may also single out and examine sets of solutions in each problem with the same order properties. For example, one may use iterative techniques, such as the steepest descent method, to obtain a sequence of solutions $\{\rho_1^n\}$, in a problem P_1, whose costs, for a given initial phase, converge to the cost of an optimal solution in P_1. If P_2 is a problem equivalent to P_1, then, under the assumptions stated in the lemma below, the image sequence of solutions $\{\rho_2^n\}$ also has the property that, for the image initial phase, the associated sequence of costs converges to the optimal cost.

Let P_1 and P_2 be two equivalent problems under the equivalence maps φ_{12} and ψ_{12} and let g_1^0 and g_2^0 be optimal laws for P_1 and P_2 respectively with $g_2^0 = \pi_{12}(g_1^0)$. Fix $x_1^* \in X_{i1}^*$, and let $x_2^* = \varphi_{12}(x_1^*)$.

LEMMA 3. *If* $F_{\mathfrak{A}_i}(x_i^*, g_i^0)$, $i = 1, 2$, *are cluster points*[2] *of the sets* $\{F_{\mathfrak{A}_i}(x_i^*, g_i) = g_i \in G_i\}$, $i = 1, 2$, *and* $\{\rho_1^n\}$ *is any sequence of solutions in* P_1, *with image sequence* $\{\rho_2^n\}$ *in* P_2, *then* $F_{\mathfrak{A}_1}(x_1^*, g_1^n) \downarrow F_{\mathfrak{A}_1}(x_1^*, g_1^0)$ *if and only if* $F_{\mathfrak{A}_2}(x_2^*, g_2^n) \downarrow F_{\mathfrak{A}_2}(x_2^*, g_2^0)$.

Proof. Necessity. The order preserving property of the isomorphism plus the existence of an optimal cost guarantee that $F_{\mathfrak{A}_2}(x_2^*, g_2^n)$ converges. If $F_{\mathfrak{A}_2}(x_2^*, g_2^n) \downarrow C > F_{\mathfrak{A}_2}(x_2^*, g_2^0)$, then there exists a $\hat{g}_2$ with $F_{\mathfrak{A}_2}(x_2^*, g_2^0) < F_{\mathfrak{A}_2}(x_2^*, \hat{g}_2) < C$ because $F_{\mathfrak{A}_2}(x_2^*, g_2^0)$ is a cluster point. Similarly, there exists an N such that

$$F_{\mathfrak{A}_1}(x_1^*, g_1^0) < F_{\mathfrak{A}_1}(x_1^*, g_1^N) < F_{\mathfrak{A}_1}(x_1^*, \pi_{12}^{-1}(\hat{g}_2)).$$

This implies that $F_{\mathfrak{A}_2}(x_2^*, g_2^N) < C$, a contradiction. The sufficiency can be proven in a similar manner.

Equivalence under optimal controls. It is reasonably clear that if one is interested in optimal control problem classification schemes depending only on the nature of the optimal solutions, then it is excessive to require that all solutions of one problem have correspondingly ordered images in the other problems belonging to the same equivalence class. We shall therefore

[2] The point x is said to be a cluster point of the set K if every neighborhood of x contains a point of K different from x.

confine our attention to the subsets formed by the optimal solutions of the problems under consideration.

Let G be the set of control laws associated with the optimal control problem P. The set $G^0 \subset G$ consisting of all the optimal control laws g^0 in G will be said to be the set of optimal control laws for the problem P. We now introduce a classification scheme depending on optimal solutions only.

DEFINITION 7. Let P_1 and P_2 be two optimal control problems and let $P_1^0 \subset P_1$, $P_2^0 \subset P_2$ be nonempty subsets consisting of all their respective optimal solutions. The problem P_1 will be said to be optimal control equivalent to the problem P_2, written $P_1 \overset{0}{\sim} P_2$, if and only if $P_1^0 \sim P_2^0$, i.e., if and only if P_1 is equivalent to P_2 when the admissible control law sets G_1, G_2 are reduced to the optimal control law sets G_1^0, G_2^0 respectively.

Remark. It is readily seen that the relation $\overset{0}{\sim}$ is symmetric, reflexive and transitive and that it is therefore a true equivalence relation. It will also be observed that condition (d) in Definition 6 is satisfied trivially in the case of optimal control equivalence and hence need not be checked.

By relaxing the conditions under which two problems will be considered equivalent, we have introduced a significantly more useful equivalence relation. To illustrate the nature of optimal control equivalence, we consider the following example.

Example 2.

Problem (a)

Given: $\dot{x}_{a1} = x_{a2}$,
$\dot{x}_{a2} = u_a$, $|u_a| \leqq 1$,
$x_a = x_{a0}$ at $t = 0$.

Find: an admissible forcing function $t \to u_a(t)$ such that $x_{a0} \to 0$ in minimum time.

Problem (b)

Given: $\dot{x}_{b1} = x_{b2}$,
$\dot{x}_{b2} = -2x_{b2} - x_{b1} + \frac{1}{2}\tan^{-1} x_{b1} + u_b$,
$|u_b| \leqq 1$, $-\pi/2 < \tan^{-1} x_{b1} < \pi/2$,
$x_b = x_{b0}$ at $t = 0$.

Find: an admissible forcing function $t \to u_b(t)$ such that $x_{b0} \to 0$ in minimum time.

It is well-known [7] that the optimal forcing functions for Problem (a) are "bang-bang" with at most one switching, and Lee and Markus [6] have proved the same to be true for Problem (b). If one examines the sets of optimal trajectories in the state plane for these two problems, one is immediately led to the idea that the optimal solutions are "equivalent."

More formally, it is clear that if $X_a^* = X_b^* = E^2 \times T$, $X_{ia}^* = X_{ib}^* = E^2 \times \{0\}$, $X_{fa}^* = X_{fb}^* = \{0\} \times T^+$, where $T^+ = \{t \mid 0 \leqq t < \infty\}$, and if

$$V_a = V_b = \{v \mid v = (t_1, t_2, \tau), -\infty < t_1 < \infty, -\infty < t_2 < \infty, \tau = |t_1| + |t_2|\},$$

with $u_j(t; v)$, for $j = a, b$, given by

$$u_j(t; v) = \begin{cases} \operatorname{sgn} t_1 & \text{if} \quad 0 \leqq t < |t_1|, \\ \operatorname{sgn} t_2 & \text{if} \quad |t_1| \leqq t < \tau, \end{cases} \tag{4}$$

then $g_a^0(X_{ia}^*) = g_b^0(X_{ib}^*) \subset V$. Consequently, $P_a \overset{0}{\sim} P_b$ under the equivalence maps $\psi_{ab} = I$ and $\varphi_{ab} = (g_b^0)^{-1} \cdot g_a^0$, where $(g_b^0)^{-1}$ exists by virtue of Lemma 2.

The same reasoning may be used to establish that a wide class of minimum time problems with second order nonlinear plants are optimal control equivalent to a "second order integrator" problem (Problem (a)). In particular, see Example 4 below.

Synthesis of optimal control laws. An inherent property of the equivalence relations exhibit so far is that the equivalence maps may be used to find an optimal solution for any problem in the equivalence class, whenever an optimal solution to one problem is known. This raises the possibility of obtaining a computational method for determining optimal control laws for a whole class of problems by solving the simplest problem in the class. It is shown below that it is indeed possible to solve certain optimal control problems in this manner. However, before demonstrating this, we first introduce a relation between optimal control problems which is still weaker than optimal control equivalence. This relation has the property that it may be used to synthesize optimal control laws in exactly the same way as the other relations.

DEFINITION 8. Let P_1 and P_2 be two optimal control problems and let $P_1^0 \subset P_1$, $P_2^0 \subset P_2$ be nonempty subsets consisting of all their respective optimal solutions. The problem P_1 will be said to be *weak-optimal-control-equivalent* to the problem P_2, written $P_1 \overset{\text{w.o.}}{\sim} P_2$, if and only if there exist nonempty subsets $P_{11}^0 \subset P_1^0$, $P_{21}^0 \subset P_2^0$, such that $P_{11}^0 \sim P_{21}^0$.

Remark. The above relation between problems, which, for lack of a better term, we shall call weak-optimal-control-equivalence, is actually not an equivalence relation. It is symmetric and reflexive, but, in general, not transitive.

Remark. It is clear from the definitions that

$$\{P_1 \sim P_2\} \Rightarrow \{P_1 \overset{0}{\sim} P_2\} \Rightarrow \{P_1 \overset{\text{w.o.}}{\sim} P_2\}.$$

The reason for introducing the concept of weak-optimal-control-equivalence in the discussion of the synthesis of optimal control laws is that it exhibits all the desirable properties of the other equivalence relations, while possessing two additional advantages. The first advantage is that there are two methods by means of which weak-optimal-control-equivalence is easily established. The first of these methods applies to the class or problems for

which the range of optimal control law is known, for example, due to the Pontryagin maximum principle. The second method applies to the class of optimal control problems for which it is relatively easy to construct isocost sets in the phase space. Second, the authors have found that it is possible to construct "prototype" problems whose optimal solutions can be determined by inspection, and which are weak-optimal-control-equivalent to a certain class of optimal control problems. Generally, these "prototype" problems are not optimal-control-equivalent to the problems to which they are weak-optimal-control-equivalent.

The application of equivalence concepts to the synthesis of optimal control laws rests essentially on the following two theorems.

THEOREM 1. *Let P_1, P_2 be two optimal control problems with identical finite dimensional Euclidean phase spaces, i.e., $X_1^* = X_2^*$. Let $\rho_1^0 \in P_1$, $\rho_2^0 \in P_2$ be optimal solutions and let g_1^0, g_2^0 be the optimal control laws associated with ρ_1^0, ρ_2^0, respectively. If for $k = 1, 2$,*

(a1) *either the terminal phase sets $X_{fk}^* = \{x_{fk}^*\}$ consist of a single point only, or*

(a2) *the initial phase sets are contained in hyperplanes $t = t_{0k}$, i.e.,*

$$X_{ik}^* = \{x^* \mid x^* = (x, t_{0k}),\ x \in X_{ik},\ t_{0k} \text{ fixed}\},$$

and the terminal phase sets consist of a halfline:

$$X_{fk}^* = \{x^* \mid x^* = (x_{fk}, t),\ x_{fk} \text{ fixed},\ t_{0k} \leqq t < \infty\};$$

(b1) *either $X_{i1}^* \cap X_{f1}^* = X_{i2}^* \cap X_{f2}^* = \varnothing$, and there exists a map $\psi_{12}: R(g_1^0) \to R(g_2^0)$, one-to-one and onto, or*

(b2) *$X_{i1}^* \cap X_{f1}^* \neq \varnothing$ and $X_{i2}^* \cap X_{f2}^* \neq \varnothing$, and there exists a map $\psi_{12}: R(g_1^0) \to R(g_2^0)$, one-to-one and onto such that $\psi_{12}(g_1^0(X_{i1}^* \cap X_{f1}^*)) = g_2^0(X_{i2}^* \cap X_{f2}^*)$,*

then $P_1 \overset{\text{w.o.}}{\sim} P_2$.

Proof. Due to the assumptions (a1) and (a2), it is clear that the conditions of either Lemma 1 or Lemma 2 are satisfied, and hence that the optimal control laws g_1^0, g_2^0 are both one-to-one. Let $(g_1^0)^{-1}: R(g_1^0) \to X_{i1}^*$, $(g_2^0)^{-1}: R(g_2^0) \to X_{i2}^*$ be their respective inverses.

Also by assumption, $X_1^* = X_2^*$ and (due to conditions (a1) and (a2) which state that the terminal phase sets are both either a point or a halfline) X_{f1}^*, X_{f2}^* can be brought into one-to-one correspondence. Hence there exists an affine map $\hat{\varphi}_{12}: X_1^* \to X_2^*$, one-to-one and onto and such that $\hat{\varphi}_{12}(X_{f1}^*) = X_{f2}^*$.

Since the map ψ_{12} exists by assumption, it is only necessary to construct a map φ_{12} such that φ_{12}, ψ_{12} are a pair of equivalence maps under which $P_1 \overset{\text{w.o.}}{\sim} P_2$. Let $\varphi_{12}: X_1^* \to X_2^*$ be defined as follows:

$$\varphi_{12}(x_1^*) = \begin{cases} (g_2^0)^{-1} \cdot \psi_{12} \cdot g_1^0(x_1^*) \text{ for all } x_1^* \in X_{i1}^*, \\ \hat{\varphi}_{12}(x_1^*) \text{ for all } x_1^* \in X_{i1}^{*c}. \end{cases} \tag{5}$$

Due to the nature of ψ_{12}, $\varphi_{12}(X_{i1}^*) = X_{i2}^*$ and due to the assumptions in (b1) and (b2), $\varphi_{12}(X_{f1}^*) = X_{f2}^*$. Clearly φ_{12} is one-to-one and onto from $X_1{}^*$ onto $X_2{}^*$. Furthermore, the image of $g_1{}^0$ under the induced map π_{12} is

$$\begin{aligned} \pi_{12}(g_1{}^0) &= \psi_{12} \cdot g_1{}^0 \cdot \varphi_{12}^{-1} \\ &= \psi_{12} \cdot g_1{}^0 \cdot (g_1{}^0)^{-1} \cdot \psi_{12}^{-1} \cdot g_2{}^0 \\ &= g_2{}^0 \text{ on } X_{i2}^* . \end{aligned} \tag{6}$$

Hence the maps φ_{12}, ψ_{12} are a pair of equivalence maps under which $P_1 \overset{\text{w.o.}}{\sim} P_2$.

We shall now give an example in which the above theorem is used to show that the problems in a class of minimum time optimal control problems with third order nonlinear plants are each weak-optimal-control-equivalent to a minimum time optimal control problem with a third order linear plant whose eigenvalues are real. Thus, many methods which have been proposed for the solution of the linear time optimal control problem can be extended to this class of nonlinear time optimal control problems.

Example 3. (Time optimal control of a class of third order nonlinear systems). Consider the class $\mathcal{P}$ of problems whose systems can be represented by the block diagram in Fig. 2, where N is a differentiable function with

$$\begin{aligned} &\text{(a)} \quad N(0) = 0, \\ &\text{(b)} \quad N'(z) > 0 \quad \text{for every} \quad z, \end{aligned} \tag{7}$$

and

$$\lambda_2 \neq \lambda_3 , \quad \lambda_1 , \lambda_2 , \lambda_3 < 0.$$

One is required in each case, to bring the system from an arbitrary state to the origin in minimum time, subject to the constraint $| u | \leqq 1$. The system can also be represented by a block diagram as shown in Fig. 3, and the state equations corresponding to this form are

$$\begin{aligned} \dot{x}_1 &= \lambda_1 x_1 + N(x_2 + x_3), \\ \dot{x}_2 &= \lambda_2 x_2 + (\lambda_2 - \lambda_3)u, \\ \dot{x}_3 &= \lambda_3 x_3 - (\lambda_2 - \lambda_3)u. \end{aligned} \tag{8}$$

Applying Pontryagin's maximum principle to this problem, it is easy to show [8] that every optimal forcing function is bang-bang, with at most two switchings. It is equally simple to show that, if the maximum principle is also a sufficient condition for optimality, then every bang-bang control with at most two switchings is optimal. (Equivalently, it is sufficient to show that no two bang-bang controls with at most two switchings bring the same initial state to the origin.) It is intuitively obvious, but very

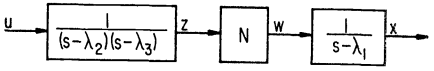

FIG. 2. *Block diagram of system of Example 3*

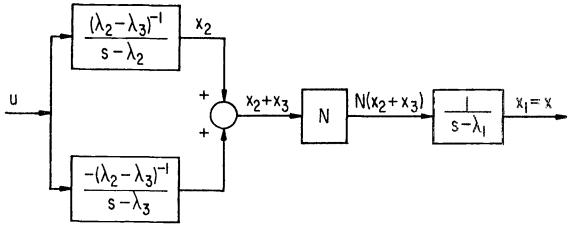

FIG. 3. *Modified block diagram of system of Example 3*

difficult to prove, that there are problems in the class $\mathcal{P}$ (with nonlinear plants) for which the range of the optimal control law is the entire set of bang-bang functions with at most two switchings. Therefore, we simply let $\hat{\mathcal{P}}$ be the subclass of $\mathcal{P}$ consisting of problems for which the range of the optimal control law is the entire class of bang-bang functions with at most two switchings.

Now, let P_1 and P_2 be any two problems in $\hat{\mathcal{P}}$. Clearly, we may take ψ_{12} to be the identity map in (b2) of Theorem 1. The rest of (b2) is satisfied since $X_{i1}^* \cap X_{f1}^* = X_{i2}^* \cap X_{f2}^* = \{0\}$, and the time optimal control is the zero control in every case. Condition (a2) is obviously satisfied in this case, and, consequently, $P_1 \overset{\text{w.o.}}{\sim} P_2$ by Theorem 1. Indeed, since the optimal control law is unique in every case, the problems in $\hat{\mathcal{P}}$ are all optimal control equivalent, and, therefore, $\hat{\mathcal{P}}$ is an equivalence class. Note that $\hat{\mathcal{P}}$ contains all the problems whose plants are described by linear third order differential equations with real eigenvalues.

Let P_1 be a problem in $\hat{\mathcal{P}}$ with a linear, real eigenvalue plant. By construction, $\varphi_{12} = (g_2^0)^{-1} \cdot g_1^0$, where $(g_2^0)^{-1}$ is the inverse of the optimal control law for P_2 in $\hat{P}$ and can be determined explicitly by solving the plant equation of P_1. Thus, if the optimal control law for P_1 can be determined by some method, then φ_{12} is known explicitly. In fact, φ_{12} can be shown to be a homeomorphism for the class of problems considered. Consequently, knowledge of the optimal control law for the problem with the linear plant determines explicitly the equivalence maps, which relate this problem to all the other problems in $\hat{\mathcal{P}}$ and, moreover, these maps have nice properties.

We shall now examine optimal control problems for which isocost phase sets are relatively easy to construct. Let P be an optimal control problem and let $\rho^0 \in P$ be an optimal solution with the associated optimal control

law g^0 and cost functional $F_{\mathfrak{A}}$. The subset of initial phases

$$X_c^* = \{x^* \mid x^* \in X_i^*, F_{\mathfrak{A}}(x^*, g^0(x^*)) = c, c \in R\}$$

will be called the c-minicost set. Clearly, X_c^* is the c-isocost set under the optimal control law g^0. The c-minicost sets of a given optimal control law g^0 are obviously independent of the particular optimal control law g^0 used for their definition. Furthermore, they can often be constructed without the knowledge of an optimal law (see [1], [2], [3]). In such cases the following theorem has been found of value.

THEOREM 2. *Let P_1, P_2 be two optimal control problems with identical phase control spaces, i.e., $X_1^* = X_2^*$, $V_1 = V_2$, and whose cost functionals have the same form: for $k = 1, 2$,*

$$F_{k\mathfrak{A}_k}(x^*, v) = \int_0^\tau f(u_k(s; w))\, ds, \quad x^* \in X_{ik}^*, v \in V_k, v = (w, \tau),$$

where u_k is the forcing function produced by the controller of the problem P_k, and f is a scalar valued cost function such that the integral is well-defined, and satisfies the condition

$$f(u_1(s; w)) = f(u_2(s; w)), \qquad v \in V = V_1 = V_2, 0 \leqq s \leqq \tau.$$

If there exists a map $\varphi_{12}: X_1^* \to X_2^*$, one-to-one and onto, such that

(a) $\varphi_{12}(X_{i1}^*) = X_{i2}^*$ (*initial phase sets*),

(b) $\varphi_{12}(X_{f1}^*) = X_{f2}^*$ (*terminal phase sets*),

(c) $\varphi_{12}(X_{c1}^*) = X_{c2}^*$ (*c-minicost sets*),

(d) *for some optimal solution $\rho_1^0 \in P_1$ with associated optimal control law g_1^0, the image control law defined by $g_2^0 = g_1^0 \cdot \varphi_{12}^{-1}$ is a control law in G_2;*

then φ_{12}, I (the identity map) are a pair of equivalence maps such that $P_1 \overset{\text{w.o.}}{\sim} P_2$, and the control law $g_2^0 = g_1^0 \cdot \varphi_{12}^{-1}$ is an optimal control law.

Proof. We only need to show that the control law g_2^0 defined in (d) is optimal, since it is then immediately obvious that the postulated maps φ_{12}, I are indeed a satisfactory pair of equivalence maps. Let x_2^* be an arbitrary point in X_{i2}^*. Hence $x_2^* \in X_{c2}^*$ for some c. It follows from condition (c) that $\varphi_{12}^{-1}(x_2^*) \in X_{c1}^*$. Let $g_1^0(\varphi_{12}^{-1}(x_2^*)) = v$. Then, by definition,

$$g_2^0(x_2^*) = g_1^0(\varphi_{12}^{-1}(x_2^*)) = v,$$

and

$$F_{2\mathfrak{A}_2}(x_2^*, v) = F_{1\mathfrak{A}_1}(\varphi_{12}^{-1}(x_2^*), v) = c,$$

i.e., the cost for any initial phase $x^* \in X_{i2}^*$, resulting from the control law

g_2^0, is equal to the optimal cost. Hence g_2^0 is an optimal control law, and $P_1 \overset{\text{w.o.}}{\sim} P_2$.

This theorem was used by one of the authors (see [1], [2], [3]) to construct an optimal control law for minimum time and minimum fuel problems with pulse-width modulation controllers from weak-optimal-control-equivalent problems with pulse-amplitude modulation controllers. The minicost sets were constructed by a method related to dynamic programming.

The second advantage mentioned previously can best be illustrated by an example. It is well-known that for a large class of second order nonlinear systems the problem of bringing the system from an arbitrary initial state to the origin in minimum time, with bounded scalar control, has the following unique solution:

> Every optimal forcing function is bang-bang with at most one switching, and every bang-bang forcing function with at most one switching is uniquely optimal for the state which it brings to the origin.

One such system was given in Example 2(b). A whole class of such problems is given in the following example.

Example 4. The problems in the class $\mathcal{P}$ considered here have plants whose state equations take the form

$$\dot{x}_1 = f(x_2), \qquad \dot{x}_2 = u, \tag{9}$$

where f is assumed to be a single-valued differentiable function with

$$\begin{aligned} &\text{(a)} && f(0) = 0, \\ &\text{(b)} && f'(z) > 0 \text{ for every } z. \end{aligned} \tag{10}$$

One is required in each case to bring the system from an arbitrary initial state to the origin in minimum time, subject to the constraint $|u| \leqq 1$.

The reader can easily verify that all the problems in Example 4 have unique optimal solutions of the type mentioned above. Now consider the following problem.

Example 5. *Problem* (c)

$$\begin{aligned} \text{Given:} \qquad & \dot{x}_{c1} = \operatorname{sgn} x_{c2} = \begin{cases} 1 & \text{if } x_{c2} > 0, \\ 0 & \text{if } x_{c2} = 0, \\ -1 & \text{if } x_{c2} < 0, \end{cases} \\ & \dot{x}_{c2} = u_c, \quad u_c \in \{1, 0, -1\}, \\ & x_c = x_{c0} \quad \text{at} \quad t = 0. \end{aligned} \tag{11}$$

Find: an admissible forcing function $t \to u_c(t)$ such that $x_{c0} \to 0$ in minimum time.

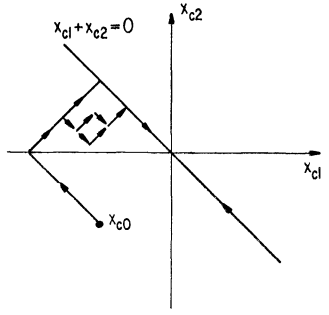

FIG. 4. *Optimal trajectories for Problem* (c) *of Example 5. The dotted paths are alternative optimal trajectories for* x_{c0} .

The possible trajectories for Problem (c) are all piecewise linear. We can determine optimal trajectories by inspection, and for almost all initial states there are infinitely many optimal trajectories. Fig. 4 illustrates the different types of optimal trajectories for a typical initial state. Among the possible optimal forcing functions for any initial state there is always exactly one which is bang-bang with at most-one switching. In fact, this forcing function is given by

$$u(t) = \begin{cases} \operatorname{sgn} t_1 & \text{if} \quad 0 \leqq t < |t_1|, \\ \operatorname{sgn} t_2 & \text{if} \quad |t_1| \leqq t < |t_1| + |t_2|, \end{cases} \tag{12}$$

where

$$t_1 = \begin{cases} -(\tfrac{1}{2}|x_{c2}| + \tfrac{1}{2}x_{c1}) - x_{c2} & \text{if} \quad x_{c1} + x_{c2} > 0, \\ (\tfrac{1}{2}|x_{c2}| - \tfrac{1}{2}x_{c1}) - x_{c2} & \text{if} \quad x_{c1} + x_{c2} < 0, \\ 0 & \text{if} \quad x_{c1} + x_{c2} = 0, \end{cases} \tag{13a}$$

$$t_2 = \begin{cases} (\tfrac{1}{2}|x_{c2}| + \tfrac{1}{2}x_{c1} & \text{if} \quad x_{c1} + x_{c2} > 0, \\ -(\tfrac{1}{2}|x_{c2}| - \tfrac{1}{2}x_{c1}) & \text{if} \quad x_{c1} + x_{c2} < 0, \\ -x_{c2} & \text{if} \quad x_{c1} + x_{c2} = 0. \end{cases} \tag{13b}$$

Furthermore, the set of forcing functions defined by the above expression for arbitrary initial states is the set of all bang-bang forcing functions with at most one switching.

We may now apply Theorem 1 to show that Problem (c) and any one of the other problems mentioned above are weak-optimal-control-equivalent. Clearly, these problems can not be optimal control equivalent since Problem (c) has infinitely many optimal solutions while the other problems have unique optimal solutions. However, weak-optimal-control-equivalence to-

gether with an explicit expression for the optimal control law of Problem (c) allows us to determine the optimal control law for any problem of the type specified in Example 4.

Let P_1 be any problem of the type specified in Example 4, and let P_2 be Problem (c). Let g_1 be the optimal control law for P_1, and let g_2 be the optimal control law for P_2 given by (13). Then, by Theorem 1, $P_1 \overset{\text{w.o.}}{\sim} P_2$ with $\varphi_{12} = g_2^{-1} \cdot g_1$ and $\psi_{12} = I$. Given any $x_1^* \in X_{i1}^*$, we have $g_1(x_1^*) = g_2[\varphi_{12}(x_1^*)]$. Therefore, if we can find $x_2^* = \varphi_{12}(x_1^*)$, we can find $g_1(x_1^*)$. To do this we need to solve

$$\varphi_{12}^{-1}(x_2^*) = g_1^{-1} \cdot g_2(x_2^*) = x_1^*$$

for x_2^*. The functions g_2 and g_1^{-1} are known and, furthermore, for all the problems in Example 4, $\varphi_{12}^{-1} = g_1^{-1} \cdot g_2$ is a homeomorphism and piecewise $C^{(1)}$. Thus, it is possible to solve for x_2^* iteratively. A complete calculation is carried out in the example below.

Example 6. Consider the particle moving in one dimension according to the equation

$$\frac{d}{dt}(m\dot{y}) = u, \tag{14}$$

where y is the position of the particle, u is the applied force, and

$$m = \frac{100}{\sqrt{10^4 - \dot{y}^2}} \quad \text{for} \quad |\dot{y}| < 100. \tag{15}$$

We assume that the force is constrained by $|u| \leqq 1$, and the initial velocity satisfies $|\dot{y}| < 100$. We are required to bring the system to rest at the origin from an arbitrary initial position and an arbitrary initial velocity in the range $|\dot{y}| < 100$ in minimum time.

If we make the substitution $x_1 = y$, $x_2 = p = m\dot{y}$, then the system is

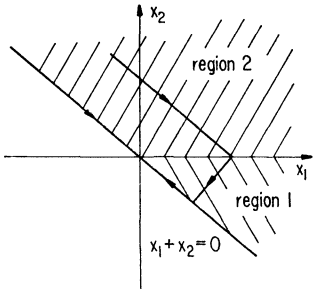

FIG. 5. *State space and minimum time switching line for prototype problem of Example 5*

TABLE 1. *Computational results for Example* 6

Desired initial state		Computed initial state		Optimal control law			Computation time (sec.)
$Z0(1)$	$Z0(2)$	$Z1(1)$	$Z1(2)$	$U1$	$T1$	$T2$	
1.0	1.0	1.0059	1.0000	−1	2.2772	1.2272	0.036
10.0	10.0	10.0743	10.0000	−1	17.8314	7.7810	0.036
0.0	50.0	0.0009	50.0000	−1	97.8204	40.0854	0.036
50.0	50.0	50.1494	50.0000	−1	98.4890	40.7545	0.018
50.0	0.0	50.0215	0.0000	−1	7.0770	7.0770	0.036
50.0	−50.0	49.5777	−50.0000	+1	97.1493	39.4143	0.036
90.0	90.0	89.7729	90.0000	−1	337.9153	131.4411	0.084
0.0	95.0	−0.0060	95.0000	−1	489.0511	184.8076	0.234
100.0	95.0	99.6597	95.0000	−1	489.6175	185.3740	0.318
100.0	0.0	100.6797	0.0000	−1	10.0466	10.0466	0.018
100.0	−95.0	100.9541	−95.0000	+1	488.4770	184.2335	0.198
1000.0	95.0	1005.3431	95.0000	−1	494.7471	190.5036	0.048
1000.0	0.0	1004.7053	0.0000	−1	32.0927	32.0927	0.048
1000.0	−95.0	1006.0459	−94.9912	+1	482.8140	178.9100	0.036

described by

$$\dot{x}_1 = f(x_2) = \frac{100\, x_2}{\sqrt{10^4 + x_2^2}}, \qquad \dot{x}_2 = u. \tag{16}$$

Inspecting the form of (16), we see that this problem falls into the class of problems considered in and following Example 4.

Instead of using this form, the authors chose as state variables the quantities $z_1 = y$, $z_2 = \dot{y}$, yielding the equations

$$\dot{z}_1 = z_2\,, \qquad \dot{z}_2 = 10^{-6}[10^4 - z_2^2]^{1.5}u, \tag{17}$$

with the initial phase set $z_i = \{z \mid -100 < z_2 < 100\}$. Optimal forcing functions for this system are independent of the choice of state variables, and hence we can still find the equivalence map φ_{12}^{-1} from the state space of Problem (c) to the set z_i according to

$$\varphi_{12}^{-1} = g_1^{-1} \cdot g_2\,,$$

where g_2 is given by (13), and g_1^{-1} is obtained by integrating (17) backward in time from the origin.

In evaluating g_1, it is sufficient to restrict our attention to the shaded regions of Fig. 5, since the map for the rest of the state space can be obtained by symmetry arguments.

Let (y_1, y_2) be an initial state for Problem (c) (P_2). Then the map $(z_1, z_2) = \varphi_{12}^{-1}(y_1, y_2)$ is given by:

(a) For (y_1, y_2) in Region 1 (see Fig. 5),

$$z_1 = \frac{1}{4}\left[\frac{(y_1 - y_2)^2}{1 + \left[1 + \left(\frac{y_1 - y_2}{100}\right)^2\right]^{1/2}} - \frac{(3y_2 - y_1)(y_2 + y_1)}{\left[1 + \left(\frac{y_2}{100}\right)^2\right]^{1/2} + \left[1 + \left(\frac{y_1 - y_2}{200}\right)^2\right]^{1/2}}\right], \tag{18}$$

$$z_2 = \frac{y_2}{\left[1 + \left(\frac{y_2}{100}\right)^2\right]^{1/2}}.$$

(b) For (y_1, y_2) in Region 2 (see Fig. 5),

$$z_1 = \frac{1}{4}\left[\frac{(y_1 + y_2)^2}{1 + \left[1 + \left(\frac{y_1 + y_2}{200}\right)^2\right]^{1/2}} - \frac{(3y_2 + y_1)(y_2 - y_1)}{\left[1 + \left(\frac{y_2}{100}\right)^2\right]^{1/2} + \left[1 + \left(\frac{y_1 + y_2}{200}\right)^2\right]^{1/2}}\right], \tag{19}$$

$$z_2 = \frac{y_2}{\left[1 + \left(\frac{y_2}{100}\right)^2\right]^{1/2}}.$$

It is not difficult to verify that φ_{12}^{-1} is indeed a homeomorphism and $C^{(1)}$ everywhere except on the lines $y_2 = 0$ and $y_1 + y_2 = 0$. The equations given here were used in conjunction with an IBM 7094 digital computer to compute $g_1(z)$ according to the formula $g_1(z) = g_2(\varphi_{12}(z))$. The inversion of φ_{12}^{-1} was accomplished numerically using a modified Newton-Raphson method. Table 1 gives results for various initial states together with the computation time required.

Conclusion. This paper has attempted to answer the question of whether optimal control problems can be classified in a manner which is both intuitively appealing and computationally useful.

For this purpose, three relations, defined either on all the admissible solution sets, or only on subsets of the optimal solutions sets, were exhibited. The first two of these selections, equivalence and optimal control equivalence, are true equivalence relations. It was shown by means of a number of examples, either worked in this paper or cited from the literature, that the optimal control problems classifications in which they result are highly nontrivial, and that the associated mathematical structure can be quite useful in the construction of algorithms for finding optimal solutions. The

third solution, weak-optimal-control-equivalence, is reflexive and symmetric, but not transitive, and hence it is not a true equivalence relation. Although it is not as useful for classification as the other two relations described, it is by far the most powerful one when applied to the construction of algorithms.

To facilitate the use of equivalence relations in the construction of algorithms, the authors introduce the concept of a prototype problem. This is usually an artificial problem, which can be solved in a very simple manner and which is related to the problem one wishes to solve. It is shown by means of an example how prototypes can be used to obtain algorithms for solving optimal control problems.

Finally, this paper has exhibited what the authors hope will be a new point of view to many who are working in the field of optimal control.

REFERENCES

[1] E. Polak, *On the equivalence of discrete systems in time-optimal control*, Trans. ASME Ser. D. J. Basic Engrg., 85D (1963), pp. 204–210.

[2] ———, *Equivalence and optimal strategies for some minimum fuel discrete systems*, J. Franklin Inst., 277 (1964), pp. 150–162.

[3] ———, *Minimal time control of a discrete system with a nonlinear plant*, IEEE Trans. Automatic Control, AC-8(1963), pp. 49–56.

[4] H. Hermes, *The equivalence and approximation of optimal control problems*, J. Differential Equations, 1 (1965), pp. 409–426.

[5] R. Liu and R. J. Leake, *Exhaustive equivalence classes of optimal systems with separable controls*, Tech. Report EE-652, Dept. of Electrical Engineering, University of Notre Dame, Notre Dame, March 1965.

[6] E. B. Lee and L. Markus, *Synthesis of optimal control for nonlinear processes with one degree of freedom*, International Union of Theoretical and Applied Mechanics, The International Symposium on Nonlinear Vibrations, Kiev, 1961.

[7] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[8] C. D. Cullum, *Equivalence relations for the classification and solution of optimal control problems*, Ph.D. Thesis, University of California, Berkeley, 1966.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

OPTIMAL TERMINAL MANEUVER AND EVASION STRATEGY*

YU-CHI HO†

Problem formulation. A typical situation in the terminal guidance of homing missile or military satellites may be described as follows: The position and velocity errors of the missile or satellite system obeys the linearized dynamic model,

$$\dot{x} = F(t)x + G(t)u + w, \tag{1}$$

where $F(t)$ and $G(t)$ are known continuous $n \times n$ and $n \times r$ time-varying matrices and $w(t)$ is a vector white gaussian random process with zero mean and covariance $Q(t)$. The system is also being tracked by enemy radar through an observation equation

$$z_1 = H_1x + v_1, \tag{2}$$

where H_1 is a known continuous $p \times n$ time-varying matrix and $v_1(t)$ is a vector white gaussian random process with zero mean and covariance matrix $R_1(t)$. The system is also making measurements on its own state through a second observation equation

$$z_2 = H_2x + v_2, \tag{3}$$

where $v_2(t)$ is another white gaussian process with zero mean and covariance matrix $R_2(t)$ and H_2 similarly defined. It is desired to have the system fly a path based on measurements $z_2(t)$ and whatever a priori information such that it not only minimizes the terminal error but also, in some sense, maximizes the estimation error of the enemy radar. The radar, on the other hand, will attempt its best to reduce the estimation error of the missile's position and velocity. This is a problem of stochastic differential games about which very little is known.[1] In this note, we shall pose and solve the above class of problems associated with (1)–(3).

Additional assumptions. In order to solve these problems, it was found necessary to make some additional assumptions. We shall not pretend that these assumptions are completely realistic and that solutions derived from

* Received by the editors November 11, 1965, and in revised form April 25, 1966.

† Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts. This work was supported in part by the Joint Services Electronics Program (United States Army, United States Navy, and the United States Air Force) under Contract NONR-1866 (16) and in part by Aerospace Corporation, Los Angeles, California.

[1] In [1, Chap. 13] there are some speculations on the subject. Johanson [2] also solved a specific stochastic differential game for simple second-order dynamic systems.

them should be used in the actual design of terminal guidance systems. However, in an unknown subject area such as this, the first task is to get some insight into the problem. This will be furnished by our solution which is presented below. Such insight can then be used in guiding the actual design of the guidance system where various realistic constraints can be introduced.

We shall assume that the form of the guidance law will be

$$u = K(t)\hat{x}_2(t), \tag{4}$$

where $K(t)$, the time-varying gain to be determined, is constrained by

$$| K(t) | \leqq 1, \tag{5}$$

and $\hat{x}_2(t)$ is the conditional mean of the state $x(t)$ given the measurements $(z_2(\tau),\ t_0 < \tau \leqq t)$ and all other a priori information. It is well-known [3] that $\hat{x}_2(t)$ can be computed from

$$\dot{\hat{x}}_2 = F\hat{x}_2 + K_2(z_2 - H_2\hat{x}_2) + Gu \quad \text{with} \quad \hat{x}_2(t_0) \text{ given}, \tag{6}$$

$$K_2 = P_2 H_2{}^T R_2^{-1},$$

$$\dot{P}_2 = FP_2 + P_2F^T - K_2R_2K_2{}^T + Q \quad \text{with} \quad P(t_0) \text{ given}, \tag{7}$$

where

$$P_2 \equiv E[(x - \hat{x}_2)(x - \hat{x}_2)^T] \equiv \operatorname{cov}(\tilde{x}_2(t)). \tag{8}$$

Combining (1) and (6) results in

$$\dot{\tilde{x}}_2 = (F - K_2H_2)\tilde{x}_2 - K_2v_2 + w. \tag{9}$$

Combining (1), (4) and the definition of $\tilde{x}_2$ results in

$$\dot{x} = (F + GK)x - GK\tilde{x}_2 + w. \tag{10}$$

Thus, as far as the radar is concerned, the dynamical equations governing the situation can be written as

$$\dot{y} = \mathcal{F}y + \mathcal{G}s, \tag{11}$$

where

$$y = \begin{bmatrix} x \\ \tilde{x}_2 \end{bmatrix}, \qquad \mathcal{F} = \begin{bmatrix} F + GK & -GK \\ 0 & F - K_2H_2 \end{bmatrix}, \qquad \mathcal{G} = \begin{bmatrix} I & 0 \\ I & -K_2 \end{bmatrix},$$

$$s = \begin{bmatrix} w \\ v_2 \end{bmatrix},$$

i.e., a linear dynamic system driven by white noise.

It is appropriate at this point to assume that the radar will attempt to estimate the state of the missile through a filter

$$\dot{\hat{y}} = \mathfrak{F}\hat{y} + K_1(z_1 - H_1\hat{x}_1), \tag{12}$$

where $K_1(t)$ is a time-varying gain matrix to be determined, and where[2]

$$\hat{y} = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix}.$$

Finally, we shall assume this is a game of perfect information in the sense that both sides know the dynamics, the constraints, the specified uncertainties, and the assumed form of the controller and the filter.

While it could be argued that these assumptions are justified by considerations of engineering reality, the real reason will become obvious presently.

The solution. A reasonable criterion of performance for this situation can be represented as[3]

$$J = E\left\{\frac{a^2}{2} \| x(T) \|^2 - \frac{1}{2}\int_0^T \| x - \hat{x}_1 \|^2 \, dt\right\}. \tag{13}$$

The objectives of the missile and the radar are then the determination of $K(t)$ and $K_1(t)$ such that J is a saddle point, i.e., a minimax. Let us define

$$\text{Cov } (y(t)) \triangleq M(t) \triangleq \begin{bmatrix} M_{11} & M_{12} \\ M_{12} & M_{22} \end{bmatrix} \tag{14}$$

and

$$\text{cov } (y(t) - \hat{y}(t)) \triangleq \text{cov } (\tilde{y}(t)) \triangleq P(t) \triangleq \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{bmatrix}. \tag{15}$$

Then it is directly verified that

$$\dot{P} = (\mathfrak{F} - K_1H)P + P(\mathfrak{F} - K_1H)^T + K_1R_1K_1^T + \mathfrak{G}S\mathfrak{G}^T, \tag{16}$$

$$\dot{M} = \mathfrak{F}M + M\mathfrak{F}^T + \mathfrak{G}S\mathfrak{G}^T, \tag{17}$$

where

$$H = [H_1 \,\vdots\, 0], \qquad S = \begin{bmatrix} Q & 0 \\ 0 & R_2 \end{bmatrix}.$$

In terms of (16) and (17), the criterion J can be rewritten as

$$J = \frac{a^2}{2} \text{tr}(M_{11}(T)) - \frac{1}{2}\int_0^T \text{tr}(P_{11}(t)) \, dt. \tag{18}$$

[2] $\hat{x}_2$ is the estimate of x by the missile, $\hat{x}_1$ is the estimate of x by the radar, $\hat{\hat{x}}_2$ is the estimate of the missile estimation error by the radar.

[3] The minus sign in (13) accounts for the fact that radar really wishes to minimize its estimation error.

Now the problem can be stated as a completely deterministic problem of determining $K(t)$ and $K_1(t)$ such that J is minimized with respect to $K(t)$ and maximized with respect to $K_1(t)$ subject to the differential constraints (16) and (17). Note that $M(t)$ and $P(t)$ now play the role of state variables, $K(t)$ and $K_1(t)$, the control variables. We shall assume that a saddle point exists for the problem. Applying standard variational procedures, we define the Hamiltonian

$$(19)\quad \mathcal{H}(M, P, \Lambda_M, \Lambda_p, K, K_1, t) = -\tfrac{1}{2}\operatorname{tr}(P_{11}) + \operatorname{tr}(\Lambda_M \dot{M}) + \operatorname{tr}(\Lambda_p \dot{P}),$$

where

$$(21)\quad \dot{\Lambda}_M = -\mathcal{H}_M = -\Lambda_M F^T - F\Lambda, \qquad \Lambda_M(T) = \begin{bmatrix} \frac{a^2}{2} I & 0 \\ 0 & 0 \end{bmatrix},$$

$$(22)\quad \dot{\Lambda}_p = -\mathcal{H}_p = -\Lambda_p(F - K_1 H)^T - (F - K_1 H)\Lambda_p + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix},$$

$$\Lambda_p(T) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Then the necessary condition for minimax is

$$(23)\quad \mathcal{H}^0(M, P, \Lambda_M, \Lambda_p, t) = \min_{|K| \leqq 1} \max_{K_1} H(M, P, \Lambda_M, \Lambda_p, K, K_1, t).$$

Since K and K_1 appear separately in (16) and (17), it is clear that the order of minimization and maximization in (23) is immaterial. Setting $\partial H/\partial K_1 = 0$, we have

$$(24)\quad -HP\Lambda_p + R_1 K_1^T \Lambda_p = 0,$$

which implies[4]

$$(25a)\quad K_1 = (\Lambda_p^{\#}\Lambda_p) P H^T R_1^{-1},$$

which reduces to

$$(25b)\quad K_1 = PH^T R_1^{-1}$$

in the case where Λ_p^{-1} exists.[5] Substituting (25b) into (16), one finds,

[4] $\Lambda_p^{\#}$ is the generalized inverse of Λ_p.

[5] Since

$$\Lambda_p(t) = \int_t^T \Phi(t, T) \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \Phi^T(t, T)\, dt,$$

where Φ is the transition matrix associated with $(F - K_1 H)$, $\Lambda_p^{-1}(t)$ will fail to exist only when $\left((F - K_1 H), \begin{bmatrix} I \\ 0 \end{bmatrix}\right)$ fails to be a controllable pair.

$$\dot{P} = +FP + PF^T - PH^T R_1^{-1} HP + \mathcal{G}S\mathcal{G}^T. \tag{26}$$

In other words, the radar should use a Kalman-Bucy filter—an expected result in view of the fact that $z_1(t)$ is still gaussian.

Similarly, $\mathcal{H}$ can be minimized with respect to K. In this case no simple expression results, however.

Nevertheless (16), (17), (21), (22) and (23) now constitute a two-point boundary value problem. The initial values of $P(0)$ and $M(0)$ will, of course, have to be given or estimated. We shall not bother to discuss the implication and the numerical methods of solution of this problem. Instead, let us consider a very special case of the general problem and attempt to get a "feel" for the nature of such problems.

A special case. Let us consider the scalar case:

$$\dot{x} = u, \tag{27}$$

$$z_1 = x + v_1, \tag{28}$$

$$z_2 = x \text{ (perfect measurement)}, \tag{29}$$

i.e., $F = 0$, $H_1 = H_2 = G = 1$, and $a^2 = 1$. Then the problem simplifies to

$$\begin{aligned} \dot{X} &= 2KX; \qquad X(0) = x_0^2; \\ \dot{P}_A &= 2KP_A - \frac{P_A^2}{R_1}; \qquad P_A(0) = P_0; \\ \dot{\Lambda}_x &= -2K\Lambda_x, \qquad \Lambda_x(T) = \frac{1}{2}; \\ \dot{\Lambda}_P &= -2K\Lambda_P + \frac{2P_A}{R_1}\Lambda_P + \frac{1}{2}, \qquad \Lambda_P(T) = 0; \\ K &= -\operatorname{sgn}(\Lambda_x X + \Lambda_P P_A). \end{aligned} \tag{30}$$

Examination of (30) immediately reveals that $X(t)$, $P_A(t)$, and $\Lambda_x(t)$ are always positive and $\Lambda_P(t)$ is always negative. Thus, the term $\Lambda_x X + \Lambda_P P_A$ can have at most one change of sign from negative to positive.[6] Hence, the controlled system behaves in two possible modes: (i) starts as an extremely unstable system ($K = +1$) and then switches to an extremely stable system ($K = -1$); (ii) operates always as an extremely stable system. This is an intuitively reasonable solution. In case (i), the system initially devotes entire effort to confound the enemy radar since small estimation error in an unstable system grows exponentially with time. However, as time goes on, the system must pay increasing attention to the

[6] It can be directly verified that the singular case where $\Lambda_x X + \Lambda_P P_A \equiv 0$ cannot be sustained.

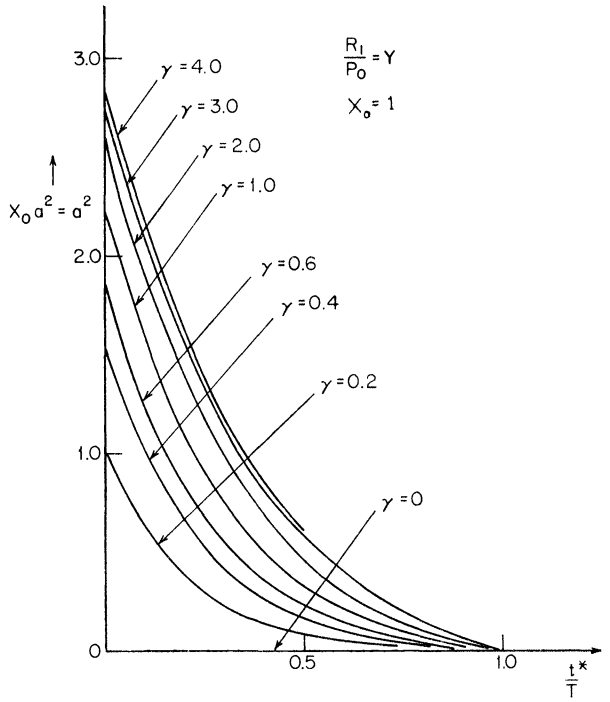

FIG. 1

objective of minimizing the expected terminal error and to forget the enemy radar, i.e., to behave like a stable system in this case. The term $\Lambda_x X + \Lambda_P P_A$ makes precise the instant at which this switching takes place. In case (ii), the initial errors are so large and the interval of control is so short as to make any attempt to confuse the enemy unprofitable. The entire effort must be devoted to minimizing the terminal error.

Because of the simplicity of the form of $K(t)$ and the associated differential equations of the two point boundary value problem, we can actually express the solution of the problem in terms of the following transcendental equation:

$$X_0 a^2 = \frac{R_1(\alpha_2 - 1)(e^{2(T-t^*)} - 1)}{\alpha_2 e^{2(T-t^*)} - 1} \left\{\frac{\alpha_1 e^{2(T-t^*)}}{\alpha_1 e^{2t^*} - 1}\right\}, \tag{31}$$

where

$$\alpha_2 = \frac{P_A(t^*) + 2R_1}{P_A(t^*)}, \qquad \alpha_1 = \frac{P_0}{P_0 - 2R_1},$$

t^* is the switching time, R_1 the variance of v_1, and $X_0 = x_0^2$.

A parametric study of a^2, the weighting factor, vs. t^*, the switching time, with the parameter $\gamma = R_1/P_0$ and x_0 normalized to one, is shown in Fig. 1. The reasonableness of the result is obvious.[7]

Discussion. From the above analysis, the general guideline for the controller design is clear. For evasive maneuver, the missile should behave as a random process (corresponding to a mixed strategy). The objective of the design is to force the system to behave as that particular random process which maximizes the estimation error. On the other hand, this must be balanced with the conflicting objective of minimizing the terminal miss criterion which requires the reduction of uncertainty of the random process. The choice of the form of the controller in (4) restricts the admissible class of random process to gaussian. This considerably simplifies the solution. However, if we had allowed a larger class of random processes for the behavior of the system, then the computational problem becomes vastly more complicated even though conceptually it is not any more so.

It is to be noted that the number of differential equations that we have to contend with even in this restricted class is $2n^2$ where n is the number of original system equations. This is a considerable increase in computational burden. However, it is the price one has to pay for a problem of this nature.

Note the solution $K(t)$ also furnishes the minimax filter that should be employed by the enemy radar for optimal estimation.

Lastly, from the above discussion a modus operandi for a more general class of problem can be contemplated:

(i) Given the physical limitation of the estimator, say (2), and the class of random processes admissible, determine the particular random process which gives a minimax solution for the criterion of performance.

(ii) Given the physical limitation of the system, say (1) and (3), determine the control law which realizes or approximates the particular random process in (i).

It is reasonably clear that such decomposition of the problem can lead to considerable easing of the computational burden.

Conclusion. The problem of stochastic differential games is still in its infancy. The above analysis suggests the nature of the problem and the more or less obvious extensions. A simple extension would be to consider instead of (4):

$$(4') \qquad u = K(t)\hat{x}_2 + \omega,$$

where $\omega(t)$ is another gaussian random process with controllable parameters. The problem can still be solved. We shall omit the details. It is hoped that this work will stimulate further researches in this area.

[7] The author is indebted to J. Daniels for the calculation connected with Fig. 1.

Acknowledgment. The author would like to acknowledge some helpful suggestions by Dr. H. S. Witsenhausen.

REFERENCES

[1] R. Isaacs, *Differential Games*, John Wiley, New York, 1965.

[2] D. Johanson, *Solution of a mean square estimation problem when process statistics are undefined*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 20–29.

[3] R. E. Kalman and R. C. Bucy, *New results in linear filtering and prediction theory*, Trans. ASME Ser. D. J. Basic Engrg., 83D (1961), pp. 95–108.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

FINITE-TURN PUSHDOWN AUTOMATA*

SEYMOUR GINSBURG† AND EDWIN H. SPANIER‡

Introduction. As is well-known, the context free languages are excellent approximations to the syntactic components of currently used programming languages (such as ALGOL). Pushdown automata (abbreviated pda) are devices used in parsing programming languages, for the most part in compiling. These two concepts are linked by the result that a set of words is a context free language if and only if it is accepted, i.e., recognized, by some pda.

To implement the construction (either by hardware or by software) and the usage of pda, it is important to have general classes of pda with particular properties. For example, it is convenient to discuss deterministic pda since they parse rapidly. In the same spirit, we investigate the class of pda having the property that the length of the pushdown tape alternatively increases and decreases at most a fixed bounded number of times during any sweep of the automaton. Such pda reject words faster than an arbitrary pda. (For a particular sweep can be halted as soon as it exceeds in number of alternations the fixed bound.)

The present paper is a study of these "finite-turn" pda and the languages they recognize. These languages are characterized both in terms of grammars and in terms of generation from finite sets by three operations. The languages turn out to coincide with the nonterminal bounded languages, a class of languages studied in another context.

After a first section on preliminary definitions, these three concepts are considered and their equivalence established in §§2, 3, and 4. §5 is concerned with decidability questions. In particular, a decision procedure is given for determining whether an arbitrary pda is a finite-turn pda. It is also proved that there is no decision procedure for determining whether an arbitrary language is accepted by some finite-turn pda. §6 contains a discussion of one-turn pda and their relationship to the class of linear languages.

1. Preliminaries. We now consider the basic concepts to be used. In particular, we define context free languages, pushdown automata, and f-transducers.

* Received by the editors December 20, 1965. This research was sponsored in part by the Air Force Cambridge Research Laboratories, Office of Aerospace Research, under Contract AF 19(628)-5166, CRL—Algorithmic Language Program.

† Research and Technology Division, System Development Corporation, Santa Monica, California.

‡ Department of Mathematics, University of California, Berkeley, California.

DEFINITION. For sets of words X and Y, $XY = \{xy \mid x \text{ in } X, y \text{ in } Y\}$, where xy denotes the concatenation of x and y. XY is called the *product* of X and Y. Let $X^0 = \{\epsilon\}$, where ϵ is the empty word, $X^{i+1} = X^iX$, and $X^* = \bigcup_{i=0}^{\infty} X^i$. Thus, for an arbitrary set E of symbols, E^* is the free semigroup with identity generated by E.

DEFINITION. A *context free grammar* (abbreviated *grammar*) is a 4-tuple $G = (V, \Sigma, P, \sigma)$, where

(i) V is a finite nonempty set,

(ii) Σ is a nonempty subset of V,

(iii) P is a finite nonempty set of pairs (ξ, v), with ξ in $V - \Sigma$ and v in V^*,

(iv) σ is an element of $V - \Sigma$.

Each element of $V - \Sigma$ is called a *variable.*

Each element of Σ is called a (*terminal*) letter.

Each element (ξ, v) in P is called a *production* (or *rewriting rule*) and is written $\xi \to v$.

Notation. Let $G = (V, \Sigma, P, \sigma)$ be a grammar. For w_1 and w_2 in V^* write $w_1 \Rightarrow w_2$ if there exist u_1, u_2, ξ, v such that $w_1 = u_1\xi u_2$, $w_2 = u_1vu_2$, and $\xi \to v$ is in P. For w and y in V^* write $w \Rightarrow^* y$ if either $w = y$ or there exist $w_0 = w, w_1, \cdots, w_k = y$ such that $w_i \Rightarrow w_{i+1}$ for each i.

A sequence of words $w_0, \cdots, w_k$ such that $w_i \Rightarrow w_{i+1}$ for each i is called a *derivation* or *generation* of w_k (from w_0) and is denoted by

$$w_0 \Rightarrow \cdots \Rightarrow w_k .$$

DEFINITION. $L \subseteq \Sigma^*$ is a *context free language* (abbreviated *language*) if there exists a grammar $G = (V, \Sigma, P, \sigma)$ such that $L = L(G)$, where $L(G) = \{w \in \Sigma^* \mid \sigma \Rightarrow^* w\}$. $L(G)$ is said to be the language *generated* by G.

It is well-known that the context free languages are excellent approximations to the syntactic classes of most currently used programming languages. As such, the mathematics of context free languages is being extensively studied [9].

We now define a type of device which is closely associated with languages, both theoretically (in characterizing languages) and practically (in the compilation procedure of data processing).

DEFINITION. A *pushdown automaton* (abbreviated *pda*) is a 6-tuple $M = (K, \Sigma, \Gamma, \delta, Z_0, q_0)$, where

(i) K is a nonempty finite set (of *states*),

(ii) Σ is a nonempty finite set (of *inputs*),

(iii) Γ is a finite nonempty set (of *pushdown* symbols),

(iv) δ is a mapping from $K \times (\Sigma \cup \{\epsilon\}) \times \Gamma$ to the finite subsets of $K \times \Gamma^*$,

(v) Z_0 is an element of Γ,

(vi) q_0 is in K (the *start* state).

Notation. Given a pda $M = (K, \Sigma, \Gamma, \delta, Z_0, q_0)$, let $\vdash^*$ be the relation on $K \times \Sigma^* \times \Gamma^*$, defined as follows. For Z in Γ and x in $\Sigma \cup \{\epsilon\}$ let $(p, xw, \alpha Z) \vdash (q, w, \alpha\gamma)$, called a *move*, if $\delta(p, x, Z)$ contains (q, γ). Let $(p, w, \alpha) \vdash^* (p, w, \alpha)$ for all p, w, α. For α, β in Γ^* and x_i in $\Sigma \cup \{\epsilon\}$, $1 \leqq i \leqq k$, let $(p, x_1 \cdots x_k w, \alpha) \vdash^* (q, w, \beta)$ if there exist $p_1 = p, \cdots, p_{k+1} = q$ in K and $\alpha_1 = \alpha, \cdots, \alpha_{k+1} = \beta$ in Γ^* such that

$$(p_i, x_i \cdots x_k w, \alpha_i) \vdash (p_{i+1}, x_{i+1} \cdots x_k w, \alpha_{i+1}) \qquad \text{for} \quad 1 \leqq i \leqq k.$$

Notation. Given a pda $M = (K, \Sigma, \Gamma, \delta, Z_0, q_0)$, let

$$\text{Null}(M) = \{w \in \Sigma^* \mid (q_0, w, Z_0) \vdash^* (q, \epsilon, \epsilon) \text{ for some } q \text{ in } K\}.$$

The fundamental connection between languages and pda is the following result [6]: A set $L \subseteq \Sigma^*$ is a language if and only if $L = \text{Null}(M)$ for some pda M.

Remark. If $(K, \Sigma, \Gamma, \delta, Z_0, q_0)$ is a pda, then $(K, \Sigma, \Gamma, \delta, Z_0, q_0, F)$, with $F \subseteq K$, is called a *pda with final states.* A pda with final states *accepts* a word w if $(q_0, w, Z_0) \vdash^* (q, \epsilon, \gamma)$ for some q in F and some γ in Γ^*. Pda with final states accept exactly the family of languages [9].

The practical significance of pda is that many algorithms used in the computer literature for recognizing programs (part of the compiling procedure) are programming implementations of specific pda [13].

We shall have occasion to use the notion of a "sequential transducer with final states". (This notion is discussed in [8] and called a *binary nondeterministic automaton.* We change the name to emphasize that the device is used to transform input words to output words. In fact, the relation of input words to output words is called a *binary transduction* in [8].)

DEFINITION. A *sequential transducer with final states* (abbreviated *f-transducer*) is a 6-tuple $S = (K, \Sigma, \Delta, H, s_0, F)$, where

(i) K is a finite nonempty set (of *states*),
(ii) Σ is a finite nonempty set (of *inputs*),
(iii) Δ is a finite nonempty set (of *outputs*),
(iv) s_0 is in K (the *start* state),
(v) $F \subseteq K$ (the set of *final* states),
(vi) H is a finite subset of $K \times \Sigma^* \times \Delta^* \times K$.

Remark. It is no loss of generality to assume that the f-transducer has the same set of inputs and outputs since we may regard both as being $\Sigma \cup \Delta$. We shall use this fact later.

An element (p, u, v, q) in H denotes the fact that applying an input word u to the f-transducer at state p results in an output word v and a next state q.

DEFINITION. Let $S = (K, \Sigma, \Delta, H, s_0, F)$ be an f-transducer. Write $(p, ux, y) \vdash (q, x, yv)$ if (p, u, v, q) is in H. Write $(p, u, y) \vdash^* (p, u, y)$ for all p, u, y. Write $(p, ux, y) \vdash^* (q, x, yv)$ if there exist $p_0 = p, p_1, \cdots, p_k = q$

in K, $u_1, \cdots, u_k$ in Σ^*, $v_1, \cdots, v_k$ in Δ^* such that $u = u_1 \cdots u_k$, $v = v_1 \cdots v_k$, and

$$(p_0, ux, y) \vdash (p_1, u_2 \cdots u_k x, yv_1) \vdash \cdots \vdash (p_k, x, yv_1 \cdots v_k).$$

For each word u, let $S_f(u) = \{v \mid (s_0, u, \epsilon) \vdash^* (q, \epsilon, v) \text{ for some } q \text{ in } F\}$.

DEFINITION. Let Σ be an abstract set. Let $\epsilon^R = \epsilon$ and for $x = x_1 \cdots x_k$, each x_i in Σ, let $x^R = x_k \cdots x_1$.

If Σ contains at least two elements, then there is no f-transducer $S = (K, \Sigma, \Delta, H, s_0, F)$ such that $S_f(w) = w^R$ for each w in Σ^*.

DEFINITION. A 1-*restricted f-transducer* is an f-transducer $S = (K, \Sigma, \Delta, H, s_0, F)$ such that $H \subseteq K \times (\Sigma \cup \{\epsilon\}) \times (\Delta \cup \{\epsilon\}) \times K$.

LEMMA 1.1. *For each f-transducer* $S = (K, \Sigma, \Delta, H, s_0, F)$ *there exists a* 1*-restricted f-transducer* $S' = (K', \Sigma, \Delta, H', s_0, F)$ *such that* $S_f(w) = S_f'(w)$ *for each* w.

Proof. Let $H = \{h_i \mid 1 \leqq i \leqq m\}$, where $h_i = (q_i, x_{i1} \cdots x_{ir(i)}, y_{i1} \cdots y_{ir(i)}, q_i')$, each x_{ij} is in $\Sigma \cup \{\epsilon\}$, each y_{ij} is in $\Delta \cup \{\epsilon\}$, for $1 \leqq i \leqq m$. For $1 \leqq j \leqq r(i) - 1$ and $1 \leqq i \leqq m$, let p_{ij} be abstract symbols not in K and let $K' = K \cup \{p_{ij} \mid i, j\}$. Let

$$H' = \bigcup_i \{(q_i, x_{i1}, y_{i1}, p_{i1}), (p_{i1}, x_{i2}, y_{i2}, p_{i2}), \cdots, (p_{ir(i)-1}, x_{ir(i)}, y_{ir(i)}, q_i')\}.$$

Clearly $S_f'(w) = S_f(w)$ for all w.

The following result on the composite of f-transducers was first proved in [8] by an involved sequence of propositions. A short (and different) proof is presented here.

LEMMA 1.2. *If* $S = (K, \Sigma, \Delta, H, s_0, F)$ *and* $S' = (K', \Delta', \Delta'', H', s_0', F')$ *are f-transducers, with* $\Delta \subseteq \Delta'$, *then there exists an f-transducer* $S'' = (K'', \Sigma, \Delta'', H'', s_0'', F'')$ *such that* $S_f'' = S_f' S_f$.

Proof. By Lemma 1.1, there exist 1-restricted f-transducers $T = (K_1, \Sigma, \Delta, H_1, s_0, F)$ and $T' = (K_1', \Delta', \Delta'', H_1', s_0', F')$ such that $S_f = T_f$ and $S_f' = T_f'$. Let H_2 be the union of H_1 and all quadruples $(q, \epsilon, \epsilon, q)$, q in K_1. Then $U = (K_1, \Sigma, \Delta, H_2, s_0, F)$ is a 1-restricted f-transducer and $U_f = T_f = S_f$. Similarly let H_2' be the union of H_1' and all quadruples $(q, \epsilon, \epsilon, q)$, q in K_1'. Then $U' = (K_1', \Delta', \Delta'', H_2', s_0', F')$ is a 1-restricted f-transducer and $U_f' = T_f' = S_f'$.

Now let $K'' = K_1 \times K_1'$, $s_0'' = (s_0, s_0')$, $F'' = F \times F'$ and H' be the set of all quadruples $((q, q'), x, y, (p, p'))$, x in $\Sigma \cup \{\epsilon\}$ and y in $\Delta'' \cup \{\epsilon\}$, such that an element z in $\Delta \cup \{\epsilon\}$ can be found satisfying (q, x, z, p) in H_2 and (q', z, y, p') in H_2'. Then $S'' = (K'', \Sigma, \Delta'', H'', s_0'', F'')$ is an f-transducer such that $S_f'' = S_f' S_f$.

DEFINITION. The *inverse* S^{-1} of an f-transducer $S = (K, \Sigma, \Delta, H, s_0, F)$ is the f-transducer $(K, \Delta, \Sigma, H^{-1}, s_0, F)$, where (q, w, w', q') is in H^{-1} if and only if (q, w', w, q') is in H.

Clearly $S_f(w)$ contains w' if and only if $S_f^{-1}(w')$ contains w.

The inverse of a 1-restricted f-transducer is a 1-restricted f-transducer.

We now present a result on f-transducers which is similar to a result on linear languages due to Chomsky and Schutzenberger [7]. We shall assume that the reader is familiar with the basic notions of regular set, automaton, and nondeterministic automaton as discussed, for example, in [14].

THEOREM 1.1. *Let $S = (K, \Sigma, \Delta, H, s_0, F)$ be an f-transducer. Then there exist a finite set Σ', a regular set $U \subseteq \Sigma'^*$, and homomorphisms τ, of Σ'^* into Σ^*, and τ', of Σ'^* into Δ^*, with the following property: $S_f(w)$ contains w' if and only if there exists w'' in U such that $\tau(w'') = w$ and $\tau'(w'') = w'$.*

Proof. Let $H = \{h_i | 1 \leqq i \leqq n\}$, with $h_i = (p_i, \alpha_i, \beta_i, q_i)$, and let Σ' consist of n abstract symbols $a_1, \cdots, a_n$. Let A be the nondeterministic automaton $(K, \Sigma', \delta, s_0, F)$, where $\delta(p_i, a_i) = \{(q_i)\}$ for $1 \leqq i \leqq n$. Let $U = T(A)$, where, for each automaton or nondeterministic automaton A, $T(A)$ denotes the set of words accepted by A. Let τ and τ' be the homomorphisms from Σ'^* to Σ^* and Σ'^* to Δ^*, respectively, defined by $\tau(a_i) = \alpha_i$ and $\tau'(a_i) = \beta_i$, $1 \leqq i \leqq n$. Clearly Σ', U, τ, and τ' have the desired properties.

The next result, a proof of which is in [8], asserts that the operations of word reversal and transduction commute in a certain sense.

LEMMA 1.3. *For each f-transducer $S = (K, \Sigma, \Delta, H, s_0, F)$ there exists an f-transducer $S' = (K', \Sigma, \Delta, H', s_0', F')$ such that w' is in $S_f'(w)$ if and only if $(w')^R$ is in $S_f(w^R)$, i.e., $[S_f'(w)]^R = S_f(w^R)$.*

COROLLARY. *Given f-transducers $S = (K, \Sigma, \Delta, H, s_0, F)$ and $S' = (K', \Sigma, \Delta, H', s_0', F')$, there exists an f-transducer $S'' = (K'', \Delta, \Delta, H'', s_0'', F'')$ such that*

$$\bigcup_{w \in \Sigma^*} S_f(w) S_f'(w^R) = \bigcup_{w' \in \Delta^*} w' S_f''[(w')^R].$$

Proof. Let S^{-1} be the inverse of S. Clearly

$$\bigcup_{w \in \Sigma^*} S_f(w) S_f'(w^R) = \bigcup_{w' \in \Delta^*} w' S_f'([S_f^{-1}(w')]^R).$$

By Lemma 1.3, there exists an f-transducer T such that $T_f((w')^R) = [S_f^{-1}(w')]^R$. By Lemma 1.1, there exists an f-transducer S'' such that $S_f'' = S_f' T_f$. Then S'' has the desired property.

2. Finite-turn pda.

DEFINITION. If the pda $(K, \Sigma, \Gamma, \delta, Z_0, q_0)$ is in the configuration (q, w, γ), with q in K, w in Σ^*, and γ in Γ^*, then γ is called the *pushdown tape*.

DEFINITION. A *sweep* of a pda$(K, \Sigma, \Gamma, \delta, Z_0, q_0)$ is a sequence of moves $(q_0, x_0 \cdots x_k, Z_0) \vdash (q_1, x_1 \cdots x_k, \gamma_1) \vdash \cdots \vdash (q_{k+1}, \epsilon, \gamma_{k+1})$, with $\gamma_{k+1} = \epsilon$.

As mentioned in the Introduction, we are concerned with those pda having an integer m such that the length of the pushdown tape for each sweep changes direction at most m times. Such a pda is of interest since a sequence of moves involved in recognition can be halted whenever the number of alternations of the length of the pushdown tape exceeds m. In this section we formally define such a pda M. We then give an explicit grammar G such that $\text{Null}(M) = L(G)$.

DEFINITION. Given a pda $M = (K, \Sigma, \Gamma, \delta, Z_0, q_0)$, a move $(p, xw, \alpha Z) \vdash (q, w, \alpha\gamma)$ is said to be *nondecreasing* (*nonincreasing*, *increasing*, *decreasing*) if $|\gamma| \geqq 1 (|\gamma| \leqq 1, |\gamma| > 1, |\gamma| = 0)$.

DEFINITION. Let $(q_0, x_0 \cdots x_k, \gamma_0) \vdash (q_1, x_1 \cdots x_k, \gamma_1) \vdash \cdots \vdash (q_{k+1}, \epsilon, \gamma_{k+1})$ be a sweep. The length† at $(q_0, x_0 \cdots x_k, \gamma_0)$ is said to be *increasing*. By induction, if the length at $(q_i, x_i \cdots x_k, \gamma_i)$ is increasing and the move $(q_i, x_i \cdots x_k, \gamma_i) \vdash (q_{i+1}, x_{i+1} \cdots x_k, \gamma_{i+1})$ is nondecreasing (decreasing), then the length at $(q_{i+1}, x_{i+1} \cdots x_k, \gamma_{i+1})$ is said to be *increasing* (*decreasing*). If the length at $(q_i, x_i \cdots x_k, \gamma_i)$ is decreasing and the move $(q_i, x_i \cdots x_k, \gamma_i) \vdash (q_{i+1}, x_{i+1} \cdots x_k, \gamma_{i+1})$ is nonincreasing (increasing), then the length at $(q_{i+1}, x_{i+1} \cdots x_k, \gamma_{i+1})$ is said to be *decreasing* (*increasing*).

DEFINITION. If the length at $(q_i, x_i \cdots x_k, \gamma_i)$ is increasing (decreasing) at $(q_i, x_i \cdots x_k, \gamma_i)$ and decreasing (increasing) at $(q_{i+1}, x_{i+1} \cdots x_k, \gamma_{i+1})$, then the length is said to have a *turn* at $(q_i, x_i \cdots x_k, \gamma_i)$.

DEFINITION. A sweep is said to be a $(2k-1)$-*turn* sweep if the length has exactly $2k-1$ turns. A pda is said to be a $(2k-1)$-*turn* pda if every sweep has at most $2k-1$ turns. A pda is said to be *finite-turn* if it is $(2k-1)$-turn for some $k \geqq 1$.

Notation. For each pda M and $k \geqq 1$, let $\text{Null}_{2k-1}(M)$ be the set of those input words accepted by some sweep with at most $2k-1$ turns.

Thus $\text{Null}(M) = \text{Null}_{2k-1}(M)$ if M is a $(2k-1)$-turn pda.

LEMMA 2.1. *For each pda M and each $k \geqq 1$ there is a $(2k-1)$-turn pda M' such that* $\text{Null}(M') = \text{Null}_{2k-1}(M)$.

Proof. Let $M = (K, \Sigma, \Gamma, \delta, Z_0, q_0)$. Let $M' = (K', \Sigma, \Gamma, \delta', Z_0, q_0')$, where $K' = K \times \{1, 2, \cdots, 2k\}$, $q_0' = (q_0, 1)$, and δ' is defined as follows:

(a) $\delta'((q, i), x, Z)$ contains $((p, i), \gamma)$ if and only if (p, γ) is in $\delta(q, x, Z)$, i is odd (even), and $|\gamma| \geqq 1$ $(|\gamma| \leqq 1)$;

(b) $\delta'((q, 2i-1), x, Z)$ contains $((p, 2i), \epsilon)$ if and only if (p, ϵ) is in $\delta(q, x, Z)$;

(c) $\delta'((q, 2i), x, Z)$ contains $((p, 2i+1), \gamma)$ if and only if (p, γ) is in $\delta(q, x, Z)$ and $|\gamma| \geqq 2$.

† For brevity, we frequently write "length" instead of "length of the pushdown tape" if no confusion arises.

Then K' is the union of the disjoint subsets $K \times \{i\}$, $1 \leqq i \leqq 2k$. Any move from a state in $K \times \{i\}$ to a state in $K \times \{i\}$ is nondecreasing if i is odd and nonincreasing if i is even. The only other moves in M' are decreasing moves from a state of $K \times \{2i - 1\}$ to a state of $K \times \{2i\}$, and increasing moves from a state of $K \times \{2i\}$ to a state of $K \times \{2i + 1\}$. Thus each turn in a sweep of M' corresponds to a move from a state of the form (q, i) to a state of the form $(p, i + 1)$. Therefore M' is a $(2k - 1)$-turn pda.

The construction of M' is such that any $(2i - 1)$-turn sweep of M, $1 \leqq i \leqq k$, corresponds to a sweep of M', and conversely. Therefore $\text{Null}_{2k-1}(M) = \text{Null}(M')$.

COROLLARY. $\text{Null}_{2k-1}(M)$ *is a language for each pda* M.

Remark. Let M be a pda with final states. For any m it can also be shown that the set of tapes accepted by M by a sweep with at most m turns is a language. Furthermore, the family of languages obtained by allowing M and m to vary is identical with the family of languages defined in the corollary by allowing M and k to vary.

Given a $(2k - 1)$-turn pda $M = (K, \Sigma, \Gamma, \delta, Z_0, q_0)$ we now construct an explicit grammar $G = (V, \Sigma, P, \sigma)$ such that $L(G) = \text{Null}(M)$. We may assume that M has the form of the lemma. Thus K is partitioned into disjoint subsets $K_1, \cdots, K_{2k}$ with q_0 in K_1. Also, every move from a state of K_{2i-1} is either a nondecreasing move to a state of K_{2i-1} or a decreasing move to a state of K_{2i}, and every move from a state of K_{2i} is either a nonincreasing move to a state of K_{2i} or an increasing move to a state of K_{2i+1}. We may also assume that M has the special form that $\delta(q, x, Z)$ contains only pairs of the form (q', γ), where $|\gamma| \leqq 2$. This is no loss of generality, for it is easily seen that given any pda M we can adjoin additional states to M to obtain a pda M' of the special form such that (i) $\text{Null}(M) = \text{Null}(M')$, and (ii) if M is a $(2k - 1)$-turn pda then so is M'.

For each state q in M let $h(q) = i$, where q is in K_i. (Thus $h(q_0) = 1$ and $h(q) \leqq 2k$ for every q.) Let

$$V = \{\sigma\} \cup \Sigma \cup \Big\{ \bigcup_{1 \leqq i < j \leqq 2k} (K_i \times \Gamma \times K_j \times (\Gamma \cup \{\epsilon\})) \Big\}.$$

The symbol $[q, Z, q', Y]$ is to denote a quadruple consisting of a state q in K_i, a symbol Z in Γ, a state q' in K_j with $j > i$, and an element Y of $\Gamma \cup \{\epsilon\}$. (Z and Y, with or without subscripts or superscripts, will denote an element of Γ and $\Gamma \cup \{\epsilon\}$, respectively. Similarly x will denote an element of $\Sigma \cup \{\epsilon\}$.) The variable $[q, Z, q', Y]$ is to have the property that the language generated from it as start variable, denoted by $L_{[q,Z,q',Y]}$, is the set of all words w in Σ^* such that $(q, w, Z) \vdash^* (q', \epsilon, Y)$. This interpreta-

tion motivates the definition of the set P of productions. P consists of the following productions.

(1) $\sigma \to [q_0, Z_0, q', \epsilon]$.
(2) $[q, Z, q', \epsilon] \to x$ if (q', ϵ) is in $\delta(q, x, Z)$.
(3) $[q, Z, q', Y] \to x[q_1, Z_1, q', Y]$ if (q_1, Z_1) is in $\delta(q, x, Z)$.
(4) $[q, Z, q', Z'] \to [q, Z, q'', Z'']x'$ if (q', Z') is in $\delta(q'', x', Z'')$.
(5) $[q, Z, q', Z'] \to x[q_1, Z_1, q', \epsilon]$ if $(q_1, Z'Z_1)$ is in $\delta(q, x, Z)$.
(6) $[q, Z, q', \epsilon] \to [q, Z, q'', Z'']x'$ if (q', ϵ) is in $\delta(q'', x', Z'')$.
(7) $[q, Z, q', Y] \to [q, Z, q_1, Z_1][q_1, Z_1, q', Y]$ for all q_1 such that $h(q) < h(q_1) < h(q')$.

If $G = (V, \Sigma, P, \sigma)$ is a grammar and $\xi \to w$ (w in Σ^*) is in P, then $\xi \to w$ is called a *terminal* production.

Intuitively, (1) starts to get all words in Null(M). Production (2) is a terminal production corresponding to a decreasing move of the pda. It applies only when $h(q)$ is odd and $h(q') - h(q) = 1$. Productions (3) and (4) correspond to length-preserving moves of the pda. Production (5) corresponds to an increasing move of the pda. It can only occur when $h(q_1)$ is odd and $h(q_1) < h(q')$. Similarly (6) corresponds to a decreasing move of the pda. It can only apply when $h(q')$ is even and $h(q) < h(q'')$. Production (7) corresponds to a sequence of moves of the pda such that there is a state q_1, $h(q) < h(q_1) < h(q')$, at which the length of the push-down tape is exactly one. A production of type (7) can only occur if $k > 1$. (If the pda is a one-turn pda, then only productions (1)–(6) are needed.)

We shall show that $L(G) = \text{Null}(M)$. Using production (1), this will follow if we show that

(8) $L_{[q,Z,q',Y]} = \{w \text{ in } \Sigma^* \mid (q, w, Z) \vdash^* (q', \epsilon, Y)\}$.

First assume that $(q, w, Z) \vdash^* (q', \epsilon, Y)$. Then there exist $x_1, \cdots, x_m$ in $\Sigma \cup \{\epsilon\}$, $q, q_1, \cdots, q_m = q'$ in K, and $Z, \alpha_1, \cdots, \alpha_m = Y$ in Γ^* such that $w = x_1 \cdots x_m$ and

$$(q, x_1 \cdots x_m, Z) \vdash (q_1, x_2 \cdots x_m, \alpha_1) \vdash \cdots \vdash (q_m, \epsilon, \alpha_m).$$

We shall show that w is in $L_{[q,Z,q',\epsilon]}$, i.e., $[q, Z, q', \epsilon] \Rightarrow^* w$. Suppose that $m = 1$. Then $(q, x_1, Z) \vdash (q', \epsilon, Y)$ is a decreasing move since $h(q) < h(q')$, so that $Y = \epsilon$. Then $[q, Z, q', \epsilon] \to x_1$ is a production of type (2). Continuing by induction suppose that $m \geqq 2$. There are four cases to consider.

(a) $|\alpha_1| = 1$. Then $h(q_1) = h(q)$, so that $h(q_1) < h(q')$. By induction, $[q_1, \alpha_1, q', Y] \Rightarrow^* x_2 \cdots x_m$. Also, $[q, Z, q', Y] \to x_1[q_1, \alpha_1, q', Y]$ is a production of type (3). Then $[q, Z, q', Y] \Rightarrow^* x_1 \cdots x_m$.

(b) $|\alpha_1| > 1$ and $|\alpha_{m-1}| = 1$. Then a decreasing move must have occurred before the last and so $h(q) < h(q_{m-1})$. By induction, $[q, Z, q_{m-1},$

$\alpha_{m-1}] \Rightarrow^* x_1 \cdots x_{m-1}$. Also, $[q, Z, q', Y] \to [q, Z, q_{m-1}, \alpha_{m-1}]x_m$ is a production of type (4) if Y is in Γ and of type (6) if $Y = \epsilon$. Thus $[q, Z, q', Y] \Rightarrow^* x_1 \cdots x_m$.

(c) $|\alpha_1| > 1$, $|\alpha_{m-1}| > 1$, and $|\alpha_i| = 1$ for some i, $1 < i < m - 1$. Then $(q, x_1 \cdots x_i, Z) \vdash (q_1, x_2 \cdots x_i, \alpha_1)$ is an increasing move and, since $|\alpha_i| = 1$, $(q_j, x_j \cdots x_i, \alpha_j) \vdash (q_{j+1}, x_{j+1} \cdots x_i, \alpha_{j+1})$ is a decreasing move for some j, $2 \leqq j \leqq i - 1$. Therefore $h(q) < h(q_i)$. By induction, $[q, Z, q_i, \alpha_i] \Rightarrow^* x_1 \cdots x_i$. Similarly $h(q_i) < h(q')$ and $[q_i, \alpha_i, q', Y] \Rightarrow^* x_{i+1} \cdots x_m$. Since

$$[q, Z, q', Y] \to [q, Z, q_i, \alpha_i][q_i, \alpha_i, q', Y]$$

is a production of type (7), $[q, Z, q', Y] \Rightarrow^* x_1 \cdots x_m$.

(d) $|\alpha_i| > 1$ for all i, $1 \leqq i \leqq m - 1$. Then $Y = Z'$ and $\alpha_1 = Z'Z''$ for some Z', Z'' in Γ. Also, $h(q_1) < h(q')$ and $(q_1, x_2 \cdots x_m, Z'') \vdash^* (q', \epsilon, \epsilon)$. By induction, $[q_1, Z'', q', \epsilon] \Rightarrow^* x_2 \cdots x_m$. Since $[q, Z, q', Z'] \to x_1[q_1, Z'', q', \epsilon]$ is a production of type (5), $[q, Z, q', Z'] \Rightarrow^* x_1 \cdots x_m$.

To complete the proof of (8), we now prove that for $h(q) < h(q')$ and w in $L_{[q,Z,q',Y]}$, $(q, w, Z) \vdash^* (q', \epsilon, Y)$. Let

$$[q, Z, q', Y] \Rightarrow w_1 \Rightarrow \cdots \Rightarrow w_m$$

be a derivation of $w_m = w$. Suppose $m = 1$. Then $[q, Z, q', Y] \to w$ is a terminal production and thus of type (2). Also, $Y = \epsilon$, w is in $\Sigma \cup \{\epsilon\}$, and (q', ϵ) is in $\delta(q, w, Z)$. Then $(q, w, Z) \vdash^* (q', \epsilon, \epsilon) = (q', \epsilon, Y)$. Continuing by induction, suppose $m \geqq 2$. Then $[q, Z, q', Y] \to w_1$ is a production of type (3), (4), (5), (6), or (7). Assume it is of type (3). Then $w_1 = x[q_1, Z_1, q', Y]$ and $w = xw'$, where (q_1, Z_1) is in $\delta(q, x, Z)$ and $[q_1, Z_1, q', Y] \Rightarrow^* w'$ by a sequence of $m - 1$ productions. By induction, $(q_1, w', Z_1) \vdash^* (q', \epsilon, Y)$. Thus

$$(q, xw', Z) \vdash (q_1, w', Z_1) \vdash^* (q', \epsilon, Y).$$

An analogous argument establishes the result in case $[q, Z, q', Y] \to w_1$ is of type (4), (5), (6), or (7). Thus $(q, w, Z) \vdash^* (q', \epsilon, Y)$ and the proof of (8), thus $\text{Null}(M) = L(G)$, is complete.

If M is a one-turn pda, then $k = 1$. Thus only productions of types (1)–(6) occur. Therefore we have the next theorem (to be stated after a definition).

DEFINITION. A grammar $G = (V, \Sigma, P, \sigma)$ is called *linear* if each production is of the form $\xi \to uv v$ or $\xi \to u$, where u and v are in Σ^* and v is in $V - \Sigma$. A language L is called *linear* if $L = L(G)$ for some linear grammar.

THEOREM 2.1. *If M is a one-turn pda, then* $\text{Null}(M)$ *is a linear language.*

If M is a finite-turn pda, then the language $\text{Null}(M)$ is a generalization of a linear language. This generalization is studied in the next section.

3. Ultralinear languages. Motivated by the form of the grammar explicitly associated with a $(2k-1)$-turn pda given in §2, we introduce the following concepts.

DEFINITION. A grammar $G = (V, \Sigma, P, \sigma)$ is said to be *ultralinear* if $V - \Sigma$ is a union of disjoint (possibly empty) sets $A_0, \cdots, A_n$ of variables with the following property: For each A_i and each variable ξ in A_i, each production with left side ξ is either of the form $\xi \to uv\nu$ with ν in A_i and u, v in Σ^*, or of the form $\xi \to w$, with w in $(\Sigma \cup A_0 \cup \cdots \cup A_{i-1})^*$. $\{A_0, \cdots, A_n\}$ is called an *ultralinear decomposition*. A language is said to be *ultralinear* if it is generated by some ultralinear grammar.

If $G = (V, \Sigma, P, \sigma)$ is ultralinear, then so is the grammar (V, Σ, P, ξ) for each variable ξ in G.

Consider the grammar $G = (V, \Sigma, P, \sigma)$ of §2 associated with a $(2k-1)$-turn pda. In the notation of §2, the variables in $V - \{\sigma\}$ are quadruples $[q, Z, q', Y]$ with $1 \leqq h(q) < h(q') \leqq 2k$. For $1 \leqq i \leqq 2k-1$ let A_i be the set of those variables $[q, Z, q', Y]$ such that $h(q') - h(q) = i$. Let $A_{2k} = \{\sigma\}$. Then $\{A_1, \cdots, A_n\}$ is an ultralinear decomposition of G. (Productions of types (3), (4), (6) are of the form $\xi \to uv\nu$, where ξ, ν are in A_1 and u, v are in Σ^*. Productions of types (1), (2), (7) are of the form $\xi \to w$, with ξ in A_i and w in $(\Sigma \cup A_1 \cup \cdots \cup A_{i-1})^*$. Productions of type (5) are sometimes of the first form, i.e., $\xi \to uv\nu$, and sometimes of the second, i.e., $\xi \to w$.) Thus we have:

LEMMA 3.1. *If M is a finite-turn pda, then* Null(M) *is an ultralinear language.*

We shall show later that the converse is also true, thereby characterizing the $(2k-1)$-turn pda in terms of the ultralinear languages. First though, we present another characterization of ultralinear languages.

Notation. Let $S = (K, \Sigma, \Sigma, H, q_0, F)$ be an f-transducer and g a function from F to subsets of Σ^*. Then $S_f(g)$ denotes the union of all sets $u_1 \cdots u_r g(q) v_r \cdots v_1$ for which there exist $r \geqq 0$, $q = q_r$ in F, and $q_1, \cdots, q_{r-1}$ in K such that each $(q_i, u_{i+1}, v_{i+1}, q_{i+1})$ is in H.

LEMMA 3.2. *Let $S = (K, \Sigma, \Sigma, H, q_0, F)$ be an f-transducer. If $g(q)$ is a language for each q in F, then $S_f(g)$ is a language. If $g(q)$ is ultralinear for each q, then $S_f(g)$ is ultralinear.*

Proof. For each q in F let $G_q = (V_q, \Sigma, P_q, \sigma_q)$ be a grammar generating $g(q)$. We may assume that $(V_p - \Sigma) \cap (V_q - \Sigma) = \varnothing$ for all p, q in F and that $K \cap (\bigcup_{q \in F} V_q) = \varnothing$. Let $G = ((\bigcup_q V_q) \cup K, \Sigma, P, q_0)$, where P contains $\bigcup_q P_q$ together with the productions (i) $q \to upv$ for each (q, u, v, p) in H, and (ii) $q \to \sigma_q$ for each q in F. Since $L(G) = S_f(g)$, $S_f(g)$ is a language.

Suppose that each G_q is ultralinear. For each q let $\{A_{q,0}, \cdots, A_{q,n_q}\}$ be an ultralinear decomposition of $V_q - \Sigma$. Let $n = \max \{n_q \mid q \text{ in } K\}$,

$A_i = \bigcup_{q \in K} A_{q,i}$ for $0 \leqq i \leqq n$, and $A_{n+1} = K$. Then $\{A_0, \cdots, A_{n+1}\}$ is an ultralinear decomposition of the variables of G. Hence G, thus $S_f(g)$, is ultralinear.

DEFINITION. A language L is said to be *bounded* if there exist words $w_1, \cdots, w_t$ such that $L \subseteq w_1^* \cdots w_t^*$.

COROLLARY (to Lemma 3.2). *Every bounded language is ultralinear.*

Proof. For u, v in Σ^* and $B \subseteq \Sigma^*$ let $(u, v)*B = \bigcup_{i \geqq 0} u^i B v^i$. Now bounded languages [11] are generated from finite sets by finite union, finite product, and the operation $(u, v)*B$. To prove the corollary it thus suffices to show that $(u, v)*B$ is an ultralinear language if B is an ultralinear language and u, v are arbitrary words in Σ^*. Let S be the one-state f-transducer $(\{s_0\}, \Sigma, \Sigma, H, s_0, \{s_0\})$, where $H = \{(s_0, u, v, s_0)\}$. Let $g(s_0) = B$. By Lemma 3.2, $S_f(g)$ is ultralinear. Since $S_f(g) = (u, v)*B$, $(u, v)*B$ is ultralinear.

THEOREM 3.1. *The ultralinear languages constitute the smallest family D of subsets of Σ^* containing the finite sets and closed with respect to $S_f(g)$,† finite union, and finite product.*

Proof. By Lemma 3.2, the family D contains the ultralinear languages. To see the converse, let $G = (V, \Sigma, P, \sigma)$ be an ultralinear grammar with $\{A_0, \cdots, A_n\}$ an ultralinear decomposition of its variables. We shall prove that $L(G)$ is in D.

For each word y in V^*, let $L_y = \{w \text{ in } \Sigma^* \mid y \Rightarrow^* w\}$. For each j and each γ in A_j let S^γ be the f-transducer $(A_j, \Sigma, \Sigma, H_j, \gamma, A_j)$, where H_j consists of all quadruples (ξ, u, v, ξ') for which there exist ξ, ξ' in A_j such that $\xi \to u\xi'v$ is in P. For each ξ let $g(\xi)$ be the union of all L_y such that $\xi \to y$, y in $(\Sigma \cup A_0 \cup \cdots \cup A_{i-1})^*$, where i is the integer for which A_i contains ξ.

We first show that for each variable ξ, $L_\xi = S_f^{\xi}(g)$. To see this let w be a word in L_ξ and let ξ be in A_j. Let

$$\xi = w_0 \Rightarrow w_1 \Rightarrow \cdots \Rightarrow w_r = w$$

be a derivation of w. Let k be the smallest integer m such that the production involved in $w_m \Rightarrow w_{m+1}$ is not of the form $\alpha \to u\beta v$, α and β in A_j, u and v in Σ^*. Since the production involved in $w_{r-1} \Rightarrow w_r$ is of the form $\alpha \to z$ (z in Σ^*), $k \geqq 0$. Thus $w_k = z_1 \xi' z_1'$ for some z_1, z_1' in Σ^* and ξ' in A_j. Starting with ξ' and applying the productions which occur in $w_i \Rightarrow w_{i+1}$, $i \geqq k$, in the same order, we obtain a derivation $\xi' \Rightarrow^* w'$. Clearly $w = z_1 w' z_1'$ and w' is in $g(\xi')$. Thus $L_\xi \subseteq S_f^{\xi}(g)$. To see the reverse inclusion let w be in $S_f^{\xi}(g)$. Then there exist $r \geqq 0$, $(\xi_i, u_i, v_i, \xi_{i+1})$, $1 \leqq i \leqq r$, in H_j, w' in $g(\xi_{r+1})$ such that $\xi = \xi_1$ and $w = u_1 \cdots u_r w' v_r \cdots v_1$. By definition

† That is, $S_f(g)$ is in D for each f-transducer $S = (K, \Sigma, \Sigma, H, q_0, F)$ and function g from F to D.

of H_j, $\xi \Rightarrow^* u_1 \cdots u_r \xi_{r+1} v_r \cdots v_1$. Since w' is in $g(\xi_{r+1})$, there exists a production $\xi_{r+1} \to y$ such that $y \Rightarrow^* w'$. Then

$$u_1 \cdots u_r \, \xi_{r+1} v_r \cdots v_1 \Rightarrow^* u_1 \cdots u_r w' v_r \cdots v_1 = w.$$

Thus w is in L_ξ and $L_\xi = S_f^{\xi}(g)$.

To prove $L(G)$ is in D we shall show that L_ξ is in D for each variable ξ. Consider $L_\xi = S_f^{\xi}(g)$ for ξ in A_0. Now $g(\nu)$, ν in A_0, is the union of all L_y such that $\nu \to y$, y in Σ^*. Thus each $g(\nu)$ is a finite set. Then L_ξ is in D since D contains the finite sets and is closed under $S_f^{\xi}(g)$. Continuing by induction, suppose that L_ξ is in D for all ξ in $A_0 \cup \cdots \cup A_i$, $i < j$. Consider L_ξ for ξ in A_j. Now $g(\nu)$, ν in A_j, is the union of all L_y such that $\nu \to y$, y in $(\Sigma \cup A_0 \cup \cdots \cup A_{j-1})^*$. Each such L_y is the finite product of words and sets L_γ, γ in $A_0 \cup \cdots \cup A_{j-1}$. By induction, D contains the L_γ. Since D is closed under finite union, it contains the $g(\nu)$. Since D is closed under $S_f^{\xi}(g)$, D contains L_ξ.

THEOREM 3.2. *A set L is an ultralinear language if and only if there is a finite-turn pda M such that $L = \mathrm{Null}(M)$.*

Proof. Let E be the family of $\mathrm{Null}(M)$, M a finite-turn pda. By Lemma 3.1, each element of E is ultralinear. To prove the converse, in view of Theorem 3.1 it suffices to show that E contains the finite sets and is closed under finite union, finite product, and $S_f(g)$.

Obviously E contains the finite sets. To see that E is closed under finite union and finite product, it suffices to show that $\mathrm{Null}(M) \cup \mathrm{Null}(M')$ and $\mathrm{Null}(M)\,\mathrm{Null}(M')$ are in E for $\mathrm{Null}(M)$, $\mathrm{Null}(M')$ in E. Let $M = (K, \Sigma, \Gamma, \delta, Z_0, q_0)$ be a $(2k-1)$-turn pda and $M' = (K', \Sigma, \Gamma', \delta', Z_0', q_0')$ a $(2k'-1)$-turn pda, with $K \cap K' = \varnothing$ and $\Gamma \cap \Gamma' = \varnothing$.

Let $M_2 = (K_2, \Sigma, \Gamma_2, \delta_2, Z_0'', q_0'')$, where q_0'' is a symbol not in $K \cup K'$, Z_0'' is a symbol not in $\Gamma \cup \Gamma'$, $K_2 = K \cup K' \cup \{q_0''\}$, $\Gamma_2 = \Gamma \cup \Gamma' \cup \{Z_0''\}$, and δ_2 is defined as follows:

(a) (q, γ) is in $\delta_2(p, x, Z)$ if (q, γ) is in $\delta(p, x, Z)$,
(b) (q', γ') is in $\delta_2(p', x, Z')$ if (q', γ') is in $\delta'(p', x, Z')$,
(c) $\delta_2(q_0'', \epsilon, Z_0'') = \{(q_0, Z_0), (q_0', Z_0')\}$.

Then M_2 is a $(2k_2 - 1)$-turn pda, where $k_2 = \max\{k, k'\}$, and $\mathrm{Null}(M_2) = \mathrm{Null}(M) \cup \mathrm{Null}(M')$.

Let $M_3 = (K_3, \Sigma, \Gamma \cup \Gamma', \delta_3, Z_0, q_0)$, where δ_3 is defined as follows:

(a') $(q, Z_0'\gamma)$ is in $\delta_3(q_0, x, Z_0)$ if (q, γ) is in $\delta(q_0, x, Z_0)$,
(b') (q, γ) is in $\delta_3(p, x, Z)$ if (q, γ) is in $\delta(p, x, Z)$,
(c') (q_0', Z_0') is in $\delta_3(p, \epsilon, Z_0')$ for each p in K,
(d') (q', γ') is in $\delta_3(p', x, Z')$ if (q', γ') is in $\delta(p', x, Z')$.

Then M_3 is a $(2(k + k') - 1)$-turn pda and

$$\mathrm{Null}(M_3) = \mathrm{Null}(M)\,\mathrm{Null}(M').$$

To see that E is closed under $S_f(g)$, let $S = (K, \Sigma, \Sigma, H, q_0, F)$ be an f-transducer, and for each q in F let $M_q = (K_q, \Sigma, \Gamma_q, \delta_q, Y_q, s_q)$ be a $(2k_q - 1)$-turn pda. We may assume that $\{K, K_q \mid q \text{ in } K\}$ are pairwise disjoint and $\{\Gamma_q \mid q \text{ in } K\}$ are pairwise disjoint. We may also assume that S is 1-restricted. We shall construct a pda M_4 which satisfies the following: When an input enters S, the next state structure of M_4 copies the next state structure of S, putting the output from S on the pushdown tape (in coded form). At a state q of F the pda M_4 moves (under ϵ) to the start state of M_q and then duplicates the moves of M_q. If the moves of M_q constitute a sweep, then M_4 moves to a special state p_* and thereafter moves only when the input symbol is the output which is represented by the rightmost symbol on the pushdown tape.

To construct M_4, let p_* be a symbol not in $K \cup (\bigcup_q K_q)$ and let Z_0 be a symbol not in $\bigcup_q \Gamma_q$. For each q in F let X_q be an abstract symbol not in $(\bigcup_q \Gamma_q) \cup \{Z_0\}$. For each x in $\Sigma \cup \{\epsilon\}$, let Z_x be an abstract symbol not in $(\bigcup_q \Gamma_q) \cup \{X_q \mid q \text{ in } F\} \cup \{Z_0\}$. Let M_4 be the pda $(K_4, \Sigma, \Gamma_4, \delta_4, Z_0, q_0)$, where $K_4 = K \cup (\bigcup_q K_q) \cup \{p_*\}$,

$$\Gamma_4 = (\bigcup_q \Gamma_q) \cup \{Z_0\} \cup \{Z_x \mid x \quad \text{in} \quad \Sigma \cup \{\epsilon\}\} \cup \{X_q \mid q \quad \text{in} \quad F\},$$

and δ_4 is defined as follows:

(a″) $(q, Z_y Z_0)$ is in $\delta_4(p, x, Z_0)$ if (p, x, y, q) is in H,
(b″) $(s_q, X_q Y_q)$ is in $\delta_4(q, \epsilon, Z_0)$ for each q in F,
(c″) for s in K_q, (s', γ) is in $\delta_4(s, x, Z)$ if (s', γ) is in $\delta_q(s, x, Z)$,
(d″) for s in K_q, (p_*, ϵ) is in $\delta_4(s, \epsilon, X_q)$,
(e″) (p_*, ϵ) is in $\delta_4(p_*, x, Z_x)$ for all x in $\Sigma \cup \{\epsilon\}$.

Then M_4 is a $(2k_4 - 1)$-turn pda, with $k_4 = \max\{k_q \mid q \text{ in } F\}$; and $\text{Null}(M_4) = S_f(g)$, where $g(q) = \text{Null}(M_q)$ for each q in F.

Using Theorem 3.2, we are able to prove several other results.

THEOREM 3.3. *If L is an ultralinear language and S an f-transducer, then $S_f(L)$ is an ultralinear language.*

Proof. Without loss of generality we may assume that $S = (K, \Sigma, \Sigma, H, s_0, F)$ is 1-restricted and that H contains $(p, \epsilon, \epsilon, p)$ for all p. Since L is ultralinear, by Theorem 3.2 there is a finite-turn pda $M = (K', \Sigma, \Gamma, \delta, Z_0, q_0)$ such that $L = \text{Null}(M)$. We may also assume that for each q in K' and Z in Γ, (q, Z) is in $\delta(q, \epsilon, Z)$. We alter M slightly to obtain a pda with the same structure as M but with the ability to detect when the length of the pushdown tape is exactly one. For each Z in Γ let $\bar{Z}$ be a new symbol and let $\Gamma' = \Gamma \cup \{\bar{Z} \mid Z \text{ in } \Gamma\}$. Let M' be the pda $(K', \Sigma, \Gamma', \delta', \bar{Z}_0, q_0)$ where

(a) (q', w) is in $\delta'(q, x, Z)$ if (q', w) is in $\delta(q, x, Z)$,
(b) (q', ϵ) is in $\delta'(q, x, \bar{Z})$ if (q', ϵ) is in $\delta'(q, x, Z)$,
(c) $(q', \bar{Z}'w')$ is in $\delta'(q, x, \bar{Z})$ if $(q', \bar{Z}'w')$ is in $\delta'(q, x, Z)$.

Clearly M' is also a finite-turn pda and $\text{Null}(M') = L$. Furthermore, M' has the property that the leftmost, and only the leftmost, symbol on the pushdown tape is marked, that is, is of the form $\bar{Z}$ for some Z in Γ.

Let M'' be the pda $(K'', \Sigma, \Gamma', \delta'', \bar{Z}_0, q_0'')$, where $K'' = K \times K'$, $q_0'' = (s_0, q_0)$, and δ'' is defined as follows:

(a′) $((p', q'), w)$ is in $\delta''((p, q), x, Z)$ if there is an element y such that (p, y, x, p') is in H and (q', w) is in $\delta'(q, y, Z)$;
(b′) $((p', q'), \epsilon)$ is in $\delta''((p, q), x, \bar{Z})$ if p' is in F, there is an element y such that (p, y, x, p') is in H, and (q', ϵ) is in $\delta'(q, y, \bar{Z})$.

Clearly $\text{Null}(M'') = S_f(\text{Null}(M'))$. Furthermore, M'' is a finite-turn pda. (For if

$$((p_0, q_0), y_1 \cdots y_m, \bar{Z}_0) \vdash_{M''} ((p_1, q_1), y_2 \cdots y_m, \gamma_1) \vdash_{M''} \cdots \vdash_{M''} (p_m, q_m), \epsilon, \gamma_m),$$

then there exist $x_1, \cdots, x_m$ such that

$$(q_0, x_1 \cdots x_m, \bar{Z}_0) \vdash_{M'} (q_1, x_2 \cdots x_m, \gamma_1) \vdash_{M'} \cdots \vdash_{M'} (q_m, \epsilon, \gamma_m).$$

Thus if M' is a $(2k' - 1)$-turn pda, then M'' is a $(2k'' - 1)$-turn pda, with $k'' \leqq k'$.)

DEFINITION. A *generalized sequential machine* (abbreviated *gsm*) is a 6-tuple $S = (K, \Sigma, \Delta, \delta, \lambda, s_0)$ where

(i) K, Σ, and Δ are finite nonempty sets (of *states*, *inputs*, and *outputs* respectively),
(ii) δ is a mapping of $K \times \Sigma$ into K (the *next state function*),
(iii) λ is a mapping of $K \times \Sigma$ into Δ^* (the *output function*), and
(iv) s_0 is in K (the *start state*).

The functions δ and λ are extended inductively to $K \times \Sigma^*$ by defining $\delta(q, \epsilon) = q$, $\lambda(q, \epsilon) = \epsilon$, $\delta(q, wx) = \delta[\delta(q, w), x]$, and $\lambda(q, wx) = \lambda(q, w)\lambda[\delta(q, w), x]$ for each q in K, w in Σ^*, and x in Σ. The mapping S of Σ^* into Δ^* defined by $S(w) = \lambda(s_0, w)$ is called a *gsm mapping*.

DEFINITION. For each element a in Σ let Σ_a be a finite nonempty set and $\tau(a)$ a subset of Σ_a^*. Let $\tau(\epsilon) = \{\epsilon\}$ and $\tau(x_1 \cdots x_r) = \tau(x_1) \cdots \tau(x_r)$ for all $x_1 \cdots x_r$, each x_i in Σ. Then the function τ, of Σ^* into the subsets of $(\bigcup_a \Sigma_a)^*$, is called a *substitution*. If each $\tau(a)$ is regular, then τ is called a *substitution by regular sets*. If each $\tau(a)$ is finite, then τ is called a *finite substitution*.

THEOREM 3.4. *Ultralinear languages are preserved by intersection with regular sets, mapping by gsm, and substitution by regular sets.*

Proof. Let L be a language and R a regular set. Let $A = (K, \Sigma, \delta, s_0, F)$ be an automaton such that $R = T(A)$. Let S be the f-transducer $(K, \Sigma, \Sigma, H, s_0, F)$, where H consists of all quadruples (q, x, x, q'), x in Σ,

such that $\delta(q, x) = q'$. Obviously $L \cap R = S_f(L)$. If L is ultralinear, then $L \cap R$ is ultralinear by Theorem 3.3.

Let $S = (K, \Sigma, \Delta, \delta, \lambda, s_0)$ be a gsm. Let S' be the f-transducer $(K, \Sigma \cup \Delta, \Sigma \cup \Delta, H, s_0, K)$, where (p, x, y, q) is in H, x in Σ, if $\delta(p, x) = q$ and $\lambda(p, x) = y$. Since $S(L) = S_f'(L)$, ultralinear languages are preserved by gsm mappings.

Suppose that τ is a substitution mapping such that $\tau(a)$ is regular for each a in Σ. For each a in Σ, let A_a be an automaton $(K_a, \Sigma_a, \delta_a, s_a, F_a)$ such that $T(A_a) = \tau(a)$. We may assume that $K_a \cap K_b = \varnothing$ for all $a \neq b$. Let s_0 be a symbol not in $\bigcup_{a\in\Sigma} K_a$. Let S be the f-transducer $(K, \Sigma', \Sigma', H, s_0, \{s_0\})$, where $K = \{s_0\} \cup (\bigcup_{a\in\Sigma} K_a)$, $\Sigma' = \Sigma \cup (\bigcup_{a\in\Sigma} \Sigma_a)$, and H consists of the following quadruples:

(a) (s_0, a, ϵ, s_a) for each a in Σ,
(b) (q, ϵ, x, q') for q, q' in K_a if $\delta_a(q, x) = q'$ (a in Σ, x in Σ_a),
(c) $(q, \epsilon, \epsilon, s_0)$ for q in F_a (a in Σ).

Clearly $S_f(L) = \tau(L)$ for each set L. Thus $\tau(L)$ is ultralinear if L is ultralinear.

4. Nonterminal bounded languages. In this section we present another characterization of ultralinear languages. This characterization will be used later in proving that it is undecidable whether a given language is ultralinear.

DEFINITION. A grammar $G = (V, \Sigma, P, \sigma)$ is called *nonterminal bounded* [1] (called *bounded* in [2], [3]) if there exists an integer k with the following property: If $\xi \Rightarrow^* w$, w in V^*, ξ in $V - \Sigma$, then w has at most k occurrences of variables. (Actually, the definition given in [1], [2], [3] requires the last condition only when $\xi = \sigma$. There is no real loss in allowing ξ to be any variable.) A language is called a *nonterminal bounded language* if it is generated by some nonterminal bounded grammar.

DEFINITION. Let $G = (V, \Sigma, P, \sigma)$ be a nonterminal bounded grammar. The *rank* $r_G(w)$, written $r(w)$ when G is understood, of a word w in V^* is defined to be the largest integer r such that there is a word u in V^*, with r occurrences of variables, such that $w \Rightarrow^* u$.

Note that $r(w) = 0$ for w in Σ^* and $r(w) = \sum_1^s r(w_i)$ for each $w = w_1 \cdots w_s$, all w_i in V.

If ξ is a variable of rank r, then there is a word w in V^*, having r occurrences of variables, such that $\xi \Rightarrow^* w$. Since the rank of each variable is at least one and since $r(w) \leqq r(\xi)$, it follows that each variable in w has rank 1. In particular, each nonterminal bounded grammar contains variables of rank 1.

THEOREM 4.1. *A grammar is ultralinear if and only if it is nonterminal bounded.*

Proof. Suppose that $G = (V, \Sigma, P, \sigma)$ is an ultralinear grammar with $\{A_0, \cdots, A_n\}$ an ultralinear decomposition. Let $m = \max \{|u| \mid v \to u \text{ is in } P\}$. Obviously w has at most n^m occurrences of variables for every w such that $\xi \Rightarrow^* w$, ξ in $V - \Sigma$.

Conversely, assume that G is a nonterminal bounded grammar. Let $A_0 = \varnothing$ and for $i \geqq 1$ let A_i be the set of variables with rank i. If k is the maximum rank of any variable, then $\{A_0, A_1, \cdots, A_k\}$ is obviously an ultralinear decomposition of the variables.

DEFINITION. A language is called *metalinear* if it is a finite union of products of linear languages.

In view of the coincidence of an ultralinear grammar and a nonterminal bounded grammar, we have the following results [1]:

Each metalinear, thus each linear, language is ultralinear.

Each finite union and finite product of ultralinear languages is ultralinear.

We now introduce the notion of "rank" of a nonterminal bounded grammar and nonterminal bounded language. In the remainder of this section we then present some facts which are not only of inherent interest but are needed to prove an unsolvability result in §5 (Theorem 5.2).

DEFINITION. For each nonterminal bounded grammar G, the *rank of* G, denoted by $r(G)$, is defined as the largest integer which is the rank of one of the variables. Let L be a nonterminal bounded language. The *rank of* L, $r(L)$, is defined as zero if L is regular. If L is nonregular, then the *rank of* L, $r(L)$, is defined as the smallest integer which is the rank of some grammar generating it.

THEOREM 4.2. *Each of the following operations preserves nonterminal bounded languages and does not increase rank*:

(a) *Intersection with a regular set.*
(b) *Finite substitution.*
(c) *Gsm mapping.*

(Parts (a) and (c) generalize portions of Theorem 3.4.)

Proof. Each of these operations is known to preserve regular sets [9]. Thus it suffices to show that if L is a nonterminal bounded language with $r(L) \leqq r$, where $r \geqq 1$, then its image L' under any of the three operations has the property that $r(L') \leqq r$.

The standard proof [4] that the intersection of a language and a regular set is a language also shows that the intersection of a nonterminal bounded language of rank $\leqq r$ and a regular set is a nonterminal bounded language of rank $\leqq r$. Thus (a) is established.

To prove (b), let L be a nonterminal bounded language. Let $G = (V, \Sigma, P, \sigma)$ be a nonterminal bounded grammar generating L, with $r(L) = r(G)$. Let τ be the substitution defined by letting $\tau(a)$ be a finite

set of words in $(\Sigma')^*$ for each a in Σ. We shall show that $r(\tau(L)) \leqq r(L)$. To this end let τ be extended to a substitution over V^* by defining $\tau(\xi) = \{\xi\}$ for each ξ in $V - \Sigma$. Let $G' = (V', \Sigma', P', \sigma)$,where $V' = (V - \Sigma) \cup \Sigma'$ and $P' = \{\xi \to u \mid \xi \to w \text{ in } P, u \text{ in } \tau(w)\}$. Obviously $\tau(L(G)) = L(G')$. Since the productions of P' differ from corresponding productions in P in terminal symbols but not in variables, it follows that each variable in G has the same rank in G as in G'. Therefore $r(G') \leqq r(G)$, so that $r(\tau(L)) \leqq r(L)$.

The proof of (c) follows in the standard way from (a) and (b) [10]. Thus the theorem is completely demonstrated.

We now consider two lemmas needed for the next theorem.

Part (b) of Theorem 4.2 can be extended to substitution by regular sets. We shall not present an argument of this but shall content ourselves with a special case in the next lemma. (The general case follows by a proof which is similar but notationally more complicated.)

LEMMA 4.1. *Let $x_1, \cdots, x_k$ be symbols not in Σ and let $L \subseteq \Sigma^* \cup \Sigma^* x_1 \cup \cdots \cup \Sigma^* x_k$ be a nonterminal bounded language. For each i let R_i be a regular subset of Σ^* and let L' be the result of substituting R_i for x_i in words of L. Then L' is a nonterminal bounded language and $r(L') \leqq r(L)$.*

Proof. Since substitution by regular sets preserves regular sets, it follows that $r(L') = 0$ if $r(L) = 0$. Hence it suffices to prove that if L is generated by a grammar G, then L' is generated by a grammar G' with $r(G') \leqq r(G)$.

Clearly there is no loss in assuming that $L \neq \varnothing$ and that L does not contain ϵ. Let $G = (V, \Sigma, P, \sigma)$ be a nonterminal bounded grammar generating L. By [4, Lemma 4.1], we may assume that

(1) no production in P is of the form $\xi \to \epsilon$.

By [4, Lemma 5.1], we may assume for each ξ in $V - \Sigma$ that

(2) $\{w \text{ in } \Sigma^* \mid \xi \Rightarrow^* w\}$ is nonempty, and
(3) there exist u_ξ and v_ξ such that $\sigma \Rightarrow^* u_\xi \xi v_\xi$.

Suppose there exists a production in P of the form $\xi \to u_1 x_i u_2$ for some i, with $u_2 \neq \epsilon$. By (1), (2), and (3) this implies that L contains a word of the form $w_1 x_i w_2$, $w_2 \neq \epsilon$, a contradiction. Thus each production in P is of the form $\xi \to w$ or $\xi \to w x_i$, w containing no x_j, $1 \leqq j \leqq k$. Let P' consist of all productions $\xi \to w$ in P with w containing no x_j and let P'' consist of all productions in P of the form $\xi \to w x_i$ for some i. Let $\pi_1, \cdots, \pi_s$ be the distinct productions in P''. For each j, let π_j be the production $\xi_j \to w_j X_j$, with $X_j = x_{i(j)}$ for some $i(j)$, $1 \leqq i(j) \leqq k$. For each j, $1 \leqq j \leqq s$, let $G_j = (V_j, \Sigma, P_j, \sigma_j)$ be a left-linear† grammar generating $R_{i(j)}$. Further-

† A grammar is called *left-linear* (*right-linear*) if each production in it is of the form $\xi \to w$ or $\xi \to \xi' w$ ($\xi \to w$ or $\xi \to w\xi'$), ξ' a variable and w containing no variable. It is known [5] that a set is regular if and only if it is generated by some left-linear (right-linear) grammar.

more, we may assume that $(V_j - \Sigma) \cap (V_{j'} - \Sigma) = \varnothing$ for $j \neq j'$. Let $\bar{V} = V \cup W$, where

$$W = \{(\pi_j, \tau) \in P'' \times (V_j - \Sigma) \mid 1 \leqq j \leqq s\}.$$

Let $\bar{G} = (\bar{V}, \Sigma, \bar{P}, \sigma)$, where $\bar{P}$ consists of the following productions ($1 \leqq j \leqq s$):

(4) $\xi \to w$ if $\xi \to w$ is in P';
(5) $\xi_j \to (\pi_j, \sigma_j)$,
(6) $(\pi_j, \tau) \to (\pi_j, \tau')w$ if $\tau \to \tau' w$ is in P_j, τ' in $V_j - \Sigma$;
(7) $(\pi_j, \tau) \to w_j w$ if $\tau \to w$ is in P_j, w in Σ^*.

By (4), $P' \subseteq \bar{P}$. The effect of (5) followed by a sequence of productions of type (6) followed by a production of type (7) is to realize a derivation $\xi_j \Rightarrow^*_{\bar{G}} w_j u$ for any u in R_j by using words with only one variable until the last word. It is easily verified that $L(\bar{G}) = L'$ and $r(\bar{G}) \leqq r(G)$.

DEFINITION. Let c be a symbol not in Σ. For each grammar $G = (V, \Sigma, P, \sigma)$ such that $L(G) \subseteq \Sigma^* c \Sigma^*$, let

$$V_L = \{\xi \in V - \Sigma \mid \sigma \Rightarrow^* u_1 \xi u_2 c u_3\},$$

$$V_c = \{\xi \in V - \Sigma \mid \xi \Rightarrow^* w, w \text{ in } V^* c V^*\},$$

and

$$V_R = \{\xi \in V - \Sigma \mid \sigma \Rightarrow^* u_1 c u_2 \xi u_3\}.$$

V_L is called the set of *left-variables*, V_c the set of *c-variables*, and V_R the set of *right-variables*.

LEMMA 4.2. *Let c be a symbol not in Σ and let $L \subseteq \Sigma^* c \Sigma^*$ be a nonterminal bounded language with $r(L) \leqq r$ and $r \geqq 1$. Then there exists a nonterminal bounded grammar $G = (V, \Sigma, P, \sigma)$ generating L such that $r(G) \leqq r$ and $\{V_L, V_c, V_R\}$ is a decomposition of $V - \Sigma$.*†

Proof. Clearly we may assume that $L \neq \varnothing$. Let $G' = (V', \Sigma, P', \sigma)$ be a nonterminal bounded grammar generating L such that $r(G) \leqq r$. Without loss of generality [4], we may assume that for each ξ in $V' - \Sigma$, (α) there exist u_ξ and v_ξ in V'^* such that $\sigma \Rightarrow^* u_\xi \xi v_\xi$, and ($\beta$) there is some word w_ξ in Σ^* such that $\xi \Rightarrow^* w_\xi$. If ξ is in V_c' and $\xi \to w$ is in P', then by (α) and (β), $w = uXv$ where (i) X is in $\{c\} \cup V_c'$ and (ii) u, v contain no occurrences of elements of $\{c\} \cup V_c'$. We shall double the number of variables not in V_c' so that each new variable occurs only to the left or only to the right of an occurrence of an element of $\{c\} \cup V_c'$ in each derivation of a word in $L(G')$.

For each variable ξ not in V_c', let ξ_L and ξ_R be new symbols. Let

† That is, $V - \Sigma = V_L \cup V_c \cup V_R$ and elements of $\{V_L, V_c, V_R\}$ are pairwise disjoint.

$V = \Sigma \cup V_c \cup \{\xi_L, \xi_R \mid \xi \text{ in } V' - (\Sigma \cup V_c')\}$. For each word w in V'^* which contains no occurrence of an element of V_c', let w_L and w_R be the words in V^* obtained by replacing in w each variable ξ not in V_c' by ξ_L and ξ_R, respectively. Let $G = (V, \Sigma, P, \sigma)$ be the grammar defined by the following set P of productions:

(1) $\xi \to u_L X v_r$ if $\xi \to uXv$ is in P', where X is in $\{c\} \cup V_c'$;
(2) $\xi_L \to w_L$ and $\xi_R \to w_R$ if $\xi \to w$ is in P' and ξ is in $V' - (\Sigma \cup V_c')$.

It is easily verified that $L(G) = L(G')$ and $r(G) = r(G') \leqq r$. Furthermore, $V_c = V_c'$, $V_L = \{\xi_L \mid \text{all } \xi_L\}$, and $V_R = \{\xi_R \mid \text{all } \xi_R\}$; and $\{V_L, V_c, V_R\}$ is a decomposition of $V - \Sigma$.

The next result shows that the rank of $L_1 c L_2$, the symbol c being foreign to L_1 and L_2, is the sum of the ranks of L_1 and L_2. This fact yields, as a corollary, a key result in a chain leading to the unsolvability of determining for an arbitrary pda M whether Null(M) is ultralinear.

THEOREM 4.3. *Let L_1, L_2 be nonterminal bounded languages in Σ^* with ranks r_1 and r_2, respectively. Let c be a symbol not in Σ. Then $L_1 c L_2$ is a nonterminal bounded language of rank $r_1 + r_2$.*

Proof. The proof is a generalization of an argument by Greibach [12] to show that LcL is linear if and only if L is regular. It is obvious that $L_1 c L_2$ is a nonterminal bounded language such that $r(L_1 c L_2) \leqq r_1 + r_2$. Since L_1 and L_2 are images of $L_1 c L_2$ under generalized sequential machine mappings, it follows from Theorem 4.2 that $r_1 \leqq r(L_1 c L_2)$ and $r_2 \leqq r(L_1 c L_2)$. Therefore, if $r_2 = 0$ (or $r_1 = 0$), then $r(L_1 c L_2) = r_1 + r_2$. Hence we may assume that $r_1 \geqq 1$ and $r_2 \geqq 1$.

Suppose that $r(L_1 c L_2) \leqq r_1 + r_2 - 1$. (We shall establish the theorem by showing that this assumption leads to a contradiction.) By Lemma 4.2, there exists a nonterminal bounded grammar $G = (V, \Sigma, P, \sigma)$ generating $L_1 c L_2$ such that $r(G) \leqq r_1 + r_2 - 1$ and $\{V_L, V_c, V_R\}$ is a decomposition of $V - \Sigma$. We shall construct another grammar $G' = (V', \Sigma, P', (0, \sigma, 0))$ generating $L_1 c L_2$, with $r(G') \leqq r_1 + r_2 - 1$, and having the property that the variables ξ in V_c' are doubly indexed to keep count of the rank of all words u, v, in V'^* such that $(0, \sigma, 0) \Rightarrow^* u\xi v$.

Let W_c' be the set of all triples (n, ξ, m), where n and m are nonnegative integers such that $n + m \leqq r_1 + r_2 - 2$ and ξ is in V_c. Let $G' = (V', \Sigma, P', (0, \sigma, 0))$, where $V' = W_c' \cup V_L \cup V_R \cup \Sigma$ and P' consists of the following productions:

(1) $\xi \to w$ if ξ is in $V_L \cup V_R$ and $\xi \to w$ is in P;
(2) $(n, \xi, m) \to u(n + r_G(u), \xi', m + r_G(v))v$ if $\xi \to u\xi' v$ is in P, where ξ' is a c-variable of G;
(3) $(n, \xi, m) \to ucv$ if $\xi \to ucv$ is in P.

Then G' is a nonterminal bounded language such that $r(G') = r(G) \leqq r_1$

$+ r_2 - 1, r_{G'}(\xi) = r_G(\xi)$ for all ξ in $V_L \cup V_R$, and $L(G') = L(G) = L_1 c L_2$.

Let P_1' be the subset of P consisting of all productions of type (1) with ξ in V_R or $r_G(\xi) \leqq r_1 - 1$, and all productions of type (2) or (3) with $n + r_G(u) \leqq r_1 - 1$. Similarly let P_2' be the subset of P consisting of all productions of type (1) with ξ in V_L or $r_G(\xi) \leqq r_2 - 1$, and all productions of type (2) or (3) with $m + r_G(v) \leqq r_2 - 1$. Since $r(G) \leqq r_1 + r_2 - 1$, every production in P belongs to P_1' or to P_2'. Clearly every derivation in G' of a word w in $L(G')$ consists of productions all of which are in P_1' or all of which are in P_2'.

Let $G_1' = (V', \Sigma, P_1', (0, \sigma, 0))$ and $G_2' = (V', \Sigma, P_2', (0, \sigma, 0))$. Let $L_1' = L(G_1')$ and $L_2' = L(G_2')$. Clearly $L_1' \cup L_2' \subseteq L_1 c L_2$. Since each derivation of a word in $L(G') = L_1 c L_2$ is obtained solely from productions in P_1' or solely from productions in P_2', $L_1 c L_2 = L_1' \cup L_2'$.

Let μ and ν be the mappings of $\Sigma^* c \Sigma^*$ onto Σ^* defined by $\mu(xcy) = x$ and $\nu(xcy) = y$. Then $\mu(L_1 c L_2) = L_1$ and $\nu(L_1 c L_2) = L_2$. We shall give explicit grammars which generate $\mu(L(G_1'))$ and $\nu(L(G_2'))$ to show that they are nonterminal bounded languages of rank $\leqq r_1 - 1$ and $\leqq r_2 - 1$, respectively.

Let $V_1'' = \Sigma \cup V_L \cup \{(n, \xi) | (n, \xi, m) \text{ in } V_c \text{ for some } m\}$. Let $G_1'' = (V_1'', \Sigma, P_1'', (0, \sigma))$, where P_1'' consists of the following productions:

(4) $\xi \to w$ if ξ is in V_L and $\xi \to w$ is in P_1';

(5) $(n, \xi) \to u(n + r_G(u), \xi')$ if there exists a production $(n, \xi, m) \to u(n + r_G(u), \xi', m + r_G(v))v$ in P_1';

(6) $(n, \xi) \to u$ if there exists a production $(n, \xi, m) \to ucv$ in P_1'.

Clearly $L(G_1'') = \mu(L(G_1')) = \mu(L_1')$. Also, if $\xi \Rightarrow^* w$ in G_1'', then w contains at most $r_1 - 1$ occurrences of V_L, at most one occurrence of $V_1'' - (\Sigma \cup V_L)$, and thus at most r_1 occurrences of variables. Hence G_1'' is nonterminal bounded and $r(G_1'') \leqq r_1$. In case $r_1 = 1$, each production is of the form $(0, \xi) \to y$ or $(0, \xi) \to y(0, \xi')$, y in Σ^*. Since this grammar is right-linear, it generates a regular set. Thus $r(\mu(L_1')) = 0 = r_1 - 1$ in this case.

Suppose that $r_1 > 1$. The only productions in G_1'' of the form $(r_1 - 1, \xi) \to w$ are of the form $(r_1 - 1, \xi) \to u(r_1 - 1, \xi')$ or $(r_1 - 1, \xi) \to u$, with u in Σ^*. Thus the set

$$A(\xi) = \{w \in \Sigma^* | (r_1 - 1, \xi) \Rightarrow^* w\}$$

is regular. Let $G_1''' = (V_1''', \Sigma_1''', P_1''', (0, \sigma))$, where $V_1''' = V_1''$, $\Sigma_1''' = \Sigma \cup \{(r_1 - 1, \xi) | (r_1 - 1, \xi) \text{ in } V_1''\}$, and P_1''' consists of all productions in P_1'' excluding those of the form $(r_1 - 1, \xi) \to w$. Then G_1''' is a nonterminal bounded grammar with $r(G_1''') \leqq r_1 - 1$. Clearly $L(G_1''') \subseteq \Sigma^* \cup (\bigcup_\xi \Sigma^*(r_1 - 1, \xi))$, and $L(G_1'')$ is obtained from $L(G_1''')$ by substituting for each $(r_1 - 1, \xi)$ the regular set $A(\xi)$ for $(r_1 - 1, \xi)$. By Lemma 4.1, $r(L(G_1'')) \leqq r_1 - 1$. Hence $r(\mu(L_1')) \leqq r_1 - 1$.

In a similar manner, we see that $r(\nu(L_2'))) \leqq r_2 - 1$. Since $r(L_1) = r_1$ and $\mu(L_1') \subseteq L_1$, there is a word w_1 in $L_1 - \mu(L_1')$. Similarly there is a word w_2 in $L_2 - \nu(L_2')$. Then w_1cw_2 is neither in L_1' nor in L_2'. This contradicts the fact that w_1cw_2 is in $L_1cL_2 = L_1' \cup L_2'$.

We now derive an important corollary.

COROLLARY 1. *Let $L \subseteq \Sigma^*$ be a nonterminal bounded language and c a symbol not in Σ. Then $(Lc)^n$ is a nonterminal bounded language with $r[(Lc)^n] = nr(L)$.*

Proof. Let $c_1, \cdots, c_n$ be distinct symbols not in Σ. Obviously there exist generalized sequential machines which map $(Lc)^n$ onto $Lc_1 \cdots Lc_n$ and $Lc_1 \cdots Lc_n$ onto $(Lc)^n$. Hence $r[(Lc)^n] = r(Lc_1 \cdots Lc_n)$ by Theorem 4.2. Using induction and Theorem 4.3, it follows that $r(Lc_1 \cdots Lc_n) = nr(L)$. Hence the corollary.

COROLLARY 2. *For every $n \geqq 0$, there exist nonterminal bounded languages with $r(L) = n$.*

5. Recognition. In [1], [2], [3], the following two results are essentially shown: it is recursively solvable to determine of an arbitrary grammar whether it is a nonterminal bounded grammar; the rank of an arbitrary nonterminal bounded grammar is effectively calculable. In this section, we consider the problems of determining (a) whether an arbitrary pda is finite-turn, and (b) whether, for an arbitrary pda M, Null(M) is an ultralinear language. (Problem (b) is equivalent to determining whether an arbitrary grammar generates a nonterminal bounded language.)

THEOREM 5.1. *It is recursively solvable to determine if an arbitrary pda is finite-turn. If a pda M is finite-turn, it is solvable to find the smallest integer k_0 such that each sweep of M has $2k - 1$ turns for some $k \leqq k_0$.*

Proof. Let $M = (K, \Sigma, \Gamma, \delta, Z_0, q_0)$ be an arbitrary pda. Let q_* be a symbol not in K. Let N be the pda $(K_N, \{a, b, c\}, \Gamma, \delta_N, q_*)$, where $K_N = K \cup \{q_*\}$ and δ_N is defined as follows: (q_0, Z_0) is in $\delta_N(q_*, a, Z_0)$. If (q', w) is in $\delta(q, x, Z)$, let

$$(q', w) \text{ be in } \begin{cases} \delta(q, b, Z) \text{ if } w = \epsilon, \\ \delta(q, c, Z) \text{ if } |w| = 1, \\ \delta(q, a, Z) \text{ if } |w| > 1. \end{cases}$$

Then to each sweep of $2k - 1$ turns in M,

$$(q_0, x_1 \cdots x_r, Z_0) \vdash_M (q_1, x_2 \cdots x_r, \gamma_1) \vdash_M \cdots \vdash_M (q_r, \epsilon, \gamma_r) = (q_r, \epsilon, \epsilon),$$

there corresponds in a one-to-one manner a sweep with $2k - 1$ turns in N,

$$(q_*, ay_1 \cdots y_r) \vdash_N (q_0, y_1 \cdots y_r, Z_0)$$
$$\vdash_N (q_1, y_2 \cdots y_r, \gamma_1) \vdash_N \cdots \vdash_N (q_r, \epsilon, \gamma_r) = (q_r, \epsilon, \epsilon),$$

and conversely. A turn occurs in applying $ay_1 \cdots y_r$ to N each time b occurs in a subword of the form ac^ib $(i \geqq 0)$ and each time a occurs in a subword of the form bc^ia $(i \geqq 0)$, and no other time.

We now construct a gsm S which, on receiving a word of Null(N) as input, deletes all occurrences of c, all occurrences of a immediately preceded by a, and all occurrences of b immediately preceded by b. Let S be the gsm $(\{p_0, p_a, p_b\}, \{a, b, c\}, \{a, b\}, \delta_S, \lambda_S, p_0)$, where $\delta_S(p_0, a) = p_a$, $\lambda_S(p_0, a) = a$, $\delta_S(p_b, a) = p_a$, $\lambda_S(p_b, a) = a$, $\delta_S(p_a, b) = p_b$, $\lambda_S(p_a, b) = b$, $\delta_S(p, x) = p$ and $\lambda_S(p, x) = \epsilon$ otherwise. Clearly S has the asserted properties. Then N, thus M, is finite-turn, if and only if $S(\text{Null}(N))$ is a finite set. If $S(\text{Null}(N))$ is finite and $2k_0$ is the length of the longest word in $S(\text{Null}(N))$, then k_0 is the smallest integer such that each sweep of N, thus M, has $2k - 1$ turns for some $k \leqq k_0$. Now Null(N) is a language, and the image of a language under a gsm is a language which is effectively calculable [10]. Moreover, it is solvable to determine whether a language is finite, and if finite, the length of its longest word [9]. This completes the proof.

We now prove that it is recursively unsolvable to determine of an arbitrary pda M whether Null(M) is an ultralinear language.

LEMMA 5.1. *$L \subseteq \Sigma^*$ be a language and c a symbol not in Σ. Then $(Lc)^* - \{\epsilon\}$ is ultralinear if and only if L is regular.*

Proof. Suppose L is regular. Then $(Lc)^* - \{\epsilon\}$ is regular and thus ultralinear.

Suppose L is not regular but $(Lc)^* - \{\epsilon\}$ is ultralinear. Then $(Lc)^* - \{\epsilon\}$ is a nonterminal bounded language and has a rank n. Since L is the image under a gsm of $(Lc)^* - \{\epsilon\}$, $r(L) \leqq n$ by Theorem 4.2. Since L is not regular, $r(L) \geqq 1$. By Corollary 1 of Theorem 4.3,

$$n < n + 1 \leqq (n + 1)r(L) = r[(Lc)^{n+1}].$$

Since

$$(Lc)^{n+1} = ((Lc)^* - \{\epsilon\}) \cap (\Sigma^*c)^{n+1},$$

$r[(Lc)^{n+1}] \leqq n$ by Theorem 4.2. This is a contradiction. Thus L is regular if $(Lc)^* - \{\epsilon\}$ is ultralinear.

THEOREM 5.2. *It is recursively unsolvable to determine of an arbitrary pda M whether* Null(M) *is an ultralinear language.*

Proof. This is an immediate consequence of Lemma 5.1 and the fact that it is recursively unsolvable to determine whether a language is regular [4].

6. One-turn pda. We now consider the relation between a one-turn pda and f-transducers. The main result, Theorem 6.1, asserts the equivalence of the concepts of linear language, Null(M) for some one-turn pda, and $\bigcup_w wS_f(w^R)$ for some f-transducer S.

LEMMA 6.1. *Let* $S = (K, \Sigma, \Sigma, H, s_0, F)$ *be an f-transducer and let*

$$L = \bigcup_{w \in \Sigma^*} wS_f(w^R).$$

Then there is a one-turn pda M *such that* Null(M) $= L$ *and each move after the turn is unique (also, there is* M *such that* Null(M) $= L$ *and each increasing move is unique, that is, if* $\delta(q, x, Z)$ *contains* (q_1, γ_1) *and* (q_2, γ_2), *with* $|\gamma_1| \geqq 1$ *and* $|\gamma_2| \geqq 1$, *then* $q_1 = q_2$ *and* $\gamma_1 = \gamma_2$).

Proof. In view of Lemma 1.1, we may assume that S is a 1-restricted f-transducer. We first show that there is a one-turn pda M such that $L =$ Null(M) and each move after the turn is unique. By Lemma 1.3, there is a 1-restricted f-transducer $S' = (K', \Sigma, \Delta, H', s_0', F')$ such that $S_f(w^R) = [S_f'(w)]^R$ for each w in Σ^*. For each symbol x in $\Sigma \cup \{\epsilon\}$ let Z_x be an abstract symbol. Let Z_0 be a symbol not in $\{Z_x \mid x \text{ in } \Sigma \cup \{\epsilon\}\}$ and let $\Gamma = \{Z_0\} \cup \{Z_x \mid x \text{ in } \Sigma \cup \{\epsilon\}\}$. Let q_0 be a symbol not in K' and $K_M = K' \cup \{q_0\}$. Let M be the pda $(K_M, \Sigma, \Gamma, \delta, Z_0, s_0')$, where δ is defined as follows:

(a) $\delta(s_0', x, Z_0)$ contains $\{(q, Z_y) \mid (s_0', x, y, q) \text{ in } H'\}$,
(b) $\delta(q, x, Z)$ contains $\{(q', ZZ_y) \mid (q, x, y, q') \text{ in } H'\}$,
(c) $\delta(q, \epsilon, Z)$ contains (q_0, Z) if q is in F',
(d) $\delta(q_0, y, Z_y)$ contains (q_0, ϵ).

Then M is one-turn and Null(M) $= \bigcup_w w[S_f'(w)]^R = \bigcup_w wS_f(w^R)$. Furthermore, each move after the turn is unique (for the length does not start to decrease until q_0 is reached, after which each move is uniquely determined).

We now construct a one-turn pda M' such that $L =$ Null(M') and each increasing move is unique. Let Γ be as above and let $K_{M'} = K \cup \{q_0\}$. Let M' be the pda $(K_{M'}, \Sigma, \Gamma, \delta', Z_0, q_0)$, where δ' is defined as follows:

(a') $\delta'(q_0, x, Z)$ contains (q_0, ZZ_x),
(b') $\delta'(q_0, y, Z_x)$ contains $\{(q', \epsilon) \mid (q', x, y, q) \text{ in } H, q \text{ in } F\}$,
(c') $\delta'(q, y, Z_x) = \{(q', \epsilon) \mid (q', x, y, q) \text{ in } H\}$,
(d') $\delta'(s_0, \epsilon, Z_0)$ contains (s_0, ϵ).

It is readily verified that M' has the desired properties.

THEOREM 6.1. *For* $L \subseteq \Sigma^*$, *the following statements are equivalent*:

(1) $L =$ Null(M) *for some one-turn pda* M;
(2) L *is a linear language*;
(3) *there is an f-transducer* S *such that* $L = \bigcup_{w \in \Sigma^*} wS_f(w^R)$.

Proof. Theorem 2.1 gives the implication (1) $\rightarrow$ (2) and Lemma 6.1 gives the implication (3) $\rightarrow$ (1). Therefore we need only show that (2) $\rightarrow$ (3).

Let L be a linear language. As is easily seen, $L = L(G)$ for some linear grammar $G = (V, \Sigma, P, \sigma)$ in which every production is either of the form $\xi \rightarrow u\xi' v$, with u and v in Σ^*, or $\xi \rightarrow \epsilon$. Let S' be the f-transducer $(V - \Sigma, \Sigma, \Sigma, H, \sigma, F)$, where $F = \{\xi \mid \xi \rightarrow \epsilon \text{ is in } P\}$ and $H = \{(\xi, u, v^R, \gamma) \mid \xi \rightarrow u\gamma v \text{ in } P\}$. Then $L = \bigcup_w w[S_f'(w)]^R$. (For, z is in $L(G)$ if and only if

$$\sigma \Rightarrow u_1\xi_1 v_1 \Rightarrow \cdots \Rightarrow u_1 \cdots u_r\xi_r v_r \cdots v_1 \Rightarrow u_1 \cdots u_r v_r \cdots v_1 = z.$$

This occurs if and only if

$$(\sigma, u_1 \cdots u_r, \epsilon) \vdash_S (\xi_1, u_2 \cdots u_r, v_1^R) \vdash_S \cdots \vdash_S (\xi_r, \epsilon, v_1^R \cdots v_r^R),$$

that is, if and only if $z = u_1 \cdots u_r[S_f'(u_1 \cdots u_r)]^R$.) By Lemma 1.3, there exists an f-transducer S such that $S_f(w^R) = [S_f'(w)]^R$ for each w. Thus $L = \bigcup_w wS_f(w^R)$, so that (2) $\rightarrow$ (3).

Combining Lemma 6.1 and Theorem 6.1, we obtain the following corollary.

COROLLARY. *For each one-turn pda M, there exists a one-turn pda M' such that* Null(M) = Null(M') *and each move after the turn is unique (each increasing move is unique).*

REFERENCES

[1] E. B. ALTMAN, *The concept of finite representability*, Systems Research Center Report SRC 56-A-64-20, Case Institute of Technology, Cleveland, 1964.

[2] E. B. ALTMAN AND R. B. BANERJI, *Some problems of finite representability*, Information and Control, 8 (1965), pp. 251–263.

[3] R. B. BANERJI, *Phrase structure languages, finite machines, and channel capacity*, Ibid., 6 (1963), pp. 153–162.

[4] Y. BAR-HILLEL, M. PERLES, AND E. SHAMIR, *On formal properties of simple phrase structure grammars*, Z. Phonetik Sprachwiss. Kommunikat., 14 (1961), pp. 143–172.

[5] N. CHOMSKY, *On certain formal properties of grammars*, Information and Control, 2 (1959), pp. 137–167.

[6] ———, *Context-free grammars and pushdown storage*, Quarterly Progress Rpt. 65, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, 1962.

[7] N. CHOMSKY AND M. P. SCHUTZENBERGER, *The algebraic theory of context-free languages*, Computer Programming and Formal Systems, P. Braffort and D. Hirschberg, eds., North-Holland, Amsterdam, 1963, pp. 118–161.

[8] C. C. ELGOT AND J. E. MEZEI, *On relations defined by generalized finite automata*, IBM J. Res. Develop., 9 (1965), pp. 47–68.

[9] S. GINSBURG, *The Mathematical Theory of Context-Free Languages*, McGraw-Hill, New York, 1966.

[10] S. GINSBURG AND G. F. ROSE, *Operations which preserve definability in languages*, J. Assoc. Comput. Mach., 10 (1963), pp. 175–195.

[11] S. GINSBURG AND E. H. SPANIER, *Bounded ALGOL-like languages*, Trans. Amer. Math. Soc., 113 (1964), pp. 333–368.

[12] S. A. GREIBACH, *The unsolvability of the recognition of linear context-free languages*, J. Assoc. Comput. Mach., to appear.

[13] T. V. GRIFFITHS AND S. R. PETRICK, *On the relative efficiencies of context-free grammars*, Air Force Cambridge Research Laboratories Report, 1964.

[14] M. O. RABIN AND D. SCOTT, *Finite automata and their decision problems*, IBM J. Res. Develop., 3 (1959), pp. 114–125.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

AN INFORMATION-THEORETIC DERIVATION OF CERTAIN LIMIT RELATIONS FOR A STATIONARY MARKOV CHAIN*

S. KULLBACK†

A limit relation for the transition probabilities of a stationary Markov chain with a countable number of states is derived by the use of concepts and properties of information measures. Rényi [7] used information-theoretic concepts to derive the limit relation for the case of a finite number of states and Kendall [5] used a different but somewhat related approach for the case of a countable number of states. The approach in this paper is based on properties of the discrimination information as defined and developed in Kullback [6]. We shall follow the notation for Markov chains in Feller [2, Chap. XV].

Fréchet [3, p. 31], using properties of stochastic matrices of finite order, gave a necessary and sufficient condition that in a Markov chain with a finite number of states the m-step transition probabilities $p_{hk}^{(m)}$ tend to a limit p_k, independent of the first index h, as m increases indefinitely. Chung [1, pp. 26–30], using probabilistic considerations, gave a complete determination of the limit or limits of the mth step transition probabilities $p_{hk}^{(m)}$ as m increases indefinitely for every h and k for stationary Markov chains with a countable number of states. Feller [2, p. 357] has shown that in a Markov chain with a countable number of states a stationary distribution can exist only in the ergodic case. The main result of this paper is the converse of Feller's theorem, a result that we now state.

THEOREM. *For a stationary Markov chain with a countable number of states,* $\lim_{m\to\infty} p_{hk}^{(m)} = \mu_k$.

Consider a stationary Markov chain with transition probabilities

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \cdot & \cdot & \cdots \\ p_{n1} & p_{n2} & \cdots \\ \cdot & \cdot & \cdots \end{bmatrix}, \tag{1}$$

where

$$\sum_j p_{ij} = 1, \quad i = 1, 2, \cdots, \qquad p_{ij} > 0, \tag{2}$$

* Received by the editors December 8, 1965, and in revised form March 25, 1966.

† Department of Statistics, George Washington University, Washington, D. C.

This work was supported in part by the National Science Foundation Grant GP-3223 and by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grant AF-AFOSR 932-65.

and with the stationary distribution

$$\mu_j = \sum_i \mu_i p_{ij}, \quad j = 1, 2, \cdots, \qquad \sum_j \mu_j = 1. \tag{3}$$

There also exist the following known relations among the m-step transition probabilities:

$$p_{jk}^{(m+1)} = \sum_h p_{jh} p_{hk}^{(m)} = \sum_h p_{jh}^{(m)} p_{hk}, \qquad j, k = 1, 2, \cdots, \tag{4}$$

$$\sum_k p_{jk}^{(m)} = 1, \qquad j = 1, 2, \cdots. \tag{5}$$

We now consider the discrimination information between the systems of probabilities,

$$P_i^{(m)} : \{p_{i1}^{(m)}, p_{i2}^{(m)}, p_{i3}^{(m)}, \cdots\}, \qquad U : \{\mu_1, \mu_2, \mu_3, \cdots\}, \tag{6}$$

$$P_i^{(m+1)} : \{p_{i1}^{(m+1)}, p_{i2}^{(m+1)}, p_{i3}^{(m+1)} \cdots\}, \qquad U : \{\mu_1, \mu_2, \mu_3, \cdots\}, \tag{7}$$

given by (natural logarithms are used)

$$I(P_i^{(m)}; U) = \sum_j p_{ij}^{(m)} \log \frac{p_{ij}^{(m)}}{\mu_j}, \tag{8}$$

$$I(P_i^{(m+1)}; U) = \sum_j p_{ij}^{(m+1)} \log \frac{p_{ij}^{(m+1)}}{\mu_j}. \tag{9}$$

We shall need the convexity property (see [6]) that

$$\sum_i a_i \log \frac{a_i}{b_i} \geqq \left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i}, \tag{10}$$

where $a_i > 0$, $b_i > 0$, with equality if and only if

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \cdots = \frac{a_n}{b_n} = \cdots.$$

Using the convexity property and (2), (3), and (4), we may write

$$\begin{aligned} \sum_j p_{ij}^{(m+1)} \log \frac{p_{ij}^{(m+1)}}{\mu_j} &= \sum_j \sum_h p_{ih}^{(m)} p_{hj} \log \frac{\sum_h p_{ih}^{(m)} p_{hj}}{\sum_h \mu_h p_{hj}} \\ &\leqq \sum_j \sum_h p_{ih}^{(m)} p_{hj} \log \frac{p_{ih}^{(m)} p_{hj}}{\mu_h p_{hj}} = \sum_h p_{ih}^{(m)} \log \frac{p_{ih}^{(m)}}{\mu_h}. \end{aligned} \tag{11}$$

Thus

$$I(P_i^{(m)}; U) \geqq I(P_i^{(m+1)}; U), \tag{12}$$

with equality if and only if $p_{ih}^{(m)}/\mu_h = \text{const.}$ for $h = 1, 2, \cdots$. Since

$\sum_h p_{ih}^{(m)} = 1 = \sum_h \mu_h$, the constant is equal to one, and the condition for equality in (12) is $p_{ih}^{(m)} = \mu_h$. The convexity also implies that

$$I(P_i^{(m)}; U) \geqq 0 \tag{13}$$

for all m, with equality if and only if $p_{ih}^{(m)} = \mu_h$. Thus

$$\begin{aligned} I(P_i^{(1)}; U) \geqq I(P_i^{(2)}; U) \geqq \cdots \geqq I(P_i^{(m)}; U) \\ \geqq I(P_i^{(m+1)}; U) \geqq \cdots \geqq 0. \end{aligned} \tag{14}$$

Let us assume that $I(P_i^{(1)}; U) < \infty$. If there is equality some place in the sequence in (14), then

$$p_{ih}^{(m)} = \mu_h, \quad p_{ih}^{(m+1)} = \sum_k p_{ik}^{(m)} p_{kh} = \sum_k \mu_k p_{kh} = \mu_h, \tag{15}$$

and thus $p_{ih}^{(N)} = \mu_h$, $I(P_i^{(N)}; U) = 0$, $N \geqq m$.

On the other hand, if the sequence in (14) is strictly monotonic, and there is no equality, we shall show that

$$\lim_{m\to\infty} p_{ih}^{(m)} = \mu_h, \tag{16}$$

and hence $\lim_{m\to\infty} I(P_i^{(m)}; U) = 0$. Indeed,

$$\begin{aligned} &I(P_i^{(m)}; U) - I(P_i^{(m+1)}; U) \\ &= \sum_j \sum_h p_{ih}^{(m)} p_{hj} \log \frac{p_{ih}^{(m)}}{\mu_h} - \sum_j \sum_h p_{ih}^{(m)} p_{hj} \log \frac{p_{ij}^{(m+1)}}{\mu_j} \\ &= \sum_j \sum_h p_{hj}\, p_{ih}^{(m)} \log \frac{p_{hj}\, p_{ih}^{(m)}}{p_{hj} \dfrac{\mu_h}{\mu_j} p_{ij}^{(m+1)}}. \end{aligned} \tag{17}$$

Note that

$$\sum_j \sum_h p_{hj}\, p_{ih}^{(m)} = \sum_h p_{ih}^{(m)} \sum_j p_{hj} = \sum_h p_{ih}^{(m)} = 1, \tag{18}$$

and

$$\begin{aligned} \sum_j \sum_h p_{hj} \frac{\mu_h}{\mu_j} p_{ij}^{(m+1)} &= \sum_j \frac{p_{ij}^{(m+1)}}{\mu_j} \sum_h \mu_h\, p_{hj} \\ &= \sum_j \frac{p_{ij}^{(m+1)}}{\mu_j} \mu_j = \sum_j p_{ij}^{(m+1)} = 1. \end{aligned} \tag{19}$$

It is seen from (14) that the sequence $I(P_i^{(m)}; U)$ tends to a finite limit as $m \to \infty$, so that from (17),

$$\sum_j \sum_h p_{hj}\, p_{ih}^{(m)} \log \frac{p_{hj}\, p_{ih}^{(m)}}{p_{hj} \dfrac{\mu_h}{\mu_j} p_{ij}^{(m+1)}} \to 0 \text{ as } m \to \infty. \tag{20}$$

We shall show later that (20) implies

$$\sum_j \sum_h \left| p_{hj}\, p_{ih}^{(m)} - p_{hj} \frac{\mu_h}{\mu_j} p_{ij}^{(m+1)} \right| \to 0 \text{ as } m \to \infty, \tag{21}$$

or

$$\sum_j \sum_h \mu_h\, p_{hj} \left| \frac{p_{ih}^{(m)}}{\mu_h} - \frac{p_{ij}^{(m+1)}}{\mu_j} \right| \to 0 \text{ as } m \to \infty\,, \qquad i = 1, 2, \cdots. \tag{22}$$

Since we have assumed that $p_{hj} > 0$, we conclude from (3) that $\mu_h > 0$. Since

$$\mu_h\, p_{hj} \left| \frac{p_{ih}^{(m)}}{\mu_h} - \frac{p_{ij}^{(m+1)}}{\mu_j} \right| \leqq \sum_j \sum_h \mu_h\, p_{hj} \left| \frac{p_{ih}^{(m)}}{\mu_h} - \frac{p_{ij}^{(m+1)}}{\mu_j} \right|,$$

it follows from (22) that

$$\frac{p_{ih}^{(m)}}{\mu_h} - \frac{p_{ij}^{(m+1)}}{\mu_j} \to 0 \quad \text{as} \quad m \to \infty, \qquad i, j = 1, 2, \cdots. \tag{23}$$

It may be shown in a similar fashion that

$$\frac{p_{ih}^{(m)}}{\mu_h} - \frac{p_{ij}^{(m+n)}}{\mu_j} \to 0 \quad \text{as} \quad m, n \to \infty, \qquad i, j = 1, 2, \cdots, \tag{24}$$

since

$$I(P_i^{(m)}; U) - I(P_i^{(m+n)}; U) = \sum_j \sum_h p_{hj}^{(n)} p_{ih}^{(m)} \log \frac{p_{hj}^{(n)} p_{ih}^{(m)}}{p_{hj}^{(n)} \frac{\mu_h}{\mu_j} p_{ij}^{(m+n)}} \to 0 \tag{25}$$

as $m \to \infty$. For $h = j$ in (24) the Cauchy mutual convergence criterion implies that there is a C_{ij} such that

$$\lim_{m \to \infty} \frac{p_{ij}^{(m)}}{\mu_j} = C_{ij}\,. \tag{26}$$

Thus letting $m \to \infty$ in (23) we get

$$C_{ih} = C_{ij} \text{ for all } i, j, h. \tag{27}$$

But (26) implies that $\sum_{j=1}^{\infty} C_{ij}\mu_j = 1$, hence by (27) that $C_{ih} \equiv 1$ for all i and h; that is,

$$\lim_{m \to \infty} p_{ih}^{(m)} = \mu_h\,, \qquad i, h = 1, 2, \cdots. \tag{28}$$

(I am indebted to Minoru Sakaguchi for very helpful comments and discussion about the foregoing.)

We shall now justify the statement that (20) implies (21). Consider

$$(29)\quad \sum a_i \log \frac{a_i}{b_i}, \qquad a_i > 0, \qquad b_i > 0, \qquad \sum a_i = \sum b_i = 1,$$

which may be written as

$$\begin{aligned}(30)\quad \sum 2a_i^{1/2}a_i^{1/2} \log \frac{a_i^{1/2}}{b_i^{1/2}} &= 2(\sum_j a_j^{1/2}) \sum_i \frac{a_i^{1/2}}{(\sum_j a_j^{1/2})} a_i^{1/2} \log \frac{a_i^{1/2}}{b_i^{1/2}} \\ &\geqq 2(\sum_j a_j^{1/2}) \left[\sum_i \frac{a_i^{1/2}}{(\sum_j a_j^{1/2})} a_i^{1/2}\right] \\ &\quad \cdot \log \left[\left(\sum_i \frac{a_i^{1/2}}{(\sum_j a_j^{1/2})} a_i^{1/2}\right) \Big/ \left(\sum_i \frac{a_i^{1/2}}{(\sum_j a_j^{1/2})} b_i^{1/2}\right)\right] \\ &= 2 \log \frac{1}{\sum_i a_i^{1/2}b_i^{1/2}} = -2 \log \sum_i a_i^{1/2}b_i^{1/2},\end{aligned}$$

using the convexity property. But $\log x \leqq x - 1$ for all $x > 0$, so that

$$(31)\quad -2 \log \sum_i a_i^{1/2}b_i^{1/2} \geqq 2(1 - \sum_i a_i^{1/2}b_i^{1/2}) = \sum_i (a_i^{1/2} - b_i^{1/2})^2.$$

Using the Cauchy-Schwarz inequality,

$$\begin{aligned}(32)\quad (\sum_i | a_i - b_i |)^2 &= (\sum_i | a_i^{1/2} - b_i^{1/2} | (a_i^{1/2} + b_i^{1/2}))^2 \\ &\leqq \sum_i | a_i^{1/2} - b_i^{1/2} |^2 \sum_i (a_i^{1/2} + b_i^{1/2})^2,\end{aligned}$$

where

$$\begin{aligned}(33)\quad \sum_i (a_i^{1/2} + b_i^{1/2})^2 &= \sum_i (a_i + b_i + 2a_i^{1/2}b_i^{1/2}) \\ &\leqq 1 + 1 + 2(\sum a_i)^{1/2} (\sum b_i)^{1/2} = 4,\end{aligned}$$

so that

$$(34)\quad (\sum | a_i - b_i |)^2 \leqq 4 \sum (a_i^{1/2} - b_i^{1/2})^2.$$

Hence there follows the chain of inequalities (cf. [4])

$$\begin{aligned}(35)\quad \sum_i a_i \log \frac{a_i}{b_i} &\geqq -2 \log \sum_i a_i^{1/2}b_i^{1/2} \\ &\geqq \sum_i (a_i^{1/2} - b_i^{1/2})^2 \geqq \frac{1}{4} (\sum_i |a_i - b_i|)^2,\end{aligned}$$

from which it is seen that

$$\sum_i a_i \log \frac{a_i}{b_i} \to 0 \quad \text{implies} \quad \sum_i | a_i - b_i | \to 0.$$

REFERENCES

[1] K. L. CHUNG, *Markov Chains with Stationary Transition Probabilities*, Springer-Verlag, Berlin, 1960.

[2] W. FELLER, *An Introduction to Probability Theory and Its Applications*, 2nd ed., John Wiley, New York, 1957.

[3] M. FRÉCHET, *Recherches théoriques modernes sur le calcul des probabilités*, vol. II, *Méthode des fonctions arbitraires. Théorie des événements en chaine dans le cas d'un nombre fini d'états possibles*, Gauthier-Villars, Paris, 1938.

[4] J. HANNAN, *Consistency of maximum likelihood estimation of discrete distributions*, Contributions to Probability and Statistics, I. Olkin, et al., eds., Stanford University Press, Stanford, 1960, pp. 249–257.

[5] D. G. KENDALL, *Information theory and the limit-theorem for Markov chains and processes with a countable infinity of states*, Ann. Inst. Statist. Math., 15 (1964), pp. 137–143.

[6] S. KULLBACK, *Information Theory and Statistics*, John Wiley, New York, 1959.

[7] A. RÉNYI, *On a measure of entropy and information*, Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, 1961, pp. 547–561.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

SOME LIAPUNOV THEOREMS*

T. A. BURTON†

1. Introduction. We consider a system of differential equations

$$X' = F(X, t), \qquad X' = \frac{dX}{dt},$$

where $F(X, t)$ is continuous for $t \geqq t_0$ and X contained in some open set. Let $F(0, t) \equiv 0$ and assume that $t_0 = 0$.

The problem is to find conditions on a Liapunov function $V(X, t)$ such that we conclude that all solutions of (1) starting in a certain region tend to some set. This problem was initiated by A. M. Liapunov [1] who proved the following theorem for $F(X, t)$ independent of t.

For terminology we refer the reader to [2].

THEOREM 1. (Liapunov) *If $V(X)$ is positive definite and if $dV/dt = (\text{grad } V) \cdot F(X)$ is negative definite, then the null solution $X = 0$ is asymptotically stable.*

The requirements that V be positive definite and dV/dt negative definite were unnecessarily stringent and were improved by J. P. La Salle [3], [4]. Again it is assumed that $F(X, t)$ is independent of t.

THEOREM 2. (La Salle) *Assume that there exists for the system (1) a scalar function $V(X)$ which has continuous first partials and that $dV/dt \leqq 0$ for all X. Let E be the set defined by $dV/dt = 0$ and let M denote the largest invariant set in E. Then every solution bounded for $t > 0$ approaches M as $t \to \infty$.*

These results were extended by Krasovskii and Barbasin [5, p. 67] and also by Yoshizawa [6] to include systems where $F(X, t)$ is time dependent.

Our purpose is to extend these results still farther. We conclude the paper with an example which extends the results of Levin and Nohel [7].

For our results it is not necessary to require uniqueness of solutions of (1) and hence we assume only continuity of $F(X, t)$, Also, when we consider Liapunov functions $V(X, t)$, we assume only the existence of upper right-hand total derivatives as in [8, p. 143].

2. Extension of periodic results. Krasovskii and Barbasin [5, p. 67] consider the system (1) and obtain the following result.

THEOREM 3. *Let $F(X, t)$ be periodic in t with period T and Lipschitzian for $\| X \| < H$ and $-\infty < t < \infty$ ($H = \text{const.}$ or $H = \infty$). Suppose:*

* Received by the editors March 16, 1966.

† Department of Mathematics, University of Alberta, Edmonton, Alberta.

(a) *there exists a function* $V(X, t)$ *which is periodic with period* T *or does not depend explicitly on time;*
(b) $V(X, t)$ *is positive definite;*
(c) $V(X, t)$ *admits an infinitely small upper bound for* $\| X \| < H$ *and* $-\infty < t < \infty$;
(d) $\sup\{V \mid \| X \| \leqq H_0, 0 \leqq t < T\} < \inf\{V \mid \| X \| = H_1\}$, *where* $H_0 < H_1 < H$;
(e) $dV/dt \leqq 0$ *for* $\| X \| < H$ *and* $-\infty < t < \infty$;
(f) *the set* M *of points at which* $dV/dt = (\text{grad } V) \cdot F(X, t) + \partial V/\partial t$ *is zero contains no nontrivial half-trajectory of* (1).

Under these conditions, the null solution $X = 0$ *is asymptotically stable and the region* $\| X \| \leqq H_0$ *lies in the region of attraction of the point* $X = 0$.

To generalize this theorem it is convenient to introduce some new terminology.

DEFINITION. A *wedge* is a continuous function $W(X) \geqq 0$ defined for $\| X \| < H$, $H = \text{const.}$ or $H = \infty$, such that
(i) for each real k, if $W(X) = k$ has a real solution, then it is homeomorphic to an n-sphere;
(ii) if $W(X) = k_1$ and $W(X) = k_2$ have real solutions and if $k_1 < k_2$, then the set of points defined by $W(X) = k_1$ lies entirely inside the set of points defined by $W(X) = k_2$. Also, let $W(0) = 0$.

A nonempty set defined by $W(X) = k$ will be called an n-sphere.

THEOREM 4. *Suppose that there exists a function* $V(X, t)$, *defined and continuous for* $\| X \| < H$ ($H = \text{const.}$ *or* $H = \infty$) *and* $0 \leqq t < \infty$, *together with wedges* $W_1(X)$ *and* $W_3(X)$ *such that* $W_3(X) \geqq V(X, t) \geqq W_1(X)$ *and* $V(0, 0) = 0$. *Suppose also that* $dV/dt \leqq 0$ *for* $\| X \| < H$ *and* $t > 0$. *Let* $M = R \times [0, \infty)$ *contain all the points where* $dV/dt = 0$, *where* R *is closed and contained in an open set* J_1 *which in turn is contained in an open set* J_2 *satisfying the following properties. There exists an* $\epsilon_0 > 0$ *such that for* $0 < \epsilon < \epsilon_0$, $J_2 - S(0, \epsilon)$ *contains no positive trajectories of* (1). *Also, let the distance between* $\bar{J}_1 - S(0, \epsilon)$ *and* $[J_2 + S(0, \epsilon)]^c$ *be positive for* $\| X \| \leqq H$. *Assume that* $F(X, t)$ *is bounded for* X *on compact sets contained in* $\| X \| < H$ *and* $0 \leqq t < \infty$. *If, in addition, there exists a wedge* $W_2(X)$ *such that* $dV/dt \leqq -W_2(X)$ *in* $J_1^c \times [0, \infty)$ *and* $\| X \| \leqq H$, *then the null solution of* (1) *is asymptotically stable.*

Proof. By well-known Liapunov theorems $X = 0$ is Liapunov stable. Let $k_0 > 0$ be sufficiently small that $W_1(X) = k_0$ is an n-sphere inside $\| X \| < H$. Since $V(0, 0) = 0$ and $V(X, t)$ is continuous, there is a $\delta > 0$ such that if $\| X_0 \| < \delta$, then $V(X_0, 0) < k_0 = W_1(X)$. Since $dV/dt \leqq 0$, the solution $f(X_0, t)$ through X_0 satisfies $V(f(X_0, t), t) = V(t) < k_0$ for any $\| X_0 \| < \delta$. For each $\| X_0 \| < \delta$, $f(X_0, t) \to 0$ as $t \to \infty$ if and only if

$V(t) \to 0$ as $t \to \infty$ since $W_3(X) \geqq V(X, t) \geqq W_1(X)$. Suppose that for some $\| X_0 \| < \delta$, $f(X_0, t) \not\to 0$ as $t \to \infty$. Now $V(f(X_0, t), t)$ is nonincreasing, so $V(t)$ has a positive limit B as $t \to \infty$. Since $V(X, t) \leqq W_3(X)$ and $V(t) \to B$, $W_3(f(X_0, t)) \geqq V(f(X_0, t), t) \geqq B$. Therefore, $f(X_0, t)$ lies outside the n-sphere $W_3(X) = B$ for all $t \geqq 0$. Hence $f(X_0, t)$ is contained in the region S between the two n-spheres $W_3(X) = B$ and $W_1(X) = k_0$ for all $t \geqq 0$. Consider $(S - J_1) \times [0, \infty)$ in which $dV/dt \leqq -W_2(X)$. By hypothesis there exists $T > 0$ such that $f(X_0, t)$ is not contained in J_2 for all $t \geqq T$; $f(X_0, t)$ is not contained in J_2^c for all $t \geqq T$ since we could then conclude that $f(X_0, t) \to 0$ as $t \to \infty$ using standard Liapunov arguments. Thus there exist three sequences $\{t_n\}$, $\{t_n'\}$, and $\{t_n''\}$, each tending monotonically to $+\infty$ and satisfying the relation $t_n' > t_n'' > t_n$, such that $f(X_0, t)$ leaves J_1 at $t = t_n$, enters J_1 at $t = t_n'$ and is in $S - J_1$ for $t_n < t < t_n'$ and such that $f(X_0, t_n'')$ is in J_2^c. Now $F(X, t)$ is bounded in $S \times [0, \infty)$, and in S the distance between J_1 and J_2^c is positive, so the differences $\{(t_n' - t_n)\}$ have a positive infimum. Hence $\sum_{n=1}^{\infty} (t_n' - t_n) = \infty$. Let $L = \bigcup_{n=1}^{\infty} [t_n, t_n']$. Then on $[J_1^c \cap S] \times L$ we have $dV/dt \leqq -\alpha$ for some $\alpha > 0$ since $dV/dt \leqq -W_2(X)$. Therefore

$$\sum_{n=1}^{\infty} \int_{t_n}^{t_n'} \frac{dV}{dt}\, dt \leqq -\alpha \sum_{n=1}^{\infty} (t_n' - t_n) = -\infty.$$

But $V(X, t) \geqq 0$ so $V(t) \to 0$ as $t \to \infty$.

COROLLARY 1. *The n-sphere $W_1(X) = k_0$ bounds a region which is contained in the region of attraction.*

COROLLARY 2. *If $H = \infty$ and if for every X_0, $V(X_0, 0) = W_1(X)$ has a real solution, then the asymptotic stability is global.* (Note. This does not necessarily imply that $W_1(X) \to \infty$ as $\| X \| \to \infty$.)

3. An extension of Yoshizawa's results. T. Yoshizawa [6] considers a system

$$X' = F(X, t) + G(X, t), \tag{2}$$

where $F(X, t)$ and $G(X, t)$ are continuous for X contained in an open set Q and $0 \leqq t < \infty$. In addition F and G satisfy the following conditions.

(a) $F(X, t)$ tends to a function $H(X)$ for X in a "certain" set P contained in Q, and on any compact subset of P this convergence is uniform.

(b) For each $\epsilon > 0$ and each Y in P, there exist positive numbers $\delta(Y)$ and $T(Y)$ such that, if $\| X - Y \| < \delta(Y)$ and $t \geqq T(Y)$, then $\| F(X, t) - F(Y, t) \| < \epsilon$.

(c) If $X(t)$ is continuous and bounded on $[0, \infty)$ and if $X(t)$ is contained in any compact subset of Q, then $\int_0^{\infty} \| G(X(t), t) \|\, dt < \infty$.

The "certain" set P in (a) is explained in a subsequent theorem. One

should notice that boundedness of solutions is assumed in these theorems rather than requiring that the Liapunov functions guarantee the boundedness.

THEOREM 5. (Yoshizawa) *Suppose that a solution* $f(X_0, t)$ *of* (2) *is bounded and approaches a closed set* P *(that of condition* (a)*) contained in* Q *and that* $F(X, t)$ *and* $G(X, t)$ *satisfy conditions* (a), (b), *and* (c). *Then the positive limiting set of* $f(X_0, t)$ *is a semiinvariant set contained in* P *of* $X' = H(X)$, X *in* P.

The following definition is needed in the next theorem.

DEFINITION. (Yoshizawa) A scalar function $f(X)$ of X in Q is *positive definite with respect to a set* A if $f(X) = 0$ for X in A and for each $\epsilon > 0$ and each compact set Q^* in Q, there exists a number $\delta(Q^*, \epsilon) > 0$ such that $f(X) \geqq \delta(Q^*, \epsilon)$ for X in $Q^* \cap [\cup (A, \epsilon)]^c$. If $-f(X)$ is positive definite with respect to A, $f(X)$ is *negative definite with respect to* A.

THEOREM 6. (Yoshizawa) *Suppose that* $F(X, t)$ *is bounded for all* t *when* X *belongs to an arbitrary compact set in* Q *and that all solutions of* (1) *are bounded. Let* $G(X, t)$ *satisfy condition* (c). *Moreover, we assume that there exists a nonnegative function* $V(X, t)$ *such that* $dV/dt \leqq -W(X)$, *where* $W(X)$ *is positive definite with respect to a closed set* P *(not necessarily that of condition* (a)*) in the set* Q. *Then every solution of* (2) *approaches* P.

These last two theorems imply the following main result.

THEOREM 7. (Yoshizawa) *Suppose that all solutions of* (1) *are bounded, that is, every solution is contained in a compact set in* Q, *and that there exists a nonnegative function* $V(X, t)$ *such that* $dV/dt \leqq -W(X)$, *where* $W(X)$ *is positive definite with respect to a closed set* P *(that of condition* (a)*) in the set* Q. *Moreover, we assume that* $F(X, t)$ *is bounded for all* t *when* X *belongs to an arbitrary compact set in* Q, *and that* $F(X, t)$ *satisfies conditions* (a) *and* (b), *while* $G(X, t)$ *satisfies condition* (c). *Then all solutions of* (1) *approach the largest semiinvariant set contained in* P *of* $X' = H(X)$ *for* X *in* P.

We have been overly cautious to emphasize the set P and conditions (a) and (b) since it is only with very careful reading of Yoshizawa's paper that one can tell what is meant. In this connection, one may see that this last theorem has already been misinterpreted in [9, p. 5]; this results in a much weaker theorem.

Our purpose here is to alter conditions (a), (b), and (c) and still retain stability results. Although the last theorem is very attractive due to its precise statement of the limiting set, it may be very difficult to determine whether or not a nonempty limiting set exists. Also, the condition that $F(X, t)$ be asymptotically autonomous on certain sets is one which seems very stringent.

THEOREM 8. *Assume that every solution of* (1) *which starts in* Q *remains in some compact subset of* Q, *and that there exists a nonnegative function* $V(X, t)$ *defined for* X *in* Q *and* $0 \leqq t < \infty$ *such that* $dV/dt \leqq -W(X)$, *where* $W(X)$

is positive definite with respect to some closed set P contained in Q. Let $F(X, t)$ be bounded when X belongs to any compact subset of Q for all $t \geqq 0$. If for every $\epsilon > 0$ there exists an open set $J(\epsilon)$ containing $P - S(0, \epsilon)$ such that any solution of (1) which enters $J(\epsilon)$ eventually leaves $J(\epsilon)$, then every solution of (1) approaches the set of all X in P satisfied by

$$\min [\sup_{X \in Q} V(X, 0), \limsup_{t \to \infty} V(0, t)] \geqq V(X, t) \geqq \liminf_{t \to \infty} V(0, t).$$

Proof. Notice that 0 belongs to P since $F(0, t) = 0$. Theorem 6 shows that every solution $f(X_0, t_0, t)$ tends to P as t tends to infinity. This means that for every $\delta > 0$ there exists $T(\delta)$ such that $d(P, f(X_0, t_0, t)) < \delta$ for $t > T(\delta)$. Since $f(X_0, t_0, t)$ tends to P and $dV/dt = 0$ for X in P, it follows that dV/dt tends to 0 as t tends to infinity. Hence, $V(X, t)$ tends to a constant as t tends to infinity. We must show that the only possible constants are those stated in the theorem. Thus we must show that for each (X_0, t_0) there exists a sequence $\{t_n\}$ tending to infinity with n such that $f(X_0, t_0, t_n)$ tends to zero as n tends to infinity. This will be done if we can show that for each $\epsilon > 0$ there exists t' such that $f(X_0, t_0, t')$ is in $S(0, \epsilon)$. Assume that for some X_0 in Q and some t_0, $f(X_0, t_0, t)$ does not tend to zero as t tends to infinity. Let $\epsilon > 0$ be given, and let J be the open set containing $P - S(0, \epsilon)$ which $f(X_0, t_0, t)$ eventually leaves if $f(X_0, t_0, t)$ is in J for some $t > t_0$. Since J is open, there exists δ satisfying $0 < \delta < \epsilon$ such that $\cup S(P - S(0, \epsilon), \delta)$ is contained in J. Take T sufficiently large that $d(P, f(X_0, t_0, t)) < \delta$ if $t > T$. Now $f(X_0, t_0, t)$ leaves J so it must enter $S(0, \epsilon)$, because $\delta < \epsilon$ and $f(X_0, t_0, t)$ is contained in $J \cup S(0, \epsilon)$.

We shall now consider Yoshizawa's example [6, p. 386] which in turn extended work of Levin and Nohel [7].

4. An example. Consider the system

$$x' = y,$$
$$y' = -h(x, y, t)y - f(x) + e(t),$$

where

(i) $xf(x) > 0$ if $x \neq 0$, f continuous;

(ii) $F(x) = \int_0^x f(s)\, ds \to \infty$ as $|x| \to \infty$;

(iii) $e(t)$ is continuous for $t \geqq 0$ and $E(t) = \int_0^t |e(s)|\, ds < \infty$;

(iv) $h(x, y, t)$ is continuous and nonnegative for all x, all y, and all $t \geqq 0$, and is bounded if $x^2 + y^2$ is bounded;

(v) there exists an open set S with four open connected components $S_1^*, \cdots, S_4^*$ which are unbounded and each of which contains $(0, 0)$ on the boundary. Also, S_i^* is in quadrant i; and the positive x-axis is on the

boundary of S_1^* and S_4^*, while the negative x-axis is on the boundary of S_2^* and S_3^*. (For example, let S_1^* consist of all points in the strip $0 < x < \infty$ and $0 < y < \alpha$ for some $\alpha > 0$; S_2^*: $-\infty < x < 0$ and $0 < y < \alpha$; S_3^*: $-\infty < x < 0$ and $-\alpha < y < 0$; S_4^*: $0 < x < \infty$ and $-\alpha < y < 0$.) For (x, y) in S we have $y \not\equiv 0$ and $h(x, y, t) \geqq k(x, y) > 0$, where k is some continuous function.

Under these conditions every solution tends to the origin as $t \to \infty$.

To prove this, we follow Yoshizawa and let

$$V(x, y, t) = e^{-2E(t)}\left[F(x) + \frac{y^2}{2} + 1\right],$$

so $dV/dt \leqq -e^{-2E(\infty)}h(x, y, t)y^2$. Yoshizawa shows that all solutions are bounded. By Theorem 8 every solution approaches S^c. Now, $x' = y$ so for $y > 0$ every solution moves from left to right, and for $y < 0$ every solution moves from right to left. Hence, if a solution remains outside $S(0, \epsilon)$, for ϵ as small as we please, it must tend to the x-axis. But for $y = 0$ we have $y' = -f(x) + e(t)$, which is not zero identically if $x \not\equiv 0$. Hence the solutions do not remain on the x-axis for $x \not\equiv 0$. Thus by Theorem 8 all solutions tend to the origin.

Acknowledgments. The author is indebted to Professor D. Bushaw for a counterexample to a previous form of Theorem 8 and to Professor J. P. LaSalle for a great deal of assistance.

REFERENCES

[1] A. M. Liapounoff, *Problème général de la stabilité du movement*, Annals of Mathematics Study 17, Princeton University Press, Princeton. (This is the 1907 French translation of Liapunov's original paper published in Russia in 1892.)

[2] W. Hahn, *Theory and Application of Liapunov's Direct Method*, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

[3] J. P. LaSalle, *Complete stability of a nonlinear control system*, Proc. Nat. Acad. Sci. U. S. A., 48 (1962), pp. 600–603.

[4] ———, *Some extensions of Liapunov's second method*, IRE Trans. Circuit Theory, CT-7 (1960), pp. 520–527.

[5] N. N. Krasovskii, *Stability of Motion*, Stanford University Press, Stanford, 1963.

[6] T. Yoshizawa, *Asymptotic behavior of solutions of a system of differential equations*, Contributions to Differential Equations, 1 (1963), pp. 371–387.

[7] J. J. Levin and J. A. Nohel, *Global asymptotic stability for nonlinear systems of differential equations and applications to reactor dynamics*, Arch. Rational Mech. Anal., 5 (1960), pp. 194–211.

[8] H. Antosiewicz, *A survey of Lyapunov's second method*, Contributions to the Theory of Nonlinear Oscillations, vol. IV, Princeton University Press, Princeton, 1958, pp. 141–166.

[9] J. P. La Salle, *Recent advances in Liapunov stability theory*, SIAM Rev., 6 (1964), pp. 1–11.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

ON THE OPTIMAL CONTROL OF DISTRIBUTIVE SYSTEMS*

WILLIAM A. PORTER†

1. Introduction. In [1] the following basic optimization problem is considered. Let T be a bounded linear transformation between Banach spaces B and D respectively. With T onto D and $\xi \in D$ arbitrary, find (if one exists) a preimage of ξ with minimum norm. Of obvious interest are the questions of existence and uniqueness and the properties of the mapping from $\xi \in D$ to a minimum norm preimage. In [2] several generalizations of this problem are treated. These results, in an expanded and extended form, provide the basis for [3, Chap. IV], which also contains several applications to lumped parameter systems.

In this earlier development many different assumptions are made about T, its domain and range; however, the assumption that T is onto is always present. In the applications to lumped parameter systems this restriction is seldom a difficulty. Distributive systems, however, have mathematical models which typically involve a transformation with dense (or dense in a proper subspace) but not closed range. The present article deals with the additional difficulties that arise in the minimum norm problem as one removes the assumption that T is onto.

The following example considers a distributive control system whose mathematical model has the properties considered in the later analysis.

Example 1. The partial differential equation

$$x_t(t, \alpha) = kx_{\alpha\alpha}(t, \alpha) + f(t, \alpha), \qquad (t, \alpha) \in \Delta,$$

defined on the fixed domain $\Delta = (t_0, t_1) \times (0, b)$, frequently arises in the study of diffusion systems (see [4] or [3, Appendix IX]). It is considered here in conjunction with the auxiliary conditions

$$x(t, 0) = x(t, b) = 0, \quad t \in [t_0, t_1];$$

$$x(t_0, \alpha) = x^0(\alpha), \quad \alpha \in [0, b];$$

and the assumption that the force f is spatially concentrated. That is, f consists of a finite number of controls $\{g_1, \cdots, g_m\}$ located at fixed spatial positions $0 < \alpha_1 < \cdots < \alpha_m < b$. Accordingly f may be expressed, using Dirac delta functions, in the form

$$f(t, \alpha) = \sum_{i=1}^{m} \delta(\alpha - \alpha_i) g_i(t), \qquad (t, \alpha) \in \Delta.$$

* Received by the editors January 19, 1966.

† Systems Engineering Laboratory, The Department of Electrical Engineering, University of Michigan, Ann Arbor, Michigan. This work was supported by the United States Army Research Office, Durham, under Contract DA 31-124-ARD-D-391.

Using the separation of variables technique, it is not difficult to show (see [3]) that this distributive system can be identified with the ordinary differential system (of infinite order)

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{g}(t), \quad \mathbf{x}(t_0) = \mathbf{x}^0, \qquad t \in [t_0, t_1], \tag{1}$$

where $\mathbf{g} = (g_1, \cdots, g_n)$ is the control tuplet and $\mathbf{x} = (x_1, \cdots, x_n, \cdots)$ has components determined by

$$x_n(t) = \sqrt{\frac{2}{b}} \int_0^b x(t, \alpha) \sin \frac{n\pi\alpha}{b}\, d\alpha, \qquad n = 1, 2, \cdots.$$

The matrix A is diagonal and of infinite rank

$$A = -k \operatorname{diag} [(\pi/b)^2, (2\pi/b)^2, \cdots, (n\pi/b)^2, \cdots],$$

while the matrix B has m columns and infinitely many rows. The jth column of B is the tuplet

$$b_j = \operatorname{col} (\sin (\pi\alpha_j/b), \sin (2\pi\alpha_j/b), \cdots, \sin (n\pi\alpha_j/b), \cdots), \qquad j = 1, 2, \cdots, m.$$

The solution of (1) can also be written in the familiar form

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}^0 + \int_{t_0}^t \Phi(t, s)B\mathbf{g}(s)\, ds,$$

where the transition matrix Φ is diagonal and has infinite rank:

$$\Phi(t, s) = \operatorname{diag} [\exp \{-k(\pi/b)^2(t - s)\}, \cdots, \exp \{-k(n\pi/b)^2(t - s)\}, \cdots]. \tag{2}$$

Consider now the transfer $(t_0, x^0(\alpha)) \to (t_1, x^1(\alpha))$. This transfer may evidently be identified with the mapping $\xi = T\mathbf{g}$, where $\xi = \mathbf{x}^1 - \Phi(t_1, t_0)\mathbf{x}^0$ and T is computed by the equation

$$T\mathbf{g} = \int_{t_0}^{t_1} \Phi(t_1, s)B\mathbf{g}(s)\, ds. \tag{3}$$

If $B_1, \cdots, B_m$ are Banach spaces from which the individual controls must be chosen, and if $B = B_1 \times B_2 \times \cdots \times B_m$ is equipped with any reasonable norm (i.e., one which maintains the product topology), then B is also a Banach space and the problem of finding the preimage of $\xi = (\xi_1, \xi_2, \cdots)$ may be properly posed.

For the present objectives it suffices to consider the single input case (i.e., $m = 1$) and for convenience we choose $\alpha_1 = b/2$. The matrix B then becomes the column vector $B_1 = (1, 0, -1, 0, 1, \cdots)$. The transformation T now has (for example) the domain $B = L_p(t_0, t_1)$, $1 \leqq p < \infty$, and has values in an infinite-dimensional space. It is easy to see that T is

not onto. Indeed since Φ is diagonal and in view of the form of b_1, any ξ with a nonzero entry in some even position (i.e., $0 \neq \xi_2$, or $0 \neq \xi_4$, or $\cdots$) is not in the range of T. To cure this deficiency T may be considered as a mapping $\tilde{T}$ into the subspace X consisting of vectors with zero entries in all even positions. In the Appendix it is shown that the range of $\tilde{T}$ is dense in X but that $\tilde{T}$ is not onto. Thus T has a range which is dense but not closed.

2. The existence of minimum norm controls. Throughout this paper B and D will denote Banach spaces and T a bounded linear transformation from B into D. The unit ball and the unit sphere of B will be denoted by U and ∂U respectively, while $C = T(U)$ denotes the image of U in D under T. The boundary of C will be denoted by ∂C. It is easily verified that C is bounded, convex and circled. It follows from the open mapping theorem (see [5]), however, that C is a neighborhood of $0 \in D$ and hence that the closure of C is a convex body (a closed convex set with nonvacuous interior) if and only if T is onto.

In the earlier studies [1], [2], [3], the fact that $C \cup \partial C$ was a convex body implied, by means of the Hahn-Banach theorem (see [5]), the property that $C \cup \partial C$ has a hyperplane of support at every boundary point. This latter property played a key role in the earlier development and it is natural in the present setting to focus attention on the set S which consists of the support points of C. Without loss of generality attention may be restricted to real spaces, in which case

$$S = \{\xi \in C \mid \langle \xi, \varphi \rangle = \sup_{\eta \in C} \langle \eta, \varphi \rangle, \quad \text{some} \quad \varphi \in D^*\}.$$

It is apparent that $S \subset \partial C$. It is convenient also to single out the set

$$M = \{\xi \in \partial C \mid T^{-1}(\xi) \text{ has a minimum (norm) element}\},$$

where $T^{-1}(\xi)$ denotes the set of all preimages of ξ for further study.

LEMMA 1. *In general,* $C \cap S \subset M$.

Proof. Suppose $\xi \in S$ and $\lambda > 1$. Then the chain

$$\langle \lambda\xi, \varphi \rangle = \lambda \langle \xi, \varphi \rangle > \langle \xi, \varphi \rangle,$$

where φ supports C at ξ, shows that $\lambda\xi \notin C$. Thus if $\xi = Tu$, then $\lambda u \notin U$, which implies $\| u \| \geqq 1/\lambda$; and since $\lambda > 1$ is arbitrary, $\| u \| \geqq 1$. However, if $\xi \in C$, then at least one $u \in U$ exists such that $\xi = Tu$. Hence every $\xi \in C \cap S$ has a preimage in ∂U and each such element is minimal for the set $T^{-1}(\xi)$; thus $C \cap S \subset M$.

A natural question is whether or not this containment is proper. To investigate this situation pick an arbitrary $\xi \in M$ with the minimum preimage u_ξ. Assume that $\| u_\xi \| \geqq 1$, since this would be implied by

$\xi \in S$. Now, since $T^{-1}(\xi) = u_\xi + N(T)$, it follows easily that

$$\| u_\xi - v \| \geqq \| u_\xi \|, \quad v \in N(T).$$

Hence, if A^0 denotes the annihilator of A (see [5, §4.6]), the Hahn-Banach theorem guarantees a $\psi \in N(T)^0 \subset B^*$ with $\| \psi \| = 1$, such that

$$\langle u_\xi , \psi \rangle = \| u_\xi \|.$$

If also $\psi \in R(T^*)$, then a $\varphi \in D^*$ exists such that $\psi = T^*\varphi$. On one hand the relationship

$$\| u_\xi \| = \langle u_\xi , \psi \rangle = \langle Tu_\xi , \varphi \rangle = \langle \xi, \varphi \rangle \geqq 1$$

holds, while on the other we have

$$\langle \varphi, \eta \rangle \leqq \sup_{\eta \in C} \langle \varphi, \eta \rangle = \sup_{u \in U} \langle T^*\varphi, u \rangle = \| T^*\varphi \| = 1.$$

Hence

$$\langle \xi, \varphi \rangle \geqq \langle \eta, \varphi \rangle \quad \text{for all} \quad \eta \in C;$$

this implies that φ supports C at ξ or that $\xi \in S$. Thus if $M' \subset M$ is the subset whose elements ξ satisfy the two additional assumptions:
(i) the minimum elements of $T^{-1}(\xi)$ have norm $\geqq 1$, and
(ii) a Hahn-Banach produced ψ exists in $R(T^*)$,
then $M' \subset S$.

For any linear manifold $N \subset B^*$ the set ${}^0(N^0)$ can be identified as the weak * closure of N; consequently $N = {}^0(N^0)$ if and only if N is weak * closed. This happens, for instance, when N is finite-dimensional or if N is norm closed and B is reflexive. In general it is true that $R(T^*) \subset N(T)^0$; this containment may be strengthened to equality whenever $R(T^*)$ is weak * closed. Thus (ii) will always be satisfied whenever T^* has a weak * closed range.

Condition (i) is equivalent to the statement that ξ is not the image of an interior point of U. This, however, cannot be proved in general; for if C has a vacuous interior, then many points of C will be images of interior points of U. (Note that $T(\text{int } U)$ must be nonvacuous if $T \neq 0$.) This happens for instance when T is compact.

LEMMA 2. *If C is closed, $M' \subset S \subset M$. If also $R(T^*) = N(T)^0$, then $M' = M \cap S$.*

So far it has been assumed that the sets M', M, and S are nonvacuous. In many cases this is true. For instance, when B is reflexive, it follows that $T^{-1}(\xi)$ has a minimum element for every $\xi \in R(T)$. C is also closed (see [1]) in this case, and hence $M = \partial C = C$. As for the set S, Bishop and Phelps [6] have shown that if C is any closed convex subset of a Banach

space then the support points of C are dense in its boundary. Thus with B reflexive, both S and $M' = S \cap M$ are dense in C.

We have mentioned explicitly the condition that B is reflexive which guarantees a closed set C. Less restrictive but much more cumbersome conditions can be imposed on the problem. The reader is referred to [3, §4.3] for a discussion of these conditions.

3. The form of minimum elements. Attention now turns to the problem of determining a representation for the minimum norm preimages of a point. In doing so the following notation will be helpful. The Minkowski functional of the set C (see [5] or [3]) will be denoted by p. The element $\bar{f} \in B$ will be called an *extremal* (see [1] or [3]) of $f \in B^*$ if $\| \bar{f} \| = 1$ and $\langle \bar{f}, f \rangle = \| f \|$ both hold. The set of all extremals for $f \in B^*$ is denoted by $\{\bar{f}\}$.

THEOREM 1. *Let Q denote the set of all $\varphi \in D^*$ such that $\{\overline{T^*\varphi}\} \subset B$ is nonempty. Then $C \cap S = \{\xi = T(\overline{T^*\varphi}) \mid \varphi \in Q\}$. Furthermore, if $\xi = T(\overline{T^*\varphi})$, $\varphi \in Q$, then φ supports C at ξ.*

Proof. Suppose $\varphi \in Q$ and $u_\varphi \in \{\overline{T^*\varphi}\}$. Clearly $\xi_\varphi = Tu_\varphi$ is in C. However, since every $\eta \in C$ has the form $\eta = Tu$, $u \in U$, the chain

$$\langle \eta, \varphi \rangle = \langle u, T^*\varphi \rangle \leqq \| T^*\varphi \| = \langle u_\varphi , T^*\varphi \rangle = \langle \xi_\varphi , \varphi \rangle$$

is valid, which implies that φ supports C at ξ_φ. This shows that

$$\{T(\overline{T^*\varphi}) \mid \varphi \in Q\} \subset C \cap S.$$

Now, if $\xi \in C \cap S$, a $\varphi \in D^*$ exists such that

$$\langle \xi, \varphi \rangle = \sup_{\eta \in C} \langle \eta, \varphi \rangle = \| T^*\varphi \|.$$

Also by Lemma 1, $T^{-1}(\xi)$ has a minimum element $u_\xi \in \partial U$, and hence

$$\langle \xi, \varphi \rangle = \langle u_\xi , T^*\varphi \rangle = \| T^*\varphi \|$$

holds. This shows that $u_\xi \in \{\overline{T^*\varphi}\}$; thus $C \cap S \subset \{T(\overline{T^*\varphi}) \mid \varphi \in Q\}$.

In this theorem two types of nonuniqueness can occur. Each point $\xi \in S$ may be supported by more than one hyperplane. Secondly, the set $\{\overline{T^*\varphi}\}$ for each $\varphi \in Q$ may contain more than one point. It was noted in [1] (where T is assumed onto) that the second situation does not occur when B is rotund, and the first situation does not occur whenever B is rotund and smooth. If T has dense range, these results still hold with the same proofs. As before these sufficient conditions become necessary conditions if uniqueness is required for all linear transformations. However, examples do exist (see [3, §4.3, Exercise 4]), where φ_ξ and u_ξ are unique although B is neither rotund nor smooth.

COROLLARY. *Assume that B is rotund and smooth. Then $T^{-1}(\xi)$ has a minimum element u_ξ if $[p(\xi)]^{-1}\xi \in C \cap S$. For each such ξ, u_ξ is uniquely given by*

$$u_\xi = p(\xi)\overline{T^*\varphi_\xi},$$

where φ_ξ defines the unique hyperplane of support at $[p(\xi)]^{-1}\xi$. The functional φ_ξ is uniquely determined by $\| \varphi_\xi \| = 1$ and either of the conditions

(a) $$\langle \eta, \varphi_\xi \rangle \leqq [p(\xi)]^{-1}\langle \xi, \varphi_\xi \rangle, \quad \eta \in C,$$

(b) $$\| T^*\varphi_\xi \| = [p(\xi)]^{-1}\langle \xi, \varphi_\xi \rangle.$$

Several related problems which are suggested by the applications are discussed in [2] or [3]. In most cases the assumption that T is onto can be removed. Since the proofs given in these references can be modified, using the present discussion in an obvious manner, we shall not deal with these matters here.

4. Conclusions. It is shown that the minimum norm control problem can be meaningfully treated where the system transformation has dense but not closed range. The recent result that the support points of every closed convex set in a Banach space are dense in its boundary preserves much of the usefulness of the Hahn-Banach theorem in analyzing this problem.

5. Acknowledgment. The several discussions with Dr. James P. Williams over the past few months have materially contributed to the growth of this paper.

Appendix. Let T be as defined in the example. Then T may be represented in the dyadic form

$$Tu = \sum_{i=1}^{\infty} e_i\langle u, f_i \rangle,$$

where $\{e_i\}$ consists of the odd members of the usual coordinate basis for the vector space l_2 and f_i is the functional

$$\langle u, f_i \rangle = (-1)^{i+1} \int_{t_0}^{t_1} \exp\{-k[(2i-1)\pi/b]^2(t_1 - s)\}u(s)\, ds,$$

$$i = 1, 2, \cdots.$$

For the purposes of the example it suffices to treat $B = L_2(t_0, t_1)$, in which case it is easily verified that

$$\| f_i \| \leqq \gamma/(2i-1), \qquad i = 1, 2, \cdots.$$

where γ is a constant. From the inequality chain

$$\| Tu \| = \left[\sum_{i=1}^{\infty} | \langle u, f_i \rangle |^2 \right]^{1/2} \leqq \gamma \left[\sum_{i=1}^{\infty} \left(\frac{1}{2i - 1} \right)^2 \right]^{1/2} < \infty,$$

where the orthonormality of the set $\{e_i\}$ is used, it follows that $R(T) \subset l_2$.

To see that $R(T)$ is dense in X (the subspace spanned by $\{e_i\}$ equipped with the l_2 norm), we note that for arbitrary $\xi \in X$ and any $\epsilon > 0$, an N exists such that $\| \xi - \xi_N \| \leqq \epsilon$, where ξ and ξ_N have the same components ξ_i for $i \leqq N$ and ξ_N has zero entries otherwise. Since the functionals $\{f_i\}$ are linearily independent it follows easily that each ξ_N has a preimage under T. Thus choosing $\epsilon = 1/n$ we conclude that every $\xi \in X$ is a limit point of the range of T.

To show that T is not onto X we observe that if $u \in T^{-1}(e_i)$, then $\| u \| \geqq \| f_i \|^{-1}$, hence

$$\| f_i \|^{-1} \leqq p(e_i),$$

where p is the Minkowski functional of $C = T(U)$. Also $\| f_i \|^{-1} \to \infty$ as $i \to \infty$; this implies that C is not a neighborhood of $0 \in X$ and hence T is not onto.

REFERENCES

[1] W. A. Porter and J. P. Williams, *A note on the minimum effort control problem*' J. Math. Anal. Appl., 13(1966), pp. 251–264.

[2] ———, *Extensions of the minimum effort control problem*, Ibid., 13(1966), pp. 536–549.

[3] W. A. Porter, *Modern Foundations of System Engineering*, Macmillan, New York, 1966, to appear.

[4] P. K. C. Wang, *Control of distributive parameter systems*, Advances in Control Systems, vol. 1, Academic Press, New York, 1964, pp. 75–171.

[5] A. E. Taylor, *Introduction to Functional Analysis*, John Wiley, New York, 1958.

[6] E. Bishop and R. R. Phelps, *The support functionals of a convex set*, Proceedings of Symposia in Pure Mathematics, vol. VII, American Mathematical Society, Providence, 1963.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

PERTURBATIONS OF OPTIMAL CONTROL PROBLEMS*

JANE CULLUM†

1. Introduction. Let P be an optimal control problem describable by a system of ordinary differential equations and with an integral cost functional. Perturb the system of differential equations associated with P, the boundary conditions, and/or the control sets in such a way that the alterations are describable by a single parameter ϵ. Denote the corresponding problem generated from P by $P(\epsilon)$. If optimal solutions exist for each $P(\epsilon)$, what is the relationship between these optimal solutions and optimal solutions of the original problem $P = P(0)$? That is, if the parameter of perturbation ϵ is reduced to zero in some continuous manner, do the optimal solutions of $P(\epsilon)$, if they exist, converge in some sense to an optimal solution of P. This is the question that is to be investigated.

2. History. Kirillova[5] considered this question for a particular linear, time optimal control problem P, namely, the transfer of a given point a_0 to the origin in minimum time along a trajectory of the system $\dot{x} = A(t)x + B(t)u$ such that each of the components of the control u is less than one in magnitude. Kirillova perturbed only the differential system associated with P, and she assumed each system was linear, $\dot{x} = A(\epsilon, t)x + B(\epsilon, t)u$. Under the assumption of continuity of the matrices $A(\epsilon, t)$ and $B(\epsilon, t)$ in ϵ and in t and the assumption that each system is normal, Kirillova proved that, given $\epsilon_n \to 0$, the optimal times for the corresponding problems $P(\epsilon_n)$ converge to the optimal time for $P = P(0)$, and the corresponding sequence of optimal controls converges in measure to the optimal control for P.

The object of this paper is to extend these results to other problems. However, in general, the families of perturbed problems do not possess the following three properties which the family considered by Kirillova does possess:

(a) The associated differential systems are linear.

(b) The optimal control for each $P(\epsilon)$ is unique.

(c) The optimal control and trajectory for each $P(\epsilon)$ can be expressed explicitly in terms of known functions.

Extensive use was made by Kirillova of each of these three properties.

* Received by the editors February 1, 1966, and in revised form March 25, 1966.

† Department of Mathematics, University of California, Berkeley, California.

This research was supported by the United States Office of Naval Research under Contract Nonr 222(88). This paper is part of a dissertation submitted in partial satisfaction of the requirements for the Ph.D. degree in applied mathematics at the University of California, Berkeley.

3. Statement of the problems. E^n denotes n-dimensional Euclidean space. Let T be a fixed interval in E^1. For each ϵ, move from $G_0(\epsilon)$ along an absolutely continuous curve $(\hat{x}, \bar{I})$, where $\bar{I} = [t_0, t_1]$ is an interval contained in T, for which there exists a measurable function u, such that, $u(t) \in U(t, \epsilon)$ a.e. in $\bar{I}$ and such that $\dot{\hat{x}}(t) = \hat{f}(\epsilon, \hat{x}(t), u(t), t)$ a.e. in $\bar{I}$, $\hat{x}(t) \in A$ on $\bar{I}$, $\hat{x}(t_1) \in G_1(\epsilon, t_1)$, and such that the integral, the cost of $\hat{x}$,

$$C(\hat{x}) = \int_I f_0^0(\hat{x}(t), u(t), t)\, dt,$$

is minimized.

This problem will be designated by $P(\epsilon) = P(f(\epsilon), G_0(\epsilon), G_1(\epsilon, t), U(t, \epsilon), T, A)$. It can be reformulated as a problem in E^{n+1} with $x = (x^0, \hat{x}) \in E^{n+1}$, where $\dot{x}^0 = f_0^0$. This formulation will also be denoted by $P(\epsilon)$. An admissible pair for $P(\epsilon)$, that is, a trajectory and its control that satisfy the differential system and the control and space variable constraints for $P(\epsilon)$, will be denoted by $(x(\epsilon), u(\epsilon), \bar{I}, a, b)$ where $\bar{I}$ is the domain of $x(\epsilon)$ and of $u(\epsilon)$, and a and b are, respectively, the initial and terminal values of $x(\epsilon)$.

The global assumptions are:

(1) f_0^0 is continuous;

(2) for each ϵ, $f(\epsilon, \hat{x}, u, t)$ is continuous in $(\hat{x}, u)$ for each t and measurable in t for each $(\hat{x}, u)$, where $(\hat{x}, u, t) \in A \times \bar{\mathfrak{U}} \times T$;

(3) $G_0(\epsilon)$, $0 \leqq \epsilon \leqq 1$, is a family of compact subsets of E^n that is upper semicontinuous with respect to inclusion in ϵ at $\epsilon = 0$;

(4) $G_1(\epsilon, t)$ is a family of compact subsets of E^n that is upper semicontinuous in (ϵ, t) at each point $(0, t) \in \{0\} \times T$;

(5) $U(t, \epsilon)$ is a family of compact subsets of E^r that is upper semicontinuous with respect to inclusion in t for each ϵ, and in ϵ at $\epsilon = 0$ for each t; and such that $U(t, \epsilon) \downarrow U(t)$ for each t as $\epsilon \downarrow 0$;

(6) A is a closed subset of E^n.

Finally, the following notation will be used throughout the paper: $U(t, 0) = U(t)$; $G_1(0, t) = G_1(t)$; $G_0(0) = G_0$; $f(0, x, u, t) = f_0(x, u, t)$; $f(\epsilon, x, u, t) = (f_0^0(\hat{x}, u, t), \hat{f}(\epsilon, \hat{x}, u, t))$; I denotes an open interval in E^1; $\bar{I}$ denotes a closed interval; $\bar{\mathfrak{U}}$ denotes the union over $(t, \epsilon) \in T \times [0, 1]$ of the sets $U(t, \epsilon)$; and given a set F and a positive number δ, the symbol $U(F, \delta)$ denotes the union of all closed spheres with centers in F and radius δ.

4. Approximation types. Following Russell's terminology [9] two definitions will be made.

DEFINITION 4.1. A sequence of functions $(x_n, \bar{I}_n, a_n, b_n)$ is said to be an *approximation of type* 1 to the function $(x_0, \bar{I}_0, a, b)$ if $I_n \to I_0$, $(a_n, b_n) \to (a, b)$ and $x_n(t)$ converges to $x_0(t)$ for each $t \in I_0$.

DEFINITION 4.2. A sequence of pairs of functions $(x_n, u_n, \bar{I}_n, a_n, b_n)$ is said to be an *approximation of type* 2 to the pair $(x_0, u_0, \bar{I}_0, a, b)$ if $(x_n, \bar{I}_n, a_n, b_n)$ is an approximation of type 1 to $(x_0, \bar{I}_0, a, b)$, and if there exist measurable extensions $\bar{u}_n$ of u_n to I_0 such that $\bar{u}_n$ converges to u_0 in the weak topology of $L_2(I_0)$.

A third definition could also be made of an approximation of type 3 by changing the word weak in Definition 4.2 to the word strong. It is a trivial matter to prove that in Kirillova's case, convergence of the optimal controls in measure to an optimal control for the original problem implies pointwise convergence of the corresponding optimal trajectories to the optimal trajectory for the original problem. Hence, Kirillova proved that any sequence of optimal pairs for the problems $P(\epsilon_n)$, where $\epsilon_n \to 0$ as $n \to \infty$, is an approximation of type 3 to the optimal pair for P.

5. Theorem 1. The following result involves approximations of type 1.

THEOREM 1. *Let* $P(\epsilon)$, $0 \leqq \epsilon \leqq 1$, *be a family of problems satisfying the assumptions in* §3 *and such that*:

(a) *For each* $(x, t, \gamma) \in (E^{n+1} \times T \times R^+)$ *there exists* $\delta = \delta(x, t, \gamma) > 0$ *such that for all* $u \in \bar{\Omega}$, $|x - x_1| + |\epsilon| < \delta$ *implies that* $|f(\epsilon, x_1, u, t) - f_0(x, u, t)| < \gamma$.

(b) *There exist functions* μ *and* g *mapping* E^1 *into* E^1 *such that* $\mu \in L_1$, $g(s) = O(s)$ *as* $s \to \infty$, g *is bounded on bounded sets and for all* $(\epsilon, x, u, t) \in [0, 1] \times A \times \bar{\Omega} \times T$, $|\hat{f}(\epsilon, \hat{x}, u, t)| \leqq \mu(t) g(|\hat{x}|)$.

(c) *For each* $(\hat{x}, t) \in A \times T$, *the set* $f_0(\hat{x}, U(t), t)$ *is convex.*

(d) f_0 *is continuous on* $A \times \bar{\Omega} \times T$.

Then, if $\epsilon_n \to 0$ *and* $(x(n), \bar{I}_n, a_n, b_n)$ *is any sequence of admissible trajectories for the corresponding problems* $P(\epsilon_n)$,

(1) *there exists a pair* $(x_0, u_0, \bar{I}_0, a, b)$ *such that* x_0 *is an admissible trajectory for* P *with control* u_0,

(2) *there exists a subsequence of* $x(n)$ *that is an approximation of type* 1 *to* x_0.

Before proceeding with the proof of Theorem 1, the lemmas and theorems needed in the proof will be listed.

LEMMA 1. *Let* $\{U(t, \epsilon) \mid (t, \epsilon) \in [0, 1] \times T\}$ *be a family of compact subsets of* E^r *such that for each* t, $U(t, \epsilon) \downarrow U(t)$ *as* $\epsilon \downarrow 0$ *and for each* ϵ, $U(t, \epsilon)$ *is upper semicontinuous in* t. *Then this family is uniformly bounded.*

Proof. Clearly, all that needs to be proved is that the family $\{U(t, 1) \mid t \in T\}$ is bounded. But the upper semicontinuity implies that for each $t \in T$ there exists a $\delta(t) > 0$ such that $|\bar{t} - t| < \delta(t)$ implies that $U(\bar{t}, 1) \subset U(U(t, 1), 1)$. Therefore, by the Heine-Borel theorem there exist $t_1, \cdots, t_N$ such that $t \in T$ implies that $U(t, 1) \subset \bigcup_{i=1}^{N} (U(t_i, 1), 1)$.

Lemmas 2, 3, and 4 are well-known lemmas and no proofs will be given for them. Actually Lemma 3 is well-known for the case $Y = E^m$, but the extension to the general case where Y is any dense subset of E^m is trivial.

LEMMA 2. *Hypothesis* (b) *in Theorem* 1 *implies that the admissible trajectories for the problems* $P(\epsilon)$ *are uniformly bounded over* ϵ.

LEMMA 3. *If* C *is a convex compact subset of* E^m *and* Y *is a dense subset of* E^m, *then* $x_0 \in C$ *if and only if for all* $y \in Y$,

$$\min_{x \in C} (y, x) \leqq (y, x_0) \leqq \max_{x \in C} (y, x).$$

LEMMA 4. *Let* $\phi_n : [a, b] \to E^m$, $n = 1, 2, \cdots$, *be a sequence of* L_1 *functions such that* $|\phi_n| \leqq \mu$ *for some* $\mu \in L_1$ *and* ϕ_n *converges in the weak topology of* L_1 *to* $\phi_0 \in L_1$. *Then for each* $y \in E^m$,

$$\limsup_n (y, \phi_n) \geqq (y, \phi_0) \geqq \liminf_n (y, \phi_n) \quad a.e.$$

LEMMA 5. *Let* U_n, $n = 1, 2, \cdots$, *be a sequence of compact subsets of* E^r *such that*

(a) $U_n \downarrow U_0$;

(b) *for every* $\delta > 0$ *there exists* $N(\delta)$ *such that* $n > N(\delta)$ *implies*

$$U_n \subset U(U_0, \delta).$$

Let $g(\epsilon, x, u)$ *be a map from* $E^1 \times E^n \times E^r$ *into* E^1 *such that*

(c) g *is continuous in* u *for each fixed* (ϵ, x);

(d) *given* $(\epsilon, x, \gamma) \in [0, 1] \times E^n \times R^+$ *there exists* $\delta = \delta(\epsilon, x, \gamma) > 0$ *such that when* $|x - x_1| + |\epsilon| < \delta$, $|g(\epsilon, x_1, u) - g(0, x, u)| < \gamma$.

Then if $\epsilon_n \to 0$ *and* $x_n \to x_0$,

$$g_n = \max_{u \in U_n} g(\epsilon_n, x_n, u) \to \max_{u \in U_0} g(0, x_0, u) = g_0.$$

Proof. Set $g_n^* = \max_{u \in U_n} g(0, x_0, u)$ and $h(u) = g(0, x_0, u) = g_0(x_0, u)$. Clearly, h is uniformly continuous on U_1. Given $\gamma > 0$ there exists $\delta > 0$ such that $u, u_1 \in U_1$, and $|u - u_1| < \delta$ imply $|h(u) - h(u_1)| < \gamma/2$. Furthermore, there exist u_n, u_n^*, u_0 such that $g_n = g(\epsilon_n, x_n, u_n)$, $g_n^* = h(u_n^*)$, and $g_0 = h(u_0)$.

Clearly, for all n, $g_n^* \geqq g_0$. By hypothesis (b) there exists N such that for any $n > N$ such that $g_n^* > g_0$ there exists $u_n^0 \in U_0$ such that

$$|u_n^0 - u_n^*| < \delta$$

and hence $|h(u_n^0) - g_n^*| < \gamma/2$. Therefore, for $n > N$,

$$g_n^* > g_0 > h(u_n^0) > g_n^* - \gamma/2.$$

That is, $|g_n^* - g_0| < \gamma/2$.

Since $x_n \to x_0$ and $\epsilon_n \to 0$, by hypothesis (d) there exists N_1 such that for all $n > N_1$ and $u \in U_1$, $|g(\epsilon_n, x_n, u) - h(u)| < \gamma/2$. In particular, $|g_n^* - g(\epsilon_n, x_n, u_n^*)| < \gamma/2$ and $|h(u_n) - g_n| < \gamma/2$. Combining

these statements one obtains for $n > N_1$,

$$g_n \geqq g(\epsilon_n, x_n, u_n^*) > g_n^* - \gamma/2 > h(u_n) - \gamma/2 > g_n - \gamma.$$

That is, $|g_n - g_n^*| < \gamma/2$.

Therefore, for all $n > \max(N, N_1)$, $|g_n - g_0| < \gamma$.

COROLLARY 5.1. *Given the hypotheses of Lemma* 5,

$$k_n = \min_{u \in U_n} g(\epsilon_n, x_n, u) \to \min_{u \in U_0} g(0, x_0, u) = k_0.$$

Proof. Set $g_n = \max_{u \in U_n} (-g(\epsilon_n, x_n, u))$ and $g_0 = \max_{u \in U_0} (-g(0, x_0, u))$. Then $h_n = -g_n$ and $h_0 = -g_0$. Apply Lemma 5.

Finally, Filippov's lemma and a theorem from Dunford and Schwartz are needed.

LEMMA 6 (Filippov). *Let f be a continuous map of $E^1 \times E^r$ into E^{n+1}. Let the sets $U(t)$ be compact subsets of E^r that are upper semicontinuous in t. Let $y(t)$ be a measurable function such that $y(t) \in f(t, U(t))$ a.e. Then there exists a measurable function $u(t) \in U(t)$ a.e. such that $y(t) = f(t, u(t))$ a.e.*

Proof. See [3].

THEOREM A. *A subset K of $L_1(I)$ is weakly sequentially compact if and only if it is bounded in the L_1-norm and the countable additivity of the integrals $\int_E f(s)\,ds$ is uniform with respect to all $f \in K$.*

Proof. See [1].

6. Proof of Theorem 1. The methods used in the following proof are generalizations of methods used by Roxin [8]. The proof consists of constructing an absolutely continuous curve x_0 on an interval $\bar{I}_0$ with initial value $a_0 \in G_0$, and then proving that x_0 is admissible for P. It is assumed without loss of generality that $\epsilon_n \downarrow 0$.

Let $\delta_m \downarrow 0$. By hypothesis there exists $N(\delta_m)$ such that for all

$$n > N(\delta_m),$$

$a_n \in U(G_0, \delta_m) = Q_m$. But by the Bolzano-Weierstrass theorem there exists a subsequence of the a_n, without loss of generality denoted by a_n, that converges to $a_0 \in Q_m$ for all m. Hence $a_0 \in G_0 = \bigcap_{i=1}^{\infty} Q_i$.

For this subsequence, by hypothesis each $\bar{I}_n = [t_n^0, t_n^1] \subset T$. Hence, there exist a subsequence and an interval $\bar{I}_0 = [t_0, t_1] \subset T$ such that $I_n \to I_0$. It should always be kept in mind that the original sequences do not necessarily converge in any sense. To simplify the notation throughout the paper any sequence considered will be denoted by the original sequence; but in general, the sequence being considered is a proper subsequence of the original sequence.

Extend each of the derivatives of the $x(n)$ to functions defined on all of $\bar{I}_0$. That is, define

$$\phi(n, t) = \begin{cases} f(\epsilon_n, x(n, t), u(n, t), t) & \text{if } t \in I_n, \\ f(\epsilon_n, x(n, t_n^1), u(n, t_n^1), t) & \text{if } t \in [t_n^1, t_1], \\ f(\epsilon_n, x(n, t_n^0), u(n, t_n^0), t) & \text{if } t \in [t_0, t_n^0]. \end{cases} \tag{6.1}$$

If $t_n^0 \leqq t^0$ and/or $t_n^1 \geqq t^1$, then obvious adjustments should be made in the preceding definitions.

By hypothesis (b) there exist R, $M > 0$ such that $|\hat{x}(\epsilon)| < R$ for any trajectory of $P(\epsilon)$, and such that on $S = \{|\hat{x}| \leqq R\} \times \bar{U} \times T$, $|\hat{f}(\epsilon, \hat{x}, u, t)| \leqq M\mu(t)$. By the continuity of f_0^0 and Lemma 2 there exists an M_0 bounding f_0^0 on S. Therefore, there exists $\mu_1 \in L_1$ such that for all n,

$$|\phi(n, t)| \leqq \mu_1(t). \tag{6.2}$$

But (6.2) implies that the $\phi(n)$ satisfy the hypotheses of Theorem A. Therefore, there exists a subsequence of $\phi(n)$ that converges in the weak topology of L_1 to $\phi_0 \in L_1$.

Define for $t \in \bar{I}_0$,

$$x_0(t) = a_0 + \int_{t_0}^{t} \phi_0,$$

and prove that

(i) $x(n, t) \to x_0(t)$ for each $t \in I_0$,
(ii) $b_n \to x_0(t_1) = b_0$,
(iii) $b_0 \in G_1(t_1)$, and
(iv) $\hat{x}_0(t) \in A$ for all $t \in \bar{I}_0$.

(i) Let $\gamma > 0$ be given and $t \in I_0$. Then there exists $N(t)$ such that for all $n > N(t)$, $t \in I_n$. Hence,

$$|x(n, t) - x_0(t)| = \left| \int_{t_n^0}^{t} f(n, x(n), u(n), s) + a_n - a_0 - \int_{t_0}^{t} \phi_0 \right|.$$

Therefore, using (6.1), (6.2), and hypothesis (b), one obtains

$$|x(n, t) - x_0(t)| \leqq \left| \int_{t_0}^{t} (\phi(n) - \phi_0) \right| + |a_n - a_0| + \left| \int_{t_n^0}^{t_0} \mu_1 \right|. \tag{6.3}$$

Now using the weak convergence of $\phi(n)$ to ϕ_0, the absolute continuity of the integral of an L_1 function, and the fact that $a_n \to a_0$, one obtains that there exists an $N^*(t) > N(t)$ such that for all $n > N^*(t)$,

$$|x(n, t) - x_0(t)| < \gamma. \tag{6.4}$$

But γ is arbitrary, so (i) is proved. The proof of (ii) is similar to the proof of (i). The only difference is that an extra term must be added on the right-

hand side of (6.3), namely, $\left| \int_{t_1}^{t_n^1} \mu_1 \right|$. The terms involving μ_1 are valid because (6.2) can be generalized to $|f(\epsilon, \hat{x}, u, t)| \leqq \mu_1(t)$ for all $(\epsilon, \hat{x}, u, t) \in [0, 1] \times \{|\hat{x}| \leqq R\} \times \bar{U} \times T$.

(iii) By the upper semicontinuity of the sets $G_1(\epsilon, t)$ at $(0, t_1)$ and the facts that $\epsilon_n \to 0$ and $t_n^1 \to t_1$, given $\delta_m \downarrow 0$ there exists $N(\delta_m)$ such that for $n > N(\delta_m)$,

$$b_n \in G_1(\epsilon_n, t_n^1) \subset U(G_1(t_1), \delta_m) = Q_m. \tag{6.5}$$

But Q_m is closed and $b_n \to b_0$, so $b_0 \in G_1(t_1) = \bigcap_{i=1}^{\infty} Q_i$. The proof of (iv) is trivial. Statement (i) and the facts that $\hat{x}(n, t) \in A$ for all $t \in \bar{I}_n$ and that A is closed immediately imply (iv).

Finally, it must be proved that there is a measurable function $u_0(t) \in U(t)$ a.e. such that $\dot{x}_0(t) = f_0(x_0(t), u_0(t), t)$ a.e. Since $\dot{x}_0(t) = \phi_0(t)$ a.e, it will first be proved that $\phi_0(t) \in f_0(x_0(t), U(t), t)$ a.e. Then hypothesis (d) and Lemma 6 finish the proof.

By Lemma 4 and the definition of the functions $\phi(n)$, for each $y \in E^{n+1}$,

$$\begin{aligned} \limsup_n (y, f(\epsilon_n, x(n), u(n), t)) &\geqq (y, \phi_0) \\ &\geqq \liminf_n (y, f(\epsilon_n, x(n), u(n), t)) \quad \text{a.e.} \end{aligned}$$

But this implies that for each $y \in E^{n+1}$,

$$\begin{aligned} \limsup_n \left(\max_{u \in U(t,\epsilon)} (y, f(\epsilon_n, x(n), u, t)) \right) &\geqq (y, \phi_0) \\ &\geqq \liminf_n \left(\min_{u \in U(t,\epsilon)} (y, f(\epsilon_n, x(n), u, t)) \right) \quad \text{a.e.} \end{aligned} \tag{6.6}$$

For each t, the functions $f(\epsilon_n, x, u, t)$ and the sets $U(t, \epsilon_n)$ satisfy the hypotheses of Lemma 5. Therefore, since $\epsilon_n \to 0$ and $x(n, t) \to x_0(t)$, (6.6), Lemma 5, and Corollary 5.1 imply that for each y,

$$\begin{aligned} \max_{u \in U(t)} (y, f_0(x_0(t), u, t) &\geqq (y, \phi_0(t)) \\ &\geqq \min_{u \in U(t)} (y, f_0(x_0(t), u, t)) \quad \text{a.e.} \end{aligned} \tag{6.7}$$

Let Y be the set of all vectors in E^{n+1} with rational coordinates. Clearly, Y is dense and countable. Let $T' = I_0 - Q_0$, where Q_0 is the union of the sets of measure zero, $Q(y)$, on which the inequality statement in (6.7) is false. Clearly, T' has full measure on T, and on T' the inequalities in (6.7) are valid for all $y \in Y$. By hypothesis (d) and the compactness of each $U(t)$, Lemma 3 is applicable to each of the sets $f_0(x, U(t), t)$. Therefore, $\phi_0(t) \in f_0(x_0(t), U(t), t)$ a.e. Finally, Lemma 6 implies that there exists a measurable $u_0(t) \in U(t)$ a.e. such that $\phi_0(t) = f_0(x_0(t), u_0(t), t)$ a.e.

Remark 1. As stated earlier, the methods used in proving Theorem 1 are generalizations of methods used by Roxin [8]. However, there is a subtle mistake in Roxin's proof. He assumed the existence of a set T' of full measure on which the inequality statement in (6.7) holds for all $y \in E^{n+1}$. However, the proof that he gave verified only that for each $y \in E^{n+1}$ such a set exists. Roxin's proof can be corrected by the insertion of the steps after (6.7) in the proof of Theorem 1.

COROLLARY 1.1. *In Theorem* 1 *replace hypotheses* (a) *and* (b) *by*:
(a′) $f(\epsilon)$ *converges uniformly to* f_0 *on compact subsets of* $A \times \bar{U} \times T$;
(b′) *the trajectories of the problems* $P(\epsilon)$ *are uniformly bounded over* ϵ. *Then the conclusions of Theorem* 1 *follow.*

Proof. The uniform boundedness of the trajectories implies that the discussion can be confined to a compact set. In this compact set, hypotheses (a′) and (b′), Lemma 1, and the continuity of f_0 imply that hypothesis (b) of Theorem 1 holds. Furthermore, hypothesis (a′) and the continuity of f_0 imply that hypothesis (a) holds.

COROLLARY 1.2. *In Theorem* 1 *omit hypothesis* (d) *and assume that the control sets do not vary with* t. *Then the conclusions of Theorem* 1 *hold.*

Proof. The proof is the same as the proof of Theorem 1 up to the construction of the control for x_0. Since the control set is not a function of t, the construction used by Roxin is applicable.

7. Theorem 2. In Theorem 1, the perturbed equations considered are not required to have the same form as the original equations. Now consider the case where the original equations are linear in the control u and where it is required that the perturbed equations also be linear in u.

THEOREM 2. *Let* $P(\epsilon)$, $0 \leqq \epsilon \leqq 1$, *be a family of problems satisfying the assumptions in* §3 *and such that*:
(a) $f(\epsilon) = g(\epsilon, \hat{x}, t) + H(\epsilon, \hat{x}, t)u$.
(b) *At each point* $(0, \hat{x}, t)$, $g(\epsilon, \hat{x}, t)$ *and* $H(\epsilon, \hat{x}, t)$ *are continuous in* $(\epsilon, \hat{x})$.
(c) *There exist functions* μ_i *and* k_i, $i = 1, 2$, *mapping* E^1 *into* E^1 *such that* $\mu_i \in L_1$, $k_i(s) = O(s)$ *as* $s \to \infty$, k_i *is bounded on bounded sets and for all* $(\epsilon, \hat{x}, t)$ *in* $[0, 1] \times A \times T$, $|g(\epsilon, \hat{x}, t)| \leqq \mu_1(t)k_1(|\hat{x}|)$ *and* $|H(\epsilon, \hat{x}, t)| \leqq \mu_2(t)k_2(|\hat{x}|)$.
(d) g_0 *and* H_0 *are continuous on* $A \times T$.
Assume each set $U(t, \epsilon)$ *is convex, and for each* ϵ *there exists a measurable function* $u(\epsilon)$ *defined on* T *such that* $u(t, \epsilon) \in U(t, \epsilon)$ *a.e. Then if* $\epsilon_n \to 0$ *and* $(x(n), u(n), \bar{I}_n, a_n, b_n)$ *is any sequence of admissible pairs for the corresponding problems* $P(\epsilon_n)$,
(1) *there exists a pair* $(x_0, u_0, \bar{I}_0, a, b)$ *such that* x_0 *is an admissible trajectory for* P *with control* u_0;
(2) *there exists a subsequence of* $(x(n), u(n))$ *that is an approximation of type* 2 *to* (x_0, u_0).

Before proceeding with the proof of Theorem 2, the lemmas and theorems needed in the proof will be listed.

THEOREM B. *A convex subset of $L_2(I)$ is closed in the norm topology on $L_2(I)$ if and only if it is closed in the weak topology on $L_2(I)$.*

Proof. See [1].

LEMMA 7. *Let $S(\epsilon)$, $0 \leqq \epsilon \leqq 1$, be the set of all measurable functions u defined on I such that $u(t) \in U(t, \epsilon)$ a.e., where the sets $U(t, \epsilon)$ are convex and satisfy the conditions in §3. Then for each ϵ, the set $S(\epsilon)$ is closed in the weak topology on $L_2(I)$; and if $\epsilon_n \downarrow 0$, $u(\epsilon_n) \in S(\epsilon_n)$ for all n, and $u(\epsilon_n)$ converges in the weak topology to a function u_0, then $u_0 \in S_0 = S(0)$.*

Proof. Lemma 1 and the convexity of each set $U(t, \epsilon)$ imply that each set $S(\epsilon)$ is a convex subset of $L_2(I)$. Fix ϵ, let $u_n \in S(\epsilon)$, $n = 1, 2, \cdots$, such that u_n converges to u^* in the strong topology. But strong convergence implies that there exists a subsequence, without loss of generality denoted by n, that converges pointwise to u^* a.e. [7]. Therefore, there exists a set T' of full measure such that for all $t \in T'$, $u_n(t) \in U(t, \epsilon)$ and $u_n(t)$ converges to $u^*(t)$. But $U(t, \epsilon)$ is closed. Hence, for all $t \in T'$, $u^*(t) \in U(t, \epsilon)$. Therefore, $S(\epsilon)$ is strongly closed and, hence, by Theorem B is weakly closed.

Let $\epsilon_n \downarrow 0$, $u(\epsilon_n) \in S(\epsilon_n)$, and $u(\epsilon_n)$ converge weakly to u_0. Since $U(t, \epsilon_n) \downarrow U(t)$ and $S_0 \subset \bigcap_{n=1}^{\infty} S(\epsilon_n)$, $S(\epsilon_n) \downarrow S_0$. This follows easily from the fact that a monotone decreasing sequence of sets converges to its common intersection. Let $u \in S(\epsilon_n)$ for all n. Then there exists a set of full measure T' on which $u(t) \in U(t, \epsilon_n)$ for all n. Hence, $u(t) \in U(t)$ a.e.; so, $u \in S_0$. Therefore, for each N, $u(\epsilon_n) \in S(\epsilon_N)$ for all $n > N$. But, $S(\epsilon_N)$ is weakly closed, so $u_0 \in S(\epsilon_N)$. Hence, $u_0 \in S_0$.

THEOREM C. *A subset K of $L_2(I)$ is weakly sequentially compact if and only if it is bounded in the L_2-norm.*

Proof. See [1].

8. Proof of Theorem 2. It is easy to verify that the hypotheses of Theorem 1 are satisfied by functions satisfying the hypotheses of Theorem 2. Therefore, by Theorem 1 there exists an admissible pair for P, $(x_0, u_0, \bar{I}_0, a, b)$, and a subsequence of $x(n)$ that is an approximation of type 1 to x_0. For this subsequence (denoted by n) consider the corresponding controls $u(n)$. Extend each of these controls to measurable functions $\bar{u}(n)$ defined on I_0 by setting

$$\bar{u}(n, t) = \begin{cases} u(n, t) & \text{if } t \in I_n, \\ u^*(\epsilon_n, t) & \text{if } t \in I_0 - I_n, \end{cases} \tag{8.1}$$

and considering $\bar{u}(n)$ on I_0. Then $\bar{u}(n) \in S(\epsilon_n)$ with $I = I_0$. By Lemma 1 and Theorem C, there exists a subsequence of $\bar{u}(n)$ (denoted by n) and a function $u_0^* \in L_2(I_0)$ such that $\bar{u}(n)$ converges weakly to u_0^*. By Lemma

7, $u_0^* \in S_0$. Hence, u_0^* is an admissible control for P. Here, it has been assumed, without loss of generality, that $\epsilon_n \downarrow 0$.

The control u_0^* generates x_0. Define for $t \in I_0$,

$$x_0^*(t) = \int_{t_0}^{t} (g_0(x_0(s), s) + H_0(x_0(s), s)u_0^*)\, ds + a. \tag{8.2}$$

Prove that for each $t \in I_0$, $x(n, t) \to x_0^*(t)$; then $x_0^*(t) = x_0(t)$ since the limit function is unique. Extend $g(\epsilon_n, x(n, t), t)$ and $H(\epsilon_n, x(n, t), t)$ to I_0 as $f(\epsilon_n)$ was extended in (6.1).

For each $t \in I_0$, by hypotheses (b) and (c) and the fact that $x(n, t) \to x_0(t)$,

$$\begin{aligned} &\bar{g}(\epsilon_n, x(n, t), t) \to g_0(x_0(t), t) \text{ and} \\ &\bar{H}(\epsilon_n, x(n, t), t) \to H_0(x_0(t), t), \end{aligned} \tag{8.3}$$

where the convergence is pointwise and dominated. The bars denote the extensions of functions $g(\epsilon_n, x(n, t), t)$ and $H(\epsilon_n, x(n, t), t)$ to I_0. Let $\gamma > 0$ be given; then by the Lebesgue dominated convergence theorem, there exists $N(t)$ such that for all $n > N(t)$,

$$\left| \int_{t_0}^{t} \{\bar{g}(\epsilon_n, x(n), s) - g_0(x_0, s)\} \right| < \frac{\gamma}{4}. \tag{8.4}$$

By hypothesis (c), there exist R, $M > 1$ such that $|\hat{x}(\epsilon)| < R$ and $k_i(|\hat{x}(\epsilon)|) < M$, $i = 1, 2$, for all trajectories of $P(\epsilon)$, $0 \leqq \epsilon \leqq 1$, and there exists $\mu_0 \in L_1$ such that $|f(\epsilon, \hat{x}, u, t)| \leqq \mu_0(t)$ for all $(\epsilon, \hat{x}, u, t) \in [0, 1] \times [|\hat{x}| < R] \times \bar{U} \times T$. Let $Q = \max(1, \sup_{u \in U} |u|)$. Since the μ_i, $i = 0, 2$, are absolutely continuous, there exists $\delta > 0$ such that for $i = 0, 2$, if meas $(E) < \delta$,

$$\int_E \mu_i < \frac{\gamma}{16MQ}.$$

By Egoroff's theorem [7], there exists a set $E_0 \subset I_0$ such that the measure of its complement E_0' is less than δ, and $\bar{H}(\epsilon_n, x(n, t), t)$ converges uniformly to $H_0(x_0(t), t)$ on E_0. Hence, there exists N_1 such that for all $n > N_1$ and all $t \in E_0$,

$$|\bar{H}(\epsilon_n, x(n, t), t) - H_0(x_0(t), t)| < \frac{\gamma}{8Q \text{ meas } (E_0)}.$$

So for $n > N_1$,

$$\begin{aligned} &\int_{t_0}^{t} |\bar{H}(\epsilon_n, x(n), s) - H_0(x_0, s)| \cdot |\bar{u}(n)| \\ &\qquad \leqq \int_{E_0} \frac{\gamma}{4Q \text{ meas } (E_0)} Q\, ds + \int_{E_0'} 2\mu_2\, QM < 3\gamma/8. \end{aligned} \tag{8.5}$$

Since $H_0(x_0(t), t)$ is continuous and $\bar{u}(n)$ converges weakly to u_0^*, there exists $N_2(t)$ such that for all $n > N_2(t)$,

$$\left| \int_{t_0}^{t} H_0(x_0, s)(\bar{u}(n) - u_0^*) \right| < \frac{\gamma}{4}. \tag{8.6}$$

Moreover, there exists $N_3(t)$ such that for all $n > N_3(t)$, $t \in I_n$; and there exists N_4 such that for all $n > N_4$, $| t_0 - t_n^0 | < \delta$ and $| a_n - a | < \gamma/16$. For $n > N_3(t)$,

$$\begin{aligned} &| x(n, t) - x_0^*(t) | \\ &= \left| a_n - a + \int_{t_n^0}^{t} \{g(\epsilon_n x(n), s) + H(\epsilon_n, x(n), s)u(n)\} \right. \\ &\qquad \left. - \int_{t_0}^{t} \{g_0(x_0, s) + H_0(x_0, s)u_0^*\} \right|. \end{aligned} \tag{8.7}$$

Therefore, from (8.4), (8.5), (8.6), (8.7) and the fact that if meas $(E) < \delta$, then $\int_E \mu_0 < \gamma/16MQ$; for $n > \max\{N(t), N_1, N_2(t), N_3(t), N_4\}$,

$$\begin{aligned} | x(n, t) - x_0^*(t) | &\leqq \left| \int_{t_0}^{t} (\bar{g}(\epsilon_n, x(n), s) - g_0(x_0, s)) \right| \\ &\quad + \left| \int_{t_0}^{t} (H_0(x_0, s)(\bar{u}(n) - u_0^*)) \right| \\ &\quad + \int_{t_0}^{t} | \{(\bar{H}(\epsilon_n, x(n), s) - H_0(x_0, s))\bar{u}(n)\} | \\ &\quad + \left| \int_{t_0}^{t_n^0} \mu_0 \right| + | a_n - a | < \gamma. \end{aligned} \tag{8.8}$$

But γ is arbitrary, so $x_0^*(t) = x_0(t)$ on $[t_0, t_1]$. Furthermore, from the continuity, $x_0^*(t_1) = x_0(t_1)$. Therefore, u_0^* generates x_0.

COROLLARY 2.1. *In Theorem 2 replace hypotheses* (b) *and* (c) *by:*
(b′) $g(\epsilon)$ *and* $H(\epsilon)$ *converge uniformly to* g_0 *and* H_0 *on compact subsets of* $A \times T$.
(c′) *The trajectories of the problems* $P(\epsilon)$ *are uniformly bounded over* ϵ. *Then the conclusions of Theorem 2 follow.*

Proof. Clearly, (b′) and hypothesis (d) of Theorem 2 imply hypothesis (b), and (b′) with (c′) and hypothesis (d) implies hypothesis (c).

9. Optimality and Theorems 1 and 2. The following theorem demonstrates a relationship between optimal solutions of $P(\epsilon)$ and P under certain conditions.

THEOREM 3. *Let* $P(\epsilon)$, $0 \leqq \epsilon \leqq 1$, *be a family of problems satisfying the hypotheses of Theorem* 1 (*Theorem* 2). *Let* C^* *be the optimal cost for* P. *Let*

$\epsilon_n \downarrow 0$ and let $(x(n), u(n))$ be a sequence of admissible pairs for the corresponding problems $P(\epsilon_n)$ such that $C(x(n)) \to C^*$. Then
(1) *there exists a subsequence of these pairs that is an approximation of type* 1 *(type* 2*) to an optimal pair for P;*
(2) *if optimal pairs exist for each $P(\epsilon_n)$, then any sequence of optimal pairs $(x_*(n), u_*(n))$ such that $x_*(n)$ is an optimal trajectory for $P(\epsilon_n)$ contains a subsequence that is an approximation of type* 1 *(type* 2*) to an optimal pair for P.*

Proof. (1) By Theorem 1 (Theorem 2), there exist an admissible pair for P, (x_0, u_0), and a subsequence of $(x(n), u(n))$ (denoted by n) that is an approximation of type 1 (type 2) to (x_0, u_0). Clearly, x_0 is optimal since $C(x(n)) = x^0(n, t_n^1) \to x_*^0(t_1) = C^*$.

(2) Given a sequence of optimal pairs $(x_*(n), u_*(n))$ for $P(\epsilon_n)$, then by Theorem 1 (Theorem 2) there exist an admissible pair for P, (x_*, u_*), and a subsequence that is an approximation of type 1 (type 2) to (x_*, u_*). But, x_* is optimal because

$$C(x_*) = x_*^0(t_1) = \lim_{n\to\infty} x(n, t_n^{1*}) = \lim_{n\to\infty} C(x_*(n)) \leqq \lim_{n\to\infty} C(x(n)) = C^*.$$

Remark 1. It is clear that Theorem 3 includes the existence theorems of Roxin [8], Filippov [3], and Lee and Markus [6]. It is also clear that in Theorem 3, the hypotheses of Theorem 1 (Theorem 2) could be replaced by the hypotheses of Corollary 1.1 or 1.2 (Corollary 2.1).

10. Penalty functions. The penalty function method is an attempt to express a problem in which the phase space constraint set is closed as the limit in some sense of a sequence of problems in which the phase space constraint set is open. In general, problems in which the phase space constraint set is open are easier to solve than those in which this set is closed.

Let $P = P(f_0, G_0, G_1(t), U(t), T, A)$ be a problem in which A is a closed set, and let p_k, $k = 1, 2, \cdots$, be a sequence of penalty functions of the first kind [9] defined on an open set $B \supset A$. Consider the sequence of problems $P(k, B)$ obtained from $P(A)$ by replacing A by B and the cost equation $\dot{x}^0 = f_0^0$ by $\dot{x}^0 = f_0^0 + p_k$. It is usually assumed that if $x(k)$ is any sequence of optimal trajectories for the problems $P(k, B)$ then these trajectories converge to an optimal trajectory for $P(A)$ as $k \to \infty$. This assumption is invalid in general. Russell [9] has proved that under certain conditions if f_0 satisfies the hypotheses of Lee and Markus [6], and it is easy to extend this result to the case where f_0 satisfies the hypotheses of Roxin [8] if the function $\mu \in L_\infty$, there exists a subsequence of any such sequence of optimal trajectories that converges to an optimal trajectory for $P(A)$. Now perturb each of the problems $P(k, B)$ and consider the

family of problems $P(\epsilon, k, B)$, $0 \leqq \epsilon \leqq 1$, $k = 1, 2, \cdots$, obtained. For such problems, the following theorem holds. In this theorem, the set B is any open set containing A on which the functions $f(\epsilon)$ satisfy the hypotheses of Theorem 1 or 2, and the p_k are penalty functions of the first kind.

THEOREM 4. *Let $P(\epsilon, A)$ be a family of problems satisfying the hypotheses of Theorem* 1 (*Theorem* 2). *Let C^* be the optimal cost for $P(A)$. Let $k_n \to \infty$ and $\epsilon_n \downarrow 0$ as $n \to \infty$. If $(x(n), u(n))$ is a sequence of admissible pairs for the corresponding problems $P(\epsilon_n, k_n, B)$ such that $C_{k_n}(x(n)) \to C^*$, then* (1) *there exists a subsequence of $(x(n), u(n))$ that is an approximation of type* 1 (*type* 2) *to an optimal pair for $P(A)$;*
(2) *if optimal pairs exist for $P(\epsilon_n, k_n, B)$, then any such sequence of optimal pairs contains a subsequence that is an approximation of type* 1 (*type* 2) *to an optimal pair for $P(A)$. ($C_{k_n}(x(n))$ denotes the cost of $x(n)$ considered as a trajectory for $P(\epsilon_n, k_n, B)$.)*

The following lemma is needed in the proof of Theorem 4.

LEMMA 8. *Let $P(\epsilon)$ be a family of problems that satisfies the hypotheses of Theorem* 1 (*Theorem* 2). *Then the convergence in* (6.4) *is locally uniform; that is, given $\bar{t} \in I_0$, there exists an interval $\bar{I} \subset I_0$ such that $\bar{t} \in \bar{I}$ and the convergence of $x(n)$ to x_0 is uniform on $\bar{I}$.*

Proof. This proof should be studied in conjunction with the proof of Theorem 1. For $t \in I_0$, set $h(n, t) = \int_{t_0}^{t} (\phi(n) - \phi_0)$. Let $\bar{t} \in I_0$ be fixed. Then there exists $\bar{I} \subset I_0$ such that $\bar{t} \in I$. Each $h(n)$ is continuous on $\bar{I}$, so there exists $t_n \in \bar{I}$ such that $M(n) = |h(n, t_n)| = \max_{t \in \bar{I}} |h(n, t)|$. By hypothesis (b) in Theorem 1, there exists Q such that $M(n) < Q$ for all n. Therefore, by the Bolzano-Weierstrass theorem there exist a subsequence of the $M(n)$ (denoted by n), an M_0, and $t^* \in \bar{I}$ such that $M(n) \to M_0$ and $t_n \to t^*$. If $M_0 > 0$, then there exists N such that for all $n > N$, $M(n) > M_0/2$; and since the $h(n)$ are equicontinuous at t^*, there exists $\gamma > 0$ such that $|t - t^*| < \gamma$ implies that for all n, $|h(n, t) - h(n, t^*)| < M_0/10$. But, there exists N_1 such that for all $n > N_1$, $|t_n - t^*| < \gamma$. Furthermore, from the weak convergence there exists N_2 such that for all $n > N_2$, $|h(n, t^*)| < M_0/10$. Therefore, for all $n > \max\{N, N_1, N_2\}$, one obtains the following contradiction:

$$M_0/2 < M(n) = |h(n, t_n)| \leqq |h(n, t^*)| + M_0/10 < M_0/5.$$

Hence, the sequence $M(n)$ has one limit point, namely, 0. So $M(n) \to 0$.

Therefore, given $\gamma > 0$, the absolute continuity of $\int \mu_1$ and the facts that $M(n) \to 0$, that $t_n^0 \to t_0$, and that eventually $\bar{I} \subset I_n$ imply that

there exists N_0 such that for all $n > N_0$ and all $t \in \bar{I}$,

$$| x(n, t) - x(t) | = \left| \int_{t_n^0}^{t} f(n) - \int_{t_0}^{t} \phi_0 + a_n - a \right|$$

$$\leqq \left| \int_{t_0}^{t} (\phi(n) - \phi_0) \right| + \left| \int_{t_n^0}^{t_0} \mu_1 \right| + | a_n - a | < \gamma.$$

11. Proof of Theorem 4. Since $P(\epsilon, k, B)$ differs from $P(\epsilon, B)$ only in the differential equation for the cost, and the cost function is not involved in $\hat{f}(\epsilon, x, u, t)$, there exists $R > 0$ such that all trajectories $\hat{x}(\epsilon)$ for $P(\epsilon, k, B)$ are contained in the sphere $\{| \hat{x} | < R\}$. Let $B_1 = \bar{B} \cap \{| \hat{x} | \leqq R\}$ and $A_1 = A \cap \{| \hat{x} | \leqq R\}$.

(1) For each n, the trajectory $\hat{x}(n)$ is admissible for $P(\epsilon_n, B_1)$. Since B_1 is a closed set, Theorem 1 (Theorem 2) is applicable and implies that there exists a subsequence (denoted by n) that is an approximation of type 1 (type 2) to an admissible pair $(\hat{x}_0, u_0)$ for $P(0, B_1)$. Two things must be proved: $\hat{x}_0(t) \in A$ for all $t \in \bar{I}_0$, and $C_0(\hat{x}_0) = C^*$.

Clearly, $\hat{x}_0(t_0) \in A$ and $\hat{x}_0(t_1) \in A$. Suppose there exists $\bar{t} \in I_0$ such that $\hat{x}_0(\bar{t}) \notin A$. Then by the continuity of $\hat{x}_0$ and the definition of B, there exist compact sets D and D_1 with nonempty interiors (in E^n) such that $D \subset \text{int } D_1 \subset B_1 - A_1$, and there exists a closed interval $\bar{I} \subset I_0$ such that $\bar{t} \in I$ and $\hat{x}_0(t) \in D$ for all $t \in \bar{I}$. By Lemma 8, $\hat{x}(n)$ converges uniformly to $\hat{x}_0$ on $\bar{I}$. Therefore, there exists N such that for all $n > N$ and all $t \in \bar{I}$, $\hat{x}(n, t) \in D_1$. By Lemmas 1 and 2 and the continuity of f_0^0, there exists Q such that $| f_0^0(\hat{x}(n, t), u(n, t), t) | \leqq Q$ for all n. Combining the above remarks and the definitions of penalty functions of the first kind, one obtains

$$C(n) = C_{k_n}(\hat{x}(n)) \geqq \int_I p_{k_n}(\hat{x}(n)) - Q, \tag{11.1}$$

$$\lim_{k \to \infty} (\min_{\hat{x} \in D_1} p_k(\hat{x})) = +\infty. \tag{11.2}$$

Hence, one obtains the contradiction

$$C^* = \lim_{n \to \infty} C(n) \geqq m(I)[\lim_{n \to \infty} \min_{\hat{x} \in D_1} p_{k_n}(\hat{x}(n))] = +\infty. \tag{11.3}$$

Therefore, $\hat{x}_0 \in A$ and is admissible for $P(0, A)$. So, $C_0(\hat{x}_0) \geqq C^*$. But,

$$C_0(\hat{x}_0) = \lim_{n \to \infty} C_0(\hat{x}(n)) \leqq \lim_{n \to \infty} C(n) = C^*.$$

(2) Let $(x_*(n), u_*(n))$ be a sequence of optimal pairs corresponding to the problems $P(\epsilon_n, k_n, B)$. Then as in (1), there exists a subsequence that is an approximation of type 1 (type 2) to an admissible pair $(x_0,$

u_0) for $P(0, B_1)$. But,

$$C_0(\hat{x}_0) = \lim_{n\to\infty} C_0(\hat{x}_*(n)) \leqq \limsup_n C_{k_n}(\hat{x}_*(n)) \leqq \lim_{n\to\infty} C(n) = C^*.$$

Therefore, $\limsup_n C_{k_n}(\hat{x}_*(n)) \leqq C^*$. But with this result, it is clear that the remainder of the proof of (2) is identical to the proof of (1).

12. Remarks. With controllability hypotheses on each of the $P(\epsilon)$, it can be proved that the optimal costs for the $P(\epsilon)$ converge to the optimal cost for P as $\epsilon \to 0$. In fact, if the function μ in hypothesis (b) of Theorem 1 is an L_∞ function, then the optimal costs for the $P(\epsilon, k, B)$ converge to the optimal cost for $P(A)$ as $\epsilon \to 0$ and $k \to \infty$ for any open set B containing A and any sequence of penalty functions of the first kind. These results are derived in [10].

REFERENCES

[1] N. Dunford and J. Schwartz, *Linear Operators*, vol. I, Interscience, New York, 1964.

[2] H. G. Eggleston, *Convexity*, Cambridge University Press, Cambridge, 1958.

[3] A. F. Filippov, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.

[4] H. Hermes, *The equivalence and approximation of optimal control problems*, Tech. Report 65-2, Center for Dynamical Systems, Brown University, Providence, 1965.

[5] F. M. Kirillova, *On the correctness of the formulation of an optimal control problem*, this Journal, 1 (1963), pp. 224–239.

[6] E. B. Lee and L. Markus, *Optimal control for non-linear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[7] M. E. Munroe, *Introduction to Measure and Integration*, Addison-Wesley, Reading, Massachusetts, 1953.

[8] E. Roxin, *The existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119.

[9] D. L. Russell, *Penalty functions and bounded phase coordinate control*, this Journal, 2 (1964), pp. 409–423.

[10] J. Cullum, *Continuous optimal control problems with phase space constraints*, Tech. Report prepared under Contract Nonr 222(88), Office of Naval Research, 1965.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

A MAXIMUM PRINCIPLE FOR OPTIMAL CONTROL PROBLEMS IN WHICH THE PHASE SPACE CONSTRAINT SET IS CLOSED*

JANE CULLUM†

1. Introduction. Let E^n denote n-dimensional Euclidean space, and T be a fixed interval in E^1. Consider the following problem P. Move from G_0 along an absolutely continuous curve $(\hat{x}, I)$ such that $\hat{x}(t) \in A$ on I and $\hat{x}(t_1) \in G_1$ and for which there exists a bounded and measurable function u on I such that $u(t) \in U$ a.e. and $\dot{\hat{x}}(t) = \hat{f}(\hat{x}(t), u(t), t)$ a.e. and such that the integral, the cost of $\hat{x}$,

$$C(\hat{x}) = \int_I f^0(\hat{x}(t), u(t), t)\, dt$$

is minimized. $I = [t_0, t_1]$ is an interval contained in T.

If the phase space constraint set A is open, if the control set U is closed, and if the function $f = (f^0, \hat{f})$ is continuous in $(\hat{x}, u, t)$ and continuously differentiable in $\hat{x}$ and in t, then an optimal pair, an optimal trajectory with its control, if such a pair exists for P, must satisfy Pontryagin's maximum principle [8]. However, if the phase space constraint set is closed, this principle is not applicable.

Several people (Gamkrelidze [8], Berkovitz [2], and Warga [12]) have obtained extensions of this principle to problems in which A is closed. Berkovitz obtained essentially the same result as Gamkrelidze. Gamkrelidze required the controls to be piecewise continuous, and the control set to be "regular" and considered a "regular" optimal trajectory lying on the boundary of A. This result of Gamkrelidze, Pontryagin's maximum principle, and the jump conditions derived by Gamkrelidze yield an overall principle that applies to any optimal trajectory that can be split into sections on the boundary of A and sections in the interior of A. Warga [10], [11], [12] derived a principle that applies to all of an optimal trajectory without splitting the trajectory into sections on the boundary of A and sections in the interior of A. The main problem with Warga's result is that it requires the sets $f(x, U, t)$ to be convex for each (x, t).

The object of this paper is to extend Gamkrelidze's result concerning

* Received by the editors February 4, 1966, and in revised form March 25, 1966.

† Department of Mathematics, University of California, Berkeley, California.
This research was supported by the United States Office of Naval Research under Contract Nonr 222(88). This paper is part of a dissertation submitted in partial satisfaction of the requirements for the Ph.D. degree in applied mathematics at the University of California, Berkeley.

sections of an optimal trajectory on the boundary of A. It will be proved that if in a neighborhood of such a trajectory the boundary of A is the C^2-diffeomorphic image of an open set in E^{n-1}, then the corresponding optimal pair satisfies a modified version of Pontryagin's maximum principle. The restrictions made by Gamkrelidze that the optimal trajectory considered be "regular" and that the controls be piecewise continuous are removed, and the condition that the control sets be "regular" is relaxed. It is proved than any "regular" problem considered by Gamkrelidze with optimal trajectory on the boundary of A, and for which the required C^2-diffeomorphic map exists, is included in the problems considered in this paper. An extension of the jump conditions obtained by Gamkrelidze for "regular" problems to the general problem considered in this paper is not obtained in this paper. Consequently, in general, the results obtained are not constructive.

2. Statement of problem. The problem to be considered was stated in the Introduction, and will be denoted by $P = P(f, G_0, G_1, U, T, A)$. Clearly, P can be reformulated as a problem in E^{n+1} with $x = (x^0, \hat{x})$ and $\dot{x}^0 = f^0$. The augmented problem will also be denoted by P. A pair for P, namely, an absolutely continuous curve and its control that satisfy the differential system and control requirements associated with P but not necessarily the boundary conditions, will be denoted by (x, u, I, a, b), where I is the interval of definition of x and u; and a and b are, respectively, the initial and terminal values of x.

The global assumptions on P are:

(2.1) The function $f = (f^0, \hat{f})$ is continuously differentiable in $\hat{x}$, u, and t.

(2.2) The phase space constraint set A is a closed subset of E^n with nonempty interior.

(2.3) The initial and terminal sets G_0 and G_1 are closed and contained in A.

(2.4) The control set U is a closed subset of E^r.

(2.5) Let ∂A denote the boundary of A. There exists an optimal solution (x_0, u_0, I_0) for P such that for all $t \in I_0$, $\hat{x}_0(t) \in \partial A$, and there exists a C^2 function g such that in a neighborhood of this trajectory $\partial A = \{\hat{x} \mid g(\hat{x}) = 0\}$ and $\operatorname{grad} g(\hat{x}) \neq 0$.

3. Theorem 1. This theorem states essentially that if the boundary of the constraint surface can be parameterized in a particular way in a neighborhood of the optimal trajectory given in (2.5), the corresponding optimal pair must satisfy a local maximum principle, local in the sense that the maximization is made over a subset of the control set.

THEOREM 1. *Let $P = P(f, \hat{x}_0, \hat{x}_1, U, T, A)$ be a problem for which the assumptions in §2 are satisfied and for which f is not a function of t. Let (x_0, u_0, I_0) be the optimal pair given in (2.5). Let $N^* \subset E^n$ be an open*

neighborhood of $(\hat{x}_0, I_0)$ *contained in the neighborhood given in* (2.5), $\hat{S} \subset E^{n-1}$ *an open set, and* $\hat{H}$ *a* C^2*-diffeomorphism mapping* $\hat{S}$ *onto* $\hat{N} = N^* \cap \partial A$. *Let* J *be the inverse of* $\hat{H}$. *Furthermore, let* $W \subset E^k$ *be a compact set and* B *be a map from* $\hat{S} \times W$ *into* U *such that*

(a) B *is continuous on* $\hat{S} \times W$, *and continuously differentiable in* $\hat{s}$ *on* $\hat{S} \times W$,

(b) *for each* $\hat{x} \in \hat{N}$, $\{u \in U \mid (\hat{f}(\hat{x}, u), \operatorname{grad} g(\hat{x})) = 0\} \supset B(J(\hat{x}), W)$,

(c) $u_0(t) \in B(J(\hat{x}_0(t)), W)$ *a.e. on* I_0.

Then there exists a nontrivial, absolutely continuous $(n+1)$*-dimensional vector function* ψ_0 *on* I_0 *such that*

$$\dot{\psi}_0 = -\left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial u}\frac{\partial B}{\partial s} K\right]^T \psi_0 \tag{1}$$

a. e., where

$$K = \begin{bmatrix} 1 & 0 \\ 0 & \dfrac{\partial J}{\partial \hat{x}} \end{bmatrix}$$

and the matrices are evaluated along the optimal pair (x_0, u_0),

$$\psi_0^0(t) = \text{const.} \leqq 0, \tag{2}$$

$$\begin{aligned}(\psi_0(t), f(\hat{x}_0(t), u_0(t))) \\ = \max_{w \in W} (\psi_0(t), f(\hat{x}_0(t), B(J(\hat{x}_0(t)), w))) = 0 \text{ a.e.}\end{aligned} \tag{3}$$

Before proceeding with the proof of Theorem 1, several lemmas will be proved.

LEMMA 1. *Let* $\partial A = \{x \mid g(x) = 0\}$ *where* g *is a* C^2 *function and* $\operatorname{grad} g(x) \neq 0$ *for* $x \in \partial A$. *An absolutely continuous curve* (x, I) *belongs to* ∂A *if and only if* $x(t_0) \in \partial A$ *and* $(\dot{x}, \operatorname{grad} g(x)) = 0$ *a.e. on* $I = [t_0, t_1]$.

Proof. Let $x(t) \in \partial A$ on I, then $h(t) = g(x(t)) \equiv 0$. Therefore, by the absolute continuity of x and the chain rule, $\dot{h} = (\dot{x}, \operatorname{grad} g(x)) = 0$ a.e. Conversely, let $x(t_0) \in \partial A$, and $(\dot{x}, \operatorname{grad} g(x)) = 0$ a.e. The function $h(t) = g(x(t))$ is absolutely continuous [7]. Therefore, by the chain rule, $\dot{h} = (\dot{x}, \operatorname{grad} g(x)) = 0$ a.e. Hence, $h(t) \equiv 0$, since an absolutely continuous function with zero derivative a.e. is a constant function.

LEMMA 2 (Filippov). *Let* f *be a continuous map from* $E^1 \times E^r$ *into* E^{n+1}. *Let the sets* $U(t)$ *be compact for each* t *and upper semicontinuous with respect to inclusion in* t (see [3]) *on* I. *Let* y *be a measurable function such that* $y(t) \in f(t, U(t))$ *a.e. on* I. *Then there exists a measurable function* u *such that* $u(t) \in U(t)$ *a.e. and* $y(t) = f(t, u(t))$ *a.e. on* I.

Proof. See [3].

LEMMA 3. *If* H *is a* C^1*-diffeomorphism mapping an open set* $S \subset E^n$ *onto*

a set $X \subset E^{n+1}$, *then*

$$(H^{-1})'(H(s))\ \frac{\partial H}{\partial s} \equiv I_n\,,$$

where I_n *is the* $n \times n$ *identity matrix and* $(H^{-1})'(H(s))$ *is the first derivative of* H^{-1} *evaluated at* $H(s)$.

Proof. For all $s \in S$, $s = H^{-1}(H(s))$. Therefore, by the generalized chain rule [9], $I_n = (H^{-1})'(H(s))\cdot H'(s)$. But, since $H \in C^1$, $H'(s) = \partial H/\partial s$.

4. Proof of Theorem 1. It is clear that $\hat{H}$ can be extended to a C^2-diffeomorphism H mapping the cylinder $S = (s^0\text{-axis}) \times \hat{S}$ onto the cylinder $N = (x^0\text{-axis}) \times \hat{N}$ by mapping s^0 onto x^0. Let $K(s) = (H^{-1})'(H(s))$, $\hat{s}_0 = J(\hat{x}_0)$, $\hat{s}_1 = J(\hat{x}_1)$, and $h(s, w) = K(s)f(\hat{H}(\hat{s}), B(\hat{s}, w))$. Consider the following problem $P(s)$. Move from $(0, \hat{s}_0)$ to the line through $(0, \hat{s}_1)$ parallel to the s^0-axis along an absolutely continuous curve (s, I) for which there exists a measurable function w on I such that $w(t) \in W$ a.e. and

$$(4.1)\quad \dot{s}(t) = h(s(t), w(t)) \quad \text{a.e. on } I \quad \text{and} \quad s(t) \in S \quad \text{on } I,$$

and such that the zeroth component of s at the terminal time is minimized. $I = [t_0\,, t_1]$ is an interval contained in T.

For each pair (s, w, I) for $P(s)$ satisfying (4.1), there exists a corresponding pair (x, u, I) for P; x satisfies the boundary conditions for P if and only if s satisfies the boundary conditions for $P(s)$. Let $(s_1\,, w_1\,, I_1)$ be any such pair for $P(s)$. Set $x_1(t) = H(s_1(t))$ and $u_1(t) = B(\hat{s}_1(t), w_1(t))$. Then $\hat{x}_1(t) \in \partial A$, and for a.e. $t \in I_1\,$, $\dot{x}_1(t)$ exists and

$$(4.2)\quad \dot{x}_1(t) = \frac{\partial H}{\partial s}\ (s_1(t))K(s_1(t))f(\hat{x}_1(t), u_1(t)) = f(\hat{x}_1(t), u_1(t)).$$

Statement (4.2) is valid because, by construction, for a.e. $t \in I_1$ there exists $c(t) \in E^n$ such that

$$f(\hat{x}_1(t), u_1(t)) = \frac{\partial H}{\partial s}\ (s_1(t))c(t)$$

and by Lemma 3,

$$K(s)\ \frac{\partial H}{\partial s}\ (s) = I_n$$

for all $s \in S$.

Set $s_0(t) = H^{-1}(x_0(t))$; then s_0 is an optimal trajectory for $P(s)$. It is clear that $s_0(t) \in S$ and satisfies the boundary conditions. Moreover, since $u_0(t) \in B(\hat{s}_0(t), W)$ a.e., by Lemma 2 there exists a measurable function w_0 such that $w_0(t) \in W$ a.e. and $u_0(t) = B(\hat{s}_0(t), w_0(t))$ a.e. Therefore,

there exists a set I' of full measure in I_0 such that for each $t \in I'$, $\dot{s}_0(t)$ exists and $\dot{s}_0(t) = h(s_0(t), w_0(t))$, so s_0 is admissible for $P(s)$. Suppose there exists an admissible trajectory (s_1, I_1) for $P(s)$ such that $C(s_1) < C(s_0)$. But, then there exists (x_1, I_1) admissible for P such that $C(x_1) = C(s_1) < C(s_0) = C(x_0)$. Hence, s_0 is optimal for $P(s)$.

Clearly, h is continuous on $S \times W$ and continuously differentiable in s on $S \times W$. Since S is open, the constructions used and the results obtained by Pontryagin are applicable to $P(s)$. Familiarity with these constructions is assumed. The variational equations for $P(s)$ are

$$\dot{\delta s} = \frac{\partial h}{\partial s}(s_0(t), w_0(t))\delta s.$$

Let $A(t, \tau_i)$ denote the fundamental solution of this system on the interval $[\tau_i, t_1]$. If τ is a regular point of w_0, a convex cone $C(s_0(\tau))$ can be generated at $s_0(\tau)$, consisting of the vectors Δs emanating from $s_0(\tau)$ such that

$$\begin{aligned}\Delta s &= h(s_0(\tau), w_0(\tau))\delta t \\ &\quad + \sum_{i=1}^{m} A(\tau, \tau_i)\{h(s_0(\tau_i), w^i) - h(s_0(\tau_i), w_0(\tau_i))\}\delta t_i ,\end{aligned} \tag{4.3}$$

where the w^i are arbitrary elements of W and τ_i is a regular point of w_0 and $t_0 < \tau_i < t_1$.

For each perturbed trajectory s^* for $P(s)$, there exists a corresponding perturbed trajectory $x^* = H(s^*)$ for P and $\Delta x^* = (\partial H/\partial s)\Delta s^*$. Hence, a corresponding cone $C(x_0(\tau))$ is generated at $x_0(\tau) = H(s_0(\tau))$ consisting of the vectors Δx emanating from $x_0(\tau)$,

$$\begin{aligned}\Delta x &= f(x_0(\tau), u_0(\tau))\delta t \\ &\quad + \sum_{i=1}^{m} \frac{\partial H}{\partial s}(s_0(\tau))A(\tau, \tau_i)\{h(s_0(\tau_i), w^i) - h(s_0(\tau_i), w_0(\tau_i))\}\delta t_i .\end{aligned} \tag{4.4}$$

The variational equations for P are

$$\dot{\delta x} = \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial u}\delta u.$$

But, each perturbed trajectory for P is the image under H of a perturbed pair (s^*, w_0) obtained from (s_0, w_0) by a change in the initial condition but not in the control. Therefore, the control for x^* is $u^*(t) = B(\hat{s}^*(t), w_0(t))$. Hence,

$$\delta u = \frac{\partial B}{\partial s}(\hat{s}_0, w_0)K(s_0)\delta x,$$

and the variational equations for these special trajectories are

$$\delta\dot{x} = \left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial u}\frac{\partial B}{\partial s}K\right]\delta x, \tag{4.5}$$

where the matrices involved are evaluated along the optimal pair.

Furthermore, since $\delta x = (\partial H/\partial s)\delta s$ and $\dot{x} = (\partial H/\partial s)\dot{s}$,

$$\delta x(t) = \frac{\partial H}{\partial s}(s_0(t))A(t, t_0)K(s_0(t_0))\delta x_0 = D(t, t_0)\delta x_0 .$$

The $(n + 1) \times (n + 1)$ matrix $D(t, t_0)$ is not a fundamental solution of system (4.5). This is clear since the rank of D can be at most n. However, it is a proper transition matrix for the transfer of vectors in the tangent plane to ∂A along the optimal trajectory, because any such vector has the form $(\partial H/\partial s)c$ for some $c \in E^n$, and the c can be thought of as a δs_0. Hence, (4.4) can be rewritten as

$$\begin{aligned}\Delta x &= f(x_0(\tau), u_0(\tau))\delta t \\ &\quad + \sum_{i=1}^{m} D(\tau, \tau_i)\{f(\hat{x}_0(\tau_i), u^i) - f(\hat{x}_0(\tau_i), u_0(\tau_i))\}\delta t_i ,\end{aligned} \tag{4.6}$$

where $u^i = B(\hat{s}_0(\tau_i), w^i)$.

Following Pontryagin [8, p. 106], the limiting cone C_s^* at $s_0(t_1)$ can be defined. Corresponding to C_s^* is the cone

$$C^* = \frac{\partial H}{\partial s}(s_0(t_1))C_s^*$$

with vertex at $x_0(t_1)$. Pontryagin proved that there exists an n-dimensional vector $\psi_s \neq 0$ such that ψ_s and C_s^* are separated. That is, there exists $a \in E^n$ such that for all $y \in C_s^*$, $(a, y) \leqq 0$ and $(a, \psi_s) \geqq 0$. Define ψ_0 to be the solution of the system

$$\dot{\psi} = -\left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial u}\frac{\partial B}{\partial s}K\right]^T\psi, \tag{4.7}$$

such that $\psi(t_1) = K^T(s_0(t_1))\psi_s$. The matrices in (4.7) are evaluated along (x_0, u_0). The projection y of the vector $\psi(t_1)$ on the tangent plane to $E^1 \times A$ at $x_0(t_1)$ is not zero. For, suppose $y = 0$; then

$$\psi(t_1) = \mu \text{ grad } g(x_0(t_1)),$$

and so

$$\psi_s = \left(\frac{\partial H}{\partial s}\right)^T(s_0(t_1)) \text{ grad } g(x_0(t_1)),$$

where grad $g(x)$ denotes the $(n + 1)$-dimensional vector $(0, \text{grad } g(\hat{x}))$.

Hence, for all $\alpha \in E^n$,

$$(4.8) \qquad (\psi_s, \alpha) = \left(\text{grad } g(x_0(t_1)), \frac{\partial H}{\partial s}(s_0(t_1))\alpha\right) = 0.$$

But, (4.8) implies $\psi_s = 0$; hence, $y \neq 0$.

Clearly, $\psi_0^0(t) = \psi_s^0(t) = \text{const.} \leqq 0$, $(\psi_0(t_1), \Delta x) \leqq 0$, and

$$(\psi_0(t_1), f(\hat{x}_0(t_1), u_0(t_1))) = (\psi_s(t_1), h(s_0(t_1), w_0(t_1))) = 0.$$

Consequently, the remainder of this proof is identical to the remainder of the proof given by Pontryagin except that, since the set W is compact, there is no need to introduce an intermediary function. That is, define

$$M(t) = \max_{w \in W} (\psi_0(t), f(\hat{x}_0(t), B(\hat{s}_0(t), w))),$$

prove that M is absolutely continuous and $\dot{M} = 0$ a.e. Finally, prove that $M(t) = (\psi_0(t), f(\hat{x}_0(t), u_0(t)))$ at regular points of w_0.

5. Transversality conditions. Consider a problem $P = P(f, G_0, G_1, U, T, A)$. For the case where A is an open set, Pontryagin has proved that not only does there exist a nonzero function ψ but that in fact this function can be chosen such that $\hat{\psi}(t_i)$ is orthogonal to G_i at $\hat{x}_0(t_i)$, $i = 0, 1$. The following theorem is an extension of this result to the case considered in §4.

THEOREM 2. *Let $P = P(f, G_0, G_1, U, T, A)$ be a problem that satisfies the hypotheses of Theorem* 1. *Let M_i, $i = 0, 1$, be C^1-maps, mapping sets $C_i \subset E^{m_i}$ onto the sets $\hat{N} \cap G_i$. Then the ψ function whose existence was established in Theorem* 1 *can be chosen such that $\hat{\psi}(t_i)$ is orthogonal to $G_i \cap \hat{N}$ at $\hat{x}_i = \hat{x}_0(t_i)$, $i = 0, 1$.*

Proof. The proof consists in checking the constructions of Pontryagin [8, pp. 108–114] to ascertain that applicability in the s-space implies applicability in the x-space.

Consider the problem $P(s)$ with $\hat{s}_0$ and $\hat{s}_1$ replaced by the smooth manifolds $\hat{S}_i = J(M_i(C_i))$, $i = 0, 1$. Pontryagin's results are applicable to $P(s)$. Therefore, there exists a vector $\psi_s \neq 0$ such that the corresponding $\psi_s(t)$ is orthogonal to $\hat{S}_i$ at $\hat{s}_i = \hat{s}_0(t_i)$, $i = 0, 1$. By the proof of Theorem 1, $\psi_0(t_1)$ may be chosen to be $K^T(s_0(t_1))\psi_s$. Set $\hat{\psi}_0(t_1) = J'(\hat{x}_1)^T\psi_s(t_1)$.

The tangent plane $T_i^{\hat{s}}$ to $\hat{S}_i$ at $\hat{s}_i$ is equal to

$$\{\hat{s} \mid \hat{s} = J'(x_i)M_i'(\alpha_i)c + \hat{s}_i, c \in E^{m_i}\},$$

where α_i is chosen such that $M_i(\alpha_i) = \hat{x}_i$, $i = 0, 1$; and the tangent plane T_i to $\hat{N} \cap G_i$ at $\hat{x}_i$ is equal to

$$\{\hat{x} \mid \hat{x} = M_i'(\alpha_i)c + \hat{x}_i, c \in E^{m_i}\}.$$

Therefore, for every vector $y \in T_1$,

$$(5.1)\quad (\hat{\psi}_0(t_1), y) = (\hat{\psi}_0(t_1), M_1'(\alpha_1)c) = (\hat{\psi}_s(t_1), J'(x_1)M_1'(\alpha_1)c) = 0.$$

Hence, $\hat{\psi}_0$ is orthogonal to $G_1 \cap \hat{N}$ at $\hat{x}_1$.

Furthermore, define $T_0^s = \{s \mid s = (0, \hat{s}),\ \hat{s} \in T_0^{\hat{s}}\}$ and $T_0 = \{x \mid x = (0, \hat{x}),\ \hat{x} \in T_0\}$ and consider the convex cones,

$$(5.2)\quad \begin{aligned} C_s^{**} &= \text{convex hull of } (A(t_1, t_0)T_0^s \cup C_s^*), \\ C^{**} &= \text{convex hull of } (D(t_0, t_1)T_0 \cup C^*). \end{aligned}$$

Clearly,

$$C^{**} = \frac{\partial H}{\partial s}(s_0(t_1))C_s^{**}.$$

By Pontryagin's construction, the ψ_s is chosen such that for all $s \in C_s^{**}$, $(s, \psi_s) \leqq 0$. In particular, the plane $A(t_1, t_0)T_0^s$ must be contained in a plane parallel to the separating plane. Therefore, for all $z \in A(t_1, t_0)T_0^s$, $(z, \psi_s) = 0$. Therefore, for all $y \in T_0^s$,

$$\left(D(t_1, t_0)\frac{\partial H}{\partial s}(s_0(t_0))y, \quad \psi_0(t_1)\right) = 0.$$

But, this inner product is constant since the functions involved are solutions of conjugate equations. Therefore, at t_0,

$$\left(\frac{\partial H}{\partial s}(s_0(t_0))y, \quad \psi_0(t_0)\right) = 0.$$

But, for each $y \in T_0^s$ there exists $c \in E^{m_0}$ such that

$$y = (0, J'(\hat{x}_0)M_0'(\alpha_0)c).$$

Therefore, for all $c \in E^{m_0}$,

$$(5.3)\quad \begin{aligned} 0 &= \left(\frac{\partial H}{\partial s}(s_0)y, \quad \psi_0(t_0)\right) \\ &= \left(\frac{\partial \hat{H}}{\partial \hat{s}}(\hat{s}_0)J'(\hat{x}_0)M_0'(\alpha_0)c, \quad \hat{\psi}_0(t_0)\right) = (M_0'(\alpha_0)c, \hat{\psi}_0(t_0)). \end{aligned}$$

Hence, $\hat{\psi}_0(t_0)$ is orthogonal to $G_0 \cap \hat{N}$ at $\hat{x}_0$.

COROLLARY 2.1. *Consider a fixed time problem* $P = P(f, \hat{x}_0, A, U, T, A)$ *that satisfies the hypotheses of Theorem* 1 *with the exception that* f *may not be autonomous. Then the conclusions of Theorem* 2 *hold with the exception that* M *need not be the zero function; and in fact,* $\psi_0(t_1)$ *may be chosen to be the vector* $(-1, 0, \cdots, 0)$.

Proof. Pontryagin [8, p. 66] demonstrated that P may be transformed into a problem P^* of the type considered in Theorem 2 by enlarging the

differential system to $\dot{x} = f$ and $\dot{x}^{n+1} = 1$ with $x^{n+1}(t_0) = t_0$ and replacing G_1 and A by $G_1 \times \{t_1\}$ and $A \times \{t_1\}$, respectively. By Theorem 2, there exists an $(n+2)$-dimensional vector $\psi^* = (\psi^0, \hat{\psi}, \psi^{n+1}) \neq 0$ such that
(1) $(\psi^*, \Delta x^*) \leqq 0$ for all $\Delta x^* \in C^{**}$,
(2) $(\hat{\psi}, \psi^{n+1})$ is orthogonal to $A \times \{t_1\}$ at the point $(\hat{x}_0(t_1), t_1)$, and
(3) the projection of ψ^* on the tangent plane to $E^1 \times A \times E^1$ at $(x_0(t_1), t_1)$ is not zero. Condition (2) implies that $(\hat{\psi}, \psi^{n+1}) = (\mu \operatorname{grad} g(\hat{x}_0(t_1)), a)$. Condition (3) implies that $(\psi^0, 0, a) \neq 0$. Consequently, ψ^0 and a cannot both vanish. If $\psi^0 = 0$, then $a \neq 0$. But, then

$$(\psi^*, f^*(\hat{x}_0(t_1), u_0(t_1))) = (\psi^0, f^0) + a = 0$$

implies that $a = 0$; this is a contradiction. Therefore, in any case, ψ^0 does not vanish and can be taken to be -1. Furthermore, by Theorem 1,

$$\begin{aligned} 0 = M^* &= \max_{w \in W} (\psi^*(t), f^*(\hat{x}_0(t), B(\hat{s}_0(t), w))) \\ &= (\psi^*(t), f^*(\hat{x}_0(t), u_0(t))) \text{ a.e.} \end{aligned} \tag{5.4}$$

Clearly, (5.4) implies that for $\psi = (\psi^0, \hat{\psi})$,

$$\begin{aligned} M &= \max_{w \in W} (\psi(t), f(\hat{x}_0(t), B(\hat{s}_0(t), w))) \\ &= (\psi(t), f(\hat{x}_0(t), u_0(t))) \text{ a.e.} \end{aligned} \tag{5.5}$$

Hence, $\psi_0(t_1)$ can be chosen to be any vector of the form $(-1, \mu \operatorname{grad} g(\hat{x}_0(t_1)))$, where μ is arbitrary. In particular, the vector $(-1, 0, \cdots, 0)$ can be chosen.

6. Extensions of Theorems 1 and 2. In this section it will be proved that the requirements of continuity and continuous differentiability of B can be relaxed slightly.

THEOREM 3. *In Theorems* 1 *and* 2 *replace the assumptions on* B *by the following assumptions. Let* $s_0(t) = H^{-1}(x_0(t))$. *Let* $\hat{S}_i$ *and* $\hat{A}_i$, $i = 1, \cdots, M$, *be open sets in* E^{n-1} *such that*
(a) $\hat{S}_i \cap \hat{S}_j = \varnothing$ *if* $j \neq i+1, i-1$,
(b) $\hat{S}_i \subset \hat{A}_i \subset \hat{S}$,
(c) *there exist* t_i^1, t_i^2 *such that* $t_i^1 < t_{i+1}^1 < t_i^2 < t_{i+1}^2$ *and*

$$\{\hat{s} \mid \hat{s} = \hat{s}_0(t), t \in [t_i^1, t_i^2]\} \subset \hat{S}_i .$$

Let $W \subset E^k$ *be a compact set and let* B_i, $i = 1, \cdots, M$, *be maps from* $\hat{A}_i \times W$ *into* U *such that*
(d) B_i *is continuous on* $\hat{A}_i \times W$ *and continuously differentiable in* $\hat{s}$ *on* $\hat{A}_i \times W$,
(e) *for each* $\hat{x} \in \hat{H}(\hat{S}_i)$, $\{u \in U \mid (\hat{f}(\hat{x}, u), \operatorname{grad} g(\hat{x})) = 0\} \supset B_i(J(\hat{x}), W)$,
(f) $u_0(t) \in B_i(J(\hat{x}_0(t)), W)$ *a.e. on* $[t_i^1, t_i^2]$.
Then the conclusions of Theorems 1 *and* 2 *follow.*

Proof. Let

$$\hat{S}_0 = \bigcup_{i=1}^{M} \hat{S}_i .$$

The proof proceeds initially as the proof of Theorem 1 with $\hat{S}$ and $\hat{N}$ replaced by $\hat{S}_0$ and $\hat{H}(\hat{S}_0)$. That is, define $S_0 = E^1 \times \hat{S}_0$, $N = E^1 \times \hat{H}(\hat{S}_0)$, and extend $\hat{H}$ to H mapping S_0 onto N. Let

$$h_i(s, w) = K(s)f(\hat{H}(\hat{s}), B_i(\hat{s}, w)),$$

and define $P(s)$ as follows. Move from $(0, \hat{s}_0)$ to the line through $(0, \hat{s}_1)$ parallel to the s^0-axis along an absolutely continuous curve (s, I) such that $s(t) \in S_0$ for all $t \in I$ and for which there exists a measurable function w such that $w(t) \in W$ a.e. and

(1) $\dot{s}(t) = h_i(s(t), w(t))$ for a.e. $t \in I$ for which there exists an i such that

$$s(t) \in (S_i - \bigcup_{\substack{j=1 \\ j \neq i}}^{M} S_j);$$

(2) for each i, there exists a unique interval $I_i = [a_i, a_{i+1}]$ such that for all $t \in I_i$, $s(t) \in S_i \cap S_{i+1}$, and for all $t \notin I_i$, $s(t) \notin S_i \cap S_{i+1}$; and on this interval, $\dot{s}(t) = h_j(s(t), w(t))$ a.e., where j is fixed and $j = i$ or $i + 1$ for all such t; and such that the value of the zeroth component of s at the terminal time is minimized.

Let $(s, w, [\bar{t}_3, \bar{t}_4])$ be a trajectory for $P(s)$, except that s may not satisfy the boundary conditions of $P(s)$. Then there exist intervals $I_k = [t_{k-1}, t_k]$, $k = 1, \cdots, j$, with $t_0 = \bar{t}_3$ and $t_j = \bar{t}_4$ such that a.e. on I_k, $\dot{s}(t) = h_k(s(t), w(t))$. Define $x(t) = H(s(t))$, then x is absolutely continuous; and if $\dot{s}(t)$ exists and equals $h_k(s(t), w(t))$ with $u(t) = B_k(\hat{s}(t), w(t))$,

$$\dot{x}(t) = \frac{\partial H}{\partial s}(s(t))K(s(t))f(\hat{x}(t), u(t)) = f(\hat{x}(t), u(t)).$$

The reasoning here is the same as in the proof of (4.2). Hence, x is a trajectory for P. Consequently, $s_0(t) = H^{-1}(x_0(t))$ is optimal for $P(s)$.

Choose $\bar{t}_i$ such that $s_0(\bar{t}_i) \in S_i \cap S_{i+1}$, $i = 1, \cdots, M - 1$. Lemma 2, with $y(t) = u_0(t)$, $U(t) = W$, and $f(t, U(t)) = B_i(\hat{s}_0(t), W)$ for $t \in J_i = [\bar{t}_{i-1}, \bar{t}_i]$, $i = 1, \cdots, M$, where $\bar{t}_0 = t_0$ and $\bar{t}_M = t_1$, implies that there exists a measurable function $w_0 \in W$ a.e. such that a.e. in $J_i u_0(t) = B_i(\hat{s}_0(t), w_0(t))$; and hence, a.e. in J_i, $\dot{s}_0(t) = h_i(s_0(t), w_0(t))$. Construct perturbed trajectories, not necessarily admissible, for $P(s)$ as follows. Let Q be the set of absolutely continuous curves $(s, [t_0, \bar{t}])$ for which there exists a measurable function $w \in W$ a.e. such that for a.e. $t \in J_i$, $i = 1, \cdots, M - 1$, and $J_M = [t_{M-1}, \bar{t}]$, $\dot{s}(t) = h_i(s(t), w(t))$; and for all $t \in J_i$, $s(t) \in S_i$. By construction, $(s_0, [t_0, t_1])$ is an element of Q.

Consider the curves $(s^*, I) \in Q$ obtained from (s_0, I_0) by changing the initial condition of s_0 to $(s_0 + \epsilon\delta s_0)$ and not changing w_0. By writing $s^* = s_0 + \epsilon\delta s + o(\epsilon)$, on J_i,

$$\delta\dot{s} = \frac{\partial h_i}{\partial s}(s_0(t), w_0(t))\delta s. \tag{6.1}$$

The right-hand side of (6.1) is well-defined and measurable on I_0. Therefore, there exists an absolutely continuous matrix function $A(t)$ that is a fundamental solution of (6.1) on I_0. Also, consider the curves $(s_*, I) \in Q$ obtained from (s_0, I_0) by altering the optimal control w_0 on a set of small measure. The perturbations in the control considered are identical to those made by Pontryagin [8, p. 87] with the additional requirement that if $\tau \in J_i$ is a regular point of w_0 at which such a perturbation is made, then all the intervals involved with τ must be in J_i. It is not difficult to verify that the formula given by Pontryagin for Δx is valid for Δs in this case. That is,

$$\begin{aligned}\Delta s = h(s_0(\tau), w_0(\tau))\delta t + \sum_{i=1}^{m} A(\tau, \tau_i)\{h(s(\tau_i), w^i) \\ - h(s(\tau_i), w(\tau_i))\}\delta t_i .\end{aligned} \tag{6.2}$$

Therefore, cones can be constructed, as in Theorem 1, at each point $x_0(\tau)$ with τ a regular point of w_0, and a limiting cone at $x_0(t_1)$. Finally, it must be verified that the lemmas Pontryagin used in the proof of the maximum principle apply to the preceding cones. This verification consists in studying the individual proofs of these lemmas and observing that each step is still valid. Lemma 3 [8, p. 94] needs to be considered only for an optimal trajectory and control. Hence, the technique used in the proof of Theorem 1 can be extended to this case.

Similarly, the lemmas and constructions used in the derivation of the transversality conditions are valid for the enlarged cones; and hence, Theorem 2 also extends to this case.

7. Regularity. The purpose of this section is to establish a relationship between Theorem 1 and the results obtained by Gamkrelidze [8, p. 267]. In order to do this, it is necessary to recall the following definitions made by Gamkrelidze. Consider a problem $P = P(f, \hat{x}_0, \hat{x}_1, U, T, A)$ satisfying the assumptions in §2, with f autonomous.

DEFINITION 7.1. The set U is *arranged in a regular manner* if $u_1 \in \partial U$ implies that there exist C^1-scalar-valued functions q_i, $i = 1, \cdots, s (s \geqq 1)$, such that

(a) there exists a neighborhood of u_1 in which the set U is defined by the inequalities $q_i(u) \leqq 0$, $i = 1, \cdots, s$;

(b) $q_i(u_1) = 0$, $i = 1, \cdots, s$; and

(c) the vectors $(\partial q_i/\partial u)(u_1)$, $i = 1, \cdots, s$, are linearly independent.

DEFINITION 7.2. A point $\hat{x} \in E^n$ is *regular* with respect to a point $u_1 \in U$ if

(a) $p(\hat{x}, u_1) = (\hat{f}(\hat{x}, u_1), \operatorname{grad} g(\hat{x})) = 0$,

(b) $\dfrac{\partial p}{\partial u}(\hat{x}, u_1) \not\equiv 0$, and

(c) if $u_1 \in \partial U$, the vectors

$$\frac{\partial p}{\partial u}(\hat{x}, u_1), \qquad \frac{\partial q_1}{\partial u}(u_1), \cdots, \qquad \frac{\partial q_s}{\partial u}(u_1),$$

where the q_i are the functions defined in Definition 7.1, are linearly independent.

DEFINITION 7.3. A pair $(x, u, [t_0, t_1])$ for P, such that $\hat{x}(t) \in \partial A$ for all t and u is piecewise continuous, is said to be *regular* if, at each point of continuity of u, $\hat{x}(t)$ is regular with respect to $u(t)$; and at each point of discontinuity of u, $\hat{x}(t)$ is regular with respect to $u(t - 0)$ and $u(t + 0)$, the left-hand and right-hand limits of u at t.

Gamkrelidze assumed that the control set for P was regular in the sense of Definition 7.1. He restricted the admissible controls to be piecewise continuous and assumed that the optimal pair given in (2.5) was regular in the sense of Definition 7.3. Under these assumptions, he proved that there exists an $(n + 1)$-dimensional absolutely continuous vector functon ψ_0 on I_0 such that:

(a) $$\dot{\psi}_0 = -\left[\frac{\partial f}{\partial x} + \Lambda \frac{\partial p}{\partial x}\right]^T \psi_0 \text{ a.e.},$$

where Λ is an $(n + 1)$-dimensional vector function obtained in the proof;

(b) $$\psi_0^0(t) = \text{const.} \leqq 0;$$

(c) $$(\psi_0(t), f(\hat{x}_0(t), u_0(t))) = \sup_{u \in \omega(x_0(t))} (\psi_0(t), f(\hat{x}_0(t), u))) \equiv 0,$$

where $\omega(x_0(t)) = \{u \in U \mid \hat{x}_0(t)$ is regular with respect to $u\}$.

THEOREM 4. *Let $P = P(f, \hat{x}_0, \hat{x}_1, U, T, A)$ be a problem that satisfies the assumptions in §2 and for which f is not a function of t. Furthermore, assume U is regular and the optimal solution given in (2.5) is regular. Then, if there exist sets $\hat{S}$ and $\hat{N}$ and a function $\hat{H}$ satisfying the conditions in Theorem 1, the conclusions of Theorem 1 follow.*

Proof. This theorem will be proved by constructing sets $\hat{S}_i$, $\hat{A}_i$ and W, and maps B_i satisfying the hypotheses of Theorem 3. Consider the optimal pair (x_0, u_0, I_0) given in (2.5). At each point of continuity $\bar{t}$ of u_0 con-

sider the system of equations:

$$
\begin{aligned}
p(\hat{x}, v) &= 0, \\
q_i(v) - q_i(u(t)) &= 0, \qquad i = 1, \cdots, s,
\end{aligned} \tag{7.1}
$$

where the q_i are the functions corresponding to $u_0(\bar{t})$. If $u_0(\bar{t}) \in \text{int } U$, then $s = 0$ and (7.1) reduces to a single equation. As before, $p(\hat{x}, v) = (f(\hat{x}, v), \text{grad } g(\hat{x}))$. At each point of discontinuity $\bar{t}$ of u_0 consider the two separate systems obtained by using the q_i corresponding to $u(\bar{t} - 0)$ in (7.1) and the q_i corresponding to $u(\bar{t} + 0)$ in (7.1). By the regularity assumptions which include differentiability assumptions, the implicit function theorem is applicable and implies that there exist a neighborhood $V \subset E^{n+r+2}$ of $(x_0(\bar{t}), \bar{t}, u_0(\bar{t}))$, an open set $R \subset E^{n+2+r-(s+1)}$ containing $(x_0(\bar{t}), \bar{t}, \hat{u}_0(\bar{t}))$, functions ϕ^i, $i = 1, \cdots, s + 1$, that are C^1 on R, such that

$$
\begin{aligned}
&\{(x, t, u) \in V \mid \text{equations (7.1) are satisfied}\} \\
&\quad = \{(x, t, u) \in V \mid (x, t, \hat{u}) \in R \text{ and } u^i = \phi^i(\hat{x}, t, \hat{u}), \\
&\qquad\qquad i = 1, \cdots, s + 1\}.
\end{aligned} \tag{7.2}
$$

Similar results are obtained at points of discontinuity of u_0. The vector $\hat{u}$ may be composed of any $r - (s + 1)$ components of u, depending upon which submatrices of the matrix

$$
\left[\frac{\partial p}{\partial u}, \frac{\partial q_1}{\partial u}, \cdots, \frac{\partial q_s}{\partial u}\right] \tag{7.3}
$$

have rank $s + 1$. However, to simplify the notation, $\hat{u}$ will always be denoted by the last $r - (s + 1)$ components of u. Clearly, the value of s may change with u. Several things are clear. Since, x^0 does not appear in (7.1), the neighborhoods V and R can be assumed to be cylindrical in the x^0-direction. Furthermore, since for each $u \in U$ there exists a neighborhood in E^r in which the set U is determined by the inequalities $q_i(u) \leqq 0$, where the q_i are the functions corresponding to u, the neighborhood V can be chosen such that if $(x, t, u) \in V$ and satisfies system (7.1) then $u \in U$. Finally, it is clear, since $(x_0(t), t, u_0(t))$ is a solution of (7.1) for each t, that if $(x_0(t), t, u_0(t)) \in V$ then there exists $(x_0(t), t, \hat{u}) \in R$ such that $u_0(t) = (\phi(\hat{x}_0(t), t, \hat{u}), \hat{u})$.

Therefore, determine an R and a V satisfying the above conditions for each point in E, where

$$
\begin{aligned}
E = &\{(x_0(t), t, u_0(t)) \mid t \text{ is a point of continuity of } u_0\} \\
&\cup \{(x_0(t), t, u_0(t - 0)), (x_0(t), t, u_0(t + 0)) \mid \\
&\qquad t \text{ is a point of discontinuity of } u_0\}.
\end{aligned} \tag{7.4}
$$

Since R is open, there exists a closed cube C contained in R with $(x_0(t), t, u_0(t))$ $[(x_0(t), t, u_0(t-0))$ or $(x_0(t), t, u_0(t+0))$ at points of discontinuity of $u_0]$ as its center point. That is, there exists $\delta_R > 0$ such that

$$C = \{y = (x, t, \hat{u}) \mid \max_{2 \leqq i \leqq n+2+r-(s+1)} |y^i - y_0{}^i(t)| \leqq \delta_R\} \subset R. \tag{7.5}$$

The image of C under ϕ, that is, the set

$$\begin{aligned} G = \{(x, t, u) \mid (x, t, \hat{u}) \in C \text{ and} \\ u^i = \phi^i(\hat{x}, t, \hat{u}), \qquad i = 1, \cdots, s+1\}, \end{aligned} \tag{7.6}$$

is a neighborhood in the relative topology of the surface in E^{n+r+2} composed of the points that are solutions to (7.1). Therefore, there exists a cylindrical neighborhood $Y \subset V$ in E^{n+r+2} of $(x_0(t), t, u_0(t))$ such that $G = \{(x, t, u) \in Y \mid$ equations (7.1 are satisfied$\}$. Clearly, there exists an open set $O \subset Y$ containing $(x_0(t), t, u_0(t))$.

Therefore, for each point in E consider the sets C, Y, and O. The set E is compact since it is the union of finitely many compact sets. Therefore, by the Heine-Borel theorem since the sets O form an open covering of E, there exists a finite subcovering $O_1, \cdots, O_M$. Let $C_1, \cdots, C_M$ correspond to $O_1, \cdots, O_M$. Consider the projection Q_i of O_i on the set $B = E^1 \times \partial A \times E^1$. Each Q_i is open in the relative topology of B, and for each $(x, t) \in Q_i$,

$$\begin{aligned} U_i(x, t) = \{u \mid u = (\phi_i(\hat{x}, t, \hat{u}), \hat{u}) \text{ such that} \\ \max_{s_i+2 \leqq k \leqq r} |u^k - u_0{}^k(t_i)| \leqq \delta_{R_i}\} \subset U. \end{aligned} \tag{7.7}$$

By construction each such u and x satisfy the first equation in (7.1).

Let k be the maximum dimension of $\hat{u}$ for all $i = 1, \cdots, M$; and let W be the closed unit cube in E^k. Set

$$\hat{u}_i = \begin{bmatrix} u^{s_i+2} \\ \vdots \\ u^r \end{bmatrix} = \delta_{R_i} \begin{bmatrix} 0 & \cdots & 0\,1\,0\,0 & \cdots & 0\,0 \\ 0 & \cdots & 0\,0\,1\,0 & \cdots & 0\,0 \\ & & \cdots & & \\ 0 & \cdots & 0\,0\,0\,0 & \cdots & 0\,1 \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix} + \hat{u}_i(t_i). \tag{7.8}$$

That is, $\hat{u}_i = P_i w + \hat{u}_i(t_i)$, where P_i is an $(r - (s_i + 1)) \times k$ dimensional matrix that picks out the last $r - (s_i + 1)$ components of w and multiplies these components by a scalar factor. It should be noted that one or more of the t_i could be a point of discontinuity of u_0. In this case, one would be considering $u_i(t_i - 0)$ or $u_i(t_i + 0)$; but, to simplify the notation it is assumed that all points considered are points of continuity of u_0. Hence, for $(x, t) \in Q_i$,

$$\begin{aligned} U_i(x, t) = \{u \mid u = (\phi_i(\hat{x}, t, P_i w + \hat{u}_i(t_i)), P_i w + \hat{u}_i(t_i)) \\ = D_i(\hat{x}, t, w), w \in W\}. \end{aligned} \tag{7.9}$$

Define $x^* = (x, x^{n+1})$, $s^* = (s, s^n)$, $H^*(s^*) = (H(s), s^n)$, where $s^n = x^{n+1} = t$, and set $S_i^* = H^{*-1}(Q_i) \cap (S \times E^1)$. Clearly, S_i^* is a cylinder with elements parallel to the s^0-axis. Set $S_0^* = \bigcup_{i=1}^M S_i^*$ and $s_0^*(t) = (s_0(t), t)$ on I_0. Also set $B_i^*(s^*, w) = D_i(H(s), t, w)$ on S_i^*, $i = 1, \cdots, M$. Clearly, the S_i^* and the B_i^* satisfy the hypotheses of Theorem 3, and it is clear that the B_i^* are actually defined on larger sets $A_i^* \subset S_i^*$. Therefore, define $P(s^*)$. Transfer the point $s_0^* = (0, \hat{s}_0, t_0)$ to the 2-dimensional plane passing through the point $(0, \hat{s}_1, 0)$ with elements parallel to the s^0-axis and to the s^n-axis, along an absolutely continuous curve (s^*, I) that satisfies conditions (1) and (2) listed in the proof of Theorem 3, where $f^* = (f, 1)$, and such that $s^*(t) \in S_0^*$ for all $t \in I$.

The extension of Theorem 2 is applicable to $P(s^*)$. Therefore, there exists a nonzero $(n+1)$-dimensional vector function ψ_{s*} satisfying the conclusions of Theorem 1 such that $\psi_{s*}(t_1)$ is perpendicular to the line through $(s_1, 0)$ parallel to the s^n-axis. Hence, the $(n+1)$-component $\psi_{s*}^n(t_1)$ of $\psi_{s*}(t_1)$ equals 0, and

$$\hat{\psi}_{s*}^0(t_1) = (\psi_{s*}(t_1), \cdots, \psi_{s*}^{n-1}(t_1)) \neq 0.$$

But, $\Delta s^* = (\Delta s, \Delta t)$; hence,

$$(\hat{\psi}_{s*}(t_1), \Delta s) \leqq 0. \tag{7.10}$$

Observe that the equations for δx are still

$$\delta \dot{x} = \left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial u}\frac{\partial B_i}{\partial s} K\right] \delta x \tag{7.11}$$

on $J_i = [\bar{t}_{i-1}, \bar{t}_i]$, where the $\bar{t}_i$ are determined in the construction of the cones considered (see the proof of Theorem 3). This follows because on J_i, $\delta t = 0$, so

$$\partial u = \frac{\partial B_i}{\partial s^*}\partial s^* = \frac{\partial B_i}{\partial s}\delta s + \frac{\partial B_i}{\partial t}\delta t = \frac{\partial B_i}{\partial s}\delta s. \tag{7.12}$$

Consequently, since f is autonomous,

$$\delta \dot{x}^* = \begin{bmatrix} \frac{\partial f}{\partial x} & \cdots & 0 \\ 0 & \cdots & 0 \end{bmatrix} \delta x^* + \begin{pmatrix} \frac{\partial f}{\partial u} \\ 0 \end{pmatrix} \left(\frac{\partial B_i}{\partial s}\right) \begin{bmatrix} K & \cdots & 0 \\ 0 & \cdots & 0 \end{bmatrix} \delta x^*. \tag{7.13}$$

Hence, the equation for the function ψ^* for $P(x^*)$ is

$$\dot{\psi}^* = -\begin{bmatrix} \left(\frac{\partial f}{\partial x}\right)^T & \cdots & 0 \\ 0 & \cdots & 0 \end{bmatrix} \psi^* - \begin{pmatrix} K^T & \cdots & 0 \\ 0 & \cdots & 0 \end{pmatrix} \left(\frac{\partial B_i}{\partial s}\right)^T \begin{pmatrix} \left(\frac{\partial f}{\partial u}\right)^T & \cdots & 0 \\ 0 & \cdots & 0 \end{pmatrix} \psi^*. \tag{7.14}$$

Consequently, $\psi^{n+1*} \equiv 0$. Define $\psi_0^*(t)$ on I_0 to be the solution of (7.14)

such that $\psi_0^*(t_1) = k^{*T}(s^*(t_1))\psi_{s*}(t_1)$. Then the $(n+2)$-component of ψ_0^* is identically zero on I_0. Therefore,

$$0 = M^*(t) = \max_{w \in W} (\psi_0^*(t), f^*(\hat{x}_0(t), B(s_0^*(t), w))) \tag{7.15}$$
$$= (\psi_0^*(t), f^*(\hat{x}_0(t), u_0(t))) \text{ a.e.},$$

where $B = B_i$ on J_i implies that

$$0 = M(t) = \max_{w \in W} (\psi_0(t), f(\hat{x}_0(t), D(\hat{x}_0(t), t, w))) \tag{7.16}$$
$$= (\psi_0(t), f(\hat{x}_0(t), u_0(t))) \text{ a.e.},$$

where $D = D_i$ on J_i. Consequently, the results of Theorem 3 hold for such problems.

8. Remarks. Gamkrelidze used his assumption that the controls were piecewise continuous to prove that the inner product $(\psi_0(t), f(\hat{x}_0(t), u_0(t))) \equiv 0$. As demonstrated, in general this result can be obtained a.e. It should also be noted that the sets $B(s_0^*(t), W)$ over which the maximization is made in Theorem 4 may be proper subsets of the sets $\omega(x_0(t))$ considered by Gamkrelidze. Moreover, it is clear that the variational equations obtained by Gamkrelidze can be chosen such that for the particular trajectories considered in the proof of Theorem 4, they reduce to the variational equations (7.11) obtained in the proof of Theorem 4.

It is interesting to observe that the proofs given demonstrate that a local maximum principle for the problem being considered is directly obtainable from Pontryagin's results [8, pp. 75–114]. One comment should be made about the statement of Pontryagin's maximum principle [8, p. 75]. Although the statement of this principle says that the control set U can be an arbitrary set, the proof given assumes otherwise. In the proof, the comparison made between the function $M(t) = \sup_{u \in U} (\psi(t), f(\hat{x}_0(t), u))$ and the function $m(t) = \max_{u \in P} (\psi(t), f(\hat{x}_0(t), u))$ is valid only if the set P is contained in U. Hence, this comparison is valid for problems in which the controls are arbitrary bounded and measurable functions, if the control set U is closed; and in problems in which the control set U is an arbitrary set, if the controls are restricted to piecewise continuous functions whose right-hand and left-hand limits at points of discontinuity are also in U. However, this comparison is obviously not valid in general.

Finally, since the work on this paper was completed, two more papers, [13] and [4], dealing with the problem considered in this paper have been published.

9. Acknowledgment. The author wishes to thank Professor S. P. Diliberto, the University of California, Berkeley, for suggesting the problem of the elucidation and extension of the results obtained by Gamkrelidze.

REFERENCES

[1] L. D. Berkovitz, *Variational methods in problems of control and programming*, J. Math. Anal. Appl., 3 (1961), pp. 145–169.

[2] ———, *On control problems with bounded state variables*, Ibid., 5 (1962), pp. 488–498.

[3] A. F. Filippov, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84 (translation from Russian).

[4] T. Guinn, *The problem of bounded space coordinates as a problem of Hestenes*, Ibid., 3 (1965), pp. 181–190.

[5] E. J. McShane, *On multipliers for Lagrange problems*, Amer. J. Math., 61 (1939), pp. 809–819.

[6] ———, *Necessary conditions in generalized curve problems of the calculus of variations*, Duke Math. J., 7 (1940), pp. 1–27.

[7] I. P. Natanson, *Theory of Functions of a Real Variable*, vol. I, Ungar, New York, 1964.

[8] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[9] W. Rudin, *Principles of Mathematical Analysis*, 2nd ed., McGraw-Hill, New York, 1964.

[10] J. Warga, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

[11] ———, *Necessary conditions for minimum in relaxed variational problems*, Ibid., 4 (1962), pp. 129–145.

[12] ———, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112 (1964), pp. 432–455.

[13] ———, *Unilateral variational problems with several inequalities*, Michigan Math. J., 12 (1965), pp. 449–480.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

AN ABSTRACT VARIATIONAL THEORY WITH APPLICATIONS TO A BROAD CLASS OF OPTIMIZATION PROBLEMS. I. GENERAL THEORY*

LUCIEN W. NEUSTADT†

1. Introduction. This article is devoted to the formulation of a very general variational problem and to the derivation of necessary conditions which solutions of this problem must satisfy.

The variational problem is formulated in §2 in the setting of a locally convex linear topological space. The basic concepts in this formulation are those of an internal cone, a first-order convex approximation, and a particular type of differential. This threesome, roughly speaking, represents a "linearization" of the constraints imposed by the problem. Of the three, the most unorthodox is the first-order, convex approximation, which turns out to be a quite natural extension of the cone of attainability introduced by Pontryagin, Boltyanskii and Gamkrelidze (see [2, Chap. II]), and of the convex set of linear variations introduced by Gamkrelidze in his work on quasiconvexity (see the set K in [3, p. 115]). The idea of such sets is originally due to McShane [4]. It appears to the author that the first-order, convex approximation, which is a convex set in the underlying linear space, is the most suitable device for handling variational problems which include constraints in the form of ordinary differential equations. This will be demonstrated in Part II of this article wherein we construct such approximations for a number of optimal control problems.

The necessary conditions satisfied by solutions of the general variational problem are in the form of a separation theorem for two convex sets, one of which is the set discussed above, and the second of which is related to the other two "linearizations" of the constraints. Thus our result is very much in the spirit of [2]–[4]. These conditions are spelled out in Theorems 2.1 and 2.2 and Corollary 3.1, whose proof is set forth in §3. In §4 we formulate a canonical optimization problem whose solutions satisfy, under suitable regularity conditions, the hypotheses of Theorems 2.1 and 2.2. The necessary conditions of these theorems are then specialized for this particular problem (see Theorems 4.1–4.6).

It will be shown in Part II that the canonical optimization problem

* Received by the editors February 23, 1966.

† Department of Electrical Engineering, University of Southern California, Los Angeles, California. This work was supported by the Joint Services Electronics Programs (United States Army, United States Navy, and United States Air Force) under Grant AF-AFOSR-496-66 and by the United States Air Force Office of Scientific Research under Grant AF-AFOSR-1029-66.

includes, as special cases, virtually all of the optimal control problems based on ordinary differential equations or difference equations which have come to the fore in recent years, including the conventional optimal control problem (both with and without restricted phase coordinates, and with fixed or variable endpoints and initial and terminal times), discrete optimal control problems, and minimax control problems. The necessary conditions for solutions of these problems, obtained on the basis of Theorems 4.1–4.6, then include as special cases all of the first-order necessary conditions in the classical calculus of variations, as well as the Pontryagin maximum principle and its various extensions and generalizations recently obtained (sometimes under hypotheses much stronger than those required here). In addition, it is now possible to obtain necessary conditions for problems which heretofore have been outside of the realm of applicability of any of the existing variational theory.

In §4 we also indicate how a particular case of the canonical optimization problem can be looked upon as a mathematical programming problem in a locally convex linear topological space. Applying Theorems 2.1 and 2.2, we obtain necessary conditions which are generalizations of the well-known Kuhn-Tucker conditions.

2. Basic definitions. Let $\mathfrak{Z}$ be a locally convex linear topological space over the real numbers such that the topology on $\mathfrak{Z}$ induces the ordinary Euclidean topology on every finite-dimensional subspace of $\mathfrak{Z}$. We shall denote by $\mathfrak{Z}^*$ the conjugate space of $\mathfrak{Z}$, i.e., the linear vector space whose elements are the linear continuous functionals defined on $\mathfrak{Z}$. Let B and Q be subsets of $\mathfrak{Z}$, and let F be a continuous function defined on a neighborhood $\bar{N}$ of 0 in $\mathfrak{Z}$, taking on values in R^m (Euclidean m-space). Let

$$Y = \{ \mathbf{x} \mid \mathbf{x} \in \bar{N}, \quad F(\mathbf{x}) = 0 \}. \tag{2.1}$$

DEFINITION 2.1. We shall say that $0 \in \mathfrak{Z}$ is a (Q, B, F)-*extremal* if there is a neighborhood N^* of 0 in $\mathfrak{Z}$ such that $Y \cap Q \cap B \cap N^* = \{0\}$.

DEFINITION 2.2. We shall say that $0 \in \mathfrak{Z}$ is a (Q, B)-*extremal* if there is a neighborhood N^* of 0 in $\mathfrak{Z}$ such that $Q \cap B \cap N^* = \{0\}$.

Thus, if 0 is a (Q, B, F)-extremal, then $0 \in Y \cap Q \cap B$; if 0 is a (Q, B)-extremal, then $0 \in Q \cap B$.

In order to obtain meaningful necessary conditions for extremality, it will be necessary to make some additional assumptions on the sets B and Q and the function F.

First, we shall assume that there exists a continuous linear transformation $\lambda(\mathbf{x})$ from $\mathfrak{Z}$ onto R^m such that

$$\frac{F(\epsilon \mathbf{y})}{\epsilon} \underset{\substack{\epsilon \to 0 \\ \mathbf{y} \to \mathbf{x}}}{\longrightarrow} \lambda(\mathbf{x}) \quad \text{for every} \quad \mathbf{x} \in \mathfrak{Z}. \tag{2.2}$$

Let $\lambda(\mathbf{x}) = (l_1(\mathbf{x}), \cdots, l_m(\mathbf{x}))$, where $l_i \in \mathfrak{Z}^*$ for each i. Since λ is onto R^m, $l_1, \cdots, l_m$ are linearly independent. It also follows from (2.2) that $F(0) = 0$.

Second, we shall assume that there exists a convex cone Z in $\mathfrak{Z}$ with vertex at 0 (and containing points other than 0) such that, if ρ is any ray in Z, there exists a cone Z_ρ with vertex at 0 and a neighborhood N_ρ of 0 in $\mathfrak{Z}$ (both possibly depending on ρ) such that (a) $Z_\rho \subset Z$, (b) Z_ρ has a non-empty interior and ρ is an interior ray of Z_ρ, (c) $Z_\rho \cap N_\rho \subset B$. If Z and B satisfy these conditions, we shall say that Z is *an internal cone* for B at 0.

Note 2.1. If B is a convex set with interior points and $0 \in B$, then $Z = \{\eta\mathbf{x} \mid \eta \geqq 0, \mathbf{x} \in (\text{interior of } B)\}$ is an internal cone for B at 0. Indeed it is clear that Z is a convex cone in $\mathfrak{Z}$ with vertex at 0. Further, let ρ be a ray in Z and let $\mathbf{y}_0 \in \rho$, $\mathbf{y}_0 \neq 0$, so that $\mathbf{y}_0 = \eta_0\mathbf{x}_0$, where $\mathbf{x}_0 \in$ (interior of B) and $\eta_0 > 0$. Also, let N_ρ be a convex neighborhood of 0 in $\mathfrak{Z}$ such that $\mathbf{x}_0 + N_\rho \subset B$ and $\mathbf{x}_0 \notin N_\rho - N_\rho$. Then if we set $Z_\rho = \{\eta\mathbf{x} \mid \eta \geqq 0, \mathbf{x} \in \mathbf{x}_0 + N_\rho\}$, it is easily seen that Z_ρ is a cone with vertex at 0, that $Z_\rho \subset Z$, that ρ is an interior ray of Z_ρ, and that $Z_\rho \cap N_\rho \subset B$.

If ν is any positive integer, we shall throughout this paper denote by I^ν the following subset of R^ν:

$$I^\nu = \{\beta = (\beta_1, \cdots, \beta_\nu) \mid \beta_i \geqq 0 \quad \text{for} \quad i = 1, \cdots, \nu, \quad \sum_{i=1}^{\nu} \beta_i = 1\}.$$

Finally, we shall assume that there is a convex set $K \subset \mathfrak{Z}$ with the following properties: (a) $0 \in K$, and K contains points other than 0; (b) if $\{\mathbf{x}_1, \cdots, \mathbf{x}_\nu\}$ is any finite subset of K, and N is an arbitrary neighborhood of 0 in $\mathfrak{Z}$, then there exists a number $\epsilon_0 > 0$ (which may depend on the $\mathbf{x}_i$ and on N) such that, for every ϵ, $0 < \epsilon < \epsilon_0$, there exists a continuous map ζ_ϵ from I^ν to $\mathfrak{Z}$ satisfying the following relation:

$$(2.3) \quad \zeta_\epsilon(\beta) = \zeta_\epsilon(\beta_1, \cdots, \beta_\nu) \in \{\epsilon(\sum_{i=1}^{\nu} \beta_i\mathbf{x}_i + N)\} \cap Q \quad \text{for all} \quad \beta \in I^\nu.$$

In this case, we shall say that K is a *first-order, convex approximation* to Q.

Note 2.2. If $0 \in Q$ and Q is convex, then Q is a first-order, convex approximation to itself.

Note 2.3. If $\mathfrak{Z} = \mathfrak{Z}_1 \times \mathfrak{Z}_2$ and $Q = Q_1 \times Q_2$, where $\mathfrak{Z}_1$ and $\mathfrak{Z}_2$ are locally convex linear topological spaces and $Q_i \subset \mathfrak{Z}_i$, $i = 1$ and 2, and if K_1 and K_2 are first-order, convex approximations to Q_1 and Q_2 respectively, then $K_1 \times K_2$ is a first-order convex approximation to Q.

The conclusions in Notes 2.2 and 2.3 follow at once from the definition of a first-order, convex approximation.

Let the sets Π and Z' in $\mathfrak{Z}$ be defined as follows:

$$(2.4) \qquad \Pi = \{\mathbf{x} \mid \mathbf{x} \in \mathfrak{Z}, \quad \lambda(\mathbf{x}) = 0\},$$

$$Z' = Z \cap \Pi. \tag{2.5}$$

Clearly, Π is a closed linear manifold in $\mathfrak{Z}$, and Z' is a convex cone in Π with vertex at 0.

We can now state our fundamental necessary condition for extremality.

THEOREM 2.1. *Let Q and B be subsets of a locally convex linear topological space $\mathfrak{Z}$, and let F be a continuous function from a neighborhood $\bar{N}$ of 0 in $\mathfrak{Z}$ into R^m. Let $0 \in \mathfrak{Z}$ be a $(Q, B, F$-$)$extremal, and suppose that (2.2) holds for some linear, continuous transformation λ from $\mathfrak{Z}$ onto R^m. Further, let K be a first-order, convex approximation to Q, and let Z be an internal cone for B at 0, with $Z \neq \mathfrak{Z}$. Then either $Z' = \{0\}$ (where Z' is defined by (2.4) and (2.5)), or there is a nonzero functional $l^* \in \mathfrak{Z}^*$ separating K and Z'; i.e.,*

$$l^*(\mathbf{x}) \leqq 0 \leqq l^*(\mathbf{y}) \quad \text{for all} \quad \mathbf{x} \in K, \quad \mathbf{y} \in Z'. \tag{2.6}$$

THEOREM 2.2. *Let Q and B be subsets of a locally convex linear topological space $\mathfrak{Z}$, let K be a first-order convex approximation to Q, and let Z be an internal cone for B at 0, with $Z \neq \mathfrak{Z}$. Then if $0 \in \mathfrak{Z}$ is a (Q, B)-extremal, there is a nonzero functional $l^* \in \mathfrak{Z}^*$ separating K and Z; i.e.,*

$$l^*(\mathbf{x}) \leqq 0 \leqq l^*(\mathbf{y}) \quad \text{for all} \quad \mathbf{x} \in K, \quad \mathbf{y} \in Z.$$

The next section is devoted to the proof of Theorems 2.1 and 2.2.

3. The proof of the necessary conditions. We first prove a lemma.

LEMMA 3.1. *Let F be a continuous function defined on a neighborhood $\bar{N}$ of 0 in a linear topological space $\mathfrak{Z}$ and taking on values in R^m, and let λ be a continuous mapping from $\mathfrak{Z}$ into R^m such that (2.2) is satisfied. Then if S is any compact set in $\mathfrak{Z}$ and $\eta > 0$ is arbitrary, there is a neighborhood N_η of 0 in $\mathfrak{Z}$ and a number $\delta > 0$ (both δ and N_η may depend on η as well as on S) such that*

$$\left| \frac{F(\epsilon \mathbf{y})}{\epsilon} - \lambda(\mathbf{x}) \right| < \eta$$

whenever $\mathbf{x} \in S$, $0 < |\epsilon| < \delta$, and $\mathbf{y} \in \mathbf{x} + N_\eta$.

Proof. Let us fix $\eta > 0$. By hypothesis, for each $\mathbf{x} \in \mathfrak{Z}$ there is a neighborhood $\bar{N}_\mathbf{x}$ of 0 in $\mathfrak{Z}$ and a number $\delta_\mathbf{x} > 0$ such that (see (2.2) and recall that λ is continuous)

$$\left| \frac{F(\epsilon \mathbf{y})}{\epsilon} - \lambda(\mathbf{x}) \right| + |\lambda(\mathbf{x}) - \lambda(\mathbf{z})| < \eta \quad \text{whenever} \tag{3.1}$$

$$0 < |\epsilon| < \delta_\mathbf{x}, \ \mathbf{y} \in \mathbf{x} + \bar{N}_\mathbf{x}, \ \mathbf{z} \in \mathbf{x} + \bar{N}_\mathbf{x}.$$

For each $\mathbf{x} \in \mathfrak{Z}$, let $N_\mathbf{x}$ be a neighborhood of 0 such that $N_\mathbf{x} + N_\mathbf{x} \subset \bar{N}_\mathbf{x}$. Since S is compact, there are points $\mathbf{x}_1, \cdots, \mathbf{x}_k$ in S such that $S \subset \bigcup_{i=1}^{k} (\mathbf{x}_i$

$+ N_{\mathbf{x}_i}$). Let $N_\eta = \bigcap_{i=1}^k N_{\mathbf{x}_i}$ and let $\delta = \min_{1 \leqq i \leqq k} \{\delta_{\mathbf{x}_i}\}$. Further, let $\mathbf{x}$ be an arbitrary point of S, so that $\mathbf{x} \in \mathbf{x}_j + N_{\mathbf{x}_j}$ for some $j = 1, \cdots, k$. If $\mathbf{y} \in \mathbf{x} + N_\eta$, then $\mathbf{y} \in \mathbf{x} + N_{\mathbf{x}_j} \subset \mathbf{x}_j + N_{\mathbf{x}_j} + N_{\mathbf{x}_j} \subset \mathbf{x}_j + \bar{N}_{\mathbf{x}_j}$. Thus if $0 < |\epsilon| < \delta \leqq \delta_{\mathbf{x}_j}$, then (see (3.1))

$$\left| \frac{F(\epsilon \mathbf{y})}{\epsilon} - \lambda(\mathbf{x}) \right| \leqq \left| \frac{F(\epsilon \mathbf{y})}{\epsilon} - \lambda(\mathbf{x}_j) \right| + |\lambda(\mathbf{x}_j) - \lambda(\mathbf{x})| < \eta.$$

This completes the proof of the lemma.

Let us now turn to the proof of Theorem 2.1, which will be by contradiction. Let us assume that $Z' \neq \{0\}$. In the sequel we shall use the word *separable* to mean: capable of being separated by a nonzero, continuous linear functional. Thus let us suppose that K and Z' are not separable. Let $\bar{K} = \lambda(K)$, so that $\bar{K}$ is a convex set in R^m, and $0 \in \bar{K}$.

Let us show that if K and Z' are not separable, then $K \cap \Pi$ and Z' are not separable in Π. Indeed, suppose that $K \cap \Pi$ and Z' are separable, so that there is a functional $l_0 \in \mathfrak{Z}^*$ such that $l_0(\mathbf{x}) \leqq 0 \leqq l_0(\mathbf{y})$ for all $\mathbf{x} \in K \cap \Pi$ and $\mathbf{y} \in Z'$, and $l_0(\mathbf{x}') \neq 0$ for some $\mathbf{x}' \in \Pi$. Let $\tilde{K} = \{(l_0(\mathbf{x}), \lambda(\mathbf{x})) \mid \mathbf{x} \in K\}$. Evidently, $\tilde{K}$ is a convex set in R^{m+1}. By hypothesis, if

$$\bar{p} = \{(\xi_0, \xi_1, \cdots, \xi_m) \mid \xi_0 \geqq 0, \quad \xi_i = 0 \quad \text{for} \quad i = 1, \cdots, m\},$$

then $\bar{p} \cap \tilde{K} = \{0\}$. Therefore, there is a hyperplane through 0 in R^{m+1} separating $\tilde{K}$ from the ray $\bar{p}$, i.e., there is a vector $\zeta = (\zeta_0, \zeta_1, \cdots, \zeta_m) \in R^{m+1}$, $\zeta \neq 0$, such that $\zeta \cdot \xi \leqq 0 \leqq \zeta \cdot \xi'$ for all $\xi \in \tilde{K}$, $\xi' \in \bar{p}$. Consequently, $\zeta_0 \geqq 0$ and $\sum_{i=0}^m \zeta_i l_i(\mathbf{x}) \leqq 0$ for all $\mathbf{x} \in K$. (Recall that $\lambda = (l_1, \cdots, l_m)$.) Let $l' = \sum_{i=0}^m \zeta_i l_i$; $l' \in \mathfrak{Z}^*$. Let us show that $l' \neq 0$. Indeed, if $l' = 0$, then $\zeta_0 \neq 0$ since $\zeta \neq 0$ and $l_1, \cdots, l_m$ are linearly independent. Hence, $l'(\mathbf{x}') = \zeta_0 l_0(\mathbf{x}') \neq 0$, and this contradiction implies that $l' \neq 0$. Furthermore, $l'(\mathbf{x}) \leqq 0$ for all $\mathbf{x} \in K$, and i $\mathbf{y} \in Z' = Z \cap \Pi$, then $l'(\mathbf{y}) = \zeta_0 l_0(\mathbf{y}) \geqq 0$ so that K and Z' are separable. This contradicts our hypothesis, so that $K \cap \Pi$ and Z' are not separable in Π.

By hypothesis, every ray of Z is an interior ray thereof, and $Z \cap \Pi \neq \{0\}$; consequently, $Z' = Z \cap \Pi$ has a nonempty interior in Π. Let $I(Z')$ denote the interior of Z' relative to Π. It is clear that $Z' = I(Z') \cup \{0\}$. We can now conclude that there is a point $2\hat{\mathbf{x}} \in I(Z') \cap (K \cap \Pi)$, for in the contrary case (see [1, p. 417, Theorem 8]) $I(Z')$ (as well as Z') could be separated from $K \cap \Pi$ in Π, contradicting the conclusion of the preceding paragraph. Let us show that $\hat{\mathbf{x}} \neq 0$. It is sufficient to prove that $0 \notin I(Z')$. To verify the latter statement note that if $0 \in I(Z')$, then there is a point $\mathbf{x} \in Z'$ such that $(-\mathbf{x}) \in Z'$ and $\mathbf{x} \neq 0$. By the definition of an internal cone, this means that $\mathbf{x}$ and $(-\mathbf{x})$ are interior points of Z, so that 0 is an interior point of Z (the interior of a convex set is convex), contradicting

the hypotheses that Z is a cone and $Z \neq \mathfrak{Z}$. Also, $0 \in Z' \cap K$ and $Z' \cap K$ is convex, so that $\hat{\mathbf{x}} \in Z' \cap K$. Further, since $\hat{\mathbf{x}} \in \Pi$ (see (2.4)),

$$\lambda(\hat{\mathbf{x}}) = 0. \tag{3.2}$$

Let us now show that 0 is an interior point of $\bar{K}$ in R^m. Indeed, if 0 is a boundary point of $\bar{K}$, then there is a vector $\zeta = (\zeta_1, \cdots, \zeta_m) \in R^m, \zeta \neq 0$, such that $\zeta \cdot \bar{k} \leqq 0$ for all $\bar{k} \in \bar{K} = \lambda(K)$, i.e., $\zeta \cdot \lambda(\mathbf{x}) \leqq 0$ for all $\mathbf{x} \in K$. Let $l'(\mathbf{x}) = \zeta \cdot \lambda(\mathbf{x})$, so that $l' \in \mathfrak{Z}^*$ and, since $l_1, \cdots, l_m$ are linearly independent, $l' \neq 0$. Also, $l'(\mathbf{x}) \leqq 0$ for all $\mathbf{x} \in K$, $l'(\mathbf{y}) = 0$ for all $\mathbf{y} \in \Pi$, and, a fortiori, $l'(\mathbf{y}) \geqq 0$ for all $\mathbf{y} \in Z'$, contradicting our assumption that K and Z' are not separable.

Thus, 0 is an interior point of $\bar{K}$, and there is an m-simplex $\bar{S} \subset \bar{K}$ such that 0 is an interior point of $\bar{S}$. Let the vertices of $\bar{S}$ be $\sigma_i = \lambda(\mathbf{k}_i)$, where $\mathbf{k}_i \in K$, $i = 0, 1, \cdots, m$, and let V be the simplex in $\mathfrak{Z}$ with vertices $\mathbf{k}_0, \mathbf{k}_1, \cdots, \mathbf{k}_m$. Since K is convex, $V \subset K$. Also

$$\lambda(V) = \bar{S}. \tag{3.3}$$

Now $\hat{\mathbf{x}} \in Z' \subset Z$, $\hat{\mathbf{x}} \neq 0$, and, since Z is an internal cone for B at 0, there are a neighborhood $\hat{N}$ of 0 in $\mathfrak{Z}$ and a cone $\hat{Z} \subset Z$ (with vertex at 0) such that $\hat{\mathbf{x}} \in$ (interior of $\hat{Z}$) and

$$\hat{Z} \cap \hat{N} \subset B. \tag{3.4}$$

Since 0 is a (Q, B, F)-extremal, there is a neighborhood N^* of 0 in $\mathfrak{Z}$ such that $Y \cap Q \cap B \cap N^* = \{0\}$, where Y is given by (2.1).

Let N_1 be a neighborhood of 0 in $\mathfrak{Z}$ such that

$$\hat{\mathbf{x}} + N_1 \subset \hat{Z} \subset Z, \tag{3.5}$$

and let N_2 be a convex neighborhood of 0 in $\mathfrak{Z}$ such that

$$N_2 + N_2 \subset N_1 \cap \hat{N} \cap N^* \cap \bar{N}. \tag{3.6}$$

(Such a neighborhood exists because $\mathfrak{Z}$ is locally convex.)

Since N_2 is convex, there is a number δ_0, $0 < \delta_0 \leqq \frac{1}{2}$, such that $\delta_0 V \subset N_2$. Let $S = \hat{\mathbf{x}} + \delta_0 V$, so that S is a simplex in $\mathfrak{Z}$ with vertices $\mathbf{x}_i = \hat{\mathbf{x}} + \delta_0 \mathbf{k}_i$, $i = 0, 1, \cdots, m$, and $S \subset \hat{\mathbf{x}} + N_2$. Consequently (see (3.6) and (3.5)),

$$S + N_2 \subset \hat{\mathbf{x}} + N_2 + N_2 \subset \hat{\mathbf{x}} + N_1 \subset \hat{Z} \subset Z. \tag{3.7}$$

Since 0 is a boundary point of Z and $S + N_2$ is open, it follows from (3.7) that

$$0 \notin S + N_2. \tag{3.8}$$

Now $2\hat{\mathbf{x}} \in K$, $0 \in K$, $V \subset K$, and therefore, since K is convex, $S \subset K$.

Let $\tilde{S} = \lambda(S) = \lambda(\hat{\mathbf{x}}) + \delta_0\lambda(V) = \delta_0\bar{S}$ (see (3.2) and (3.3)) so that $0 \in$ (interior of $\tilde{S}$) $\subset R^m$. Let $\bar{\eta} > 0$ be such that if $\xi \in R^m$ and $|\xi| < \bar{\eta}$, then $\xi \in \tilde{S}$.

Since S is compact, it follows from Lemma 3.1 that there are a number $\bar{\delta} > 0$ and a neighborhood N_3 of 0 in $\mathfrak{Z}$ such that

$$\left|\frac{F(\epsilon\mathbf{y})}{\epsilon} - \lambda(\mathbf{x})\right| < \bar{\eta} \quad \text{whenever} \quad 0 < \epsilon < \bar{\delta}, \quad \mathbf{x} \in S, \quad \text{and} \quad \mathbf{y} \in \mathbf{x} + N_3. \tag{3.9}$$

Let

$$\tilde{N} = N_3 \cap N_2, \tag{3.10}$$

and let $\tilde{\eta} > 0$ be such that

$$\eta S \subset N_2 \quad \text{whenever} \quad 0 \leqq \eta \leqq \tilde{\eta}. \tag{3.11}$$

(Such an $\tilde{\eta}$ exists since N_2 is convex.) Also note that, because N_2 is convex and $0 \in N_2$,

$$\eta N_2 \subset N_2 \quad \text{whenever} \quad 0 \leqq \eta \leqq 1. \tag{3.12}$$

Since K is a first-order convex approximation to Q at 0, there are a continuous map $\tilde{\xi}$ from P^{m+1} to Q, and a number $\tilde{\eta}_1$ such that (see (2.3))

$$0 < \tilde{\eta}_1 < \min\{\tilde{\eta}, \bar{\delta}, 1\}, \tag{3.13}$$

$$\tilde{\xi}(\beta) = \tilde{\xi}(\beta_0, \beta_1, \cdots, \beta_m) \in \{\tilde{\eta}_1(\sum_{i=0}^{m} \beta_i\mathbf{x}_i + \tilde{N})\} \cap Q \quad \text{for all} \quad \beta \in P^{m+1}. \tag{3.14}$$

It follows from (3.10)–(3.14) and (3.6) that, for all $\beta \in P^{m+1}$,

$$\tilde{\xi}(\beta) \in \tilde{\eta}_1 S + \tilde{\eta}_1\tilde{N} \subset N_2 + N_2 \subset \bar{N} \cap \hat{N} \cap N^*. \tag{3.15}$$

If $\sigma \in \tilde{S} = \lambda(S)$, σ has a unique representation of the form

$$\sigma = \lambda(\sum_{i=0}^{m} \beta_i\mathbf{x}_i) \quad \text{with} \quad (\beta_0, \cdots, \beta_m) \in P^{m+1}. \tag{3.16}$$

Let us denote the mapping which assigns $\beta \in P^{m+1}$ to each $\sigma \in \tilde{S}$ in this manner by q. Clearly, q is a continuous mapping from $\tilde{S}$ onto P^{m+1}. Further, for each $\sigma \in \tilde{S}$, let

$$\gamma(\sigma) = -\frac{F(\tilde{\xi}(q(\sigma)))}{\tilde{\eta}_1} + \sigma, \tag{3.17}$$

so that γ is a continuous map from $\tilde{S}$ into R^m. Now, for each $\beta \in P^{m+1}$

(see (3.14), (3.13), and (3.10)), $\tilde{\zeta}(\beta) = \tilde{\eta}_1 \mathbf{y}$, where $\mathbf{y} \in \sum_{i=0}^{m} \beta_i \mathbf{x}_i + N_3$ and $0 < \tilde{\eta}_1 < \delta$. Hence, by virtue of (3.9), (3.16) and (3.17) $|\gamma(\sigma)| < \bar{\eta}$ for all $\sigma \in \tilde{S}$, or, by definition of $\bar{\eta}$, $\gamma(\sigma) \in \tilde{S}$. Hence, γ is a continuous map of $\tilde{S}$ into itself, and by the Brouwer fixed point theorem (recall that $\tilde{S} = \delta_0 \bar{S}$ is a simplex in R^m) there is a point $\tilde{\sigma} \in \tilde{S}$ such that $\gamma(\tilde{\sigma}) = \tilde{\sigma}$. Let $q(\tilde{\sigma}) = (\tilde{\beta}_0, \cdots, \tilde{\beta}_m) = \tilde{\beta}$, so that (see (3.17)) $F(\tilde{\zeta}(\tilde{\beta})) = 0$, or (see (2.1) and (3.14)),

$$\tilde{\zeta}(\tilde{\beta}) \in Y \cap Q. \tag{3.18}$$

Now (see (3.14), (3.10) and (3.7)),

$$(\tilde{\eta}_1)^{-1} \tilde{\zeta}(\tilde{\beta}) \in \sum_{i=0}^{m} \tilde{\beta}_i \mathbf{x}_i + \tilde{N} \subset S + N_2 \subset \hat{Z}. \tag{3.19}$$

Because $\hat{Z}$ is a cone with vertex at 0, it follows from (3.19) that $\tilde{\zeta}(\tilde{\beta}) \in \hat{Z}$. Thus we have shown that (see (3.15), (3.4), and (3.18)) $\tilde{\zeta}(\tilde{\beta}) \in Y \cap Q \cap B \cap N^*$. Finally, it follows from (3.19) and (3.8) that $\tilde{\zeta}(\tilde{\beta}) \neq 0$. But this contradicts the fact that $Y \cap Q \cap B \cap N^* = \{0\}$. This completes the proof of Theorem 2.1.

We shall use the following notation. If K is an arbitrary set in a linear topological space $\mathfrak{Z}$, we shall denote the cone with vertex at 0 generated by K by (cone K), i.e., cone $K = \{\eta \mathbf{x} \mid \mathbf{x} \in K, \quad \eta \geqq 0\}$. The closure of (cone K) will be denoted by $\overline{\text{cone}}\, K$.

Note 3.1. It is obvious that if $l \in \mathfrak{Z}^*$ and $l(\mathbf{x}) \leqq 0$ for all $\mathbf{x} \in K$, then $l(\mathbf{x}) \leqq 0$ for all $\mathbf{x} \in \overline{\text{cone}}\, K$.

COROLLARY 3.1. *If the hypotheses of Theorem* 2.1 *are satisfied, there exists a vector* $\hat{\alpha} \in R^m$ *and a functional* $\bar{l} \in \mathfrak{Z}^*$ *such that*

$$\bar{l}(\mathbf{x}) + \hat{\alpha} \cdot \lambda(\mathbf{x}) \leqq 0 \quad \textit{for all} \quad \mathbf{x} \in \overline{\text{cone}}\, K; \tag{3.20}$$

$$\bar{l}(\mathbf{y}) \geqq 0 \quad \textit{for all} \quad \mathbf{y} \in Z; \tag{3.21}$$

$$\begin{aligned} &\bar{l} + \hat{\alpha} \cdot \lambda \neq 0 \quad \textit{if} \quad Z' \neq \{0\}, \\ &\bar{l} + \hat{\alpha} \cdot \lambda = 0 \quad \textit{and} \quad \bar{l} \neq 0 \quad \textit{if} \quad Z' = \{0\}. \end{aligned} \tag{3.22}$$

Proof. First suppose that $Z' \neq \{0\}$, and consider the set $\bar{Z} = \{(l^*(\mathbf{y}), \lambda(\mathbf{y})) \mid \mathbf{y} \in Z\} \subset R^{m+1}$, where $l^* \in \mathfrak{Z}^*$ is such that $l^* \neq 0$ and (2.6) holds. Since l^* and λ are linear and Z is convex, $\bar{Z}$ is convex. If $\bar{p}$ denotes the ray

$$\{(\xi_0, \xi_1, \cdots, \xi_m) \mid \xi_0 \leqq 0, \quad \xi_i = 0 \quad \text{for} \quad i = 1, \cdots, m\},$$

it follows from (2.6) that $\bar{Z} \cap \bar{p} = \{0\}$. Hence, there is a hyperplane through 0 in R^{m+1} separating $\bar{Z}$ from $\bar{p}$, i.e., there is a nonzero vector $(\beta_0, \beta) \in R^{m+1}$ such that $\beta_0 \leqq 0$, $\beta = (\beta_1, \cdots, \beta_m) \in R^m$ and $\beta_0 l^*(\mathbf{y}) + \beta \cdot \lambda(\mathbf{y}) \leqq 0$ for all $\mathbf{y} \in Z$. If $\beta_0 < 0$, we set $\hat{\alpha} = (\alpha_1, \cdots, \alpha_m)$, where $\alpha_i = -\beta_i/\beta_0$ for $i = 1, \cdots, m$, and $\bar{l} = l^* - \sum_{i=1}^{m} \alpha_i l_i$ (recall that

$\lambda = (l_1, \cdots, l_m)$); (3.20) and (3.21) follow at once. Let us show that the relation $\beta_0 = 0$ is impossible. Indeed, if $\beta_0 = 0$, then $\beta \cdot \lambda(\mathbf{y}) \leqq 0$ for all $\mathbf{y} \in Z$ where $\beta \neq 0$, $\beta \in R^m$. Since $l_1, \cdots, l_m$ are linearly independent, $\beta \cdot \lambda \not\equiv 0, \beta \cdot \lambda \in \mathfrak{Z}^*$. Also, $\beta \cdot \lambda(\mathbf{x}) = 0$ for all $\mathbf{x} \in \Pi$ (see (2.4)). By hypothesis, there is a vector $\mathbf{y}_0 \in \Pi \cap Z$, $\mathbf{y}_0 \neq 0$. Because Z is an internal cone for B at 0, $\mathbf{y}_0$ is an interior point of Z, which contradicts the relations $\beta \cdot \lambda \not\equiv 0$, $\beta \cdot \lambda(\mathbf{y}_0) = 0$, and $\beta \cdot \lambda(\mathbf{y}) \leqq 0$ for all $\mathbf{y} \in Z$. Finally, $\bar{l} + \hat{\alpha} \cdot \lambda = l^* \neq 0$ by hypothesis.

If $Z' = \{0\}$, Π does not meet the interior of Z (which is nonempty by definition) inasmuch as 0 is on the boundary of Z (because $Z \neq \mathfrak{Z}$). Hence, (see [1, p. 417, Theorem 8]), there is a nonzero functional $\bar{l} \in \mathfrak{Z}^*$ such that $\bar{l}(\mathbf{y}) \geqq 0$ whenever $\mathbf{y} \in Z$ and $\bar{l}(\mathbf{x}) \leqq 0$ whenever $\mathbf{x} \in \Pi$. Since Π is a subspace this means that $\bar{l}(\mathbf{x}) = 0$ whenever $\mathbf{x} \in \Pi$, i.e., whenever $l_i(\mathbf{x}) = 0$ for every $i = 1, \cdots, m$. But according to [1, p. 421, Lemma 10], this means that $\bar{l} = \sum_{i=1}^{m} (-\alpha_i l_i)$ for some real numbers $\alpha_1, \cdots, \alpha_m$. Setting $\hat{\alpha} = (\alpha_1, \cdots, \alpha_m)$, (3.20)–(3.22) now follow at once.

Let us now turn to the proof of Theorem 2.2, which will also be by contradiction. Thus let us suppose that K and Z are not separable. Let $I(Z)$ denote the interior of Z. Clearly, $Z = I(Z) \cup \{0\}$, and since $Z \neq \mathfrak{Z}$, $0 \notin I(Z)$. There is a point $\hat{\mathbf{x}} \in I(Z) \cap K$, for in the contrary case (see [1, p. 417, Theorem 8]), $I(Z)$, as well as Z, could be separated from K.

Since $\hat{\mathbf{x}} \in I(Z)$, $\hat{\mathbf{x}} \neq 0$, and it follows from the definition of an internal cone that there are a neighborhood $\hat{N}$ of 0 in $\mathfrak{Z}$ and a cone $\hat{Z} \subset Z$ (with vertex at 0) such that $\hat{\mathbf{x}} \in$ (interior of $\hat{Z}$) and (3.4) holds. Let N_1 be a neighborhood of 0 in $\mathfrak{Z}$ such that (3.5) holds. Because 0 is a (Q, B)-extremal, there is a neighborhood N^* of 0 in $\mathfrak{Z}$ such that

$$Q \cap B \cap N^* = \{0\}. \tag{3.23}$$

Finally, let N_2 be a convex neighborhood of 0 in $\mathfrak{Z}$ such that

$$N_2 + N_2 \subset \hat{N} \cap N^* \quad \text{and} \quad N_2 \subset N_1. \tag{3.24}$$

Thus, (3.12) is satisfied. Also, $\hat{\mathbf{x}} + N_2 \subset \hat{\mathbf{x}} + N_1 \subset Z$, so that, since $\hat{\mathbf{x}} + N_2$ is open and 0 is on the boundary of Z,

$$0 \notin \hat{\mathbf{x}} + N_2. \tag{3.25}$$

Let $\tilde{\eta} > 0$ be such that

$$\eta \hat{\mathbf{x}} \in N_2 \quad \text{whenever} \quad 0 \leqq \eta \leqq \tilde{\eta}. \tag{3.26}$$

Now K is a first-order, convex approximation to Q. Consequently, there are a positive number $\tilde{\eta}_1 \leqq \min\{1, \tilde{\eta}\}$ and a vector $\bar{\mathbf{x}} \in [\tilde{\eta}_1(\hat{\mathbf{x}} + N_2)] \cap Q$. It now follows from (3.12), (3.26), and (3.24) that

$$\bar{\mathbf{x}} \in \hat{N} \cap N^* \cap Q. \tag{3.27}$$

Also (see (3.24) and (3.5)),

$$\tilde{\eta}_1^{-1}\bar{\mathbf{x}} \in \hat{\mathbf{x}} + N_2 \subset \hat{\mathbf{x}} + N_1 \subset \hat{Z},$$

so that, since $\hat{Z}$ is a cone with vertex at 0, $\bar{\mathbf{x}} \in \hat{Z}$, which, by virtue of (3.27), (3.25) and (3.4), means that $\bar{\mathbf{x}} \in Q \cap B \cap N^*$, $\bar{\mathbf{x}} \neq 0$, contradicting (3.23), and thereby completing the proof of Theorem 2.2.

4. Optimization and mathematical programming problems as extremal problems. In this section we shall define a canonical optimization problem and shall show how solutions of this problem, under suitable regularity conditions, give rise to sets Q, Z, and B as well as functions F and λ that satisfy the hypotheses of Theorem 2.1 or of Theorem 2.2. This makes it possible to apply Corollary 3.1 or Theorem 2.2 and obtain necessary conditions for solutions of this optimization problem. We shall also consider a special case of the canonical optimization problem which is a generalized mathematical programming problem.

We begin with three lemmas.

LEMMA 4.1. *Let* $B = \bigcap_{i=0}^{\nu} B_i$, *where each* B_i, $i = 0, 1, \cdots, \nu$, *is a subset of a linear topological space* $\mathfrak{Z}$, *let* Z_i *for* $i = 0, 1, \cdots, \nu$, *be an internal cone for* B_i *at* 0, *and let* $Z = \bigcap_{i=0}^{\nu} Z_i$. *Then, if* $Z \neq \{0\}$, Z *is an internal cone for* B *at* 0.

Lemma 4.1 is an immediate consequence of the definition of an internal cone.

LEMMA 4.2. *Let* φ *be a function from a set* W *in a locally convex, linear topological space* $\mathfrak{Z}$ *into* R^1, *let* $\mathbf{z}$ *be an interior point of* W, *and suppose that there exists a continuous, convex functional* c *defined on* $\mathfrak{Z}$ *such that* $c(\mathbf{x}) < 0$ *for some* $\mathbf{x} \in \mathfrak{Z}$ *and*

$$(4.1) \quad \epsilon^{-1}[\varphi(\mathbf{z} + \epsilon\mathbf{y}) - \varphi(\mathbf{z})] \xrightarrow[\substack{\epsilon \to 0+ \\ \mathbf{y} \to \mathbf{x}}]{} c(\mathbf{x}) \quad \textit{for every} \quad \mathbf{x} \in \mathfrak{Z}.$$

Then $c(\eta\mathbf{x}) = \eta c(\mathbf{x})$ *for all* $\eta > 0$ *and all* $\mathbf{x} \in \mathfrak{Z}$, $c(0) = 0$, *and the set*

$$Z = \{\mathbf{x} \mid \mathbf{x} \in \mathfrak{Z}, \quad c(\mathbf{x}) < 0\} \cup \{0\}$$

is an internal cone at 0 *for the set*

$$B = \{\mathbf{x} \mid \mathbf{x} \in W - \mathbf{z}, \quad \varphi(\mathbf{z} + \mathbf{x}) < \varphi(\mathbf{z})\} \cup \{0\}.$$

Proof. It is an immediate consequence of (4.1) that $c(\eta\mathbf{x}) = \eta c(\mathbf{x})$ for all $\eta > 0$ and $\mathbf{x} \in \mathfrak{Z}$. Since c is continuous, $c(0) = 0$. Because c is also convex, Z is a convex cone with vertex at 0 containing interior points.

Now let ρ be an arbitrary ray in Z, and let $\mathbf{x}_0 \in \rho \subset Z$, $\mathbf{x}_0 \neq 0$, so that $c(\mathbf{x}_0) = -\zeta < 0$. Because of (4.1) and the continuity of c, there exist a convex neighborhood N of 0 in $\mathfrak{Z}$ and a number $\delta > 0$ such that $\epsilon(\mathbf{x}_0 + N) \subset W - \mathbf{z}$, $| \epsilon^{-1}[\varphi(\mathbf{z} + \epsilon\mathbf{y}) - \varphi(\mathbf{z})] - c(\mathbf{x}_0) | < \zeta/2$, $| c(\mathbf{y}) - c(\mathbf{x}_0) | < \zeta/2$,

and $c(\mathbf{w}) > -\zeta/2$, whenever $0 < \epsilon < \delta$, $\mathbf{w} \in N$, and $\mathbf{y} \in \mathbf{x}_0 + N$. Consequently,

$$c(\mathbf{y}) < -\zeta/2 \quad \text{and} \quad \epsilon^{-1}[\varphi(\mathbf{z} + \epsilon\mathbf{y}) - \varphi(\mathbf{z})] < -\zeta/2 < 0, \tag{4.2}$$
$$\text{whenever} \quad \mathbf{y} \in \mathbf{x}_0 + N \quad \text{and} \quad 0 < \epsilon < \delta.$$

Let Z_ρ be the (convex) cone with vertex at 0 generated by $\mathbf{x}_0 + N$. It is clear that $Z_\rho \subset Z$, that Z_ρ has a nonempty interior, and that ρ is an interior ray of Z_ρ. Further, $N \cap (\mathbf{x}_0 + N)$ is empty.

We shall show that $Z_\rho \cap \delta N \subset B$, which will complete the proof of the lemma. Indeed, let $\mathbf{x} \in Z_\rho \cap \delta N$, $\mathbf{x} \neq 0$, so that $\mathbf{x} = \epsilon_0(\mathbf{x}_0 + \mathbf{x}')$, where $\epsilon_0 > 0$ and $\mathbf{x}' \in N$. Thus, $\epsilon_0^{-1}\mathbf{x} = (\mathbf{x}_0 + \mathbf{x}') \in (\mathbf{x}_0 + N) \cap (\epsilon_0^{-1}\delta N)$. Since N is convex and $0 \in N$, $\epsilon N \subset N$ whenever $0 \leqq \epsilon \leqq 1$, so that $\epsilon N \cap (\mathbf{x}_0 + N)$ is empty whenever $0 \leqq \epsilon \leqq 1$, which implies that $\epsilon_0^{-1}\delta > 1$, or $\epsilon_0 < \delta$, i.e., $\mathbf{x} \in \epsilon_0(\mathbf{x}_0 + N) \cap (W - \mathbf{z})$, where $0 < \epsilon_0 < \delta$. Consequently (see (4.2)), $\varphi(\mathbf{z} + \mathbf{x}) - \varphi(\mathbf{z}) < 0$, i.e., $\mathbf{x} \in B$. Thus, $Z_\rho \cap \delta N \subset B$.

LEMMA 4.3. *Let Z_i, $i = 0, 1, \cdots, \nu$, be convex cones with vertex at 0 in a Banach space $\mathfrak{Z}$ such that $Z = \bigcap_{i=0}^{\nu} Z_i$ has a nonempty interior. Let*

$$L_i = \{l \mid l \in \mathfrak{Z}^*, \quad l(\mathbf{y}) \geqq 0 \quad \textit{for every} \quad \mathbf{y} \in Z_i\}, \qquad i = 0, \cdots, \nu,$$

$$L = \{l \mid l \in \mathfrak{Z}^*, \quad l(\mathbf{y}) \geqq 0 \quad \textit{for all} \quad \mathbf{y} \in Z\}.$$

Then $L = \{l \mid l = \sum_{i=0}^{\nu} l_i, \quad l_i \in L_i \quad \textit{for} \quad i = 0, \cdots, \nu\}$.

Lemma 4.3 is an immediate consequence of Corollary 2 in [9, p. 51].

Consider the following problem, which we shall refer to as the *canonical optimization problem.*

Given two sets Q' and W in a locally convex, linear topological space $\mathfrak{Z}$, and real-valued functions φ_i, $i = 1, \cdots, m, 0, -1, \cdots, -\mu$, defined on W, find an element $\mathbf{x} \in W \cap Q'$ that (a) satisfies the equations $\varphi_i(\mathbf{x}) = 0$ for $i = 1, \cdots, m$, (b) satisfies the inequalities $\varphi_{-i}(\mathbf{x}) \leqq 0$ for $i = 1, \cdots, \mu$, and (c) in so doing, minimizes the value of the functional φ_0.

If $\mathbf{x} \in W$, let $\mathcal{J}_\mathbf{x}$ denote the set of those indices $i = 1, \cdots, \mu$ for which $\varphi_{-i}(\mathbf{x}) < 0$, and let $\mathcal{I}_\mathbf{x}$ denote the set of those indices $i = 0, 1, \cdots, \mu$ such that $i \notin \mathcal{J}_\mathbf{x}$. Note that $0 \in \mathcal{I}_\mathbf{x}$ for all $\mathbf{x} \in W$.

t DEFINITION 4.1. An element $\mathbf{z} \in \mathfrak{Z}$ is a *local solution* of the canonical optimization problem if $\mathbf{z} \in W \cap Q'$, $\varphi_i(\mathbf{z}) = 0$ for $i = 1, \cdots, m$, $\varphi_{-i}(\mathbf{z}) \leqq 0$ for $i = 1, \cdots, \mu$, and if there is a neighborhood N of 0 in $\mathfrak{Z}$ such that $\varphi_0(\mathbf{x}) \geqq \varphi_0(\mathbf{z})$ for all $\mathbf{x} \in (\mathbf{z} + N) \cap W \cap Q'$ that in addition satisfy the relations $\varphi_i(\mathbf{x}) = 0$ for $i = 1, \cdots, m$ and $\varphi_{-i}(\mathbf{x}) \leqq 0$ for $i = 1, \cdots, \mu$.

Clearly, every solution of the canonical optimization problem is also a local solution thereof.

DEFINITION 4.2. A local solution $\mathbf{z}$ of the canonical optimization problem is *regular* if:

(1) $\mathbf{z}$ is an interior point of W and the functionals $\varphi_1, \cdots, \varphi_m$ are continuous in a neighborhood of $\mathbf{z}$;

(2) there exist functionals $l_i \in \mathfrak{Z}^*$, $i = 1, \cdots, m$, such that, for every $i \geqq 1$,

$$\frac{\varphi_i(\mathbf{z} + \epsilon\mathbf{y})}{\epsilon} \xrightarrow[\substack{\epsilon \to 0 \\ \mathbf{y} \to \mathbf{x}}]{} l_i(\mathbf{x}) \quad \text{for every} \quad \mathbf{x} \in \mathfrak{Z}; \tag{4.3}$$

(3) the functionals $l_1, \cdots, l_m$ are linearly independent;

(4) either $\mathcal{J}_\mathbf{z}$ is empty or there is a neighborhood N^* of 0 in $\mathfrak{Z}$ such that $\varphi_{-i}(\mathbf{z} + \mathbf{x}) \leqq 0$ for all $\mathbf{x} \in N^*$ and every $i \in \mathcal{J}_\mathbf{z}$;

(5) for every $i \in \mathcal{J}_\mathbf{z}$, there exists a cone $Z_i \subset \mathfrak{Z}$ which is an internal cone at 0 for the set

$$B_i = \{\mathbf{x} \mid (\mathbf{z} + \mathbf{x}) \in W, \quad \varphi_{-i}(\mathbf{z} + \mathbf{x}) < \varphi_{-i}(\mathbf{z})\} \cup \{0\} \tag{4.4}$$

such that if

$$Z = \bigcap_{i \in \mathcal{J}_\mathbf{z}} Z_i, \tag{4.5}$$

then $Z \neq \mathfrak{Z}$ and $Z \neq \{0\}$.

Note 4.1. If $\mathfrak{Z}$ is a Banach space and if φ_i, $i = 1, \cdots, m$, has a Fréchet differential $l_i \in \mathfrak{Z}^*$ at $\mathbf{z}$, it is easily seen that (4.3) is satisfied. (Recall that $\varphi_i(\mathbf{z}) = 0$.) However, (4.3) does not necessarily imply that φ_i has a Fréchet differential at $\mathbf{z}$.

Note 4.2. If the functionals φ_{-i}, for $i \in \mathcal{J}_\mathbf{z}$, are upper semicontinuous at $\mathbf{z}$, it is evident that condition (4) in Definition 4.2 will be satisfied.

DEFINITION 4.3. A local solution $\mathbf{z}$ of the canonical optimization problem is *totally regular* if (i) for each $i \in \mathcal{J}_\mathbf{z}$,

$$\epsilon^{-1}[\varphi_{-i}(\mathbf{z} + \epsilon\mathbf{y}) - \varphi_{-i}(\mathbf{z})] \xrightarrow[\substack{\epsilon \to 0+ \\ \mathbf{y} \to \mathbf{x}}]{} c_i(\mathbf{x}) \quad \text{for every} \quad \mathbf{x} \in \mathfrak{Z}, \tag{4.6}$$

where the c_i are certain continuous, convex functionals defined on $\mathfrak{Z}$; (ii) conditions (1)–(4) of Definition 4.2 are satisfied; (iii) $c_j(\mathbf{x}) > 0$ for some $\mathbf{x} \in \mathfrak{Z}$ and some $j \in \mathcal{J}_\mathbf{z}$; (iv) there is an $\mathbf{x} \in \mathfrak{Z}$ such that $c_i(\mathbf{x}) < 0$ for every $i \in \mathcal{J}_\mathbf{z}$.

DEFINITION 4.4. A local solution $\mathbf{z}$ of the canonical optimization problem is *smoothly regular* if, for every $i \in \mathcal{J}_\mathbf{z}$, (4.6) holds with some $c_i \in \mathfrak{Z}^*$, if conditions (1), (2), and (4) of Definition 4.2 are satisfied, and if the relations

$$\sum_{i \in \mathcal{J}_\mathbf{z}} \alpha_{-i} c_i + \sum_{i=1}^{m} \alpha_i l_i = 0, \quad \alpha_{-i} \leqq 0 \quad \text{for every} \quad i \in \mathcal{J}_\mathbf{z}, \tag{4.7}$$

imply that $\alpha_i = 0$ for all $i = 1, \cdots, m$ and $-i \in \mathcal{J}_\mathbf{z}$.

The following lemma is an immediate consequence of Lemma 4.2.

LEMMA 4.4. *Every totally regular local solution of a canonical optimization problem is a regular local solution, where the sets Z_i in condition* (5) *of Definition* 4.2 *are defined by*

$$Z_i = \{\mathbf{x} \mid \mathbf{x} \in \mathfrak{I},\quad c_i(\mathbf{x}) < 0\} \cup \{0\}. \tag{4.8}$$

LEMMA 4.5. *Every smoothly regular local solution of a canonical optimization problem is a totally regular (and consequently regular) local solution. Further, if*

$$\Pi = \{\mathbf{x} \mid \mathbf{x} \in \mathfrak{I},\quad l_i(\mathbf{x}) = 0 \quad \text{for} \quad i = 1, \cdots, m\},\quad Z' = Z \cap \Pi, \tag{4.9}$$

where Z is given by (4.5) *and the Z_i by* (4.8), *then $Z' \neq \{0\}$.*

Proof. Let $\mathbf{z}$ be a smoothly regular local solution of a canonical optimization problem, and let the sets Z_i (for $i \in \mathcal{I}_\mathbf{z}$), Z and Z' be defined by (4.8), (4.5), and (4.9), respectively.

Let us show that $Z' \neq \{0\}$. Suppose the contrary. For ease of notation, and without loss of generality, we shall suppose that $\mathcal{I}_\mathbf{z} = \{0, 1, \cdots, \nu\}$, $\mathcal{J}_\mathbf{z} = \{\nu + 1, \cdots, \mu\}$, where $\nu \geqq 0$ (if $\mathcal{J}_\mathbf{z}$ is empty, then $\nu = \mu$). By hypothesis, the convex set $\{[c_\nu(\mathbf{x}), \cdots, c_0(\mathbf{x}), l_1(\mathbf{x}), \cdots, l_m(\mathbf{x})) \mid \mathbf{x} \in \mathfrak{I}\}$ in $R^{m+\nu+1}$ has an empty intersection with the convex set

$$\{(\xi_{-\nu}, \cdots, \xi_0, \xi_1, \cdots, \xi_m) \mid \xi_{-i} < 0 \quad \text{for} \quad i = 0, 1, \cdots, \nu,\ \xi_i = 0 \quad \text{for} \quad i = 1, \cdots, m\},$$

so that these two sets can be separated; i.e., there are numbers α_i, $i = -\nu, \cdots, 0, \cdots, m$, not all zero, such that $\alpha_{-i} \leqq 0$ for $i \in \mathcal{I}_\mathbf{z}$ and

$$\sum_{i=0}^{\nu} \alpha_{-i} c_i(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i l_i(\mathbf{x}) \leqq 0$$

for all $\mathbf{x} \in \mathfrak{I}$. But since $c_i \in \mathfrak{I}^*$ for each $i = 0, \cdots, \nu$, this is only possible if (4.7) holds, which by Definition 4.4 implies that $\alpha_i = 0$ for all i. This contradiction shows that $Z' \neq \{0\}$, and, a fortiori, $Z \neq \{0\}$.

It is clear from Definition 4.4 that $c_0 \neq 0$. Since $c_0 \in \mathfrak{I}^*$, there is an $\mathbf{x} \in \mathfrak{I}$ with $c_0(\mathbf{x}) > 0$. Finally, condition (3) in Definition 4.2 follows at once from Definition 4.4, and we conclude that $\mathbf{z}$ is totally regular.

LEMMA 4.6. *Let $\mathbf{z}$ be a local solution of a canonical optimization problem such that, for every $i \in \mathcal{I}_\mathbf{z}$,* (4.6) *holds with $c_i = l_{-i} \in \mathfrak{I}^*$, and suppose that conditions* (1), (2), *and* (4) *of Definition* 4.2 *are satisfied. Then if the functionals l_i, $i = 1, \cdots, m$ and $(-i) \in \mathcal{I}_\mathbf{z}$, are linearly independent, $\mathbf{z}$ is a smoothly regular local solution.*

The proof of Lemma 4.6 is obvious.

Now let $\mathbf{z}$ be a regular local solution of a canonical optimization problem, so that $\varphi_i(\mathbf{z}) = 0$ for $i = 1, \cdots, m$. Let $\bar{N}$ be a neighborhood of 0 in $\mathfrak{I}$

such that $\mathbf{z} + \bar{N} \subset W$ and $\varphi_1, \cdots, \varphi_m$ are continuous in $\mathbf{z} + \bar{N}$. Further, let F be the function from $\bar{N}$ into R^m defined by the relation

$$F(\mathbf{x}) = (\varphi_1(\mathbf{z} + \mathbf{x}), \cdots, \varphi_m(\mathbf{z} + \mathbf{x})). \tag{4.10}$$

Clearly, F is continuous and $F(0) = 0$. Further, by virtue of (4.3), we see that (2.2) holds, where $\lambda(\mathbf{x}) = (l_1(\mathbf{x}), \cdots, l_m(\mathbf{x}))$. Inasmuch as $l_i \in \mathfrak{Z}^*$ for $i = 1, \cdots, m$, λ is linear and continuous, and, since $l_1, \cdots, l_m$ are linearly independent, λ is onto R^m. Also, the set Π defined by (2.4) coincides with the set Π defined by (4.9). Finally if

$$B = \bigcap_{i \in \mathcal{I}_\mathbf{z}} B_i \quad \text{and} \quad Q = Q' - \mathbf{z}, \tag{4.11}$$

(where the B_i are defined by (4.4)), it is easily verified that $0 \in \mathfrak{Z}$ is a (Q, B, F)-extremal. We can now prove two basic theorems.

THEOREM 4.1. *Let $\mathbf{z}$ be a regular local solution of a canonical optimization problem, and let K be a first-order, convex approximation to $Q' - \mathbf{z}$. Then there exist real numbers $\alpha_1, \cdots, \alpha_m$ and a functional $\bar{l} \in \mathfrak{Z}^*$ such that*

$$\bar{l}(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i l_i(\mathbf{x}) \leqq 0 \quad \textit{for all} \quad \mathbf{x} \in \overline{\text{cone}}\, K; \tag{4.12}$$

$$\bar{l}(\mathbf{y}) \geqq 0 \quad \textit{for all} \quad \mathbf{y} \in Z; \tag{4.13}$$

$$\begin{aligned} &\bar{l} + \sum_{i=1}^{m} \alpha_i l_i \neq 0 \quad \textit{if} \quad Z' \neq \{0\}, \\ &\bar{l} + \sum_{i=1}^{m} \alpha_i l_i = 0 \textit{ and } \bar{l} \neq 0 \quad \textit{if} \quad Z' = \{0\}, \end{aligned} \tag{4.14}$$

(where the Z_i, for $i \in \mathcal{I}_\mathbf{z}$, are as indicated in condition (5) of Definition 4.2, and Z and Z' are given by (4.5) and (4.9), respectively).

Further, if $\mathfrak{Z}$ is a Banach space, or if $\mathcal{I}_\mathbf{z} = \{0\}$, or if Z_i is given by (4.8), with $c_i \in \mathfrak{Z}^*$, for all but one (or all) $i \in \mathcal{I}_\mathbf{z}$, then (4.13) implies that

$$\bar{l} = \sum_{i \in \mathcal{I}_\mathbf{z}} l_{-i}, \tag{4.15}$$

where, for each $i \in \mathcal{I}_\mathbf{z}$,

$$l_{-i} \in \mathfrak{Z}^*, \quad \textit{and} \quad l_{-i}(\mathbf{y}) \geqq 0 \quad \textit{for all} \quad \mathbf{y} \in Z_i. \tag{4.16}$$

If Z_i is given by (4.8) for some $i \in \mathcal{I}_\mathbf{z}$, where $c_i \in \mathfrak{Z}^*$, then (4.16) implies that $l_{-i} = \alpha_{-i} c_i$, where $\alpha_{-i} \leqq 0$.

Proof. Let us define B, Q, and F by means of (4.4), (4.11), and (4.10). It follows from Definition 4.2 and Lemma 4.1 that the set Z defined by (4.5) is an internal cone at 0 for B. We can therefore conclude, on the basis of Corollary 3.1, that there exist real numbers $\alpha_1, \cdots, \alpha_m$ and a vector $\bar{l} \in \mathfrak{Z}^*$ such that (4.12)–(4.14) hold.

For ease of notation, let us again suppose that $\mathcal{I}_{\mathbf{z}} = \{0, 1, \cdots, \nu\}$, where $\nu \geqq 0$.

If $\mathfrak{Z}$ is a Banach space, (4.15) and (4.16) follow from (4.13) and Lemma 4.3. (Note that Z has a nonempty interior by definition of an internal cone.) If $\mathcal{I}_{\mathbf{z}} = \{0\}$, so that $\nu = 0$ and $Z = Z_0$, (4.15) and (4.16) are clearly equivalent to (4.12).

Let us now prove that (4.13) implies (4.15) and (4.16) under the assumption that Z_i is given by (4.8) with $c_i \in \mathfrak{Z}^*$ for all but one (or all) $i \in \mathcal{I}_{\mathbf{z}}$. For ease of notation, and without loss of generality, let us suppose that (4.8) holds with $c_i \in \mathfrak{Z}^*$ for $i = 0, 1, \cdots, \nu - 1$ Clearly $\{(c_0(\mathbf{y}), c_1(\mathbf{y}), \cdots, c_{\nu-1}(\mathbf{y}), \bar{l}(\mathbf{y})) \mid \mathbf{y} \in Z_\nu\}$ is a convex subset of $R^{\nu+1}$,which, by (4.13), does not meet the open convex set

$$\{(\xi_0, \xi_1, \cdots, \xi_\nu) \mid \xi_i < 0 \quad \text{for} \quad i = 0, 1, \cdots, \nu\} \subset R^{\nu+1}.$$

Now, 0 is a limit point of both of these sets, so that there is a hyperplane through $0 \in R^{\nu+1}$ which separates them, i.e., there is a nonzero vector $(\beta_0, \beta_1, \cdots, \beta_\nu) \in R^{\nu+1}$ such that

$$\beta_\nu \bar{l}(\mathbf{y}) + \sum_{i=0}^{\nu-1} \beta_i c_i(\mathbf{y}) \leqq 0 \quad \text{for all} \quad \mathbf{y} \in Z_\nu, \tag{4.17}$$

$$\text{and} \quad \beta_i \leqq 0 \quad \text{for} \quad i = 0, \cdots, \nu.$$

By hypothesis (see condition (5) of Definition 4.2) there is an element $\mathbf{y}_0 \in Z = \bigcap_{i=0}^{\nu} Z_i$ such that $\mathbf{y}_0 \neq 0$. Consequently $\mathbf{y}_0 \in Z_\nu$ and $c_i(\mathbf{y}_0) < 0$ for $i = 0, 1, \cdots, \nu - 1$. By virtue of (4.17), this implies that $\beta_\nu < 0$. If we now set $l_{-i} = (-\beta_i/\beta_\nu)c_i$ for $i = 0, 1, \cdots, \nu - 1$ and $l_{-\nu} = \bar{l} - \sum_{i=0}^{\nu-1} l_{-i}$, (4.15) and (4.16) follow at once from (4.17).

It only remains to prove that (4.16) and (4.8) with $c_i \in \mathfrak{Z}^*$ imply that $l_{-i} = \alpha_{-i} c_i$, where $\alpha_{-i} \leqq 0$. But it is easy to see that these hypotheses imply that $l_{-i}(\mathbf{x}) = 0$ whenever $c_i(\mathbf{x}) = 0$, so that (see [1, p. 421, Lemma 10]), $l_{-i} = \alpha_{-i} c_i$ for some real number α_{-i}. If $\mathbf{x}_0 \in Z_i$ and $\mathbf{x}_0 \neq 0$, then $c_i(\mathbf{x}_0) < 0$ and $0 \leqq l_{-i}(\mathbf{x}_0) = \alpha_{-i} c_i(\mathbf{x}_0)$, which means that $\alpha_{-i} \leqq 0$.

THEOREM 4.2. *Let* $\mathbf{z}$ *be a totally regular, or smoothly regular local solution of a canonical optimization problem. Then if* K *is a first-order, convex approximation to* $Q' - \mathbf{z}$, *there exist real numbers* $\alpha_1, \cdots, \alpha_m$, *and* α_{-i} *for* $i \in \mathcal{I}_{\mathbf{z}}$, *not all zero, such that*

$$\sum_{i=1}^{m} \alpha_i l_i(\mathbf{x}) + \sum_{i \in \mathcal{I}_{\mathbf{z}}} \alpha_{-i} c_i(\mathbf{x}) \leqq 0 \quad \textit{for all} \quad \mathbf{x} \in \overline{\text{cone}}\, K, \tag{4.18}$$

$$\alpha_{-i} \leqq 0 \quad \textit{for} \quad i \in \mathcal{I}_{\mathbf{z}}.$$

If $Z' = \{0\}$ (*where* Z' *is given by* (4.5), (4.8), *and* (4.9)), *the inequality in* (4.18) *holds for all* $\mathbf{x} \in \mathfrak{Z}$. *If* $\mathbf{z}$ *is smoothly regular, then, in addition,*

$$\sum_{i=1}^{m} \alpha_i l_i + \sum_{i \in \vartheta_z} \alpha_{-i} c_i \neq 0.$$

Proof. Without loss of generality, and for ease of notation, we shall again suppose that $\vartheta_z = \{0, 1, \cdots, \nu\}$, where $\nu \geqq 0$. Let us first suppose that $Z' \neq \{0\}$. It follows from Lemmas 4.4 and 4.5 and Theorem 4.1 that there is a nonzero functional $l^* \in \mathfrak{Z}^*$ such that

$$(4.19) \quad l^*(\mathbf{x}) \leqq 0 \leqq l^*(\mathbf{y}) \quad \text{for all} \quad \mathbf{x} \in \overline{\text{cone}}\, K \quad \text{and} \quad \mathbf{y} \in Z'.$$

Consider the following two subsets of $R^{m+\nu+2}$:

$$S_1 = \{(l^*(\mathbf{x}), l_1(\mathbf{x}), \cdots, l_m(\mathbf{x}), c_0(\mathbf{x}) + \gamma_0, \cdots, c_\nu(\mathbf{x}) + \gamma_\nu) \mid \mathbf{x} \in \mathfrak{Z},$$

$$\gamma_i \geqq 0 \quad \text{for} \quad i = 0, 1, \cdots, \nu\},$$

$$S_2 = \{(\xi^*, \xi_1, \cdots, \xi_m, \xi_0, \xi_{-1}, \cdots, \xi_{-\nu}) \mid \xi^* < 0;$$

$$\xi_i = 0 \quad \text{for} \quad i = 1, \cdots, m; \xi_{-i} < 0 \quad \text{for} \quad i = 0, 1, \cdots, ,\nu\}.$$

Recalling that the c_i are convex functionals, we see at once that S_1 and S_2 are convex sets. Further, it follows from (4.5), (4.8), (4.9) and (4.19) that S_1 and S_2 have an empty intersection. Since 0 is a limit point of both S_1 and S_2, there is a nonzero vector $\tilde{\alpha} = (\alpha^*, \alpha_1, \cdots, \alpha_m, \alpha_0, \alpha_{-1}, \cdots, \alpha_{-\nu}) \in R^{m+\nu+2}$ such that $\tilde{\alpha} \cdot \tilde{\xi} \leqq 0 \leqq \tilde{\alpha} \cdot \tilde{\xi}'$ for all $\tilde{\xi} \in S_1$, $\tilde{\xi}' \in S_2$, or

$$(4.20) \quad \alpha^* l^*(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i l_i(\mathbf{x}) + \sum_{i=0}^{\nu} \alpha_{-i} c_i(\mathbf{x}) \leqq 0 \quad \text{for all} \quad \mathbf{x} \in \mathfrak{Z},$$

$$\alpha^* \leqq 0, \quad \alpha_{-i} \leqq 0 \quad \text{for} \quad i = 0, 1, \cdots, \nu.$$

If $\alpha^* = 0$, (4.18) is an obvious consequence of (4.20). If $\alpha^* < 0$, we shall suppose, without loss of generality, that $\alpha^* = -1$. Then, (4.20) can be written in the form

$$(4.21) \quad \sum_{i=1}^{m} \alpha_i l_i(\mathbf{x}) + \sum_{i=0}^{\nu} \alpha_{-i} c_i(\mathbf{x}) \leqq l^*(\mathbf{x}) \quad \text{for all} \quad \mathbf{x} \in \mathfrak{Z},$$

and we conclude, on the basis of (4.19) and (4.21), that (4.18) holds. The numbers α_i, $i = -\nu, \cdots, m$, cannot all vanish, for if $\alpha_i = 0$ for all i, then (see (4.21)) $l^*(\mathbf{x}) \geqq 0$ for all $\mathbf{x} \in \mathfrak{Z}$, which is absurd.

Now consider the case $Z' = \{0\}$. It follows from Theorem 4.1 that there are numbers $\tilde{\alpha}_1, \cdots, \tilde{\alpha}_m$, not all zero, such that $\sum_{i=1}^{m} \tilde{\alpha}_i l_i(\mathbf{y}) \geqq 0$ for all $\mathbf{y} \in Z$. Consequently, the convex subsets

$$\{(\sum_{i=1}^{m} \tilde{\alpha}_i l_i(\mathbf{x}), c_0(\mathbf{x}) + \gamma_0, \cdots, c_\nu(\mathbf{x}) + \gamma_\nu) \mid \mathbf{x} \in \mathfrak{Z},$$

$$\gamma_i \geqq 0 \quad \text{for} \quad i = 0, 1, \cdots, \nu\}$$

and $\{(\xi_1, \xi_0, \xi_{-1}, \cdots, \xi_{-\nu}) \mid \xi_i < 0 \text{ for } i = 1, 0, \cdots, -\nu\}$ of $R^{\nu+2}$, which have 0 as a common limit point, are disjoint. Consequently, there are nonpositive numbers $\bar{\alpha}_i$, $i = 1, 0, \cdots, -\nu$, not all zero, such that

$$\sum_{i=1}^{m} \bar{\alpha}_1 \bar{\alpha}_i l_i(\mathbf{x}) + \sum_{i=0}^{\nu} \bar{\alpha}_{-i} c_i(\mathbf{x}) \leqq 0 \quad \text{for all} \quad \mathbf{x} \in \mathfrak{I},$$

and, if we set $\alpha_i = \bar{\alpha}_1 \bar{\alpha}_i$ for $i = 1, \cdots, m$ and $\alpha_{-i} = \bar{\alpha}_{-i}$ for $i = 0, \cdots, \nu$, we have the desired conclusion. The last statement of Theorem 4.2 follows at once from Definition 4.4.

Note 4.3. If $\mathbf{z}$ is a totally regular, but not smoothly regular local solution, the necessary conditions expressed by Theorems 4.1 and 4.2 are generally quite distinct.

Note 4.4. The necessary conditions of Theorems 4.1 and 4.2 do not distinguish between φ_0 and the functionals φ_{-i} for $i \geqq 1$ and $i \in \mathcal{I}_{\mathbf{z}}$. This means that the form of these necessary conditions is unchanged if, in the original problem statement, the roles of the functional (φ_0) to be minimized and a functional $(\varphi_{-j}, j \geqq 1)$ defining an inequality constraint are interchanged. Further, if, for some $j = -\mu, \cdots, m, j \neq 0$,

$$(4.22) \quad \epsilon^{-1}[\varphi_j(\mathbf{x}_0 + \epsilon\mathbf{y}) - \varphi_j(\mathbf{x}_0)] \xrightarrow[\mathbf{y} \to \mathbf{x}]{\epsilon \to 0} l_j(\mathbf{x}\,; \mathbf{x}_0), \quad \text{where} \quad l_j(\cdot\,; \mathbf{x}_0) \in \mathfrak{I}^*,$$

for every $\mathbf{x}_0 \in$ (interior of W), the form of the necessary conditions (assuming that (4.15) and (4.16) hold in Theorem 4.1) will be unchanged if the role of φ_j in the original problem statement is changed from that of defining an equality constraint to that of defining an inequality constraint, or vice versa (except that the requirement $\alpha_j \leqq 0$ is present only if φ_j defines an inequality constraint). In particular, if (4.22) holds for every $\mathbf{x}_0 \in$ (interior of W) and $j = -\mu, \cdots, m$, the form of the necessary conditions (except for the sign of certain α_j) is invariant under an arbitrary change or interchange in the roles of the φ_j in the problem statement.

DEFINITION 4.5. A canonical optimization problem is *convex* if W is a convex set and the functionals $\varphi_0, \varphi_{-1}, \cdots, \varphi_{-\mu}$ are convex.

THEOREM 4.3. *Let* $\mathbf{z}$ *be a local solution of a convex canonical optimization problem such that conditions* (1)–(4) *of Definition* 4.2 *are satisfied, and suppose, in addition, that the functionals* φ_{-i}, *for* $i \in \mathcal{I}_{\mathbf{z}}$, *are continuous in* W *and that there are elements* $\mathbf{y}_1$ *and* $\mathbf{y}_2$ *in* W *such that* $\varphi_{-i}(\mathbf{y}_1) < \varphi_{-i}(\mathbf{z})$ *for every* $i \in \mathcal{I}_{\mathbf{z}}$ *and* $\varphi_{-j}(\mathbf{z} + \theta(\mathbf{y}_2 - \mathbf{z})) > \varphi_{-j}(\mathbf{z})$ *for every* θ, $0 \leqq \theta \leqq 1$, *and some* $j \in \mathcal{I}_{\mathbf{z}}$. *Then if* K *is a first-order, convex approximation to* $Q' - \mathbf{z}$, *there exist real numbers* $\alpha_1, \cdots, \alpha_m$, *and* α_{-i} *for* $i \in \mathcal{I}_{\mathbf{z}}$, *not all zero, such that* $\alpha_{-i} \leqq 0$ *for* $i \in \mathcal{I}_{\mathbf{z}}$ *and*

$$(4.23) \quad \sum_{i=1}^{m} \alpha_i l_i(\mathbf{x}) + \sum_{i \in \mathcal{I}_{\mathbf{z}}} \alpha_{-i} \varphi_{-i}(\mathbf{z} + \mathbf{x}) \leqq \alpha_0 \varphi_0(\mathbf{z})$$

for all $\mathbf{x} \in (\overline{\text{cone}}\, K) \cap (W - \mathbf{z})$.

If $B \cap \Pi = \{0\}$ (*where* B *is given by* (4.4) *and* (4.11) *and* Π *by* (4.9)), *then the inequality in* (4.23) *holds for all* $\mathbf{x} \in (W - \mathbf{z})$.

Proof. It follows at once from our hypotheses that $\mathbf{z}$ is a regular local solution where for each $i \in \mathcal{I}_{\mathbf{z}}$, the set Z_i in condition (5) of Definition 4.2 is cone (interior of B_i). (It is easily seen that B_i is convex and has a nonempty interior, so that Note 2.1 is applicable.)

The remainder of the proof parallels that of Theorem 4.2. Indeed, it is only necessary to replace $c_i(\mathbf{x})$ by $[\varphi_{-i}(\mathbf{z} + \mathbf{x}) - \varphi_{-i}(\mathbf{z})]$ and $\mathfrak{Z}$ by $(W - \mathbf{z})$ in the arguments. Also, note that $B \cap \Pi = \{0\}$ if and only if $Z \cap \Pi = \{0\}$, where Z is given by (4.5). We point out that if the hypotheses of Theorem 4.3 are satisfied, then the necessary conditions of Theorem 4.1 also hold with $Z_i =$ cone (interior of B_i) for $i \in \mathcal{I}_{\mathbf{z}}$.

Now consider the following variant of the canonical optimization problem, which we shall refer to as a *simple optimization problem.*

Given two sets Q' and W in a locally convex, linear topological space $\mathfrak{Z}$, and real-valued functions φ_i, $i = 0, -1, \cdots, -\mu$, defined on W, find an element $\mathbf{x} \in W \cap Q'$ that satisfies the inequalities $\varphi_{-i}(\mathbf{x}) \leqq 0$ for $i = 1, \cdots, \mu$, and which, in so doing, minimizes the value of φ_0.

The simple optimization problem differs from the canonical one in that there are no equality constraints in the former.

For $\mathbf{x} \in W$, we define the sets $\mathcal{I}_{\mathbf{x}}$ and $\mathcal{J}_{\mathbf{x}}$ as before. A local solution of a simple optimization problem is defined by an obvious modification of Definition 4.1.

Definition 4.6. A local solution $\mathbf{z}$ of a simple optimization problem is *regular* if $\mathbf{z} \in$ (interior of W) and conditions (4) and (5) of Definition 4.2 are satisfied.

Definition 4.7. A local solution $\mathbf{z}$ of a simple optimization problem is *totally regular* if $\mathbf{z} \in$ (interior of W) and condition (4) of Definition 4.2 as well as conditions (i), (iii), and (iv) of Definition 4.3 are satisfied.

Definition 4.8. A local solution $\mathbf{z}$ of a simple optimization problem is *smoothly regular* if, for every $i \in \mathcal{I}_{\mathbf{z}}$, (4.6) holds for some $c_i \in \mathfrak{Z}^*$, if $\mathbf{z} \in$ (interior of W), if condition (4) of Definition 4.2 is satisfied, and if the relations

$$\sum_{i \in \mathcal{I}_{\mathbf{z}}} \alpha_{-i} c_i = 0, \quad \alpha_{-i} \leqq 0 \quad \text{for} \quad \text{every} \quad i \in \mathcal{I}_{\mathbf{z}},$$

imply that $\alpha_{-i} = 0$ for every $i \in \mathcal{I}_{\mathbf{z}}$.

The following lemma follows in the same way as Lemmas 4.4 and 4.5.

Lemma 4.7. *Every smoothly regular local solution of a simple optimization problem is totally regular, and every totally regular local solution is regular, where the sets* Z_i *in condition* (5) *of Definition* 4.2 *are given by* (4.8).

If $\mathbf{z}$ is a regular local solution of a simple optimization problem, it follows

at once that $0 \in \mathfrak{Z}$ is a (Q, B)-extremal, where the sets Q and B are given by (4.11) and (4.4).

The following theorems can now be proved on the basis of Theorem 2.2 in the same way that Theorem 4.1 and 4.2 were proved.

THEOREM 4.4. *Let* $\mathbf{z}$ *be a regular local solution of a simple optimization problem, and let* K *be a first-order, convex approximation to* $Q' - \mathbf{z}$. *Then there exists a nonzero functional* $l^* \in \mathfrak{Z}^*$ *such that* $l^*(\mathbf{x}) \leqq 0 \leqq l^*(\mathbf{y})$ *for all* $\mathbf{x} \in \overline{\text{cone}}\, K$ *and* $\mathbf{y} \in Z$ *(where* Z *is given by* (4.5) *and the* Z_i *for* $i \in \mathcal{I}_\mathbf{z}$ *are as indicated in condition* (5) *of Definition* 4.2). *Further, if* $\mathfrak{Z}$ *is a Banach space, or if* $\mathcal{I}_\mathbf{z} = \{0\}$, *or if* Z_i *is given by* (4.8), *with* $c_i \in \mathfrak{Z}^*$, *for all but one (or all)* $i \in \mathcal{I}_\mathbf{z}$, *then* $l^* = \sum_{i \in \mathcal{I}_\mathbf{z}} l_{-i}$, *where, for each* $i \in \mathcal{I}_\mathbf{z}$, (4.16) *holds. If* Z_i *is given by* (4.8) *for some* $i \in \mathcal{I}_\mathbf{z}$, *where* $c_i \in \mathfrak{Z}^*$, *then* (4.16) *implies that* $l_{-i} = \alpha_{-i} c_i$ *where* $\alpha_{-i} \leqq 0$.

THEOREM 4.5. *Let* $\mathbf{z}$ *be a totally regular, or smoothly regular local solution of a simple optimization problem. Then if* K *is a first-order, convex approximation to* $Q' - \mathbf{z}$, *there exist nonpositive numbers* α_{-i}, *for* $i \in \mathcal{I}_\mathbf{z}$, *not all zero, such that* $\sum_{i \in \mathcal{I}_\mathbf{z}} \alpha_{-i} c_i(\mathbf{x}) \leqq 0$ *for all* $\mathbf{x} \in \overline{\text{cone}}\, K$. *If* $\mathbf{z}$ *is smoothly regular, then, in addition,* $\sum_{i \in \mathcal{I}_\mathbf{z}} \alpha_{-i} c_i \neq 0$.

Note 4.5. The remarks of Notes 4.3 and 4.4 are also pertinent for the simple optimization problem and Theorems 4.4 and 4.5.

DEFINITION 4.9. A simple optimization problem is *convex* if W is a convex set and the functionals $\varphi_0, \varphi_{-1}, \cdots, \varphi_{-\mu}$ are convex.

The following theorem follows in the same way as Theorem 4.3.

THEOREM 4.6. *Let* $\mathbf{z}$ *be a local solution of a convex simple optimization problem such that* $\mathbf{z} \in$ *(interior of* W*) and condition* (4) *of Definition* 4.2 *is satisfied, and suppose in addition that the functionals* φ_{-i}, *for* $i \in \mathcal{I}_\mathbf{z}$, *satisfy the hypotheses of Theorem* 4.3. *Then if* K *is a first-order, convex approximation to* $Q' - \mathbf{z}$, *there exist nonpositive real numbers* α_{-i}, *for* $i \in \mathcal{I}_\mathbf{z}$, *not all zero, such that*

$$\sum_{i \in \mathcal{I}_\mathbf{z}} \alpha_{-i} \varphi_{-i}(\mathbf{z} + \mathbf{x}) \leqq \alpha_0 \varphi_0(\mathbf{z}) \quad \textit{for all} \quad \mathbf{x} \in (\overline{\text{cone}}\, K) \cap (W - \mathbf{z}).$$

We point out that if the hypotheses of Theorem 4.6 are satisfied, then the necessary conditions of Theorem 4.4 also hold with $Z_i =$ cone (interior of B_i) for $i \in \mathcal{I}_\mathbf{z}$.

Note 4.6. The canonical optimization problem may be generalized to the following:

Let there be given locally convex, linear topological spaces $\mathfrak{Z}$, $\mathfrak{Z}'$, and $\mathfrak{Z}''$; sets Q' and W in $\mathfrak{Z}$; convex cones Z_2 and Z_3 in $\mathfrak{Z}'$ and $\mathfrak{Z}''$, respectively, with both Z_2 and Z_3 having vertex at 0 and a nonempty interior, and functions φ_1, φ_0, and φ_{-1} defined on W and taking on values in R^m, $\mathfrak{Z}'$, and $\mathfrak{Z}''$, respectively. Then find an element $\mathbf{x} \in W \cap Q'$ such that $\varphi_1(\mathbf{x}) = 0$,

$\varphi_{-1}(\mathbf{x}) \in Z_3$, and such that the relations $\mathbf{x}' \in W \cap Q'$, $\varphi_1(\mathbf{x}') = 0$, $\varphi_{-1}(\mathbf{x}') \in Z_3$, $[\varphi_0(\mathbf{x}') - \varphi_0(\mathbf{x})] \in Z_2$ imply that $\varphi_0(\mathbf{x}') = \varphi_0(\mathbf{x})$.

We can define a local solution of the generalized canonical optimization problem by means of an obvious modification of Definition 4.1. Such a local solution will be called regular if (1) $\mathbf{z}$ is an interior point of W and φ_1 is continuous is a neighborhood of $\mathbf{z}$, (2) relation (4.3) holds for $i = 1$ with l_1 a linear, continuous map from $\mathfrak{Z}$ onto R^m, and (3) there are cones Z_0 and Z_1 which are internal cones at 0 for the sets

$$B_0 = \{\mathbf{x} | \mathbf{z} + \mathbf{x} \in W, [\varphi_0(\mathbf{z} + \mathbf{x}) - \varphi_0(\mathbf{z})] \in Z_2, \varphi_0(\mathbf{z} + \mathbf{x}) \neq \varphi_0(\mathbf{z})\} \cup \{0\}$$

and

$$B_1 = \{\mathbf{x} | \mathbf{z} + \mathbf{x} \in W, [\varphi_{-1}(\mathbf{z} + \mathbf{x}) - \varphi_{-1}(\mathbf{z})] \in Z_3\} \cup \{0\},$$

respectively, such that $Z_1 \cap Z_0 \neq \mathfrak{Z}$ and $Z_1 \cap Z_0 \neq \{0\}$. A local solution will be called totally regular if the preceding hypotheses (1) and (2) are satisfied, if (4.6) holds for $i = 0$ and $i = 1$, where c_0 and c_1 are certain continuous functions from $\mathfrak{Z}$ to $\mathfrak{Z}'$ and $\mathfrak{Z}''$, respectively, such that $[c_i(\mathbf{x} + \mathbf{y}) - c_i(\mathbf{x}) - c_i(\mathbf{y})] \in Z_{2+i}$, $i = 0$ and 1, whenever $\mathbf{x}$ and $\mathbf{y} \in \mathfrak{Z}$, $c_0^{-1}(Z_2) \cap c_1^{-1}(Z_3) \neq \mathfrak{Z}$, and $[c_0^{-1}(\text{interior of } Z_2)] \cap [c_1^{-1}(\text{interior of } Z_3)]$ is not empty. Finally, a local solution will be called smoothly regular if it is totally regular with c_0 and c_1 linear. A generalized canonical optimization problem will be called convex if W is convex and if

$$[\varphi_{-i}(\alpha\mathbf{x} + \beta\mathbf{y}) - \alpha\varphi_{-i}(\mathbf{x}) - \beta\varphi_{-i}(\mathbf{y})] \in Z_{i+2}$$

for $i = 0$ and 1, all $\mathbf{x}$ and $\mathbf{y} \in W$, and all nonnegative numbers α and β such that $\alpha + \beta = 1$.

If $\mathbf{z}$ is a totally regular local solution, it is not difficult to show (the arguments are almost identical to those used in the proof of Lemma 4.2) that

$$Z_i = [c_i^{-1}(\text{interior of } Z_{i+2})] \cup \{0\}$$

is an internal cone at 0 for B_i, $i = 0$ or 1, as defined in condition (3), from which it follows at once that $\mathbf{z}$ is also a regular local solution. Further, if $\mathbf{z}$ is a regular local solution, it is easily seen that 0 is a (Q, B, F)-extremal, where $Q = Q' - \mathbf{z}$, $B = B_1 \cap B_0$, and $F(\mathbf{x}) = \varphi_1(\mathbf{z} + \mathbf{x})$. Then, if K is a first-order, convex approximation to Q, one can obtain, on the basis of Corollary 3.1, necessary conditions for solutions of the generalized canonical optimization problem which generalize those in Theorems 4.1 and 4.2. Finally, if $\mathbf{z}$ is a local solution of a convex generalized canonical optimization problem satisfying suitable hypotheses (in essence the same as those in Theorem 4.3), then $\mathbf{z}$ can be shown to satisfy necessary conditions which generalize those of Theorem 4.3.

Note 4.7. In an entirely analogous manner to that indicated in Note 4.6, it is possible to define a generalized simple optimization problem, and to obtain necessary conditions which generalize those of Theorems 4.4–4.6.

Note 4.8. The special case of the canonical or simple optimization problems where Q' is convex is of particular interest, and is usually referred to as a mathematical programming problem. (This name is most commonly applied to simple problems, i.e., to problems with no equality constraints.) If Q' is convex and $\mathbf{z} \in Q'$, then $Q' - \mathbf{z}$ is also convex, and is consequently a first-order convex approximation to itself (see Note 2.2). Thus, all of the necessary conditions in Theorems 4.1–4.6 hold (under the appropriate hypotheses) with $K = Q' - \mathbf{z}$.

The mathematical programming problem wherein $Q' = \mathfrak{Z}$ is an important special case which exhibits some particularly interesting features. If $\mathbf{z}$ is a regular local solution of such a problem which is canonical in form, $\mathbf{z}$ satisfies all of the necessary conditions of Theorem 4.1 with $K = \overline{\text{cone}}\, K = \mathfrak{Z}$. Then (4.12) and (4.14) imply that

$$\bar{l} = -\sum_{i=1}^{m} \alpha_i l_i \neq 0;$$

(4.15) and (4.16) also hold under the indicated additional hypotheses. If $\mathbf{z}$ is totally regular, then the necessary conditions of Theorem 4.2, with $K = \overline{\text{cone}}\, K = \mathfrak{Z}$, are also satisfied, and if, in addition $c_i \in \mathfrak{Z}^*$ for each $i \in \mathcal{I}_\mathbf{z}$, then (4.18) implies that

$$\sum_{i=1}^{m} \alpha_i l_i + \sum_{i \in \mathcal{I}_\mathbf{z}} \alpha_{-i} c_i = 0.$$

On the other hand, if the problem is simple in form, there can be no regular local solutions (otherwise, according to Theorem 4.4, there would be a nonzero functional $l^* \in \mathfrak{Z}^*$ such that $l^*(\mathbf{x}) \leqq 0$ for all $\mathbf{x} \in \mathfrak{Z}$, which is absurd). Thus, if $\mathbf{z}$ is a solution of such a problem such that $\mathbf{z} \in$ (interior of W), and such that conditions (4) and (5) of Definition 4.2—with the exception of the requirement that $Z \neq \{0\}$—are satisfied, then $Z = \bigcap_{i \in \mathcal{I}_\mathbf{z}} Z_i = \{0\}$. This means that for some $j \in \mathcal{I}_\mathbf{z}$ and some subset $\mathcal{I}_\mathbf{z}'$ of $\mathcal{I}_\mathbf{z}$, the cones Z_j and $(\bigcap_{i \in \mathcal{I}_\mathbf{z}'} Z_i)$ have a nonempty interior and have only 0 in common, and consequently are separable; i.e., there is a nonzero functional $l_{-j} \in \mathfrak{Z}^*$ such that $l_{-j}(\mathbf{x}) \leqq 0 \leqq l_{-j}(\mathbf{y})$ for all $\mathbf{x} \in \bigcap_{i \in \mathcal{I}_\mathbf{z}'} Z_i$ and $\mathbf{y} \in Z_j$. If $\mathfrak{Z}$ is a Banach space it then follows from Lemma 4.3 that there are functionals l_{-i} for every $i \in \mathcal{I}_\mathbf{z}$, not all zero, such that $\sum_{i \in \mathcal{I}_\mathbf{z}} l_{-i} = 0$ and (4.16) holds (the requirement that $\mathfrak{Z}$ is a Banach space can be replaced by other hypotheses as indicated in Theorem 4.1). If, in addition, conditions (i) and (iii) of Definition 4.3 are satisfied, we can show, arguing as in Theorem 4.2, that there are nonpositive numbers α_{-i}, not all zero, such

that $\sum_{i\in\mathcal{J}_Z} \alpha_{-i}c_i(\mathbf{x}) \leqq 0$ for all $\mathbf{x} \in \mathfrak{Z}$. Finally, if the programming problem is convex, then there are nonpositive numbers α_{-i}, not all zero, such that

$$\sum_{i\in\mathcal{J}_Z} \alpha_{-i}\varphi_{-i}(\mathbf{z}+\mathbf{x}) \leqq \alpha_0\varphi_0(\mathbf{z})$$

for all $\mathbf{x} \in (W - \mathbf{z})$.

Note 4.9. The programming problems described in Note 4.8 can, of course, be generalized as indicated in Notes 4.6 and 4.7.

The necessary conditions of Theorems 4.1–4.6, for solutions of the mathematical programming problems described in Notes 4.8 and 4.9, generalize the well-known Kuhn-Tucker conditions satisfied by solutions of mathematical programming problems in finite-dimensional spaces, as well as the Lagrange multiplier rule in the ordinary calculus.

5. Relation to earlier work. Variational problems in infinite-dimensional spaces were first formulated almost thirty years ago. Indeed, Goldstine, in 1937 [5], obtained a Lagrange multiplier rule valid in Banach spaces. Kuhn-Tucker conditions in Banach spaces were obtained by Hurwicz [6, p. 99, Theorems V.3.3.4 and V.3.3.5] for (what we referred to as) simple mathematical programming problems, generalized as indicated in Note 4.6 (indeed the definitions in Note 4.6 were motivated by Hurwicz's problem statement), but with the requirement of Fréchet differentiability or convexity for the constraint and minimizing functionals.

Simple convex mathematical programming problems (in our terminology) in Banach spaces were considered in [9]. The necessary conditions presented in [9, §2] are essentially included in our necessary conditions as discussed in Note 4.8.

The concept of an internal cone was introduced in [7, §2] by Dubovitskii and Milyutin under the name of "cone of forbidden variations". The idea of a convex differential defined by (4.6) is also found in [7, §6] in a slightly different, but equivalent form (with the term "uniform differentiability"). As will be seen in Part II of this article, and as was also shown in [7], these differentials arise when one wishes to find necessary conditions for solutions of optimal control problems with restricted phase coordinates, or of minimax control problems. Lemma 4.3 (in a slightly more general form) was stated in [7, Theorem 3.1] without proof; a proof was presented in [9] (see [9, Corollary 2] from which Lemma 4.3 follows at once).

Indeed, in [7] a problem very similar to our canonical optimization problem was considered. The problem in [7] was less general in that the underlying space $\mathfrak{Z}$ was a Banach space, and in that the set Q' was the entire space $\mathfrak{Z}$. It was more general in that the equality constraint neither had to be finite-dimensional nor had to be differentiable (in the sense that (4.3) had to hold). However, no counterpart of Theorems 2.1 and 2.2 was proved in

[7]. Roughly speaking, a theorem analogous to Theorem 4.1 for the case $Z \cap \Pi = \{0\}$ was obtained, under the assumption that the points in $\mathfrak{Z}$ which satisfy the equality constraints lie on a "surface" which has a tangent "cone" (analogous to our Π) at the solution point $\mathbf{z}$. General conditions under which such a surface and tangent cone exist were not given.

As will be seen in Part II, the set Q', and its first-order convex approximation K, make it possible to handle a very broad class of differential equation constraints in a very simple and natural manner. In [7], such constraints were considered to be equality-type constraints (analogous to our φ_i for $i > 0$) which turned out to be very difficult to take into account even in conventional optimal control problems.

Preliminary results on the subject matter of this paper were presented in [8].

6. Acknowledgments. The author would like to acknowledge many stimulating conversations with R. V. Gamkrelidze, which gave the impetus to the research described in this paper. Further, it is due to a suggestion of A. V. Balakrishnan that the author's vision was broadened from normed spaces to locally convex spaces. Finally, it was as a result of some discussions with E. Polak, C. D. Cullum, M. D. Canon, R. J. B. Wets, and R. Van Slyke that the author became aware that his results could be applied to mathematical programming problems.

REFERENCES

[1] N. Dunford and J. T. Schwartz, *Linear Operators, Part I. General Theory*, Interscience, New York, 1958.

[2] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[3] R. V. Gamkrelidze, *On some extremal problems in the theory of differential equations with applications to the theory of optimal control*, this Journal, 3(1965), pp. 106–128.

[4] E. J. McShane, *On multipliers for Lagrange problems*, Amer. J. Math., 61(1939), pp. 809–819.

[5] H. H. Goldstine, *A multiplier rule in abstract spaces*, Bull. Amer. Math. Soc., 44(1938), pp. 388–394.

[6] L. Hurwicz, *Programming in linear spaces*, Studies in Linear and Nonlinear Programming, K. J. Arrow, L. Hurwicz, and H. Uzawa, eds., Stanford University Press, Stanford, 1958, pp. 38–102.

[7] A. Ya. Dubovitskii and A. A. Milyutin, *Extremum problems in the presence of constraints*, Ž. Vyčisl. Mat. i Mat. Fiz., 5(1965), pp. 395–453.

[8] L. W. Neustadt, *Optimal control problems as extremal problems in a Banach space*, Proceedings of Polytechnic Institute of Brooklyn Symposium on System Theory, 1965, pp. 215–224.

[9] B. N. Pshenichniy, *Convex programming in a normed space*, Kibernetika, 1(1965), No. 5, pp. 46–54.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

CONSTRAINED MINIMIZATION PROBLEMS IN FINITE-DIMENSIONAL SPACES*

M. CANON, C. CULLUM, AND E. POLAK†

Introduction. The entire approach to constrained minimization problems in finite-dimensional spaces, as found in the field of optimal control, is substantially different from the approach to these problems found in mathematical programming. Furthermore, within each of these fields, one finds a diversity of methods and points of view. The purpose of this paper is to exhibit a unified approach to constrained minimization problems in finite-dimensional spaces and to show that most of the known necessary conditions for optimality are straightforward consequences of a fairly simple, but all-encompassing theorem.

Section 1 is devoted to formulating the Basic Problem, i.e., the form into which most of the known finite-dimensional constrained minimization problems can be transcribed. A necessary condition for the optimality of a solution to this Basic Problem is then derived by a geometric method, first used by McShane [1] in the calculus of variations and subsequently greatly popularized by Pontryagin, Boltyanskii, Gamkrelidze, and Mishchenko [2] in their derivation of the maximum principle. The necessary condition for the Basic Problem is stated as an inequality which must hold for all the elements in a cone which is a suitable linearization of the constraint set. The wide range of applicability of this theorem is substantially due to the fact that one has a great deal of freedom in choosing this linearization cone.

Section 2 is devoted to transcribing a wide variety of minimization problems into the form of the Basic Problem, to rederiving many classical necessary conditions, and to obtaining several new ones. In particular, it is shown that classical Lagrange multiplier theory, the results of Fritz John [3], Kuhn and Tucker [4], and Mangasarian and Fromovitz [5], in nonlinear programming theory, and the results of Jordan and Polak [6], Halkin [7], and Holtzman [8], in discrete optimal control theory, can all be obtained from the necessary condition for the Basic Problem. In addition, several new results are obtained for bounded state space, discrete optimal control problems. Presently known necessary conditions for certain bounded state space problems, such as those obtained by Rosen [9], can

* Received by the editors February 9, 1966, and in revised form May 10, 1966.

† Department of Electrical Engineering, University of California, Berkeley, California. This research was supported by the National Science Foundation under Grant GK-569 and by the National Aeronautics and Space Administration under Grant NsG-354 (supp. 2).

be seen to be special cases of the more general results presented in this paper.

It is the author's hope that the unified approach to constrained minimization problems in E^n, presented in this paper, will facilitate the mastery of the subject and will lead to a deeper and more fruitful understanding of minimization problems in general.

1. The Basic Problem.

1.1. Statement of the Basic Problem. Let $f: E^n \to E^1$ and $r: E^n \to E^m$ be continuously differentiable functions, and let $\Omega \subset E^n$ be a subset of E^n. The *Basic Problem* can be stated as follows:

Find a vector $\hat{z} \in E^n$ *such that*

(i) $\hat{z} \in \Omega$, $r(\hat{z}) = 0$,
(ii) *for all* $z \in \Omega$ *with* $r(z) = 0$, $f(\hat{z}) \leqq f(z)$.

We shall call a vector $\hat{z}$ satisfying (i) and (ii) an *optimal solution* to the Basic Problem.

1.2. Necessary condition for optimality. The necessary condition to be derived will be stated in the form of an inequality which is valid for all $\delta z = (\delta z^1, \delta z^2, \cdots, \delta z^n)$ in a convex cone "approximation" or "linearization" of the set Ω. We shall make use of two kinds of "linearizations" of the set Ω at a point z. The first one will be defined after a review of needed terminology and notation; the second one will be defined after the proof of Theorem 1, to obtain an extension.

A set C is a cone with vertex x_0 if for every $x \in C$, $x \neq x_0$, $x_0 + \lambda(x - x_0) \in C$ for all $\lambda > 0$. Since the vertex x_0 of the cone C will normally be obvious, we shall omit mentioning it. The notation co $\{z, z + \epsilon\delta z^1, \cdots, z + \epsilon\delta z^k\}$ denotes the convex hull of $z, z + \epsilon\delta z^1, \cdots, z + \epsilon\delta z^k$, i.e., the set of all points y of the form

$$y = \mu_0 z + \mu_1(z + \epsilon\delta z^1) + \cdots + \mu_k(z + \epsilon\delta z^k),$$

where $\sum_{i=0}^{k} \mu_i = 1$, $\mu_i \geqq 0$ for all i.

DEFINITION. A convex cone $C(z, \Omega) \subset E^n$ will be called a *linearization of the first kind* of the constraint set Ω at z if for any finite collection $\{\delta z^1, \delta z^2, \cdots, \delta z^k\}$ of linearly independent vectors in $C(z, \Omega)$ there exists an $\epsilon > 0$, possibly depending on $z, \delta z^1, \delta z^2, \cdots, \delta z^k$, such that co $\{z, z + \epsilon\delta z^1, \cdots, z + \epsilon\delta z^k\} \subset \Omega$.

If the cone $C(z, \Omega)$ is a linearization of the first kind, then for every $\delta z \in C(z, \Omega)$ there exists an $\epsilon_1 > 0$ such that $z + \epsilon\delta z \in \Omega$ for all ϵ such that $0 \leqq \epsilon \leqq \epsilon_1$. The largest cone having this property is given a special name.

DEFINITION. The *radial cone* to the set Ω at a point $z \in \Omega$ will be denoted by $RC(z, \Omega)$ and is defined by

$$RC(z, \Omega) = \{\delta z: \text{ there exists an } \epsilon_1(z, \delta z) > 0 \text{ such that } z + \epsilon\delta z \in \Omega \text{ whenever } 0 \leqq \epsilon \leqq \epsilon_1\}.$$

Whenever the radial cone $RC(\hat{z}, \Omega)$ is a linearization of the first kind, it contains all the other linearizations of the first kind of the set Ω at $\hat{z}$. Consequently, in the various theorems to follow, the radial cone $RC(\hat{z}, \Omega)$ should always be used if possible, since this will result in stronger necessary conditions.

Next, we define the $C^{(1)}$ map $F: E^n \to E^{m+1}$ by

$$F(z) = (f(z), r(z)).$$

We shall number the components of E^{m+1} from 0 to m, i.e., $y \in E^{m+1}$ is given by $y = (y^0, y^1, \cdots, y^m)$. The Jacobian matrix $(\partial F^i(z)/\partial z^j)$ of the map $F(z)$ will be denoted by $\partial F(z)/\partial z$.

For the Basic Problem stated above, the following theorem gives a necessary condition for optimality.

THEOREM 1. *If $\hat{z}$ is an optimal solution to the Basic Problem, and $C(\hat{z}, \Omega)$ is a linearization of the first kind of Ω at $\hat{z}$, then there exists a nonzero vector $\psi = (\psi^0, \psi^1, \cdots, \psi^m) \in E^{m+1}$, with $\psi^0 \leqq 0$, such that for all $\delta z \in \overline{C(\hat{z}, \Omega)}$ (the closure of $C(\hat{z}, \Omega)$ in E^n),*

$$\langle \psi, \frac{\partial F(\hat{z})}{\partial z} \delta z \rangle \leqq 0. \tag{1}$$

Proof. Let $K(\hat{z}) \subset E^{m+1}$ be the cone defined by

$$K(\hat{z}) = \frac{\partial F(\hat{z})}{\partial z} C(\hat{z}, \Omega). \tag{2}$$

$K(\hat{z})$ is convex because $C(\hat{z}, \Omega)$ is convex and $\partial F(\hat{z})/\partial z$ is a linear map. Let $\hat{y} = F(\hat{z})$. We shall now show that the cone $K(\hat{z})$ must be separated from the ray

$$R = \{y: y = \beta(-1, 0, \cdots, 0), \beta \geqq 0\}, \tag{3}$$

i.e., that there must exist a nonzero vector $\psi \in E^{m+1}$ such that

$$\begin{aligned} &\text{(i)} && \langle \psi, y \rangle \leqq 0 \quad \text{for every} \quad y \in K(\hat{z}), \\ &\text{(ii)} && \langle \psi, y \rangle \geqq 0 \quad \text{for every} \quad y \in R. \end{aligned} \tag{4}$$

Suppose that the cone $K(\hat{z})$ and the ray R are not separated. Then the cone $K(\hat{z})$ must be of dimension $m + 1$ and R must be an interior ray of $K(\hat{z})$ (i.e., all points of R except the origin are interior points of $K(\hat{z})$).

Let us now construct in the cone $K(\hat{z})$ a simplex Σ with vertices 0, δy^1, δy^2, $\cdots$, δy^{m+1} such that

(i) there exists a point on R, $\delta y^0 = \gamma(-1, 0, \cdots, 0)$ with $\gamma > 0$, which lies in the interior of Σ,

(ii) there exists a set of vectors $\delta z^i \in C(\hat{z}, \Omega)$ satisfying

$$\delta y^i = \frac{\partial F(\hat{z})}{\partial z}\,\delta z^i, \qquad i = 1, \cdots, m+1, \tag{5}$$

and such that

$$\text{co}\,\{\hat{z}, \hat{z} + \delta z^1, \cdots, \hat{z} + \delta z^{m+1}\} \subset \Omega. \tag{6}$$

It is possible to satisfy (i) because R is an interior ray of the $(m+1)$-dimensional cone $K(\hat{z})$, and it is possible to satisfy (ii) because $C(\hat{z}, \Omega)$ is a linearization of the first kind. Note that the vectors δz^i, $i = 1, \cdots, m+1$, are linearly independent since the vectors $\delta y^1, \delta y^2, \cdots, \delta y^{m+1}$ are linearly independent.

Since δy^0 is an interior point of Σ, there is a number $r > 0$ such that the sphere of radius r with center at δy^0 is contained in Σ. For $0 < \alpha \leqq 1$, let S_α be the sphere of radius αr with center at $\alpha\delta y^0$. Clearly $S_\alpha \subset \Sigma$ whenever $0 < \alpha \leqq 1$. For each fixed α, $0 < \alpha \leqq 1$, we now define the map G_α from $S_\alpha - \{\alpha\delta y^0\}$ into E^{m+1} as follows. For any $x \in S_\alpha - \{\alpha\delta y^0\}$, let

$$G_\alpha(x) = F(\hat{z} + ZY^{-1}(\alpha\delta y^0 + x)) - (\hat{y} + \alpha\delta y^0), \tag{7}$$

where Y is an $(m+1) \times (m+1)$ matrix whose ith column is δy^i, $i = 1, \cdots, m+1$, and Z is an $n \times (m+1)$ matrix whose ith column is δz^i. The matrix Y is invertible because the δy^i form a basis for E^{m+1} by construction.

Expanding the right-hand side of (7) about $\hat{z}$, we get

$$G_\alpha(x) = \hat{y} + \frac{\partial F(\hat{z})}{\partial z}\,ZY^{-1}(\alpha\delta y^0 + x) - (\hat{y} + \alpha\delta y^0) + o(ZY^{-1}(\alpha\delta y^0 + x)), \tag{8}$$

where $o(\cdot)$ is a continuous function such that $\lim_{\|y\|\to 0} \|o(y)\|/\|y\| = 0$. By definition, $(\partial F(\hat{z})/\partial z)Z = Y$, and hence (8) simplifies to

$$G_\alpha(x) = x + o(ZY^{-1}(\alpha\delta y^0 + x)). \tag{9}$$

Now, for $x \in \partial(S_\alpha - \{\alpha\delta y^0\})$ (the boundary of the sphere), $\|x\| = \alpha r$, and we may write $x = \alpha\rho_1$, where $\|\rho_1\| = r$. Hence for $x \in \partial(S_\alpha - \{\alpha\delta y^0\})$,

$$G_\alpha(\alpha\rho_1) = \alpha\rho_1 + o(\alpha ZY^{-1}(\delta y^0 + \rho_1)). \tag{10}$$

By definition of $o(\cdot)$, there exists an α^*, $0 < \alpha^* \leqq 1$, such that for all $\rho_1 \in E^{m+1}$, with $\|\rho_1\| = r$,

$$\|o(\alpha^* ZY^{-1}(\delta y^0 + \rho_1))\| < \alpha^* r. \tag{11}$$

We now conclude from Brouwer's fixed point theorem (see Appendix) that there exists an $\tilde{x} \in S_{\alpha^*} - \{\alpha^* \delta y^0\}$ such that

$$G_{\alpha^*}(\tilde{x}) = 0, \tag{12}$$

i.e.,

$$F(\hat{z} + ZY^{-1}(\alpha^* \delta y^0 + \tilde{x})) = \hat{y} + \alpha^* \delta y^0. \tag{13}$$

Now $\hat{y} + \alpha^* \delta y^0 = \text{col}\,(f(\hat{z}) - \alpha^* \gamma, 0, 0, \cdots, 0)$, where $\gamma > 0$. Thus, expanding (13),

$$r(\hat{z} + ZY^{-1}(\alpha^* \delta y^0 + \tilde{x})) = 0 \tag{14}$$

and

$$f(\hat{z} + ZY^{-1}(\alpha^* \delta y^0 + \tilde{x})) = f(\hat{z}) - \alpha^* \gamma < f(\hat{z}). \tag{15}$$

Furthermore, because of (6) and the fact that for any δy in the simplex Σ, the vector $z = \hat{z} + ZY^{-1}\delta y$ belongs to co $\{\hat{z}, \hat{z} + \delta z^1, \cdots, \hat{z} + \delta z^{m+1}\}$,

$$\hat{z} + ZY^{-1}(\alpha^* \delta y^0 + \tilde{x}) \in \Omega. \tag{16}$$

Hence $\hat{z}$ is not optimal, which is a contradiction. We therefore conclude that the cone $K(\hat{z})$ and the ray R must be separated, i.e., there must exist a nonzero vector $\psi \in E^{m+1}$ such that

$$\langle \psi, y \rangle \leqq 0 \quad \text{for every} \quad y \in K(\hat{z}) \tag{17}$$

and

$$\langle \psi, y \rangle \geqq 0 \quad \text{for every} \quad y \in R. \tag{18}$$

Substituting (2) in (17), we have

$$\langle \psi, \frac{\partial F(\hat{z})}{\partial z} \delta z \rangle \leqq 0 \quad \text{for every} \quad \delta z \in C(\hat{z}, \Omega). \tag{19}$$

Clearly, (19) must also hold for every $\delta z \in \overline{C(\hat{z}, \Omega)}$. Substituting for y from (3) into (18), we have

$$\langle \psi, (-1, 0, \cdots, 0) \rangle = -\psi^0 \geqq 0. \tag{20}$$

This completes the proof.

It has been pointed out by Neustadt [10] that Theorem 4 remains valid under the relaxed assumption that $C(\hat{z}, \Omega)$ is a linearization of the second kind of Ω at $\hat{z}$, defined as follows.

DEFINITION. A convex cone $C(z, \Omega) \subset E^n$ will be called a *linearization of the second kind* of the constraint set Ω at z, if, for any finite collection $\{\delta z^1, \delta z^2, \cdots, \delta z^k\}$ of linearly independent vectors in $C(z, \Omega)$, there exists an $\epsilon > 0$, possibly depending on $z, \delta z^1, \cdots, \delta z^k$, and a continuous map ζ from

co $\{z, z + \epsilon\delta z^1, \cdots, z + \epsilon\delta z^k\}$ into Ω, such that $\zeta(z + \delta z) = z + \delta z + o(\delta z)$, where $\lim_{\|\delta z\| \to 0} \| o(\delta z)\| / \| \delta z \| = 0$.

Remark. We observe that if $C(z, \Omega)$ is a linearization of the first kind of Ω at z, then it is also a linearization of the second kind of Ω at z, with the map ζ being the identity. Thus, unless we have specific cause to indicate whether a cone $C(z, \Omega)$ is a linearization of the first or second kind, we shall refer to it simply as a *linearization* of Ω at z. We now restate Theorem 1 in this form.

THEOREM 1′. *If $\hat{z}$ is an optimal solution to the Basic Problem and $C(\hat{z}, \Omega)$ is a linearization of Ω at $\hat{z}$, then there exists a nonzero vector $\psi = (\psi^0, \psi^1, \cdots, \psi^m) \in E^{m+1}$ with $\psi^0 \leqq 0$, such that for all $\delta z \in \overline{C(\hat{z}, \Omega)}$, (the closure of $C(\hat{z}, \Omega)$ in E^n),*

$$\langle \psi, \frac{\partial F(\hat{z})}{\partial z} \delta z \rangle \leqq 0.$$

The reader may easily modify the proof of Theorem 1 so as to apply to Theorem 1′. Finally, it should be pointed out that all conditions such as continuity, differentiability, etc., imposed on the various functions need only hold in a neighborhood of the optimal point.

2. Applications. We shall now show how a number of classical optimization problems can be cast in the form of the Basic Problem, and we shall then apply Theorem 1 or Theorem 1′ to rederive several classical conditions for optimality, as well as to obtain some new ones.

2.1. Classical theory of Lagrange multipliers. The classical constrained minimization problem admits equality constraints only. Thus, it is the Basic Problem with $\Omega = E^n$, the entire space. Clearly, E^n is a linearization of the first kind for E^n at any point $z \in E^n$.

Thus, we conclude from Theorem 1 that if $\hat{z}$ is an optimal solution of the Basic Problem, with $\Omega = E^n$, then there exists a nonzero vector $\psi \in E^{m+1}$ such that

$$\langle \psi, \frac{\partial F(\hat{z})}{\partial z} \delta z \rangle \leqq 0 \quad \text{for all} \quad \delta z \in E^n. \tag{21}$$

This may be rewritten as

$$\left\langle \left(\frac{\partial F(\hat{z})}{\partial z} \right)^T \psi, \delta z \right\rangle \leqq 0 \quad \text{for all} \quad \delta z \in E^n. \tag{22}$$

Since for any $\delta z \in E^n$, $-\delta z$ is also in E^n, we conclude from (22) that

$$\left(\frac{\partial F(\hat{z})}{\partial z} \right)^T \psi = 0. \tag{23}$$

Now, $(\partial F(\hat{z})/\partial z)^T$ is an $n \times (m+1)$ matrix with columns $\nabla f(\hat{z})$, $\nabla r^1(\hat{z}), \cdots \nabla r^m(\hat{z})$, where

$$\nabla f(\hat{z}) = \left(\frac{\partial f(\hat{z})}{\partial z^1}, \cdots, \frac{\partial f(\hat{z})}{\partial z^n}\right), \quad \nabla r^i(\hat{z}) = \left(\frac{\partial r^i(\hat{z})}{\partial z^1}, \cdots, \frac{\partial r^i(\hat{z})}{\partial z^n}\right).$$

We may therefore expand (23) into the form

$$\psi^0 \nabla f(\hat{z}) + \sum_{i=1}^{m} \psi^i \nabla r^i(\hat{z}) = 0. \tag{24}$$

We have thus reproved the following classical result.

THEOREM 2. *Let $f, r^1, r^2, \cdots, r^m$ be real valued, continuously differentiable functions on E^n. If $\hat{z} \in E^n$ minimizes $f(z)$ subject to the constraints $r^i(z) = 0$, $i = 1, 2, \cdots, m$, then there exist scalar multipliers, $\psi^0, \psi^1, \cdots, \psi^m$, not all zero, such that the function H on E^n, which they define by*

$$H(z) = \psi^0 f(z) + \sum_{i=1}^{m} \psi^i r^i(z), \tag{25}$$

has a stationary point at $z = \hat{z}$, i.e., (24) is satisfied.

It is usual to assume that the gradient vectors $\nabla r^i(z)$, $i = 1, 2, \cdots, m$, are linearly independent for all z such that $r(z) = 0$. This precludes $\sum_{i=1}^{m} \psi^i \nabla r^i(\hat{z}) = 0$ and hence in (24), $\psi^0 \neq 0$. Multiplying (24) by $1/\psi^0$ and letting $\hat{\lambda}^i = \psi^i/\psi^0$, $i = 1, 2, \cdots, m$, we now deduce the more commonly seen condition.

THEOREM 2′. *Let $f, r^1, r^2, \cdots, r^m$ be real valued, continuously differentiable functions on E^n. If $\hat{z}$ minimizes $f(z)$ subject to $r^i(z) = 0$ for $i = 1, 2, \cdots, m$, and the gradient vectors $\nabla r^i(\hat{z})$, with $i = 1, 2, \cdots, m$, are linearly independent, then there exists a vector $\hat{\lambda} \in E^m$ such that the real valued Langrangian L on $E^n \times E^m$, defined by*

$$L(z, \lambda) = f(z) + \sum_{i=1}^{m} \lambda^i r^i(z) \tag{26}$$

has a stationary point at $(\hat{z}, \hat{\lambda})$.

We note that by (24), $\partial L(\hat{z}, \hat{\lambda})/\partial z = 0$ and that $\partial L(\hat{z}, \hat{\lambda})/\partial \lambda = r(\hat{z}) = 0$, by assumption.

2.2. Nonlinear programming. Let $f: E^n \to E^1$, $r: E^n \to E^m$, and $q: E^n \to E^k$ be continuously differentiable functions. The standard nonlinear programming problem is that of minimizing $f(z)$ subject to the constraints that $r(z) = 0$ and $q(z) \leqq 0$.

This corresponds to the special case of the Basic Problem, with $\Omega = \{z: q(z) \leqq 0\}$. We shall now show how Theorem 1 can be used to obtain various commonly known necessary conditions for $\hat{z}$ to be optimal. The

presentation is divided into two parts. It should be noted that the necessary conditions obtained in Part I are stronger than those obtained in Part II.

Given a particular point $z \in \Omega$, we shall often have occasion to divide the components of the inequality constraints functions, q^i, $i = 1, \cdots, k$, into two sets; those for which $q^i(z) = 0$ and those for which $q^i(z) < 0$. To simplify notation we introduce the following definition.

DEFINITION. For $z \in \Omega$, let the index set $I(z)$ be defined by

$$I(z) = \{i : q^i(z) = 0\}. \tag{27}$$

The constraints q^i, $i \in I(z)$, will be called the *active constraints* at z. We shall denote by $I(z)^c$ the complement of $I(z)$ in $\{1, \cdots, k\}$.

Part I. The set $\Omega = \{z : q(z) \leqq 0\}$ introduced above is assumed to satisfy the following condition:

ASSUMPTION (A1).[1] Let $\hat{z} \in \Omega$ be an optimal solution of the nonlinear programming problem. Then there exists a vector $h \in E^n$ such that

$$\langle \nabla q^i(\hat{z}), h \rangle < 0 \quad \text{for all} \quad i \in I(\hat{z}).$$

A sufficient condition for (A1) to be satisfied is that the vectors $\nabla q^i(\hat{z})$, $i \in I(\hat{z})$, be linearly independent (see Corollary to Lemma 3).

DEFINITION. For any $z \in \Omega$, the *internal cone* of Ω at z, denoted by $IC(z, \Omega)$, is defined by

$$IC(z, \Omega) = \{\delta z : \langle \nabla q^i(z), \delta z \rangle < 0 \quad \text{for all} \quad i \in I(z)\}.$$

By Assumption (A1), the convex cone $IC(\hat{z}, \Omega)$ is nonempty. It is a simple exercise in the use of Taylor's theorem to prove the following lemma.

LEMMA 1. *If* $IC(z, \Omega) \neq \varnothing$, *the empty set, then*

(i) $IC(z, \Omega)$ *is a linearization of the first kind of* Ω *at* z,

(ii) $\overline{IC(z, \Omega)} = \{\delta z : \langle \nabla q^i(z), \delta z \rangle \leqq 0$ *for all* $i \in I(z)\}$.

When specialized to the nonlinear programming problem, Theorem 1 assumes the following form.

THEOREM 3. *If* $\hat{z}$ *is an optimal solution to the nonlinear programming problem, with* (A1) *satisfied, then there exists a nonzero vector* $\psi \in E^{m+1}$, *with* $\psi^0 \leqq 0$, *such that for all*

$$\delta z \in \overline{IC(\hat{z}, \Omega)} = \{\delta z : \langle \nabla q^i(\hat{z}), \delta z \rangle \leqq 0 \quad \text{for all} \quad i \in I(\hat{z})\},$$

$$\langle \frac{\partial H(\hat{z})}{\partial z}, \delta z \rangle \leqq 0,$$

where $H(z) = \psi^0 f(z) + \sum_{i=1}^{m} \psi^i r^i(z)$.

[1] When some of the functions q^i, $i \in I(z)$, are linear, it suffices to require that there exist a vector $h \in E^n$ such that $\langle \nabla q^i(z), h \rangle \leqq 0$ for these functions and $\langle \nabla q^i(z), h \rangle < 0$ for the remaining functions q^i, $i \in I(z)$.

Using Theorem 3 and Farkas' lemma (see [16]) we obtain the following necessary condition for optimality, which is in a form more familiar to specialists in mathematical programming.

THEOREM 4. *If $\hat{z}$ is an optimal solution to the nonlinear programming problem, with* (A1) *satisfied, then there exist a nonzero vector $\psi \in E^{m+1}$, with $\psi^0 \leqq 0$, and a vector $\mu \in E^k$, with $\mu \leqq 0$, such that*

$$\text{(i)} \qquad \psi^0 \nabla f(\hat{z}) + \sum_{i=1}^{m} \psi^i \nabla r^i(\hat{z}) + \sum_{i=1}^{k} \mu^i \nabla q^i(\hat{z}) = 0$$

and

$$\text{(ii)} \qquad \sum_{i=1}^{k} \mu^i q^i(\hat{z}) = 0.$$

Proof. From Theorem 3,

$$\langle \frac{\partial H(\hat{z})}{\partial z}, \delta z \rangle \leqq 0$$

for all δz such that $\langle \nabla q^i(\hat{z}), \delta z \rangle \leqq 0$, $i \in I(\hat{z})$. By Farkas' lemma, there exist scalars $\mu^i \leqq 0$, $i \in I(\hat{z})$, such that

$$\frac{\partial H(\hat{z})}{\partial z} + \sum_{i \in I(\hat{z})} \mu^i \nabla q^i(\hat{z}) = 0.$$

Let $\mu^i = 0$ for $i \in I(\hat{z})^c$. This completes the proof.

Most of the other well-known necessary conditions for nonlinear programming problems can be obtained from Theorem 4 by making additional assumptions on the functions r and q. For example, the following corollaries to Theorem 4 are immediate consequences of that theorem.

COROLLARY 1. *If Assumption* (A1) *is satisfied and the vectors $\nabla r^i(\hat{z})$, $i = 1, \cdots, m$, are linearly independent, then any vector $\psi \in E^{m+1}$, $\mu \in E^k$, which satisfy the conditions of Theorem* 4 *are such that $\langle \psi^0, \mu \rangle \neq 0$.*

COROLLARY 2. *If $\nabla r^i(\hat{z})$, $i = 1, \cdots, m$, together with $\nabla q^i(\hat{z})$, $i \in I(\hat{z})$, are linearly independent vectors, then any vector $\psi \in E^{m+1}$ satisfying the conditions of Theorem* 4 *also satisfies $\psi^0 < 0$.*

The assumption in Corollary 2 is a well-known [11] sufficient condition for the Kuhn-Tucker constraint qualification to be satisfied. When it is added to Theorem 4 we obtain a slightly restricted form[2] of the Kuhn-Tucker theorem [4].

COROLLARY 3. *If there exists a vector $h \in E^n$ such that $\langle \nabla q^i(\hat{z}), h \rangle < 0$ for all $i \in I(\hat{z})$, $\langle \nabla r^i(\hat{z}), h \rangle = 0$ for $i = 1, \cdots, m$, and the vectors $\nabla r^i(\hat{z})$,*

[2] In practice, the Kuhn-Tucker constraint conditions can rarely be shown to be satisfied unless the restrictions imposed in Corollaries 2 and 3 hold.

$i = 1, \cdots, m$, *are linearly independent, then any vector* $\psi \in E^{m+1}$ *satisfying the conditions of Theorem* 4 *also satisfies* $\psi^0 < 0$.

Proof. Assume $\psi^0 = 0$. Taking the scalar product of both sides of (i) in Theorem 4 with the hypothesized vector h, one concludes that $\mu = 0$ and hence $\psi = 0$, a contradiction.

The assumption in this corollary is a sufficient condition for the weakened constraint qualification [13] to be satisfied. Augmented by this assumption, Theorem 4 becomes a slightly restricted form of the Kuhn-Tucker theorem with the weakened constraint qualification.

Part II. We shall now derive a necessary condition for the nonlinear programming problem which is not based on Assumption (A1) and hence is weaker than the necessary condition stated in Theorem 4. This condition was first proved by Mangasarian and Fromovitz [5] using the implicit function theorem and a lemma by Motzkin [12].

Whenever Assumption (A1) is not satisfied, it is possible to show that the vectors $\nabla q^i(\hat{z})$, $i \in I(\hat{z})$, can be summed to zero with nonpositive scalars. This is established in the following lemma.

LEMMA 2. *Suppose that Assumption* (A1) *is not satisfied for the set* $\Omega = \{z : q(z) \leqq 0\}$. *Then there exists a nonzero vector* $\mu \in E^k$, *with* $\mu \leqq 0$, *such that*

$$\text{(i)} \qquad \sum_{i=1}^{k} \mu^i \nabla q^i(\hat{z}) = 0,$$

$$\text{(ii)} \qquad \sum_{i=1}^{k} \mu^i q^i(\hat{z}) = 0.$$

Proof. Consider the linear subspace of E^α,

$$L = \{v : v = (\langle h, \nabla q^{i_1}(\hat{z})\rangle, \cdots, \langle h, \nabla q^{i_\alpha}(\hat{z})\rangle) \quad \text{with} \quad h \in E^n \quad \text{and where} \quad \{i_1, i_2, \cdots, i_\alpha\} = I(\hat{z})\}.$$

By hypothesis L has no rays in common with the convex cone

$$C = \{v : v = (v^1, \cdots, v^\alpha), \quad \text{with} \quad v^j < 0 \quad \text{for} \quad j = 1, \cdots, \alpha\}.$$

Hence L can be separated from $\bar{C}$, i.e., there exists a nonzero vector $\beta \in E^\alpha$ such that

$$\text{(i)} \qquad \langle \beta, v\rangle \geqq 0 \quad \text{for all} \quad v \in \bar{C},$$

and

$$\text{(ii)} \qquad \langle \beta, v\rangle \leqq 0 \quad \text{for all} \quad v \in L.$$

It is obvious from (i) that $\beta \leqq 0$. Since L is a linear subspace, $\langle \beta, v\rangle = 0$ for all $v \in L$, which implies that

$$\langle \sum_{j=1}^{\alpha} \beta^j \nabla q^{i_j}(\hat{z}), h \rangle = 0 \quad \text{for all} \quad h \in E^n,$$

and therefore

$$\sum_{j=1}^{\alpha} \beta^j \nabla q^{i_j}(\hat{z}) = 0.$$

If we now let $\mu^i = \beta^j$ when $i = i_j \in I(\hat{z})$, and $\mu^i = 0$ when $i \in I(\hat{z})^c$, then $\mu = (\mu^1, \cdots, \mu^k)$ is the desired vector.

COROLLARY. *A sufficient condition for the Assumption* (A1) *to be satisfied is that the vectors* $\nabla q^i(\hat{z})$, $i \in I(\hat{z})$, *be linearly independent.*

Lemma 2 may be combined with Theorem 4 to give a necessary condition for optimality which does not require that (A1) be satisfied. For this, the most general case of the nonlinear programming problem, we obtain the following necessary condition for optimality.

THEOREM 5. *If* $\hat{z}$ *is an optimal solution to the nonlinear programming problem, then there exist a vector* $\psi \in E^{m+1}$ *and a vector* $\mu \in E^k$, *with* $\psi^0 \leqq 0$ *and* $\mu \leqq 0$, ψ *and* μ *not both zero, such that*

$$\text{(i)} \qquad \psi^0 \nabla f(\hat{z}) + \sum_{i=1}^{m} \psi^i \nabla r^i(\hat{z}) + \sum_{i=1}^{k} \mu^i \nabla q^i(\hat{z}) = 0$$

and

$$\text{(ii)} \qquad \sum_{i=1}^{k} \mu^i q^i(\hat{z}) = 0.$$

Proof. If (A1) is satisfied, Theorem 5 is a slightly weaker statement of Theorem 4. If (A1) is not satisfied, let μ be the vector specified in Lemma 3, and let $\psi = 0$.

Finally, we note that if we let $r \equiv 0$, Theorem 5 becomes the well-known Fritz John necessary condition for optimality [3].

We have thus shown that most of the known necessary conditions for nonlinear programming problems, previously derived by diverse and often unrelated techniques, can now be obtained simply by applying Theorem 1 and Farkas' lemma.

2.3. Optimal control. In the field of optimal control of discrete time systems, necessary conditions for optimality have been developed by Jordan and Polak [6], Halkin [7], Holtzman [8], and Rosen [9]. By recasting the optimal control problem in the form of the Basic Problem, it is possible to obtain from Theorem 1 and 1′ essentially all of the above mentioned results in a unified manner. Furthermore, the derivation given in this paper is significantly simpler in most cases. In addition, Theorems 1 and 1′ together with Farkas' lemma yield necessary conditions for opti-

mality for a class of bounded state space problems, a result which is new with this paper.

The general *optimal control problem* that we will consider takes the following form:

Given a system described by the difference equation

$$x_{i+1} - x_i = f_i(x_i, u_i), \quad x_i \in E^n, \quad u_i \in E^m, \quad i = 0, \cdots, k-1, \tag{28}$$

find a *control sequence* $(u_0, u_1, \cdots, u_{k-1})$ and a *corresponding trajectory* $(x_0, x_1, \cdots, x_k)$ such that

(i) $u_i \in U_i \subset E^m$ for $i = 0, \cdots, k-1$ (control constraints),

(ii) $x \in \Omega_i = \{x : q_i(x) \leqq 0\}$, $q_i : E^n \to E^{n_i}$, $i = 0, \cdots, k$ (state space constraints),

(29) and, in addition, the initial and terminal states, x_0 and x_k, satisfy

(iii) $g_0(x_0) = 0$, $g_0 : E^n \to E^{l_0}$ (initial manifold constraint),

(iv) $g_k(x_k) = 0$, $g_k : E^n \to E^{l_k}$ (terminal manifold constraint),

and such that $\sum_{i=0}^{k-1} f_i^0(x_i, u_i)$ is minimized.

We make the following assumptions on the various sets and functions appearing above.

ASSUMPTIONS.

(a) $f_i : E^n \times E^m \to E^n$ is a $C^{(1)}$ function for $i = 0, \cdots, k-1$.

(b) For every $u_i \in U_i$, and for all $i = 0, \cdots, k-1$, the radial cone $RC(u_i, U_i)$ is a linearization of the first kind for U_i at u_i.

(30) (c) g_0 and g_k are $C^{(1)}$ functions whose Jacobian matrices have rank l_0 and l_k, respectively.

(d) For all $x \in \Omega_i$, $i = 0, \cdots, k$, the gradients of the active constraints, $\nabla q_i^j(x)$, $j \in I(x)$, (see (27)) are linearly independent vectors.

(e) $f_i^0 : E^n \times E^m \to E^1$ is a $C^{(1)}$ function for $i = 0, \cdots, k-1$.

This problem may be reformulated in the form of the Basic Problem, i.e., $\{\min f(z) : r(z) = 0, z \in \Omega\}$, by making the following identifications. Let $z = (x_0, x_1, \cdots, x_k, u_0, \cdots, u_{k-1}) \in E^{(k+1)n+km}$; and let f, r, and Ω be defined by

(i) $f(z) = \sum_{i=0}^{k-1} f_i^0(x_i, u_i)$,

$$(31)\quad \text{(ii)}\ r(z) = \begin{bmatrix} x_1 - x_0 - f_0(x_0, u_0) \\ \vdots \\ x_k - x_{k-1} - f_{k-1}(x_{k-1}, u_{k-1}) \\ g_0(x_0) \\ g_k(x_k) \end{bmatrix},$$

$$\text{(iii)}\quad \Omega = \Omega_0 \times \Omega_1 \times \cdots \times \Omega_k \times U_0 \times U_1 \times \cdots \times U_{k-1}.$$

Clearly f and r have the required differentiability properties. The cone

$$(32)\quad C(z, \Omega) = IC(x_0, \Omega_0) \times \cdots \times IC(x_k, \Omega_k) \times RC(u_0, U_0) \times \cdots \times RC(u_{k-1}, U_{k-1}),$$

where $IC(x_i, \Omega_i)$ and $RC(u_i, U_i)$ were defined earlier, is obviously a linearization of the first kind for Ω at z since assumption (d) and Lemma 3 guarantee that $IC(x_i, \Omega_i)$ is nonempty for every $i = 0, \cdots, k$, and by Lemma 1 it is a linearization of the first kind for Ω_i at x_i, while $RC(u_i, U_i)$, for $i = 0, \cdots, k-1$, is a linearization of the first kind by assumption (b). Therefore, we may apply Theorem 1, from which we conclude that if $\hat{z}$ is an optimal solution to the optimal control problem, then there exists a nonzero vector $\psi = (p^0, \pi)$, with $p^0 \leqq 0$ and $\pi = (-p_1, \cdots, -p_k, \mu_0, \mu_k)$, where $p_i \in E^n$, $\mu_0 \in E^{l_0}$, $\mu_k \in E^{l_k}$, such that

$$(33)\quad \langle p^0 \frac{\partial f(\hat{z})}{\partial z}, \delta z\rangle + \langle \pi, \frac{\partial r(\hat{z})}{\partial z}\rangle \delta z \leqq 0$$

for all $\delta z \in \overline{C(\hat{z}, \Omega)}$. Substituting for f and r in (33) and expanding, we get

$$(34)\quad \begin{aligned} & p^0 \left(\sum_{i=0}^{k-1} \frac{\partial f_i{}^0(\hat{x}_i, \hat{u}_i)}{\partial x} \delta x_i + \sum_{i=0}^{k-1} \frac{\partial f_i{}^0(\hat{x}_i, \hat{u}_i)}{\partial u} \delta u_i \right) \\ & + \sum_{i=0}^{k-1} \langle -p_{i+1}, \delta x_{i+1} - \delta x_i - \frac{\partial f_i(\hat{x}_i, \hat{u}_i)}{\partial x} \delta x_i - \frac{\partial f_i(\hat{x}_i, \hat{u}_i)}{\partial u} \delta u_i \rangle \\ & + \langle \mu_0, \frac{\partial g_0(\hat{x}_0)}{\partial x} \delta x_0 \rangle + \langle \mu_k, \frac{\partial g_k(\hat{x}_k)}{\partial x} \delta x_k \rangle \leqq 0 \end{aligned}$$

for every $\delta x = (\delta x_0, \cdots, \delta x_k, \delta u_0, \cdots, \delta u_{k-1}) \in \overline{C(\hat{z}, \Omega)}$.

The usual form of the necessary conditions in terms of a Hamiltonian, adjoint equation, transversality conditions, etc., are obtained by considering special forms of δz. The conditions obtainable by this procedure are summarized in Theorem 6 below.

THEOREM 6. *If $\hat{z} = (\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_k, \hat{u}_0, \cdots, \hat{u}_{k-1})$ is an optimal solution to the optimal control problem, then there exist vectors $p_0, p_1, \cdots, p_k$ in*

E^n, λ_0, λ_1, $\cdots$, λ_k, $\lambda_i \in E^{n_i}$, *with* $\lambda_i \leqq 0$, $\mu_0 \in E^{l_0}$, $\mu_k \in E^{l_k}$, *and a scalar* $p^0 \leqq 0$, *such that*

(i) *not all of the quantities* p^0, p_0, p_1, $\cdots$, p_k, *are zero*;

(ii) $$p_i - p_{i+1} = \left[\frac{\partial f_i(\hat{x}_i, \hat{u}_i)}{\partial x}\right]^T p_{i+1} + \left[\frac{\partial f_i^0(\hat{x}_i, \hat{u}_i)}{\partial x}\right]^T p^0 + \left[\frac{\partial q_i(\hat{x}_i)}{\partial x}\right]^T \lambda_i, \qquad i = 0, 1, 2, \cdots, k-1;$$

(iii) $$p_k = \left[\frac{\partial g_k(\hat{x}_k)}{\partial x}\right]^T \mu_k + \left[\frac{\partial q_k(\hat{x}_k)}{\partial x}\right]^T \lambda_k;$$

(iv) $$p_0 = -\left[\frac{\partial g_0(\hat{x}_0)}{\partial x}\right]^T \mu_0;$$

(v) $$\langle \lambda_i, q_i(\hat{x}_i)\rangle = 0, \qquad i = 0, \cdots, k;$$

(vi) $$\left\langle \left[\frac{\partial f_i^0(\hat{x}_i, \hat{u}_i)}{\partial u}\right]^T p^0 + \left[\frac{\partial f_i(\hat{x}_i, \hat{u}_i)}{\partial u}\right]^T p_{i+1}, \delta u \right\rangle \leqq 0$$

for all $\delta u \in \overline{RC(\hat{u}_i, U_i)}$ *and all* $i = 0, 1, \cdots, k-1$.

To prove all of the above conditions would be somewhat laborious. Therefore, we will only derive (vi) to demonstrate how one proceeds.

Let $\delta z = (0, \cdots, 0, \delta u_i, 0, \cdots, 0)$ with $\delta u_i \in \overline{RC(\hat{u}_i, U_i)}$. Clearly $\delta z \in \overline{C(\hat{z}, \Omega)}$, and (34) reduces to

$$p^0 \frac{\partial f_i^0(\hat{x}_i, \hat{u}_i)}{\partial u} \delta u_i + \langle p_{i+1}, \frac{\partial f_i(\hat{x}_i, \hat{u}_i)}{\partial u} \delta u_i \rangle \leqq 0.$$

Simple rearrangement yields (vi).

It should be remarked at this point that the derivation of conditions (ii), (iii), and (v) requires the use of Farkas' lemma (see [16]), while (iv) is simply a definition.

To the authors' knowledge the above, quite general, necessary conditions have not been obtained previously, although Rosen [9] did obtain a similar result under substantially more restrictive assumptions on the sets U_i.

In the special case where there are no state space constraints, i.e., $q_i \equiv 0$ for $i = 0, \cdots, k$, Theorem 6 reduces to the following.

COROLLARY. *If the functions* $q_i \equiv 0$ *for* $i = 0, \cdots, k$, *and* $\hat{z}$ *is an optimal solution to the optimal control problem, then there exist vectors* $p_0, \cdots, p_k$ *in* E^n, $\mu_0 \in E^{l_0}$, $\mu_k \in E^{l_k}$, *and a scalar* $p^0 \leqq 0$, *such that*

(i) *not all of the quantities* p^0, $p_0, \cdots, p_k$ *are zero*;

(ii) $$p_i - p_{i+1} = \left[\frac{\partial f_i(\hat{x}_i, \hat{u}_i)}{\partial x}\right]^T p_{i+1} + \left[\frac{\partial f_i^0(\hat{x}_i, \hat{u}_i)}{\partial x}\right]^T p^0,$$

$$i = 0, \cdots, k-1;$$

(iii) $$p_k = \left[\frac{\partial g_k(\hat{x}_k)}{\partial x}\right]^T \mu_k;$$

(iv) $$p_0 = -\left[\frac{\partial g_0(\hat{x}_0)}{\partial x}\right]^T \mu_0;$$

(v) $$\left\langle \left[\frac{\partial f_i^0(\hat{x}_i, \hat{u}_i)}{\partial u}\right]^T p^0 + \left[\frac{\partial f_i(\hat{x}_i, \hat{u}_i)}{\partial u}\right]^T p_{i+1}, \delta u \right\rangle \leqq 0$$

for all $\delta u \in \overline{RC(\hat{u}_i, U_i)}$ *and all* $i = 0, 1, \cdots, k-1$.

This is the condition derived by Jordan and Polak [6].

A maximum principle. Halkin [7] and Holtzman [8] have shown that by making some additional assumptions, condition (v) in the above corollary may be replaced by a stronger condition, which is usually called a maximum principle. Both Halkin's and Holtzman's results can be obtained from Theorem 1′, but, for simplicity, we shall only show how Halkin's results are obtained.

The optimal control problem considered by Halkin differs from the optimal control problem stated at the beginning of this section in the following way:

(35)

(i) There are no state space constraints other than the initial and terminal manifold constraints, i.e., $q_i \equiv 0$ for $i = 0, 1, \cdots, k$.

(ii) Assumptions (a) and (e) for the optimal control problem are replaced by the following, respectively:

(a′) For every $u_i \in U_i$, the functions $f_i(\cdot, u_i)$, $i = 0, \cdots, k-1$, are continuously differentiable on E^n.

(e′) For every $u_i \in U_i$, the functions $f_i^0(\cdot, u_i)$, are continuously differentiable functions on E^n.

(iii) Assumption (b) is replaced by the following:

(b′) For every $x \in E^n$ and every $i = 0, 1, \cdots, k-1$, the sets $\mathbf{f}_i(x, U_i)$ are convex, where $\mathbf{f}_i: E^n \times E^m \to E^{n+1}$ is defined by $\mathbf{f}_i(x, u) = [f_i^0(x, u), f_i(x, u)]$.

The reformulation of the Halkin problem as a Basic Problem differs only slightly from that used for the optimal control problem. First, we introduce new variables $\mathbf{v}_i = (v_i^0, v_i) \in E^{n+1}$ with $v_i = (v_i^1, v_i^2, \cdots, v_i^n) \in E^n$ where $i = 0, \cdots, k-1$. Then we let $z = (x_0, x_1, \cdots, x_k, \mathbf{v}_0, \cdots, \mathbf{v}_{k-1}) \in E^{(k+1)n+k(n+1)}$, and we define the functions f and r and the set Ω by

$$\text{(i)}\quad f(z) = \sum_{i=0}^{k-1} v_i^0,$$

$$\text{(ii)}\quad r(z) = \begin{bmatrix} x_1 - x_0 - v_0 \\ \vdots \\ x_k - x_{k-1} - v_{k-1} \\ g_0(x_0) \\ g_k(x_k) \end{bmatrix}, \tag{36}$$

$$\text{(iii)}\quad \Omega = \{z = (x_0, \cdots, x_k, \mathbf{v}_0, \cdots, \mathbf{v}_{k-1}) : x_i \in E^n,$$
$$\mathbf{v}_i \in \mathbf{f}_i(x_i, U_i) \text{ for all } i = 0, 1, \cdots, k-1\}.$$

Let $\hat{z} = (\hat{x}_0, \cdots, \hat{x}_k, \hat{\mathbf{v}}_0, \cdots, \hat{\mathbf{v}}_{k-1})$, where $\hat{\mathbf{v}}_i = \mathbf{f}_i(\hat{x}_i, \hat{u}_i)$, $\hat{u}_i \in U_i$, be the optimal solution. For the linearization of the set Ω at $\hat{z}$, we take the cone

$$C(\hat{z}, \Omega) = \Big\{\delta z : \delta z = (\delta x_0, \cdots, \delta x_k, \delta \mathbf{v}_0, \cdots, \delta \mathbf{v}_{k-1}), \quad \text{with} \quad \delta x_i \in E^n$$
$$\text{and} \left(\delta \mathbf{v}_i - \frac{\partial \mathbf{f}_i(\hat{x}_i, \hat{u}_i)}{\partial x}\, \delta x_i\right) \in RC(\hat{\mathbf{v}}_i, \mathbf{f}_i(\hat{x}_i, U_i)) \tag{37}$$
$$\text{for every} \quad i = 0, \cdots, k-1\Big\}.$$

Clearly $C(\hat{z}, \Omega)$ is a convex cone. We shall now show that $C(\hat{z}, \Omega)$ is a linearization of the second kind of Ω at $\hat{z}$.

Let $\delta z^1, \cdots, \delta z^r$ be any finite collection of linearly independent vectors in $C(\hat{z}, \Omega)$, with $\delta z^i = (\delta x_0{}^i, \cdots, \delta x_k{}^i, \delta \mathbf{v}_0{}^i, \cdots, \delta \mathbf{v}_{k-1}^i)$. For each $i = 1, \cdots, r$, and for each $j = 0, \cdots, k-1$, there exists an $\epsilon_j{}^i > 0$ such that

$$\hat{\mathbf{v}}_j + \epsilon_j{}^i \left(\delta \mathbf{v}_j{}^i - \frac{\partial \mathbf{f}_j(\hat{x}_j, \hat{u}_j)}{\partial x}\, \delta x_j{}^i\right) \in \mathbf{f}_j(\hat{x}_j, U_j)\ .$$

As a consequence, there exist an $\epsilon > 0$ and vectors $u_j{}^i \in U_j$ such that

$$\hat{\mathbf{v}}_j + \epsilon \delta \mathbf{v}_j{}^i = \epsilon \frac{\partial \mathbf{f}_j(\hat{x}_j, \hat{u}_j)}{\partial x}\, \delta x_j{}^i + \mathbf{f}_j(\hat{x}_j, u_j{}^i) \tag{38}$$

for every $i = 1, \cdots, r$, and $j = 0, \cdots, k-1$. Let $C_0 = \text{co}\,\{\hat{z}, \hat{z} + \epsilon\delta z^1, \cdots, \hat{z} + \epsilon\delta z^r\}$. Let $z \in C_0$ be arbitrary, and let $\delta z = z - \hat{z}$. Then we may write

$$\delta z = \sum_{i=1}^{r} \mu^i \epsilon \delta z^i, \quad \text{where} \quad \mu^i \geqq 0, \ \sum_{i=1}^{r} \mu^i \leqq 1;$$

or $\delta z = Z\mu$, where $Z = (\epsilon\delta z^1, \epsilon\delta z^2, \cdots, \epsilon\delta z^r)$ is a matrix with columns $\epsilon\delta z^i$, and $\mu = (\mu^1, \cdots, \mu^r)$ is an r-vector. For every $z \in C_0$, the vector μ is

uniquely determined by the expression $\mu = Y\delta z$, where Y is a matrix whose rows y_i, $i = 1, \cdots, r$, satisfy $\langle y_i, \epsilon\delta z^j \rangle = \delta_{ij}$, the Kronecker delta, for $i, j = 1, \cdots, r$.

The map $\zeta: C_0 \to \Omega$ is defined as follows. For every $z = (x_0, \cdots, x_k, \mathbf{v}_0, \cdots, \mathbf{v}_{k-1}) \in C_0$ and corresponding $\delta z = (\delta x_0, \cdots, \delta x_k, \delta\mathbf{v}_0, \cdots, \delta\mathbf{v}_{k-1}) = z - \hat{z}$, let $\zeta(z) = (x_0, \cdots, x_k, \mathbf{w}_0, \cdots, \mathbf{w}_{k-1})$ with

$$\mathbf{w}_j = \mathbf{f}_j(x_j, \hat{u}_j) + \sum_{i=1}^{r} \mu^i(\delta z)[\mathbf{f}_j(x_j, u_j{}^i) - \mathbf{f}_j(x_j, \hat{u}_j)], \quad j = 0, \cdots, k-1, \tag{39}$$

where $\mu(\delta z) = (\mu^1(\delta z), \cdots, \mu^r(\delta z)) = Y\delta z$, and $u_j{}^i$, $i = 1, \cdots, r$, $j = 0, \cdots, k-1$, are defined in (38). The range of $\zeta(z)$ is contained in Ω because of the convexity of $\mathbf{f}_i(x_i, U_i)$. Since it is clear that $\zeta(z)$ is continuously differentiable, the reader may verify that $\zeta(z)$ is the identity map plus $o(z - \hat{z})$, as required in the definition of a linearization of the second kind, by expanding $\zeta(z)$ about $\hat{z}$.

Theorem 1′ may now be applied to this problem to obtain the usual separation results, i.e., if $\hat{z} = (\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_k, \hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{k-1})$ is an optimal solution to the Halkin problem, then there exists a nonzero vector $\psi = (p^0, \pi)$ with $p^0 \leqq 0$ and $\pi = (-p_1, \cdots, -p_k, \mu_0, \mu_k)$, where $p_i \in E^n$, for each $i = 0, 1, \cdots, k$, $\mu_0 \in E^{l_0}$, $\mu_k \in E^{l_k}$, such that

$$p_0 \sum_{i=0}^{k-1} \delta v_i{}^0 + \sum_{i=0}^{k-1} \langle -p_{i+1}, \delta x_{i+1} - \delta x_i - \delta v_i \rangle + \langle \mu_0, \frac{\partial g_0(\hat{x}_0)}{\partial x} \delta x_0 \rangle + \langle \mu_k, \frac{\partial g_k(\hat{x}_k)}{\partial x} \delta x_k \rangle \leqq 0 \tag{40}$$

for all $\delta z = (\delta x_0, \cdots, \delta x_k, \delta\mathbf{v}_0, \cdots, \delta\mathbf{v}_{k-1}) \in C(\hat{z}, \Omega)$. By taking appropriate perturbations we can obtain Halkin's necessary conditions [7].

THEOREM 7. *If* $(\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{k-1})$ *is an optimal control sequence and* $(\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_k)$ *is a corresponding optimal trajectory for the Halkin problem, then there exist vectors* $p_0, \cdots, p_k \in E^n$, $\mu_0 \in E^{l_0}$, $\mu_k \in E^{l_k}$, *and a scalar* $p^0 \leqq 0$, *such that*

(i) *not all of the quantities* $p^0, p_0, \cdots, p_k$, *are zero*;

(ii) $p_i - p_{i+1} = \left[\frac{\partial f_i(\hat{x}_i, \hat{u}_i)}{\partial x}\right]^T p_{i+1} + \left[\frac{\partial f_i{}^0(\hat{x}_i, \hat{u}_i)}{\partial x}\right]^T p^0, \quad i = 0, \cdots, k-1$;

(iii) $p_k = \left[\frac{\partial g_k(\hat{x}_k)}{\partial x}\right]^T \mu_k$;

(iv) $p_0 = -\left[\frac{\partial g_0(\hat{x}_0)}{\partial x}\right]^T \mu_0$;

(v) $p^0 f_i^0(\hat{x}_i, u) + \langle p_{i+1}, f_i(\hat{x}_i, u)\rangle \leqq p^0 f_i^0(\hat{x}_i, \hat{u}_i) + \langle p_{i+1}, f_i(\hat{x}_i, \hat{u}_i)\rangle$ *for all* $u \in U_i$ *and all* $i = 0, 1, \cdots, k-1$.

The reader can obtain most of the results by straightforward substitution of appropriate perturbations, δz, in (40). We shall prove only (v).

Let $\delta z = (0, \cdots, \delta \mathbf{v}_i, \cdots, 0)$, with $\delta \mathbf{v}_i \in RC(\hat{\mathbf{v}}_i, \mathbf{f}_i(\hat{x}_i, U_i))$. This is certainly an admissible perturbation, and, for such δz, (40) reduces to

(41) $p^0 \delta v_i^0 + \langle p_{i+1}, \delta v_i\rangle \leqq 0$ for all $\delta \mathbf{v}_i \in \overline{RC(\hat{\mathbf{v}}_i, \mathbf{f}_i(\hat{x}_i, U_i))}$.

Since $\mathbf{f}_i(\hat{x}_i, U_i)$ is a convex set, the vector $\mathbf{f}_i(\hat{x}_i, u) - \mathbf{f}_i(\hat{x}_i, \hat{u}_i)$ belongs to $RC(\hat{\mathbf{v}}_i, \mathbf{f}_i(\hat{x}_i, U_i))$ for every $u \in U_i$. Therefore, from (41), we get

$$p^0[f_i^0(\hat{x}_i, u) - f_i^0(\hat{x}_i, \hat{u}_i)] + \langle p_{i+1}, f_i(\hat{x}_i, u) - f_i(\hat{x}_i, \hat{u}_i)\rangle \leqq 0$$

for every $u \in U_i$. Condition (v) follows immediately.

Holtzman obtained exactly the same result as Halkin, (i.e., Theorem 7), with (iii b′) replaced by the less restrictive assumption that the sets $\mathbf{f}_i(x, U_i)$ are only directionally convex (see [8]). The derivation of this result from Theorem 1′ proceeds in essentially the same manner as the derivation of Theorem 7 above.

Remark. It has already been pointed out that Theorem 7 differs from the corollary to Theorem 6 only in the condition (v). In fact, using the method outlined above, a maximum principle can be derived in the presence of state space constraints of the type considered in (29 ii), provided all the other assumptions of Halkin or Holtzman are satisfied. One then gets a theorem identical to Theorem 6 except that condition (vi) is replaced by the maximum principle, i.e., condition (v) of Theorem 7. Theorem 7 then becomes a corollary to this more general result.

Conclusion. We have shown that a wide class of constrained minimization problems can be reduced to a common canonical form, the so-called Basic Problem, for which we have derived necessary conditions of optimality. It is rather clear that the present paper does not exhaust all the possible permutations and combinations of necessary conditions or minimization problems that can be treated by reduction to the Basic Problem. To name but a few, not discussed herein explicitly, we can point out optimal control problems with nonseparable constraints, such as total energy, total fuel, or else involving products of trajectory and control variables, which can also be reduced to the Basic Problem. However, one gets for these problems a necessary condition which applies to the entire trajectory and which does not necessarily break down into a series of conditions applicable at each sampling instant. One can also consider optimal control or nonlinear programming problems in which the "trajectory" constraint

sets are specified in more general form than equalities or inequalities. The necessary conditions derived in this paper can be suitably modified to cover such cases, yielding transversality conditions in terms of polar cones rather than in terms of gradient vectors.

Although nothing has been said in this paper about sufficient conditions, it is clear that under assumptions such as convexity, it is possible to show that some of the necessary conditions given here are also sufficient.

Finally, it should be pointed out that the general approach presented in this paper is the result of hindsight, an irritation with fragmentation, and the authors' conviction that in terms of problem solving, the geometric approach taken has great conceptual and intuitive advantages.

Appendix. The Brouwer fixed point theorem. In proving Theorem 1′ the authors have used a modified version of the Brouwer fixed point theorem. The conventional form of the theorem, which is stated and proved in [14], is worded as follows.

BROUWER FIXED POINT THEOREM. *If f is a continuous map from the unit sphere in E^n into the unit sphere in E^n, then f has a fixed point.*

The version used in this paper is stated without proof by Dieudonné [15]. Since the proof is very short, it is included here.

THEOREM. *If f is a continuous map from the unit sphere in E^n into E^n with $f(x) = x + g(x)$, where $\| g(x) \| \leqq 1$ for all x with $\| x \| = 1$, then the origin is contained in the range of f.*

Proof. To say that the origin is contained in the range of f is equivalent to saying that the function $h(x) = -g(x)$ has a fixed point. Let us define the function h_1 by

$$h_1(x) = \begin{cases} -g(x) & \text{if } \| g(x) \| \leqq 1, \\ -g(x)/\| g(x) \| & \text{if } \| g(x) \| > 1. \end{cases}$$

Clearly, h_1 is a continuous function from the unit sphere in E^n into the unit sphere in E^n. Therefore, by the Brouwer fixed point theorem, h_1 has a fixed point, say x_1. If $\| h_1(x_1) \| < 1$, then $h_1(x_1) = -g(x_1)$, and x_1 is a fixed point of $-g$. Suppose $\| h_1(x_1) \| = 1$. Then $\| x_1 \| = 1$ and consequently $\| g(x_1) \| \leqq 1$. Again $h_1(x_1) = -g(x_1)$ and x_1 is a fixed point of $-g$.

The Brouwer fixed point theorem follows immediately from this theorem, so they are in fact equivalent.

Acknowledgment. The authors wish to thank L. W. Neustadt for his critical perusal of this paper and for his comments.

REFERENCES

[1] E. J. MCSHANE, *On multipliers for Lagrange problems*, Amer. J. Math., 61 (1939), pp. 809–819.

[2] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[3] F. JOHN, *Extremum problems with inequalities as side conditions*, Studies and Essays, Courant Anniversary Volume, K. O. Friedrichs, O. E. Neugebauer and J. J. Stoker, eds., Interscience, New York, 1948, pp. 187–204.

[4] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1951, pp. 481–492.

[5] O. L. MANGASARIAN AND S. FROMOVITZ, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, Shell Development Company Paper P1433, 1965.

[6] B. W. JORDAN AND E. POLAK, *Theory of a class of discrete optimal control systems*, J. Electronics Control, 17 (1964), pp. 697–713.

[7] H. HALKIN, *A maximum principle of the Pontryagin type for systems described by nonlinear difference equations*, this Journal, 4 (1966), pp. 90–111.

[8] J. M. HOLTZMAN, *On the maximum principle for nonlinear discrete-time systems*, IEEE Trans. Automatic Control, to appear.

[9] J. B. ROSEN, *Optimal control and convex programming*, MRC Tech. Report 547, Mathematics Research Center, University of Wisconsin, Madison, 1965.

[10] L. W. NEUSTADT, personal communication.

[11] K. ARROW AND L. HURWICZ, *Reduction of constrained maxima to saddle-point problems*, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1956.

[12] T. S. MOTZKIN, *Two consequences of the transportation theorem of linear inequalities*, Econometrica, 19 (1951), pp. 184–185.

[13] K. ARROW, L. HURWICZ AND H. UZAWA, *Constraint qualifications in maximization problems*, Naval Res. Logist. Quart., 8 (1961) pp. 175–191.

[14] W. HUREWICZ AND H. WALLMAN, *Dimension Theory*, Princeton University Press, Princeton, 1948, pp. 40–41.

[15] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1960, p. 269.

[16] C. BERGE, *Topological Spaces*, Oliver and Boyd, Edinburgh and London, 1963, p. 164.

J. SIAM CONTROL
Vol. 4, No. 3, 1966
Printed in U.S.A.

LINEAR OPTIMAL SYSTEMS WITH TIME DELAYS*

D. H. CHYUNG AND E. BRUCE LEE†

Introduction. In the study of economic, biological, and physiological systems, as well as electromechanical systems composed of subsystems interconnected by hydraulic, mechanical and various other linkages [1], [2], we encounter phenomena which cannot be readily modeled unless relations involving time delays are admitted. In many of these systems, the interest is not just in a model for describing the evolution, but in selecting parameters and situations within the system to obtain the best evolution, an optimum system. Such parameters will be called controllers and we develop methods for selecting the best controllers subject to various limitations for systems involving time delays. The results are limited to systems describable by linear differential equations involving time delays, control parameters, and a variety of integral type criterion for optimality.

We consider linear controlled delay systems of the form

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t).$$

Various remarks and a few results for the more general delay systems (systems with nonlinearities, time varying delays, etc.) are stated in [3].

The problem of optimum control, as considered here for the above system, is to select a measurable controller $u(t)$ from some restraint set $\Omega \subset R^r$ to steer the response $x(t)$ from the initial continuous function $\phi(t)$, $-h_m \leqq t \leqq 0$, to a target set $G \subset R^n$ minimizing the real cost functional

$$C(u) = \int_0^{t_1} f(t, x(t), x(t - h_1), \cdots, x(t - h_m), u(t))\, dt.$$

R^k is the k-dimensional real number space.

This problem has been considered in [6], [13], [14], where certain necessary conditions were obtained, and in [11], where time optimal problems ($f \equiv 1$) for linear systems of the above form were considered. Also, see [5] for a discussion of a different class of control problems of the delay type, namely the class of hereditary processes with control. The above system is different in many respects from the class of hereditary systems whose evolution does not require a whole initial function. The necessary conditions of [5] and [6]

* Received by the editors February 28, 1966.

† Center for Control Sciences, University of Minnesota, Minneapolis, Minnesota. This research was sponsored by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under Grant AF-AFOSR-571-64.

are the same as those obtained here for our problem. Because of the special form for our problem, sufficiency results are also obtained and we present a complete theory. First we consider control problems where f is essentially quadratic in the system state $x(t)$ and controller $u(t)$. Both necessary and sufficient conditions are obtained by consideration of geometric properties of the set of attainability. The questions of existence of the optimal control and its uniqueness are also dealt with. The other cost functionals considered involve certain convexity hypotheses. An example and summary of other results are given.

Before dealing with the special control problem, we present certain needed preliminaries on linear differential equations with time delays. The notation of [9] will be used when applicable.

Consider the linear controlled system with time delays

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t) + v(t),$$

with continuous initial function $x(t) = \phi(t)$ on $[t_0 - h_m , t_0]$, where

$0 = h_0 < h_1 < \cdots < h_m < \infty$ are real constants,
$x(t)$, the system state, is an n-vector,
$u(t)$, the controller, is a measurable r-vector on $[t_0 , t_1]$,
$v(t)$ is an n-vector,
$A_i(t)$ is an $n \times n$ real continuous matrix for each i,
$B(t)$ is an $n \times r$ real continuous matrix.

It is well-known [1] that the above system has a unique continuous solution for $t \geqq t_0$ and the response can be written as

$$x(t) = x(t, \phi) + \int_{t_0}^{t} Y(s, t)[B(s)u(s) + v(s)]\, ds.$$

Here, $x(t, \phi)$ is the solution of the homogeneous equation

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i),$$

with the initial function $\phi(t)$ on $[t_0 - h_m , t_0]$; and $Y(s, t)$ is the fundamental solution of the adjoint equation defined by

$$\frac{\partial}{\partial s} Y(s, t) = -\sum_{i=0}^{m} Y(s + h_i , t)A_i(s + h_i), \qquad t_0 \leqq s \leqq t - h_m ,$$

$$\frac{\partial}{\partial s} Y(s, t) = -\sum_{i=0}^{k} Y(s + h_i, t)A_i(s + h_i),$$

$$t - h_{k+1} \leqq s \leqq t - h_k, \quad k = 0, 1, \cdots , (m - 1),$$

with endpoint condition $Y(t, t) = I$, where I is the $n \times n$ identity matrix. It is also well-known that such a $Y(s, t)$ exists and is continuous for s and t on $t_0 \leqq s \leqq t$, $t_0 \leqq t \leqq t_1$.

1. Integral quadratic cost functionals. Consider the system

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t)$$

on $[t_0, T]$ with continuous initial function $\phi(t)$ as above. The controller $u(t) \in R^r$ is measurable on $[t_0, T]$. Let the cost functional of control be

$$C(u) = g(x(T)) + \int_{t_0}^{T} \{x(s)'W(s)x(s) + u(s)'U(s)u(s)\}\, ds,$$

where $g(x)$ is a real continuous function in R^n, and $W(s)$ and $U(s)$ are real symmetric continuous $n \times n$ and $r \times r$ matrices, respectively, on $t_0 \leqq s \leqq T$. We also assume that $W(s)$ is positive semidefinite, and $U(s)$ is positive definite, i.e., $W(s)' = W(s)$, $U(s)' = U(s)$, and

$$\| x \|_W^2 \equiv x'W(s)x \geqq 0 \quad \text{for all} \quad x \in R^n,$$
$$\| u \|_U^2 \equiv u'U(s)u > 0 \quad \text{for all} \quad u \neq 0.$$

A controller $u(s)$ on $[t_0, T]$ is admissible if and only if $u(s) \in L_2[t_0, T]$, the space of square integrable functions, i.e.,

$$\int_{t_0}^{T} u(s)'u(s)\, ds = \int_{t_0}^{T} \| u(s)\|_I^2\, ds < \infty.$$

Then obviously $u(s)$ is integrable on $[t_0, T]$, so the system has a unique continuous solution $x(t)$ on $[t_0, T]$. Also, since $W(s)$ and $U(s)$ are continuous, $u(s) \in L_2[t_0, T]$ and $x(t)$ is continuous, for any admissible control $u(s)$ we have

$$\int_{t_0}^{T} (\| x(s)\|_W^2 + \| u(s)\|_U^2)\, ds < \infty$$

and

$$C(u) = g(x(T)) + \int_{t_0}^{T} (\| x(s)\|_W^2 + \| u(s)\|_U^2)\, ds < \infty.$$

The problem is to minimize the cost functional $C(u)$ while steering the response to a target set $G \subset R^n$. G may be the entire space $G = R^n$, the free endpoint problem.

Define the new variable $x_u^0(t)$ by

$$\dot{x}_u^0(t) = x_u(t)'W(t)x_u(t) + u(t)'U(t)u(t), \quad x_u^0(t_0) = 0,$$

where $x_u(t)$ is the response of the original system corresponding to the controller $u(t)$. Let[1]

$$\hat{x}_u(t) = (x_u{}^0(t), x_u(t)) \in R^{n+1}.$$

DEFINITION. Consider the system

$$\dot{x}^0(t) = x(t)'W(t)x(t) + u(t)'U(t)u(t),$$

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

with initial function $\phi(t)$, $t_0 - h_m \leqq t \leqq t_0$, and $x^0(t_0) = 0$. The set of attainability $\hat{K}(T)$ is the set of all response endpoints $\hat{x}_u(T) = (x_u{}^0(T), x_u(T)) \in R^{n+1}$ for all admissible controllers $u(t) \in L_2[t_0, T]$.

Properties of the set of attainability are now established, because of their usefulness in consideration of questions of the existence of the optimum controller and characterization of optimum controllers as controllers which steer to the boundary of this set.

THEOREM 1. *Consider the system*

$$\dot{x}^0(t) = x(t)'W(t)x(t) + u(t)'U(t)u(t),$$

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

with the continuous initial function $\phi(t)$, $t_0 - h_m \leqq t \leqq t_0$, *and* $x^0(t_0) = 0$. *The set of attainability* $\hat{K}(T) \subset R^{n+1}$ *is convex and its orthogonal projection on the hyperplane* $x^0 = 0$ *is a linear manifold* $K(T)$. *Also, if* $\hat{y} = (y^0, y)$ *is in* $\hat{K}(T)$, *then the halfline* $x^0 \geqq y^0$, $x = y$, *is in* $\hat{K}(T)$.

A proof of this result when there are no time delays is given in [9, Chap. III] and requires no essential modification in the above delay case.

THEOREM 2. *The set of attainability* $\hat{K}(T)$, *of Theorem* 1, *is closed in* R^{n+1}.

Proof. Define a set $\tilde{K}(T) \subset R^{n+1}$ by

$$\tilde{K}(T) = \{(\sqrt{x^0}, x) \mid (x^0, x) \in \hat{K}(T)\}.$$

Then clearly $\tilde{K}(T)$ is also convex and contains the entire halfline $x^0 \geqq \sqrt{y^0}$, $x = y$, for all $(\sqrt{y^0}, y)$ in $\tilde{K}(T)$. Also, $\tilde{K}(T)$ lies above the hyperplane $x^0 = 0$, and its orthogonal projection on the hyperplane $x^0 = 0$ is the same linear manifold $K(T)$. Since $x^0 \geqq 0$ for all (x^0, x) in $\hat{K}(T)$, from the definition of $\tilde{K}(T)$ it is obvious that $\tilde{p} = (\sqrt{p^0}, p)$ is a boundary point of $\tilde{K}(T)$ if and only if $\hat{p} = (p^0, p)$ is a boundary point of $\hat{K}(T)$. Therefore, if

[1] By $(x^0, x) \in R^{n+1}$ we mean an $n + 1$ column vector with components $(x^0, x^1, \cdots, x^n)$.

we can show that for all the boundary points $\tilde{p} = (\sqrt{p^0}, p)$ of $\tilde{K}(T)$ the corresponding point $\hat{p} = (p^0, p)$ is in $\hat{K}(T)$, then $\hat{K}(T)$ is closed.

We assume that $\hat{K}(T)$, and hence $\tilde{K}(T)$, has a nonempty interior in R^{n+1}, for otherwise we could define our subsequent constructions within the linear manifold spanned by $\hat{K}(T)$.

Each boundary point $\tilde{p} = (\sqrt{p^0}, p)$ of $\tilde{K}(T)$ has a support hyperplane with exterior normal extending towards the hyperplane $x^0 = 0$, as follows from the properties of $\tilde{K}(T)$ given above. Therefore there exists a point $\hat{q} = (0, q)$ on the hyperplane $x^0 = 0$ such that $\tilde{p} = (\sqrt{p^0}, p)$ is the unique point in $\overline{\tilde{K}}(T)$, which is the closest point to $\hat{q}$ in terms of the usual metric, where $\overline{\tilde{K}}(T)$ is the closure of $\tilde{K}(T)$. That is, $\tilde{p} = (\sqrt{p^0}, p)$ is the unique point in $\overline{\tilde{K}}(T)$ such that

$$\begin{aligned} | p^0 | + \| p - q \|^2 &= \inf_{(\sqrt{r^0}, r) \in \tilde{K}(T)} \{ | r^0 | + \| r - q \|^2 \} \\ &= \inf_{(r^0, r) \in \hat{K}(T)} \{ | r^0 | + \| r - q \|^2 \}. \end{aligned}$$

Here $\| p - q \|^2 = (p - q)'(p - q)$.

Thus, if we show that for each given $\hat{q} = (0, q)$ of the plane $x^0 = 0$ there exists a point $\tilde{p} = (\sqrt{p^0}, p)$ in $\tilde{K}(T)$, i.e., $\hat{p} = (p^0, p)$ in $\hat{K}(T)$, which satisfies the above condition, then each boundary point $\tilde{p} = (\sqrt{p^0}, p)$ of $\tilde{K}(T)$ is in $\tilde{K}(T)$ and so $\tilde{K}(T)$ is closed. This in turn proves that $\hat{K}(T)$ is closed.

Consider a sequence of controls $\{u_i(s)\}$ such that

$$\lim_{i \to \infty} \left\{ \int_{t_0}^{T} (\| x_i(s) \|_W^2 + \| u_i(s) \|_U^2) \, ds + \| x_i(T) - q \|^2 \right\} = \alpha,$$

where $x_i(s)$ is the response corresponding to $u_i(s)$ and

$$\alpha = \inf_{\hat{r} \in \hat{K}(T)} \{ | r^0 | + \| r - q \|^2 \}.$$

Let $H(t) = x(t, \phi)$,

$$P_i(t) = \int_{t_0}^{t} Y(s, t) B(s) \, u_i(s) \, ds.$$

Then $x_i(t) = x_{u_i}(t) = H(t) + P_i(t)$. Now define

$$J(u) = \int_{t_0}^{T} (\| x(s) \|_W^2 + \| u(s) \|_U^2) \, ds + \| x(T) - q \|^2 - \int_{t_0}^{T} \| H(s) \|_W^2 \, ds - \| H(T) - q \|^2.$$

Then

$$J(u) = 2P(T)'(H(T) - q) + \| P(T)\|^2 + \int_{t_0}^{T} \{\| P(s)\|_W^2 + 2H(s)'W(s)P(s) + \| u(s)\|_U^2\}\, ds.$$

It is straightforward to show that

$$J\left(\frac{u_i - u_j}{2}\right) + J\left(\frac{u_i + u_j}{2}\right) = \frac{1}{2}J(u_i) + \frac{1}{2}J(u_j) + (H(T) - q)'(P_i(T) - P_j(T)) + \int_{t_0}^{T} H(s)'W(s)(P_i(s) - P_j(s))\, ds$$

and

$$\frac{1}{2}\left[J(u_i) + J(u_j) - 2J\left(\frac{u_i + u_j}{2}\right)\right] = J\left(\frac{u_i - u_j}{2}\right) - (H(T) - q)'(P_i(T) - P_j(T)) - \int_{t_0}^{T} H(s)'W(s)(P_i(s) - P_j(s))\, ds$$

$$= \left\| \frac{P_i(T) - P_j(T)}{2} \right\|^2 + \int_{t_0}^{T} \left[\left\| \frac{P_i(s) - P_j(s)}{2} \right\|_W^2 + \left\| \frac{u_i(s) - u_j(s)}{2} \right\|_U^2 \right] ds.$$

Now, $J(u_i) \to \inf_u J(u) = \beta$ and

$$J\left(\frac{u_i + u_j}{2}\right) \geqq \beta$$

as $i \to \infty$; therefore

$$\frac{1}{2}[J(u_i) + J(u_j) - 2\beta] \geqq \left\| \frac{P_i(T) - P_j(T)}{2} \right\|^2 + \int_{t_0}^{T} \left[\left\| \frac{P_i(s) - P_j(s)}{2} \right\|_W^2 + \left\| \frac{u_i(s) - u_j(s)}{2} \right\|_U^2 \right] ds.$$

Since $J(u_i) + J(u_j) - 2\beta > 0$ and $J(u_i) + J(u_j) - 2\beta \to 0$ as $i, j \to \infty$,

$$\lim_{i,j \to \infty} \int_{t_0}^{T} \| u_i(s) - u_j(s)\|_U^2\, ds = 0,$$

for $U(s)$ is positive definite. Then, by the Riesz-Fischer Theorem, $\{u_i\}$

converges strongly to $u^*(s) \in L_2[t_0, T]$ and

$$\int_{t_0}^{T} (\| x^*(s)\|_W^2 + \| u^*(s)\|_U^2)\, ds + \| x^*(T) - q \|^2 = \alpha,$$

where $x^*(t)$ is the response corresponding to $u^*(t)$. Hence $\hat{p} = (p^0, p) = (x^{0*}(T), x^*(T)) \in \hat{K}(T)$ and so $\hat{K}(T) = \bar{\hat{K}}(T)$ is closed.

In the above proof, we have also proved the following corollary.

COROLLARY. *Let*

$$x^0(t) = \left[\int_{t_0}^{t} (\| x(s)\|_W^2 + \| u(s)\|_U^2)\, ds\right]^{1/2}$$

in Theorem 1, *and let* $\tilde{K}(T)$ *be the corresponding set of attainability. Then* $\tilde{K}(T)$ *is closed in* R^{n+1}.

THEOREM 3. (Existence) *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

with cost functional

$$C(u) = g(x(T)) + \int_{t_0}^{T} (\| x(t)\|_W^2 + \|u(t)\|_U^2)\, dt.$$

If either

(a) $g(x)$ *is bounded below, i.e.,* $g(x) > a$,

or

(b) $g(x)$ *is a convex function,*

then there exists a (minimal cost) optimal control.

Proof. Let

$$x^0(T) = \int_{t_0}^{T} (\| x(t)\|_W^2 + \| u(t)\|_U^2)\, dt.$$

Then $\hat{K}(T)$ is closed and convex by Theorems 1 and 2. We have to show that the function $g(x) + x^0$ has a minimum in $\hat{K}(T)$. If $g(x) > a$, then

$$\lim_{x^0 \to \infty} (g(x) + x^0) = +\infty$$

uniformly in $\hat{K}(T)$. Thus there exists a bound $\alpha > 0$ such that the minimum of $g(x) + x^0$ in $\hat{K}(T)$ is assumed on the compact set $\hat{K}(T) \cap [x^0 \leqq \alpha]$.

If $g(x)$ is a convex function, then for each real number C_1, the set $\hat{G} \subset R^{n+1}$ defined by

$$\hat{G} = \{(x^0, x) \mid g(x) + x^0 \leqq C_1\}$$

is closed and has nonempty interior. Furthermore, this set is convex, for

$g(x_1) + x_1^0 \leqq C_1$ and $g(x_2) + x_2^0 \leqq C_1$ imply

$$g(\lambda x_1 + (1 - \lambda)x_2) + \lambda x_1^0 + (1 - \lambda)x_2^0 \leqq C_1, \qquad 0 \leqq \lambda \leqq 1.$$

Now consider a constant C_1 such that the corresponding set meets $\hat{K}(T)$, i.e., $\hat{G} \cap \hat{K}(T) \neq \varnothing$. It is easy to show that $\hat{G} \cap \hat{K}(T)$ is compact. (See Fig. 1. A proof can be found in [9].) But then, since $g(x) + x^0$ is continuous in (x^0, x), it assumes a minimum value on the compact set and so there exists an optimal control. This completes the proof.

Actually the minimum occurs at a boundary point $(x^0, x) \in \partial\hat{K}(T)$ of $\hat{K}(T)$. For otherwise, there exists a point $(y^0, x) \in \hat{K}(T)$ with $y^0 < x^0$; but then $g(x) + x^0 > g(x) + y^0$, that is, $g(x) + x^0$ is not a minimum. Therefore $(x^0, x) \in \partial\hat{K}(T)$. This is an important property. For this reason we now study properties of the boundary $\partial\hat{K}(T)$ of $\hat{K}(T)$ in more detail.

DEFINITION. Consider the system

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

$$\dot{x}^0(t) = x(t)'W(t)x(t) + u(t)'U(t)u(t),$$

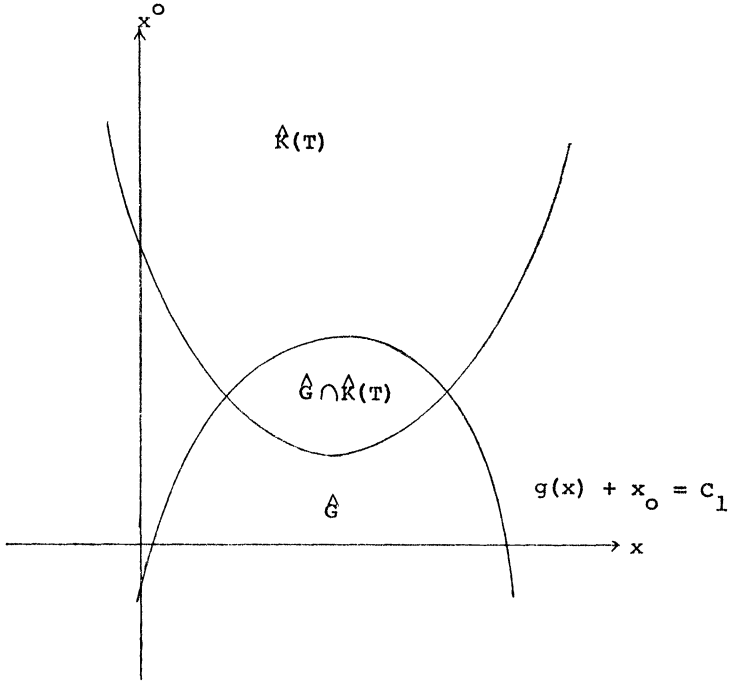

FIG. 1

with initial function $x(t) = \phi(t)$, $t_0 - h_m \leqq t \leqq t_0$, $x^0(t_0) = 0$. A controller $u(t) \in L_2[t_0, T]$ is called *extremal* if and only if it steers the response $\hat{x}(t)$ to the boundary $\partial\hat{K}(T)$ of $\hat{K}(T)$ at $t = T$.

The following theorem is equivalent to Pontryagin's maximum principle [13] for the present problem, but will also provide sufficiency conditions for optimal control.

THEOREM 4. (Maximum principle) *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

$$\dot{x}^0(t) = x(t)'W(t)x(t) + u(t)'U(t)u(t),$$

with initial function $x(t) = \phi(t)$, $t_0 - h_m \leqq t \leqq t_0$, $x^0(t_0) = 0$. *A controller* $\bar{u}(t) \in L_2[t_0, T]$ *with response* $\bar{x}(t)$ *is extremal if and only if there exists a nontrivial solution* $\hat{\eta}(t) = (\eta_0, \eta(t)) \in R^{n+1}$ *of the adjoint equation*

$$\dot{\eta}(t) = -\sum_{i=0}^{m} \eta(t + h_i)A_i(t + h_i) - 2\eta_0\bar{x}(t)'W(t), \quad t_0 \leqq t \leqq T - h_m,$$

$$\dot{\eta}(t) = -\sum_{i=0}^{k} \eta(t + h_i)A_i(t + h_i) - 2\eta_0\bar{x}(t)'W(t),$$

$$T - h_{k+1} \leqq t \leqq T - h_k, \quad k = 0, 1, \cdots, (m - 1),$$

with $\eta_0 < 0$ *a constant,*[2] *such that*

$$\eta_0 \| \bar{u}(t) \|_U^2 + \eta(t)B(t)\bar{u}(t) = \max_{u \in R^r} \{\eta_0 \| u \|_U^2 + \eta(t)B(t)u\}$$

or

$$\bar{u}(t) = -\frac{1}{2\eta_0} U(t)^{-1}B(t)'\eta(t)' \qquad \text{a.e. on } [t_0, T]$$

for

$$\| u \|_U^2 - \frac{\eta}{|\eta_0|} Bu = \left\| u - \frac{U^{-1}B'\eta'}{2|\eta_0|} \right\|_U^2 - \left\| \frac{U^{-1}B'\eta'}{2|\eta_0|} \right\|_U^2.$$

Proof. Suppose

$$\bar{u}(t) = -\frac{1}{2\eta_0} U(t)^{-1}B(t)'\eta(t)'.$$

Since only the ratio η/η_0 enters the hypothesis, we can choose $\eta_0 = -1/2$ without loss of generality. Then

$$\bar{u}(t) = U^{-1}(t)B(t)'\eta(t)'.$$

[2] Recall that all constructions assume that $\hat{K}(T)$ has an interior in R^{n+1}, or we reduce the problem to the linear manifold spanned by $\hat{K}(T)$.

Let $\hat{x}(t) = (\bar{x}^0(t), \bar{x}(t))$ be the corresponding response, and let $\hat{y}(t) = (y^0(t), y(t))$ be the general response to any control $u(t) \in L_2[t_0, T]$. Consider

$$\frac{d}{dt}(\hat{\eta}(t)\hat{y}(t)) = -\frac{1}{2}\dot{y}^0(t) + \eta(t)\dot{y}(t) + \dot{\eta}(t)y(t).$$

Integration from t_0 to T yields

$$\hat{\eta}(T)\hat{y}(T) - \hat{\eta}(t_0)\hat{y}(t_0) = -\frac{1}{2}\int_{t_0}^{T}\dot{y}^0(t)\,dt + \int_{t_0}^{T}\eta(t)\dot{y}(t)\,dt + \int_{t_0}^{T}\dot{\eta}(t)y(t)\,dt.$$

Now

$$\int_{t_0}^{T}\eta(t)\dot{y}(t)\,dt = \int_{t_0}^{T}\eta(t)\left\{\sum_{i=0}^{m}A_i(t)y(t-h_i) + B(t)u(t)\right\}dt$$

and

$$\begin{aligned}\int_{t_0}^{T}&\dot{\eta}(t)y(t)\,dt \\ &= \int_{t_0}^{T-h_m}\left\{-\sum_{i=0}^{m}\eta(t+h_i)A_i(t+h_i)y(t) + \bar{x}(t)'W(t)y(t)\right\}dt \\ &+ \sum_{i=0}^{m-1}\int_{T-h_{k+1}}^{T-h_k}\left\{-\sum_{i=0}^{k}\eta(t+h_i)\,A_i(t+h_i)y(t) + \bar{x}(t)'W(t)y(t)\right\}dt \\ &= -\sum_{i=0}^{m}\int_{t_0+h_i}^{T}\eta(t)A_i(t)y(t-h_i)\,dt + \int_{t_0}^{T}\bar{x}(t)'W(t)y(t)\,dt.\end{aligned}$$

Therefore

$$\begin{aligned}&\hat{\eta}(T)\hat{y}(T) - \hat{\eta}(t_0)\hat{y}(t_0) \\ &= \int_{t_0}^{T}\left\{-\frac{1}{2}\dot{y}_0(t) + \bar{x}(t)'W(t)y(t) + \eta(t)B(t)u(t)\right\}dt \\ &\qquad + \sum_{i=0}^{m}\int_{t_0}^{t_0+h_i}\eta(t)A_i(t)\phi(t-h_i)\,dt\end{aligned}$$

for $y(t-h_i) = \phi(t-h_i)$ on $t_0 \leqq t \leqq t_0 + h_m$. Also

$$\begin{aligned}&\int_{t_0}^{T}\left\{-\frac{1}{2}\dot{y}^0(t) + \bar{x}(t)'W(t)y(t) + \eta(t)B(t)u(t)\right\}dt \\ &= \int_{t_0}^{T}\left\{-\frac{1}{2}\left\|y(t)\right\|_W^2 - \frac{1}{2}\left\|u(t)\right\|_U^2 + \bar{x}(t)'W(t)y(t) + \eta(t)B(t)u(t)\right\}dt.\end{aligned}$$

If $u(t) = \bar{u}(t) = U^{-1}B(t)'\eta(t)'$, then $y(t) = \bar{x}(t)$ and

$$\int_{t_0}^{T} \left\{ -\frac{1}{2}\dot{\bar{x}}^0(t) + \bar{x}(t)'W(t)\bar{x}(t) + \eta(t)B(t)\bar{u}(t) \right\} dt$$

$$= \int_{t_0}^{T} \left\{ \frac{1}{2} \left\| \bar{x}(t) \right\|_W^2 - \frac{1}{2} \left\| \bar{u}(t) \right\|_U^2 + \eta(t)B(t)\bar{u}(t) \right\} dt.$$

Then, since

$$-\tfrac{1}{2}\| \bar{u}(t) \|_U^2 + \eta(t)B(t)\bar{u}(t) = \max_{u \in R^r} (-\tfrac{1}{2}\| u \|_U^2 + \eta(t)B(t)u),$$

and $\| \bar{x}(t) - y(t) \|_W^2 \geqq 0$ implies that

$$\tfrac{1}{2}\| \bar{x}(t) \|_W^2 \geqq -\tfrac{1}{2}\| y(t) \|_W^2 + x(t)'W(t)y(t),$$

we have

$$\int_{t_0}^{T} \left\{ \frac{1}{2} \left\| \bar{x}(t) \right\|_W^2 - \frac{1}{2} \left\| \bar{u}(t) \right\|_U^2 + \eta(t)B(t)\bar{u}(t) \right\} dt$$

$$> \int_{t_0}^{T} \left\{ -\frac{1}{2} \left\| y(t) \right\|_W^2 - \frac{1}{2} \left\| u(t) \right\|_U^2 + \bar{x}(t)'W(t)y(t) + \eta(t)B(t)u(t) \right\} dt,$$

unless $u(t) = \bar{u}(t)$ almost everywhere on $[t_0, T]$. Therefore

$$\hat{\eta}(T)\hat{x}(T) - \hat{\eta}(t_0)\hat{x}(t_0) > \hat{\eta}(T)\hat{y}(T) - \hat{\eta}(t_0)\hat{y}(t_0).$$

Since $\hat{x}(t_0) = \hat{y}(t_0) = (0, \phi(t_0))$, we have $\hat{\eta}(T)\hat{x}(T) > \hat{\eta}(T)\hat{y}(T)$ for all $\hat{y}(T) \neq \hat{x}(T)$ in $\hat{K}(T)$. But this shows that there is a supporting hyperplane to $\hat{K}(T)$ at $\hat{x}(T)$ having the exterior normal vector $\hat{\eta}(T)$. Since $\eta_0 < 0$, the supporting hyperplane cannot meet $\hat{K}(T)$ in its interior and so it must meet on the boundary $\partial\hat{K}(T)$. Hence $\bar{u}(t)$ is extremal.

Conversely suppose $\bar{u}(t)$ steers the response $\hat{x}(t) = (\bar{x}^0(t), \bar{x}(t))$ to the boundary, i.e., $\hat{x}(T) \in \partial\hat{K}(T)$. Let $\hat{\eta}(T) = (-\frac{1}{2}, \bar{\eta}(T))$ be an exterior normal to $\hat{K}(T)$ at $\hat{x}(T)$ and let $\bar{\eta}(t)$ be the solution of the adjoint equation with the above endpoint condition $\bar{\eta}(T)$ at $t = T$. We want to show that $\bar{u}(t)$ satisfies the maximum principle.

Suppose

$$-\tfrac{1}{2}\| \bar{u} \|_U^2 + \bar{\eta}(t)B(t)\bar{u}(t) + \delta \leqq \max_{u \in R^r} (-\tfrac{1}{2}\| u \|_U^2 + \bar{\eta}(t)B(t)u)$$

for some $\delta > 0$ on some compact subset E of $[t_0, T]$ of positive measure. For each small $\epsilon > 0$, define a new controller $u_\epsilon(t)$,

$$u_\epsilon(t) = \begin{cases} U^{-1}(t)B(t)'\bar{\eta}(t)' & \text{if } t \in E(\epsilon), \\ \bar{u}(t) & \text{if } t \notin E(\epsilon), \end{cases}$$

where $E(\epsilon)$ is a subset of E with measure ϵ. Let $\hat{x}_\epsilon(t)$ be the corresponding

response. Then, for some constant $C_1 > 0$,

$$| \hat{x}_\epsilon(t) - \hat{x}(t) | \leqq C_1\epsilon$$

on $[t_0, T]$, for $\hat{x}_\epsilon(t)$ is continuous in the parameter ϵ. Now compute as before:

$$\hat{\eta}(T)\hat{x}(T) - \hat{\eta}(T)\hat{x}_\epsilon(T) \leqq \int_{t_0}^{T} \frac{1}{2} \left\| \bar{x}(t) - x_\epsilon(t) \right\|_W^2 dt - \int_{E(\epsilon)} \delta \, dt$$

$$\leqq C_2\epsilon^2 - \epsilon$$

for some constant $C_2 > 0$. Thus for sufficiently small $\epsilon > 0$,

$$\hat{\eta}(T)\hat{x}_\epsilon(T) > \hat{\eta}(T)\hat{x}(T).$$

This is impossible, for $\hat{\eta}(T)$ is an exterior normal to $\hat{K}(T)$ at $\hat{x}(T)$. Therefore $\bar{u}(t)$ must satisfy the maximum principle.

Since $g(x) + x^0$ assumes its minimum on the boundary $\partial\hat{K}(T)$ as discussed after Theorem 4, the following corollary is an immediate result of Theorem 4.

COROLLARY. *Consider the cost functional as in Theorem* 3:

$$C(u) = g(x(T)) + \int_{t_0}^{T} \{ \| x(t) \|_W^2 + \| u(t) \|_U^2 \} \, dt.$$

Then an optimal controller is an extremal controller, that is, $u(t)$ is of the form $u(t) = U(t)^{-1}B(t)'\eta(t)'$ a.e. on $[t_0, T]$, where $\eta(t)$ is a solution of the adjoint equation.

THEOREM 5. (Uniqueness) *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t)$$

with cost functional

$$C(u) = x^0(T) = \int_{t0}^{T} \{ \| x(t) \|_W^2 + \| u(t) \|_U^2 \} \, dt.$$

Let $u_1(t)$ and $u_2(t)$ be extremal controllers with corresponding responses $\hat{x}_1(t)$ and $\hat{x}_2(t)$. If $\hat{x}_1(T) = \hat{x}_2(T) = \hat{x}_1$, then $u_1(t) = u_2(t)$ almost everywhere on $[t_0, T]$.

Proof. Let $\hat{\eta}(t)$ be the solution of the adjoint equation with $\hat{\eta}(T) = (-\frac{1}{2}, \eta(T))$, an exterior normal to $\hat{K}(T)$ at $\hat{x}_1 \in \partial\hat{K}(T)$. Then, as in the proof of Theorem 4,

$$u_1(t) = u_2(t) = U(t)^{-1}B(t)'\eta(t)'$$

almost everywhere on $[t_0, T]$.

THEOREM 6. *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

$$\dot{x}^0(t) = x(t)'W(t)x(t) + u(t)'U(t)u(t),$$

with initial condition $x(t) = \phi(t)$, $t_0 - h_m \leqq t \leqq t_0$, and $x^0(t_0) = 0$. *Suppose the cost functional is given by* $C(u) = g(x(T)) + x^0(T)$, *where* $g(x)$ *is a* C^1 *convex function. Then there exists a unique hypersurface* S_m *among the family*

$$S_c : g(x) + x^0 = c,$$

such that S_m *is tangent to* $\hat{K}(T)$, *and hence* m *is the optimal cost. Also, there exists a unique optimal controller, i.e., the extremal controller* $u^*(t)$ *which steers the response to the single point at which* S_m *touches* $\hat{K}(T)$ *is unique. Furthermore, there is a unique solution of the equations*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)U(t)^{-1}B(t)'\eta(t)',$$

$$\dot{\eta}(t) = -\sum_{i=0}^{m} \eta(t + h_i)A_i(t + h_i) + x(t)'W(t), \quad t_0 \leqq t \leqq T - h_m,$$

$$\dot{\eta}(t) = -\sum_{i=0}^{k} \eta(t + h_i)A_i(t + h_i) + x(t)'W(t),$$

$$T - h_{k+1} \leqq t \leqq T - h_k, \quad k = 0, 1, \cdots, (m - 1),$$

satisfying the boundary conditions

$$x(t) = \phi(t),\ t_0 - h_m \leqq t \leqq t_0, \quad \textit{and} \quad \eta(T) = -\tfrac{1}{2} \operatorname{grad} g(x(T));$$

i.e., the optimal response $x^*(t)$ *and* $\eta^*(t)$ *are such that*

$$u^*(t) = U(t)^{-1}B(t)'\eta^*(t)'$$

is the optimal control on $[t_0, T]$.

Proof. By Theorem 3, the intersection of $\hat{K}(T)$ with $g(x) + x^0 \leqq c$ is compact for large c. Let m be the infimum of all c such that the intersection is nonempty. For $c > m$ the hypersurface S_c meets the interior of $\hat{K}(T)$, and for $c < m$, S_c does not meet $\hat{K}(T)$. Therefore S_c can be tangent to $\hat{K}(T)$ only if $c = m$ and m is the minimum cost. Now let $p \in \hat{K}(T) \cap S_m$ and let π be the tangent hyperplane to S_m at p. Then π fails to separate $\hat{K}(T)$ and S_m only in the case where π meets the interior of $\hat{K}(T)$. But if an open set N of the interior of $\hat{K}(T)$ lies below π, then the cone with base N and vertex p lies interior to $\hat{K}(T)$ and so below π. However, S_m is tangent to π at p which implies that S_m meets the interior of $\hat{K}(T)$. This is impossible, for m is the minimum cost. Thus π separates S_m and $\hat{K}(T)$.

Now suppose there are two distinct points, p_1 and p_2, in $\hat{K}(T) \cap S_m$ and hence on $\partial\hat{K}(T)$. Then $\frac{1}{2}(p_1 + p_2) \in \partial\hat{K}(T)$. Let $u_1(t)$ and $u_2(t)$ be the two extremal controls with responses $\hat{x}_1(t)$ and $\hat{x}_2(t)$ such that $\hat{x}_1(T) = p_1$ and $\hat{x}_2(T) = p_2$. Since $p_1 \neq p_2$, $u_1(t) \neq u_2(t)$ on some positive interval. Consider the control $\frac{1}{2}(u_1(t) + u_2(t))$ with response $\hat{x}(t) = (x^0(t), x(t))$. Then $x(T) = \frac{1}{2}(x_1(T) + x_2(T))$. Now we want to show that $x^0(T) < \frac{1}{2}(x_1^0(T) + x_2^0(T))$. By definition,

$$x^0(T) = \int_{t_0}^{T} \left\{ \left\| \frac{x_1(t) + x_2(t)}{2} \right\|_W^2 + \left\| \frac{u_1(t) + u_2(t)}{2} \right\|_U^2 \right\} dt$$

$$= \int_{t_0}^{T} \left\{ \frac{1}{4} \| x_1 \|_W^2 + \frac{1}{2} x_1' W x_2 + \frac{1}{4} \| x_2 \|_W^2 \right.$$

$$\left. + \frac{1}{4} \| u_1 \|_U^2 + \frac{1}{2} u_1' U u_2 + \frac{1}{4} \| u_2 \|_U^2 \right\} dt.$$

Since $2x_1'Wx_2 \leqq \| x_1 \|_W^2 + \| x_2 \|_W^2$ and since

$$2u_1'Uu_2 < \| u_1 \|_U^2 + \| u_2 \|_U^2 \quad \text{whenever} \quad u_1(t) \neq u_2(t),$$

we have

$$x^0(T) < \frac{1}{2} \int_{t_0}^{T} \left\{ \| x_1 \|_W^2 + \| u_1 \|_U^2 \right\} dt + \frac{1}{2} \int_{t_0}^{T} \left\{ \| x_1 \|_W^2 + \| u_2 \|_U^2 \right\} dt.$$

Therefore

$$x^0(T) < \tfrac{1}{2}(x_1^0(T) + x_2^0(T)).$$

Then the halfline $x^0 > x^0(T)$, $x = x(T)$, is in the interior of $\hat{K}(T)$, and so $\frac{1}{2}(p_1 + p_2)$ is also in the interior of $\hat{K}(T)$. But $\frac{1}{2}(p_1 + p_2) \in \hat{K}(T) \cap S_m \subset \partial\hat{K}(T)$. This contradiction establishes that $\hat{K}(T) \cap S_m$ is a single point.

By Theorem 5, there is a unique extremal controller $u^*(t)$ steering the response to $p = \hat{K} \cap S_m$; therefore $u^*(t)$ is the optimal controller and $p = (x^{*0}(T), x^*(T))$.

The normal vector to S_m at $p = \hat{x}^*(T)$ is $\hat{\eta}^*(T) = (-\frac{1}{2}, \eta^*(T))$, where $\eta^*(T) = -\frac{1}{2}$ grad $g(p)$. By Theorem 4, $x^*(t)$ and $\eta^*(t)$ satisfy

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)U(t)^{-1}B(t)'\eta(t)',$$

$$\dot{\eta}(t) = -\sum_{i=0}^{m} \eta(t + h_i)A_i(t + h_i) + x(t)'W(t), \quad t_0 \leqq t \leqq T - h_m,$$

$$\dot{\eta}(t) = -\sum_{i=0}^{k} \eta(t + h_i)A_i(t + h_i) + x(t)'W(t),$$

$$T - h_{k+1} \leqq t \leqq T - h_k, \quad k = 0, 1, \cdots, (m - 1),$$

with $x^*(t) = \phi(t)$, $t_0 - h_m \leqq t \leqq t_0$, and $\eta^*(T) = -\frac{1}{2}\operatorname{grad} g(x^*(T))$. Now let $x(t)$, $\eta(t)$ be any solution of the above equations with the same boundary conditions. Then $\hat{x}(t) = (x^0(t), x(t))$ is the response corresponding to the control $u(t) = U(t)^{-1}B(t)'\eta(t)'$, and

$$\hat{\eta}(T)\hat{x}(T) = \tfrac{1}{2}x^0(T) + \eta(T)x(T) > \hat{\eta}(T)\hat{y}$$

for all $\hat{y} \neq \hat{x}(T)$ in $\hat{K}(T)$. Thus $\hat{\eta}(T)$ is the exterior normal to the supporting hyperplane π to $\hat{K}(T)$ at $\hat{x}(T)$. Also $\hat{\eta}(T)$ is the inward normal to the hypersurface S_c through $\hat{x}(T)$, for $\eta(T) = -\frac{1}{2}\operatorname{grad} g(x(T))$. Thus S_c is tangent to $\hat{K}(T)$ at $\hat{x}(T)$ and π is the common supporting hyperplane. But then $S_c = S_m$ and $\hat{x}(T) = \hat{x}^*(T)$. Therefore by the uniqueness, $u(t) = u^*(t)$ a.e., and $\hat{x}(t) = \hat{x}^*(t)$. Finally, $\eta(t)$ is the unique solution of the adjoint equation with $\eta(T) = -\frac{1}{2}\operatorname{grad} g(x^*(T))$, and so $\eta(t) = \eta^*(t)$.

Remark. It is easy to show that all the theorems in this section are equally valid if $x^0(t)$ is given by

$$x^0(t) = \left\{\int_{t_0}^{T} [\| x(t) \|_W^2 + \| u(t) \|_U^2]\, dt\right\}^{1/2}.$$

2. Controllability. In the previous section we assumed that $\hat{K}(T)$ has an interior in R^{n+1}. This property is important when the target G is given by a compact set. In this section we study the problem of controllability.

Let B be the set of all continuous functions $\phi(t)$ on $[t_0 - h_m, t_0]$. Define the norm $\| \phi(t) \|$ by

$$\| \phi(t) \| = \max_{t_0 - h_m \leqq t \leqq t_0} | \phi(t) |.$$

Then B is a Banach space with respect to this norm.

DEFINITION. Consider the system

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t)$$

in R^n. The system is said to be *controllable on* $[t_0, T]$ if and only if for any given initial function $\phi(t)$ in the Banach space B on $[t_0 - h_m, t_0]$ and any given target point x_1 in R^n, there exists a bounded measurable controller $u(t)$ on $[t_0, T]$ which steers the given initial function to the target at $t = T$ along the solution curves of

$$\dot{x} = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t).$$

THEOREM 7. *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t)$$

in R^n. The system is controllable on $[t_0, T]$ if and only if the matrix

$$M = \int_{t_0}^{T} Y(s, T)B(s)B(s)'Y(s, T)' \, ds$$

has rank n.

The proof is exactly the same as the proof for systems without time delay [7]. For a proof see [3, Chap. 2].

COROLLARY. *The set of attainability $\hat{K}(T)$ has nonempty interior if and only if the system is controllable on $[t_0, T]$.*

Proof. If the system is controllable, then clearly the set of attainability $K(T)$ is all of R^n. Then, since $K(T)$ is the orthogonal projection of $\hat{K}(T)$ on the hyperplane, $x^0 = 0$; $\hat{K}(T)$ has nonempty interior, because it contains the vertical rays. See Theorem 1.

If $\hat{K}(T)$ has nonempty interior, then by Theorem 1, $K(T) = R^n$, and hence the system must be controllable.

3. Closed and convex target sets. Often we wish to steer the initial function to a closed convex target set $G \subset R^n$. The target set G may be a compact set. The problem is to minimize the cost functional $C(u)$ while steering the response to G at time T.

Consider the system

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

with cost functional

$$C(u) = x^0(T) = \int_{t_0}^{T} \{\| x(t)\|_W^2 + \| u(t)\|_U^2\} \, ds.$$

For simplicity we assume that the system is controllable and hence the set of attainability $\hat{K}(T)$ is a closed convex set with nonempty interior. If the system is not controllable, then a similar analysis can be made in the linear manifold spanned by $\hat{K}(T)$ in R^{n+1} provided G meets $K(T)$.

Define $\hat{G} = G \times R^1$ in the (x^0, x)-space R^{n+1}. Since $\hat{K}(T)$ has nonempty interior, $\hat{G}$ intersects $\hat{K}(T)$ and the intersection is closed and convex. We seek a point of $\hat{G} \cap \hat{K}(T)$ with minimum cost coordinate $x^0(T)$. Since the projection of $\hat{G} \cap \hat{K}(T)$ on the x^0-axis is bounded below and closed, there exists a point $\hat{x}^*(T)$ in $\hat{G} \cap \hat{K}(T)$ attaining this minimum value; the control $u^*(t)$ which steers the response to this point is an optimal controller.

If $\hat{x}^*(T)$ is an inner point of $\hat{G}$, then $x^{0*}(T)$ is the minimum value of $x^0(T)$ for all points in $\hat{K}(T)$, and this problem is the same problem as the free endpoint problem, when $G = R^n$, a case which is covered in §1. Therefore we assume that $\hat{x}^*(T)$ is a boundary point of $\hat{G} \cap \hat{K}(T)$. Then $u^*(t)$

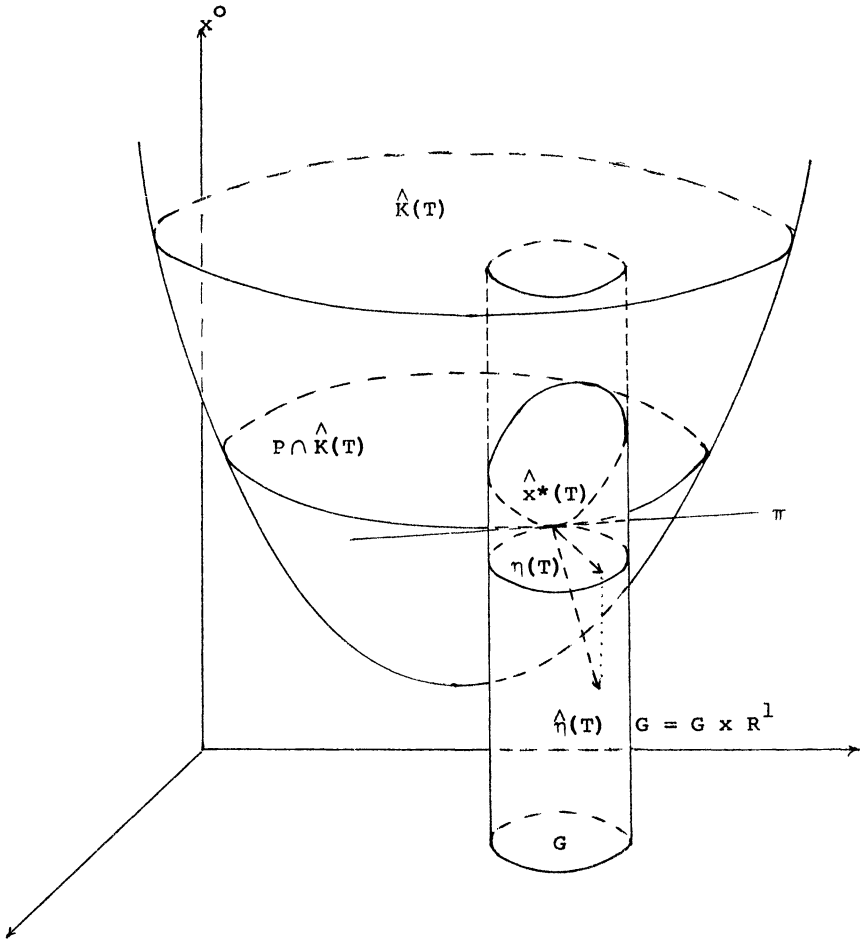

FIG. 2

is an extremal control and is given by $u^*(t) = U(t)^{-1}B(t)'\eta(t)'$ (see Figs. 2 and 3). Here $\hat{\eta}(t) = (-\frac{1}{2}, \eta(t))$, and $\eta(t)$ is the solution of the adjoint equation:

$$\dot{\eta}(t) = -\sum_{i=0}^{m} \eta(t + h_i)A_i(t + h_i) + x^*(t)'W(t), \quad t_0 \leqq t \leqq T - h_m,$$

$$\dot{\eta}(t) = -\sum_{i=0}^{k} \eta(t + h_i)A_i(t + h_i) + x^*(t)'W(t),$$

$$T - h_{k+1} \leqq t \leqq T - h_k, \quad k = 0, 1, \cdots, (m - 1),$$

with $\hat{\eta}(T)$ an exterior normal to $\hat{K}(T)$ at $\hat{x}^*(T)$.

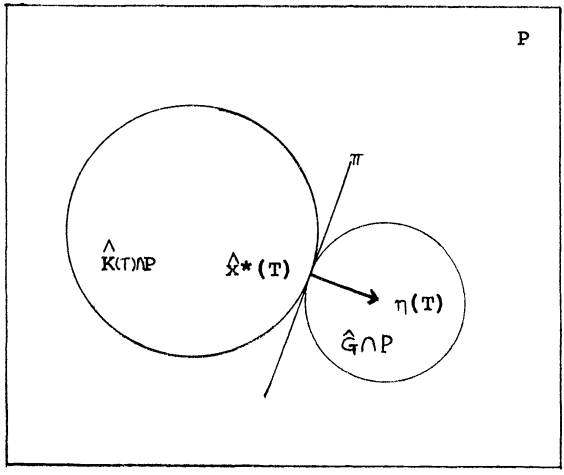

FIG. 3

Consider the cross section P: $x^0 = x^{0*}(T)$ in R^{n+1}. Then $\hat{G} \cap P$ and $\hat{K}(T) \cap P$ are separated by a common supporting $(n-1)$-plane π, and $\eta(T)$ is normal to π and directed into $\hat{G} \cap P$ in the plane P (see Fig. 3). By arguments similar to those for Theorem 6, we can show that there exists a unique optimal controller $u(t)$ and that $x(t)$ and $\eta(t)$ are the solutions of

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t-h_i) + B(t)U(t)^{-1}B(t)'\eta(t)'$$

$$\dot{\eta}(t) = -\sum_{i=0}^{m} \eta(t+h_i)A_i(t+h_i) + x(t)'W(t), \quad t_0 \leqq t \leqq T - h_m,$$

$$\dot{\eta}(t) = -\sum_{i=0}^{k} \eta(t+h_i)A_i(t+h_i) + x(t)'W(t),$$
$$T - h_{k+1} \leqq t \leqq T - h_k, \quad k = 0, 1, \cdots, (m-1),$$

with $x(t) = \phi(t), t_0 - h_m \leqq t \leqq t_0, x(T) \in \partial G$ and $\eta(T)$ an interior normal to G at $x(T)$.

If the target G is given by $f(x) \leqq 0$, where $f(x)$ is a convex C^1 function with grad $f(x) \neq 0$ for $x \in \partial G$, then the boundary condition is:

$$x(t) = \phi(t), t_0 - h_m \leqq t \leqq t_0, \quad f(x(T)) = 0, \quad \eta(T) = -k \operatorname{grad} f(x(T))$$

for some constant $k > 0$.

4. Integral convex cost functionals. Consider the system in R^{n+1}:

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t-h_i) + B(t)u(t),$$

$$\dot{x}^0(t) = f^0(t, x(t)) + h^0(t, u(t)),$$

with initial function $x(t) = \phi(t), t_0 - h_m \leqq t \leqq t_0, x^0(t_0) = 0$ and $u(t) \in R^r$. Here $A_i(t)$, $B(t)$, $f^0(t, x)$, $h^0(t, u)$ are continuous for all t, $t_0 \leqq t \leqq T$, and all $x \in R^n$, $u \in R^r$. Also assume $f^0(t, x(t))$ and $h^0(t, u(t))$ are convex for each t, and

$$f^0(t, x) \geqq 0, \quad h^0(t, u) \geqq a|u|^p$$

for some constants $a > 0$ and $p > 1$. An admissible controller $u(t)$ is a measurable controller such that the corresponding cost functional

$$C_0(u) = x^0(T) = \int_{t_0}^{T} \{f^0(t, x(t)) + h^0(t, u(t))\}\, dt \geqq 0$$

is finite. Therefore every bounded measurable controller $u(t)$ is admissible, and $u(t) \in L_p[t_0, T]$ for

$$C_0(u) \geqq a \int_{t_0}^{T} |u(t)|^p\, dt,$$

so that $u(t) \in L_1[t_0, T]$. From convexity of $f^0(t, x)$ and $h^0(t, u)$, every convex combination of admissible controls is also admissible.

We assume that the system is controllable, for otherwise we can always consider the problem in the linear manifold which is spanned by the projection of $\hat{K}(T)$ on x-space. Here $\hat{K}(T)$ is the set of attainability as defined in §1. Since we assumed that the system is controllable, the projection of $\hat{K}(T)$ on x-space is the entire space R^n.

The response of the system is given by

$$x(t) = x(t, \phi) + \int_{t_0}^{t} Y(s, t)B(s)u(s)\, ds,$$

$$x^0(t) = \int_{t_0}^{t} \{f^0(s, x(s)) + h^0(s, u(s))\}\, ds.$$

The following theorems are direct generalizations of the theorems in the previous sections. The proofs are similar with obvious modifications, and hence are omitted.

THEOREM 8. *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

$$\dot{x}^0(t) = f^0(t, x(t)) + h^0(t, u(t)),$$

with cost functional

$$C_0(u) = x^0(T) = \int_{t_0}^{T} \dot{x}^0(t)\, dt.$$

Assume the basic part of the system is controllable, that is, the system without

x^0. *Then the orthogonal projection of* $\hat{K}(T)$ *on the hyperplane* $x^0 = 0$ *is the whole x-space* R^n. *Also, if* $\hat{y} = (y^0, y) \in \hat{K}(T)$, *then the entire halfline* $x^0 \geqq y^0$, $x = y$, *is in* $\hat{K}(T)$. *Furthermore,* $\hat{K}(T)$ *lies above the hypersurface* $x^0 = k|\, x\,|^p$ *for all large* $|\, x\,|$ *and for some constant* $k > 0$.

THEOREM 9. *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

$$\dot{x}^0(t) = f^0(t, x(t)) + h^0(t, u(t)),$$

with cost functional

$$C_0(u) = x^0(T) = \int_{t_0}^{T} \dot{x}^0(t)\, dt.$$

Then the set of attainability $\hat{K}(T) \subset R^{n+1}$ *is closed and convex.*

COROLLARY. *Consider the system in Theorem* 9 *with cost functional*

$$C(u) = g(x(T)) + C_0(u) = g(x(T)) + x^0(T).$$

If either $g(x) > b$, *i.e.,* $g(x)$ *is bounded below in* R^n *or* $g(x)$ *is convex in* R^n, *then there exists an optimal control.*

Now, as in §1, we call the controller $u(t)$ an extremal controller if and only if it steers the corresponding response $\hat{x}(t) = (x^0(t), x(t))$ to the boundary $\partial\hat{K}(T)$ of $\hat{K}(T)$ in R^{n+1}. Also we define the adjoint equation of the system by

$$\dot{\eta}^0 = 0, \text{ i. e., } \eta^0(t) = \eta_0 = \text{const.},$$

$$\dot{\eta}(t) = -\sum_{i=0}^{m} \eta(t + h_i)A_i(t + h_i) - \eta^0 \frac{\partial f^0}{\partial x(t)}\,(t, x(t)), \quad t_0 \leqq t \leqq T - h_m,$$

$$\dot{\eta}(t) = -\sum_{i=0}^{k} \eta(t + h_i)A_i(t + h_i) - \eta^0 \frac{\partial f^0}{\partial x(t)}\,(t, x(t)),$$

$$T - h_{k+1} \leqq t \leqq T - h_k, \quad k = 0, 1, \cdots, (m - 1),$$

and denote the solution by $\hat{\eta}(t) = (\eta_0, \eta(t))$. Assume $f^0(t, x) \in C^1$ on $[t_0, T]$; then the convexity of $f^0(t, x)$ implies that

$$f^0(t, x) - f^0(t, \bar{x}) \geqq \frac{\partial f^0}{\partial x}\,(t, \bar{x})(x - \bar{x}).$$

THEOREM 10. (Maximum principle) *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

$$\dot{x}^0(t) = f^0(t, x(t)) + h^0(t, u(t)),$$

with initial conditions $x(t) = \phi(t)$, $t_0 - h_m \leqq t \leqq t_0$, $x^0(t_0) = 0$, *and cost functional*

$$C_0(u) = x^0(T) = \int_{t_0}^{T} \{f^0(t, x(t)) + h^0(t, u(t))\}\, dt.$$

A control $\bar{u}(t)$ *with response* $\bar{x}(t)$ *is extremal if and only if there exists a nontrivial adjoint solution* $\hat{\eta}(t) = (\eta_0, \eta(t))$ *satisfying*

$$\eta^0(t) = \eta_0 < 0,$$

$$\eta(t) = -\sum_{i=0}^{m} \eta(t + h_i)A_i(t + h_i) - \eta_0 \frac{\partial f^0}{\partial x(t)}(t, \bar{x}(t)), \quad t_0 \leqq t \leqq T - h_m,$$

$$\eta(t) = -\sum_{i=0}^{k} \eta(t + h_i)A_i(t + h_i) - \eta_0 \frac{\partial f^0}{\partial x(t)}(t, \bar{x}(t)),$$

$$T - h_{k+1} \leqq t \leqq T - h_k, \quad k = 0, 1, \cdots, (m - 1),$$

and such that the maximum principle

$$\eta_0 h^0(t, \bar{u}(t)) + \eta(t)B(t)\bar{u}(t) = \max_{u \in R^r} [\eta_0 h^0(t, u) + \eta(t)B(t)u]$$

holds almost everywhere.

COROLLARY. *Consider the same system as in Theorem* 10. *Suppose the cost functional* $C(u)$ *is given by*

$$C(u) = g(x(T)) + \int_{t_0}^{T} \{f^0(t, x(t)) + h^0(t, u(t))\}\, dt,$$

where $g(x)$ *is convex and* $h(t, u)$ *is strictly convex; i.e., for* $0 < \lambda < 1$,

$$h^0(t, \lambda u_1 + (1 - \lambda)u_2) < \lambda h^0(t, u_1) + (1 - \lambda)h^0(t, u_2).$$

Then any two extremal controls steering the responses to the same boundary point of $\hat{K}(T)$ *must coincide almost everywhere. Furthermore, there exists a unique optimal control.*

THEOREM 11. *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

with

$$C(u) = g(x(T)) + \int_{t_0}^{T} \{f^0(t, x(t)) + h^0(t, u(t))\}\, dt.$$

Assume $g(x) \in C^1$ *is convex. Then there exists a solution* $x^*(t)$, $\eta^*(t)$ *of*

the system

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)\bar{u}(t, \eta(t)),$$

$$\dot{\eta}(t) = \frac{\partial f^0}{\partial x(t)}(t, x(t)) - \sum_{i=0}^{m} \eta(t + h_i)A_i(t + h_i), \qquad t_0 \leqq t \leqq T - h_m,$$

$$\dot{\eta}(t) = \frac{\partial f^0}{\partial x(t)}(t, x(t)) - \sum_{i=0}^{k} \eta(t + h_i)A_i(t + h_i),$$

$$T - h_{k+1} \leqq t \leqq T - h_k, \quad k = 0, 1, \cdots, (m-1),$$

with $x(t) = \phi(t), t_0 - h_m \leqq t \leqq t_0, \eta(T) = -\text{grad}\,(x(T))$. *Hence* $\bar{u}(t, \eta(t))$ *is defined by the maximum principle*

$$\eta(t)B(t)\bar{u}(t) - h^0(t, \bar{u}(t)) = \max_u [\eta(t)B(t)u - h^0(t, u)].$$

An optimal control is $u^*(t) = \bar{u}(t, \eta^*(t))$ *with the corresponding response* $x^*(t)$. *If* $h^0(t, u)$ *is strictly convex for each* t, *then the solution* $x^*(t)$, $\eta^*(t)$ *is unique, and* $u^*(t)$ *is the unique optimal control.*

5. Remarks on the restricted endpoint problem with integral convex cost functional. Consider the system

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t)$$

with cost functional

$$C(u) = g(x(T)) + C_0(u) = g(x(T)) + x^0(T),$$

as in Theorem 11. Assume $h^0(t, u)$ is strictly convex. Then, as in §3, consider the problem of steering the response to a closed convex fixed target set $G \subset R^n$ while minimizing the cost functional. For simplicity, assume the system is controllable. Then, as in §3, there exists a unique optimal control $u^*(t)$ on $[t_0, T]$. If $\hat{x}^*(T)$ is an inner point of $\hat{G} = G \times R^1 \subset R^{n+1}$, then the problem is the same as the unrestricted endpoint problem. Therefore we assume that $\hat{x}^*(T)$ is a boundary point of $\hat{G}$. Assume $C(u) = C_0(u) = x^0(T)$. Then, by the same argument as in §3, we can show that the minimum of $x^0(T)$ in $\hat{G} \cap \hat{K}(T)$ occurs at just one common boundary point $\hat{x}^*(T) = \partial\hat{G} \cap \partial\hat{K}(T)$. Therefore there exists a unique optimal control $u^*(t)$, and $u^*(t)$ can be obtained from any solution $x^*(t)$, $\eta^*(t)$ of the system

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u^*(t, \eta(t)),$$

$$\dot{\eta}(t) = \frac{\partial f^0}{\partial x}(t, x(t)) - \sum_{i=0}^{m} \eta(t + h_i)A_i(t + h_i), \qquad t_0 \leqq t \leqq T - h_m,$$

$$\dot{\eta}(t) = \frac{\partial f^0}{\partial x}(t, x(t)) - \sum_{i=0}^{k} \eta(\ + h_i)A_i(t + h_i)),$$

$$T - h_{k+1} \leqq t \leqq T - h_k, \quad k = 0, 1, \cdots, (m-1),$$

with $x(t) = \phi(t), t_0 - h_m \leqq t \leqq t_0, x(T) \in \partial G$, and $\eta(T)$ an interior normal to G at $x(T)$. In fact the optimal control $u^*(t)$ is given by $u^*(t) = u^*(t, \eta^*(t))$.

6. Compact controller restraint set. So far no restriction has been imposed on the controller $u(t)$ (except for measurability conditions). In this section, we restrict $u(t)$ to a compact convex restraint set $\Omega \subset R^r$.

Consider the system

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

$$\dot{x}^0(t) = f^0(t, x(t)) + h^0(t, u(t)),$$

with initial condition $x(t) = \phi(t)$, $t_0 - h_m \leqq t \leqq t_0$, $x^0(t_0) = 0$. Here $u(t) \subset \Omega$ is measurable on $[t_0, T]$, $A_i(t)$ and $B(t)$ are continuous real matrices on each compact interval, $f^0(t, x) \in C^1$, $h^0(t, u) \in C^0$ for all values of their arguments $x \in R^n$, $u \in R^r$. Assume that $f^0(t, x)$ and $h^0(t, u)$ are convex for each $t \in [t_0, T]$. The cost functional is

$$C(u) = g(x(T)) + x^0(T)$$

$$= g(x(T)) + \int_{t_0}^{T} \{f^0(t, x(t)) + h^0(t, u(t))\}\, dt,$$

where $g(x)$ is a convex C^1 function. For simplicity, assume that the system is normal[3] on $[t_0, T]$, and hence the set of attainability $K(T)$ in R^n is a strictly convex, compact set with nonempty interior. Here, also assume that Ω contains more than one point. Then we know that each boundary point of $K(T)$ is reached by a unique extremal controller [9].

Let $\hat{x}(t) = (x^0(t), x(t)) \in R^{n+1}$ be the response of the system, and let $\hat{K}(T)$ be the set of attainability in R^{n+1}. Then since Ω is compact, $\hat{K}(T)$ is bounded in R^{n+1}.

DEFINITION. Define $\hat{K}_v$ to be the *vertical saturation* of $\hat{K}(T)$, that is, $\hat{K}_v$ consists of all points $(x^0, x) \in R^{n+1}$ for which there exists a point $(y^0, x) \in \hat{K}(T)$ with $y^0 \leqq x^0$.

From the above definition, it is clear that the lower boundary of $\hat{K}_v$ is just the lower boundary of $\hat{K}(T)$.

[3] The system is normal if no component of $r_0 Y(t, T)B(t)$ is identically zero on any subinterval of $[t^0, T]$ for any nonzero constant n-vector r_0, [8].

THEOREM 12. *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t)$$

with cost functional

$$C_0(u) = x^0(T) = \int_{t_0}^{T} \{f^0(t, x(t)) + h^0(t, u(t))\}\, dt$$

and compact convex restraint set $\Omega \subset R^r$. *Then the vertical saturation* $\hat{K}_v$ *of* $\hat{K}(T)$ *is a closed convex set in* R^{n+1}. *Thus the lower boundary of* $\hat{K}_v$ *belongs to* $\hat{K}(T)$ *and this consists of a convex hypersurface defined over* $K(T) \subset R^n$.

Proof. Suppose $\hat{y}_k = (y_k^0, y_k) \in \hat{K}_v$ converges to (y^0, y) in R^{n+1}. Then there exists a sequence $u_k(t) \subset \Omega$ such that $x_k(T) = y_k$ and $x_k^0(T) \leqq y_k^0$, where $x_k(t)$ is the response corresponding to $u_k(t)$. Then there exists a subsequence, still called $u_k(t)$, which converges weakly to an admissible control $u(t) \subset \Omega$ with response $x(t)$ so that $x_k(t) \to x(t)$. It is easy to show that

$$y^0 \geqq \liminf_{k\to\infty} x_k^0(T) \geqq x^0(T).$$

Hence the response $(x^0(t), x(t))$ for $u(t)$ leads to the end point $(x^0(T), y)$ in $\hat{K}(T)$. Thus (y^0, y) is in $\hat{K}_v$ and so $\hat{K}_v$ is closed in R^{n+1}. If (y^0, y) is on the lower boundary of $\hat{K}(T)$, then $x^0(T) = y^0$ and $x(T) = y$ so that $u(t)$ steers the response to (y^0, y). Hence the lower boundary of $\hat{K}_v$ belongs to $\hat{K}(T)$. The proof of convexity is even more trivial.

COROLLARY. (Existence) *Consider the same system as in Theorem* 12 *with cost functional*

$$C(u) = g(x(T)) + \int_{t_0}^{T} \{f^0(t, x(t)) + h^0(t, u(t))\}\, dt.$$

Then there exists an optimal controller.

Proof. Since $g(x) + x^0$ increases monotonically in x^0 for each fixed x, the infimum of $g(x) + x^0$ is just the minimum of $g(x) + x^0$ on the lower boundary of $\hat{K}(T)$. Now the entire lower boundary of $\hat{K}(T)$ is compact, for $\hat{K}(T)$ is bounded; the lower boundary of $\hat{K}(T)$ is the lower boundary of $\hat{K}_v$, and $\hat{K}_v$ is closed. Since $g(x) + x^0$ is continuous in (x^0, x) on the compact lower boundary of $\hat{K}(T)$, there exists a minimum.

DEFINITION. A control $u(t)$ is called *extremal* if and only if it steers the response to the lower boundary of $\hat{K}(T)$.

THEOREM 13. (Maximal principle) *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

$$\dot{x}^0(t) = f^0(t, x(t)) + h^0(t, u(t)),$$

with cost functional

$$C_0(u) = x^0(T) = \int_{t_0}^{T} \{f^0(t, x(t)) + h^0(t, u(t))\}\, dt$$

and compact convex restraint set $\Omega \subset R^r$. *A control* $\bar{u}(t)$ *with response* $\hat{x}(t) = (\bar{x}^0(t), \bar{x}(t))$ *is extremal if and only if there exists a nontrivial adjoint solution* $\hat{\eta}(t) = (\eta_0, \eta(t))$ *satisfying*:

$$\dot{\eta}_0 = 0, \quad \eta_0 \leqq 0,$$

$$\dot{\eta}(t) = -\sum_{i=0}^{m} \eta(t + h_i)A_i(t + h_i) - \eta_0 \frac{\partial f^0}{\partial x}(t, \bar{x}(t)), \qquad t_0 \leqq t \leqq T - h_m,$$

$$\dot{\eta}(t) = -\sum_{i=0}^{k} \eta(t + h_i)A_i(t + h_i) - \eta_0 \frac{\partial f^0}{\partial x}(t, \bar{x}(t)),$$

$$T - h_{k+1} \leqq t \leqq T - h_k, \quad k = 0, 1, \cdots, (m - 1);$$

and satisfying the maximum principle:

$$\eta_0 h^0(t, \bar{u}(t)) + \eta(t)B(t)\bar{u}(t) = \max_{u \in \Omega} \{\eta_0 h^0(t, u) + \eta(t)B(t)u\}$$

almost everywhere on $[t_0, T]$.

Proof. Suppose $\bar{u}(t)$ with $\hat{x}(t) = (\bar{x}^0(t), \bar{x}(t))$ and $\hat{\eta}(t) = (\eta_0, \eta(t))$ satisfies the maximum principle. Then, as for Theorem 10, we have

$$\hat{\eta}(T)\hat{x}(T) \geqq \hat{\eta}(T)\hat{y}(t)$$

for all $\hat{y}(T) \in \hat{K}(T)$. Therefore, if $\eta_0 < 0$, then $\hat{x}(T)$ is on the lower boundary of $\hat{K}(T)$. If $\eta_0 = 0$, then it is on the lateral boundary of $\hat{K}_v$. But then $\bar{x}(T)$ is on the boundary of $K(T)$ in R^n; and since the system is normal, $\bar{u}(t)$ is the only control which steers the response to $\bar{x}(T) \in \partial K(T)$. Therefore $\hat{x}(T) = (\bar{x}^0(T), \bar{x}(T))$ is the unique point of $\hat{K}(T)$ with $x = \bar{x}(T)$. Thus $\hat{x}(T)$ is on the lower boundary of $\hat{K}(T)$ and so $\bar{u}(t)$ is extremal.

Conversely, assume $\bar{u}(t)$ is extremal so that $\hat{x}(T) = (\bar{x}^0(T), \bar{x}(T))$ is on the lower boundary of $\hat{K}(T)$. Let $\hat{\eta}(t) = (\eta_0, \eta(t))$ be the solution of the adjoint equation with $\hat{\eta}(T)$ exterior normal to the convex set $\hat{K}_v$ at $\hat{x}(T)$. Clearly $\eta_0 \leqq 0$. If $\eta_0 = 0$, then $x(T)$ is on $\partial K(T)$ and hence $\bar{u}(t)$ satisfies the maximum principle; and if $\eta_0 < 0$, then the proof of Theorem 10 applies.

COROLLARY. (Uniqueness) *Consider the same system as in Theorem* 13. *If* $g(x)$ *is convex and* $h^0(t, u)$ *is strictly convex for each* t, *then any two extremal controllers steering the responses to the same lower boundary point of* $\hat{K}(T)$ *must coincide almost everywhere. Furthermore, there exists a unique optimal control.*

Proof. Suppose $\hat{\eta} = (\eta_0, \eta)$ is an exterior normal vector to $\hat{K}_v$ at $\hat{x}(T) \in \partial\hat{K}(T)$. Let $u_1(t)$ and $u_2(t)$ be the two controls which steer the re-

sponses to $\hat{x}(T)$. If $\eta_0 = 0$, then since the system is normal, $u_1(t) = u_2(t)$ a.e. (see [3, Chap. 1]). If $\eta_0 < 0$, then by the corollary of Theorem 10, $u_1(t) = u_2(t)$ a.e. The uniqueness follows, as in Theorem 6, from the fact that $g(x) + x^0$ assumes its minimum on $\partial\hat{K}(T)$ at just one single point.

As in §4, define $u^* = u^*(t, \eta)$ by

$$-h^0(t, u^*) + \eta B(t)u^* = \max_{u \in \Omega} \{-h^0(t, u) + \eta B(t)u\}.$$

If $\eta(t)$ is continuous, then obviously $u^* = u^*(t, \eta)$ is admissible in Ω. Here we assumed $\eta_0 = -1$.

THEOREM 14. *Consider the system*

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u(t),$$

$$\dot{x}^0(t) = f^0(t, x(t)) + h^0(t, u(t)), \quad x^0(t_0) = 0,$$

with cost functional

$$C(u) = g(x(T)) + x^0(T)$$

and compact convex restraint set $\Omega \subset R^r$. Assume $g(x) \in C^1$ is convex in R^n. Then there exists a solution $x^*(t)$, $\eta^*(t)$ of the system

$$\dot{x}(t) = \sum_{i=0}^{m} A_i(t)x(t - h_i) + B(t)u^*(t, \eta(t)),$$

$$\dot{\eta}(t) = \frac{\partial f^0}{\partial x}(t, x(t)) - \sum_{i=0}^{m} \eta(t + h_i)A_i(t + h_i), \qquad t_0 \leqq t \leqq T - h_m,$$

$$\dot{\eta}(t) = \frac{\partial f^0}{\partial x}(t, x(t)) - \sum_{i=0}^{k} \eta(t + h_i)A_i(t + h_i),$$

$$T - h_{k+1} \leqq t \leqq T - h_k, \quad k = 0, 1, \cdots, (m-1).$$

An optimal control is $u^*(t) = u^*(t, \eta^*(t))$ with response $x^*(t)$.

If $h^0(t, u)$ is strictly convex for each t, then $x^*(t)$, $\eta^*(t)$ is unique, and $u^*(t)$ is the unique optimal control.

Proof. Let S_c be the hypersurface defined by $x^0 + g(x) = c$ in R^{n+1}. Then there is only one value $c = m$ so that S_m is tangent to $\hat{K}_v$, and m is the optimal cost. Also, S_m meets $\hat{K}_v$ at some point p on the boundary of $\hat{K}(T)$. The tangent hyperplane to S_m is also a supporting hyperplane to $\hat{K}_v$ at p, and hence we can choose a solution of the adjoint equation such that $\hat{\eta}^*(T) = (-1, \eta^*(T))$ is normal to this hyperplane. Let $u^*(t)$ be an extremal control such that $x^*(T) = p$. Then by Theorem 13, $u^*(t) \subset \Omega$ satisfies the maximum principle and

$$u^*(t) = u^*(t, \eta^*(t)).$$

If $h^0(t, u)$ is strictly convex, then S_m meets $\hat{K}(T)$ at just one point, and so $u^*(t)$, $x^*(t)$, $\eta^*(t)$ are unique.

7. An example. Consider the system given by the scalar equation

$$\dot{x}(t) = -x(t-1) + u(t)$$

with initial function $\phi(t) = 1$ on $[-1, 0]$. The problem is to find an optimal controller $u(t)$ on $[0, 2]$ steering the response $x(t)$ to the origin 0 at $T = 2$ while minimizing the cost functional

$$C(u) = \int_0^2 u(t)^2\, dt.$$

Here $u(t) \in R^1$ is bounded and measurable on $[0, 2]$.

The adjoint equation is:

$$\dot{\eta}(t) = \begin{cases} 0 & \text{if } 1 \leqq t \leqq 2, \\ \eta(t+1) & \text{if } 0 \leqq t \leqq 1. \end{cases}$$

By Theorem 6, the optimal controller $u(t)$ is of the form

$$u(t) = -\frac{1}{2\eta_0} U^{-1} B\eta(t) = \eta(t)$$

for $U = 1$ and $B = 1$. Take $\eta_0 = -1/2$ for simplicity, and let $\eta(2) = k$, then

$$\eta(t) = \begin{cases} k & \text{if } 1 \leqq t \leqq 2, \\ kt & \text{if } 0 \leqq t \leqq 1. \end{cases}$$

So

$$u(t) = \begin{cases} k & \text{if } 1 \leqq t \leqq 2, \\ kt & \text{if } 0 \leqq t \leqq 1. \end{cases}$$

The solution of the system corresponding to the controller $u(t)$ is:

$$x(t) = \begin{cases} \frac{1}{2}kt^2 - t + 1 & \text{if } 0 \leqq t \leqq 1, \\ -\frac{1}{6}kt^3 + \frac{1}{2}(1+k)t^2 - (2 - \frac{1}{2}k)t + (\frac{3}{2} - \frac{1}{3}k) & \text{if } 1 \leqq t \leqq 2. \end{cases}$$

Now $x(2) = 0$, for the target is the origin 0 and $T = 2$. From the solution above,

$$x(2) = -\tfrac{1}{2} + \tfrac{4}{3}k = 0.$$

Therefore, $k = \frac{3}{8}$ and the optimal control $u(t)$ is given by

$$u(t) = \begin{cases} \frac{3}{8}t & \text{if } 0 \leqq t \leqq 1, \\ \frac{3}{8} & \text{if } 1 \leqq t \leqq 2. \end{cases}$$

The corresponding cost functional $C(u)$ is:

$$C(u) = \int_0^2 u(t)^2\,dt = \frac{3}{16}.$$

REFERENCES

[1] R. Bellman and K. L. Cooke, *Differential-Difference Equations*, Academic Press, New York, 1963.

[2] N. H. Choksy, *Time lag systems—a bibliography*, IRE Trans. Automatic Control, AC5 (1960), pp. 66–70.

[3] D. H. Chyung, *Optimal control systems with time delays*, Ph.D. Thesis, University of Minnesota, Minneapolis, 1965.

[4] D. H. Chyung and E. B. Lee, *Control of linear time delay systems with essentially linear cost functionals*, Differential Equations and Dynamical Systems, J. P. LaSalle, ed., Academic Press, New York, 1966.

[5] A. Friedman, *Optimal control for hereditary processes*, Arch. Rational Mech. Anal., 15 (1963), pp. 396–416.

[6] G. L. Haratisville, *The maximum principle in the theory of optimal processes involving delay*, Dokl. Akad. Nauk SSSR, 136 (1961), pp. 39–42.

[7] R. E. Kalman, *On the general theory of control systems*, IFAC-I, 4 (1960), pp. 2020–2030.

[8] J. P. LaSalle, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. V, Princeton University Press, Princeton, 1960, pp. 1–24.

[9] E. B. Lee and L. Markus, *Foundations of Optimum Control Theory*, John Wiley, New York, 1966.

[10] ———, *Optimal control of nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[11] M. N. Oguztoreli, *A time optimal control problem for systems described by differential difference equations*, this Journal, 1 (1963), pp. 290–310.

[12] ———, *Time-Lag Control Systems*, Academic Press, New York, 1966.

[13] L. S. Pontryagin, V.G. Boltyanskii, R. V. Gamkrelidze and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[14] V. M. Popov and A. Khalanay, *A problem in the theory of time delay optimum systems*, Automat. Remote Control, 24 (1963), pp. 129–131.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

LINEAR OPTIMAL CONTROL PROBLEMS*

B. N. PSHENICHNIY†

The most completely developed area of optimal control theory, from the viewpoint of developing effective computational algorithms, is that of linear optimal control problems. If we classify the papers in this area according to their conceptual approach, they can be subdivided into three basis classes:

(1) methods utilizing the maximum principle, which consist in finding the initial values of the adjoint system,

(2) methods utilizing the method of steepest descent in control space,

(3) methods based on the theory of moments.

The present article is concerned only with the first class of methods. The first paper in this direction was written by L. W. Neustadt [1]. This work was subsequently developed in the papers of L. W. Neustadt himself [2] and of others [3], [4], [5]. The proof of the convergence of these algorithms was carried out in [4] and [5] by certain geometric arguments. This proof made it possible to draw definite conclusions on the nature of the convergence of the iterative process by relating, at each step of the process, the estimate for the stepwidth and the magnitude of the increment of the function being optimized to the geometric properties of the set of attainability.

If one carefully examines the analysis in [1]–[5], one sees that the convergence of the algorithms presented therein depends only partially on the particular linear optimal control problems, and is based on a single very general assumption. Therefore, these algorithms are, in fact, applicable for the solution of a considerably wider class of problems [6]. In this connection, the close relationship of these algorithms with the Kuhn-Tucker theorem [7] and the dual methods of solving extremal problems [8] was clarified.

Finally, the theoretical analysis carried out in developing these algorithms made it possible to construct a general theory of convex programming [9] by showing how a differential form of the Kuhn-Tucker theorem could be formulated, and by relating this theorem to important results in the general theory of extremal problems obtained in [10] and [11].

This article represents a brief survey of work carried out by the author. Naturally, it was not possible to include everything or to give complete

* Submitted in Russian on February 1, 1966. The translation into English was carried out by N. H. Choksy and J. R. La Frieda and was edited by L. W. Neustadt. The translation was supported in part through a grant-in-aid by the National Science Foundation.

† Institute of Cybernetics, Kiev, USSR.

proofs of all theorems. Therefore, in the subsequent presentation, many proofs are either omitted, or are only outlined in their basic features. The reader will find complete proofs in the literature, to which the corresponding references are given.

1. The Kuhn-Tucker theorem and necessary optimality conditions. The Kuhn-Tucker theorem [7] can serve as a source of necessary and sufficient conditions in linear optimal control problems. This theorem may be formulated in a quite general form as follows:

Let B be a Banach space, and let $\mu_i(x)$, $i \in I_1 = \{1, \cdots, m\}$, be convex continuous functionals defined on B, let $\mu_i(x)$, $i \in I_2 = \{m+1, \cdots, n\}$, be linear continuous functionals defined on B, and let X be a closed convex subset of B.

THEOREM 1.1. *In order that x^0 yield a minimum for the convex functional $\mu_0(x)$, subject to the conditions*

$$\begin{aligned} \mu_i(x) &\leqq 0 \quad \text{for} \quad i \in I_1, \\ \mu_i(x) &= 0 \quad \text{for} \quad i \in I_2, \\ x &\in X, \end{aligned} \tag{1.1}$$

it is necessary that there exist a nonzero vector $\psi^0 \in E^{n+1}$ such that

$$\psi_0^0 \mu_0(x^0) + \sum_{i=1}^{n} \psi_i^0 \mu_i(x^0) \leqq \psi_0^0 \mu_0(x) + \sum_{i=1}^{n} \psi_i^0 \mu_i(x) \quad \text{for all } x \in X, \tag{1.2}$$

and, in addition, that

$$\begin{aligned} \psi_i^0 &\geqq 0 \quad \text{and} \quad \psi_i^0 \mu_i(x^0) = 0 \quad \text{for} \quad i \in I_1, \\ \psi_0^0 &\geqq 0. \end{aligned} \tag{1.3}$$

If $\psi_0^0 > 0$, then these conditions are also sufficient.

The proof of this theorem can be found in [7] and we shall not dwell on it here. It is interesting to consider the application of this theorem to optimal control problems; however, in order to do this we must reformulate Theorem 1.1 into a differential form. Let $\mu(x)$ be a convex functional, defined on B, which is continuous and bounded on every bounded subset of B. Let us define a set of support functionals to $\mu(x)$ at x^0 as follows:

$$M(x^0) = \{x^*: x^* \in B^*, \mu(x) - \mu(x^0) \geqq x^*(x - x^0) \text{ for all } x \in B\}. \tag{1.4}$$

Here, B^* is the conjugate space of B. It is possible [9] to prove the following theorem.

THEOREM 1.2. *$M(x^0)$ is nonempty, bounded, convex and closed in the weak* topology of B^*.*

The convexity and weak* closure are obvious here. To prove that $M(x^0)$

is nonempty, one uses methods which are the same as the ones used to prove the following theorem.

THEOREM 1.3. *Let $e \in B$ and let*

$$\frac{\partial \mu(x^0)}{\partial e} = \lim_{\lambda \to 0+} \frac{\mu(x^0 + \lambda e) - \mu(x^0)}{\lambda}. \tag{1.5}$$

Then the limit in (1.5) *exists for every e, and*

$$\frac{\partial \mu(x^0)}{\partial e} = \max_{x^* \in M(x^0)} x^*(e). \tag{1.6}$$

Proof. Since $\mu(x^0 + \lambda e)$ is a convex function of the single scalar argument λ, the existence of $\partial\mu/\partial e$ follows from [12]. In [12], it is also shown that

$$\varphi(\lambda) = \frac{\mu(x^0 + \lambda e) - \mu(x^0)}{\lambda}$$

is a nondecreasing function of λ. Let us now verify (1.6). By definition, for every $x^* \in M(x^0)$, we have, for $\lambda > 0$,

$$\frac{\mu(x^0 + \lambda e) - \mu(x^0)}{\lambda} \geqq x^*(e).$$

Therefore,

$$\frac{\partial \mu}{\partial e} \geqq \max_{x^* \in M(x^0)} x^*(e).$$

Let us assume that, for some e^0,

$$\frac{\partial \mu}{\partial e^0} > \max_{x^* \in M(x^0)} x^*(e^0). \tag{1.7}$$

In the product space $B_1 = E^1 \times B$, consider the convex set

$$Z = \{(\alpha, x) : \alpha \geqq \mu(x)\}$$

and the ray

$$L = \left\{(\alpha, x) : \alpha = \mu(x^0) + \lambda \frac{\partial \mu}{\partial e^0},\ x = x^0 + \lambda e^0,\ \lambda \geqq 0\right\}.$$

Since $\varphi(\lambda)$ is nondecreasing, $\partial\mu/\partial e^0 \leqq \varphi(\lambda)$ for $\lambda \geqq 0$, and

$$\mu(x^0 + \lambda e^0) \geqq \mu(x^0) + \lambda \frac{\partial \mu}{\partial e^0}.$$

From this it follows that Z and L have no common interior points, and, since Z has interior points, Z and L can be separated; i.e., there exist a number c and a functional $y^* \in B^*$ (c and y^* do not both vanish) such

that

$$(1.8)\quad c\alpha + y^*(x) \geqq c\left(\mu(x^0) + \lambda\frac{\partial\mu}{\partial e^0}\right) + y^*(x^0 + \lambda e^0) \quad \text{for all} \quad x \in B,$$

whenever $\alpha \geqq \mu(x)$ and $\lambda \geqq 0$. It is not difficult to show that $c > 0$. Setting $\alpha = \mu(x)$ and $\lambda = 0$, we obtain

$$\mu(x) - \mu(x^0) \geqq -\frac{1}{c}\, y^*(x - x^0) \quad \text{for all} \quad x \in B,$$

i.e., $x^* = -(1/c)y^* \in M(x^0)$. (By the same token, we have proved that $M(x^0)$ is nonempty.)

Further, it follows from (1.8) that

$$(1.9)\quad \mu(x) - \mu(x^0) \geqq x^*(x - x^0) + \lambda\left[\frac{\partial\mu}{\partial e^0} - x^*(e^0)\right] \quad \text{for all} \quad x \in B \quad \text{and} \quad \lambda \geqq 0.$$

Setting $x = x^0$ in (1.9), we have $\partial\mu/\partial e^0 \leqq x^*(e^0)$, which contradicts (1.7). This completes the proof of the theorem.

Theorem 1.3 can serve as a basis for the development of numerical methods for the minimization of nonsmooth, convex functionals. However, it is most useful in that it makes it possible to actually construct the set $M(x^0)$. The detailed questions that arise in the actual construction of $M(x^0)$ are considered in [9]. Here we shall cite, without proof, only certain facts which are necessary for the subsequent presentation.

THEOREM 1.4. (a) *If* $\mu(x) = c_1\mu_1(x) + c_2\mu_2(x)$, *where* $c_1 \geqq 0$ *and* $c_2 \geqq 0$, *then* $M(x^0) = c_1M_1(x^0) + c_2M_2(x^0)$, *where* $M_i(x^0)$, $i = 1, 2$, *is the set of support functionals for the function* $\mu_i(x)$.

(b) *If* $\mu(x) = \max_{1 \leqq i \leqq m} \mu_i(x)$, *then*

$$M(x^0) = \{x^*: x^* = \sum_{i \in I(x^0)} \lambda_i x_i^*,\ x_i^* \in M_i(x^0),\ \sum_{i \in I(x^0)} \lambda_i = 1,\ \lambda_i \geqq 0\},$$

where

$$I(x^0) = \{i: 1 \leqq i \leqq m,\ \mu_i(x^0) = \mu(x^0)\}.$$

(c) *If* $\mu(x) = \max_{x^* \in M} x^*(x)$, *and* M *is a bounded, weak* closed, convex set in* B^*, *then*

$$M(x^0) = \{x^*: x^* \in M,\ x^*(x^0) = \mu(x^0)\}.$$

As an example of an application of Theorem 1.4, consider the functional $\mu(x) = \| x \|$. It is well-known [13] that $\| x \| = \max_{x^* \in S^*} x^*(x)$, where S^* is the unit sphere in B^*. Applying Theorem 1.4(c), we see that

$$M(x^0) = \{x^*: \| x^* \| \leqq 1,\ x^*(x^0) = \| x^0 \|\}.$$

Thus, in this case, $M(x^0)$ coincides with the set of extremal functionals of x^0.

Let us now turn to the question of what are the conditions for a minimum of the convex functional $\mu(x)$ in a closed convex region Ω.

Let $x^0 \in \Omega$, and let

$$\Gamma_{x^0} = \{e: e \in B, x^0 + \lambda e \in \Omega \quad \text{for some} \quad \lambda > 0\}.$$

It is clear that Γ_{x^0} is the cone of directions which lead into the interior of Ω. Let $\Gamma^*_{x^0}$ denote the dual cone [14] of the cone Γ_{x^0}.

THEOREM 1.5. *The convex functional $\mu(x)$ attains its minimum at the point $x^0 \in \Omega$ if and only if*

$$\Gamma^*_{x^0} \cap M(x^0) \neq \varnothing. \tag{1.10}$$

Proof. Necessity. Let $\mu(x^0) \leqq \mu(x)$ for all $x \in \Omega$, and suppose that (1.10) does not hold. Since $\Gamma^*_{x^0}$ and $M(x^0)$ are weak* closed and convex, they are regularly convex [7], [13]. In addition, $M(x^0)$ is bounded and hence weak* compact. From this it follows that the set $\Gamma^*_{x^0} - M(x^0)$ is convex and weak* closed [13]. Therefore it is also regularly convex [7], and there exists an element $e \in B$ such that

$$\inf_{x^* \in \Gamma^*_{x^0}} x^*(e) \geqq \delta + \max_{x^* \in M(x^0)} x^*(e) \quad \text{for some} \quad \delta > 0. \tag{1.11}$$

But, since $\Gamma^*_{x^0}$ is a cone, $\inf x^*(e) = 0$. From this it follows [14] that $e \in \bar{\Gamma}_{x^0}$, and that

$$\frac{\partial \mu}{\partial e} \leqq -\delta. \tag{1.12}$$

It is easy to show, on the basis of (1.12), that there exists a direction leading into Ω along which μ is decreasing, which contradicts our hypothesis.

Sufficiency. Let (1.10) hold. Then there exists an x_0^* such that $x_0^* \in \Gamma^*_{x^0}$, and, in particular,

$$x_0^*(x - x^0) \geqq 0 \quad \text{for all} \quad x \in \Omega,$$

and such that $x_0^* \in M(x^0)$.

But, by definition of $M(x^0)$, we have

$$\mu(x) - \mu(x^0) \geqq x_0^*(x - x^0) \quad \text{for all} \quad x. \tag{1.13}$$

Comparing (1.13) with the previous inequality, we see that the proof of the theorem is complete.

The following corollary is an immediate consequence of Theorem 1.5.

COROLLARY. *The convex functional $\mu(x)$ attains its minimum on Ω at the point $x^0 \in \Omega$ if and only if there exists a functional $x_0^* \in M(x^0)$ such*

that

$$x_0^*(x^0) \leqq x_0^*(x) \quad \textit{for all} \quad x \in \Omega.$$

Proof. If the point $x^0 \in \Omega$ provides $\mu(x)$ with its minimum on Ω, then it is obvious that for x_0^* it is sufficient to take the functional appearing in (1.13). Conversely, if $x_0^* \in M(x^0)$ such that $x_0^*(x - x^0) \geqq 0$ for all $x \in \Omega$, then our result follows from (1.13).

The results we have obtained now enable us to formulate the differential form of Theorem 1.1.

THEOREM 1.6. (The differential form of the Kuhn-Tucker theorem). *In order that x^0 yield a minimum for the convex functional $\mu_0(x)$, subject to conditions* (1.1), *it is necessary that there exist a vector $\psi^0 \in E^{n+1}$ and functionals $x_i^* \in M_i(x^0)$, $i = 0, 1, \cdots, n$, such that*

$$(1.14) \quad \psi_0^0 x_0^*(x^0) + \sum_{i=1}^{n} \psi_i^0 x_i^*(x^0) \leqq \psi_0^0 x_0^*(x) + \sum_{i=1}^{n} \psi_i^0 x_i^*(x) \quad \textit{for all } x \in X,$$

and

$$(1.15) \qquad \psi_i^0 \geqq 0 \quad \textit{and} \quad \psi_i^0 \mu_i(x^0) = 0 \quad \textit{for} \quad i \in I_1, \quad \psi_0^0 \geqq 0.$$

if $\psi_0^0 > 0$, then (1.14) *and* (1.15) *are also sufficient.*

Theorem 1.6 follows immediately from Theorem 1.1, the corollary to Theorem 1.5, and Theorem 1.4 (a).

In [9] techniques for operating with convex functionals are developed in more detail. Some examples are also considered therein. Here we shall consider the possibility of applying the theory developed above in order to obtain necessary conditions for one optimal control problem with linear systems.

Let us consider an object whose behavior is described by the system of differential equations

$$(1.16) \qquad \frac{dx}{dt} = Ax + Bu, \quad 0 \leqq t \leqq T,$$

$$x(0) = x^0,$$

where $x \in E^n$, A is an $n \times n$ matrix, and B is an $n \times r$ matrix. The control $u(t)$ is a measurable function, and $u(t)$, for each t, belongs to a convex, bounded, closed set U, where $U \subset E^r$. Such controls will be called admissible.

From among all the admissible controls $u(t)$, we are to find one whose corresponding trajectory $x(t)$ satisfies the conditions

$$(1.17) \qquad x(T) - x^1 = 0,$$

$$\mu_1(x(t)) = \max_{0 \leqq t \leqq T} g_1(x(t)) \leqq 0, \tag{1.18}$$

and such that the functional

$$\mu_0(x(t)) = \max_{0 \leqq t \leqq T} g_0(x(t))$$

attains its minimum value. Here $g_i(x)$, $i = 0, 1$, are convex, twice continuously differentiable functions. We note that similar problems were considered in [10], [15], [16].

Let us study the functionals $\mu_i(x(t))$, $i = 0, 1$. Since, for any measurable function $u(t)$, $x(t)$ is a continuous vector function, the $\mu_i(x(t))$ can be considered to be functionals on the space C of continuous functions. Let M^* denote the set of all functions $\sigma(t)$, with $\sigma(0) = 0$ and $\sigma(T) = 1$, which are nondecreasing over the interval $[0, T]$. Every such function defines a measure on $[0, T]$. In what follows, we shall often identify $\sigma(t)$ with the corresponding measure.

Then it is obvious that

$$\mu_i(x(t)) = \max_{\sigma \in M^*} \int_0^T g_i(x(t))\, d\sigma(t), \qquad i = 0, 1. \tag{1.19}$$

Appealing to Theorem 1.4(c), we conclude that

$$\mu_i(x(t)) - \mu_i(x^0(t)) \geqq \int_0^T [g_i(x(t)) - g_i(x^0(t))]\, d\sigma(t) \tag{1.20}$$

for every continuous function $x(t)$ and for those (and only those) $\sigma(t)$ such that $\sigma(t) \in M^*$ and

$$\int_0^T g_i(x^0(t))\, d\sigma(t) = \mu_i(x^0(t)) = \max_{0 \leqq t \leqq T} g_i(x^0(t)). \tag{1.21}$$

It follows from (1.21) that the measure defined by $\sigma(t)$ must be concentrated on the set of those t_0 for which $g_i(x^0(t_0)) = \max_{0 \leqq t \leqq T} g_i(x^0(t))$. Making use of the convexity and the continuous differentiability of $g_0(x)$ and $g_1(x)$, we conclude, by virtue of (1.20), that

$$\mu_i(x(t)) - \mu_i(x^0(t)) \geqq \int_0^T (\partial g_i(x^0(t)), x(t) - x^0(t)) d\sigma(t), \tag{1.22}$$

where

$$\partial g_i(x) = \left\{ \frac{\partial g_i}{\partial x_1}, \cdots, \frac{\partial g_i}{\partial x_n} \right\}.$$

Thus, we can now convince ourselves that the set $M_i^*(x^0(t))$ of all support functionals to $\mu_i(x(t))$ at the point $x^0(t)$ (see (1.4)) includes all

functionals of the form:

$$x^*(x) = \int_0^T (\partial g_i(x^0(t)), x(t))\, d\sigma(t), \tag{1.23}$$

where $\sigma(t) \in M^*$, and the measure defined by σ is concentrated on the set of t_0 for which $g_i(x^0(t_0)) = \max_{0 \leqq t \leqq T} g_i(x^0(t))$. It is possible to show that these functionals exhaust the entire set $M_i^*(x^0(t))$.

Let us now turn to the derivation of the necessary conditions for optimality. To make it possible to apply Theorem 1.6, we let X denote the set of all trajectories $x(t)$, $0 \leqq t \leqq T$, obtained from all possible admissible controls. Then the above formulated problem reduces to that of finding a trajectory $x^0(t) \in X$ which minimizes the functional $\mu_0(x(t))$ subject to conditions (1.17) and (1.18). Applying Theorem 1.6, with due regard to the actual form of the support functional, we arrive at the following conclusion.

In order that the trajectory $x^0(t)$ minimizes the functional $\mu_0(x(t))$ subject to conditions (1.17) and (1.18), it is necessary that there exist numbers $\psi_0 \geqq 0$ and $\psi_1 \geqq 0$, and a vector $\psi \in E^n$ such that the trajectory $x^0(t)$ minimizes on X the linear functional

$$\psi_0 \int_0^T (\partial g_0(x^0(t)), x(t))\, d\sigma_0(t) + \psi_1 \int_0^T (\partial g_1(x^0(t)), x(t))\, d\sigma_1(t) + (\psi, x(T) - x^1), \tag{1.24}$$

where $\sigma_0(t)$ and $\sigma_1(t)$ are nondecreasing functions with $\sigma_i(0) = 0$ and $\sigma_i(T) = 1$ for $i = 0, 1$, and

$$\int_0^T g_i(x^0(t))\, d\sigma_i(t) = \mu_i(x^0(t)), \qquad i = 0, 1, \tag{1.25}$$

$$\psi_1 \mu_1(x^0(t)) = 0.$$

If $\psi_0 > 0$, then this condition is also sufficient.

Let us now make use of the fact that the trajectory $x(t)$ of (1.16) can be written in the form

$$x(t) = \Phi(t)x^0 + \Phi(t) \int_0^t \Phi^{-1}(\tau) B u(\tau)\, d\tau,$$

where the matrix $\Phi(t)$ satisfies the system of equations $d\Phi/dt = A\Phi$ and $\Phi(0) = I$, where I is the identity matrix; and let us substitute this expression into (1.24).

Omitting the rather tedious, but quite elementary, calculations which

involve changing the order of integration in the first and second terms in (1.24), we conclude that the expression in (1.24) is equal to

$$\int_0^T (\psi(\tau), Bu)\, d\tau + \cdots, \tag{1.26}$$

where

$$\psi(\tau) = \Phi^{-1*}(\tau)\left[\psi_0 \int_\tau^T \Phi^*(t)\partial g_0(x^0(t))\, d\sigma_0(t) + \psi_1 \int_\tau^T \Phi^*(t)\partial g_1(x^0(t))\, d\sigma_1(t) + \Phi^*(T)\psi\right], \tag{1.27}$$

and the dots denote terms which are independent of the control.

It is clear from (1.26) that in order that $u^0(\tau)$ minimize (1.26), and hence that the corresponding trajectory $x^0(t)$ minimize (1.24), it is necessary and sufficient that the following condition hold almost everywhere:

$$(\psi(\tau), Bu^0(\tau)) = \min_{v \in U} (\psi(\tau), Bv), \tag{1.28}$$

where $u^0(t)$ is the control corresponding to the trajectory $x^0(t)$.

The obtained results can now be formulated in the form of the following theorem.

THEOREM 1.7. *In order that the admissible control* $u^0(t)$, $0 \leqq t \leqq T$, *be the solution of the problem formulated above, it is necessary that there exist constants* $\psi_0 \geqq 0$ *and* $\psi_1 \geqq 0$, *a vector* $\psi \in E^n$, *and functions* $\sigma_0(t)$ *and* $\sigma_1(t)$ *which are nondecreasing on the interval* $[0, T]$, *such that*:

(1) $\sigma_i(0) = 0$, $\sigma_i(T) = 1$, $i = 0, 1$,

(2) $\int_0^T g_i(x^0(t))\, d\sigma_i(t) = \mu_i(x^0(t))$, $i = 0, 1$,

(3) $\psi_1\mu_1(x^0(t)) = 0$,

and that (1.28) *be satisfied almost everywhere, where the function* $\psi(\tau)$ *is given by* (1.27).

If $\psi_0 > 0$, *then the conditions are also sufficient.*

Note. Let us point out certain properties of the function $\psi(\tau)$. At each point τ where the $\sigma_i(\tau)$, $i = 0, 1$, are differentiable, i.e., almost everywhere, $\psi(\tau)$ has a derivative which is given by the formula

$$\frac{d\psi}{d\tau} = -A^*\psi - \psi_0\partial g_0(x^0(\tau))\frac{d\sigma_0}{d\tau} - \psi_1\partial g_1(x^0(\tau))\frac{d\sigma_1}{d\tau}. \tag{1.29}$$

At the points of discontinuity of the $\sigma_i(\tau)$, $\psi(\tau)$ also undergoes a discon-

tinuity which is given by the formula

$$\psi(\tau - 0) - \psi(\tau + 0) = \psi_0 \Delta\sigma_0(\tau)\partial g_0(x^0(\tau)) + \psi_1 \Delta\sigma_1(\tau)\partial g_1(x^0(\tau)), \tag{1.30}$$

where $\Delta\sigma_i(\tau)$ is the magnitude of the jump of $\sigma_i(\tau)$ at the point τ.

It should be mentioned that one must not consider (1.29) as a system of equations that defines the function $\psi(\tau)$, for $\psi(\tau)$ is discontinuous, and is not, in the generally accepted sense, a solution of (1.29).

Finally, it is not difficult to see that any additional conditions at the right-hand endpoint of the trajectory, or any additional phase constraints of the same type as those considered above, can be taken into account in an analogous way, without any particular complications.

2. Algorithms. From the viewpoint of developing computational algorithms, the main content of Theorem 1.1 is in essence that the original problem with constraints can be reduced to solving a problem without constraints, once we determine the constants ψ_i, $i = 0, \cdots, n$, whose existence is guaranteed by Theorem 1.1.

The algorithm described below is based on successively approximating these constants.

Let us consider the problem of minimizing $\mu_0(x)$, where $x \in B$, subject to the constraints

$$\begin{gathered} \mu_i(x) \leqq 0 \quad \text{for} \quad i \in I_1, \\ \mu_i(x) = 0 \quad \text{for} \quad i \in I_2, \\ x \in X. \end{gathered} \tag{2.1}$$

We note that for the time being no assumptions have been made relative to the functionals $\mu_i(x)$ or the set X. In particular, the $\mu_i(x)$ are not assumed to be convex.

Let us define the set $M \subset E^{n+1}$ of vectors $z \in E^{n+1}$, with components z_i, $i = 0, 1, \cdots, n$, as follows:

$$M = \{z : z_i = \mu_i(x),\ i = 0, 1, \cdots, n, \quad \text{for some} \quad x \in X\}. \tag{2.2}$$

It is now possible to formulate the requirements which we shall impose on the functionals $\mu_i(x)$ and the set X, indirectly, through the conditions which we shall impose on M.

Basic Assumption. (a) *M is a bounded and closed subset of E^{n+1}*; (b) *for any $\psi \in E^{n+1}$ that satisfies the conditions $\psi_0 = -1$ and $\psi_i \leqq 0$ for $i \in I_1$, the maximum value of the inner product (ψ, z) for $z \in M$, is attained at a unique point $z(\psi) \in M$.*

We note that the original problem (2.1) can be reformulated as follows:

from among all $z \in M$ that satisfy the conditions

$$\begin{aligned} z_i &\leqq 0 \quad \text{for} \quad i \in I_1, \\ z_i &= 0 \quad \text{for} \quad i \in I_2, \end{aligned} \tag{2.3}$$

find one for which the coordinate z_0 is minimal.

The carry over from the original problem (2.1) to the problem (2.3) is convenient at a certain stage, since it permits us, at that time, to digress from the actual properties of the functionals $\mu_i(x)$ and to construct a general scheme for the algorithm. In the process of testing, in a given specific problem, for the conditions under which the algorithm can be applied, to verify whether or not the conditions of the basic assumption are satisfied.

We shall now temporarily digress from the optimization problem and shall examine the function

$$\varphi(\psi) = \max_{z \in M} (\psi, z). \tag{2.4}$$

We shall continue to assume that M is closed and bounded.

THEOREM 2.1. *The function $\varphi(\psi)$ is continuous and convex, and its derivative in any direction $e \in E^{n+1}$ is given by the formula*

$$\frac{\partial \varphi(\psi^0)}{\partial e} = \max_{z \in M(\psi^0)} (e, z), \tag{2.5}$$

where

$$M(\psi^0) = \{z : z \in M, (\psi^0, z) = \varphi(\psi^0)\}. \tag{2.6}$$

Proof. The convexity of $\varphi(\psi)$ is easily verified. By definition,

$$(\psi, z) = \varphi(\psi) \geqq (\psi, z^0),$$

$$(\psi^0, z) \leqq \varphi(\psi^0) = (\psi^0, z^0),$$

whenever

$$z \in M(\psi) \quad \text{and} \quad z^0 \in M(\psi^0).$$

Subtracting these two inequalities, we obtain

$$(\psi - \psi^0, z) \geqq \varphi(\psi) - \varphi(\psi^0) \geqq (\psi - \psi^0, z^0), \tag{2.7}$$

from which the continuity of $\varphi(\psi)$ follows. Now let $\psi = \psi^0 + \lambda e$, where $\lambda \geqq 0$. Substituting this expression into (2.7), we obtain

$$\lambda(e, z) \geqq \varphi(\psi^0 + \lambda e) - \varphi(\psi^0) \geqq \lambda(e, z^0).$$

Since z and z^0 are arbitrary elements of $M(\psi)$ and $M(\psi^0)$, respectively, we

have

$$\max_{z\in M(\psi)} (e, z) \geqq \frac{\varphi(\psi^0 + \lambda e) - \varphi(\psi^0)}{\lambda} \geqq \max_{z\in M(\psi^0)} (e, z). \tag{2.8}$$

For any $\delta > 0$, it is not difficult to show that if λ is sufficiently small, then $M(\psi^0 + \lambda e)$ is contained in the δ-neighborhood of $M(\psi^0)$. Therefore,

$$\max_{z\in M(\psi)} (e, z) \leqq \max_{z\in M(\psi^0)} (e, z) + \omega(\lambda),$$

where $\omega(\lambda) \to 0$ as $\lambda \to 0$.

The theorem now follows from the preceding inequality and (2.8).

COROLLARY. *If $M(\psi^0)$ consists of the single point $z(\psi^0)$, then*

$$\frac{\partial\varphi(\psi^0)}{\partial e} = (e, z(\psi^0)). \tag{2.9}$$

Moreover, if $M(\psi)$ consists of a single point $z(\psi)$ for ψ in a certain region, then $\varphi(\psi)$ is continuously differentiable in this region.

Equation (2.9) is a consequence of (2.5), and the continuous differentiability of $\varphi(\psi)$ follows from (2.9) and the previously noted fact that $M(\psi)$ is contained in the δ-neighborhood of $M(\psi^0)$ if $\| \psi - \psi^0 \|$ is sufficiently small.

The next theorem shows that the problem of minimizing z_0 subject to conditions (2.3) can be reduced to that of minimizing $\varphi(\psi)$ over the region Ω, where

$$\Omega = \{\psi: \psi \in E^{n+1}, \psi_0 = -1, \psi_i \leqq 0 \quad \text{for} \quad i \in I_1\}.$$

THEOREM 2.2. *Let the set M satisfy the conditions of the Basic Assumption. Then, if $\psi^0 \in \Omega$ and $\varphi(\psi^0) \leqq \varphi(\psi)$ for all $\psi \in \Omega$, we have that*

$$\begin{aligned} z_i(\psi^0) &\leqq 0 \quad \textit{for} \quad i \in I_1, \\ z_i(\psi^0) &= 0 \quad \textit{for} \quad i \in I_2, \\ \psi_i^0 z_i(\psi^0) &= 0 \quad \textit{for} \quad i \in I_1, \end{aligned} \tag{2.10}$$

and, for every vector $z \in M$ satisfying (2.3), $z_0 \geqq z_0(\psi^0)$.

Proof. By virtue of the basic assumption and the corollary to Theorem 2.1, the derivative of $\varphi(\psi)$ in any direction e, evaluated at the point ψ^0, is given by the formula

$$\frac{\partial\varphi(\psi^0)}{\partial e} = (e, z(\psi^0)).$$

Since $\varphi(\psi)$ is a convex function, it follows that ψ^0 will yield a minimum for the function $\varphi(\psi)$ in Ω if and only if

$$(e, z(\psi^0)) \geqq 0 \tag{2.11}$$

for all e such that $\psi^0 + \lambda e \in \Omega$ for some $\lambda > 0$. But one can easily convince oneself that conditions (2.11) and (2.10) are equivalent. Further, if $z \in M$ satisfies conditions (2.3), then

$$\varphi(\psi^0) = -z_0(\psi^0) + \sum_{i=1}^{n} \psi_i^0 z_i(\psi^0) = -z_0(\psi^0) \geqq -z_0 + \sum_{i=1}^{n} \psi_i^0 z_i \geqq -z_0,$$

i.e., $z_0(\psi^0) \leqq z_0$, which was to be proved.

In the preceding chain of inequalities, we in turn made use of (2.10), the definition of $\varphi(\psi)$, and (2.3).

Thus, if one finds the minimum of $\varphi(\psi)$ in Ω, then at the same time the original optimization problem has been solved. But, as was shown above, $\varphi(\psi)$ is a smooth function in the region under consideration, and its gradient, on the basis of the corollary to Theorem 2.1, coincides with $z(\psi)$. Therefore, to obtain the minimum of φ, one of the methods of descent [18], [19] may be applied.

Note. In the proof of Theorem 2.2, the fact that the set $M(\psi)$, for $\psi \in \Omega$, consists of a single point played an essential role. Theorem 2.2 can be generalized to the case when this assumption is not satisfied; however, it is then necessary to require that the set M be convex. The relevant algorithms for this case [8], [17] are complicated and can be applied to optimal control problems only with difficulty.

Let us dwell on certain peculiarities of the developed algorithm. It is clear that the algorithm can effectively be applied only if $\varphi(\psi)$ can be computed quite simply. But

$$\varphi(\psi) = \max_{z \in M} (\psi, z) = \max_{x \in X} \sum_{i=0}^{n} \psi_i \mu_i(x). \tag{2.12}$$

Therefore, everything reduces to the question of how simple it is to compute the maximum in the right-hand side of (2.12). Fortunately, in linear optimal control problems $\varphi(\psi)$ is quite easy to calculate.

We shall now consider how to apply the approach described above to the solution of optimal control problems with linear systems, by means of the following example.

Let there be given an object whose behavior is described by the linear differential equation

$$\frac{dx}{dt} = A(t)x + G(t, u). \tag{2.13}$$

Here, $A(t)$ is an $n \times n$ matrix and $G(t, u)$ is a vector-function, and $A(t)$ and $G(t, u)$ depend continuously on their arguments. For the admissible controls we take the set of all measurable functions $u(t)$, defined on the interval $[0, T]$, such that $u(t) \in U$ for each $t \in [0, T]$, where U is some compact set. From among all the admissible controls $u(t)$ we are to find

one such that the corresponding solution $x(t)$ of (2.13), with initial condition $x(0) = x^0$, satisfies the constraints:

$$z^k = A^k x(t_k) - b^k \leqq 0, \qquad k = 1, \cdots, m,$$
$$z^{m+1} = A^{m+1} x(t_{m+1}) - b^{m+1} = 0, \tag{2.14}$$

and such that the functional

$$z_0 = \int_0^T g(u(t))\, dt \tag{2.15}$$

assumes its minimum value.

In (2.14) and (2.15), the A^k, $k = 1, \cdots, m+1$, are $n_k \times n$ matrices, the b^k are vectors of dimension n_k, the t_i are fixed times with $0 < t_1 < t_2 < \cdots < t_m < t_{m+1} = T$, and $g(u)$ is a continuous function.

For the set X considered above, we shall choose the set of all admissible controls $u(t)$. Thus, X is the set of all measurable functions $u(t)$ such that $u(t) \in U$ for each t. Then the constraints (2.14) and the functional (2.15) being minimized define, according to (2.2), the set M in a space of dimension $n_1 + n_2 + \cdots + n_{m+1} + 1$. (It is obvious that in all the preceding arguments the dimension of M is equal to the number of constraints $\mu_i(x)$, which in this particular case is $n_1 + n_2 + \cdots + n_{m+1}$, plus 1 for the functional being minimized.)

The set M is bounded, and, on the basis of [20], it is closed. The function $\varphi(\psi)$ for this particular problem has the form

$$\varphi(\psi) = \max_{u(\cdot) \in X} \left[\psi_0 \int_0^T g(u(t))\, dt + \sum_{k=1}^{m+1} (\psi^k, A^k x(t_k) - b^k) \right], \tag{2.16}$$

where $\psi = (\psi_0, \psi^1, \cdots, \psi^m, \psi^{m+1})$ is a vector of dimension $n_1 + n_2 + \cdots + n_{m+1} + 1$, and the vectors ψ^k, $k = 1, \cdots, m+1$, have dimension n_k. Finally, the region Ω for this particular problem is defined by the conditions $\psi_0 = -1$ and $\psi^k \leqq 0$ for $k = 1, \cdots, m$.

Let us consider the problem of computing the maximum in (2.16). If $\Phi(t)$ is the matrix satisfying the equations

$$\frac{d\Phi(t)}{dt} = A(t)\Phi(t),$$
$$\Phi(0) = I,$$

where I is the identity matrix, then the solution $x(t)$ of (2.13) is given by the expression

$$x(t) = \Phi(t)x^0 + \Phi(t) \int_0^t \Phi^{-1}(\tau) G(\tau, u(\tau))\, d\tau.$$

Substituting this expression into (2.16), we obtain, after some tedious but

not complicated transformations, that

$$(2.17)\quad \varphi(\psi) = \max_{u(\cdot)\in X} \int_0^T [\psi_0 g(u(t)) + (\psi(t), G(t, u(t)))]\, dt + \cdots,$$

where the dots denote terms which are independent of $u(\cdot)$, and the function $\psi(t)$ is uniquely specified by the following conditions:

(a) on the interval $t_i < t \leqq t_{i+1}$, $\psi(t)$ satisfies the system of equations

$$(2.18)\quad \frac{d\psi}{dt} = -A^*(t)\psi;$$

(b) $\psi(t)$ is discontinuous from the right at the points t_i, $i = 1, \cdots, m$; in particular,

$$(2.19)\quad \begin{aligned} \psi(t_{m+1}) &= (A^{m+1})^*\psi^{m+1}, \\ \psi(t_i) &= \psi(t_i + 0) + (A^i)^*\psi^i. \end{aligned}$$

It follows from (2.17) that the right-hand side of (2.16) attains its maximum when $u(t)$ satisfies the following condition almost everywhere on the interval $[0, T]$:

$$(2.20)\quad \psi_0 g(u(t)) + (\psi(t), G(t, u(t))) = \max_{v\in U} [\psi_0 g(v) + (\psi(t), G(t, v))].$$

In order for our algorithm to be applicable, it is necessary and sufficient that (2.20) uniquely determine the control $u(t)$. In particular, this will be true if $g(u)$ is a strictly convex function and $G(t, u)$ is linear in u.

Let us now assume that (2.20) uniquely determines $u(t)$, and let us describe one step of the algorithm. As we already noted, the original problem reduces to that of minimizing the function $\varphi(\psi)$ subject to the conditions $\psi_0 = -1$ and $\psi^k \leqq 0$ for $k = 1, \cdots, m$. Therefore, one step of the algorithm coincides with a step in a direction of descent for the minimization of $\varphi(\psi)$. Thus, suppose that the vector $\psi^{(N)}$ has already been constructed, where $\psi^{k(N)} \leqq 0$ for $k = 1, \cdots, m$.

Step 1. The computation of $\varphi(\psi^{(N)})$ and $z^{k(N)}$.

For the vector $\psi^{(N)}$, we determine the function $\psi^{(N)}(t)$ from (2.18) and (2.19). From (2.20), we determine $u^{(N)}(t)$. Integrating (2.13), with $u(t) = u^{(N)}(t)$, we find $x^{(N)}(t)$, and then, from (2.14), we determine $z^{k(N)}$ for $k = 1, \cdots, m + 1$.

Step 2. The construction of $\psi^{(N+1)}$.

In accordance with the method of steepest descent [18], we set

$$\psi^{(m+1)(N+1)} = \psi^{(m+1)(N)} - \lambda_N z^{(m+1)(N)},$$

$$\psi_j^{k(N+1)} = \begin{cases} \psi_j^{k(N)} - \lambda_N z_j^{k(N)} & \text{if } \psi_j^{k(N)} < 0 \text{ or } z_j^{k(N)} \geqq 0, \\ \psi_j^{k(N)} & \text{if } \psi_j^{k(N)} = 0 \text{ and } z_j^{k(N)} < 0. \end{cases}$$

The stepsize λ_N is calculated from the condition

$$\lambda_N = \min \{\lambda_0 , \kappa_N\},$$

where

$$\kappa_N = \min_{\substack{1 \leqq k \leqq m \\ 1 \leqq j \leqq n_k}} \frac{\psi_j^{k(N)}}{z_j^{k(N)}},$$

and the minimum is taken only over those values of k and j for which the ratio $\psi_j^{k(N)}/z_j^{k(N)}$ is positive. Also, λ_0 is a sufficiently small positive number, and ψ_j^k and z_j^k denote the jth components of the vectors ψ^k and z^k, respectively.

The convergence of the indicated method, in connection with optimal control problems, was investigated in more detail in [4], [5], [6]. The relationship between the magnitude λ_N of the step and the geometric structure of the region of attainability is also given there.

From all that has been presented, it becomes clear that the algorithms originally developed in [2] are of a quite general nature and may be applied to a wide class of problems.

Conclusion. Among the papers written in the U.S.S.R. which touch on the systematic approach described above, we should take note of [21], [22]. There, a time-optimal problem where the trajectories are chosen from an arbitrary convex set, and also problems with retardation, were considered. Furthermore, the development of an algorithm, when there is not necessarily a uniquely determined element that maximizes $\varphi(\psi)$, is considered in [22].

In conclusion, I consider it my pleasant duty to express my gratitude to Professor L. W. Neustadt for his kind invitation to write this article, and for the work, which he took upon himself, with respect to its editing.

REFERENCES

[1] L. W. Neustadt, *Synthesizing time-optimal control systems*, J. Math. Anal. Appl., 1 (1960), pp. 484–493.

[2] L. W. Neustadt and B. Paiewonsky, *On synthesizing optimal controls*, Proceedings of Second IFAC Congress, Butterworths, London, 1965.

[3] J. H. Eaton, *An iterative solution to time-optimal control*, J. Math. Anal. Appl., 5 (1962), pp. 287–305. Also see Ibid., 9 (1964), pp. 147–152 for errata.

[4] B. N. Pshenichniy, *A numerical method of computing time-optimal controls in linear systems*, Zh. Vychisl. Mat. i Mat. Fiz., 4 (1964), pp. 52–60.

[5] ———, *A numerical method for solving some optimal control problems*, Ibid., 4 (1964), pp. 292–305.

[6] ———, *A duality principle in convex programming problems*, Ibid., 5 (1965), pp. 98–106.

[7] K. J. Arrow, L. Hurwicz and H. Uzawa, *Studies in Linear and Nonlinear Programming*, Stanford University Press, Stanford, 1958.

[8] B. N. PSHENICHNIY, *A dual method in extremal problems I and II*, Kibernetika, 1 (1965), no. 3, pp. 89–95; no. 4, pp. 64–69.

[9] ———, *Convex programming in a normed space*, Ibid., 1 (1965), no. 5, pp. 46–54.

[10] A. YA. DUBOVITSKIY AND A. A. MILYUTIN, *Extremum problems in the presence of constraints*, Dokl. Akad. Nauk SSSR, 149 (1963), pp. 759–762; English transl. in Soviet Math., 4 (1963), pp. 452–455.

[11] ———, *Extremum problems in the presence of constraints*, Zh. Vychisl. Mat. i Mat. Fiz., 5 (1965), pp. 395–453.

[12] M. A. KRASNOSEL'SKIY AND YA. B. RUTITSKIY, *Convex functions and Orlicz spaces*, Fizmatgiz, Moscow, 1958.

[13] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I: General Theory*, Interscience, New York, 1958.

[14] M. G. KREIN AND M. A. RUTMAN, *Linear operators which leave a cone in a Banach space invariant*, Uspehi Mat. Nauk, 3 (1948), pp. 3–95.

[15] A. YA. DUBOVITSKIY AND A. A. MILYUTIN, *Some optimal control problems for linear systems*, Avtomat. i Telemeh., 24 (1963), pp. 1616–1625; English transl. in Automat. Remote Control, 24 (1964), pp. 1471–1481.

[16] L. W. NEUSTADT, *Optimal control problems as extremal problems in a Banach space*, Proceedings of Polytechnic Institute of Brooklyn Symposium on System Theory, 1965, pp. 215–224.

[17] J. WARGA, *A convergent procedure for convex programming*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 579–587.

[18] G. ZOUTENDIJK, *Methods of Feasible Directions: A Study in Linear and Nonlinear Programming*, Elsevier, Amsterdam, 1960.

[19] S. I. ZUKHOVITSKIY ET AL., *An algorithm for solving convex Chebyshev approximation problems*, Dokl. Akad. Nauk SSSR, 151 (1963), pp. 27–30; English transl. in Soviet Math., 4 (1963), pp. 901–904.

[20] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

[21] N. E. KIRIN, *On the solution of a general time-optimal linear problem*, Avtomat. i Telemeh., 25 (1964), pp. 16–22; English transl. in Automat. Remote Control, 25 (1964), pp. 15–21.

[22] ———, *Programmed optimization for linear systems with holdover*, Ibid., 26 (1965), pp. 3–10; English transl. in Automat. Remote Control, 26 (1965), pp. 1–8.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

ON THE DUALITY BETWEEN ESTIMATION AND CONTROL*

J. D. PEARSON†

Abstract. The problem of smoothing is shown to be a variational dual to a problem of control or regulation. Two fundamental equations occur in both contexts which relate to an earlier duality principle.

1. Introduction. This paper investigates the relationship between the problems of statistical estimation of the state of a linear dynamical system under Gaussian disturbances and the design of a linear regulator with a quadratic performance index. By using an equivalent deterministic smoothing formulation of statistical estimation, the two problems are shown to be variational duals. This duality results in the equivalence of their solutions, which also appear as the optimal estimate and variance equations. The apparently novel formulation of these well-known results explains the close connection between the two problems, and relates to an earlier "duality" equivalence principle [1]. However the duality employed here is fundamental to variational problems [7].

2. The estimation problem. Given:

(i) a partially observed linear dynamic system disturbed by Gaussian white noise inputs,

$$\frac{dx}{dt} = Fx(t) + Gu(t),$$

$$z(t) = Hx(t) + v(t),$$

where $x(t)$ is an n-component state vector, $u(t)$ and $v(t)$ are respectively α- and β-component vectors from independent Gaussian white noise sources, F, G and H are constant matrices;

(ii) a priori estimates of $x(0)$, $u(t)$, $v(t)$ for $0 \leqq t \leqq T$,

$$E(x(0)) = \bar{x}_0, \qquad E(\tilde{x}(0)\tilde{x}(0)') = P_0,$$

$$E(u(t)) = \bar{u}(t), \qquad E(\tilde{u}(t)\tilde{u}(t)') = Q,$$

$$E(v(t)) = 0, \qquad E(\tilde{v}(t)\tilde{v}(t)') = R,$$

where P_0, Q, R are constant positive definite symmetric matrices, $\tilde{x} \triangleq (x - \bar{x})$, prime indicates transpose, $E(\cdot)$ is the expectation operator;

(iii) observations for $0 \leqq t \leqq T$ of $z(t)$;

* Received by the editors August 10, 1965, and in final revised form January 21, 1966.

† Systems Research Center, Case Institute of Technology, Cleveland, Ohio. Now at Research Analysis Corporation, McLean, Virginia.

(iv) that the model and observation process is inherently completely controllable and observable; i.e.,

$$\text{rank } [G, FG, F^2G, \cdots, F^{n-1}G] = n,$$
$$\text{rank } [H', F'H', \cdots, (F')^{n-1}H'] = n.$$

The problem is to find a maximum likelihood estimate $\bar{x}(t)$ of $x(t)$ for $0 \leqq t \leqq T$, and $\hat{x}(T)$ of $x(T)$.

3. The primal and dual smoothing problems. It is reasonably well-known that the maximum likelihood estimate of the smoothed trajectory is given by solving an equivalent primal minimization problem [5], [6], [8].

In this section both primal and dual formulations of the solution will be developed in detail to uncover two fundamental equations which play a central role in smoothing, estimation and regulation.

3.1. Primal smoothing problem. Given $z(t)$, $\bar{u}(t)$, $\bar{x}_0$, P_0, Q and R for $0 \leqq t \leqq T$, minimize the functional

$$\frac{1}{2}\left\{ \| x(0) - \bar{x}_0 \|^2_{P_0^{-1}} + \int_0^T (\| z(t) - Hx(t) \|^2_{R^{-1}} + \| u(t) - \bar{u}(t) \|^2_{Q^{-1}})\, dt \right\}, \tag{3.1}$$

subject to the following controllable dynamic constraint:

$$\frac{dx}{dt} = Fx(t) + Gu(t), \tag{3.2}$$

where $\| x \|^2_Q$ denotes $x'Qx$, and where a forcing function $u(t)$, $0 \leqq t \leqq T$, and $x(0)$ are to be found, which generate the smoothed trajectory $\bar{x}(t)$ over this interval.

Associated with this problem there is a second dual maximization problem.

3.2. Dual smoothing problem. Given are the following dynamic constraints:

$$Q^{-1}(u(t) - \bar{u}(t)) + G'p(t) = 0, \tag{3.3}$$

$$\frac{dp}{dt} + H'R^{-1}Hx(t) + F'p(t) = H'R^{-1}z(t), \tag{3.4}$$

$$\begin{aligned} p(0) + P_0^{-1}(x(0) - \bar{x}_0) &= 0, \\ p(T) &= 0, \end{aligned} \tag{3.5}$$

where $p(t)$ is an n-component multiplier or co-state associated with the

dynamic constraints (3.2); maximize the following functional [7]:

$$\frac{1}{2}\left\{\| x(0) - \bar{x}_0 \|^2_{P_0^{-1}} + \int_0^T \left(\| z(t) - Hx(t) \|^2_{R^{-1}} + \| u(t) - \bar{u}(t) \|^2_{Q^{-1}} + 2p(t)' \left(Fx(t) + Gu(t) - \frac{dx}{dt} \right) \right) dt \right\}. \tag{3.6}$$

This latter equation is the Lagrangian form of the primal problem subject to constraints (3.3)–(3.5), which are the first order necessary conditions for a minimal primal solution. Observability plays a more obvious role in a later formulation (5.1)–(5.2).

Equations (3.2)–(3.5) define a minimizing curve $\bar{x}(t)$, $p(t)$ for (3.1) and a maximizing curve for (3.6) as can be shown using the convexity properties of the integrands. It follows that the roles of state x and co-state p for the primal problem become state p and co-state x for the dual problem; also, that the minimal value of (3.1) is the maximal value of (3.6).

An alternative "feedback" solution can be developed for both problems as opposed to the "open loop" solution given by solving (3.2)–(3.5).

In (3.4) the following substitution for $p(t)$ is employed:

$$p(t) = -P^{-1}(t)(x(t) - e(t)), \tag{3.7}$$

where $e(t)$ is an n-component vector of time functions and $P(t)$ is required to be an $n \times n$ symmetric positive definite matrix. The forcing function $u(t)$ in (3.2) and (3.3) then assumes the form:

$$u(t) = \bar{u}(t) + Q^{-1}G'P^{-1}(t)(x(t) - e(t)). \tag{3.8}$$

Performing the substitution, (3.7) requires that $P(t)$, $e(t)$ satisfy the following differential equations and boundary conditions:

$$\frac{de}{dt} = Fe(t) + G\bar{u}(t) + P(t)H'R^{-1}(z(t) - He(t)), \tag{3.9}$$

$$e(0) = \bar{x}_0,$$

$$\frac{dP}{dt} = FP(t) + P(t)F' + GQG' - P(t)H'R^{-1}HP(t), \tag{3.10}$$

$$P(0) = P_0.$$

R. E. Kalman has shown that (3.10) has a positive semidefinite solution for positive definite P_0, Q, R. Since the same form of (3.10) holds for $P^{-1}(t)$, it follows that $P^{-1}(t)$ exists and $P(t)$ is positive definite [2].

A feedback solution for the dual problem can obviously be found by using (3.7) in its alternate form, where $x(t)$ is the dual co-state to be

eliminated in terms of the dual state $p(t)$,

$$x(t) = e(t) - P(t)p(t). \tag{3.11}$$

Clearly, P and e will satisfy the same equations and boundary conditions. The equations for e and P are fundamental to what follows, for they will appear as the estimator and variance equations in the estimation problem, and as the feedforward and feedback equations in the regulator problem. In addition they solve the smoothing problem.

4. The optimal estimation solution. The best estimate $\hat{x}(T)$ of $x(T)$, given past observations $z(t)$ for $0 \leqq t \leqq T$, coincides in the linear case with the optimal smoothed terminal state $\bar{x}(T)$. Equations for the rate of change of the estimate $\hat{x}(T)$ can be found directly from (3.9).

The terminal state $\bar{x}(T)$ is given by (3.7) and the second equation of (3.5):

$$\bar{x}(T) = e(T) = \hat{x}(T). \tag{4.1}$$

Since (3.9) and (3.10) satisfy all the required initial boundary value problems, and (4.1) satisfies the terminal boundary value problem in (3.5), it follows that (3.9) is the optimal estimator equation when $t = T$. In fact it is true for any time T.

$$\begin{aligned} \frac{d\hat{x}}{dT} &= F\hat{x}(T) + G\bar{u}(T) + P(T)H'R^{-1}(z(T) - H\hat{x}(T)), \\ \hat{x}(0) &= \bar{x}_0. \end{aligned} \tag{4.2}$$

The matrix $P(T)$ is then the variance of $\tilde{x}(T)$ with (3.10) being the variance equation.

It is interesting to note that this result now yields a uniformly simple approach to smoothing given the optimal estimate $\hat{x}(T)$ of covariance $P(T)$ and using (3.8)–(3.10).

5. The optimal regulator problem [7]. The dual smoothing problem can be transformed into a regulator problem. To do this (3.6) is rewritten to eliminate all the primal variables (see Appendix) using partial integration and substitutions from (3.3)–(3.5). An equivalent more interesting problem follows.

Given is a controllable dynamic system,

$$\begin{aligned} \frac{dp}{dt} + F'p(t) + H'm(t) &= 0, \\ p(T) &= 0, \end{aligned} \tag{5.1}$$

where $m(t)$ is a β-component forcing function, $p_0 = P_0^{-1}x_0$, and

$$c(T) = \frac{1}{2}\int_0^T (\| \bar{u}(t) \|^2_{Q^{-1}} + \| z(t) \|^2_{R^{-1}})\, dt + \frac{1}{2}\| x_0 \|^2_{P_0^{-1}}$$

is a constant; maximize the functional

$$\begin{aligned} &c(T) - \frac{1}{2}\| p(0) - p_0 \|^2_{P_0} \\ &\quad - \frac{1}{2}\int_0^T (\| m(t) + R^{-1}z(t) \|^2_R + \| G'p(t) - Q^{-1}\bar{u}(t) \|^2_Q)\, dt. \end{aligned} \tag{5.2}$$

Note that observability of the basic process in §2 implies controllability of this dual problem, which resembles one of making the observed state $G'p(t)$ follow a reference function $Q^{-1}u(t)$. The following equivalence transformation reveals this similarity more clearly:

$$\begin{aligned} t^* &= -t, \\ F^* &= F', \\ G^* &= H', \\ H^* &= G'. \end{aligned} \tag{5.3}$$

To these are added two reference functions:

$$\begin{aligned} m^*(t) &= -R^{-1}z(t), \\ z^*(t) &= Q^{-1}\bar{u}(t). \end{aligned} \tag{5.4}$$

5.1. Optimal regulator problem. Given a controllable dynamic system,

$$\begin{aligned} \frac{dp}{dt^*} &= F^*p(t^*) + G^*m(t^*), \\ p(T) &= 0, \end{aligned} \tag{5.5}$$

minimize the performance index

$$\begin{aligned} &\frac{1}{2}\| p(0) - p_0 \|^2_{P_0} - c(T) \\ &\quad + \frac{1}{2}\int_T^0 (\| z^*(t^*) - H^*p(t^*) \|^2_Q + \| m^*(t^*) - m(t^*) \|^2_R)\, dt^*. \end{aligned} \tag{5.6}$$

The solution to this problem is obtained simply by transforming the dual smoothing solution.

$$\begin{aligned} \frac{dp}{dt^*} &= F^*p(t^*) + G^*m^*(t^*) + G^*R^{-1}G^{*\prime}(e(t^*) - P(t^*)p(t^*)), \\ p(T) &= 0, \end{aligned} \tag{5.7}$$

$$\frac{de}{dt^*} + F^{*\prime}e(t^*) + H^{*\prime}Qz^*(t^*) - P(t^*)G^*(m^*(t^*) + R^{-1}G^{*\prime}e(t^*)) = 0, \tag{5.8}$$

$$e(0) = \bar{x}_0,$$

$$\frac{dP}{dt^*} + P(t^*)F^* + F^{*\prime}P(t^*) + H^{*\prime}QH^* - P(t^*)G^*R^{-1}G^{*\prime}P(t^*) = 0, \tag{5.9}$$

$$P(0) = P_0.$$

The regulator problem operates in reverse time: $T \geqq t^* \geqq 0$. Here $e(t^*)$ is a "feedforward" function depending on the reference functions $z^*(t^*)$ and $m^*(t^*)$, and $P(t^*)$ is a "feedback gain" determined by the relative weighting Q and R.

6. The duality equivalence principles. Clearly there is a connection between the problems of smoothing, estimation and regulation, which can be summarized as follows.

DUALITY EQUIVALENCE THEOREM. *The problems of optimal smoothing defined by* (3.1)–(3.2) *and optimal regulation defined by* (5.5)–(5.6) *are variational duals under the equivalence transformations* (5.3)–(5.4). *The equations for estimation* (3.9) *and variance* (3.10) *are equivalent to those for feedforward* (5.8) *and feedback* (5.9) *under transformations* (5.3)–(5.4).

Practically speaking, the two problems have the following equivalence:
(i) The primal state and co-state are the dual co-state and state respectively.
(ii) The minimal performance indices have the same value, by definition.
(iii) The e and P equations are the same under the equivalence transformation.
(iv) Primal controllability and observability are revealed as the essential conditions that dual formulations of the *same* problem are controllable [2].

7. Conclusions. The central role of two differential equations for e and P in the solution of the dual smoothing and regulator problems has been demonstrated.

An earlier equivalence between estimation and regulation is not strictly applicable to continuous formulations of both problems [1], [3], [4]. This earlier principle applied here states that solutions of the estimation and regulation problems are the same under the equivalence transformation (5.3). The variance equations (3.10) and (5.9) are equivalent but only the homogeneous parts of the estimation equations (4.2) correspond to (5.7). Equations (5.8) play no role whatever. The results of this paper thus provide a conventional justification of a very useful concept.

8. Appendix. The derivation of (5.2) is performed by substituting the dual constraints (3.3)–(3.5) directly into the Lagrangian form (3.6),

with the aid of partial integration of the $p'(dx/dt)$ term,

$$-\frac{1}{2}(x(0) + \bar{x})'P_0^{-1}(x(0) - \bar{x}_0)$$

$$-\frac{1}{2}\int_0^T [(Hx(t) + z(t))'R^{-1}(Hx(t) - z(t)) \tag{8.1}$$

$$+ (u(t) + \bar{u}(t))'Q^{-1}(u(t) - \bar{u}(t))]\, dt.$$

In order to eliminate the primal variables $x(t)$, $x(0)$, $u(t)$, the following transformations are obtained from (3.3)–(3.5).

$$x(0) = \bar{x}_0 - P_0 p(0), \tag{8.2}$$

$$u(t) = \bar{u}(t) - QG'p(t), \tag{8.3}$$

$$Hx(t) = Rm(t) + z(t). \tag{8.4}$$

The resulting dual functional now takes on the following form:

$$\frac{1}{2}\left\{\| \bar{x}_0 \|^2_{P_0^{-1}} + \int_0^T (\| z(t) \|^2_{R^{-1}} + \| \bar{u}(t) \|^2_{Q^{-1}})\, dt\right\}$$

$$-\frac{1}{2}\left\{\| P_0^{-1}\bar{x}_0 - p(0) \|^2_{P_0}\right. \tag{8.5}$$

$$\left. + \int_0^T \| m(t) + R^{-1}z(t) \|^2_R + \| G'p(t) - Q^{-1}\bar{u}(t) \|^2_Q)\, dt\right\}.$$

Equation (5.2) follows directly from this result, and the dual constraint (5.1) is obtained using (3.4) and (8.4).

REFERENCES

[1] R. E. Kalman and R. S. Bucy, *New results in linear filtering and prediction theory*, Trans. ASME Ser. D. J. Basic Engrg., 83D (1961), pp. 95–108.

[2] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[3] ———, *On the general theory of optimal control*, IFAC Congress, Moscow, 1960.

[4] ———, *A new approach to linear filtering and prediction problems*, Trans. ASME Ser. D. J. Basic Engrg., 82D (1960), pp. 35–45.

[5] H. Cox, *On the estimation of state variables and parameters for noisy dynamic systems*, IEEE Trans. Automatic Control, 9 (1964), pp. 5–12.

[6] A. E. Bryson and M. Frazier, *Smoothing for linear and nonlinear dynamic systems*, Tech. Report ASD-TOR-63-119, Aeronautical Systems Division, Wright-Patterson Air Force Base, Dayton, 1963.

[7] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, vol. I, Interscience, New York, 1962, pp. 231–242.

[8] Y. C. Ho, *The method of least squares and optimal filter theory*, Memo RM-3329 PR, The RAND Corporation, Santa Monica, California, 1963.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

OPTIMAL PROCESSES IN DISTRIBUTED PARAMETER SYSTEMS AND CERTAIN PROBLEMS IN INVARIANCE THEORY*

A. I. EGOROV

Abstract. In this paper we investigate optimal processes in systems whose behavior is described by various boundary value problems for partial differential equations.

The majority of physical processes with which an engineer has to deal in his practice are controlled processes and, consequently, in realizing them it is important to obtain variants which are optimal (in some sense or other). The maximum principle of L. S. Pontryagin [1] has proven to be an effective mathematical method for the solution of optimal control problems when the process can be described by ordinary differential equations. However, many controlled processes are described by partial differential equations with supplementary (boundary and initial) conditions. These equations may be of various types (equations of mass- and heat-transfer, of hydro- and aerodynamics, of heat conduction, of kinetics of chemical reactions, etc.). If the behavior of the control system is described by equations among which there are partial differential equations, then it is called a distributed-parameter system (see [2]). In a number of the simplest cases these systems can be described by differential-difference equations and, consequently, the maximum principle can be applied (see [3]).

Optimal control problems for more complex systems cannot be solved directly with the aid of the maximum principle of L. S. Pontryagin (see [4, pp. 516–518]). Therefore, attempts were made to generalize this principle so that controlled processes in more complex distributed-parameter systems could be investigated (see [5]–[15]). In particular, a method based on the application of differential equations in Banach spaces was proposed in reference [10]. In many cases such a method allows us to treat partial differential equations as ordinary differential equations and to solve the optimal control problem when the functional

$$I = \int_{t_0}^{T} f(t, x(t), u(t))\, dt \tag{1}$$

* Originally published in Izv. Akad. Nauk SSSR Ser. Mat., 29(1965), pp. 1205–1260. Submitted on March 13, 1964. This translation into English has been prepared by N. H. Choksy.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

The basic contents of the paper were presented in the Seminar of L. S. Pontryagin on the Theory of Optimal Processes on February 13, 1964.

is chosen as the criterion of optimality. Although this method has definite merit, it also has intrinsic deficiencies since the introduction of Banach spaces imposes auxiliary constraints on the class of admissible controls not called for by the nature of the problems. Moreover, the choice of functional (1) as the optimality criterion for problems with partial differential equations is not as successful a one as for problems with ordinary differential equations. In particular, the indicated method cannot solve the problem, important in practice, when the functional (1) is replaced by an integral computed over the surface bounding the region in which the equation is considered.

Of definite interest is the method (see [13]) which is based on the representation of the controlled quantities by means of integral relations. However, it is not possible to consider it as satisfactorily substantiated. Moreover, the application of this method to processes which are described by boundary value problems for partial differential equations is not sufficiently effective for the following reasons. Firstly, the reduction of boundary value problems to integral equations cannot always be effected in practice, although the problem can be solved by other methods. Secondly, it is always desirable to have the optimality criteria expressed in terms of quantities occurring in the equations and in the supplementary conditions.

In the present paper a method of solution is used which is equally applicable in cases of hyperbolic, parabolic, and elliptic equations. Using the same method, L. I. Rozonoer [16] has studied the case when the controlled process is described by ordinary differential and finite-difference equations. In succeeding papers (see [17]) he obtained the condition for the invariance of systems relative to external excitations; moreover, the starting point in these investigations was the formula for the increment of the functional from [16]. Analogous results (however, for particular cases only) were successfully obtained also for distributed-parameter systems.

The paper consists of five sections. In §§1 and 2 various optimal control problems are considered for processes which are described by boundary value problems for hyperbolic equations with data on the characteristics. Necessary optimality conditions are formulated in the form of the maximum principle.

In §3 connections are established between the problems being investigated and the problems of the calculus of variations. It is shown that the Euler-Ostrogradskiĭ equations can be obtained from the maximum principle when the control region is the whole space. However, if this region is closed, the Legendre condition need not be satisfied along the optimal surface.

Section 4 studies the optimal control problem when the processes are described by boundary value problems for parabolic systems. A formula for the increment of the functional is obtained, with the help of which the

optimality conditions are found. These results are carried over to the analogous problems connected with elliptic and hyperbolic systems.

Section 5 deals with problems in invariance theory. Necessary and sufficient invariance conditions relative to external excitations are obtained for linear equations when functionals analogous to those considered in §§1–4 are chosen as the criteria of invariance.

The author takes this opportunity to express his thanks to L. S. Pontryagin and to the participants in his seminar for their attention to the present paper. Furthermore, the author sincerely acknowledges V. G. Boltyanskii, O. A. Oleinik and Yu. V. Egorov for very useful discussions of the results obtained in the paper.

1. Optimal processes in systems whose behavior is described by hyperbolic equations.

1.1. Statement of the problem. Optimality conditions. Let the controlled process be described by the system of equations

$$z_{ixy} = f_i(x, y, z_1, \cdots, z_m, z_{1x}, \cdots, z_{mx}, z_{1y}, \cdots, z_{my}, v), \quad i = 1, \cdots, m, \tag{1.1}$$

where the functions f_i have continuous first-order derivatives with respect to x and y and are twice continuously differentiable with respect to the remaining arguments in the region G, $0 \leqq x \leqq X$, $0 \leqq y \leqq Y$. As the class of admissible controls we shall take the set of piecewise-continuous functions $v = v(x, y)$ defined in the region G and with values in some bounded convex region V (open or closed) of the r-dimensional Euclidean space. It is assumed that the lines of discontinuity of the admissible controls are sufficiently smooth. On the function z_i defined by (1.1) are imposed the boundary conditions (the Goursat conditions)

$$z_i(0, y) = \varphi_i(y), \qquad z_i(x, 0) = \psi_i(x), \quad i = 1, \cdots, m, \tag{1.2}$$

where φ_i and ψ_i are continuous, piecewise-continuously differentiable functions defined in the region G and satisfying the conjugate conditions

$$\varphi_i(0) = \psi_i(0).$$

To every admissible control there corresponds a unique solution $z(x, y) = \{z_1(x, y), \cdots, z_m(x, y)\}$ of the problem (1.1)–(1.2) having derivatives z_{ixy} which are integrable in the region G (see [18]). Here, however, we have to distinguish two cases.

(1) If the line of discontinuity of the function $v(x, y)$ is parallel to one of the coordinate axes, the boundary value problem (1.1)–(1.2) splits up into two analogous problems in regions which abut each other along this line.

Having solved these problems in sequence we determine the solution of the original problem, which will be continuous in the region G and will have everywhere, except on the points of the line of discontinuity of the control $v(x, y)$, continuous derivatives $z_{ix}(x, y)$, $z_{iy}(x, y)$ and $z_{ixy}(x, y)$ (see [18]).

(2) Let the line of discontinuity Γ of the function $v(x, y)$ not coincide with a characteristic of system (1.1) on any nonzero segment. By a solution of the boundary-value problem (1.1)–(1.2) we shall mean a function $z(x, y)$ which satisfies the system (1.1) at all points of region G not lying on Γ, the conditions (1.2), and certain preassigned smoothness conditions on Γ (see [19]). Such a solution is determined uniquely; it is continuous in the region G and has piecewise-continuous derivatives z_{ix} , z_{iy} , z_{ixy} .

Therefore, in what follows we shall assume that to every admissible control there corresponds a class of functions in which the boundary-value problem (1.1)–(1.2) is solvable uniquely.

Let A_i , $i = 1, \cdots, m$, be a given system of real numbers. We shall take an arbitrary control $v(x, y)$, denote by $z(x, y)$ the solution of problem (1.1)–(1.2) corresponding to it, and consider the functional

$$S = \sum_{i=1}^{m} A_i z_i(X, Y), \tag{1.3}$$

where X and Y are the constants occurring in the definition of region G.

We pose the problem: from among all the admissible controls find the control $v(x, y)$ (if it exists) such that the solution $z(x, y)$ of the Goursat problem corresponding to it makes the functional S attain its largest (smallest) value.

The admissible control which realizes the minimum (maximum) of functional S will be called min-optimal (max-optimal) with respect to S (see [16]).

Let us remark that the problem (1.1)–(1.2) being considered is of great theoretical and practical interest. The investigation of the solvability of this problem under various assumptions relative to the functions f_i , φ_i , and ψ_i has an extensive literature devoted to it (for example, see [20]–[25]). It is also well-known (see [26]–[28]) that the study of gas sorption and desorption processes, drying processes, etc., leads to such problems. The presence of the control parameters in (1.1) makes it possible to handle the process and in many cases to choose the best mode which (from a mathematical point of view) leads to the determination of the maximum or the minimum of a certain functional. In a number of cases the problem can reduce to the investigation of (1.3). Let us consider some examples.

Example 1. It is required to minimize the functional

$$I = \iint_G f_0(x, y, z, z_x , z_y , v)\, dx\, dy.$$

If we introduce the new variable z_0 by setting

$$(1.4) \qquad z_{0xy} = f_0(x, y, z, z_x, z_y, v), \qquad z_0(0, y) = z_0(x, 0) = 0,$$

then the problem leads to the determination of the minimum of the functional $S = z_0(X, Y)$, which is a special case of (1.3) and is defined on the functions $z_0, \cdots, z_m$ given by the set of relations (1.1)–(1.2) and (1.4).

Example 2. It is required to minimize the functional $I = \Phi(z_1(X, Y), \cdots, z_m(X, Y))$, where Φ is a twice continuously differentiable function.

We introduce the new function $z_0(x, y)$ with the aid of the equation

$$z_{0xy} = \sum_{i,k=1}^{m} \frac{\partial^2 \Phi(z_1(x, y), \cdots, z_m(x, y))}{\partial z_i \, \partial z_k} z_{ix} z_{ky} + \sum_{i=1}^{m} \frac{\partial \Phi}{\partial z_i} f_i(x, y, z, z_x, z_y, v)$$

and of the supplementary conditions

$$z_0(0, y) = \Phi(\varphi_1(y), \cdots, \varphi_m(y)), \qquad z_0(x, 0) = \Phi(\psi_1(x), \cdots, \psi_m(x)).$$

By the same token the problem is reduced to the study of the functional $S = z_0(X, Y)$.

Example 3. It is required to minimize the functional

$$I = \int_0^X F(x, z(x, Y), z_x(x, Y)) \, dx.$$

We introduce the auxiliary function $z_0(x, y)$ with the aid of the equation

$$z_{0yx} = \sum_{i=1}^{m} \left[\frac{\partial F}{\partial z_i} z_{iy} + \frac{\partial F}{\partial z_{ix}} f_i(x, y, z, z_x, z_y, v) \right]$$

and of the supplementary conditions

$$z_0(0, y) = 0, \qquad z_0(x, 0) = \int_0^x F(x, \psi(x), \psi'(x)) \, dx.$$

Example 4. Analogously we consider the problem of minimizing the functional

$$I = \int_0^Y F(y, z(X, y), z_y(X, y)) \, dy.$$

To solve the optimization problem we have formulated let us introduce the auxiliary functions $u_1, \cdots, u_m$ with the aid of the equations

$$(1.5) \qquad u_{ixy} = \frac{\partial H(x, y, p, v)}{\partial z_i} - \frac{d}{dx}\left(\frac{\partial H(x, y, p, v)}{\partial z_{ix}}\right) - \frac{d}{dy}\left(\frac{\partial H(x, y, p, v)}{\partial z_{iy}}\right)$$

and of the supplementary conditions

$$(1.6)\qquad \begin{aligned} u_{ix}(x, Y) &= -\frac{\partial H(x, Y, p, v)}{\partial z_{iy}}, \\ u_{iy}(X, y) &= -\frac{\partial H(X, y, p, v)}{\partial z_{ix}}, \qquad u_i(X, Y) = A_i, \end{aligned}$$

where the A_i are the constants occurring in the definition of functional S,

$$p = (z_1, \cdots, z_m, u_1, \cdots, u_m, z_{1x}, \cdots, z_{mx}, z_{1y}, \cdots, z_{my}),$$
$$H = \sum u_i f_i(x, y, z, z_x, z_y, v).$$

Equations (1.6) are linear differential equations in the ordinary derivatives with initial data.

In the general case the functions z_{ixx}, z_{iyy}, v_x and v_y occur on the right-hand sides of (1.5). However, the existence of these derivatives does not follow from the conditions imposed on (1.1) and on the admissible controls. Therefore, in what follows we shall assume that the functions f_i are of the form

$$f_i = \sum_{j,k=1}^{m} a_{ijk}(x, y, z) z_{kx} z_{jy} + \sum_{j=1}^{m} b_{ij}(x, y, z) z_{jx} + \sum_{j=1}^{m} c_{ij}(x, y, z) z_{jy} + d_i(x, y, z, v),$$

where the functions a_{ijk}, b_{ij}, c_{ij} and d_i are continuously differentiable with respect to x and y and twice continuously differentiable with respect to the remaining arguments. If it turns out that a_{ijk}, b_{ij} and c_{ij} depend on v, then it is necessary to require that the admissible controls have the piecewise-continuous derivatives $v_x(x, y)$ and $v_y(x, y)$.

When these conditions are satisfied the system of linear equations (1.5) will have piecewise-continuous coefficients and, together with the supplementary conditions (1.6), will uniquely determine the functions $u_1(x, y), \cdots, u_m(x, y)$ for each admissible control. Therefore, in what follows we shall assume that the functions f_i and the admissible controls are such that the boundary value problem (1.5)–(1.6) is uniquely solvable for each admissible control.

We shall say that an admissible control $v(x, y)$ satisfies a maximum condition if the relation

$$(1.7)\qquad H(x, y, p(x, y), v(x, y)) \quad ((=)) \quad \sup_{v \in V} H(x, y, p(x, y), v)$$

is fulfilled, where $z(x, y)$ and $u(x, y)$ are solutions of problems (1.1)–(1.2) and (1.5)–(1.6) corresponding to the control $v(x, y)$, while the symbol

$((=))$ denotes an equality which is valid at all points of the region G, $0 \leqq x \leqq X$, $0 \leqq y \leqq Y$, except perhaps on a set of points lying on a finite number of lines with zero area. A minimum condition is defined analogously.

THEOREM 1. (The Maximum Principle). *In order that an admissible control $v(x, y)$ be min-optimal (max-optimal) with respect to S, it is necessary that it satisfy a maximum (minimum) condition.*

Although this theorem does not give sufficient conditions for the existence of optimal controls, it can be used for a practical solution of the optimal problem. Indeed, according to the maximum principle a solution of this problem leads to the necessity of determining the $2n + 1$ unknowns z_i, u_i and v from the $2n + 1$ equations (1.1), (1.5) and (1.7). The first $2n$ relations are second-order differential equations and, generally speaking, $4n$ arbitrary functions will appear in their solutions. To eliminate them we have the $4n$ supplementary conditions (1.2) and (1.6). By the same token we have determined, generally speaking, the isolated solutions of problem (1.1)–(1.2) satisfying the conditions of the maximum principle. If within the meaning of the problem it turns out that the optimization problem necessarily has a solution, then at least one of the isolated solutions we have found will also be the desired one.

1.2. Formula for the increment of functional S. To prove Theorem 1 let us consider the functional

$$I[p, v] = \iint_G \left[\sum_{i=1}^{m} u_i z_{ixy} - H(x, y, p, v) \right] dx\, dy.$$

If v is some admissible control and $z = z(x, y)$ is the solution of problem (1.1)–(1.2) corresponding to this control, then the functional I equals zero for an arbitrary function $u = (u_1, \cdots, u_m)$.

Let $v = v(x, y)$ be some admissible control and let $z(x, y)$ and $u(x, y)$ be the solutions of the boundary value problems (1.1)–(1.2) and (1.5)–(1.6) corresponding to it. Let us give the function v an admissible increment Δv, and let $z + \Delta z$ and $u + \Delta u$ denote the solutions of the same problems but corresponding to the control $v + \Delta v$. It is obvious that the functions Δz_i and Δu_i satisfy the equations

$$\begin{aligned} \Delta z_{ixy} &= \Delta \frac{\partial H}{\partial u_i}, \\ \Delta u_{ixy} &= \Delta \frac{\partial H}{\partial z_i} - \frac{d}{dx}\left(\Delta \frac{\partial H}{\partial z_{ix}}\right) - \frac{d}{dy}\left(\Delta \frac{\partial H}{\partial z_{iy}}\right), \quad i = 1, \cdots, m, \end{aligned} \tag{1.8}$$

and the supplementary conditions

$$\Delta z_i(0, y) = \Delta z_i(x, 0) = 0, \tag{1.9}$$

$$\Delta u_{ix}(x, Y) = -\Delta \frac{\partial H(x, Y, p, v)}{\partial z_{iy}},$$

$$\Delta u_{iy}(X, y) = -\Delta \frac{\partial H(X, y, p, v)}{\partial z_{ix}}, \tag{1.10}$$

$$\Delta u_i(X, Y) = 0, \qquad i = 1, \cdots, m, \tag{1.11}$$

where

$$\Delta \frac{\partial H}{\partial p_i} = \frac{\partial H(x, y, p + \Delta p, v + \Delta v)}{\partial p_i} - \frac{\partial H(x, y, p, v)}{\partial p_i}. \tag{1.12}$$

Equations (1.10) are differential equations in ordinary derivatives and, moreover, for linear f_i the functions

$$\Delta u_i(x, Y) \equiv 0, \qquad \Delta u_i(X, y) \equiv 0 \tag{1.13}$$

are their solutions and satisfy the supplementary conditions (1.11). By virtue of the uniqueness theorem, the functions (1.13) form a unique solution of the boundary value problem (1.10)–(1.11).

Furthermore, according to the remark made above,

$$\Delta I = I[p + \Delta p, v + \Delta v] - I[p, v] = 0. \tag{1.14}$$

On the other hand,

$$\Delta I = \iint_G \left\{ \sum_{i=1}^{m} [\Delta u_i \, \Delta z_{ixy} + u_i \, \Delta z_{ixy} + \Delta u_i \, z_{ixy}] - [H(x, y, p + \Delta p, v + \Delta v) - H(x, y, p, v)] \right\} dx \, dy. \tag{1.15}$$

With the help of Green's formula (see [29, p. 196]) we transform the expression under the integral sign:

$$\iint_G (qs_{xy} - sq_{xy}) \, dx \, dy = \int_L (qs_y - sq_y) \, dy - (qs_x - sq_x) \, dx,$$

where L is the contour bounding the region G and q and s are arbitrary functions having piecewise-continuous first and second order derivatives. Since G is a rectangle, the Green's formula can be reduced to the form

28/*CU*

$$\iint_G (qs_{xy} - sq_{xy}) \, dx \, dy = \{[q(x, y)s(x, y)]_{x=0}^{X}\}_{y=0}^{Y} - \int_0^X (sq_x)_{y=0}^{Y} \, dx - \int_0^Y (sq_y)_{x=0}^{X} \, dy. \tag{1.16}$$

We set $q = \Delta u_i$, $s = \Delta z_i$ in this equality. Taking into account (1.8) with the supplementary conditions (1.9), (1.10) and (1.11), after elementary

manipulations we obtain:

$$\iint_G \sum_{i=1}^{m} \Delta u_i \Delta z_{ixy}\, dx\, dy = \iint_G \sum_{i=1}^{m} \left[\Delta \frac{\partial H}{\partial z_i} \Delta z_i + \Delta \frac{\partial H}{\partial z_{ix}} \Delta z_{ix} + \Delta \frac{\partial H}{\partial z_{iy}} \Delta z_{iy} \right] dx\, dy.$$

On the other hand, by virtue of the first m equations from (1.8) we have:

$$\iint_G \sum_{i=1}^{m} \Delta u_i \Delta z_{ixy}\, dx\, dy = \iint_G \sum_{i=1}^{m} \Delta \frac{\partial H}{\partial u_i} \Delta u_i\, dx\, dy.$$

From the last two equalities we obtain:

$$\iint_G \sum_{i=1}^{m} \Delta u_i \Delta z_{ixy}\, dx\, dy = \frac{1}{2} \iint_G \sum_{i=1}^{4m} \Delta \frac{\partial H}{\partial p_i} \Delta p_i\, dx\, dy. \tag{1.17}$$

We set $q = u_i$, $s = \Delta z_i$ in (1.16). Then by virtue of (1.5) and (1.8) and of (1.6) and (1.9) we have:

$$\begin{aligned} \iint_G \sum_{i=1}^{m} u_i \Delta z_{ixy}\, dx\, dy &= -\sum_{i=1}^{m} A_i \Delta z_i(X, Y) \\ &+ \iint_G \sum_{i=1}^{m} \left[\frac{\partial H}{\partial z_i} \Delta z_i + \frac{\partial H}{\partial z_{ix}} \Delta z_{ix} + \frac{\partial H}{\partial z_{iy}} \Delta z_{iy} \right] dx\, dy. \end{aligned} \tag{1.18}$$

Furthermore, since the functions z_i form the solution of the system (1.1),

$$\iint_G \sum_{i=1}^{m} \Delta u_i z_{ixy}\, dx\, dy = \iint_G \sum_{i=1}^{m} \frac{\partial H}{\partial u_i} \Delta u_i\, dx\, dy. \tag{1.19}$$

Applying Taylor's formula we get the equality

$$\begin{aligned} H(x, y, p + \Delta p, v + \Delta v) - H(x, y, p, v) &= \sum_{i=1}^{4m} \frac{\partial H(x, y, p, v + \Delta v)}{\partial p_i} \Delta p_i \\ &+ \frac{1}{2} \sum_{i,k=1}^{4m} \frac{\partial^2 H(x, y, p + \theta \Delta p, v + \Delta v)}{\partial p_i \partial p_k} \Delta p_i \Delta p_k \\ &+ H(x, y, p, v + \Delta v) - H(x, y, p, v), \qquad 0 \leqq \theta \leqq 1. \end{aligned} \tag{1.20}$$

From (1.14), (1.15) and (1.17)–(1.20) it follows that

$$\begin{aligned} \Delta I = &- \sum_{i=1}^{m} A_i \Delta z_i(X, Y) \\ &- \iint_G [H(x, y, p, v + \Delta v) - H(x, y, p, v)]\, dx\, dy \\ &+ \frac{1}{2} \iint_G \sum_{i=1}^{m} \left\{ \left[\frac{\partial H(x, y, p + \Delta p, v + \Delta v)}{\partial p_i} - \frac{\partial H(x, y, p, v + \Delta v)}{\partial p_i} \right] \right. \end{aligned}$$

$$-\left[\frac{\partial H\ (x, y, p, v + \Delta v)}{\partial p_i} - \frac{\partial H\ (x, y, p, v)}{\partial p_i}\right]\Delta p_i\, dx\, dy$$

$$-\frac{1}{2}\iint_G \sum_{i,k=1}^{4m} \frac{\partial^2 H(x, y, p + \theta\Delta p, v + \Delta v)}{\partial p_i \partial p_k}\, \Delta p_i \Delta p_k\, dx\, dy.$$

Applying Taylor's formula to the functions $\partial H/\partial p_i$ and taking (1.14) into account, we finally obtain:

$$(1.21)\quad \Delta S = -\iint_G [H(x, y, p, v + \Delta v) - H(x, y, p, v)]\, dx\, dy - \eta,$$

where $\Delta S = \sum A_i \Delta z_i(X, Y)$ is the increment of functional S, $\eta = \eta_1 + \eta_2$,

$$\eta_1 = \frac{1}{2}\sum_{i=1}^{4m} \iint_G \left[\frac{\partial H(x, y, p, v + \Delta v)}{\partial p_i} - \frac{\partial H(x, y, p, v)}{\partial p_i}\right]\Delta p_i\, dx\, dy,$$

$$(1.22)\quad \eta_2 = \frac{1}{2}\sum_{i,k=1}^{4m} \iint_G \left[\frac{\partial^2 H(x, y, p + \theta\Delta p, v + \Delta v)}{\partial p_i \partial p_k} - \frac{\partial^2 H(x, y, p + \theta_1 \Delta p, v + \Delta v)}{\partial p_i \partial p_k}\right]\Delta p_i \Delta p_k\, dx\, dy.$$

1.3. Estimate of the remainder term η in (1.21). To obtain the necessary estimates of the quantities η we introduce the auxiliary functions $\alpha_i(x, y)$ and $\beta_i(x, y)$ by setting

$$\alpha_i = \Delta z_{ix}, \qquad \beta_i = \Delta z_{iy}.$$

Since the functions f_i satisfy a Lipschitz condition, from the first m equations of (1.8) and conditions (1.9) we obtain:

$$|\alpha_i| \leqq N \int_0^y \sum_{i=1}^{m} (|\Delta z_i| + |\alpha_i| + |\beta_i|)\, dy + N_1 \int_0^y \sum_{k=1}^{r} |\Delta v_k|\, dy,$$

$$(1.23)\qquad |\Delta z_i| \leqq \int_0^x |\alpha_i|\, dx,$$

$$|\beta_i| \leqq N \int_0^x \sum_{i=1}^{m} (|\Delta z_i| + |\alpha_i| + |\beta_i|)\, dx + N_1 \int_0^x \sum_{k=1}^{r} |\Delta v_k|\, dx,$$

$$|\Delta z_i| \leqq \int_0^y |\beta_i|\, dy,$$

where N and N_1 are specific positive constants. By introducing the notations

$$(1.24)\quad \alpha = \sum_{i=1}^{m} |\alpha_i|, \qquad \beta = \sum_{i=1}^{m} |\beta_i|, \qquad \gamma = \sum_{i=1}^{m} |\Delta z_i|, \qquad \Delta v = \sum_{k=1}^{r} |\Delta v_k|,$$

from (1.23) we have:

$$\begin{aligned}
\alpha(x,y) &\leqq Nm\int_0^y \alpha(x,y)\,dy + Nm\int_0^\eta [\beta(x,\eta)+\gamma(x,\eta)]\,d\eta \\
&\quad + N_1 m\int_0^Y \Delta v\,dy, \qquad \gamma \leqq \int_0^x \alpha(x,y)\,dx, \\
\beta(x,y) &\leqq Nm\int_0^x \beta(x,y)\,dx + Nm\int_0^\xi [\alpha(x,y)+\beta(x,y)]\,dx \\
&\quad + N_1 m\int_0^X \Delta v\,dx, \qquad \gamma \leqq \int_0^y \beta(x,y)\,dy,
\end{aligned} \tag{1.25}$$

where $0 \leqq x \leqq \xi \leqq X$, $0 \leqq y \leqq \eta \leqq Y$. Hence, by virtue of a well-known lemma (see [30, p. 19]) it follows that

$$\alpha(x,y) \leqq M\int_0^\eta [\gamma(x,y)+\beta(x,y)]\,dy + M_1\int_0^Y \Delta v(x,y)\,dy,$$

$$\beta(x,y) \leqq P\int_0^\xi [\gamma(x,y)+\alpha(x,y)]\,dx + P_1\int_0^X \Delta v(x,y)\,dx,$$

where M, M_1, P, P_1 are positive constants. Taking the estimates for the function γ in (1.25) into account we get:

$$\alpha(x,y) \leqq M_2\int_0^\eta \beta(x,y)\,dy + M_1\int_0^Y \Delta v(x,y)\,dy,$$

$$\beta(x,y) \leqq P_2\int_0^\xi \alpha(x,y)\,dy + P_1\int_0^X \Delta v(x,y)\,dx.$$

Hence we find that

$$\begin{aligned}
\alpha(\xi,\eta) &\leqq M_3\int_0^\eta\int_0^\xi \alpha(x,y)\,dx\,dy + M_4\int_0^Y\int_0^X \Delta v(x,y)\,dx\,dy \\
&\quad + M_1\int_0^Y \Delta v(\xi,y)\,dy, \\
\beta(\xi,\eta) &\leqq P_3\int_0^\eta\int_0^\xi \beta(x,y)\,dx\,dy + P_4\int_0^Y\int_0^X \Delta v(x,y)\,dx\,dy \\
&\quad + P_1\int_0^X \Delta v(x,\eta)\,dx.
\end{aligned} \tag{1.26}$$

Integrating the first of these inequalities with respect to ξ in the limits from 0 to ξ and applying the lemma mentioned above we get:

$$\int_0^\xi \alpha(x,y) \leqq M_5\int_0^X\int_0^Y \Delta v(x,y)\,dy\,dx.$$

Hence also from the first inequality in (1.26) we have:

$$\alpha(x, y) \leqq M_6 \int_0^X \int_0^Y \Delta v(x, y)\, dy\, dx + M_7 \int_0^Y \Delta v(x, y)\, dy,$$
$$0 \leqq x \leqq X, \qquad 0 \leqq y \leqq Y.$$

Analogously, we find:

$$\beta(x, y) \leqq P_6 \int_0^X \int_0^Y \Delta v(x, y)\, dx\, dy + P_7 \int_0^X \Delta v(x, y)\, dx.$$

Hence also from (1.25) we obtain:

$$\gamma(x, y) \leqq Q \int_0^X \int_0^Y \Delta v(x, y)\, dy\, dx.$$

Thus, the inequalities

$$|\Delta z_i(x, y)| \leqq Q \iint_G \Delta v(x, y)\, dx\, dy,$$

$$|\Delta z_{ix}(x, y)| \leqq Q_1 \iint_G \Delta v(x, y)\, dx\, dy + R_1 \int_0^Y \Delta v(x, y)\, dy, \tag{1.27}$$

$$|\Delta z_{iy}(x, y)| \leqq Q_2 \iint_G \Delta v(x, y)\, dx\, dy + R_2 \int_0^X \Delta v(x, y)\, dx,$$

are valid for all x and y, $0 \leqq x \leqq X$, $0 \leqq y \leqq Y$, by virtue of (1.24).

By applying analogous methods to the last m equations of (1.8) we get:

$$|\Delta u_i(x, y)| \leqq Q_3 \iint_G \Delta v(x, y)\, dx\, dy. \tag{1.28}$$

Since the functions $\partial H/\partial p_i$ satisfy a Lipschitz condition, from the first formula in (1.22) we have by virtue of (1.27) and (1.28),

$$|\eta_1| \leqq T \left(\iint_G \Delta v(x, y)\, dx\, dy \right)^2 + T_1 \int_0^X \left[\int_0^Y \Delta v(x, y)\, dy \right]^2 dx + T_3 \int_0^Y \left[\int_0^X \Delta v(x, y)\, dx \right]^2 dy.$$

Consequently,

$$|\eta_1| \leqq (T_1 XY + T_2 Y + T_3 X) \iint_G [\Delta v(x, y)]^2\, dx\, dy,$$

where the T_i are specific positive constants.

The functions $\partial^2 H/\partial p_i \partial p_k$ are bounded in the region G. Therefore,

$$|\eta_2| \leqq (T_4 XY + T_5 Y + T_6 X) \iint_G [\Delta v(x, y)]^2\, dx\, dy.$$

Thus, the remainder term in (1.21) satisfies the inequality

$$(1.29) \qquad |\eta| \leqq (A \text{ Meas } G + BX + CY) \iint_G [\Delta v(x, y)]^2 \, dx \, dy,$$

where A, B and C are specific positive constants. If the function Δv is nonzero in the circle G_ϵ of radius ϵ, then from (1.29) it follows that

$$(1.30) \qquad |\eta| \leqq L\epsilon \iint_{G_\epsilon} \Delta v^2(x, y) \, dx \, dy,$$

where L does not depend on ϵ.

1.4. Proof of Theorem 1. The case of a linear system of equations. It is easy to obtain the proof of Theorem 1 from formula (1.21) for the increment of the functional and from estimate (1.30) of the remainder term in this formula.

Indeed, for the sake of definiteness let $v(x, y)$ be a control which is minoptimal with respect to S, and let $z(x, y)$ and $u(x, y)$ be the solutions of the boundary value problems (1.1)–(1.2) and (1.5)–(1.6) corresponding to it. Then, the inequality $\Delta S \geqq 0$ is valid for any admissible increment $\Delta v(x, y)$. Let us suppose that in the region G there exists a point (ξ, η) at which the maximum condition is not satisfied, i.e., there exists a control v^1 such that

$$(1.31) \qquad H(\xi, \eta, p(\xi, \eta), v^1) > H(\xi, \eta, p(\xi, \eta), v(\xi, \eta)).$$

Since the functions $z(x, y)$ and $u(x, y)$ are continuous, while z_x, z_y and $v(x, y)$ are piecewise continuous, there exists a closed region $G^1 \in G$ containing the point (ξ, η), in which the left- and right-hand sides of (1.31) are continuous and, consequently, uniformly continuous. If (ξ, η) is a point of discontinuity of control v, then it is obvious that it can be put on the boundary of region G^1. It follows from (1.31) that we can find a number $\delta > 0$ for which

$$(1.32) \qquad H(x, y, p(x, y), v^1) - H(x, y, p(x, y), v(x, y)) > \delta$$

at all points $(x, y) \in G_\epsilon \subset G^1$, where G_ϵ is a circle of radius ϵ. Let us take the control

$$v^2(x, y) = \begin{cases} v(x, y) & \text{if} \quad (x, y) \notin G_\epsilon, \\ v^1 & \text{if} \quad (x, y) \in G_\epsilon. \end{cases}$$

Then, by virtue of relations (1.21), (1.30) and (1.32),

$$\Delta S = -\iint_{G_\epsilon} [H(x, y, p(x, y), v^1) - H(x, y, p(x, y), v(x, y))] \, dx \, dy - \eta$$

$$< -\iint_{G_\epsilon} \delta \, dx \, dy + |\eta| \leqq -\iint_{G_\epsilon} \{\delta - \epsilon L[\Delta v(x, y)]^2\} \, dx \, dy,$$

where $\Delta v = v^1 - v(x, y)$. Since the function Δv is bounded, the number ϵ can be chosen so small that the expression within the square brackets on the right-hand side of the last inequality is positive. Then ΔS will be negative, which contradicts the min-optimality with respect to S of the control $v(x, y)$. The theorem is proved.

The formula (1.21) for the increment of the functional together with the formulas (1.22) for the remainder term allows us to obtain more general results for a linear boundary value problem.

Indeed, let the controlled process be described by the boundary value problem

$$z_{ixy} = \sum_{k=1}^{m} [c_{ik}(x, y)z_{kx} + d_{ik}(x, y)z_{ky} + g_{ik}(x, y)z_k] + f_i(v),$$

$$z_i(0, y) = \varphi_i(y), \qquad z_i(x, 0) = \psi_i(x), \qquad \varphi_i(0) = \psi_i(0), \tag{1.33}$$

$$i = 1, \cdots, m,$$

and let it be required to determine the control by which the functional S attains its minimal (maximal) value. In this case,

$$H(x, y, p, v) = \sum_{i,k=1}^{m} u_i[c_{ik}z_{kx} + d_{ik}z_{ky} + g_{ik}z_k] + \sum_{i=1}^{m} u_i f_i(v),$$

and the functions u_i form the solution of the boundary value problem

$$u_{ixy} = \sum_{k=1}^{m} \left[g_{ki}\, u_k - \frac{d}{dx}(c_{ki}\, u_k) - \frac{d}{dy}(d_{ki}\, u_{ki}) \right],$$

$$u_{ix}(x, Y) = -\sum_{k=1}^{m} d_{ki}(x, Y)u_k(x, Y), \tag{1.34}$$

$$u_{iy}(X, y) = -\sum_{k=1}^{m} c_{ki}(X, y)u_k(X, y),$$

$$u_i(X, Y) = -A_i, \qquad i = 1, \cdots, m.$$

From what was proved earlier, since

$$\Delta u_i(x, Y) = \Delta u_i(X, y) \equiv 0$$

(see (1.13)),

$$\Delta u_i(x, y) \equiv 0.$$

Furthermore,

$$\frac{\partial H(x, y, p, v + \Delta v)}{\partial w_i} - \frac{\partial H(x, y, p, v)}{\partial w_i} = 0,$$

$$w = (z_1, \cdots, z_m, \cdots, z_{1y}, \cdots, z_{my}).$$

Therefore, $\eta_1 = 0$. Let us now compute η_2. We have:

$$\frac{\partial^2 H(x, y, p, v)}{\partial w_i \partial w_k} \equiv 0, \qquad i, k = 1, \cdots, 3m.$$

Hence,

$$\eta_2 = \sum_{i=1}^{m} \sum_{k=1}^{3m} \iint_G \left[\frac{\partial^2 H(x, y, p + \theta \Delta p, v + \Delta v)}{\partial u_i \partial w_k} - \frac{\partial^2 H(x, y, p + \theta_1 \Delta p, v + \Delta v}{\partial u_i \partial w_k} \right] \Delta u_i \, \Delta w_k \, dx \, dy.$$

Since $\Delta u_i(x, y) \equiv 0$, it follows that $\eta_2 = 0$. Consequently, in the case being considered, (1.21) takes the form:

$$\Delta S = -\iint_G [H(x, y, p, v + \Delta v) - H(x, y, p, v)] \, dx \, dy. \tag{1.35}$$

With the aid of the latter formula it is easy to prove the following theorem.

THEOREM 2. *In order that an admissible control $v(x, y)$ in the boundary value problem* (1.33) *be locally min-optimal* (*max-optimal*) *with respect to S, it is necessary and sufficient that it satisfy the maximum* (*minimum*) *condition.*

1.5. Control of a system with the aid of boundary conditions. Up to now we have assumed that the control is realized with the aid of the functions v occurring in (1.1) or (1.33). The boundary values (1.2) of the functions z_i were fixed. However, the method we have presented allows us to solve a more general problem.

Let the controlled process be described by the system (1.1), but let the boundary values of the functions z_i be given not by conditions (1.2) but with the aid of the differential equations

$$z_{iy}(0, y) = \varphi_i(y, z_1, \cdots, z_m, v^1), \quad z_{ix}(x, 0) = \psi_i(x, z_1, \cdots, z_m, v^2), \tag{1.36}$$

and the initial conditions

$$z_i(0, 0) = z_i^0, \qquad i = 1, \cdots, m, \tag{1.37}$$

where the functions φ_i and ψ_i are continuous in y and x and are twice continuously differentiable in the remaining arguments; v^1 and v^2 are the control parameters taking values from the region V^1 and V^2, respectively, in s- and t-dimensional Euclidean spaces.

The presence of parameters in (1.36) allows us to control the process with the aid of the boundary conditions. As the admissible controls in (1.36) we take the piecewise-continuous functions $v^1(y)$ and $v^2(x)$ with

values in the regions V^1 and V^2, respectively. It is well-known (for example, see [31, pp. 16–17]) that every pair of admissible controls $v^1(y)$, $v^2(x)$, with the help of (1.36) and (1.37), determines a unique pair of absolutely continuous functions $z(0, y)$, $z(x, 0)$. Everywhere in the following, by an admissible control in the boundary value problem (1.1)–(1.36)–(1.37) we shall mean the function

$$\omega(x, y) = (v(x, y), v^1(y), v^2(x)),$$

whose components are piecewise-continuous functions with values in the regions V, V^1 and V^2, respectively. Therefore, to each admissible control $\omega(x, y)$ there corresponds a unique solution of the boundary value problem (1.1)–(1.36)–(1.37) with the same smoothness conditions as were introduced for the boundary value problem (1.1)–(1.2).

We introduce the notation:

$$q = (z_1, \cdots, z_m, u_1, \cdots, u_m),$$

$$H_1(y, q, v^1) = \sum_{i=1}^{m} u_i \varphi_i(y, z, v^1),$$

$$H_2(x, q, v^2) = \sum_{i=1}^{m} u_i \psi_i(x, z, v^2).$$

We determine the functions u_i with the aid of (1.5) and (1.6).

The optimal problem with boundary conditions (1.36)–(1.37) has not yet been successfully solved in the general form. However, it can be solved by the method proposed above if the following conditions are satisfied:

$$\left.\frac{\partial f_k(x, y, z, z_x, z_y, v)}{\partial z_{iy}}\right|_{y=0} = \frac{\partial \psi_k(x, z, v^2)}{\partial z_i},$$

$$\left.\frac{\partial f_k(x, y, z, z_x, z_y, v)}{\partial z_{ix}}\right|_{x=0} = \frac{\partial \varphi_k(y, z, v^1)}{\partial z_i}, \tag{1.38}$$

$$k, i = 1, \cdots, m, \qquad v \in V, \qquad v^1 \in V^1, \qquad v^2 \in V^2.$$

Thus, in what follows we shall assume that conditions (1.38) are fulfilled and, consequently, the equalities

$$\left.\frac{\partial H(x, y, p, v)}{\partial z_{iy}}\right|_{y=0} = \frac{\partial H_2(x, q, v^2)}{\partial z_i},$$

$$\left.\frac{\partial H(x, y, p, v)}{\partial z_{ix}}\right|_{x=0} = \frac{\partial H_1(y, q, v^1)}{\partial z_i}, \tag{1.38'}$$

are valid for any function $u(u_1, \cdots, u_m)$ and for all v, v^1 and v^2 from the regions V, V^1 and V^2, respectively.

We shall say that an admissible control $\omega(x, y)$ in the boundary value

problem (1.1)–(1.36)–(1.37) satisfies the maximum condition if

$$H(x, y, p(x, y), v(x, y)) \quad ((=)) \quad \sup_{v \in V} H(x, y, p(x, y), v),$$

$$H_1(y, q(0, y), v^1(y)) \quad (=) \quad \sup_{v^1 \in V^1} H_1(y, q(0, y), v^1), \tag{1.39}$$

$$H_2(x, q(x, 0), v^2(x)) \quad (=) \quad \sup_{v^2 \in V^2} H(x, q(x, 0), v),$$

where $z(x, y)$ and $u(x, y)$ are the solutions of boundary value problems (1.1)–(1.36)–(1.37) and (1.5)–(1.6) corresponding to the control $\omega(x, y) = (v(x, y), v^1(y), v^2(x))$, while the symbol $(=)$ denotes that the equality is valid almost everywhere in the ranges of the arguments. The minimum is defined analogously.

THEOREM 3. *In order that an admissible control $\omega(x, y)$ in the boundary value problem* (1.1)–(1.36)–(1.37) *be min-optimal (max-optimal) with respect to S, it is necessary that it satisfy the maximum (minimum) condition.*

The proof of this theorem is carried out by exactly the same scheme as was the proof of Theorem 1: we start by deriving the formula for the increment of the functional, next we estimate the remainder term, and finally we prove the theorem.

To obtain the formula for the increment of the functional we take an arbitrary admissible control $\omega(x, y)$ and we denote by $z(x, y)$ and $u(x, y)$ the solutions of boundary value problems (1.1)–(1.36)–(1.37) and (1.5)–(1.6) corresponding to it. Then the following equality is valid:

$$I[p, \omega] = \iint_G \left[\sum_{i=1}^{m} u_i z_{ixy} - H(x, y, p, v)\right] dx\, dy + \int_0^X \left[\sum_{i=1}^{m} u_i(x, 0) z_{ix}(x, 0) - H_2(x, q(x, 0), v^2)\right] dx + \int_0^Y \left[\sum_{i=1}^{m} u_i(0, y) z_{iy}(0, y) - H_1(y, q(0, y), v^1)\right] dy = 0.$$

Let $\Delta\omega$ denote an arbitrary admissible increment of the control $\omega(x, y)$, and Δz and Δu, the increments of the functions $z(x, y)$ and $u(x, y)$ corresponding to it. Obviously,

$$\Delta I = I[p + \Delta p, \omega + \Delta\omega] - I[p, \omega] = 0.$$

For the transformation of ΔI we shall start with the equality

$$\iint_G pq_{xy}\, dy\, dx + \int_0^X p(x, 0) q_x(x, 0)\, dx + \int_0^Y p(0, y) q_y(0, y)\, dy = \iint_G qp_{xy}\, dy\, dx - \int_0^X q(x, Y) p_x(x, Y)\, dx - \int_0^Y q(X, y) p_y(X, y)\, dy + p(X, Y) q(X, Y) + p(0, 0) q(0, 0), \tag{1.40}$$

valid for any twice piecewise-continuously differentiable functions p and q, which can be obtained from the Green's formula (1.16).

Setting $p = \Delta u_i$, $q = \Delta z_i$ in (1.40) and taking conditions (1.38′) into account, we obtain:

$$\begin{aligned}
&\iint_G \sum_{i=1}^{m} \Delta u_i \Delta z_{ixy}\, dx\, dy + \int_0^X \sum_{i=1}^{m} \Delta u_i(x, 0)\Delta z_{ix}(x, 0)\, dx \\
&\qquad + \int_0^Y \sum_{i=1}^{m} \Delta u_i(0, y)\Delta z_{iy}(0, y)\, dy \\
&= \iint_G \sum_{i=1}^{m} \left[\Delta \frac{\partial H}{\partial z_i} \Delta z_i + \Delta \frac{\partial H}{\partial z_{ix}} \Delta z_{ix} + \Delta \frac{\partial H}{\partial z_{iy}} \Delta z_{iy} \right] dx\, dy \\
&\qquad - \int_0^X \sum_{i=1}^{m} \left\{ \left[\Delta u_{ix}(x, Y) + \Delta \frac{\partial H(x, Y, p(x, Y), v)}{\partial z_{iy}} \right] \Delta z_i(x, Y) \right. \\
&\qquad \left. - \Delta \frac{\partial H(x, 0, p(x, 0), v)}{\partial z_{iy}} \Delta z_i(x, 0) \right\} dx \\
&\qquad - \int_0^Y \sum_{i=1}^{m} \left\{ \left[\Delta u_{iy}(X, y) + \Delta \frac{\partial H(X, y, p(X, y), v)}{\partial z_{ix}} \right] \Delta z_i(X, y) \right. \\
&\qquad \left. - \Delta \frac{\partial H(0, y, p(0, y), v)}{\partial z_{ix}} \Delta z_i(0, y) \right\} dy \\
&= \iint_G \sum_{i=1}^{m} \left[\Delta \frac{\partial H}{\partial z_i} \Delta z_i + \Delta \frac{\partial H}{\partial z_{ix}} \Delta z_{ix} + \Delta \frac{\partial H}{\partial z_{iy}} \Delta z_{iy} \right] dx\, dy \\
&\qquad + \int_0^X \sum_{i=1}^{m} \Delta \frac{\partial H_2(x, q(x, 0), v^2)}{\partial z_i} \Delta z_i(x, 0)\, dx \\
&\qquad + \int_0^Y \sum_{i=1}^{m} \Delta \frac{\partial H_1(y, q(0, y), v^1)}{\partial z_i} \Delta z_i(0, y)\, dy.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\iint_G \sum_{i=1}^{m} \Delta u_i \Delta z_{ixy}\, dx\, dy + \int_0^X \sum_{i=1}^{m} \Delta u_i(x, 0)\Delta z_{ix}(x, 0)\, dx \\
&\qquad + \int_0^Y \sum_{i=1}^{m} \Delta u_i(0, y)\Delta z_{iy}(0, y)\, dy \\
&= \iint_G \sum_{i=1}^{m} \Delta \frac{\partial H}{\partial u_i} \Delta u_i\, dx\, dy + \int_0^X \sum_{i=1}^{m} \Delta \frac{\partial H_2(x, q(x, 0), v^2)}{\partial u_i} \Delta u_i(x, 0)\, dx \\
&\qquad + \int_0^Y \sum_{i=1}^{m} \Delta \frac{\partial H_1(y, q(0, y), v^1)}{\partial u_i} \Delta u_i(0, y)\, dx.
\end{aligned}$$

From the last two equalities we obtain:

$$\iint_G \sum_{i=1}^{m} \Delta u_i \Delta z_{ixy}\, dx\, dy + \int_0^X \sum_{i=1}^{m} \Delta u_i(x, 0)\Delta z_{ix}(x, 0)\, dx$$

$$+ \int_0^Y \sum_{i=1}^{m} \Delta u_i(0, y)\Delta z_{iy}(0, y)\, dy \tag{1.41}$$

$$= \frac{1}{2}\left[\iint_G \sum_{i=1}^{4m} \Delta \frac{\partial H}{\partial p_i} \Delta p_i\, dx\, dy + \int_0^X \sum_{i=1}^{2n} \Delta \frac{\partial H_2(x, q(x, 0), v^2)}{\partial q_i} \Delta q_i(x, 0)\, dx\right.$$

$$\left. + \int_0^Y \sum_{i=1}^{2m} \Delta \frac{\partial H_1(y, q(0, y), v^1)}{\partial q_i} \Delta q_i(0, y)\, dy\right].$$

Furthermore, by the same method that we used to derive (1.18) and (1.19), we find:

$$\iint_G \sum_{i=1}^{m} u_i \Delta z_{ixy}\, dx\, dy + \int_0^X \sum_{i=1}^{m} u_i(x, 0)\Delta z_{ix}(x, 0)\, dx$$

$$+ \int_0^Y \sum_{i=1}^{m} u_i(0, y)\Delta z_{iy}(0, y)\, dy = -\sum_{i=1}^{m} A_i \Delta z_i(X, Y)$$

$$+ \iint_G \sum_{i=1}^{m} \left[\frac{\partial H}{\partial z_i} \Delta z_i + \frac{\partial H}{\partial z_{ix}} \Delta z_{ix} + \frac{\partial H}{\partial z_{iy}} \Delta z_{iy}\right] dx\, dy \tag{1.42}$$

$$+ \int_0^X \sum_{i=1}^{m} \frac{\partial H_2(x, q(x, 0), v^2)}{\partial z_i} \Delta z_i(x, 0)\, dx$$

$$+ \int_0^Y \sum_{i=1}^{m} \frac{\partial H_1(y, q(0, y), v^1)}{\partial z_i} \Delta z_i(0, y)\, dy,$$

$$\iint_G \sum_{i=1}^{m} \Delta u_i z_{ixy}\, dx\, dy + \int_0^X \sum_{i=1}^{m} \Delta u_i(x, 0) z_{ix}(x, 0)\, dx$$

$$+ \int_0^Y \sum_{i=1}^{m} \Delta u_i(0, y) z_{iy}(0, y)\, dy = \iint_G \sum_{i=1}^{m} \frac{\partial H}{\partial u_i} \Delta u_i\, dx\, dy$$

$$+ \int_0^X \sum_{i=1}^{m} \frac{\partial H_2(x, q(x, 0), v^2)}{\partial u_i} \Delta u_i(x, 0)\, dx \tag{1.43}$$

$$+ \int_0^Y \sum_{i=1}^{m} \frac{\partial H_1(y, q(0, y), v^1)}{\partial u_i} \Delta u_i(0, y)\, dy.$$

Taking into account (1.41), (1.42), (1.43), and the fact that $\Delta I = 0$, by the same method we use in the proof of Theorem 1, we obtain the formula

for the increment of the functional:

$$\Delta S = -\iint_G [H(x, y, p, v + \Delta v) - H(x, y, p, v)]\, dx\, dy$$

$$- \int_0^X [H_2(x, q(x, 0), v^2 + \Delta v^2) - H_2(x, q(x, 0), v^2)]\, dx \tag{1.44}$$

$$- \int_0^Y [H_1(y, q(0, y), v^1 + \Delta v^1) - H_1(y, q(0, y), v^1)]\, dy - \eta,$$

where $\eta = \eta_1 + \eta_2 + \eta_3$,

$$\eta_1 = \frac{1}{2} \iint_G \sum_{i=1}^{4m} \left\{ \left[\frac{\partial H(x, y, p, v + \Delta v)}{\partial p_i} - \frac{\partial H(x, y, p, v)}{\partial p_i} \right] \Delta p_i \right.$$

$$+ \sum_{k=1}^{4m} \left[\frac{\partial^2 H(x, y, p + \theta \Delta p, v + \Delta v)}{\partial p_i\, \partial p_k} \right.$$

$$\left. \left. - \frac{\partial^2 H(x, y, p + \theta_1 \Delta p, v + \Delta v)}{\partial p_i\, \partial p_k} \right] \Delta p_i\, \Delta p_k \right\} dx\, dy,$$

$$\eta_2 = \frac{1}{2} \int_0^X \sum_{i=1}^{2m} \left\{ \left[\frac{\partial H_2(x, q(x, 0), v^2 + \Delta v^2)}{\partial q_i} - \frac{\partial H_2(x, q(x, 0), v^2)}{\partial q_i} \right] \right.$$

$$\cdot \Delta q_i(x, 0) + \sum_{k=1}^{2m} \left[\frac{\partial^2 H_2(x, q(x, 0) + \theta_2 \Delta q, v^2 + \Delta v^2)}{\partial q_i\, \partial q_k} \right. \tag{1.44'}$$

$$\left. \left. - \frac{\partial^2 H_2(x, q(x, 0) + \theta_3 \Delta q, v^2 + \Delta v^2)}{\partial q_i\, \partial q_k} \right] \Delta q_i\, \Delta q_k \right\} dx,$$

$$\eta_3 = \frac{1}{2} \int_0^Y \sum_{i=1}^{m} \left\{ \left[\frac{\partial H_1(y, q(0, y), v^1 + \Delta v^1)}{\partial q_1} \right. \right.$$

$$\left. - \frac{\partial H_1(y, q(0, y), v^1)}{\partial q_i} \right] \Delta q_i(0, y)$$

$$+ \sum_{k=1}^{2m} \left[\frac{\partial^2 H_1(y, q(0, y) + \theta_4 \Delta q, v^1 + \Delta v^1)}{\partial q_i \partial q_k} \right.$$

$$\left. \left. - \frac{\partial^2 H_1(y, q(0, y) + \theta_5 \Delta q, v^1 + \Delta v^1)}{\partial q_i \partial q_k} \right] \Delta q_i \Delta q_k \right\} dy.$$

Let us now estimate the remainder term η in (1.44). The quantities η_2 and η_3 are determined in terms of the values of the functions z and u on the boundary of region G. By virtue of the Lipschitz condition, from (1.36) and (1.37) we get:

$$\sum_{i=1}^{m} |\, \Delta z_i(0, y)\, | \leqq N \int_0^y \sum_{i=1}^{m} |\, \Delta z_i(0, y)\, |\, dy + P \int_0^Y \sum_{k=1}^{s} |\, \Delta v_k^1(y)\, |\, dy,$$

$$\sum_{i=1}^{m} | \Delta z_i(x, 0) | \leqq N_1 \int_0^x \sum_{i=1}^{m} | \Delta z_i(x, 0) | \, dx + P_1 \int_0^X \sum_{k=1}^{t} | \Delta v_k^2(x) | \, dx.$$

Hence, in accordance with the lemma referred to above it follows that

$$\begin{aligned} | \Delta z_i(0, y) | &\leqq M_0 \int_0^Y \sum_{i=1}^{s} | \Delta v_i^1(y) | \, dy, \\ | \Delta z_i(x, 0) | &\leqq M_1 \int_0^X \sum_{i=1}^{t} | \Delta v_i^2(x) | \, dx. \end{aligned} \tag{1.45}$$

We introduce the notation:

$$\alpha(x, y) = \sum_{i=1}^{m} | \Delta z_{ix} |, \qquad \beta = \sum_{i=1}^{m} | \Delta z_{iy} |, \qquad \gamma = \sum_{i=1}^{m} | \Delta z_i |,$$

$$| \Delta v | = \sum_{k=1}^{r} | \Delta v_k |, \qquad \Delta v^1 = \sum | \Delta v_k^1(y) |, \qquad | \Delta v^2 | = \sum | \Delta v_k^2(x) |.$$

Since the functions f_i satisfy a Lipschitz condition, then just as in the derivation of (1.23) we obtain:

$$\begin{aligned} \alpha(x, y) &\leqq N_2 \int_0^y \alpha(x, y) \, dy + N_2 \int_0^\eta [\beta(x, y) + \gamma(x, y)] \, dy \\ &\quad + N_3 \int_0^Y \Delta v(x, y) \, dy + N_4 \int_0^X \Delta v^2(x) \, dx + N_5 \Delta v^2(x), \\ \beta(x, y) &\leqq M_2 \int_0^x \beta(x, y) \, dx + M_2 \int_0^\xi [\alpha(x, y) + \gamma(x, y)] \, dx \\ &\quad + M_3 \int_0^X \Delta v(x, y) \, dx + M_4 \int_0^Y \Delta v^1(y) \, dy + M_5 \Delta v^1(y), \\ \tau(x, y) &\leqq \int_0^x \alpha(x, y) \, dx, \qquad \gamma(x, y) \leqq \int_0^y \beta(x, y) \, dy, \end{aligned}$$

where $0 \leqq x \leqq \xi \leqq X$, $0 \leqq y \leqq \eta \leqq Y$. Hence we find:

$$\begin{aligned} \alpha(x, y) &\leqq N_6 \int_0^\eta [\beta(x, y) + \gamma(x, y)] \, dy + N_7 \int_0^Y \Delta v(x, y) \, dy \\ &\quad + N_8 \int_0^X \Delta v^2(x) \, dx + N_9 \Delta v^2(x), \\ \beta(x, y) &\leqq M_6 \int_0^\xi [\alpha(x, y) + \gamma(x, y)] \, dx + M_7 \int_0^X \Delta v(x, y) \, dx \\ &\quad + M_8 \int_0^Y \Delta v^1(y) \, dy + M_9 \Delta v^1(y). \end{aligned}$$

Taking the estimate for the function γ into account we shall have:

$$(*)\quad \begin{aligned} \alpha(x,y) &\leqq N_{10}\int_0^{\eta}\beta(x,y)\,dy + N_7\int_0^{Y}\Delta v(x,y)\,dy \\ &\quad + N_8\int_0^{X}\Delta v^2(x)\,dx + N_9\Delta v^2(x), \\ \beta(x,y) &\leqq M_{10}\int_0^{\xi}\alpha(x,y)\,dx + M_7\int_0^{X}\Delta v(x,y)\,dx \\ &\quad + M_8\int_0^{Y}\Delta v^1(y)\,dy + M_9\Delta v^1(y). \end{aligned}$$

From these inequalities we get:

$$\begin{aligned} \alpha(\xi,\eta) &\leqq N_{11}\int_0^{\xi}\int_0^{\eta}\alpha(x,y)\,dx\,dy + N_{12}\int_0^{X}\int_0^{Y}\Delta v(x,y)\,dx\,dy \\ &\quad + N_7\int_0^{Y}\Delta v(\xi,y)\,dy + N_{13}\int_0^{Y}\Delta v^1(y)\,dy + N_8\int_0^{X}\Delta v^2(x)\,dx + N_9\Delta v^2(\xi), \\ \beta(\xi,\eta) &\leqq M_{11}\int_0^{\eta}\int_0^{\xi}\beta(x,y)\,dx\,dy + M_{12}\int_0^{X}\int_0^{Y}\Delta v(x,y)\,dx\,dy \\ &\quad + M_7\int_0^{X}\Delta v(x,\eta)\,dx + M_8\int_0^{Y}\Delta v^1(y)\,dy + M_{13}\int_0^{X}\Delta v^2(x)\,dx + M_9\Delta v^1(\eta). \end{aligned}$$

Integrating the first of these inequalities with respect to ξ in the limits from 0 to ξ and applying the lemma mentioned above, we obtain:

$$\begin{aligned} \gamma \leqq \int_0^{\xi}\alpha(\xi,\eta)\,d\xi &\leqq N_{14}\iint_G \Delta v(x,y)\,dx\,dy + N_{15}\int_0^{Y}\Delta v^1(y)\,dy \\ &\quad + N_{16}\int_0^{X}\Delta v^2(x)\,dx. \end{aligned}$$

Analogously, from the second inequality we obtain:

$$\begin{aligned} \int_0^{\eta}\beta(\xi,\eta)\,d\eta &\leqq M_{14}\iint_G \Delta v(x,y)\,dx\,dy + M_{15}\int_0^{X}\Delta v^2(x)\,dx \\ &\quad + M_{16}\int_0^{Y}\Delta v^1(y)\,dy. \end{aligned}$$

Hence also from (*) we get:

$$\begin{aligned} |\,\Delta z_i(x,y)\,| &\leqq N_{14}\iint_G \Delta v(x,y)\,dx\,dy + N_{15}\int_0^{Y}\Delta v^1(y)\,dy \\ &\quad + N_{16}\int_0^{X}\Delta v^2(x)\,dx, \end{aligned}$$

$$
\begin{aligned}
|\Delta z_{ix}(x,y)| &\leqq N_{17}\iint_G \Delta v(x,y)\,dx\,dy + N_{18}\int_0^Y \Delta v(x,y)\,dy \\
&\quad + N_{19}\int_0^X \Delta v^2\,dx + N_{20}\int_0^Y \Delta v^1\,dy + N_9\Delta v^2(x), \\
|\Delta z_{iy}(x,y)| &\leqq M_{17}\iint_G \Delta v(x,y)\,dx\,dy + M_{18}\int_0^X \Delta v(x,y)\,dx \\
&\quad + M_{19}\int_0^Y \Delta v^1\,dy + M_{20}\int_0^X \Delta v^2\,dx + M_9\Delta v^1(y).
\end{aligned}
\tag{1.46}
$$

Analogously we have:

$$
\begin{aligned}
|\Delta u_i(x,y)| &\leqq M_{21}\iint_G \Delta v(x,y)\,dx\,dy + M_{22}\int_0^Y \Delta v^1(y)\,dy \\
&\quad + M_{23}\int_0^X \Delta v^2(x)\,dx.
\end{aligned}
\tag{1.47}
$$

If $\Delta v_i^1(y) = \Delta v_i^2(x) \equiv 0$, but $\Delta v_i(x,y) \not\equiv 0$, then from (1.46) and (1.47) we obtain (1.27) and (1.29). Having established this fact we proceed directly to the proof of the theorem.

For definiteness let the admissible control $\omega(x,y) = (v(x,y), v^1(y), v^2(x))$ be min-optimal with respect to S. Then, the inequality $\Delta S \geqq 0$ is valid for an arbitrary $\Delta\omega$. Let us assume that the theorem is false. Then in the closed region G we can find either a subregion G_1 in which the first equality in (1.39) is not satisfied, or a line segment lying on the boundary of G on which one of the last two equalities in (1.39) is not satisfied.

In the first of these cases we can find an admissible control $\tilde{v} \in V$ such that

$$H(x,y,p(x,y),\tilde{v}) - H(x,y,p(x,y),v) > 0 \quad \text{if} \quad (x,y) \in G_1 .$$

Then, there exists a $\delta > 0$ for which

$$H(x,y,p(x,y),\tilde{v}) - H(x,y,p(x,y),v) > \delta$$

if $(x,y) \in G_\epsilon \subset G_1$, where G_ϵ is a circle of radius ϵ, lying together with its boundary inside the region G_1. Setting $\Delta v^1 = \Delta v^2 \equiv 0$ and repeating the argument used in the proof of Theorem 1, we obtain $\Delta S < 0$ with the aid of estimates (1.46) and (1.47). But this contradicts the hypothesis, and hence the first equality in the maximum conditions is satisfied.

Let us consider the second case. For the sake of definiteness we assume that it is the last equality in (1.39) that is not satisfied. Then, there exist a control $\tilde{v}^2 \in V^2$ and a segment l of the boundary $y = 0$ of region G such that

$$H_2(x,q(x,0),\tilde{v}^2) - H_2(x,q(x,0),v^2) > 0$$

if $x \in l$. Consequently, we can find a number $\delta > 0$ such that

$$H_2(x, q(x, 0), \bar{v}^2) - H_2(x, q(x, 0), v^2) > \delta$$

if $x \in l_\epsilon \subset l$, where l_ϵ is a segment of length ϵ. Let us set

$$\Delta v_i = \Delta v_i^1 = 0$$

and consider the auxiliary control

$$\bar{\omega}^1(x, y) = (v, v^1, \bar{v}^2),$$

where

$$\bar{v}^2 = \begin{cases} v^2 & \text{if } x \notin l_\epsilon, \\ \bar{v}^2 & \text{if } x \in l_\epsilon. \end{cases}$$

Then, the remainder term η in (1.44) coincides with η_2 (see (1.44′)), where $\Delta v^2 = \bar{v}^2 - v^2$ and, consequently, Δv^2 is nonzero only if $x \in l_\epsilon$.

Since the functions $\partial H_2/\partial q_i$ satisfy a Lipschitz condition, and the $\partial^2 H_2/\partial q_i \partial q_k$ are bounded, by virtue of estimates (1.46) and (1.47) we get:

$$|\eta| \leqq M\epsilon \int_{l_\epsilon} \left(\sum_{k=1}^{t} |\Delta v_k^2(x)| \right)^2 dx,$$

where M is a constant not depending on ϵ. With the help of this estimate it is easy to establish that $\Delta S > 0$, but this contradicts the assumption of min-optimality with respect to S of the control $\omega(x, y)$.

Thus Theorem 3 is completely proved.

Now let the controlled process be described by the system of linear equations

$$z_{ixy} = \sum_{k=1}^{m} [c_{ik}(x, y) z_{kx} + d_{ik}(x, y) z_{ky} + g_{ik}(x, y) z_k] + f_i(v), \qquad i = 1, \cdots, m, \tag{1.48}$$

with the supplementary conditions:

$$\begin{aligned} z_{iy}(0, y) &= \sum_{k=1}^{m} c_{ik}(0, y) z_k + \varphi_i(v^1), \\ z_{ix}(x, 0) &= \sum_{k=1}^{m} d_{ik}(x, 0) z_k(x, 0) + \psi_i(v^2), \end{aligned} \tag{1.49}$$

$$z_i(0, 0) = z_i^0, \qquad i = 1, \cdots, m. \tag{1.50}$$

The requirements (1.38) imply that the coefficients in (1.49) have to be specially chosen. Just as in the proof of Theorem 2 we find that in the case being considered the remainder term η in (1.45) equals zero and, consequently,

$$\Delta S = -\iint_G [H(x, y, p, v + \Delta v) - H(x, y, p, v)]\, dx\, dy$$
$$- \int_0^X [H_2(x, q, v^2 + \Delta v^2) - H_2(x, q, v^2)]\, dx$$
$$- \int_0^Y [H_1(y, q, v^1 + \Delta v^1) - H_1(y, q, v^1)]\, dy.$$

From this formula there follows the validity of the following theorem.

THEOREM 4. *In order that an admissible control $\omega(x, y)$ in the boundary value problem* (1.48)–(1.49)–(1.50) *be locally min-optimal (max-optimal) with respect to the functional $S = \sum A_i z_i(X, Y)$, it is necessary and sufficient that it satisfy the maximum (minimum) condition.*

2. Other optimal control problems for hyperbolic systems. Let us consider the same problem of minimizing the functional $S = \sum A_i z_i(X, Y)$, in which the controlled process is described by the boundary value problem (1.1)–(1.36)–(1.37), where z_i^0, $i = 1, \cdots, m$, are given numbers. The admissible controls determined in §1 were constrained by the requirement that the numbers $z_i(X, Y)$ corresponding to them should belong to the convex set D in the space of the variables $z_1, \cdots, z_m$. Thus, in the problem being considered the admissible controls transfer, by (1.1) and (1.36), the point $(z_1^0, \cdots, z_m^0)$ to a point in region D. In what follows we shall assume that the convex region D contains an interior point and is closed.

To solve the problem, just as in [16] we introduce the function

$$A(z) = \langle A, z \rangle = \sum A_i z_i,$$

and we denote by D^* the set of points $z^* \in D^*$ at which

$$A(z^*) = \min_{z \in D} A(z).$$

If the set D^* is not empty, then

$$A(z^*) \leqq A(z), \qquad z^* \in D^*, \qquad z \in D,$$

and, consequently, the functional

$$S = \sum A_i z_i(X, Y),$$

defined on the solutions of the boundary value problem (1.1)–(1.36)–(1.37), cannot take values less than $A(z^*)$. If an admissible control exists which transfers the point z^0 to any point of the set D^*, then this control is min-optimal with respect to S. In this case the problem is reduced to seeking the controls which transfer z^0 to the given region. We shall not consider such problems in what follows, i.e., we shall assume that there are no admissible controls transferring z^0 onto D^*.

2.1. Necessary optimality conditions. We shall say that an admissible control $\omega(x, y)$ satisfies the maximum condition relative to a given func-

tion $u(x, y)$ if the conditions (1.39) are satisfied, where $z(x, y)$ is the solution of the boundary value problem (1.1)–(1.36)–(1.37).

THEOREM 5. *If $\omega(x, y)$ is a control which is min-optimal with respect to S, and $z(x, y)$ is the solution of the problem* (1.1)–(1.36)–(1.37) *corresponding to it, then there exists a vector-function $u(x, y)$ relative to which the control $\omega(x, y)$ satisfies the maximum condition.*

Let $\omega(x, y) = (v(x, y), v^1(y), v^2(x))$ be a control which is min-optimal with respect to S, and let $z(x, y)$ be the solution of the boundary value problem (1.1)–(1.36)–(1.37) corresponding to it. We denote by $D^-(D^+)$ the part of region D for which $A(z) \leqq \sum A_i z_i(X, Y)$, $z \in D^-$, $(A(z) \geqq \sum A_i z_i(X, Y)$, $z \in D^+)$. The common part of these closed convex regions is the plane

$$\sum_{i=1}^{m} A_i(z_i - z_i(X, Y)) = 0$$

on which the point $z(X, Y)$ lies. Admissible controls transferring the point z^0 to the interior of region D^- do not exist since the control $\omega(x, y)$ is min-optimal with respect to S. Having noted this fact we introduce the variational equations by assuming that all the admissible controls are piecewise continuous. We choose arbitrary points (x_i, y_j), $i, j \geqq 0$ $(x_0 = 0, y_0 = 0)$, in region G, and by G_{ij} denote the rectangle, formed by the neighboring points (x_ν, y_μ), whose lower left corner lies at the point (x_i, y_j). We construct the squares I_{ij}, $x_{i+1} - \tau \leqq x \leqq x_{i+1}$, $y_{j+1} - \tau \leqq y \leqq y_{j+1}$, where the number τ is chosen so small that for the given set of points (x_i, y_j) these squares have no points in common.[1]

Let us take arbitrary piecewise-continuous vector functions $\alpha_{ij}(x, y)$, $\beta_i(x)$ and $\gamma_j(y)$, defined for $x, y \in [0, 1]$ and taking values, respectively, in the ranges V, V^2 and V^1 of the control parameters v, v^2 and v^1. We introduce the functions

$$v_b(x, y, \alpha_{ik}) = \begin{cases} v(x, y) & \text{if } (x, y) \notin I_{i-1,k-1}, \\ \alpha_{ik}\left(\dfrac{x_i - x}{\tau}, \dfrac{y_k - y}{\tau}\right) & \text{if } (x, y) \in I_{i-1,k-1}, \end{cases}$$

$$v_b^1(y, \gamma_k) = \begin{cases} v^1(y) & \text{if } y \notin [y_k - \tau, y_k], \\ \gamma_k\left(\dfrac{y_k - y}{\tau}\right) & \text{if } y \in [y_k - \tau, y_k), \end{cases}$$

$$v_b^2(x, \beta_i) = \begin{cases} v^2(x) & \text{if } x \notin [x_i - \tau, x_i), \\ \beta_i\left(\dfrac{x_i - x}{\tau}\right) & \text{if } x \in [x_i - \tau, x_i). \end{cases}$$

[1] In the case when the number of points (x_i, y_j) is finite, there is no doubt about the existence of such a τ.

We shall call the function

$$\omega_b(x, y, \alpha_{ik}, \beta_i, \gamma_k) = (v_b(x, y, \alpha_{ik}), v_b^1(y, \gamma_k), v_b^2(x, \beta_i))$$

the variational control and denote by Ω the collection of all possible variational controls corresponding to all possible squares I_{ij} and all possible functions α_{ik}, β_i and γ_k of the type indicated above. We denote by $z(x, y, \omega_b)$ the solution of boundary value problem (1.1)–(1.36)–(1.37) corresponding to the control $\omega_b \in \Omega$. Then the function

$$\Delta z(x, y, \omega) = z(x, y, \omega_b) - z(x, y, \omega)$$

is the solution of the boundary value problem

$$\Delta z_{ixy}(x, y, \omega) = \Delta \frac{\partial H}{\partial u_i}, \qquad (x, y) \in G, \tag{2.1}$$

$$\begin{aligned} \Delta z_{iy}(0, y, \omega) &= \Delta \frac{\partial H_1}{\partial u_i}, \qquad y \in [0, Y], \\ \Delta z_i(x, 0, \omega) &= \Delta \frac{\partial H_2}{\partial u_i}, \qquad x \in [0, X], \end{aligned} \tag{2.2}$$

$$\Delta z_i(0, 0, \omega) = 0, \qquad i = 1, \cdots, m. \tag{2.3}$$

Since equations (2.2) are equations in ordinary derivatives, according to the results of [16] it follows from (2.2) and (2.3) that

$$\delta z_i(0, y, \omega) = \int_0^y \sum_{k=1}^m \frac{\varphi_i(y, z(0, y, \omega), \omega)}{\partial z_k} \delta z_k(0, y, \omega)\, dy + \sum_{j=1}^k R_i[y_j, \gamma_j], \qquad y_k < y < y_{k+1},$$

$$\delta z_i(x, 0, \omega) = \int_0^x \sum_{k=1}^m \frac{\psi_i(x, z(x, 0, \omega), \omega)}{\partial z} \delta z_k(x, 0, \omega)\, dx + \sum_{j=1}^l Q_i[x_j, \beta_j], \qquad x_i < x < x_{l+1},$$

where

$$\delta z_i = \lim_{\tau \to 0} \frac{\Delta z_i}{\tau},$$

$$R_i[y_j, \gamma_j] = \int_0^1 [\varphi_i(y_j, z(0, y_j, v^1), \gamma_j(y)) - \varphi_i(y_j, z(0, y_j, v^1), v^1(y))]\, dy,$$

$$Q_i(x_j, \beta_j) = \int_0^1 [\psi_i(x_j, z(x_j, 0, v^2), \beta_j(x)) - \psi_i(x_j, z(x_j, 0, v^2), v^2(x))]\, dx.$$

In the same reference it is shown that

$$\begin{aligned}\delta z_i(x, 0, \omega) &= \sum_{j=1}^{l}\sum_{s=1}^{m} A_{is}(x, x_j)Q_s(x_j, \omega_b),\\ \delta z_i(0, y, \omega) &= \sum_{j=1}^{k}\sum_{s=1}^{m} B_{is}(y, y_j)R_s(y_j, \omega_b),\end{aligned} \tag{2.4}$$

where the matrices A_{is} and B_{is} are independent of the choice of the functions β_i and γ_i.

From (2.1) it follows that

$$\begin{aligned}\Delta z_i(x, y, \omega) &= \Delta z_i(x, 0, \omega) + \Delta z_i(0, y, \omega)\\ &+ \int_0^x\int_0^y \sum_{s=1}^{3m} \frac{\partial f_i(x, y, w, v)}{\partial w_s}\Delta w_s\, dx\, dy + \sum_{j=1}^{l}\sum_{\nu=1}^{k} I_{ij\nu}[\omega_b] + E_i,\\ &(x, y) \in G_{l,k} - I_{l,k}.\end{aligned} \tag{2.5}$$

Here we have introduced the following notation:

$$I_{ijk}[\omega_b] = \int_{x_j-\tau}^{x_j}\int_{y_k-\tau}^{y_k} F_i(x, y, w, \alpha_{jk}, v)\, dy\, dx,$$

$$\begin{aligned}E_i &= \frac{1}{2}\sum_{s,q=1}^{3m}\int_0^x\int_0^y \frac{\partial^2 f_i(x, y, w + \theta\Delta w, v_b)}{\partial w_s\,\partial w_q}\Delta w_s\Delta w_q\, dy\, dx\\ &+ \sum_{j=1}^{l}\sum_{\nu=1}^{k}\int_{x_j-\tau}^{x_j}\int_{y_\nu-1}^{y_\nu}\sum_{s=1}^{3m}\frac{\partial F_i(x, y, w, \alpha_{i\nu}, v)}{\partial w_s}\Delta w_s\, dy\, dx,\end{aligned}$$

where

$$F_i(x, y, w, \alpha_{j\nu}, v) = f_i\left(x, y, w, \alpha_{j\nu}\left(\frac{x_j - x}{\tau}, \frac{y_\nu - y}{\tau}\right)\right) - f_i(x, y, w, v).$$

Analogously we find that

$$\begin{aligned}\Delta z_{ix}(x, y, \omega) &= \Delta z_{ix}(x, 0, \omega)\\ &+ \int_0^y \sum_{s=1}^{3m}\frac{\partial f_i(x, y, w, v)}{\partial w_s}\Delta w_s\, dy + \sum_{\nu=1}^{k} E_{i\nu}[\omega_b] + \bar{E}_i,\end{aligned} \tag{2.6}$$

$$\begin{aligned}\Delta z_{iy}(x, y, \omega) &= \Delta z_{iy}(0, y, \omega)\\ &+ \int_0^x \sum_{s=1}^{3m}\frac{\partial f_i(x, y, w, v)}{\partial w_s}\Delta w_s\, dx + \sum_{j=1}^{l} F_{ij}[\omega_b] + \bar{F}\end{aligned} \tag{2.7}$$

if $(x, y) \in G_{l,k} - I_{l,k}$, where

$$E_{ip} = \begin{cases} 0 & \text{when } x_l < x < x_{l+1} - \tau,\\ \displaystyle\int_{y_p-\tau}^{y_p} F_i(x, y, w, \alpha_{l,p}, v)\, dy & \text{when } x_{l+1} - \tau \leqq x < x_{l+1},\end{cases}$$

$$F_{ij} = \begin{cases} 0 & \text{when} \quad y_k < y < y_{k+1} - \tau, \\ \displaystyle\int_{x_j-\tau}^{x_j} F_i(x, y, w, \alpha_{jk}, v)\, dx & \text{when} \quad y_{k+1} - \tau \leqq y < y_{k+1}, \end{cases}$$

$$\bar{E}_i = \begin{cases} J_i = \dfrac{1}{2}\displaystyle\int_0^y \sum_{s,q=1}^{3m} \frac{\partial^2 f_i(x, y, w + \theta\Delta w, v_b)}{\partial w_s\, \partial w_q}\, \Delta w_s\, \Delta w_q\, dy \\ \qquad\qquad \text{when} \quad x_l < x < x_{l+1} - \tau, \\ J_i + \displaystyle\sum_{p=1}^{k} \int_{y_p-\tau}^{y_p} \sum_{s=1}^{3m} \frac{\partial F_i(x, y, w, \alpha_{l,p}, v)}{\partial w_s}\, \Delta w_s\, dy \\ \qquad\qquad \text{when} \quad x_{l+1} - \tau \leqq x < x_{l+1}, \end{cases}$$

$$\bar{F}_i = \begin{cases} L_i = \dfrac{1}{2}\displaystyle\sum_{s,q=1}^{3m} \int_0^x \frac{\partial^2 f_i(x, y, w + \theta\Delta w, v_b)}{\partial w_s\, \partial w_q}\, \Delta w_s\, \Delta w_q\, dx \\ \qquad\qquad \text{when} \quad y_k < y < y_{k+1} - \tau, \\ L_i + \displaystyle\sum_{j=1}^{l} \int_{x_j-\tau}^{x_j} \sum_{s=1}^{3m} \frac{\partial F_i(x, y, w, \alpha_{ij}, v)}{\partial w_s}\, \Delta w_s\, dx \\ \qquad\qquad \text{when} \quad y_{k+1} - \tau \leqq y < y_{k+1}. \end{cases}$$

From what we have proved earlier (see (1.46)) we can find a positive number N such that $|\Delta w_i(x, y)| \leqq N\tau$ and, consequently,

$$|\bar{E}_i| \leqq N_1\tau^2, \qquad |E_i| \leqq N_2\tau^2, \qquad |\bar{F}_i| \leqq N_3\tau^2,$$

and, uniformly in the x and y,

$$\lim_{\tau\to 0} \frac{E_i}{\tau} = \lim_{\tau\to 0} \frac{\bar{E}_i}{\tau} = \lim_{\tau\to 0} \frac{F_i}{\tau} = 0.$$

Making the substitutions $\xi\tau = x_j - x$, $\eta\tau = y_\nu - y$ and going to the limit we obtain:

$$R_{ijp}[x_j, y_p, \omega_b] = \lim_{\tau\to 0} \frac{I_{ijp}}{\tau}$$

$$= \int_0^1 \int_0^1 [f_i(x_j, y_p, w(x_j, y_p), \alpha_{jp}(\xi, \eta)) - f_i(x_j, y_p, w(x_j, y_p)v)]\, d\xi\, d\eta.$$

We can show that the collection of equations (2.5), (2.6) and (2.7) is solvable and that for all (x, y) not lying on the mesh $x = x_j$, $y = y_\nu$, the limits

$$\lim_{\tau\to 0} \frac{\Delta z_i(x, y, \omega)}{\tau} = \delta z_i,$$

$$\lim_{\tau\to 0} \frac{\Delta z_{ix}(x, y, \omega)}{\tau} = \delta z_{ix}, \qquad \lim_{\tau\to 0} \frac{\Delta z_{iy}(x, y, \omega)}{\tau} = \delta z_{iy},$$

exist and, moreover,

$$\delta z_{ix} = \frac{\partial \delta z_i}{\partial x}, \qquad \delta z_{iy} = \frac{\partial \delta z_i}{\partial y}.$$

Having divided these equations by τ and by going to the limit as $\tau \to 0$, we find:

$$\begin{aligned}
\delta z_i(x, y, \omega) &= \delta z_i(x, 0, \omega) + \delta z_i(0, y, \omega) \\
&\quad + \int_0^y \int_0^x \sum_{s=1}^{3m} \frac{\partial f_i(x, y, w, v)}{\partial w_s} \delta w_s \, dx \, dy + \sum_{j=1}^{l} \sum_{p=1}^{k} R_{ijp}[x_j, y_p, \omega], \\
\delta z_{ix}(x, y, \omega) &= \delta z_{ix}(x, 0, \omega) + \int_0^y \sum_{s=1}^{3m} \frac{\partial f_i(x, y, w, v)}{\partial w_s} \delta w_s \, dy, \\
\delta z_{iy}(x, y, \omega) &= \delta z_{iy}(0, y, \omega) + \int_0^x \sum_{s=1}^{3m} \frac{\partial f_i(x, y, w, v)}{\partial w_s} \delta w_s \, dx
\end{aligned} \tag{2.8}$$

when $x_l < x < x_{l+1} = X$, $y_k < y < y_{k+1} = Y$.

By the way in which the functions $\delta z_i(x, 0, \omega)$ and $\delta z_i(0, y, \omega)$ were determined it follows that

$$\delta z_i(x, 0, \omega) = \delta z_i(0, y, \omega) = 0$$

when $0 \leqq x \leqq x_1$, $0 \leqq y \leqq y_1$ and, consequently, (2.8) implies that

$$\delta z_i(x, y, \omega) = \delta z_{ix}(x, y, \omega) = \delta z_{iy}(x, y, \omega) \equiv 0$$

when $0 \leqq x \leqq x_1$, $0 \leqq y \leqq y_1$. Further, from (2.4) and (2.8) we obtain

$$\begin{aligned}
\delta z_i(x, y, \omega) &= \sum_{s=1}^{m} A_{is}(x, x_1) Q_s(x_1 \omega_b) + \int_{x_1}^x \int_0^y \sum_{s=1}^{3m} \frac{\partial f_i(x, y, w, v)}{\partial w_s} \delta w_s \, dy \, dx, \\
\delta z_{ix}(x, y, \omega) &= \sum_{s=1}^{m} A'_{is}(x, x_1) Q_s(x_1 \omega_b) + \int_0^y \sum_{s=1}^{3m} \frac{\partial f_i(x, y, w, v)}{\partial w_s} \delta w_s \, dy, \\
\delta z_{iy}(x, y, \omega) &= \int_{x_1}^x \sum_{s=1}^{3m} \frac{\partial f_i(x, y, w, v)}{\partial w_s} \delta w_s \, dx,
\end{aligned}$$

when $x_1 < x < x_2$, $0 < y < y_1$. Solving this system, for example, by the method of successive approximations, we get that the functions δz_i have the form

$$\delta z_i(x, y, \omega) = \sum_{s=1}^{m} A^1_{is}(x, y, x_1) Q_s(x_1, \omega_b), \qquad x_1 < x < x_2, \qquad 0 \leqq y \leqq y_1,$$

where $A^1_{is}(x, y, x_1)$ is a completely determinate function independent of the choice of the functions α_{ij}, β_i and γ_j. By continuing this reasoning we

determine:

$$\delta z_i(x, y, \omega) = \sum_{s=1}^{m} \sum_{j=1}^{l} A_{is}^{j}(x, y, x_j) Q_s(x_j, \omega_b), \qquad 0 \leqq y \leqq y_1, \tag{2.9}$$
$$x_l < x \leqq x_{l+1} = X.$$

Analogously we find that

$$\delta z_i(x, y, \omega) = \sum_{s=1}^{m} \sum_{p=1}^{k} B_{is}^{p}(x, y, y_p) R_s(y_p, \omega_b), \qquad 0 \leqq x \leqq x_1, \tag{2.10}$$
$$y_k < y \leqq y_{k+1} = Y.$$

From these relations it follows, in particular, that

$$\delta z_i(x_l + 0, y, \omega) - \delta z_i(x_l - 0, y, \omega)$$
$$= \sum_{s=1}^{m} [A_{is}^{l}(x_l + 0, y, x_l) Q_s(x_l, \omega_l) - A_{is}^{l-1}(x_l - 0, y, x_{l-1}) Q_s(x_{l-1}, \omega_b)],$$

$$\delta z_i(x, y_k + 0, \omega) - \delta z_i(x, y_k - 0, \omega)$$
$$= \sum_{s=1}^{m} [B_{is}^{k}(x, y_k + 0, y_k) R_s(y_k, \omega_b) - B_{is}^{k-1}(x, y_k - 0, y_{k-1}) R_s(y_{k-1}, \omega_b)].$$

Consequently, the functions $\delta z_i(x, y, \omega)$ defined by (2.9) and (2.10), generally speaking, are discontinuous on the lines $x = x_i$ and $y = y_k$.

Continuing by analogous reasoning, we have:

$$\delta z_i(x, y, \omega) = \sum_{s=1}^{m} \sum_{j=1}^{l} \sum_{p=1}^{k} [C_{ijps}(x, y, x_j, y_p) S(x_j, y_p, \omega_b)$$
$$+ D_{is}^{j}(x, y, x_j) Q_s(x_j, \omega_b) + F_{is}^{p}(x, y, y_p) R_s(y_p, \omega_b)],$$
$$x_l < x \leqq x_{l+1} = X, \qquad y_k < y \leqq y_{k+1} = Y.$$

Setting $x = X$, $y = Y$ in this equality we finally obtain:

$$\delta z_i(X, Y, \omega) = \sum_{s=1}^{m} \sum_{j=1}^{l} \sum_{p=1}^{k} [C_{ijps}(x_j, y_p) S(x_j, y_p, \omega_b) \tag{2.11}$$
$$+ D_{is}^{j}(x_j) Q_s(x_j, \omega_b) + F_{is}^{p}(y_p) R_s(y_p, \omega_b)],$$

where the constants C_{ijps}, D_{is}^{j}, F_{is}^{p} do not depend on the choice of α_{ij}, β_i, γ_j.

The point $z(X, Y) + \delta z(X, Y, \omega)$ corresponding to an arbitrary variation $\omega_b \in \Omega$ of control ω, varies through a certain set Π in the space of the variables $z_1, \cdots, z_m$. By the same method as in [16] we can show that Π is convex and that none of its interior points can belong to the interior of the set D^{-}. Hence it follows that through the point $z(X, Y)$ we can draw

the plane

$$\sum_{i=1}^{m} a_i(z_i - z_i(X, Y)) = 0, \tag{2.12}$$

separating the sets Π and D, and, moreover, the signs of the coefficients a_i can be chosen so that Π lies in the halfspace

$$\sum_{i=1}^{m} a_i(z_i - z_i(X, Y)) \geqq 0.$$

Therefore, for any ω_b

$$\sum a_i \delta z_i(X, Y, \omega) \geqq 0,$$

i.e.,

$$\lim_{\tau \to 0} \frac{\sum a_i \Delta z_i(X, Y, \omega)}{\tau} \geqq 0.$$

We introduce the auxiliary functions u_i with the help of (1.5) and the supplementary conditions:

$$\begin{aligned} u_{ix}(x, Y) &= -\frac{\partial H(x, Y, p(x, Y), v)}{\partial z_{iy}}, \\ u_{iy}(X, y) &= -\frac{\partial H(X, y, p(X, y), v)}{\partial z_{ix}}, \quad u_i(X, Y) = -a_i. \end{aligned} \tag{2.13}$$

By the same method as was applied above we can obtain a formula for the increment of the functional $\bar{S} = \sum a_i z_i(X, Y)$ in the form (1.44) and, consequently, by the same method show that the maximum condition (1.39) is necessary in order for the admissible control $\omega(x, y)$ to realize the minimum of the functional $\bar{S}$. But $\bar{S}$ attains its minimum by a control which is min-optimal with respect to S.

Theorem 5 is completely proved.

It is obvious that the assertion that we have proved remains valid also in the case when the controlled process is described by the boundary value problem (1.1)–(1.2).

The following theorem is valid if the controlled process is described by the linear boundary value problem (1.48)–(1.49)–(1.50).

THEOREM 6. *Let $z(x, y)$ be the solution of boundary value problem* (1.48)–(1.49)–(1.50), *corresponding to the control $\omega(x, y)$ and satisfying the condition $z(X, Y) = z^1$. Then, if $\omega(x, y)$ satisfies the maximum (minimum) condition relative to the functions $u_i(x, y)$ which take the boundary values*

$$u_i(X, Y) = -\lambda A_i - \mu B_i(z^1), \qquad \mu \geqq 0, \qquad \lambda > 0,$$

where $B_i(z^1)$ are the coordinates of the normal to the support hyperplane of D, then the control $\omega(x, y)$ is min-optimal with respect to the functional $S = \sum_{i=1}^{n} A_i z_i(X, Y)$.

The proof of this theorem almost literally coincides with the proof of the corresponding theorem (see [16, Theorem 4]) for ordinary differential equations.

2.2. Application of Theorem 5 to the solution of some actual problems. Generally speaking, the results obtained do not give us a means for constructing the vector $u(x, y)$. However, this problem may be solved in a number of special cases. Let us consider some of them.

(1) The point $z(X, Y)$ is located inside the region D. Then $a_i = A_i$ since any plane, except the plane (2.12), passing through the point $z(X, Y)$ intersects the region D^- and, consequently, cannot separate D^- and Π.

(2) The point $z(X, Y)$ lies on the boundary of the region D, which is given by the inequality $F(z) \leqq 0$. Then, the boundary is given by the equation $F(z) = 0$. If the function $F(z)$ is differentiable, the equation of the tangent plane at the point $z(X, Y)$ is

$$\sum_{i=1}^{m} B_i(z_i - z_i(X, Y)) = 0, \quad B_i = \left[\frac{\partial F}{\partial z_i}\right]_{z=z(X,Y)}.$$

Since the plane $\sum a_i(z_i - z_i(X, Y)) = 0$ also passes through the point $z(X, Y)$,

$$a_i = \lambda A_i + \mu B_i,$$

where without loss of generality we can take $\lambda \geqq 0, \mu \geqq 0$ $(\lambda^2 + \mu^2 \neq 0)$. Since the a_i are determined up to a constant factor, only one of the quantities λ and μ is independent. Since according to (2.13), $u_i(X, Y) = a_i$ while $F(z(X, Y)) = 0$, we obtain $m + 1$ relations

$$(2.14) \quad u_i(X, Y) = -\lambda A_i - \mu B_i, \qquad F(z(X, Y)) = 0, \quad i = 1, \cdots, m,$$

for the determination of the $u_i(X, Y)$ and of one of the quantities λ and μ. Adding the conditions (1.37) onto (2.14) we obtain $2m$ boundary conditions for the $2m$ functions $z_1, \cdots, z_m, u_1, \cdots, u_m$. These conditions together with (1.1), (1.5), (1.36), (1.39) and (2.13) form a "complete" system of relations for the determination of the optimal control and of the vector-functions $z(x, y)$ and $u(x, y)$ corresponding to it.

For example, let it be required to determine the minimum of the functional

$$I = \int_0^X \int_0^Y f_0(x, y, z, z_x, z_y, v)\, dy\, dx$$

under the condition that the function $z(x, y)$ is a solution of the boundary value problem (1.1)–(1.2), while the point $z(X, Y)$ belongs to a certain convex region D in the space of the variables $z_1, \cdots, z_m$. By introducing the auxiliary function z_0 by means of (1.4) we reduce the problem to seeking the minimum of the functional $S = z_0(X, Y)$ under the condition that the point

$$Z(X, Y) = (z_0(X, Y), z_1(X, Y), \cdots, z_m(X, Y))$$

lie on a cylinder with axis parallel to the z_0-axis. Since the variable z_0 does not enter into the right-hand side of (1.1) and (1.4), $B_0 = 0$ in (2.14). For the functional being considered, $A_1 = \cdots = A_m = 0$, $A_0 = 1$, and hence from (1.5) and (1.6) it follows that $u_0(x, y) = -1$. Thus, for the problem being considered the differential equations and the boundary conditions take the form of the relations

$$z_{ixy} = \frac{\partial H}{\partial u_i}, \quad z_i(0, y) = \varphi_i(y), \quad z_i(x, 0) = \psi_i(x),$$

$$u_{ixy} = \frac{\partial H}{\partial z_i} - \frac{d}{dx}\left(\frac{\partial H}{\partial z_{ix}}\right) - \frac{d}{dy}\left(\frac{\partial H}{\partial z_{iy}}\right), \quad u_{ix}(x, Y) = -\left.\frac{\partial H}{\partial z_{iy}}\right|_{y=Y},$$

$$u_{iy}(X, y) = -\left.\frac{\partial H}{\partial z_{ix}}\right|_{x=X}, \quad u_i(X, Y) = 0, \quad H = \sum_{i=1}^{m} u_i f_i - f_0,$$

from which the auxiliary equality (1.4) is eliminated.

2.3. Generalization to the case of an arbitrary number of independent variables. The formula for the increment of functional S and its corollaries can be generalized to the case when the controlled process is described by a Goursat problem with an arbitrary number of independent variables (see [18]). However, so as not to burden the formulas with unnecessary details we shall assume that the number of independent variables equals three.

Thus, let the functions $z_i(x)$, $x = (x_1, x_2, x_3)$, $i = 1, \cdots, m$, be given by the equations

$$\frac{\partial^3 z_i}{\partial x_1 \partial x_2 \partial x_3} = f_i\left(x, z_1, \cdots, z_m, \frac{\partial z_1}{\partial x_1}, \cdots, \frac{\partial z_m}{\partial x_3}, \cdots, \frac{\partial^2 z_m}{\partial x_2 \partial x_3}, v\right), \tag{2.15}$$

$$i = 1, \cdots, m, \qquad 0 \leqq x_k \leqq X_k, \qquad k = 1, 2, 3,$$

and the supplementary conditions

$$z_i(0, x_2, x_3) = \varphi_i^1(x_2, x_3), \qquad z_1(x_1, 0, x_3) = \varphi_i^2(x_1, x_3), \tag{2.16}$$

$$z_i(x_1, x_2, 0) = \varphi_i^3(x_1, x_2),$$

where the functions f_i contain the mixed derivatives of the variables z_j of order not exceeding two. These functions are twice continuously differentiable with respect to the set of all the arguments. The control param-

eters are subject to the same conditions as before. The functions φ_i^k are twice piecewise-continuously differentiable with respect to their own arguments and satisfy the natural conditions of conjugacy. Just as before we assume that to each admissible control there corresponds a class of functions in which the boundary value Goursat problem which is posed is solvable uniquely.

As the optimality criterion we select the functional

$$S = \sum_{i=1}^{m} A_i z_i(X_1, X_2, X_3), \tag{2.17}$$

where the A_i are given real numbers.

We introduce the auxiliary variables u_i and the function $H(x, w, v) = \sum_{i=1}^{m} u_i f_i$, where

$$w = \left(u_1, \cdots, u_m, z_1, \cdots, z_m, \frac{\partial z_1}{\partial x_1}, \cdots, \frac{\partial z_m}{\partial x_3}, \cdots, \frac{\partial^2 z_m}{\partial x_2 \partial x_3}\right)$$

is a vector, the number of whose components we shall denote by N. The functions $u_i(x)$ are determined with the help of the equations

$$\frac{\partial^3 u_i}{\partial x_1 \partial x_2 \partial x_3} = \frac{\partial H}{\partial z_i} - \sum_{k=1}^{3} \frac{\partial}{\partial x_k}\left(\frac{\partial H}{\partial z_{ix_k}}\right) + \frac{1}{2}\sum_{j \neq k}^{3} \frac{\partial^2}{\partial x_j \partial x_k}\left(\frac{\partial H}{\partial z_{ix_j x_k}}\right), \tag{2.18}$$

$$i = 1, \cdots, m,$$

and the supplementary conditions:

$$\frac{\partial^2 u_i}{\partial x_1 \partial x_2} = -\frac{\partial H}{\partial z_{ix_3}} + \frac{\partial}{\partial x_1}\left(\frac{\partial H}{\partial z_{ix_1x_3}}\right) + \frac{\partial}{\partial x_2}\left(\frac{\partial H}{\partial z_{ix_2x_3}}\right) \quad \text{when} \quad x_3 = X_3,$$

$$\frac{\partial^2 u_i}{\partial x_1 \partial x_3} = -\frac{\partial H}{\partial z_{ix_2}} + \frac{\partial}{\partial x_1}\left(\frac{\partial H}{\partial z_{ix_1x_2}}\right) + \frac{\partial}{\partial x_3}\left(\frac{\partial H}{\partial z_{ix_2x_3}}\right) \quad \text{when} \quad x_2 = X_2, \tag{2.19}$$

$$\frac{\partial^2 u_i}{\partial x_2 \partial x_3} = -\frac{\partial H}{\partial z_{ix_1}} + \frac{\partial}{\partial x_2}\left(\frac{\partial H}{\partial z_{ix_1x_2}}\right) + \frac{\partial}{\partial x_3}\left(\frac{\partial H}{\partial z_{ix_1x_3}}\right) \quad \text{when} \quad x_1 = X_1,$$

$$\frac{\partial u_i}{\partial x_1} = \frac{\partial H}{\partial z_{ix_3x_2}} \quad \text{when} \quad x_2 = X_2, \quad x_3 = X_3,$$

$$\frac{\partial u_i}{\partial x_2} = \frac{\partial H}{\partial z_{ix_1x_3}} \quad \text{when} \quad x_3 = X_3, \quad x_1 = X_1, \tag{2.20}$$

$$\frac{\partial u_i}{\partial x_3} = \frac{\partial H}{\partial z_{ix_1x_2}} \quad \text{when} \quad x_1 = X_1, \quad x_2 = X_2,$$

$$u_i(X_1, X_2, X_3) = -A_i, \qquad i = 1, \cdots, m. \tag{2.21}$$

Equations (2.20) are equations in ordinary derivatives. Therefore, for every admissible control they, together with conditions (2.21), uniquely determine the functions $u_i(x_1, X_2, X_3)$, $u_i(X_1, x_2, X_3)$ and $u_i(X_1, X_2, x_3)$. Let us now solve (2.19) with the supplementary conditions

$$\left.\begin{aligned} u_i(x_1, x_2, X_3)\,|_{x_1=X_1} &= u_i(X_1, x_2, X_3), \\ u_i(x_1, x_2, X_3)\,|_{x_2=X_2} &= u_i(x_1, X_2, X_3) \end{aligned}\right\} \quad \text{when} \quad x_3 = X_3;$$

$$\left.\begin{aligned} u_i(x_1, X_2, x_3)\,|_{x_1=X_1} &= u_i(X_1, X_2, x_3), \\ u_i(x_1, X_2, x_3)\,|_{x_3=X_3} &= u_i(x_1, X_2, X_3) \end{aligned}\right\} \quad \text{when} \quad x_2 = X_2;$$

$$\left.\begin{aligned} u_i(X_1, x_2, x_3)\,|_{x_2=X_2} &= u_i(X_1, X_2, x_3), \\ u_i(X_1, x_2, x_3)\,|_{x_3=X_3} &= u_i(X_1, x_2, X_3) \end{aligned}\right\} \quad \text{when} \quad x_1 = X_1.$$

By virtue of the assumptions made above, the functions $u_i(x_1, x_2, X_3)$, $u_i(x_1, X_2, x_3)$ and $u_i(X_1, x_2, x_3)$ are determined uniquely. Thus, at the final count the problem is reduced to the Goursat problem: find the solution of (2.18) in the region $0 \leqq x_k \leqq X_k$, which satisfies the boundary conditions:

$$\begin{aligned} &u_i(x_1, x_2, x_3)\,|_{x_2=X_2} = u_i(x_1, X_2, x_3), \\ &u_i(x_1, x_2, x_3)\,|_{x_3=X_3} = u_i(x_1, x_2, X_3), \\ &u_i(x_1, x_2, x_3)\,|_{x_1=X_1} = u_i(X_1, x_2, x_3), \qquad i = 1, 2, \cdots, m. \end{aligned} \tag{2.22}$$

Here we should keep in mind that in (2.18) and (2.19) the functions $\partial H/\partial w_k$ are differentiable with respect to x_1, x_2 and x_3. Thus, if it is assumed that the class of admissible controls consists of piecewise-continuous functions, it is necessary to satisfy the condition: the right-hand sides of these equations are independent of the derivatives of the functions v and of $z_{x_1x_1}$, $z_{x_2x_2}$, $z_{x_3x_3}$. However, if the right-hand sides of (2.18) and (2.19) do depend on these quantities, then as the class of admissible controls we should choose the functions $v(x)$ having piecewise-continuous derivatives.

Assuming that these conditions are satisfied, by the same method which was used above we can obtain a formula for the increment of the functional

$$\Delta S = -\int_0^{X_1}\int_0^{X_2}\int_0^{X_3} [H(x, w, v + \Delta v) - H(x, w, v)]\, dx_3\, dx_2\, dx_1 - \eta,$$

where $\eta = \eta_1 + \eta_2$,

$$\eta_1 = \frac{1}{2}\sum_{i=1}^{N}\int_0^{X_1}\int_0^{X_2}\int_0^{X_3}\left[\frac{\partial H(x, w, v + \Delta v)}{\partial w_i} - \frac{\partial H(x, w, v)}{\partial w_i}\right]\Delta w_i\, dx_3\, dx_2\, dx_1,$$

$$\eta_2 = \frac{1}{2}\sum_{i,k=1}^{N}\int_0^{X_1}\int_0^{X_2}\int_0^{X_3}\left[\frac{\partial^2 H(x, w + \theta_1 \Delta w, v + \Delta v)}{\partial w_i\, \partial w_k} - \frac{\partial^2 H(x, w + \theta_2 \Delta w, v + \Delta v)}{\partial w_i\, \partial w_k}\right] \Delta w_i\, \Delta w_k\, dx_3\, dx_2\, dx_1 .$$

From this formula we can derive the optimality conditions which may be formulated in the form of Theorems 1 and 2. If it is proposed to realize the control with the aid of boundary conditions, then we can obtain results analogous to Theorems 3 and 4.

2.4. Control of a process with the aid of "point steering". In all the problems considered above it was assumed that all the components of the vector $v(x, y)$ were functions of two variables: x and y. However, the proposed method allows us to solve the problem when all the admissible controls $v(x, y)$ can be represented in the form

$$v(x, y) = (v^1(x), v^2(x, y), v^3(y)).$$

(Some components of this vector are functions of only one independent variable x or y.)

For the sake of definiteness let us consider the problem of minimizing (1.3) when the process is described by the boundary value problem (1.1)–(1.2). Formula (1.21) for the increment of the functional remains valid also in this case. The estimate (1.29) of the remainder term in this formula is also valid. Therefore, by the same method we prove Theorem 1′.

THEOREM 1′. *In order that the admissible control* $v(x, y) = (v^1(x), v^2(x, y), v^3(y))$ *in the boundary value problem* (1.1)–(1.2) *be min-optimal with respect to functional* (1.3), *it is necessary that the condition*

$$\iint_G [H(x, y, p(x, y), v(x, y) + \Delta v) - H(x, y, p(x, y), v(x, y)]\, dx\, dy \leqq 0 \tag{2.23}$$

be fulfilled for any admissible increment Δv, *where* $p(x, y)$ *is a vector corresponding to the control* $v(x, y)$ *and is determined from* (1.1) *and* (1.5) *and the supplementary conditions* (1.2) *and* (1.6).

In particular, if the admissible controls depend on only one variable (say, x) *and if*

$$f_i(x, y, z, z_x, z_y, v) \equiv f_i^0(x, y, z, z_x, z_y) + f_i^1(x, v)$$

in (1.1), *then* (2.23) *takes the form*

$$\iint_G \sum_{i=1}^{m} u_i(x, y)[f_i^1(x, v(x) + \Delta v) - f_i^1(x, v)]\, dx\, dy \leqq 0.$$

By introducing the notation

$$H^1(x, u(x), v) = \sum_{i=1}^{m} f_i^1(x, v) \int_0^Y u_i(x, y)\, dy,$$

we obtain the optimality condition in the following form.

THEOREM 1″. *In order that an admissible control $v(x)$ in the boundary value problem* (1.1)–(1.2) *be min-optimal (among the controls depending only on x) with respect to functional* (1.3), *it is necessary that*

$$H^1(x, u(x), v(x)) \ (=) \sup_{v \in V} H^1(x, u(x), v),$$

where the symbol $(=)$ denotes equality which is valid for almost all x in the interval $0 \leqq x \leqq X$.

3. Calculus of variations and optimal control problems. The problems considered in the present paper are essentially problems in the calculus of variations. However, the classical methods are inapplicable here since, generally speaking, the control parameters may take values from a closed region. In the case where the range of the control parameters is open, from the maximum principle we obtain the necessary conditions of the classical calculus of variations for functionals with partial derivatives.

Let it be required to find the minimum of the functional

$$I = \int_0^X \int_0^Y f(x, y, z, z_x, z_y, v)\, dy\, dx,$$

which is defined on the functions $z = (z_1, \cdots, z_m)$ given by the relations

$$z_{ixy}(x, y) = v_i, \qquad v = (v_1, \cdots, v_m),$$

$$z_i(0, y) = \varphi_i(y), \qquad z_i(x, 0) = \psi_i(x), \qquad i = 1, \cdots, m,$$

where the control parameters v are chosen in the class of all piecewise-continuous vector-functions.

By an optimal control we shall mean an admissible control by which the functional I attains its minimum among the functions lying in a small neighborhood of the function $z(x, y)$ corresponding to this control. It is obvious that the optimal control defined in such a way is a special case of the optimal control in the previous sense. Therefore, the maximum principle remains in force and every optimal control is an extremal one. The converse is also valid: every extremal solution is an optimal solution.

To seek such a solution we introduce the auxiliary variable z_0 :

$$z_{0xy} = f(x, y, z, z_x, z_y, v), \qquad z_0(x, 0) = z_0(0, y) = 0,$$

and we construct the function H:

$$H = u_0 f + \sum u_p v_p .$$

Then the auxiliary functions $u_i(x, y)$ are determined with the aid of the boundary value problem

$$
\begin{gathered}
u_{ixy} = \frac{\partial f}{\partial z_i} u_0 - \frac{d}{dx}\left(\frac{\partial f}{\partial z_{ix}} u_0\right) - \frac{d}{dy}\left(\frac{\partial f}{\partial z_{iy}} u_0\right), \\
u_{ix}(x, Y) = -\left[\frac{\partial f}{\partial z_{iy}} u_0\right]_{y=Y}, \qquad (3.1) \\
u_{iy}(X, y) = -\left[\frac{\partial f}{\partial z_{ix}} u_0\right]_{x=X}, \\
u_i(X, Y) = 0, \qquad i = 1, \cdots, m, \qquad u_0(x, y) = -1.
\end{gathered}
$$

Hence we find that $H = \sum u_p v_p - f$. Since the function H attains its maximum by the optimal control $v(x, y)$,

$$
\left(\frac{\partial H}{\partial v_i}\right)_{v=v(x,y)} = \left(u_i - \frac{\partial f}{\partial v_i}\right)_{v=v(x,y)} = 0.
$$

Consequently, $u_{ixy} = d^2(\partial f/\partial z_{ixy})/dx\,dy$, and by virtue of (3.1) we find that the solution $z(x, y)$ of the optimal problem posed satisfies the system of Ostrogradskiĭ-Euler equations (for example, see [32, p. 122]):

$$
\frac{\partial f}{\partial z_i} - \frac{d}{dx}\left(\frac{\partial f}{\partial z_{ix}}\right) - \frac{d}{dy}\left(\frac{\partial f}{\partial z_{iy}}\right) + \frac{d^2}{dx\,dy}\left(\frac{\partial f}{\partial z_{ixy}}\right) = 0.
$$

By assumption the function f has continuous second derivatives with respect to the variables $v_1, \cdots, v_m$. Since the control $v(x, y)$ realizes the maximum of the function H, the quadratic form

$$
\sum_{i,k=1}^{m} \frac{\partial^2 H}{\partial v_i\,\partial v_k} \lambda_i \lambda_k = -\sum_{i,k=1}^{m} \frac{\partial^2 f}{\partial v_i\,\partial v_k} \lambda_i \lambda_k
$$

is nonpositive. Therefore, from the maximum condition (1.7) it follows that everywhere in the region G, $0 \leqq x \leqq X$, $0 \leqq y \leqq Y$, with the possible exception of points lying on a finite number of lines with zero area, the following inequality (the Legendre condition) is satisfied:

$$
\sum_{i,k=1}^{m} \frac{\partial^2 f(x, y, z, z_x, z_y, z_{xy})}{\partial v_i\,\partial v_k} \lambda_i \lambda_k \geqq 0, \qquad \sum_{i=1}^{m} \lambda_i^2 \neq 0, \qquad (3.2)
$$

which is a necessary condition for the function $z(x, y)$ to be an extremal which minimizes the functional I.

In the case when the range of the control parameter is closed, the derivatives $\partial H/\partial v_i$ may not vanish on the optimal trajectory $z(x, y)$ and, consequently, (3.2) may not be satisfied. As a confirmation of what we have said we consider the simplest example.

Let the controlled process be described by the boundary value problem

$$z_{xy} = v^2, \qquad z(x,0) = z(0,y) = 0, \qquad 0 \leqq x, y \leqq 1,$$

where v is the control parameter, $|v| \leqq 1$. As the optimality criterion we shall take the functional

$$S = -\int_0^1 \int_0^1 z_{xy}\, dx\, dy = -z(1,1), \qquad (f(x,y,z,z_x,z_y,v) \equiv -v^2).$$

It is easily shown that the control which is min-optimal with respect to D will be $v(x,y) = 1$ and, consequently, at this control

$$\frac{\partial^2 f}{\partial v^2} < 0,$$

and (3.2) is not satisfied.

4. Optimal processes in systems whose behaviour is described by parabolic equations.

4.1. Statement of the problem. The Maximum Principle. Let E^n be the Euclidean space of the vectors $x = (x_1, \cdots, x_n)$, let G be a bounded region in E^n with boundary Γ of class $A^{(2)}$ (see [33, p. 10]), and let $X_i(x)$ be direction cosines of the outward normal to the boundary Γ.

Further, in the region G let an elliptic operator $L = (L_1, \cdots, L_m)$ be defined by the formula

$$L_i y = \sum_{p=1}^{m} \sum_{j,k=1}^{n} a_{jk}^{ip} \frac{\partial^2 y_p}{\partial x_j\, \partial x_k}, \tag{4.1}$$

where the functions $a_{jk}^{ip}(x_1, \cdots, x_n)$ in the region $G + \Gamma$ are of class $C^{(2)}$. We denote by $M = (M_1, \cdots, M_n)$ the operator defined by the formula

$$M_i z = \sum_{p=1}^{m} \sum_{j,k=1}^{n} \frac{\partial}{\partial x_j}\left(a_{jk}^{pi} \frac{\partial z_p}{\partial x_k}\right) + \sum_{j=1}^{n} \frac{\partial}{\partial x_j}(l_j^{pi} z_p), \qquad i = 1, \cdots, m,$$

where

$$l_j^{pi} = -\sum_{k=1}^{n} \frac{\partial a_{jk}^{pi}}{\partial x_k}.$$

A direct verification can convince us of the validity of the following equality:

$$\sum_{i=1}^{m} \int_G (z_i L_i y - y_i M_i z)\, dx$$

$$= \sum_{i,p=1}^{m} \sum_{j=1}^{n} \int_\Gamma \left[\sum_{k=1}^{n} a_{jk}^{ip}\left(z_i \frac{\partial y_p}{\partial x_k} - y_p \frac{\partial z_i}{\partial x_k}\right) + l_j^{ip} y_p z_i\right] X_j(x)\, d\sigma.$$

By the same method which is used for a single equation of elliptic type, this formula can be transformed to the form:

$$(4.2)\quad \sum_{i=1}^{m} \int_G (z_i L_i y - y_i M_i z)\, dx = \sum_{i=1}^{m} \int_\Gamma (z_i P_i y - y_i Q_i z)\, d\sigma,$$

where

$$(4.3)\quad P_i y = \sum_{p=1}^{m} \left[a_l^{ip} \frac{dy_p}{dl_{ip}} + b_{ip} y_p \right], \qquad Q_i z = \sum_{p=1}^{m} \left[a_\lambda^{pi} \frac{dz_p}{d\lambda_{ip}} + d_{ip} z_p \right].$$

In (4.3) the paths l_{ip} are chosen arbitrarily except that $\cos(n, l_{ip}) > 0$ (n is the outward normal to Γ), and their direction cosines belong to class $C^{(1)}$ on Γ. The paths λ_{ip} are chosen as functions of l_{ip}.

We assume that the coefficients in operator L still depend on t, $0 \leqq t \leqq T$, and we shall study the controlled system whose behavior is described by a system of equations of the parabolic type

$$(4.4)\quad L_i y = f(t, x, y, y_x, u), \qquad 0 \leqq t \leqq T, \qquad x \in G, \qquad \left(L_{it} y = \frac{\partial y_i}{\partial t} - L_i y \right),$$

where the function $f = (f_1, \cdots, f_m)$ is continuous in t and twice continuously differentiable in the remaining arguments, while the parameter u takes values from some convex (open or closed) region U of p-dimensional Euclidean space.

We further assume that the function $y(t, x) = (y_1, \cdots, y_m)$, determined by (4.4), satisfies the conditions

$$(4.5)\quad \begin{aligned} P_i(t, x)y &= \varphi_i(t, x, y, v), \qquad x \in \Gamma, \qquad 0 \leqq t \leqq T, \\ y(0, x) &= a(x), \qquad x \in G, \end{aligned}$$

where the operators P_i are determined by (4.3) in which the functions $a_l^{is}(t, x)$, $b_{ip}(t, x)$ and $a(x)$ are continuous, and the φ_i satisfy the same conditions as the f_i, while the parameter v takes values from a convex (open or closed) region V of a q-dimensional Euclidean space.

We shall call the function $\omega(t, x) = (u(t, x), v(t, x))$ an admissible control if all its components are piecewise continuous and $u(t, x)$ and $v(t, x)$ take values from the regions U and V, respectively. Moreover, we shall assume that the surfaces of discontinuity of the admissible controls are smooth and each of them either is orthogonal to the t-axis or in the neighborhood of any point in it we can carry out the nonsingular transformation of coordinates

$$\tau = t, \qquad \xi_i = \xi_i(t, x), \qquad i = 1, \cdots, n,$$

so that the surface of discontinuity becomes a piece of the plane $\xi_n = 0$.

If the discontinuities of a certain admissible control satisfy the first condition, then the boundary value problem (4.4)–(4.5) corresponding to this control splits up into several problems of the same kind but in regions which abut on each other along the surfaces of discontinuity of the control. In this case the problem (4.4)–(4.5) has a unique continuous solution (for example, see [34]) and, moreover, this solution is not subject to any supplementary smoothness conditions on the surfaces of discontinuity of the control.

However, if these surfaces satisfy the second condition, then by a solution of problem (4.4)–(4.5) we shall mean the vector-function $y(t, x)$ which satisfies (4.4), (4.5), and certain smoothness conditions on the surfaces of discontinuity of the control. Apparently, this problem has not been studied in its general form; however, special cases of it have been considered in a number of papers (for example, see [35]–[40]) where various theorems on the existence and uniqueness of the solutions are obtained. Therefore, everywhere in the following we shall assume that the given functions in (4.4) and (4.5) satisfy, in addition to the properties listed above, further conditions under which a unique solution of problem (4.4)–(4.5) corresponds to each admissible control.

Let $\omega(t, x)$ be some admissible control while $y(t, x)$ is the solution of problem (4.4)–(4.5) corresponding to it, and let there be given the functional

$$S = \sum_{i=1}^{m} \left[\int_G \alpha_i(x) y_i(T, x)\, dx + \int_0^T \int_G \beta_i(t, x) y_i(t, x)\, dx\, dt + \int_0^T \int_\Gamma \gamma_i(t, x) y_i(t, x)\, d\sigma\, dt \right], \tag{4.6}$$

where α_i, β_i and γ_i are given continuous functions.

We pose the problem: among all the admissible controls find the control $\omega(t, x)$ (if it exists) such that the solution of problem (4.4)–(4.5) corresponding to it realizes the minimum of functional S.

The admissible control $\omega(t, x)$ at which the functional S attains its maximal (minimal) value will be called max-optimal (min-optimal) with respect to S. Functionals of more general form will be considered in the final section.

As was noted above, the problem of optimal control of processes described by parabolic equations is of definite theoretical and practical interest. A number of papers (see [5], [6], [11]) have considered certain problems when the control is effected with the help of initial or boundary conditions and as the optimality criterion is chosen time-optimality or a functional of the form

$$I = \int_0^1 [u(T, x) - u_0(x)]^2\, dx + \gamma \int_0^T p^2(t)\, dt,$$

where $u_0(x)$ is a given function from $L_2(0, 1)$, $p(t)$ is the control, and γ is a nonnegative constant. Here we still consider the problem when the control of the process can be effected simultaneously with the help of controls occurring both in the equation and in the boundary conditions. It is obvious that the functional

$$S_1 = \sum_{i=1}^{m} \int_0^T \int_G \left[\gamma_i(t, x) y_i(t, x) + \sum_{k=1}^{n} \alpha_{ik}(t, x) \frac{\partial y_i}{\partial x_k} + \beta_i(t, x) \frac{\partial y_i}{\partial t} \right] dx\, dt,$$

where α_{ik} and β_i are continuously differentiable functions, can be brought to the form (4.6).

In order to formulate the optimality conditions we introduce the auxiliary function $z(t, x) = (z_1, \cdots, z_m)$ with the aid of the boundary value problem "adjoint" to (4.4)–(4.5):

$$M_{it} z = -\sum_{s=1}^{m} \left[\frac{\partial f_s(t, x, y, y_x, u)}{\partial y_i} z_s - \sum_{k=1}^{n} \frac{d}{dx_k} \left(\frac{\partial f_s(t, x, y, y_x, u)}{\partial y_{ix_k}} z_s \right) \right] + \beta_i(t, x), \qquad x \in G, \tag{4.7}$$

$$Q_i(t, x) z = \sum_{s=1}^{m} \left[\frac{\partial \varphi_s(t, x, y, v)}{\partial y_i} + \sum_{k=1}^{n} \frac{\partial f_s(t, x, y, y_x, v)}{\partial y_{ix_k}} X_k(x) \right] z_s - \gamma_i(t, x), \qquad x \in \Gamma, \tag{4.8}$$

$$z_i(T, x) = -\alpha_i(x), \qquad x \in G, \qquad i = 1, \cdots, m,$$

where $M_{it} z = (\partial z_i / \partial t) + M_i z$, the Q_i are defined by (4.3), the functions α_i, β_i and γ_i occur in the definition of functional S, and the $X_k(x)$ are the direction cosines of the normal to the boundary Γ external to G. In order that the boundary value problem (4.7)–(4.8) be solvable it is necessary that the functions α_i and γ_i be connected by consistency relations. In what follows it is assumed that these relations are fulfilled.

We introduce the notation:

$$w = \left(z_1, \cdots, z_m, y_1, \cdots, y_m, \frac{\partial y_1}{\partial x_1}, \cdots, \frac{\partial y_m}{\partial x_n} \right),$$

$$p = (z_1, \cdots, z_m, y_1, \cdots, y_m), \qquad H(t, x, w, u) = \sum_{i=1}^{m} z_i f_i(t, x, y, y_x, u),$$

$$h(t, x, p, v) = \sum_{i=1}^{m} z_i \varphi_i(t, x, y, v).$$

Then the boundary value problems (4.4)–(4.5) and (4.7)–(4.8) can be

written in the following form:

$$
\begin{aligned}
L_{it}\, y &= \frac{\partial H(t, x, w, u)}{\partial z_i}, \qquad y_i(0, x) = a_i(x), \qquad x \in G, \\
P_i\, y &= \frac{\partial h(t, x, p, v)}{\partial z_i}, \qquad x \in \Gamma,
\end{aligned} \tag{4.9}
$$

$$
\begin{aligned}
M_{it}\, z &= -\frac{\partial H(t, x, w, u)}{\partial y_i} + \sum_{k=1}^{n} \frac{d}{dx_k}\left(\frac{\partial H(t, x, w, u)}{\partial y_{ix_k}}\right) + \beta_i(t, x), \\
z_i(T, x) &= -\alpha_i(x),
\end{aligned} \tag{4.10}
$$

$$
Q_i\, z = \frac{\partial h(t, x, p, v)}{\partial y_i} + \sum_{k=1}^{n} \frac{\partial H(t, x, w, u)}{\partial y_{ix_k}} X_k(x) - \gamma_i(t, x), \quad x \in \Gamma.
$$

By using (4.2) it is easy to establish that the Ostrogradskii-Green formula,

$$
\begin{aligned}
&\sum_{i=1}^{m} \int_0^T \int_G (z_i\, L_{it}\, y + y_i\, M_{it}\, z)\, dx\, dt \\
&\quad = -\sum_{i=1}^{m} \left[\int_0^T \int_\Gamma (z_i\, P_i\, y - y_i\, Q_i\, z)\, d\sigma\, dt - \int_G y_i\, z_i \Big|_{t=0}^{T} dx \right],
\end{aligned} \tag{4.11}
$$

is valid for any twice piecewise-continuously differentiable functions $y_i(t, x)$ and $z_i(t, x)$. Let $\omega(t, x) = (u(t, x), v(t, x))$ be some admissible control and let $y(t, x)$ and $z(t, x)$ be the solutions of boundary value problems (4.9) and (4.10) corresponding to it. We shall say that the admissible control $\omega(t, x)$ satisfies the maximum condition if

$$
H(t, x, w(t, x), u(t, x)) \,((=))\, \sup_{u \in U} H(t, x, w(t, x), u), \quad x \in G, \quad 0 \leqq t \leqq T,
$$

$$
h(t, x, p(t, x), v(t, x)) \,(=)\, \sup_{v \in V} h(t, x, p(t, x), v), \quad x \in \Gamma, \quad 0 \leqq t \leqq T,
$$

where the symbol $((=))$ denotes equality valid everywhere in the region C, $0 \leqq t \leqq T$, $x \in G$, with the possible exception of points lying on a finite number of n-dimensional surfaces whose $(n + 1)$-dimensional volumes are zero. The symbol $(=)$ is defined analogously, only instead of n and G we should choose $n - 1$ and Γ, respectively. The minimum condition is defined analogously.

THEOREM 7. (The Maximum Principle). *In order that the admissible control* $\omega(t, x) = (u(t, x), v(t, x))$ *be min-optimal (max-optimal) with respect to* S, *it is necessary that it satisfy the maximum (minimum) condition.*

This theorem, although it does not give sufficient conditions for optimality, can still serve as a practical means for the determination of the optimal controls and of the solutions of boundary value problem (4.4)–

(4.5) corresponding to them. We can convince ourselves of this by repeating the reasoning carried out in §1.

4.2. Formula for the increment of functional S. Proof of Theorem 7. Let $\omega(t, x)$ be an arbitrary admissible control and let $y(t, x)$ and $z(t, x)$ be the solutions of boundary value problems (4.9) and (4.10) corresponding to it. Then

$$I = \int_C \left[\sum_{i=1}^m z_i L_{it}\, y - H(t, x, w, u(t, x))\right] dx\, dt + \int_\sigma \left[\sum_{i=1}^m z_i P_i\, y - h(t, x, p, v(t, x))\right] d\sigma = 0,$$

where $C = (0 \leqq t \leqq T, x \in G)$, $\sigma = (0 \leqq t \leqq T, x \in \Gamma)$. We shall take a certain admissible increment $\Delta\omega = (\Delta u, \Delta v)$ of the control $\omega(t, x)$ and denote by $y + \Delta y$ and $z + \Delta z$ the solutions of the same problems (4.9) and (4.10) but corresponding to the control $\omega + \Delta\omega$. Then

$$\begin{aligned}\Delta I &= I[w + \Delta w, \omega + \Delta\omega] - I[w, \omega] \\ &= \int_C \left\{\sum_{i=1}^m (\Delta z_i L_{it}\, \Delta y + \Delta z_i L_{it}\, y + z_i L_{it}\, \Delta y)\right. \\ &\quad \left. - [H(t, x, w + \Delta w, u + \Delta u) - H(t, x, w, u)]\right\} dx\, dt \qquad (4.12) \\ &\quad + \int_\sigma \left\{\sum_{i=1}^m (\Delta z_i P_i\, \Delta y + \Delta z_i P_i\, y + z_i P_i\, \Delta y)\right. \\ &\quad \left. - [h(t, x, w + \Delta w, v + \Delta v) - h(t, x, w, v)]\right\} d\sigma = 0,\end{aligned}$$

and the functions Δy_i and Δz_i, $i = 1, \cdots, m$, form the solutions, respectively, of the boundary value problems:

$$\left.\begin{aligned} L_{it}\, \Delta y &= \Delta \frac{\partial H(t, x, w, u)}{\partial z_i}, \qquad \Delta y_i(0, x) = 0, \qquad x \in G, \\ P_i\, \Delta y &= \Delta \frac{\partial h(t, x, p, v)}{\partial z_i}, \qquad x \in \Gamma, \end{aligned}\right\} \qquad (4.13)$$

$$\left.\begin{aligned} M_{it}\, \Delta z &= -\Delta \frac{\partial H(t, x, w, u)}{\partial y_i} + \sum_{k=1}^n \frac{d}{dx_k}\left(\Delta \frac{\partial H(t, x, w, u)}{\partial y_{ix_k}}\right), \\ \Delta z_i(T, x) &= 0, \qquad x \in G, \\ Q_{it}\, \Delta z &= \Delta \frac{\partial h(t, x, p, v)}{\partial y_i} + \sum_{k=1}^n \Delta \frac{\partial H(t, x, w, u)}{\partial y_{ix_k}} X_k(x), \qquad x \in \Gamma, \end{aligned}\right\} \qquad (4.14)$$

where

$$\Delta \frac{\partial h}{\partial w_k} = \frac{\partial H(t, x, w + \Delta w, u + \Delta u)}{\partial w_k} - \frac{\partial H(t, x, w, u)}{\partial w_k},$$

$$\Delta \frac{\partial h}{\partial p_k} = \frac{\partial h(t, x, p + \Delta p, v + \Delta v)}{\partial p_k} - \frac{\partial h(t, x, p, v)}{\partial p_k}.$$

We transform (4.12) with the aid of (4.11). Since the functions Δy and Δz are, respectively, the solutions of boundary value problems (4.13) and (4.14),

$$\sum_{i=1}^{m} \left[\int_C \Delta z_i \, L_{it} \, \Delta y \, dx \, dt + \int_\sigma \Delta z_i \, P_i \, \Delta y \, d\sigma \right]$$

$$= \sum_{i=1}^{m} \left\{ \int_C \left[\Delta \frac{\partial H(t, x, w, u)}{\partial y_i} \Delta y_i - \sum_{k=1}^{n} \frac{d}{dx_k} \left(\Delta \frac{\partial H(t, x, w, u)}{\partial y_i} \right) \Delta y_i \right] dx \, dt \right.$$

$$\left. + \int_\sigma \left[\Delta \frac{\partial h(t, x, p, v)}{\partial y_i} + \sum_{k=1}^{n} \Delta \frac{\partial H(t, x, w, v)}{\partial y_{ix_k}} X_k(x) \right] \Delta y_i \, d\sigma \right\}$$

$$= \sum_{i=1}^{m} \left\{ \int_C \left[\Delta \frac{\partial H(t, x, w, u)}{\partial y_i} \Delta y_i + \sum_{k=1}^{n} \Delta \frac{\partial H(t, x, w, u)}{\partial y_{ix_k}} \Delta y_{ix_k} \right] dx \, dt \right.$$

$$\left. + \int_\sigma \Delta \frac{\partial h(t, x, p, v)}{\partial y_i} \Delta y_i \, d\sigma \right\}.$$

On the other hand

$$\sum_{i=1}^{m} \left[\int_C \Delta z_i \, L_{it} \, \Delta y \, dx \, dt + \int_\sigma \Delta z_i \, P_i \, \Delta y \, d\sigma \right]$$

$$= \sum_{i=1}^{m} \left[\int_C \Delta \frac{\partial H}{\partial z_i} \Delta z_i \, dx \, dt + \int_\sigma \Delta \frac{\partial h}{\partial z_i} \Delta z_i \, d\sigma \right].$$

Consequently,

$$\sum_{i=1}^{m} \left[\int_C \Delta z_i \, L_{it} \, \Delta y \, dx \, dt + \int_\sigma \Delta z_i \, P_i \, \Delta y \, d\sigma \right]$$

$$= \frac{1}{2} \left[\sum_{i=1}^{N} \int_C \Delta \frac{\partial H}{\partial w_i} \Delta w_i \, dx \, dt + \sum_{i=1}^{2m} \int_\sigma \Delta \frac{\partial h}{\partial p_i} \Delta p_i \, d\sigma \right], \tag{4.15}$$

where $N = 2m + nm$ is the dimension of the vector w.

Analogously, we find:

$$\sum_{i=1}^{m} \left[\int_C \Delta z_i \, L_{it} \, y \, dx \, dt + \int_\sigma \Delta z_i \, P_i \, y \, d\sigma \right]$$

$$= \sum_{i=1}^{m} \left[\int_C \frac{\partial H}{\partial z_i} \Delta z_i \, dx \, dt + \int_\sigma \frac{\partial h}{\partial z_i} \Delta z_i \, d\sigma \right], \tag{4.16}$$

$$
\begin{aligned}
&\sum_{i=1}^{m}\left[\int_C z_i\, L_{it}\,\Delta y\,dx\,dt+\int_\sigma z_i\, P_i\,\Delta y\,d\sigma\right]\\
&\quad=-\sum_{i=1}^{m}\left[\int_G \alpha_i(x)\Delta y_i(T,x)\,dx+\int_C \beta_i(t,x)\Delta y_i(t,x)\,dx\,dt\right.\\
&\qquad\left.+\int_\sigma \gamma_i(t,x)\Delta y_i(t,x)\,d\sigma\right]\\
&\qquad+\sum_{i=1}^{m}\left[\int_C\left(\frac{\partial H}{\partial y_i}\Delta y_i+\sum_{k=1}^{n}\frac{\partial H}{\partial y_{ix_k}}\Delta y_{ix_k}\right)dx\,dt+\int_\sigma\frac{\partial h}{\partial y_i}\Delta y_i\,d\sigma\right].
\end{aligned}
\tag{4.17}
$$

The first sum on the right-hand side of (4.17) is the increment ΔS of (4.6) when the control $\omega(t, x)$ goes over to the control $\omega(t, x) + \Delta\omega$. Therefore, from (4.12), (4.15), (4.16) and (4.17), it follows that

$$
\begin{aligned}
\Delta S=&-\int_C\left[H(t,x,w+\Delta w,u+\Delta u)-H(t,x,w,u)\right.\\
&\qquad\left.-\sum_{i=1}^{N}\left(\frac{\partial H}{\partial w_i}+\frac{1}{2}\Delta\frac{\partial H}{\partial w_i}\right)\Delta w_i\right]dx\,dt\\
&-\int_\sigma\left[h(t,x,p+\Delta p,v+\Delta v)-h(t,x,p,v)\right.\\
&\qquad\left.-\sum_{i=1}^{2m}\left(\frac{\partial h}{\partial p_i}+\frac{1}{2}\Delta\frac{\partial h}{\partial p_i}\right)\Delta p_i\right]d\sigma.
\end{aligned}
$$

Applying Taylor's formula to the functions h, H, $\partial H/\partial w_i$ and $\partial h/\partial p_i$, and restricting ourselves to the second-order terms in the expansions, we obtain, in just the same way as in §1,

$$
\begin{aligned}
\Delta S=&-\int_C\left[H(t,x,w,u+\Delta u)-H(t,x,w,u)\right]dx\,dt\\
&-\int_\sigma\left[h(t,x,p,v+\Delta v)-h(t,x,p,v)\right]d\sigma-\eta,
\end{aligned}
\tag{4.18}
$$

where $\eta=\eta_1+\eta_2$,

$$
\begin{aligned}
\eta_1=&\frac{1}{2}\sum_{i=1}^{N}\int_C\left(\frac{\partial H(t,x,w,u+\Delta u)}{\partial w_i}-\frac{\partial H(t,x,w,u)}{\partial w_i}\right)\Delta w_i\,dx\,dt\\
&+\frac{1}{2}\sum_{i=1}^{2m}\int_\sigma\left(\frac{\partial h(t,x,p,v+\Delta v)}{\partial p_i}-\frac{\partial h(t,x,p,v)}{\partial p_i}\right)\Delta p_i\,d\sigma,\\
\eta_2=&\frac{1}{2}\left\{\sum_{i,k=1}^{N}\int_C\left[\frac{\partial^2 H(t,x,w+\theta_1\Delta w,u+\Delta u)}{\partial w_i\,\partial w_k}\right.\right.\\
&\left.\left.-\frac{\partial^2 H(t,x,\ w+\theta_2\Delta w,v+\Delta v)}{\partial w_i\,\partial w_k}\right]\Delta w_i\,\Delta w_k\,dx\,dt\right.
\end{aligned}
\tag{4.19}
$$

$$+\sum_{i,k=1}^{2n}\int_\sigma\left[\frac{\partial^2 h(t,x,p+\theta_3\,\Delta p, v+\Delta v)}{\partial p_i\,\partial p_k}\right.$$
$$\left.-\frac{\partial^2 h(t,x,p+\theta_4\,\Delta p, v+\Delta v)}{\partial p_i\,\partial p_k}\right]\Delta p_i\,\Delta p_k\,d\sigma.$$

To obtain the necessary estimates of the remainder term η in (4.18) we reduce the boundary value problem to a system of integro-differential equations (for example, see [34, pp. 90–96]):

$$\Delta y(t,x) = \int_0^t\int_G K_{11}(t,x,\tau,\xi)\Delta\frac{\partial H}{\partial z}\,d\xi\,d\tau$$
$$+\int_0^t\int_\Gamma K_{12}(t,x,\tau,\xi)\psi(\tau,\xi)\,d_\xi\,\sigma\,d\tau, \qquad x\in G, \tag{4.20}$$

$$\psi(t,X) = -\Delta\frac{\partial h}{\partial z}+\int_0^t\int_G K_{21}(t,X,\tau\,\xi)\Delta\frac{\partial H}{\partial z}\,d\xi\,d\tau$$
$$+\int_0^t\int_\Gamma K_{22}(t,X,\tau,\xi)\psi(\tau,\xi)\,d_\xi\,\sigma\,d\tau, \qquad X\in\Gamma, \tag{4.21}$$

where

$$\Delta\frac{\partial H}{\partial z}=\left(\Delta\frac{\partial H}{\partial z_1},\cdots,\Delta\frac{\partial H}{\partial z_m}\right), \qquad \Delta\frac{\partial h}{\partial z}=\left(\Delta\frac{\partial h}{\partial z_1},\cdots,\Delta\frac{\partial h}{\partial z_m}\right),$$

and K_{ik} is a matrix of the Green type. By inserting the values of $\psi(t,X)$ obtained from (4.21) into the right-hand side of the same relation and by successively repeating this operation, we get:

$$\psi(t,X) = -\Delta\frac{\partial h}{\partial z}+\int_0^t\int_G K_n(t,X,\tau,\eta)\Delta\frac{\partial H}{\partial z}\,d\eta\,d\tau$$
$$+\int_0^t\int_\Gamma K^n(t,X,\tau,\eta)\psi(\tau,\eta)\,d_\eta\,\sigma\,d\tau \tag{4.22}$$
$$-\int_0^t\int_\Gamma\sum_{i=0}^{n-1}K^i(t,X,\tau,\eta)\Delta\frac{\partial h}{\partial z}\,d_\eta\,\sigma\,d\tau,$$

where

$$K_n(t,X,\tau,\eta)=K_{n-1}(t,X,\tau,\eta)+\int_\tau^t\int_\Gamma K_{n-1}(t,X,\alpha,\beta)K_0(\alpha,\beta,\tau,\eta)\,d_\beta\,\sigma\,d\alpha,$$

$$K^n(t,X,\tau,\eta)=\int_\tau^t\int_\Gamma K^{n-1}(t,X,\alpha,\beta)K^0(\alpha,\beta,\tau,\eta)\,d_\beta\,\sigma\,d\alpha,$$

$$K_0=K_{21}, \qquad K^0=K_{22}, \qquad n=1,2,\cdots.$$

The number n is chosen so large that the kernel K^n is bounded. This can

be done by virtue of the known estimates for the Green matrix and its derivatives (for example, see [34, p. 92]). Then from (4.22) we have:

$$w(t) \leqq P\int_0^t w(\tau)\,d\tau + \int_0^t\int_G Q_n(t,\tau,\eta)\left|\Delta\frac{\partial H}{\partial z}\right| d\eta\,d\tau + \int_0^t\int_\Gamma R_n(t,\tau,\eta)\left|\Delta\frac{\partial h}{\partial z}\right| d_\eta\,\sigma\,d\tau + \int_\Gamma\left|\Delta\frac{\partial h}{\partial z}\right| d\sigma, \tag{4.23}$$

where P is a specific positive constant,

$$w(t) = \int_\Gamma |\psi(t,X)|\,d\sigma, \qquad Q_n = \int_\Gamma |K_n(t,X,\tau,\eta)|\,d_x\,\sigma,$$
$$R_n = \int_\Gamma \sum_{i=0}^{n-1} |K^i(t,X,\tau,\eta)|\,d_x\,\sigma.$$

We introduce the notations:

$$w_k(t) = \int_0^t w_{k-1}(\tau)\,d\tau, \qquad w_0(t) = w(t),$$
$$Q_{nk} = \int_0^t Q_{nk-1}(t,\tau,\eta)\,d\tau, \qquad Q_{n0} = Q_n, \tag{4.24}$$
$$R_{nk} = \int_0^t R_{nk-1}(t,\tau,\eta)\,d\tau, \quad R_{n1} = R_n(t,\tau,\eta) + 1, \quad k = 2,3,\cdots.$$

By sequentially integrating (4.23) we find:

$$w_k(t) \leqq P\int_0^t w_k(\tau)\,d\tau + \int_0^t\int_G Q_{nk}(t,\tau,\eta)\left|\Delta\frac{\partial H}{\partial z}\right| d\eta\,d\tau + \int_0^t\int_\Gamma R_{nk}(t,\tau,\eta)\left|\Delta\frac{\partial h}{\partial z}\right| d\sigma\,d\tau. \tag{4.25}$$

We choose the number k so large that the functions Q_{nk} and R_{nk} are bounded when $0 \leqq \tau \leqq t \leqq T$, $x \in G$, and we set

$$Q(t) = \int_0^t\int_G \left|\Delta\frac{\partial H}{\partial z}\right| \max_{0\leqq\theta\leqq t} Q_{nk}(\theta,\tau,\eta)\,d\eta\,d\tau,$$
$$R(t) = \int_0^t d\tau\int_\Gamma \left|\Delta\frac{\partial h}{\partial z}\right| \max_{0\leqq\theta\leqq t} R_{nk}(\theta,\tau,\eta)\,d\sigma.$$

Then, when $0 \leqq \theta \leqq t$, from (4.25) we have:

$$w_k(\theta) \leqq P\int_0^\theta w_k(\theta)\,d\theta + Q(t) + R(t).$$

Hence, by a well-known lemma (see [30, p. 19]) we get that

$$w_k(\theta) \leqq A[Q(t) + R(t)]$$

when $0 \leqq \theta \leqq t \leqq T$ and, consequently,

$$w_k(t) \leqq A[Q(t) + R(t)],$$

where A is a specific positive constant. Therefore, from (4.23), (4.24) and (4.25) it follows that

$$w_k(t) \leqq \int_0^t d\tau \left[\int_G M_1(t, \tau, \eta) \left| \Delta \frac{\partial H}{\partial z} \right| d\eta + \int_\Gamma N_1(t, \tau, \eta) \left| \Delta \frac{\partial h}{\partial z} \right| d\sigma \right], \tag{4.26}$$

where M_1 and N_1 are functions of the same type as Q_n and R_n, respectively.

Since the number n is chosen sufficiently large, from (4.22) and (4.26) we have:

$$\begin{aligned} |\psi(t, X)| \leqq \int_0^T d\tau \Bigg[& \int_G M_2(t, \tau, X, \eta) \left| \Delta \frac{\partial H}{\partial z} \right| d\eta \\ & + \int_\Gamma N_2(t, \tau, X, \eta) \left| \Delta \frac{\partial h}{\partial z} \right| d_\eta \sigma \Bigg], \end{aligned} \tag{4.27}$$

where M_2 and N_2 are scalar functions of the Green function type.

The functions $\partial H/\partial z$ and $\partial h/\partial z$ are continuous in t and twice continuously differentiable in the rest of the arguments. Therefore, under every admissible control we can differentiate (4.20) with respect to $x_1, \cdots, x_n$ and by the method set forth above obtain the inequality:

$$\begin{aligned} |\Delta g_i(t, x)| \leqq & \int_G M_3(t, x, \tau, \eta) \sum_{s=1}^{r} |\Delta u_s(\tau, \eta)| \, d\tau \, d\eta \\ & + \int_\sigma N_3(t, x, \tau, \eta) \sum_{j=1}^{q} |\Delta v_j(\tau, \eta)| \, d_{\tau,\eta} \sigma, \end{aligned} \tag{4.28}$$

where

$$g = \left(y_1, \cdots, y_m, \frac{\partial y_1}{\partial x_1}, \cdots, \frac{\partial y_m}{\partial x_n} \right), \qquad x \in G, \qquad 0 \leqq t \leqq T.$$

Since the functions $\Delta z_i(t, x)$, $i = 1, \cdots, m$, form the solution of boundary value problem (4.14), analogously we find that

$$\begin{aligned} |\Delta z_i(t, x)| \leqq & \int_G M_4(t, x, \tau, \eta) \sum_{s=1}^{r} |\Delta u_s(\tau, \eta)| \, d\eta \, d\tau \\ & + \int_\sigma N_4(t, x, \tau, \eta) \sum_{j=1}^{q} |\Delta v_j(\tau, \eta)| \, d_{\tau,\eta} \sigma, \qquad x \in G, \\ & 0 \leqq t \leqq T, \quad i = 1, \cdots, m. \end{aligned} \tag{4.29}$$

By virtue of these inequalities, from (4.19) we have:

$$
\begin{aligned}
|\eta_1| \leqq B_1 \int_C \sum_{j=1}^{r} |\Delta u_j(t,x)| \Big\{ \int_C M_{11}(t,x,\tau,\eta) \sum_{j=1}^{r} |\Delta u_j(\tau,\eta)|\, d\eta\, d\tau \\
+ \int_\sigma N_{11}(t,x,\tau,\eta) \sum_{k=1}^{q} |\Delta v_k|\, d_{\tau,\eta}\,\sigma \Big\}\, dx\, dt \\
+ B_2 \int_\sigma \sum_{i=1}^{q} |\Delta v_i(t,x)| \Big\{ \int_C M_{11}(t,x,\tau,\eta) \sum_{j=1}^{r} |\Delta u_j|\, d\eta\, d\tau \\
+ \int_\sigma N_{11}(t,x,\tau,\eta) \sum_{k=1}^{q} |\Delta v_k|\, d_{\tau,\eta}\,\sigma \Big\}\, d_{t,x}\,\sigma,
\end{aligned}
\tag{4.30}
$$

where the B_i are positive constants, $M_{11} = M_3 + M_4$, $N_{11} = N_3 + N_4$. Since $\partial^2 H/\partial w_i \partial w_j$ and $\partial^2 h/\partial p_i \partial p_j$ are bounded by hypothesis,

$$
\begin{aligned}
|\eta_2| \leqq B_3 \int_C \Big[\int_C M_{11}(t,x,\tau,\eta) \sum_{k=1}^{r} |\Delta u_k|\, d\tau\, d\eta \\
+ \int_\sigma N_{11}(t,x,\tau,\eta) \sum_{j=1}^{q} |\Delta v_j|\, d_{\tau,\eta}\,\sigma \Big]^2 dx\, dt \\
+ B_4 \int_\sigma \Big[\int_C M_{11}(t,x,\tau,\eta) \sum_{k=1}^{r} |\Delta u_k|\, d\eta\, d\tau \\
+ \int_\sigma N_{11}(t,x,\tau,\eta) \sum_{j=1}^{q} |\Delta v_j|\, d_{\tau,\eta}\,\sigma \Big]^2 d_{t,x}\,\sigma.
\end{aligned}
\tag{4.31}
$$

If we consider that the constant B can be chosen so that

$$
\sum_{k=1}^{r} \int_C |\Delta u_k(t,x)|^2\, dx\, dt \leqq B^2, \qquad \sum_{j=1}^{q} \int_\sigma |\Delta v_j(t,x)|^2\, d\sigma \leqq B^2,
$$

then from (4.30) and (4.31) it will follow that the estimate

$$
\begin{aligned}
|\eta| \leqq \int_C \Big\{ \int_C P(t,x,\tau,\eta) \sum_{k=1}^{r} |\Delta u_k|\, d\tau\, d\eta \\
+ \int_\sigma Q(t,x,\tau,\eta) \sum_{j=1}^{q} |\Delta v_j|\, d_{\tau,\eta}\,\sigma \Big\}^2 dx\, dt \\
+ \int_\sigma \Big\{ \int_C P(t,x,\tau,\eta) \sum_{k=1}^{r} |\Delta u_k|\, d\tau\, d\eta \\
+ \int_\sigma Q(t,x,\tau,\eta) \sum_{j=1}^{q} |\Delta v_j|\, d_{\tau,\eta}\,\sigma \Big\}^2 d_{t,x}\,\sigma,
\end{aligned}
\tag{4.32}
$$

where the functions P and Q are of the same type as M_{11} and N_{11}, is valid for the remainder term η in (4.18).

Formula (4.18) and (4.32) are analogous to the corresponding relations in §1. Therefore, the proof of Theorem 7 coincides almost word for word with the proof of Theorem 1.

If (4.4) is linear and has the form

$$L_{it}y = \sum_{k=1}^{m} d_{ik}(t, x)y_k + f_i(v), \qquad i = 1, \cdots, m, \tag{4.33}$$

and if the boundary conditions are given as

$$\begin{aligned} P_i(t, x)y &= \sum_{k=1}^{m} c_{ik}(t, x)y_k + \varphi_i(v), \\ x \in \Gamma, \qquad y(0, x) &= a(x), \qquad x \in G, \end{aligned} \tag{4.34}$$

then the following theorem is valid.

THEOREM 8. *If to every admissible control there corresponds a unique solution of the boundary value problem* (4.33)–(4.34), *then in order that the control* $\omega(t, x) = (u(t, x), v(t, x))$ *be min-optimal (max-optimal) with respect to functional* (4.6), *it is necessary and sufficient that it satisfy the maximum (minimum) conditions.*

The proof of this theorem follows immediately from the fact that in the case being considered formula (4.18) for the increment of functional S takes the form

$$\begin{aligned} \Delta S = &-\int_G [H(t, x, w, u + \Delta u) - H(t, x, w, u)]\, dx\, dt \\ &- \int_\sigma [h(t, x, p, v + \Delta v) - h(t, x, p, v)]\, d\sigma. \end{aligned} \tag{4.35}$$

4.3. Problems with other optimality criteria. The results we have obtained can be used to solve optimal control problems with other optimality criteria.

For example, let the controlled process be described by the boundary value problem (4.4)–(4.5) in which the region G is the rectangle $0 \leqq x_i \leqq X_i$ and in which the optimality criterion is chosen to be the functional

$$S = \int_0^T \int_0^{X_1} \int_0^{X_2} f_0(t, x, y, y_x, u)\, dx_2\, dx_1\, dt. \tag{4.36}$$

We introduce the auxiliary variable y_0 by means of the relations

$$\frac{\partial^3 y_0}{\partial x_1\, \partial x_2\, \partial t} = f_0(t, x, y, y_x, u), \quad y_0(x_1, x_2, 0) = y_0(x_1, 0, t) = y_0(0, x_2, t) = 0.$$

Then the problem reduces to a search for the minimum of the functional $S = y_0(X_1, X_2, T)$. We construct the function $\bar{H}$:

$$\bar{H}(t, x, w, u) = \sum_{i=1}^{m} z_i f_i(t, x, y, y_x, u) + z_0 f_0(t, x, y, y_x, u).$$

The functions $z_i(t, x)$ are determined from the equations

$$M_{it}\, z = -\frac{\partial \bar{H}}{\partial y_i} + \sum_{k=1}^{2} \frac{d}{dx_k}\left(\frac{\partial \bar{H}}{\partial y_{ixk}}\right), \qquad i = 1, \cdots, m,$$

$$\frac{\partial^3 z_0}{\partial x_1\, \partial x_2\, \partial t} = 0$$

and from the subsidiary conditions (see (2.19) and (4.10)):

$$Q_{it}\, z = \frac{\partial h(t, x, p, v)}{\partial y_i} + \sum_{k=1}^{2} \frac{\partial H(t, x, w, u)}{\partial y_{ixk}} X_k(x), \qquad x \in \Gamma,$$

$$\frac{\partial z_0}{\partial t} = 0 \quad \text{when} \quad x_1 = X_1, \qquad x_2 = X_2,$$

$$\frac{\partial z_0}{\partial x_1} = 0 \quad \text{when} \quad t = T, \qquad x_2 = X_2,$$

$$\frac{\partial z_0}{\partial x_2} = 0 \quad \text{when} \quad t = T, \qquad x_1 = X_1,$$

$$\frac{\partial^2 z_0}{\partial x_1\, \partial x_2} = 0 \quad \text{when} \quad t = T, \qquad \frac{\partial^2 z_0}{\partial x_1\, \partial t} = 0 \quad \text{when} \quad x_2 = X_2,$$

$$\frac{\partial^2 z_0}{\partial x_2\, \partial t} = 0 \quad \text{when} \quad x_1 = X_1,$$

$$z_0(X_1, X_2, T) = -1, \qquad z_i(x_1, x_2, T) = 0, \qquad i = 1, \cdots, m.$$

Thus, $z_0(x_1, x_2, t) = -1$, and the function $\bar{H}$ takes the form

$$\bar{H} = H(t, x, w, u) - f_0(t, x, y, y_x, u),$$

where

$$H = \sum_{i=1}^{m} z_i f_i(t, x, y, y_x, u).$$

Consider the functional

$$I = \int_0^T \int_0^{X_1} \int_0^{X_2} \left[\sum_{i=1}^{m} z_i L_{it} y + z_0 \frac{\partial^3 y_0}{\partial x_1 \partial x_2 \partial t} - \bar{H}(t, x, w, u)\right] dx_2\, dx_1\, dt$$

$$+ \int_0^T \int_\Gamma \left[\sum_{i=1}^{m} z_i P_i y - h(t, x, p, v)\right] d\sigma\, dt = I_1 + I_2,$$

where

$$I_1 = \int_0^T \int_0^{X_1} \int_0^{X_2} \left[\sum_{i=1}^{m} z_i L_{it} y - H(t, x, w, u)\right] dx_2\, dx_1\, dt$$

$$+ \int_0^T \int_\Gamma \left[\sum_{i=1}^m z_i P_i(t, x) y - h(t, x, p, v) \right] d\sigma \, dt,$$

$$I_2 = \int_0^T \int_0^{X_1} \int_0^{X_2} z_0 \left[\frac{\partial^3 y_0}{\partial x_1 \partial x_2 \partial x_3} - f_0(t, x, y, y_x, u) \right] dx_2 \, dx_1 \, dt.$$

By transforming integrals I_1 and I_2 in the same way as we did in §§1 and 4, we obtain the formula for the increment of (4.36) in the following form:

$$\Delta S = - \int_0^T \int_0^{X_1} \int_0^{X_2} [\bar{H}(t, x, w, u + \Delta u) - \bar{H}(t, x, w, u)] \, dx_2 \, dx_1 \, dt$$

$$- \int_0^T \int_\Gamma [h(t, x, p, v + \Delta v) - h(t, x, p, v)] \, d\sigma \, dt - \eta,$$

where the remainder term η is determined by formulas analogous to (4.19).

Consequently, the necessary conditions for optimality in the problem being considered can be formulated as Theorem 7 where the function H is replaced by $\bar{H}$ in the maximum (minimum) conditions.

By using the results of §1 we can study, analogously, other optimal process control problems when as optimality criteria we choose various nonlinear functionals. In particular, the results obtained can be applied to the investigation of the problems treated in references [5], [6].

4.4. Optimal problems in the theory of elliptic systems. Control problems analogous to those we have considered above arise during the study of diffusion processes (see [3], [8]). However here we have to consider boundary value problems for elliptic equations. Problems of the same type arise in the investigation of the optimal distribution of thermal and electromagnetic fields in various power installations.

In this subsection we briefly state the minimax problem for elliptic systems and obtain a formula for the increment of the functional, with the aid of which we find the optimality conditions.

Thus, let there be given the elliptic system of equations

$$(4.37) \qquad Ly = f(x, y, y_x, u), \qquad x = (x_1, \cdots, x_n) \in G,$$

where the operator L is defined by (4.1) and the function $f = (f_1, \cdots, f_m)$ is twice continuously differentiable in all its arguments. The control parameter u takes values from a bounded region U (closed or open) of an r-dimensional Euclidean space.

Further, let the function $y(x)$ satisfy the boundary conditions

$$(4.38) \qquad P_i(x)y = \varphi_i(x, y, v), \qquad i = 1, \cdots, m, \qquad x \in \Gamma,$$

where the φ_i satisfy the very same conditions as the f_i, and the parameter v takes values from a bounded region V of a q-dimensional Euclidean space.

The admissible control $\omega(x) = (u(x), v(x))$ is defined in the same way as in §4.1, and we shall assume that the desired function satisfies certain smoothness conditions on the surfaces of discontinuity of the control (see [35]). We shall assume that in addition to the conditions listed above, some other constraints are imposed on the known functions in the boundary value problem, under which a unique solution of this problem corresponds to each admissible control.

We pose the problem: among all the admissible controls determine the control $\omega(x)$ (if it exists) such that the solution of boundary value problem (4.37)–(4.38) corresponding to it realizes the minimum (maximum) of the functional

$$S = \sum_{i=1}^{m} \left[\int_G \alpha_i(x) y_i(x)\, dx + \int_\Gamma \gamma_i(x) y_i(x)\, d\sigma \right], \tag{4.39}$$

where $\alpha_i(x)$ and $\gamma_i(x)$ are given continuous functions.

We introduce the functions $H = \sum z_i f_i$ and $h = \sum z_i \varphi_i$. The functions $z_i(x)$ are determined as the solution of the boundary value problem

$$M_i z = \frac{\partial H}{\partial y_i} - \sum_{k=1}^{n} \frac{d}{dx_k}\left(\frac{\partial H}{\partial y_{ix_k}}\right) - \alpha_i(x), \qquad x \in G, \tag{4.40}$$

$$Q_i z = -\frac{\partial h}{\partial y_i} - \sum_{k=1}^{n} \frac{\partial H}{\partial y_{ix_k}} X_k(x) + \gamma_i(x), \qquad x \in \Gamma, \tag{4.41}$$

(the operators M_i and Q_i were defined at the beginning of the section). If it happens that the right-hand side of (4.40) contains the derivatives $v_{x_1}(x), \cdots, v_{x_n}(x)$, then we should require that the admissible controls have piecewise-continuous derivatives with sufficiently smooth discontinuity boundaries. Then, the boundary value problem (4.40)–(4.41) has a unique solution for each admissible control.

By the same method as was used above we can obtain the formula for the increment of (4.39) in the following form:

$$\begin{aligned} \Delta S = &- \int_G [H(x, w, u + \Delta u) - H(x, w, u)]\, dx \\ &- \int_\Gamma [h(x, p, v + \Delta v) - h(x, p, v)]\, d\sigma - \eta, \end{aligned} \tag{4.42}$$

where the remainder term η is determined by formulas analogous to (4.19). If boundary value problem (4.37)–(4.38) is linear, $\eta = 0$ and, consequently, the following theorem is valid.

THEOREM 9. *In order that an admissible control be locally min-optimal (max-optimal) with respect to functional* (4.39) *in the linear boundary value problem*

(4.37)–(4.38) (*the functions f_i and φ_i are linear in y and y_x*), *it is necessary and sufficient that this control satisfy the maximum (minimum) conditions.*

In conclusion let us remark that analogous problems can be considered (with analogous results) also for systems of hyperbolic equations with initial and boundary conditions.

5. Certain problems in invariance theory. Let the controlled process be described by a system of partial differential equations

$$Az = f(x_1, \cdots, x_k, z, u), \tag{5.1}$$

where A is a linear differential operator of the parabolic, elliptic or hyperbolic type, $z = (z_1, \cdots, z_m)$ is a vector characterizing the state of the controlled system, and u is a vector characterizing the external excitations. Let there also be given subsidiary conditions in which the vector v, determining the external excitations on the system, occurs. It is assumed that the vector $\omega = (u, v)$ is subject to the same conditions as the admissible control in the optimal control problems considered above and to subsidiary conditions such that to each vector ω there corresponds a unique solution of (5.1) with these subsidiary conditions.

Further, let there be given a certain functional $I[z]$, defined on the solutions of (5.1). The fundamental problem in invariance theory consists of finding the conditions which when satisfied make the functional I independent of the external excitation. In [17] it was shown that the invariance problem can be studied by the methods of the calculus of variations in the case when the controlled process was described by ordinary differential equations. Analogously, we can study the invariance problem also for systems with distributed parameters.

Let us consider the control system whose behavior is described by boundary value problem (4.9) with certain smoothness conditions on the surfaces of discontinuity of the function $u(t, x)$. We shall assume that to every admissible vector $\omega(t, x) = (u(t, x), v(t, x))$ there corresponds a unique solution of this boundary value problem, and that

$$\begin{aligned} f_i(t, x, y, y_x, u) &= \sum_{k=1}^{m} d_{ik}(t, x)y_k + g_i(t, x)u, \\ \varphi_i(t, x, y, v) &= p_i(t, x)v, \end{aligned} \tag{5.2}$$

where for the sake of simplifying the succeeding formulas, u and v are taken to be scalar quantities.

As the functional I we shall take (4.6) in which the time T and the region G are taken as fixed. In this case the "adjoint" boundary value problem (4.10) will have the form:

$$M_{it}z = -\sum_{k=1}^{m} d_{ki}z_k + \beta_i(t, x), \qquad z_i(T, x) = -\alpha_i(x), \qquad x \in G, \tag{5.3}$$
$$Q_i z = -\gamma_i(t, x), \qquad x \in \Gamma.$$

Formula (4.35) for the increment of (4.6) takes the form:

$$\Delta S = -\int_C \Delta u \left(\sum_{i=1}^{m} g_i z_i\right) dx\, dt - \int_\sigma \Delta v \sum p_i z_i\, d\sigma.$$

Consequently, if

$$\sum_{i=1}^{m} g_i(t, x)z_i(t, x) \equiv 0, \qquad x \in G,$$
$$\sum_{i=1}^{m} p_i(t, x)z_i(t, x) \equiv 0, \qquad x \in \Gamma, \qquad 0 \leqq t \leqq T, \tag{5.4}$$

the functional S is independent of the external excitation $\omega(t, x)$. By the method of contradiction it is easy to establish that these conditions are also necessary for the functional to be independent of ω (for example, see [17]). To check (5.4) we must find the solution of boundary value problem (5.3). However, for the special case presented below we can successfully obtain the necessary and sufficient invariance conditions expressed in terms of the coefficients of the equations of boundary value problem (4.9).

Let the controlled process be described by the equations

$$Ly_i = \sum_{k=1}^{m} d_{ik}\, y_k + g_i u$$
$$\left(Ly_i \equiv \frac{\partial y_i}{\partial t} - \sum_{j,k=1}^{n} a_{jk} \frac{\partial^2 y_i}{\partial x_j \partial x_k}, \quad d_{ik}, g_i = \text{const.}\right) \tag{5.5}$$

with the subsidiary conditions

$$y_i(0, x) = a_i(x), \qquad x \in G, \qquad Py_i = \psi_i(t, x), \qquad x \in \Gamma; \tag{5.6}$$

here P is a linear differential operator defined on the boundary Γ, where the differentiation is carried out in the direction external relative to G.

Consider the functional

$$S = \sum_{i=1}^{m} \left[\int_G \alpha_i(x)y_i(T, x)\, dx + \int_0^T \int_\Gamma \gamma_i(t, x)y_i(t, x)\, d\sigma\, dt\right]. \tag{5.7}$$

Then, the functions $z_i(t, x)$ occurring in (5.4) are determined from the equations

$$Mz_i = -\sum_{k=1}^{m} d_{ki} z_k$$

$$\left(Mz_i = \frac{\partial z_i}{\partial t} + \sum_{j,k=1}^{n} \frac{\partial}{\partial x_j}\left(a_{jk} \frac{\partial z_i}{\partial x_k} \right) + \sum_{j=1}^{n} \frac{\partial}{\partial x_j} (l_j z_i) \right) \tag{5.8}$$

with the subsidiary conditions

$$\begin{aligned} z_i(T, x) &= -\alpha_i(x), \qquad x \in G, \\ Q(t, x) z_i &= -\gamma_i(t, x), \qquad x \in \Gamma, \qquad 0 \leqq t \leqq T, \end{aligned} \tag{5.9}$$

where, in accordance with (4.3), the operator Q is defined as the adjoint to P.

By D and D^* we denote the matrix of coefficients d_{ik} and the matrix adjoint to it, and by $\langle r, s \rangle$, the scalar product of the vectors r and s. Since in the problem being considered the boundary conditions (5.6) do not contain v, (5.4) can be written as

$$R(t, x) \equiv \langle z, g \rangle = 0, \qquad x \in G, \qquad 0 \leqq t \leqq T.$$

Applying operator M to this equality and taking into account that z satisfies (5.8), we have:

$$MR(t, x) = \langle Mz, g \rangle = -\langle D^* z, g \rangle = \langle z, Dg \rangle = 0.$$

Analogously we find that

$$M^k R(t, x) = (-1)^k \langle z, D^k g \rangle = 0, \qquad k = 0, 1, \cdots, m - 1. \tag{5.10}$$

Hence, by virtue of (5.9) it follows that

$$M^k R(T, x) = -(-1)^k \langle \alpha(x), D^k g \rangle = 0, \quad k = 0, 1, \cdots, m - 1. \tag{5.11}$$

Assuming in (5.10) that $x \in \Gamma$ and applying operator Q, with due regard to conditions (5.9), we obtain:

$$QM^k R(t, x) = (-1)^k \langle Qz, D^k g \rangle = -(-1)^k \langle \gamma(t, x), D^k g \rangle = 0, \qquad k = 0, 1, \cdots, m - 1. \tag{5.12}$$

Conditions (5.11) and (5.12) are necessary for the invariance of (5.7) relative to the external excitation u in boundary value problem (5.5)–(5.6). Let us show that these conditions are also sufficient.

Since by hypothesis at least one of the vectors $\alpha = (\alpha_1, \cdots, \alpha_m)$ and $\gamma = (\gamma_1, \cdots, \gamma_m)$ is nonzero, there exist numbers $\lambda_0, \cdots, \lambda_{m-1}$ such that

$$\sum_{k=0}^{m-1} \lambda_k D^k g = 0.$$

Multiplying the kth equality in (5.10) by $(-1)^k \lambda_k$ and summing over all k we obtain:

$$\sum_{k=0}^{m-1} (-1)^k \lambda_k M^k R(t, x) = 0, \qquad x \in G, \qquad 0 \leqq t \leqq T.$$

By introducing the notation $R_k = M^k R$, $k = 0, 1, \cdots, m-2$, from this equation and from (5.10) and (5.11) we obtain a homogeneous boundary value problem for the determination of R_k :

$$(-1)^{m-1}\lambda_{m-1} M R_{m-2} + (-1)^{m-2}\lambda_{m-2} R_{m-2} + \cdots - \lambda_1 R_1 + \lambda_0 R_0 = 0,$$

$$R_{m-2}(T, x) = 0, \qquad x \in G; \qquad Q R_{m-2}(t, x) = 0, \qquad x \in \Gamma,$$

$$(5.13) \quad M R_{m-3} - R_{m-2} = 0, \qquad R_{m-3}(T, x) = 0, \qquad x \in G;$$

$$Q R_{m-3}(t, x) = 0, \qquad x \in \Gamma,$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$M R_0 - R_1 = 0, \quad R_0(T, x) = 0, \quad x \in G; \quad Q R_0(t, x) = 0, \quad x \in \Gamma.$$

Since it is assumed that the coefficients of operators M and Q are sufficiently smooth, boundary value problem (5.13) has only the trivial solution (for example, see [34, pp. 97–103]):

$$R_{m-2}(t, x) = R_{m-3}(t, x) = \cdots = R_0(t, x) \equiv 0,$$

and hence follows the fulfillment of the condition

$$\sum_{i=1}^{m} g_i(t, x) z_i(t, x) = 0, \qquad x \in G.$$

By the same token we have proved Theorem 10.

THEOREM 10. *For the invariance of* (5.7) *relative to the external excitation* u *in boundary value problem* (5.5)–(5.6), *it is necessary and sufficient that the following conditions be satisfied*:

$$\langle \alpha(x), D^k g \rangle = 0, \qquad x \in G,$$

$$\langle \gamma(t, x), D^k g \rangle = 0, \qquad x \in \Gamma,$$

$$0 \leqq t \leqq T, \qquad k = 0, 1, \cdots, m-1.$$

From the method by which this theorem was proved it is seen that analogous results can be obtained for the boundary value problems which were studied in §1. In particular, for boundary value problem (1.33) we should make use of (1.35) for the increment of (1.3), where the functions u_i are determined with the aid of boundary value problem (1.34).

REFERENCES

[1] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *Mathematical Theory of Optimal Processes*, Fizmatgiz, Moscow, 1961.

[2] E. P. Popov, *Dynamics of Automatic Control Systems*, GITTL, Moscow, 1954.
[3] *Theory of discrete, optimal and adaptive systems*, Proceedings of the First International Congress of I.F.A.C., Akad. Nauk SSSR, 1961.
[4] G. L. Haratishvili, *The maximum principle in the theory of optimal processes with lag*, Dokl. Akad. Nauk SSSR, 136 (1961), pp. 39–42.
[5] R. Bellman and H. Osborn, *Dynamic programming and the variation of Green's functions*, J. Math. Mech., 7 (1958), pp. 81–85.
[6] A. G. Butkovskii and A. Ya. Lerner, *Optimal control in distributed-parameter systems*, Dokl. Akad. Nauk SSSR, 134 (1960), pp. 778–781.
[7] I. V. Girsanov, *Minimax problems in the theory of diffusion processes*, Ibid., 136 (1961), pp. 761–764.
[8] Yu. V. Egorov, *Certain problems in optimal control theory*, Ibid., 145 (1962), pp. 720–723.
[9] W. H. Fleming, *Some Markovian optimization problems*, J. Math. Mech., 12 (1963), pp. 131–140.
[10] Yu. V. Egorov, *Optimal control in a Banach space*, Dokl. Akad. Nauk SSSR, 150 (1963), pp. 241–244.
[11] ———, *Optimal control in a Banach space*, Uspehi Mat. Nauk, 18 (1963), pp. 211–213.
[12] ———, *Certain problems in optimal control theory*, Zh. Vychisl. Mat. i Mat. Fiz., 3 (1963), pp. 887–904.
[13] A. G. Butkovskii, *Automatic Control Theory in Distributed-Parameter Systems*, Dissertation, Institute of Automation and Remote Control, 1963.
[14] A. I. Egorov, *Optimal control processes in distributed plants*, Prikl. Mat. Meh., 27 (1963), pp. 688–696.
[15] K. A. Lur'e, *The Mayer-Bolza problem for multiple integrals and the optimization of the behavior of distributed-parameter systems*, Ibid., 27 (1963), pp. 842–853.
[16] L. I. Rozonoer, *The maximum principle of L. S. Pontryagin in the theory of optimal systems. I–III*, Avtomat. i Telemeh., 20 (1959), pp. 1320–1334, 1441–1458, 1561–1578.
[17] ———, *Variational approach to the invariance problem in automatic control systems. I–II*, Ibid., 24 (1963), pp. 744–756, 861–870.
[18] B. M. Budak and A. D. Gorbunov, *Difference method of solving the nonlinear Gauss problem*, Dokl. Akad. Nauk SSSR, 117 (1957), pp. 559–562.
[19] Yu. V. Egorov, *Hyperbolic equations with discontinuous coefficients*, Ibid., 134 (1960), pp. 514–517.
[20] Francesco Guglielmino, *Sul problema di Darboux*, Ricerche Mat., 8 (1959), pp. 180–196.
[21] N. P. Salihov, *Solution of the Goursat problem by multipoint difference methods. I–II*, Vestnik Moskov. Univ. Ser. I Mat. Meh., (1961), no. 4, pp. 25–34; no. 6, pp. 3–16.
[22] M. Hukuhara, *Le problème de Darboux pour l'équation* $s = f(x, y, z, p, q)$, Ann. Mat. Pura Appl., 51 (1960), pp. 39–54.
[23] M. Lees, *The Goursat problem*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 518–530.
[24] Jan Kisynski, *Solutions généralisées du problème de Cauchy-Darboux pour l'équation* $s = f(x, y, z, p, q)$, Ann. Univ. Mariae Curie-Sklodowska Sect. A, 14 (1960), pp. 87–109.
[25] Carlo Ciliberto, *Sul'approssimazione delle soluzioni del problema di Darboux per l'equation* $s = f(x, y, z, p, q)$, Ricerche Mat., 10 (1961), pp. 106–139.

[26] A. I. Tihonov and A. A. Samarskii, *The Equations of Mathematical Physics*, GITTL, Moscow, 1953.

[27] A. I. Tihonov, A. A. Zhukovskii and Ya. L. Zabezhinskii, *Absorption of a gas from the flow of air in a layer of grainy material*, Zh. Fiz. Himii, 20 (1946), pp. 1113–1126.

[28] A. V. Lukov, *Transfer Phenomenon in Capillary-Porous Bodies*, GITTL, Moscow, 1954.

[29] F. Tricomi, *Lectures on Partial Differential Equations*, IL, Moscow, 1957.

[30] V. V. Nemyckii and V. V. Stepanov, *Qualitative Theory of Differential Equations*, GITTL, Moscow, 1949.

[31] J. Sansone, *Ordinary Differential Equations II*, IL, Moscow, 1954.

[32] N. I. Ahiezer, *Lectures on the Calculus of Variations*, GITTL, Moscow, 1955.

[33] K. Miranda, *Elliptic Partial Differential Equations*, IL, Moscow, 1957.

[34] T. Ya. Zagorskii, *Mixed Boundary Value Problems for Systems of Parabolic Partial Differential Equations*, L'vovsk. Un-ta, 1961.

[35] O. A. Oleinik, *Boundary value problems for linear elliptic and parabolic equations with discontinuous coefficients*, Izv. Akad. Nauk SSSR Ser. Mat., 25 (1961), pp. 3–20.

[36] I. V. Girsanov, *The solutions of certain boundary value problems for parabolic and elliptic equations with discontinuous coefficients*, Dokl. Akad. Nauk SSSR, 135 (1960), pp. 1311–1313.

[37] L. I. Kamynin and V. N. Maslennikova, *Certain properties of the solutions of mixed boundary value problems for parabolic equations with discontinuous coefficients*, Ibid., 133 (1960), pp. 1003–1006.

[38] L. I. Kamynin, *Solutions of boundary value problems for a parabolic equation with discontinuous coefficients*, Ibid., 139 (1961), pp. 1048–1051.

[39] ———, *Temperature potential methods for parabolic equations with discontinuous coefficients*, Sibirsk. Mat. Zh., 4 (1963), pp. 1071–1105.

[40] G. K. Namazov, *Boundary value problems for a second-order parabolic equation with discontinuous coefficients*, Izv. Akad. Nauk Azerbaidzhan. SSR. Ser. Fiz.-Mat. Tehn. Nauk, (1961), no. 3, pp. 39–46.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

ON THE OPTIMALITY OF A TOTALLY SINGULAR VECTOR CONTROL: AN EXTENSION OF THE GREEN'S THEOREM APPROACH TO HIGHER DIMENSIONS*

GEORGE W. HAYNES†

1. Introduction. We are concerned with the optimality of totally singular vector controls governing dynamical systems of the form

$$\dot{x}_\alpha = A_\alpha(x) + B_{\alpha r}(x)u_r\,, \quad \alpha = 1, \cdots, n, \quad r = 1, \cdots, (n-1), \tag{1.1}$$

and the extension of the Green's theorem approach [1], [2] to higher dimensions to evaluate the optimality of such totally singular vector controls. Adopting the definition due to Hermes [3] a vector control is said to be totally singular when the maximum principle yields no information in the time optimal problem for any components of the optimal control. The usual summation convention on repeated indices is used. Greek letters will assume the values 1 to n, and Roman letters 1 to $(n-1)$; the exceptions to this rule are noted where they occur.

The problem of defining a control set for the dynamical system (1.1) is of paramount importance, because the singularity of a control is an inherent feature of the dynamical system and the function or functional to be extremized, and not the control set per se. It is not the intent here to rule out a singular control because of the limitations on the control imposed by a given control set. Therefore, the maximal control set which overcomes these limitations must necessarily include distributions. It should be noted that Kreindler [4] and Neustadt [5] have considered such control sets in their treatment of linear systems. However, for the nonlinear system considered, we shall effectively circumvent a difficult problem by replacing the dynamical system (1.1) by the equivalent pfaffian system

$$dx_\alpha = A_\alpha(x)\,dt + B_{\alpha r}(x)\,dy_r\,, \tag{1.2}$$

where the control has the representation $u_r = dy_r/dt$ when it exists. The solutions to the pfaffian system (1.2) will be parameterized by $x(\sigma)$, $y(\sigma)$ and $t(\sigma)$ with $t(\sigma)$ monotone such that

$$dx_\alpha(\sigma) \overset{\sigma}{\equiv} A_\alpha(x(\sigma))\,dt(\sigma) + B_{\alpha r}(x(\sigma))\,dy_r(\sigma). \tag{1.3}$$

It is assumed that the vector $x(\sigma)$ has values confined to some simply connected region $D \subset R^n$, also $A_\alpha(x)$ and $B_{\alpha r}(x)$ are twice continuously differ-

* Received by the editors February 4, 1966, and in revised form April 18, 1966.

† The Martin Company, Denver, Colorado. This work was supported by the National Aeronautics and Space Administration, Ames Research Center, under Contract NAS 2-2351.

entiable in D. Furthermore, it is assumed that the system (1.1) is controllable, which implies that there does not exist a scalar function $W(t, x)$ such that the hypersurface $W(t, x) = \text{const.}$ contains all the solutions to the system (1.1) independent of the controls. From this can be inferred [6] that the system of n partial differential equations

$$\frac{\partial W(t, x)}{\partial t} + \frac{\partial W(t, x)}{\partial x_\alpha} A_\alpha(x) = 0, \qquad \frac{\partial W(t, x)}{\partial x_\alpha} B_{\alpha r}(x) = 0, \tag{1.4}$$

is not complete, so that another independent partial differential equation can be determined by the Poisson operator to yield the nonexistence of a nontrivial $W(t, x)$, namely $[\partial W(t, x)]/\partial t = [\partial W(t, x)]/\partial x_\alpha = 0$.

The problem posed is to determine the control which steers the state from some initial point x^0 to some final point x^f in minimum time. We shall now state a further condition which in essence is the sine qua non of the Green's theorem approach to higher dimensions.

CONDITION A. *For each x, the columns of the $B_{\alpha r}(x)$ matrix are $(n - 1)$ linearly independent tangent vectors; furthermore the system of partial differential equations formed with the tangent vectors*

$$\frac{\partial V(x)}{\partial x_\alpha} B_{\alpha r}(x) = 0 \tag{1.5}$$

is a complete system of order $(n - 1)$.

This condition has three important implications which will be developed in detail in later sections; however, for the purposes of motivation we shall briefly describe what these implications are.

(1) Condition A guarantees the existence of a single unique pfaffian to system (1.1).

(2) It provides a necessary condition for the existence of an optimal totally singular vector control. The sufficiency condition for the existence of an optimal totally singular vector control follows from the Green's theorem application.

(3) On applying the n-dimensional Green's theorem to the single pfaffian, there result $[n(n - 1)]/2$ hypersurfaces whose interpretation as singular hypersurfaces (assuming an analogy with the 2-dimensional Green's theorem approach) is doubtful since we need only $(n - 1)$ such hypersurfaces to specify the totally singular vector control. However, Condition A enables an integrability argument to be invoked, and from this it can be shown that no more than $(n - 1)$ hypersurfaces are obtained which can then be interpreted as singular hypersurfaces.

2. Existence of a totally singular vector control. Let $\psi_\alpha(x)$ be a nonzero vector orthogonal to the columns of $B_{\alpha r}(x)$; that is,

$$\psi_\alpha(x)B_{\alpha r}(x) \overset{x}{\equiv} 0. \tag{2.1}$$

Hence the pfaffian system (1.2) can be expressed as a single pfaffian, which is unique to within an arbitrary multiplicative factor by virtue of the linear independence of the columns of $B_{\alpha r}(x)$:

$$\psi_\alpha(x)\, dx_\alpha = \psi_\alpha(x)A_\alpha(x)\, dt. \tag{2.2}$$

Let $x = \phi^s(t)$, $t \in [t_0, t_f]$, represent parametrically a totally singular arc in state space satisfying (1.1), which transfers the state from $x^0 = \phi^s(t_0)$ to $x^f = \phi^s(t_f)$. This singular arc automatically satisfies the pfaffian (2.2); in fact, any solution of the dynamical system (1.1) satisfies the pfaffian (2.2). The question now arises whether it is possible to obtain a parametric representation in state space of the same transfer (but not necessarily the same arc) by $x = x(\sigma)$, $\sigma \in [\sigma_0, \sigma_f]$ and $t = \text{const.}$, which satisfies the pfaffian

$$\psi_\alpha(x(\sigma))\, dx_\alpha(\sigma) \overset{\sigma}{\equiv} 0 \tag{2.3}$$

with $x^0 = x(\sigma_0)$ and $x^f = x(\sigma_f)$. If this is possible, then the totally singular vector control will not be optimum, since the transfer of the state vector from x^0 to x^f can be synthesized by suitable impulses to achieve the transfer in zero time. Therefore, the problem resolves down to the question of accessibility of points by trajectories satisfying the pfaffian

$$\psi_\alpha(x)\, dx_\alpha = 0. \tag{2.4}$$

The resolution of this question leads to the following lemma.

LEMMA 2.1. *A necessary condition that an optimal totally singular vector control exists is that the pfaffian* $\psi_\alpha(x)\, dx_\alpha = 0$ *be integrable.*

Proof. The proof makes use of the following theorem [7] and its contrapositive which we shall state formally.

THEOREM (Carathéodory). *If a pfaffian* $\psi_\alpha(x)\, dx_\alpha = 0$ *has the property that in every arbitrarily close neighborhood of a given point* $\bar{x}$ *there exist points which are inaccessible from* $\bar{x}$ *by trajectories satisfying the pfaffian, then the pfaffian is integrable.*

CONTRAPOSITIVE. *If the pfaffian* $\psi_\alpha(x)\, dx_\alpha = 0$ *is not integrable, then there exists some neighborhood of a given point* $\bar{x}$ *in which all points are accessible by trajectories satisfying the pfaffian.*

The following concise form of the proof is due to the referee.

Proof. Assuming the pfaffian is not integrable, then with each point $\phi^s(t)$ of the singular arc we can associate an open neighborhood of accessibility $\mathfrak{N}(\phi^s(t))$. This gives an open cover of the compact arc; by the Heine-Borel theorem there is a finite subcover. Let $\{\mathfrak{N}(\phi^s(t_0)), \mathfrak{N}(\phi^s(t_1)), \cdots,$

$\mathfrak{N}(\phi^s(t_k), \cdots, \mathfrak{N}(\phi^s(t_f))\}$ be the finite subcover, and assume without loss of generality that $t_0 < t_1 < \cdots < t_k < \cdots < t_f$. Now the point $x^0 = \phi^s(t_0)$ can be joined to a point $\hat{x}$ in $\mathfrak{N}(\phi^s(t_0)) \cap \mathfrak{N}(\phi^s(t_1))$ by an arc satisfying $\psi_\alpha(x)\, dx_\alpha = 0$. Then $\hat{x}$ can be joined to $x^1 = \phi^s(t_1)$ by an arc satisfying $\psi_\alpha(x)\, dx_\alpha = 0$. Continuing this procedure we can construct a zero time "polygonal" arc joining x^0 and x^f, which completes the proof of the lemma.

If the pfaffian $\psi_\alpha(x)\, dx_\alpha = 0$ is integrable, then there exist a nonzero integrating factor $\mu(x)$ and a function $V(x)$ such that

$$\mu(x)\psi_\alpha(x) \overset{x}{\equiv} \frac{\partial V(x)}{\partial x_\alpha}. \tag{2.5}$$

Therefore from (2.1) we have

$$\frac{\partial V(x)}{\partial x_\alpha} B_{\alpha r}(x) \overset{x}{\equiv} 0; \tag{2.6}$$

and by Condition A we are assured that such a $V(x)$ exists, so that the pfaffian $\psi_\alpha(x)\, dx_\alpha = 0$ is integrable. It should be noted that the pfaffian

$$\psi_\alpha(x)\, dx_\alpha - \psi_\tau(x) A_\tau(x)\, dt = 0$$

is not integrable [3]; otherwise this would contradict the assumption that the system (1.1) is controllable.

3. Generalized Green's theorem. The problem of extremizing two-dimensional line integrals of the form

$$I = \int_{x^0}^{x^f} [a_1(x_1, x_2)\, dx_1 + a_2(x_1, x_2)\, dx_2] \tag{3.1}$$

is a fairly simple one, since the relative optimality of two distinct trajectories may be compared directly under certain smoothness conditions by an application of Green's theorem. The unique feature of this approach when applied to two-dimensional nonlinear systems [2] in which the control (single component) appears linearly and the cost functional can be given the representation (3.1) is that the projection of the singular arc in state space is obtained immediately by $\omega(x_1, x_2) = 0$, where

$$\omega(x_1, x_2) = \frac{\partial a_2(x_1, x_2)}{\partial x_1} - \frac{\partial a_1(x_1, x_2)}{\partial x_2}. \tag{3.2}$$

Since only simple algebraic manipulations are required to generate $\omega(x_1, x_2)$, the utility of the method is immediately obvious and motivates this extension to higher dimensions. Before describing the Green's theorem approach to the problem posed, we shall briefly review the extension of Green's theorem to higher dimensions pertinent to the problem. Stoke's theorem in particular is the extension of Green's theorem from two to three dimensions;

and from the theory of exterior calculus [8], [9], [10], Stoke's theorem has been generalized to higher dimensions. Specializing the form of Stoke's theorem [8, Theorem 9.50] to our purpose we have the following theorem.

THEOREM 3.1. *Let $\bar{S}$ be an orientable two-surface of class C'' in an open set $D \subset R^n$; also let the edge or boundary of $\bar{S}$ be a Jordan curve Γ. If $\pi = a_\alpha(x)\, dx_\alpha$ is a pfaffian (or one-form) of class C' in D, then*

$$\int_{\bar{S}} d\pi = \int_\Gamma \pi, \tag{3.3}$$

where $d\pi$ is called the exterior derivative of π and is defined to be the two-form

$$d\pi = \frac{\partial a_\alpha(x)}{\partial x_\beta}\, dx_\beta\, dx_\alpha . \tag{3.4}$$

Here we have adopted the notation due to Flanders [9] in omitting the exterior multiplication sign. Since the integrals (3.3) are oriented integrals, the multiplication of differentials satisfies the rules

$$dx_\alpha\, dx_\beta = -dx_\beta\, dx_\alpha$$

so that

$$dx_\alpha\, dx_\alpha = 0 \qquad \text{(no sum)}.$$

Hence the exterior derivative of π may be expressed as

$$d\pi = \omega_{\alpha\beta}\, dx_\beta\, dx_\alpha , \quad \alpha = 1, \cdots, (n-1), \quad \beta = (\alpha+1), \cdots, n, \tag{3.5}$$

where

$$\omega_{\alpha\beta} = \frac{\partial a_\alpha(x)}{\partial x_\beta} - \frac{\partial a_\beta(x)}{\partial x_\alpha} . \tag{3.6}$$

The proof of (3.3) may be found in such standard texts as Rudin [8], Flanders [9] and Guggenheim [10]. However, we shall sketch one method of proof, since the construction employed therein is essentially the procedure we shall adopt for applying Green's theorem to the control problem stated.

Let f denote a smooth mapping of $P \subset R^2$ into $S \subset D \subset R^n$, so that P, which is described by coordinates (z_1, z_2), is the parameter domain of the two-surface S. Using f^* to denote the induced mapping of the differential forms from S to P then the line integral appearing in (3.3) can be transformed into

$$\int_\Gamma \pi = \int_B f^*\pi, \tag{3.7}$$

where B is so defined that Γ is the image of B under f. Since $f^*\pi$ is a pfaffian in z_1 and z_2, then $P \supset B$ is a region to which Green's theorem applies. On

applying Green's theorem there results

$$\int_B f^*\pi = \int_{\bar{P}} d(f^*\pi), \tag{3.8}$$

where $\bar{P} \subset P$ is the parameter domain of the two-surface $\bar{S}$ under the mapping f, and has B as its boundary. From the properties of the exterior derivative we have

$$d(f^*\pi) = f^*(d\pi); \tag{3.9}$$

and it should be noted that since π is a one-form and the dimension of P is two, this is not a vacuous result.

Using (3.9) with (3.7) and (3.8) completes the proof, namely,

$$\int_\Gamma \pi = \int_B f^*\pi = \int_{\bar{D}} d(f^*\pi) = \int_{\bar{D}} f^* d(\pi) = \int_{\bar{S}} d\pi.$$

Having delineated in principle the manner in which Green's theorem will be applied to the control problem posed, we have to resolve the analogy with the two-dimensional control problems of the projection of the singular arc in state space. If we adopt the procedure of the two-dimensional Green's theorem approach and equate each coefficient of the basic two-forms of $d\pi$ to zero, then from (3.5) each $\omega_{\alpha\beta}(x) = 0$ would be interpreted as a singular hypersurface. But there will exist $[n(n-1)]/2$ such hypersurfaces whereas we need only $(n-1)$ to determine the $(n-1)$ components of the totally singular control. When n is equal to two we obtain the required number of hypersurfaces, while for n greater than two we obtain too many hypersurfaces; however, since for optimality it is necessary that the pfaffian $\psi_\alpha(x)\, dx_\alpha = 0$ is integrable, it will be shown that this implies no more than $(n-1)$ of the $[n(n-1)]/2$ hypersurfaces $\omega_{\alpha\beta} = 0$ are independent.

4. On the optimality of a totally singular vector control. By virtue of Condition A there exists a unique pfaffian to the control system (1.1) which can be expressed as

$$dt = \frac{\psi_\alpha(x)\, dx_\alpha}{\psi_\tau(x) A_\tau(x)}. \tag{4.1}$$

Equivalently the pfaffian could be expressed by (2.5) as

$$dt = \frac{\dfrac{\partial V(x)}{\partial x_\alpha}\, dx_\alpha}{\dfrac{\partial V(x)}{\partial x_\tau} A_\tau(x)}. \tag{4.2}$$

However, the determination of $V(x)$ is inconsequential to the analysis;

what is important is to generate the pfaffian (4.1) from the system of pfaffians (1.2) by the elimination of the differentials dy_τ.

It has been assumed in (4.1) that $\psi_\tau(x)A_\tau(x) \neq 0$ in D. From the controllability requirements it is known that $\psi_\tau(x)A_\tau(x) \not\equiv 0$ in D; otherwise the hypersurface $V(x) = \text{const.}$ would contain all the solutions to (1.1) independent of the controls.

By (4.1) the time required to transfer the state from x^0 to x^f through (1.1) can be expressed as a line integral by

$$I = t_f - t_0 = \int_{x^0}^{x^f} \frac{\psi_\alpha(x)\, dx_\alpha}{\psi_\tau(x)A_\tau(x)}. \tag{4.3}$$

We now perform the usual ritual of comparing two trajectories joining x^0 to x^f that project a Jordan curve Γ in state space.

It is assumed that the two trajectories bound an orientable two-surface. Denoting by I_1 and I_2 the respective costs to traverse the trajectories, and accordingly associating a sense of direction to Γ, we have

$$I_1 - I_2 = \int_\Gamma \frac{\psi_\alpha(x)\, dx_\alpha}{\psi_\tau(x)A_\tau(x)}.$$

On applying the n-dimensional Green's theorem we obtain

$$I_1 - I_2 = \int_{S_{\alpha\beta}} \omega_{\alpha\beta}\, dx_\alpha\, dx_\beta, \qquad \alpha = 1,\cdots,n-1, \quad \beta = \alpha+1,\cdots,n, \tag{4.4}$$

where

$$\omega_{\alpha\beta}(x) = \frac{\partial}{\partial x_\beta}\left\{\frac{\psi_\alpha(x)}{\psi_\tau(x)A_\tau(x)}\right\} - \frac{\partial}{\partial x_\alpha}\left\{\frac{\psi_\beta(x)}{\psi_\tau(x)A_\tau(x)}\right\}. \tag{4.5}$$

From the form of $\omega_{\alpha\beta}$ we have the next lemma.

LEMMA 4.1. *No more than* $(n-1)$ *of the* $[n(n-1)]/2$ *hypersurfaces* $\omega_{\alpha\beta}(x) = 0$ *are independent.*

Proof. The proof follows directly from the following theorem [7].

THEOREM. *A necessary and sufficient condition that the pfaffian* $a_\alpha(x)\, dx_\alpha = 0$ *be integrable is*

$$a_\alpha(x)\left(\frac{\partial a_\beta(x)}{\partial x_\gamma} - \frac{\partial a_\gamma(x)}{\partial x_\beta}\right) + a_\beta\left(\frac{\partial a_\gamma(x)}{\partial x_\alpha} - \frac{\partial a_\alpha(x)}{\partial x_\gamma}\right) + a_\gamma\left(\frac{\partial a_\alpha(x)}{\partial x_\beta} - \frac{\partial a_\beta(x)}{\partial x_\alpha}\right) \overset{x}{\equiv} 0.$$

Since the pfaffian $\psi_\alpha(x)\, dx_\alpha = 0$ is integrable, the pfaffian $[\psi_\alpha(x)\, dx_\alpha]/[\psi_\tau(x)A_\tau(x)] = 0$ also is integrable. Applying the integrability

test yields

$$\psi_\alpha(x)\omega_{\beta\gamma}(x) + \psi_\beta(x)\omega_{\gamma\alpha}(x) + \psi_\gamma(x)\omega_{\alpha\beta}(x) \overset{x}{\equiv} 0, \tag{4.6}$$

from which it follows that only $(n-1)$ of the $[n(n-1)]/2$ hypersurfaces $\omega_{\alpha\beta} = 0$ are independent.

The hypersurfaces $\omega_{\alpha\beta} = 0$ can now be interpreted as singular hypersurfaces, since their common intersection (assuming it exists) yields the totally singular arc. This relation is demonstrated in the next section. The importance of the n-dimensional Green's theorem approach is in the simple algorithms it provides for the determination of the singular hypersurfaces and the totally singular arc. Once this has been accomplished, then it is a relatively simple matter to construct a family of two surfaces as indicated in §3, which contains the totally singular arc for some values of the parameters, and then to use the two-dimensional Green's theorem to evaluate the optimality of the totally singular arc. Since we are primarily interested in evaluating the optimality of the totally singular arc, it is tacitly assumed that the singular hypersurfaces have a common intersection that can be represented in terms of a single parameter $x = x(\sigma)$ so that $\omega_{\alpha\beta}(x(\sigma)) \overset{\sigma}{\equiv} 0$. We shall illustrate the method with an obvious example. Consider the system

$$\dot{x}_1 = u_1\,, \qquad \dot{x}_2 = u_2\,, \qquad \dot{x}_3 = \frac{1}{x_1^2 + x_2^2 + x_3^2}\,; \tag{4.7}$$

the problem is to transfer the state from [0, 0, 1] to [0, 0, 2] in minimum time. It is obvious that the system (1.5) of partial differential equations is a complete system of order 2 thus satisfying Condition A. The line integral (4.3) is

$$I = \int_{[0,0,1]}^{[0,0,2]} (x_1^2 + x_2^2 + x_3^2)\, dx_3\,, \tag{4.8}$$

so that $\omega_{12} = 0$, $\omega_{13} = 2x_1$, $\omega_{23} = 2x_2$. The singular hypersurfaces are given by the planes $x_1 = 0$, $x_2 = 0$; and the singular arc can be parameterized by $x_1 = 0$, $x_2 = 0$, $x_3 = \sigma$. Let the representation of the surface S (see Fig. 1) containing the totally singular arc, in terms of the two parameters z_1 and z_2, be

$$\begin{aligned} x_1 &= z_1 \cos\phi, \\ x_2 &= z_1 \sin\phi, \qquad z_1 \geqq 0, \\ x_3 &= z_2\,, \end{aligned} \tag{4.9}$$

so that $z_1 = 0$, $z_2 = \sigma$ are the values of the parameters yielding the singular arc.

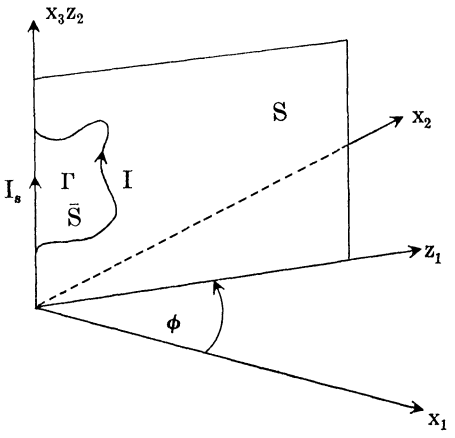

FIG. 1

Denoting by I_s the cost along the totally singular arc, and by I the cost along any other arc contained in S, then

$$I - I_s = \int_\Gamma (x_1^2 + x_2^2 + x_3^2)\, dx_3 = \int_\Gamma (z_1^2 + z_2^2)\, dz_2 . \tag{4.10}$$

Hence applying the two-dimensional Green's theorem yields

$$I - I_s = \int_S 2z_1\, dS \geqq 0,$$

so that the totally singular arc is optimum relative to the comparison trajectories contained in the family of surfaces given by (4.9).

A more general family of surfaces, similar to (4.9), can be described in terms of a vector valued parameter ϕ by

$$\begin{aligned} x_1 &= z_1 f_1(z_2\,;\phi), \\ x_2 &= z_1 f_2(z_2\,;\phi), \\ x_3 &= z_2\,, \end{aligned} \tag{4.11}$$

where f_1 and f_2 are scalar functions of the vector ϕ. Hence, for $z_1 \geqq 0$ we have

$$\begin{aligned} I - I_s &= \int_\Gamma \{z_1^2(f_1^2(z_2\,;\phi) + f_2^2(z_2\,;\phi)) + z_2^2\}\, dz_2 \\ &= \int_S 2z_1(f_1^2(z_2\,;\phi) + f_2^2(z_2\,;\phi))\, dS \geqq 0; \end{aligned}$$

and for $z_1 \leqq 0$,

$$I - I_s = \int_\Gamma \{z_1^2(f_1^2(z_2\,;\phi) + f_2^2(z_2\,;\phi)) + z_2^2\}\, dz_2$$

$$= -\int_{\bar{S}} 2z_1(f_1^2(z_2\,;\phi) + f_2^2(z_2\,;\phi))\, dS \geqq 0,$$

so that once again the totally singular arc is optimum relative to the comparison trajectories contained in the family of surfaces given by (4.11). The existence and construction of a family of surfaces containing all possible comparison trajectories will be left as an open question.

5. The relation between the hypersurfaces $\omega_{\alpha\beta} = 0$ and the totally singular problem. Treating the time optimal problem (1.1) by the conventional methods of optimization [11], the Hamiltonian is

$$H(x, p, u) = 1 + p_\alpha[A_\alpha(x) + B_{\alpha r}(x)u_r], \tag{5.1}$$

where p is the co-state and is determined by the Euler-Lagrange equations:

$$\dot{p}_\alpha = -\frac{\partial H(x, p, u)}{\partial x_\alpha} = -p_\gamma\left[\frac{\partial A_\gamma(x)}{\partial x_\alpha} + \frac{\partial B_{\gamma r}(x)u_r}{\partial x_\alpha}\right]. \tag{5.2}$$

The singular problem giving rise to the totally singular control $u_r^s(t)$ occurs when

$$p_\alpha(t)B_{\alpha r}(\varphi^s(t)) = 0 \tag{5.3}$$

is satisfied together with

$$\frac{d\varphi_\alpha^s(t)}{dt} = A_\alpha(\varphi^s(t)) + B_{\alpha r}(\varphi^s(t))\, u_r^s(t), \tag{5.4}$$

$$\frac{dp_\alpha(t)}{dt} = -p_\gamma(t)\left[\frac{\partial A_\gamma(\varphi^s(t))}{\partial x_\alpha} + \frac{B_{\gamma r}(\varphi^s(t))}{\partial x_\alpha}\, u_r^s(t)\right]. \tag{5.5}$$

Furthermore, the Hamiltonian is a constant along the extremals, and this constant is zero by virtue of the transversality condition to yield

$$1 + p_\alpha(t)A_\alpha(\varphi^s(t)) = 0. \tag{5.6}$$

Differentiating (5.3) with respect to time and using (5.4) and (5.5) to simplify we obtain the following set of $(n - 1)$ equations, which also has to be satisfied:

$$p_\alpha(t)\left[\frac{\partial A_\alpha(\varphi^s(t))}{\partial x_\gamma} B_{\gamma r}(\varphi^s(t)) - \frac{\partial B_{\alpha r}(\varphi^s(t))}{\partial x_\gamma} A_\gamma(\varphi^s(t))\right] = 0. \tag{5.7}$$

It follows that the coefficients

$$\left[\frac{\partial A_\alpha(\varphi^s(t))}{\partial x_\gamma} B_{\gamma r_1}(\varphi^s(t)) - \frac{\partial B_{\alpha r_1}(\varphi^s(t))}{\partial x_\gamma} A_\gamma(\varphi^s(t))\right]$$

are either zero or some linear combination of $B_{\alpha r_2}(\varphi^s(t))$; otherwise (5.3) and (5.7) would imply a trivial result for $p(t)$. To demonstrate the relation between the hypersurfaces $\omega_{\alpha\beta} = 0$ and the singular problem, we shall for convenience take the pfaffian in the equivalent form (4.2),

$$dt = \frac{\dfrac{\partial V(x)}{\partial x_\alpha} dx_\alpha}{\dfrac{\partial V(x)}{\partial x_\tau} A_\tau(x)}, \tag{5.8}$$

and recall that $V(x)$ satisfies the complete system of partial differential equations

$$\frac{\partial V(x)}{\partial x_\alpha} B_{\alpha r}(x) \overset{x}{\equiv} 0 \tag{5.9}$$

of order $(n - 1)$. From the definition of the hypersurface $\omega_{\alpha\beta} = 0$ we have for the equivalent pfaffian form (5.8),

$$\begin{aligned}\omega_{\alpha\beta}(\varphi^s(t)) = {} & \frac{1}{\left\{\dfrac{\partial V(\varphi^s(t))}{\partial x_\tau} A_\tau(\varphi^s(t))\right\}^2} \\ & \cdot\left\{\frac{\partial V(\varphi^s(t))}{\partial x_\beta}\left[\frac{\partial^2 V(\varphi^s(t))}{\partial x_\tau\, \partial x_\alpha} A_\tau(\varphi^s(t)) + \frac{\partial V(\varphi^s(t))}{\partial x_\tau}\frac{\partial A_\tau(\varphi^s(t))}{\partial x_\alpha}\right]\right. \\ & \left. - \frac{\partial V(\varphi^s(t))}{\partial x_\alpha}\left[\frac{\partial^2 V(\varphi^s(t))}{\partial x_\tau\, \partial x_\beta} A_\tau(\varphi^s(t)) + \frac{\partial V(\varphi^s(t))}{\partial x_\tau}\frac{\partial A_\tau(\varphi^s(t))}{\partial x_\beta}\right]\right\}.\end{aligned} \tag{5.10}$$

The factor $1/\{[\partial V(\varphi^s(t))/\partial x_\tau]A_\tau(\varphi^s(t))\}^2$ may be neglected since by assumption

$$\frac{\partial V(x)}{\partial x_\tau} A_\tau(x) \neq 0.$$

From (5.9) we have

$$\frac{\partial V(\varphi^s(t))}{\partial x_\alpha} B_{\alpha r}(\varphi^s(t)) = 0, \tag{5.11}$$

which can be identified with (5.3) by defining

$$p_\alpha(t) = \lambda(t) \frac{\partial V(\varphi^s(t))}{\partial x_\alpha}, \tag{5.12}$$

where $\lambda(t)$ is a nonzero multiplier that has to be determined. Substituting (5.12) into the Euler-Lagrange equations (5.5) yields

$$\frac{d\lambda(t)}{dt}\frac{\partial V(\varphi^s(t))}{\partial x_\alpha} + \lambda(t)\left[\frac{\partial^2 V(\varphi^s(t))}{\partial x_\alpha\,\partial x_\gamma} A_\gamma(\varphi^s(t)) + \frac{\partial V(\varphi^s(t))}{\partial x_\gamma}\frac{\partial A_\gamma(\varphi^s(t))}{\partial x_\alpha}\right] = 0. \tag{5.13}$$

Using this result we find from (5.10) that

$$\omega_{\alpha\beta}(\varphi^s(t)) = \frac{1}{\left\{\dfrac{\partial V(\varphi^s(t))}{\partial x_\tau} A_\tau(\varphi^s(t))\right\}^2} \cdot\left[-\frac{\partial V(\varphi^s(t))}{\partial x_\beta}\frac{\partial V(\varphi^s(t))}{\partial x_\alpha}\frac{d\lambda(t)}{dt}\frac{1}{\lambda(t)} + \frac{\partial V(\varphi^s(t))}{\partial x_\alpha}\frac{\partial V(\varphi^s(t))}{\partial x_\beta}\frac{d\lambda(t)}{dt}\frac{1}{\lambda(t)}\right] = 0.$$

This shows that $\varphi^s(t)$ totally singular implies $\omega_{\alpha\beta}(\varphi^s(t)) = 0$.

Some further consequences of this relationship are the derivation of (5.7) and the first integral (5.6) as follows. Multiplying (5.13) by $B_{\alpha r}$ and summing and invoking (5.11) yields

$$\lambda(t)\left[\frac{\partial^2 V(\varphi^s(t))}{\partial x_\alpha\,\partial x_\gamma} A_\gamma(\varphi^s(t)) B_{\alpha r}(\varphi^s(t)) + \frac{\partial V(\varphi^s(t))}{\partial x_\gamma}\frac{\partial A_\gamma(\varphi^s(t))}{\partial x_\alpha} B_{\alpha r}(\varphi^s(t))\right] = 0. \tag{5.14}$$

Differentiating (5.9), which is an identity in x, with respect to x_γ and multiplying by $A_\gamma(x)$ and summing gives

$$\frac{\partial^2 V(x)}{\partial x_\alpha\,\partial x_\gamma} B_{\alpha r}(x) A_\gamma(x) + \frac{\partial V(x)}{\partial x_\alpha}\frac{\partial B_{\alpha r}(x)}{\partial x_\gamma} A_\gamma(x) \overset{x}{\equiv} 0. \tag{5.15}$$

By virtue of this result, (5.14) becomes

$$\lambda(t)\frac{\partial V(\varphi^s(t))}{\partial x_\gamma}\left[\frac{\partial A_\gamma(\varphi^s(t))}{\partial x_\alpha} B_{\alpha r}(\varphi^s(t)) - \frac{\partial B_{\gamma r}(\varphi^s(t))}{\partial x_\alpha} A_\alpha(\varphi^s(t))\right] = 0, \tag{5.16}$$

which is easily recognized as (5.7).

Finally, to complete the equivalence, if we multiply (5.13) by $A_\alpha(\varphi^s(t))$

and sum, we obtain

$$\frac{d\lambda(t)}{dt}\frac{\partial V(\varphi^s(t))}{\partial x_\alpha}A_\alpha(\varphi^s(t)) + \lambda(t)A_\alpha(\varphi^s(t))\frac{\partial}{\partial x_\alpha}\left[A_\gamma(\varphi^s(t))\frac{\partial V(\varphi^s(t))}{\partial x_\gamma}\right] = 0.$$

Using (5.4) the above equation becomes

$$\frac{d}{dt}\left\{\lambda(t)\frac{\partial V(\varphi^s(t))}{\partial x_\alpha}A_\alpha(\varphi^s(t))\right\}$$

$$- \lambda(t)B_{\alpha r}(\varphi^s(t))u_r^s(t)\frac{\partial}{\partial x_\alpha}\left[A_\gamma(\varphi^s(t))\frac{\partial V(\varphi^s(t))}{\partial x_\gamma}\right] = 0$$

However, the second term is zero by (5.14) so that the above equation can be integrated directly to yield

$$\lambda(t)\frac{\partial V(\varphi^s(t))}{\partial x_\alpha}A_\alpha(\varphi^s(t)) = \text{const.}$$

This result is equivalent to (5.6), the constancy of the Hamiltonian, and determines the multiplier $\lambda(t)$.

6. Minimization of a functional. The n-dimensional Green's theorem approach described in the previous sections can be applied to minimizing functionals of the form

$$I = \int_{t_0}^{t_f} L(x(t))\,dt. \tag{6.1}$$

The problem is to determine a control $u_r(t)$ which by (1.1) transfers the state from x_0 to x_f with no restrictions on t_f (t_f free), such that I is minimized.

Since there is no precise statement about the reachable set for the system (1.1) given, some restrictions must be placed on $L(x)$. This is necessary because it could transpire that if for some $u_r(t)$ the solution to (1.1) formed a closed curve in a region of state space where $L(x)$ is negative, then I could assume any negative value whatsoever simply by traversing the closed curve an arbitrary number of times. The existence theorem of Markus and Lee [12] circumvents this problem by placing a restriction on t_f. However, we cannot include such a restriction without destroying the equivalence between the hypersurfaces $\omega_{\alpha\beta}(x) = 0$ and the singular problem. We shall assume that $L(x) > 0$ in D so that the problem becomes equivalent to one of minimum time.

By use of the pfaffian (4.1), (6.1) can be expressed as

$$I = \int_{x_0}^{x_f}\frac{L(x)\psi_\tau(x)\,dx_\tau}{\psi_\alpha(x)A_\alpha(x)}. \tag{6.2}$$

For this form of the line integral the singular hypersurfaces are given by

$$\omega_{\alpha\beta}(x) = \frac{\partial}{\partial x_\beta}\left\{\frac{L(x)\psi_\alpha(x)}{\psi_\tau(x)A_\tau(x)}\right\} - \frac{\partial}{\partial x_\alpha}\left\{\frac{L(x)\psi_\beta(x)}{\psi_\tau(x)A_\tau(x)}\right\} = 0;$$

and the arguments given in §4 regarding the number of hypersurfaces still apply, since the pfaffian

$$\frac{L(x)\psi_\alpha(x)}{\psi_\tau(x)A_\tau(x)}\,dx_\alpha = 0$$

is integrable.

Similarly, the equivalence between the hypersurfaces $\omega_{\alpha\beta}(x) = 0$ and the singular problem follows from §5 with minor modifications. The totally singular arc $\varphi^s(t)$ with the totally singular control $u_r^s(t)$ satisfies

$$\frac{d\varphi_\alpha^s(t)}{dt} = A_\alpha(\varphi^s(t)) + B_{\alpha r}(\varphi^s(t))u_r^s(t),$$

$$\frac{dp_\alpha(t)}{dt} = -\frac{\partial L(\varphi^s(t))}{\partial x_\alpha} - p_\gamma(t)\left[\frac{\partial A_\gamma(\varphi^s(t))}{\partial x_\alpha} + \frac{\partial B_{\gamma r}(\varphi^s(t))u_r^s(t)}{\partial x_\alpha}\right]$$

and

$$p_\alpha(t)B_{\alpha r}(\varphi^s(t)) \overset{t}{\equiv} 0.$$

The Hamiltonian is a constant along the extremals, and the constant is zero by virtue of the transversality condition and the final time t_f being unspecified, so that

$$L(\varphi^s(t)) + p_\alpha(t)A_\alpha(\varphi^s(t)) \overset{t}{\equiv} 0.$$

From these equations it can be shown that

$$\omega_{\alpha\beta}(\varphi^s(t)) \overset{t}{\equiv} 0,$$

and hence the methods described can be used to evaluate the optimality of the totally singular arc.

7. Some examples. In applying the Green's theorem technique to a specific example, it is not necessary to determine beforehand if the system (1.1) is controllable, because if the system (1.1) is not controllable, then the pfaffian (4.1) is integrable, and the integrability conditions are given by

$$\omega_{\alpha\beta}(x) \equiv 0. \tag{7.1}$$

Consider the following system:

$$\begin{aligned} \dot{x}_1 &= x_1 + x_2u_1, \\ \dot{x}_2 &= x_2 - x_1u_1 + x_3u_2, \\ \dot{x}_3 &= x_3 - x_2u_2. \end{aligned} \tag{7.2}$$

The system of partial differential equations (1.5) associated with (7.2), namely,

$$x_2 \frac{\partial V}{\partial x_1} - x_1 \frac{\partial V}{\partial x_2} = 0,$$

$$x_3 \frac{\partial V}{\partial x_2} - x_2 \frac{\partial V}{\partial x_3} = 0,$$

is a complete system of order 2, thus satisfying Condition A. The corresponding pfaffian (4.1) is

$$dt = \frac{x_1\,dx_1 + x_2\,dx_2 + x_3\,dx_3}{x_1^2 + x_2^2 + x_3^2}, \tag{7.3}$$

and it is immediately obvious that $\omega_{12}(x) \equiv \omega_{13}(x) \equiv \omega_{23}(x) \equiv 0$ so that (7.3) is integrable. Therefore (7.2) is not controllable, since the hypersurface $W(t, x) \equiv (x_1^2 + x_2^2 + x_3^2)e^{-2t} = \text{const.}$ contains all the solutions independent of the controls.

On the other hand, it is most important to check whether Condition A is satisfied before applying the Green's theorem technique. It does not follow that, if the required number of hypersurfaces are obtained, then Condition A is automatically satisfied, as demonstrated by the following counterexample. The system equations are

$$\dot{x}_1 = x_2 + u_1 + x_3 u_2\,,$$

$$\dot{x}_2 = x_3^2 + x_2 u_2\,,$$

$$\dot{x}_3 = -x_2 x_1 + x_1 u_2\,;$$

and the pfaffian (4.1) is

$$dt = \frac{dx_2}{(x_2^2 + x_3^2)} - \frac{x_2\,dx_3}{x_1(x_2^2 + x_3^2)},$$

so that

$$\omega_{12} = 0,$$

$$\omega_{13} = \frac{x_2}{x_1^2(x_2^2 + x_3^2)},$$

$$\omega_{23} = \frac{x_2^2 - x_3^2 + 2x_3 x_1}{x_1(x_2^2 + x_3^2)}.$$

Therefore, it would appear that if $x_1 \neq 0$ and $x_3 \neq 0$, then the singular hypersurfaces are given by

$$x_2 = 0,$$

$$x_3 - 2x_1 = 0,$$

thus yielding the correct number of hypersurfaces despite the fact that Condition A is not satisfied. However, the fallacy of this result is readily apparent, since the hypersurface $x_2 = 0$ implies $x_3 = 0$ thus contradicting the requirement that $x_3 \neq 0$.

REFERENCES

[1] A. MIELE, *Application of Green's theorem to the extremization of linear integrals*, Symposium on Vehicle Systems Optimization, Garden City, New York, 1961, pp. 26–35.

[2] H. HERMES AND G. HAYNES, *On the nonlinear control problem with control appearing linearly*, this Journal, 1 (1963), pp. 85–108.

[3] H. HERMES, *Controllability and the singular problem*, Ibid., 2 (1964), pp. 241–260.

[4] E. KREINDLER, *Contributions to the theory of time optimal control*, J. Franklin Inst., 275 (1963), pp. 314–344.

[5] L. W. NEUSTADT, *Optimization, a moment problem, and nonlinear programming*, this Journal, 2 (1964), pp. 33–53.

[6] R. HERMANN, *On the accessibility problem in control theory*, International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, J. P. LaSalle and S. Lefschetz, ed., Academic Press, New York, 1963.

[7] I. N. SNEDDON, *Elements of Partial Differential Equations*, McGraw-Hill, New York, 1957.

[8] W. RUDIN, *Principles of Mathematical Analysis*, 2nd ed., McGraw-Hill, New York, 1964.

[9] H. FLANDERS, *Differential Forms*, Academic Press, New York, 1963.

[10] H. W. GUGGENHEIMER, *Differential Geometry*, McGraw-Hill, New York, 1963.

[11] R. E. KALMAN, *The theory of optimal control and the calculus of variations*, Mathematical Optimization Techniques, R. Bellman, ed., University of California Press, Berkeley, 1963, Ch. 16.

[12] E. B. LEE AND L. MARKUS, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

EXHAUSTIVE EQUIVALENCE CLASSES OF OPTIMAL SYSTEMS WITH SEPARABLE CONTROLS*

RUEY-WEN LIU† AND R. JEFFREY LEAKE†

Abstract. For a given optimal feedback control problem two systems which have the same optimal feedback control laws and have identical Hamilton-Jacobi equations are said to be equivalent. A necessary and sufficient condition for the equivalence of two systems with separable controls is obtained; with this condition, one can generate exhaustive equivalence classes of optimal systems.

1. Introduction. It is well known that a wide variety of optimal feedback control problems can be reduced to solving a Hamilton-Jacobi partial differential equation. However, very few solutions of optimal control problems are known. Those who doubt this may try to solve the minimum time problem for the following simple linear system:

$$\dot{x} = x + u_1 ,$$

$$\dot{y} = 2y + u_2 ,$$

where $u_1^2 + u_2^2 \leqq 1$. Therefore, any broadening of the class of solvable optimal control problems should be worthwhile.

One way of improving the situation is to make the best use of known results. More specifically, if the solution of a particular Hamilton-Jacobi equation is known, one might attempt to establish a procedure for generating all other optimal control problems having the same Hamilton-Jacobi equation, and thus the same solutions.

Another related approach consists of formulating an inverse problem by choosing a suitable function $V(x, t)$ and directing attention toward a search for all optimal control problems having $V(x, t)$ as a solution of their respective Hamilton-Jacobi equations.

This paper gives positive information to each of the above-mentioned approaches, and the final results are given in §5. Applications and examples are given in §6. In §4 two important lemmas on inner products are presented.

2. Formulation of the optimal feedback control problem [1]. Let G be a region of $R^n \times R^1$, and $S \subset G$ a closed n-dimensional smooth manifold called the *target set*. Consider the dynamical system with separable controls

$$\dot{x} = f(x, t) + g(u), \tag{1}$$

* Received by the editors March 23, 1965, and in revised form April 29, 1966.

† Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana 46556. This work was supported in part by the National Science Foundation, Grant GK-91.

where the n-vector x is the plant state, f and g are continuously differentiable n-vector functions, and

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

is a continuously differentiable m-vector function ($m \leqq n$) from $R^n \times R^1$ to R^m. For each u let $\phi_u(t) = \phi(t; x_0, t_0, u)$ be the unique solution of (1) satisfying the initial condition $\phi_u(t_0) = x_0$ when (2) is substituted into (1); and let $u(t) = k(\phi_u(t), t)$. The function $k(x, t)$ will be called an *admissible feedback control law* if:

(i) $k: R^n \times R^1 \to R^m$ is continuously differentiable;

(ii) its values lie in a set $U \subset R^m$;

(iii) given any $(x_0, t_0) \in G$, when (2) is substituted into (1) the motion $(\phi_u(t), t)$ becomes a member of S for some $t \geqq t_0$; let t_1 be the first such instant. Further, $(\phi_u(t), t) \in G$ for all $t \in [t_0, t_1]$.

The system performance index is

$$J(x_0, t_0; S, u) = \lambda(\phi_u(t_1), t_1) + \int_{t_0}^{t_1} L(\phi_u(t), u(t), t)\, dt, \tag{3}$$

where L and λ are continuously differentiable scalar functions. The *optimal feedback control problem* is to find a particular admissible feedback control law such that the functional (3) assumes its infimum over the class of admissible feedback control laws for every initial phase $(x_0, t_0) \in G$.

3. The Hamilton-Jacobi equation. Subject to certain smoothness conditions, it is possible to solve optimal feedback control problems by finding an appropriate solution of the related Hamilton-Jacobi equation. Following Kalman [1], a formulation of this approach is summarized below. In addition, it is shown that although the solution of the Hamilton-Jacobi equation may not be unique, there is at most one solution which leads to an admissible feedback control law. As such, sufficiency conditions can be stated in a manner which does not require uniqueness of the solution of the Hamilton-Jacobi equation itself.

Define the scalar function H by

$$H(x, p, t, u) = L(x, u, t) + \langle f(x, t), p\rangle + \langle g(u), p\rangle. \tag{4}$$

Assume that for every x, p and t, H has a unique absolute minimum with respect to $u \in U$ and let the associated value of u be denoted as $u^0 = c(x, p, t)$. Assume further that the function c is unique and continuously differentiable in each argument. Define the Hamiltonian H^0 as

$$\begin{aligned} H^0(x, p, t) &= \min_{u \in U} H(x, p, t, u) \\ &= L(x, c(x, p, t), t) + f\langle (x, t), p\rangle + \langle g(c(x, p, t)), p\rangle. \end{aligned} \tag{5}$$

Let there exist a function $V(x, t) = \lambda(x, t)$ on S, which in addition satisfies the Hamilton-Jacobi equation

$$(6) \quad \langle f(x, t), V_x \rangle = V_t - L(x, c(x, V_x, t), t) + \langle g(c(x, V_x, t)), V_x \rangle,$$

or

$$(7) \quad V_t + H^0(x, V_x, t) = 0,$$

for all $(x, t) \in G$. Then if $c(x, V_x, t)$ is an admissible feedback control law (with emphasis on condition (iii)), the lemma of Carathéodory [1] may be applied to $V_t + H^0(x, V_x, t)$, yielding: $V(x_0, t_0) = J(x_0, t_0, S; u^0) \leqq J(x_0, t_0, S; u)$ for all admissible controls u and for all $(x_0, t_0) \in G$. Thus, if $V^*(x, t)$ is any other twice continuously differentiable solution of (7), if $V^*(x, t) = \lambda(x, t)$ on S and $c(x, V_x^*, t)$ is admissible, repeated application of Carathéodory's lemma shows that $V^*(x, t) \leqq V(x, t)$ and thus $V^*(x, t) = V(x, t)$ on G.

To summarize, for a given optimal feedback control problem, if one can find a function $V(x, t)$ which satisfies the stated boundary and smoothness conditions, which is a solution of (7), and for which $c(x, V_x, t)$ is admissible, then he has at once a solution to the optimal feedback control problem.

4. Two lemmas on inner products. It is noted that the inner products of vector functions are essential elements of the Hamilton-Jacobi equation (6). Two lemmas on inner products given here will be employed in the next section. Although the results presented seem to be classical in nature, the authors have been able to find neither these lemmas nor any of their direct applications to engineering problems in the literature.

LEMMA 1. *Let y and z be any two real r-vectors and assume $y \neq 0$. Then z satisfies the condition $\langle z, y \rangle = 0$ if and only if there exists a real, skew-symmetric, $r \times r$ matrix A, such that $z = Ay$.*

Proof. Sufficiency is trivial. To show necessity, let B be a real $r \times r$ orthogonal matrix and $\hat{y}$ be a column vector such that $\hat{y} = By = \text{col } (\| y \|, 0, \cdots, 0)$. Let $\hat{z} = Bz = \text{col } (\hat{z}_1, \cdots, \hat{z}_r)$, where $\hat{z}_i$ is the ith component of $\hat{z}$. Since B is orthogonal we have $\langle z, y \rangle = \langle \hat{z}, \hat{y} \rangle = 0$, which implies that $\hat{z}_1 = 0$. Now let

$$Q = \begin{bmatrix} 0 & -\hat{z}_2 & -\hat{z}_3 & \cdots & -\hat{z}_r \\ \hat{z}_2 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \hat{z}_r & 0 & 0 & \cdots & 0 \end{bmatrix},$$

and define $A = B^T Q B / \| y \|$. Since $Q + Q^T = 0$, we have $A + A^T = B^T(Q + Q^T)B/\| y \| = 0$. Furthermore, $\| y \| Ay = B^T Q B y = B^T Q \hat{y} = B^T \| y \| \hat{z} = \| y \| B^T B z = \| y \| z$. Thus, A is a skew-symmetric $r \times r$ matrix such that $Ay = z$.

LEMMA 2. *Let y and z be any two real r-vectors and assume $y \neq 0$. If α is any scalar, then $\langle z, y\rangle = \alpha$ if and only if there exists a real, skew-symmetric, $r \times r$ matrix A such that $z = [\beta I + A]y$, where $\beta = \alpha/\| y \|^2$ and I is the unit matrix.*

Proof. As in Lemma 1, the sufficiency condition follows by direct evaluation. To show the condition is necessary, let z be decomposed as $z = z^a + z^b$, where $\langle z^a, y\rangle = 0$ and $z^b = \gamma y$, with γ being some scalar. Since $\langle z, y\rangle = \alpha = \langle z^b, y\rangle = \gamma \| y \|^2$, it follows that $\gamma = \alpha/\| y \|^2 = \beta$. Furthermore, Lemma 1 implies $z^a = Ay$, where A is a skew-symmetric $r \times r$ matrix, so $z = [\beta I + A]y$.

The lemmas give explicit (non-unique) solutions of the implicit algebraic equation $\langle z, y\rangle = \alpha$ for $y \neq 0$. When $y = 0$, α must be zero for a solution to exist and z is arbitrary in this case. Of course, A is not unique.

5. Exhaustive equivalence classes. Let P be the set (g, U, J, S, G) where all quantities are defined in §2. Note that for any given pair (f, P) a specific optimal feedback control problem is defined. Assume henceforth that all of the stated continuity and differentiability requirements are satisfied. Because of the separable control, observe that the expression of $c(x, p, t)$ and thus that of the right-hand side of the Hamilton-Jacobi equation (6) is completely determined by a given P. Further, for any P, in order that a scalar function $W(x, t)$ is a solution of (6) and satisfies the boundary condition, it is necessary that $W(x, t) = \lambda(x, t)$ on S, and $W_t(x, t) = L(x, c(x, 0, t), t)$ on the set $\{(x, t) \mid (x, t) \in G, W_x(x, t) = 0\}$. We will say that a scalar function $W(x, t)$ is *compatible* with P if it satisfies the above two conditions. If W is compatible with P, $f(x, t)$ is said to be of class $\mathcal{C}(W, P)$ if by substituting W for V, (6) is satisfied for all $(x, t) \in G$ and $c(x, W_x, t)$ is an admissible control law for system (1). Consequently, $W(x, t)$ is the solution of (6) for all $f \in \mathcal{C}(W, P)$ and $c(x, W_x, t)$ is the admissible control, thus, the optimal feedback control law for the entire class $\mathcal{C}(W, P)$. For the sake of notational convenience, henceforth we will use V instead of W.

DEFINITION 1. Given P and a compatible function $V(x, t)$, $f(x, t)$ and $f^*(x, t)$ are said to be *equivalent* if they belong to the same class $\mathcal{C}(V, P)$.

The following theorem gives an explicit expression for all $f \in \mathcal{C}(V, P)$. In this sense, the formulation of the equivalence class is exhaustive. Let D be the subset of G on which $V_x(x, t) \neq 0$.

THEOREM 1. *Given P and a compatible function $V(x, t)$, then $f \in \mathcal{C}(V, P)$ if and only if $c(x, V_x, t)$ is an admissible feedback control law for f, and there exists a skew-symmetric $n \times n$ matrix $A(x, t)$ such that*

$$f(x, t) = [\eta(x, t)I + A(x, t)]V_x(x, t), \qquad (x, t) \in D, \tag{8}$$

where

$$\eta(x, t) = \frac{-[V_t(x, t) + L(x, c(x, V_x\, t), t) + \langle g(c(x, V_x, t)), V_x(x, t)\rangle]}{\| V_x(x, t) \|^2}.$$

Proof. The assertion follows directly from (6) and Lemma 2.

THEOREM 2. *Given P and a compatible function $V(x, t)$, let $f^* \in \mathcal{C}(V, P)$. Then f and f^* are equivalent if and only if $c(x, V_x, t)$ is an admissible control law for f, and there exists a skew-symmetric $n \times n$ matrix $A(x, t)$ such that*

$$f(x, t) = f^*(x, t) + A(x, t)V_x(x, t), \qquad (x, t) \in D. \tag{9}$$

Proof. If f and f^* are equivalent, Theorem 1 implies that $f^* = [\eta I + B^*]V_x$, $f = [\eta I + B]V_x$ on D, where B and B^* are skew-symmetric. Equation (9) follows by subtracting f^* from f and letting $A = B - B^*$. Sufficiency follows by taking the inner product of V_x with (9), applying Lemma 1, and noting that $V_x = 0$ for $(x, t) \notin D$. Thus, one obtains $\langle f, V_x\rangle = \langle f^*, V_x\rangle$ for $(x, t) \in G$. Therefore, f and f^* yield the same Hamilton-Jacobi equation (6). Since $c(x, V_x, t)$ is admissible for f, $f \in \mathcal{C}(V, P)$.

Consequently, for any P and compatible V, one may formulate the inverse problem by specifying the form of all possible system functions $f \in \mathcal{C}(V, P)$. Also, if a particular problem is solved, the equivalence class may be generated using the algorithm (9).

In order that $c(x, V_x, t)$ be admissible for the class $\mathcal{C}(V, P)$, one has to show that with the control $u = c$ all motions $(\phi_u(t), t)$ initiated in G will reach S for every system with $f \in \mathcal{C}(V, P)$. The following result can be easily proved by use of (6) and is useful in this respect.

PROPERTY 1. *Given P and a compatible $V(x, t)$, let $u = c(x, V_x, t)$. Then the Eulerian derivative of V is given by $\dot{V} = -L(x, c(x, V_x, t), t)$ for all f defined by* (8) *or* (9).

Thus if one can conclude from the pair V and $\dot{V}$ that all motions initiated from G will reach S in a finite time (see Example 6.2), then this is true for *all* f defined by (8) or (9).

6. Examples.

6.1. Linear regulator problem [1], [3]. The purpose of this example is to demonstrate an extension of the well-known linear regulator theory to a broader class of (possibly nonlinear) systems. Consider the system

$$\dot{x} = Fx + Gu, \tag{10}$$

where F is an $n \times n$ matrix, G is an $n \times n$ matrix, and the performance index is specified[1] by $L = \| Hx \|_Q^2 + \| u \|_R^2$, $\lambda = \| x \|_B^2$, where Q, R are positive

[1] $\| x \|_A^2 \triangleq \langle x, Ax\rangle$.

definite symmetric, B is nonnegative definite symmetric. Let $S = R^n \times \{t_1\}$, $G = \{(x, t) | \, t \leqq t_1\}$; then it is known that the unique solution of Hamilton-Jacobi equation is $V(x, t) = \frac{1}{2} \| x \|^2_{P(t)}$ (where $P(t)$ satisfies a Ricatti equation), $P(t_1) = B$, and $c(x, V_x, t) = \frac{1}{2} R^{-1} G^t P(t) x$. By Theorem 2, all systems $\dot{x} = f(x, t) + Gu$ with

$$f(x, t) = Fx + A(x, t)P(t)x, \qquad A + A^T = 0, \tag{11}$$

are equivalent to (10), provided that A is sufficiently well-behaved to ensure that the smoothness requirements on f are satisfied and a finite escape time condition is not present in the interval $t_0 < t \leqq t_1$.

6.2. Norm invariant systems [4]. In this example it is shown how properties of norm invariant systems may be derived from the study of a trivial problem. Consider the simple system $\dot{x} = u$, $\| u \| \leqq 1$, and the associated optimization problem with $G = \{(x, t) | \, \| x \| \geqq a\}$, $S = \{(x, t) | \, \| x \| = a\}$, $L = 1$, $\lambda = 0$, where a is a "small" positive scalar. The Hamilton-Jacobi equation for this problem is

$$V_t - \| V_x \| + 1 = 0, \qquad V(x, t) = 0 \quad \text{for} \quad \| x \| = a. \tag{12}$$

Possible solutions are $V(x) = \pm \{\| x \| - a\}$. From § 2, there is only one solution which leads to an admissible control law, and this is evidently $V(x) = \| x \| - a$, with $c(x, V_x, t) = -x/\| x \|$. By Theorem 2, all systems in the associated equivalence class must be of the form $\dot{x} = f(x, t) + u$, where

$$f(x, t) = B(x, t)x, \qquad B + B^T = 0. \tag{13}$$

Since $\dot{V} = -1$ in $G - S$, c is admissible for all such f if B is smooth enough. Lemma 1 further implies that *all* norm invariant systems must be expressible in the form $\dot{x} = B(x, t)x$, $B + B^T = 0$, a fact which is not in agreement with the corresponding statement of Athans, Falb and Lacoss [4].

So far, we have shown how an equivalence class can be generated from known solutions. Observing that (8) gives an explicit relation between the Hamilton-Jacobi equation solution and system function, one may approach the optimal feedback control problem from another (inverse) viewpoint.

6.3. An inverse asymptotic problem [5]. Starting with a prespecified function $V(x, t)$ it is demonstrated here how one might proceed to find a meaningful class of systems for which V is the solution. Consider the system

$$\dot{x} = f(x) + bu,$$

where b is an n-vector and u a scalar. Let

$$J = \int_0^\infty (L(x) + u^2) \, dt.$$

If the problem is meaningful, it is easily shown formally that

$$u^0 = c(x, V_x, t) = -\tfrac{1}{2}\langle b, V_x\rangle \quad \text{and} \quad V_t = 0.$$

Therefore, the Hamilton-Jacobi equation has the form

$$\langle f, V_x\rangle = \tfrac{1}{4}\langle b, V_x\rangle^2 - L(x).$$

Consequently, any given $V(x)$ will satisfy this equation if

$$f(x) = [\eta(x)I + A(x)]V_x(x), \tag{14}$$

where

$$\eta(x) = [\tfrac{1}{4}\langle b, V_x\rangle^2 - L(x)]/\| V_x \|^2 \quad \text{and} \quad A + A^T = 0.$$

In fact, f may be a function of both x and t since the exhaustive equivalence class will be of the form

$$f(x, t) = [\eta(x)I + A(x, t)]V_x(x).$$

If $L(x) = \frac{1}{4}\langle b, V_x\rangle^2$, then (14) reduces to $f(x) = A(x)V_x$, or by Lemma 1, $\langle f(x), V_x(x)\rangle = 0$. The insight gained here leads to the following example.

6.4. Systems with a known integral. Let an integral $\phi(x)$ of the system

$$\dot{x} = f(x, t) \tag{15}$$

be known, i.e., $\langle f(x, t), \phi_x(x)\rangle = 0$. Consequently, all solutions are on a constant $\phi(x)$ surface. For example, if (15) is a conservative system, then $\phi(x)$ may be the total system energy; or if (15) is norm invariant $\phi(x) = \| x \|$ is an integral.

Now consider the optimal control problem

$$\dot{x} = f(x, t) + bu, \qquad J = \int_0^\infty \left[\frac{1}{4}\langle b, \phi_x\rangle^2 + u^2\right] dt.$$

Then by simple verification, a solution of the Hamilton-Jacobi equation is $V(x) = \phi(x)$ and the corresponding control law is

$$k(x, t) = -\tfrac{1}{2}\langle b, \phi_x\rangle. \tag{16}$$

Thus, we use the information from the inverse problem of §6.3 to obtain a formal solution of the Hamilton-Jacobi equation of §6.4. In order that this solution give the optimal feedback control law, other problems have to be studied, such as the conditions on P and the admissibility of the control law. Furthermore, questions of controllability and convergence would have to be investigated on a separate basis due to the asymptotic nature of the problem [5]. We will not attempt to solve these problems here.

Finally, we note that one of the major restrictions of these results is the

differentiability requirement of $V(x, t)$, the solution of the Hamilton-Jacobi equation, because it excludes a class of interesting problems which require discontinuous controls. However, Hermes [6] has shown that a class of optimal feedback control problems with discontinuous optimal controls can be approximated arbitrarily closely by other problems with continuous optimal controls. Therefore, the results obtained in this paper may be applied to the latter approximate systems.

Acknowledgment. The authors wish to express their thanks for the constructive suggestions from the referee.

REFERENCES

[1] R. E. KALMAN, *The theory of optimal control and the calculus of variations*, Mathematical Optimization Techniques, University of California Press, Berkeley, 1963, Chap. 16.

[2] T. F. BRIDGLAND, JR., *On the existence of optimal feedback controls*, this Journal, 1 (1963), pp. 261–274.

[3] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[4] M. ATHANS, P. L. FALB AND R. T. LACOSS, *Time, fuel, and energy optimal control of nonlinear norm-invariant systems*, IEEE Trans. Automatic Control, AC-8 (1963), pp. 196–202.

[5] R. BELLMAN AND R. BUCY, *Asymptotic control theory*, this Journal, 2 (1964), pp. 11–18.

[6] H. HERMES, *The equivalence and approximation of optimal control problems*, J. Differential Equations, 1 (1965), pp. 409–426.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

SOME REMARKS ON COMPLETE CONTROLLABILITY*

H. O. FATTORINI†

1. Introduction. The aim of this paper is to generalize a well-known criterion for complete controllability of the finite-dimensional linear control system $u' = Au + Bf$ (as in [1, p. 201]) to a class of control systems in Banach space that includes, among others, the case where A is an elliptic partial differential operator in a bounded domain of Euclidean space.

A similar generalization was considered in [6] for the case where A is a self-adjoint operator in a separable Hilbert space.

2. Notations and preliminary results. Throughout this paper E, F will be complex Banach spaces; the norm in any of them will be written $|\cdot|$. If E is any Banach space and $u^* \in E^*$, the dual space of E, we shall denote by (u^*, u) or (u, u^*) the value of the functional u^* at the point $u \in E$.[1] If $N \subseteq E$, we define $N^\perp = \{u^* \in E^* \mid (u^*, u) = 0 \text{ for all } u \in N\}$. It is easy to see that $N^\perp$ is a closed subspace of E^*. If N is a subspace of E then, by the Hahn-Banach theorem N is dense in E if and only if $N^\perp = \{0\}$.

We shall consider the linear control system

$$u'(t) = Au(t) + Bf(t), \tag{2.1}$$

and various other systems derived from it. The E-valued function $u(\cdot)$ is the *output* of the system; the F-valued function $f(\cdot)$, the *input* or *control*, determines in some sense its behavior. We shall assume the closed linear operator A with dense domain $D(A) \subseteq E$ to be the infinitesimal generator of a strongly continuous semigroup $T(t)$, $t \geqq 0$, in the space E [2, Chap. VIII]; B will be a linear bounded operator from F to E.[2] We shall always denote by L the control system (2.1).

If f is sufficiently smooth in $[0, \infty)$, for instance continuously (strongly) differentiable,[3] then (2.1) has a solution. That is, there exists an E-valued function $u(t)$, $t \geqq 0$, strongly continuously differentiable and such that

* Received by the editors December 3, 1965, and in revised form April 4, 1966.

† Consejo Nacional de Investigaciones Científicas y Técnicas and Departamento de Matematicas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina.

This work was supported in part by a Ford Foundation Pre-Doctoral Fellowship at the Courant Institute of Mathematical Sciences, New York University.

[1] Adjoints of linear operators will be defined with respect to the bilinear form $(\cdot, \cdot)$, except in the examples after Proposition 2.3 and Corollary 3.2, where E is a Hilbert space and we use instead the corresponding scalar product.

[2] See [6], [7] for further details.

[3] We can weaken somewhat the requirements on f; however, for the purposes of this paper we can consider f as smooth as we please.

$u(t) \in D(A)$ and (2.1) is satisfied for all $t \geqq 0$. Moreover, we can assume $u(0) = u$, where u is any element in $D(A)$. With this initial condition the solution is unique and given by

$$u(t) = T(t)u + \int_0^t T(t - s)Bf(s)\, ds, \tag{2.2}$$

[4, especially pp. 215–218].

As in [6], we shall say that the control system L is *completely controllable*[4] for f in a given linear class $\mathfrak{L}$ of controls defined in $[0, \infty)$ if, given $u \in E$ and an arbitrarily small number $\epsilon > 0$, there exists $f \in \mathfrak{L}$ such that the solution of (2.1) with initial condition $u(0) = 0$ satisfies

$$| u(t_0) - u | \leqq \epsilon,$$

for some $t_0 \geqq 0$, depending in general on u, ϵ. If t_0 can be chosen independently of u, ϵ, we shall say that L is *completely controllable in time* t_0.[5]

Let us denote by K_{t_0} the subspace of E consisting of the values (for $t = t_0$) of all solutions of (2.1) with $u(0) = 0$ and $f \in \mathfrak{L}$. Then it is clear that L will be completely controllable if and only if $\mathrm{Cl}\, K = E$, where $K = \bigcup_{t>0} K_t$, it will be completely controllable in time t_0 if and only if $\mathrm{Cl}\, K_{t_0} = E$. Sometimes we shall write $K_t(L)$ or $K(L)$ to emphasize dependence on the system L. If $0 \leqq t_0 \leqq t_1$ we have $K_{t_0} \subseteq K_{t_1} \subseteq K$. Thus complete controllability in time t_0 implies complete controllability in time t_1, which in turn implies complete controllability. The reverse implications are in general false. (See [6] for a class of systems where all three notions coincide.)

In the remainder of this paper, we shall assume $\mathfrak{L}$ to be the class of all F-valued functions defined for $t \geqq 0$ and continuously strongly differentiable there. (The results do not change if we consider, for instance, the functional classes used in [7], and change conveniently the meaning of "solution".) For the sake of simplicity we shall assume E to be *reflexive*. Then the adjoint semigroup $T^*(\cdot)$ is strongly continuous and has A^* as an infinitesimal generator. If E is not reflexive our results still hold: we only have to replace E^*, A^*, $T^*(\cdot)$ by the Phillips adjoints $E^\odot$, $A^\odot$, $T^\odot(\cdot)$. (See [5, XIV] for details on these adjoints.)

Our first result is a slight modification of Proposition 1 in [6].

PROPOSITION 2.1. *$u^* \in K(L)^\perp (K_t(L)^\perp)$ if and only if $B^*T^*(s)u^* = 0$, $0 \leqq s$, $(0 \leqq s \leqq t)$.*

Proof. It is easy to deduce from (2.2) by means of elementary computations that $u^* \in K(L)^\perp$ if and only if

$$\int_0^t (B^*T^*(s)u^*, f(s))\, ds = 0,$$

[4] Or *null controllable* to emphasize the fact that we start at $u = 0$ for $t = 0$.

[5] See [8], where a problem similar to ours is formulated, as well as other control problems in Banach space.

for all $t \geqq 0, f \in \mathfrak{L}$. If $a(\cdot)$ is any scalar-valued, differentiable function defined for $t \geqq 0$ and u is any element of $F, f(\cdot) = a(\cdot)u \in \mathfrak{L}$. Then

$$\int_0^t (B^*T^*(s)u^*, u)a(s)\, ds = 0.$$

Since $a(\cdot)$ is arbitrary, $(B^*T^*(s)u^*, u) = 0$ for $s \geqq 0$, which implies the desired result. The proof is similar for K_t.

Recall that, since A is the infinitesimal generator of a strongly continuous semigroup $T(\cdot)$, $\sigma(A)$, the spectrum of A, is contained in the halfplane $\operatorname{Re}\lambda \leqq \omega_0$, where

$$\omega_0 = \lim_{t\to\infty} t^{-1} \log |T(t)| < \infty,$$

[2, Chap. VII, §1.11]. Let $\rho(A)$ (the resolvent set of A) be the complement of $\sigma(A)$ and call $\rho_0(A)$ the connected component of $\rho(A)$ that contains the halfplane $\operatorname{Re}\lambda > \omega_0$. We now have the following corollary.

COROLLARY 2.2. *Let $u^* \in K(L)^\perp$. Then*

$$B^*R(\lambda; A^*)^n u^* = B^*(\lambda I - A^*)^{-n}u^* = 0, \quad \lambda \in \rho_0(A), \quad n = 0, 1, \cdots.$$

Conversely, assume $B^*R(\lambda; A^*)u = 0$ for $\lambda \in \rho_0(A)$. Then $u^* \in K(L)^\perp$.

Proof. Assume $u^* \in K(L)^\perp$. Then by Proposition 2.1, $B^*T^*(s)u^* = 0$, $s \geqq 0$. By the formula

$$(2.3) \quad R(\lambda; A^*)^n u^* = \frac{1}{(n-1)!}\int_0^\infty t^{n-1}e^{-\lambda t}T^*(t)u^*\, dt, \qquad \operatorname{Re}\lambda > \omega_0, \quad n = 1, 2, \cdots$$

[2, Chap. VII, §1.12], $B^*R(\lambda; A^*)^n u^* = 0$ for $\operatorname{Re}\lambda > \omega_0$, and thus, by analytic continuation for $\lambda \in \rho_0(A)$. The case $n = 0$ follows from $R(\lambda; A^*)^0 u^* = u^* = T^*(0)u^*$. Conversely, assume $B^*R(\lambda; A^*)u^* = 0$ for $\operatorname{Re}\lambda > \omega_0$. Applying B^* to both sides of (2.3) and applying the functionals thus obtained to any element $u \in E$, we obtain

$$\int_0^\infty e^{-\lambda t}(B^*T^*(t)u^*, u)\, dt = 0,$$

for $\operatorname{Re}\lambda > \omega_0$. By uniqueness of Laplace transforms, $(B^*T^*(t)u^*, u) = 0$ for $t \geqq 0$; since u is arbitrary, $B^*T^*(t)u^* = 0$, $t \geqq 0$. By Proposition 2.1, $u^* \in K(L)^\perp$, which ends the proof.

The following result, which will not be used in the sequel, shows that as far as complete controllability is concerned, we only need to consider the case where A is bounded.

PROPOSITION 2.3. *Let $\lambda_0 \in \rho_0(A)$ and let L_{λ_0} be the linear control system*

$$u'(t) = R(\lambda_0; A)u(t) + Bf(t).$$

Then $\operatorname{Cl} K(L) = \operatorname{Cl} K(L_{\lambda_0})$.

Proof. Clearly we only have to prove that $K(L)^{\perp} = K(L_{\lambda_0})^{\perp}$ in E^*. Assume $u^* \in K(L)^{\perp}$. Then, by Corollary 2.2,

$$B^*R(\lambda_0 ; A^*)^n u^* = B^*R(\lambda_0 ; A)^{*n} u^* = 0, \qquad n = 0, 1, 2, \cdots .$$

This and the series representation for the exponential function imply

$$B^* \exp (sR(\lambda_0 ; A)^*)u^* = 0, \qquad -\infty < s < \infty,$$

thus $u^* \in K(L_{\lambda_0})^{\perp}$. Conversely, assume $B^*\exp\ (sR(\lambda_0;\ A^*))u^* = 0$. Differentiating this last expression repeatedly and setting $s = 0$ we see that

$$B^*R(\lambda_0 ; A^*)^n u^* = 0, \qquad n = 0, 1, \cdots .$$

Now if $|\lambda - \lambda_0| < |R(\lambda_0 ; A^*)|^{-1}$, then $\lambda \in \rho(A)$, a fortiori $\lambda \in \rho_0(A)$, and

$$R(\lambda; A^*) = \sum_{n=0}^{\infty} (\lambda_0 - \lambda)^n R(\lambda_0 ; A^*)^{n+1},$$

[2, Chap. VII, §3], which shows that $B^*R(\lambda; A^*)u^* = 0$ for λ near λ_0 and thus, by analytic continuation for all $\lambda \in \rho_0(A)$. An application of Corollary 2.2 then proves our assertion.

Remark. If we replace $\rho_0(A)$ by $\rho(A)$, then the conclusion of Proposition 2.3 (and thus also that of Corollary 2.2) may become false. For instance, let E be the space $L^2(0, 2\pi)$, the interval $(0, 2\pi)$ endowed with ordinary Lebesgue measure, and let A be the (unitary) operator $Au(x) = e^{-ix}u(x)$. It is easy to see that

$$\begin{aligned} \sigma(A) &= \{\lambda \mid |\lambda| = 1\}, \\ \rho_0(A) &= \{\lambda \mid |\lambda| > 1\}, \\ A^*u(x) &= e^{ix}u(x), \\ R(\lambda; A^*)u(x) &= (\lambda - e^{ix})^{-1}u(x). \end{aligned}$$

Let now B be any bounded operator with range in E such that the nullspace H^2 of B^* is given by

$$H^2 = \left\{ u \in L^2 \mid u \sim \sum_{n=0}^{\infty} a_n e^{inx} \right\}.$$

If $u \in H^2$, then $R(\lambda; A^*)u \in H^2$ for $|\lambda| > 1$. But if $R(\lambda; A^*)u \in H^2$ for $|\lambda| < 1$, $u = 0$.

Remark. The conclusion of Proposition 2.3 cannot be improved, in general, to $K(L) = K(L_{\lambda_0})$.

3. Complete controllability and spectral sets. Recall [2, Chap. VIII, §3.7] that a *spectral set* in $\sigma(A)$ is any subset σ_0 of $\sigma(A)$ which is both open

and closed in $\sigma(A)$. Let σ_0 be a bounded spectral set in $\sigma(A)$ and consider the (bounded) operator

$$P_0 = \frac{1}{2\pi i}\int_C R(\lambda; A)\, d\lambda, \tag{3.1}$$

where C consists of a finite number of smooth curves, boundary of an open set V such that $\sigma_0 \subseteq V$, and $V \cap (\sigma(A) - \sigma_0) = \varnothing$. Then, P_0 is a *projector* [2, Chap. VII, §9], i.e.,

$$P_0P_0 = P_0\,. \tag{3.2}$$

(We shall sometimes write $P_0 = P(\sigma_0\,; A)$ to indicate explicitly the dependence of P on A and σ_0.) The equality (3.2) implies that the subspace

$$E_0 = P_0E = \{u \in E \mid P_0u = u\}$$

is closed. It is easy to see that

$$E_0 \subseteq D(A^m), \qquad m = 1, 2, \cdots,$$

and that $A^mE_0 \subseteq E_0$; moreover the restriction of A^m to E_0 is a bounded operator. Let us call A_0 the restriction of A to E_0. We have $\sigma(A_0) = \sigma_0$ [2, Chap. VII, §3.2]; in particular, if $\lambda \in \rho(A)$, $R(\lambda; A_0)$ is the restriction of $R(\lambda; A)$ to E_0. As for $T(t)$, the semigroup generated by A, we have

$$T(t)u_0 = \exp(tA_0)u_0\,, \qquad u_0 \in E_0\,.$$

Let us observe next that, as $\sigma(A) = \sigma(A^*)$, σ_0 is also a spectral set in $\sigma(A^*)$ and all the preceding considerations apply. We have

$$P(\sigma_0\,; A^*) = P(\sigma_0\,; A)^*.$$

The space E_0^*, dual of E_0, is isometrically isomorphic to the quotient space $E^*/E_0^{\perp}$; let us write $u^* + E_0^{\perp}$ for the equivalence classes in this space. If $u_0 \in E_0$,

$$(P_0^*u^* - u^*, u_0) = (u^* - u^*, u_0) = 0,$$

then

$$u^* + E_0^{\perp} = P_0^*u^* + E_0^{\perp}.$$

If $v^*, w^* \in P_0^*E^*$, $v^* - w^* \in E_0^{\perp}$, then

$$(v^* - w^*, u) = (v^* - w^*, P_0u) = 0$$

for all $u \in E$; and thus $v^* = w^*$. Consequently, each equivalence class in $E^*/E_0^{\perp}$ contains a unique element of $P_0^*E^*$, which allows us to identify, at least algebraically, the dual space E_0^* with $P_0^*E^*$. Although we shall make no use of this fact, let us observe that this identification is also topo-

logic; in fact, the norm in $E^*/E_0^\perp$ is equivalent to the norm that $P_0{}^*E^*$ inherits from E^*.

Finally, let us state two identities concerning adjoints. We have $A_0{}^* = A^* \mid E_0{}^* = (\text{restriction of } A^* \text{ to } E_0{}^*) = P_0{}^*E^*$. If M is any bounded operator with range in E, $(P_0M)^* = M^* \mid E_0{}^* = (\text{restriction of } M^* \text{ to } E_0{}^*)$. Our first result is the following proposition.

PROPOSITION 3.1. *Let $u^* \in E^*$ and let σ_0 be a bounded spectral set in $\sigma(A)$ such that the set of curves C used to define P_0 in (3.1) can be chosen entirely contained in $\rho_0(A)$. Assume $u^* \in K(L)^\perp$. Then $P_0{}^*u^* \in K(L_0)^\perp$, where L_0 is the control system (in E_0),*

$$u'(t) = A_0u(t) + P_0Bf(t).$$

Proof. We shall use the characterization of elements of $K(L)^\perp$, $K(L_0)^\perp$ given by Corollary 2.2. Let μ lie in the halfplane $\operatorname{Re}\mu > \omega_0$ and outside the contours C. Then, by the first resolvent equation and Cauchy's theorem,

$$B^*R(\mu; A^*)P_0{}^*u^* = B^*R(\mu; A^*)\frac{1}{2\pi i}\int_C R(\lambda; A^*)u^*\,d\lambda$$

$$= \frac{1}{2\pi i}\int_C \frac{B^*R(\lambda; A^*)u^*}{\mu - \lambda}\,d\lambda.$$

As $B^*R(\lambda; A^*)u^* = 0$ for $\lambda \in C \subseteq \rho_0(A)$,

$$B^*R(\mu; A^*)P_0{}^*u^* = B^*R(\mu; A_0{}^*)P_0{}^*u^* = 0.$$

By analytic continuation, $B^*R(\mu; A_0{}^*)P_0{}^*u^* = 0$ for all $\mu \in \rho_0(A_0)$.

COROLLARY 3.2. *Let $\{\sigma_n\}$ be a family of bounded, pairwise disjoint spectral sets in $\sigma(A)$. Let P_n be the projector associated with σ_n defined by (3.1). It is assumed that for each σ_n the set C of curves used in (3.1) to define P_n is entirely contained in $\rho_0(A)$. Let $E_n = P_nE$ and A_n be the restriction of A to E_n. Assume that the smallest closed subspace of E containing all the E_n coincides with E. Then the linear control system L,*

$$u'(t) = Au(t) + Bf(t),$$

is completely controllable if and only if each of the linear control systems,

$$u'(t) = A_nu(t) + P_nBf(t),$$

is completely controllable.

Proof. The easily verifiable relation $K(L_n) = P_nK(L)$ shows that if $K(L)$ is dense in E then $K(L_n)$ is dense in E_n. Conversely, assume that each L_n is completely controllable and let $u^* \in K(L)^\perp$. Then, by Proposition 3.1, $P_n{}^*u^* \in K(L_n)^\perp$ for all n. Since each L_n is completely controllable, $P_n{}^*u^* = 0$. But this means that $(u^*, u) = 0$ for all u in the subspace generated by all the E_n, and thus $u^* = 0$.

Remark. If for some spectral set σ_n there does not exist a finite set C of smooth curves which is contained in $\rho_0(A)$ and which is the boundary of an open set V such that $\sigma_n \subseteq V$ and $V \cap (\sigma(A) - \sigma_n) = \varnothing$, then the conclusion of Corollary 3.2, and thus that of Proposition 3.1, may fail to hold. In fact, let

$$E = L^2(0, 2\pi) \oplus L^2(0, 2\pi) = \{[u(x), v(x)], \cdots\}$$

(orthogonal sum), let $0 < r < 1$, and let A be the (normal) operator

$$A[u(x), v(x)] = [e^{-ix}u(x), re^{-ix}v(x)].$$

Plainly $\sigma(A) = C_0 \cup C_1$, where $C_0 = \{\lambda \mid |\lambda| = r\}$ and $C_1 = \{\lambda \mid |\lambda| = 1\}$; both C_0 and C_1 are spectral sets in $\sigma(A)$, but for neither does there exist the finite set of smooth curves entirely contained in $\rho_0(A)$ to be used to define the projections. The projections are $P_0[u, v] = [u, 0]$ and $P_1[u, v] = [0, v]$. Let $N = \{[u, v] \in E$ such that $u \sim \sum_{-\infty}^{\infty} a_n e^{inx}$ and $v \sim \sum_{-\infty}^{\infty} r^n a_n e^{inx}\}$, and let B be any operator with range in E such that the nullspace of B^* is N. Since the nullspace of $(P_iB)^*$ is $N \cap E_i = \{0\}$, $i = 1, 2$, it is easy to deduce from Proposition 2.1 that both control systems $u'(t) = A_iu(t) + P_iBf(t)$ are completely controllable. However, the control system $u'(t) = Au(t) + Bf(t)$ is *not* completely controllable. For if we let $[u, v] \in N$, $[u, v] \neq 0$, we have

$$A^{*m}[u, v] = [e^{imx}u(x), r^m e^{imx}v(x)],$$

$$e^{imx}u(x) \sim \sum_{-\infty}^{\infty} a_{n-m}e^{inx},$$

$$r^m e^{imx}v(x) \sim \sum_{-\infty}^{\infty} r^n a_{n-m}e^{inx},$$

then $A^{*m}[u, v] \in N$ for all $m \geqq 0$. But then $\exp(sA^*)[u, v] \in N$ for all s, which shows, in view of Proposition 2.1, that $u' = Au + Bf$ is not completely controllable.

Let us now consider the important particular case of Corollary 3.2 in which each σ_n reduces to a point λ_n and the corresponding subspaces E_n are finite-dimensional. Choose a basis in $E_n{}^*$ such that the matrix of $A_n{}^*$ has Jordan canonical form, i.e., a basis

$$u^*_{1,1}, \cdots, u^*_{1,n(1)}, u^*_{2,1}, \cdots, u^*_{2,n(2)}, \cdots, u^*_{p,1}, \cdots, u^*_{p,n(p)},$$

such that

$$A_n{}^*u^*_{k,1} = \lambda_n u^*_{k,1}, \qquad 1 \leqq k \leqq p,$$

$$A_n{}^*u^*_{k,j} = \lambda_n u^*_{k,j} + u^*_{k,j-1}, \qquad 2 \leqq j \leqq n(k), \quad 1 \leqq k \leqq p.$$

Assume that $B^*u^*_{1,1}, B^*u^*_{2,1}, \cdots, B^*u^*_{p,1}$ are linearly independent in F^*.

Now let $u^* = \sum_{k=1}^{p} \sum_{j=1}^{n(k)} a_{k,j} u_{k,j}^*$ be any nonzero element of E_n^* such that $B^*T^*(s)u^* = 0$, $s \geqq 0$. By Corollary 2.2, $B^*R(\lambda; A_n^*)u^* = 0$ for $\lambda \in \rho_0(A_n) = \rho(A_n)$; thus $B^*f(A_n^*)u^* = 0$ for any function f defined and analytic in a neighborhood of λ_n.

If $n_0 = \max\{n(1), \cdots, n(p)\}$, $(\lambda_n I - A_n^*)^{n_0} = 0$, then there exists a number m, $1 \leqq m \leqq n_0 - 1$, such that $(\lambda_n I - A_n^*)^m u^* \neq 0$, $(\lambda_n I - A_n^*)^{m+1} u^* = 0$. Let us now take $f(\lambda) = (\lambda_n - \lambda)^m$. Plainly we have

$$(\lambda_n I - A_n^*) f(A_n^*) u^* = 0,$$

thus

$$f(A_n^*)u^* = \sum_{k=1}^{p} b_k u_{k,1}^*, \qquad B^*f(A_n^*)u^* = \sum_{k=1}^{p} b_k B^* u_{k,1}^*.$$

Using the assumption that the $B^*u_{i,1}^*$, $i = 1, \cdots, p$, are linearly independent in F^*, we get $b_1 = b_2 = \cdots = b_k = 0$, then $(\lambda_n I - A_n^*)^m u^* = 0$, which is impossible. Conversely, assume that the vectors $B^*u_{i,1}^*$, $i = 1, \cdots, p$, are not linearly independent in F^*, and let $b_1, b_2, \cdots, b_p$ be complex numbers, not all zero and such that $\sum_{k=1}^{p} b_k B^* u_{k,1}^* = 0$. Then it is easy to see that $B^*T^*(s)u^* = 0$, where $u^* = \sum_{k=1}^{p} b_k u_{k,1}^* \neq 0$. Consequently we have the following corollary.

COROLLARY 3.3. *Let $\lambda_1, \lambda_2, \cdots$ be a sequence of isolated points in $\sigma(A)$ such that the spectral sets $\{\lambda_1\}, \{\lambda_2\}, \cdots$ satisfy the assumptions in Corollary 3.2. Assume moreover that $E_n = P_n E$ is finite-dimensional for all n. Then the linear control system*

$$u'(t) = Au(t) + Bf(t)$$

is completely controllable if and only if B^* is one-to-one in all the subspaces $D_n^* \subseteq E_n^*$ of eigenvectors of A^* corresponding to each eigenvalue λ_n.

Let us say, as in [6], that A is *m-controllable* if there exists an m-dimensional space F and a bounded operator $B: F \to E$ such that $u'(t) = Au(t) + Bf(t)$ is completely controllable. It follows from Corollary 3.3 that A cannot be m-controllable unless $\sup_n \dim D_n^* \leqq m$, i.e., unless the multiplicity of each eigenvalue of A^* (of A) does not exceed m. On the other hand, it is easy to see that if $\sup_n \dim D_n^* \leqq m$ then A is m-controllable.

Let us now apply our results to a concrete example. Let

$$\tau = \sum_{|J| \leqq 2p} a_J(x) \partial^J$$

be a real, negative elliptic formal partial differential operator of order $2p$ defined in a bounded domain I of Euclidean n-space E^n. (For notations and definitions see [3, Chap. XIV, especially §2 and §6.1].) Assume I is smoothly bounded and let V be the operator in $E = L^2(I)$ obtained from τ by impo-

sition of the boundary conditions $u = \partial_\nu u = \cdots = \partial_\nu^{p-1} = 0$ at the boundary of I, ∂_ν is the normal derivative. Then, by [3, Chap. XIV, §3.1], V is the infinitesimal generator of a strongly continuous semigroup. The spectrum $\sigma(V)$ of V consists of a countable set of points with no finite accumulation point [3, Chap. XIV, §6.23]; if $\lambda \in \rho(V)$, then $R(\lambda; V)$ is a compact operator, which shows that the subspaces $E_n = P_n E$ corresponding to each eigenvalue are finite-dimensional. The fact that the E_n span the entire space E follows from the Browder completeness theorem [3, Chap. XIV, §6.23]. Consequently all the assumptions required in Corollary 3.3 are satisfied, furnishing us with a criterion for complete controllability of the parabolic equation $u'(t) = Vu(t) + Bf(t)$. If $n = 1$, it follows from the theory of ordinary differential equations that V cannot have more than p linearly independent eigenfunctions corresponding to a given eigenvalue. Then V is p-controllable. This is in general false if $n > 1$.(See the corresponding example in [6].)

Let us observe that in most applications the spaces E, F are *real* Banach spaces. In this case we embed E, F into complex Banach spaces in the customary manner (in symbols, $E^c = E \oplus iE$, $F^c = F \oplus iF$) and extend A, B to the new spaces

$$A^c(u + iv) = Au + iAv,$$

$$B^c(u + iv) = Bu + iBv.$$

If L^c is the control system obtained in this way from L, $K(L^c) = K(L) \oplus iK(L)$, $K_t(L^c) = K_t(L) \oplus iK_t(L)$, thus L is completely controllable if and only if L^c is, allowing us to apply our result also to the real case.

REFERENCES

[1] R. E. Kalman, Y. C. Ho and K. S. Narendra, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1963), pp. 189–213.

[2] N. Dunford and J. T. Schwartz, *Linear Operators*, Part I, Interscience, New York, 1957.

[3] ———, *Linear Operators*, Part II, Interscience, New York, 1963.

[4] R. S. Phillips, *Perturbation theory for semi-groups of linear operators*, Trans. Amer. Math. Soc., 74 (1953), pp. 199–221.

[5] E. Hille and R. S. Phillips, *Functional analysis and semi-groups*, American Mathematical Society Colloquium Publications, vol. XXXI, American Mathematical Society, Providence, 1957.

[6] H. O. Fattorini, *On complete controllability of linear systems*, J. Differential Equations, to appear.

[7] ———, *Control in finite time of differential equations in Banach space*, Comm. Pure Appl. Math., 19 (1966), pp. 17–34.

[8] A. V. Balakrishnan, *Optimal control problems in Banach spaces*, this Journal, 3 (1965), pp. 153–180.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

ADDENDUM: ON EXPONENTIAL STABILITY OF LINEAR DIFFERENTIAL SYSTEMS*

NAM P. BHATIA

The proof of Lemma 2.1 in [1] is incorrect. In fact, the lemma is false except in the trivial case $n = 1$, as may easily be seen. This was pointed out to us by Dr. K. M. Das, to whom the author wishes to express his sincere thanks.

The purpose of this note is to provide correct proofs of Theorems 2.3, 2.4, and 2.5, which in [1] were made to depend on Lemma 2.1. As Theorem 2.5 is a restatement of Theorems 2.3 and 2.4, only the latter will be proved.

The same notation, definitions, and numbering as in [1] are used here.

LEMMA A. *If $X(t)$ is a fundamental matrix solution of* (1.1) *and if there are functions $\alpha(\tau, t)$, $\beta(\tau, t)$ such that*

$$\alpha(\tau, t)x'x \leqq x'Y'(\tau, t)Y(\tau, t)x \leqq \beta(\tau, t)x'x,$$

where $Y(\tau, t) = X(\tau)X^{-1}(t)$, and if $\lambda(\tau, t)$ and $\Lambda(\tau, t)$ are the smallest and largest eigenvalues of the matrix $Y'Y$ for any τ, t, then

$$\alpha(\tau, t) \leqq \lambda(\tau, t) \leqq \Lambda(\tau, t) \leqq \beta(\tau, t).$$

The truth of this lemma may be ascertained from the fact that for any symmetric matrix B, if λ and Λ are the least and greatest characteristic roots of B, then

$$\lambda = \inf \frac{x'Bx}{x'x} \quad \text{and} \quad \Lambda = \sup \frac{x'Bx}{x'x}.$$

(See, for example, [3, p. 110].)

Proof of Theorem 2.3. Exponential decay of solutions of (1.1) means that

$$x'x\ \alpha\ \exp\ [-\beta(\tau - t)] \leqq x'Y'(\tau, t)Y(\tau, t)x \leqq x'x\ a\ \exp\ [-b(\tau - t)]$$

holds for some positive constants a, b, α, β, and $\tau \geqq t \geqq 0$. From Lemma A we conclude that

$$\alpha \exp\ [-\beta(\tau - t)] \leqq \lambda(\tau, t) \leqq \Lambda(\tau, t) \leqq a \exp\ [-b(\tau - t)].$$

Now

$$[\det\ Y(\tau, t)]^2 \equiv \exp\left(2\int_t^\tau \operatorname{Tr} A(s)\, ds\right) \equiv \det\ [Y'(\tau, t)Y(\tau, t)]$$

$$\equiv \text{product of all characteristic roots of } Y'Y.$$

* This Journal, 2(1964), pp. 181–191. Received by the editors May 20, 1966.

Hence, as all characteristic roots of $Y'(\tau, t)Y(\tau, t)$ are positive, we have

$$\alpha^n \exp[-n\beta(\tau - t)] \leqq \exp\left(2\int_t^\tau \operatorname{Tr} A(s)\, ds\right) \leqq a^n \exp[-nb(\tau - t)],$$

which on integration yields

$$\frac{\alpha^n}{n\beta} \leqq \int_t^\infty \exp\left(2\int_t^\tau \operatorname{Tr} A(s)\, ds\right) d\tau \leqq \frac{a^n}{nb},$$

which is (2.6) in [1], and the theorem is proved.

LEMMA B. *If the solution $x = 0$ of (1.1) is exponentially stable and $X(t)$ denotes any fundamental matrix solution of (1.1), then there is a positive constant a such that*

$$(*) \qquad 0 < x'Z'Zx \leqq ax'x, \qquad x \neq 0, \qquad \tau \geqq t \geqq 0,$$

where $Z \equiv Z(\tau, t) = \operatorname{adj} Y$, $Y \equiv Y(\tau, t) = X(\tau)X^{-1}(t)$. Consequently,

$$(**) \qquad \frac{1}{a}\, x'x \leqq x'Z'^{-1}Z^{-1}x.$$

Proof. The matrix Y is nonsingular and therefore Z is nonsingular. Hence for fixed τ and t, $y = Zx = 0$ if and only if $x = 0$. Thus $y'y = x'Z'Zx > 0$ if $x \neq 0$. Now exponential stability implies that the coefficients of $Y(\tau, t)$ are uniformly bounded in $\tau \geqq t \geqq 0$. Since the coefficients of $Z(\tau, t)$ are sums of products of coefficients of $Y(\tau, t)$, they are also uniformly bounded. The characteristic roots of $Z'Z$ are thus uniformly bounded implying the existence of a positive constant a such that (*) holds; see, for example, [3, p. 110].

Lastly, replacing x by $Z^{-1}x$ in (*), we get (**). This proves the lemma.

Proof of Theorem 2.4 *in* [1]. Exponential stability of the solution $x = 0$ of (1.1) in [1] means that there are positive constants α and β such that

$$\| X(\tau)X^{-1}(t)x \| \leqq \alpha \| x \| \exp[-\beta(\tau - t)], \qquad \tau \geqq t \geqq 0.$$

Hence the quadratic form

$$V = x'\left[\int_t^\infty Y'Y\, d\tau\right]x$$

satisfies

$$V \leqq \frac{\alpha^2}{\beta}\, x'x.$$

Again

$$V = x'\left[\int_t^\infty Y'Y\,d\tau\right]x = x'\left[\int_t^\infty Y'Z'(Z')^{-1}Z^{-1}ZY\,d\tau\right]x$$

$$\geqq \frac{1}{a}\,x'\left[\int_t^\infty Y'Z'ZY\,d\tau\right]x$$

$$= \frac{1}{a}\,x'\left[\int_t^\infty (\det Y)^2\,d\tau\right]x$$

$$= \frac{1}{a}\,x'x\int_t^\infty \exp\left(2\int_t^\tau \operatorname{Tr} A(s)\,ds\right)d\tau.$$

Thus, if (2.7) in [1] holds, V will satisfy property P. Since $V^*_{(1.1)} = -x'x$, i.e., $-V^*_{(1.1)}$ has property P, we conclude by Theorem 2.1 in [1] that the solution of (1.1) decays exponentially. This completes the proof.

Remark. Theorem 2.4 is an improvement on the well-known Theorem 1.1 of Malkin. For, notice that (2.7) is satisfied whenever the elements of the matrix $A(t)$ are uniformly bounded. Our proof above is really a simplification of the proof given by Malkin and reproduced by Antosiewicz and Davis [2].

REFERENCES

[1] N. P. Bhatia, *On exponential stability of linear differential systems*, this Journal, 2 (1964), pp. 181–191.

[2] H. A. Antosiewicz and P. Davis, *Some implications of Lyapunov's conditions of stability*, J. Rational Mech. Anal., 3 (1954), pp. 447–457.

[3] Richard Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

BOUNDED-INPUT BOUNDED-OUTPUT STABILITY OF NONLINEAR TIME-VARYING DIFFERENTIAL SYSTEMS*

P. P. VARAIYA† AND R. LIU‡

The problem of boundedness of solutions of the differential equation

$$\dot{x} = f(x, t) \tag{1}$$

has been studied by Yoshizawa [1]. He obtains necessary and sufficient conditions for various kinds of stability of (1) using the techniques of the Lyapunov direct method. We have extended the definitions and the methods of Yoshizawa to the study of the bounded-input bounded-output stability of the differential system

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

Here $x \in R^n$ is the state of (S), $u \in R^n$ is the input or control and $t \in I = [0, \infty)$ is the time. $f: R^n \times R^m \times I \to R^n$ is the instantaneous velocity function which satisfies the following conditions: For fixed $t \in I$, f is continuous in the pair (x, u), whereas for fixed (x, u) it is measurable in t. Moreover, for bounded sets $X \subseteq R^n$ and $U \subseteq R^m$ there exist measurable functions $L(t)$ and $M(t)$ (dependent on X, U) which are summable over every finite interval and such that

$$|f(x, u, t)| \leqq M(t) \tag{2}$$

and

$$|f(x, u, t) - f(x', u, t)| \leqq L(t)|x - x'|, \tag{3}$$

for every x, x' in X and u in U. In general, $|x|$ and $|u|$ denote the Euclidean norm of x and u, respectively. Also, if $x(t)$ and $u(t)$ are measurable functions of time, then

$$\|x\| = \sup_t |x(t)| \quad \text{and} \quad \|u\| = \sup_t |u(t)|,$$

where the supremum is taken in each case over those values of t for which

* Received by the editors February 8, 1966, and in final revised form June 22, 1966.
The research reported herein was supported in part by the National Aeronautics and Space Administration under Grant NsG-354 (S-2) and by the Joint Services Electronics Program (Air Force Office of Scientific Research, Army Research Office), Office of Naval Research, under Grant AF-AFOSR-139-65.

† Department of Electrical Engineering, College of Engineering, University of California, Berkeley, California.
‡ On leave of absence from the Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana.

the function is defined. The solutions of (S) are to be interpreted in the sense of Carathéodory [2], [3]. Thus let $u(t)$, $t \in I$, be any bounded measurable function and let (x_0, t_0) be any initial condition. Then a function

$$x(\tau) = x_u(\tau; x_0, t_0) \tag{4}$$

is a solution of (S) if it is absolutely continuous in τ, satisfies the initial condition

$$x(t_0) = x_u(t_0; x_0, t_0) = x_0,$$

and satisfies (S) almost everywhere in the domain of definition of (4). Because of the conditions (2), (3) imposed on f, $x(\tau)$ is defined on a nonvanishing interval containing t_0 and furthermore it is unique [2], [3].

For each $r \geqq 0$ we define the set

$$\Delta_r = \{x: x \in R^n, |x| \geqq r\}. \tag{5}$$

Following Yoshizawa [1] we shall need to consider Lyapunov functions $V(t, x)$ defined continuously on $I \times \Delta_r$ for some r and such that $V \in C_0(x)$. That is to say, for each $\alpha \geqq 0$, there is a continuous function $L(t) = L_\alpha(t)$ such that

$$|V(t, x) - V(t, x')| \leqq L(t)|x - x'| \tag{6}$$

for every x, x' with norm less than α. We also say that $V(t, x)$ is absolutely continuous in t uniformly at a point (x_0, t_0) if there is a positive number ρ (depending on x_0, t_0) such that for each $\epsilon > 0$ there is a number $\delta = \delta(\epsilon) > 0$ such that for every m,

$$\sum_{k=1}^{m} |V(t_k', x_k) - V(t_k, x_k)| < \epsilon,$$

whenever

$$\sum_{k=1}^{m} |t_k' - t_k| < \delta, \qquad t_0 - \rho \leqq t_1' \leqq t_1 \leqq \cdots \leqq t_m' \leqq t_m \leqq r_0 + \rho,$$

and

$$|x_k - x_0| < \delta \quad \text{for each } k.$$

We will always suppose that the Lyapunov functions have this property, so that if $x(t)$ is an absolutely continuous function, $V(t, x(t))$ is also absolutely continuous in a neighborhood of t. Then corresponding to each bounded, measurable function $u(t)$, $t \in I$, we can define

$$V_u'(t, x) = \lim_{h \to 0+} \sup \frac{1}{h} \{V(t + h, x + hf(x, u(t), t)) - V(t, x)\}.$$

DEFINITION 1. The system (S) is *bounded-input bounded-output stable* (BIBO) if for every $\alpha \geqq 0$, for every $a \geqq 0$ there is a number $\beta = \beta(\alpha, a)$ such that

$$| x_u(\tau; x_0, t_0) | \leqq \beta \quad \text{for all} \quad \tau \geqq t_0, \tag{7}$$

for every initial condition (x_0, t_0) with $| x_0 | \leqq \alpha$ and every measurable function $u(t)$, $t \in I$, with $\| u \| \leqq a$.

Remarks. Since (7) depends on the solution (4) to the system (S), it is useful for a large class of dynamical systems. Of course the nature of the results is such as to be particularly useful for differential systems.

Various weaker notions of boundedness can also be introduced. In some cases analogous results can be obtained. The reader is referred to Yoshizawa [1] for a thorough discussion of the behavior of (1).

DEFINITION 2. We say that the Lyapunov function $V(t, x)$ has property A if there is a positive continuously increasing function $a(r)$ such that $V(t, x) \leqq a(| x |)$. It has property B if there is a nonnegative continuously increasing function $b(r)$ with $b(r) \to \infty$ as $r \to \infty$ and such that $b(| x |) \leqq V(t, x)$.

A trivial refinement of the proof of Theorem 3 of Yoshizawa [1] yields Theorem 1.

THEOREM 1. *Suppose for each $a \geqq 0$ there is a positive Lyapunov function $V(t, x) = V_a(t, x)$, defined in $\Delta = \Delta_{r(a)}$, and possessing properties* A *and* B. *Then, if*

$$V_u'(t, x) \leqq 0 \tag{8}$$

for (t, x) in Δ and for each measurable function $u(t)$, $t \in I$, with $\| u \| \leqq a$, the system (S) is BIBO *stable.*

The following lemma will be very useful to prove the converse of Theorem 1.

LEMMA 1. *Let (x_0, t_0) and (x_1, t_1) be two initial conditions with $t_0 \leqq t_1$, let $u(t)$ be an arbitrary measurable function on I with $\| u \| \leqq a$. Suppose that the two solutions*

$$x_u(t; x_0, t_0) \quad \text{and} \quad x_u(t; x_1, t_1)$$

can be defined to the left over the interval $t^* \leqq t \leqq t_0$, $t^* \leqq t \leqq t_1$. (We assume that $0 \leqq t^*$.) Also suppose that $| x_u(t; x_0, t_0) |$ and $| x_u(t; x_1, t_1) |$ are less than α over these intervals. Then

$$\begin{aligned} &| x_u(t^*; x_0, t_0) - x_u(t^*; x_1, t_1) | \\ &\qquad \leqq \left[| x_1 - x_0 | + \int_{t_0}^{t_1} M(\tau)\, d\tau \right] \left[\exp \int_{t^*}^{t_0} L(\tau)\, dt \right], \end{aligned} \tag{9}$$

where the functions L and M are the same as those in (2) *and* (3).

The proof of this lemma is very similar to the proof of the (generalized) Gronwall's lemma given in [2].

THEOREM 2. *If* (S) *is* BIBO *stable, for each* $a \geqq 0$ *there is a Lyapunov function* $V(t, x) = V_a(t, x)$ *defined on* $\Delta = \Delta_{r(a)}$ *such that* V *has properties* A *and* B *and* $V'(t, x) \leqq 0$.

Proof. Fix $a \geqq 0$. Since (S) is BIBO stable, for each u with $\| u \| \leqq a$, for each (x_0, t_0) with $| x_0 | = \alpha$,

$$| x_u(\tau; x_0, t_0) | \leqq \beta(\alpha) = \beta(a, \alpha), \tag{10}$$

for all $\tau \geqq t_0$. We can assume that β is a continuous strictly monotonically increasing function and $\beta(\alpha) \to \infty$ as $\alpha \to \infty$. Hence the inverse function $\alpha = \alpha(\beta)$ is defined for $\alpha \geqq \beta(0)$ and has the same properties as β.

Let $r(a) = \beta(0)$. Let $\Delta = \Delta_{r(a)}$. Then, for each (x_0, t_0) in $\Delta \times I$ and each u with $\| u \| \leqq a$, define

$$K_u(t_0, x_0) = \min \{ | x_u(\tau; x_0, t_0) | : 0 \leqq \tau \leqq t_0 \}, \tag{11}$$

where the region of τ is that for which the solution $x_u(\tau; x_0, t_0)$ exists. The required Lyapunov function is

$$V(t_0, x_0) = V_a(t_0, x_0) = \inf \{ K_u(t_0, x_0) : \| u \| \leqq a \}. \tag{12}$$

Clearly $0 \leqq K_u(t_0, x_0) \leqq | x_0 |$ so that

$$V(t_0, x_0) \leqq | x_0 |. \tag{13}$$

Hence V has the property A. We also claim that

$$0 < \alpha(| x_0 |) \leqq V(t_0, x_0). \tag{14}$$

If this is not the case, then there is a u, $\| u \| \leqq a$, such that for some x_0

$$K_u(t_0, x_0) < \alpha(| x_0 |).$$

Hence for some τ, $0 \leqq \tau \leqq t_0$, we must have

$$| x_u(\tau; x_0, t_0) | < \alpha(| x_0 |).$$

Therefore

$$\beta(| x_u(\tau; x_0, t_0) |) < \beta(\alpha(| x_0 |)) = | x_0 |.$$

But $\beta(| x_u(\tau; x_0, t_0) |) \geqq | x_0 |$, which is a contradiction. Hence (14) is true, so that V has property B. It remains to be shown that V has the required smoothness properties.

Let (x_0, t_0) be any element of $\Delta \times I$. Then, using the lemma, the same arguments as in the proof of Theorem 4 of Yoshizawa [1] yield

$$| K_u(t, x) - K_u(t', x') | \leqq A \left(| x - x' | + \int_t^{t'} M(\tau)\, dt \right)$$

for every u, $\| u \| \leqq a$, and every (x, t), (x', t') in some δ-neighborhood N of (x_0, t_0). Therefore by definition (12) of V_a we have

$$(22) \qquad | V_a(t, x) - V_a(t', x') | \leqq A \left(| x - x' | + \int_t^{t'} M(\tau)\, dt \right)$$

in a δ-neighborhood N of (t_0, x_0). Trivially from above $V \in C_0(x)$. Also for every m, and

$$t_1' \leqq t_1 \leqq \cdots \leqq t_m' \leqq t_m,$$

and every $x_1, x_2, \cdots, x_m$ with (x_k, t_k) and (x_k, t_k') in N, we have from (22),

$$\sum_{k=1}^{m} | V_a(t_k', x_k) - V_a(t_k, x_k) | \leqq A \sum_{k=1}^{m} \int_{t_k}^{t_k'} M(\tau)\, dt.$$

Since $M(\tau)$ is an integrable function, its indefinite integral is absolutely continuous. Hence $V(t, x)$ is absolutely continuous in t uniformly at each point. V therefore has the required smoothness properties. Also by definition (11) of $K_u(t, x)$ we see that $K_u(\tau; x_u(\tau; x, t))$ is nonincreasing in τ. Hence $V_a(t, x)$ is nonincreasing along every solution of (S) for each u with $\| u \| \leqq a$. Therefore $V_a'(t, x) \leqq 0$. The theorem is proved.

A simple application. We close this paper by a simple application of Theorem 1. Let the differential system [4], [5] be given by

$$(15) \qquad \begin{aligned} \dot{x} &= Ax + bf(\sigma), \\ \dot{\sigma} &= d^T x - rf(\sigma) + u, \end{aligned}$$

where A is an $n \times n$ matrix with all its eigenvalues having negative real parts, b, d and x are n-vectors whereas σ and u are scalars, f is a locally integrable function of σ such that

$$f(\sigma) \to \pm\infty \quad \text{as} \quad \sigma \to \pm\infty.$$

Consider the function

$$V(x, \sigma) \triangleq x^T Q x + \int_0^{\sigma} f(\sigma')\, d\sigma',$$

where $Q > 0$. Clearly $V(x, \sigma)$ is positive for $| x | + | \sigma |$ sufficiently large, and V enjoys properties A and B. Then, if $y = (x, f(\sigma))$, we have

$$\dot{V} = -y^T F y + f(\sigma) u,$$

where F is an $(n + 1) \times (n + 1)$ matrix with

$$F = \left[\begin{array}{c|c} G & g \\ \hline g^T & r \end{array}\right],$$

where $-G = A^T Q + QA$, $-g = Qb + \frac{1}{2}d$.

The following result is a straightforward application of Theorem 1.

THEOREM 3. *If* $F > 0$, *then the system* (15) *is* BIBO *stable.*

The previous example serves to illustrate the situation where, if the zero-input system

$$x = f(x, 0, t) \tag{16}$$

is uniformly asymptotically stable in the whole (u.a.s.w.), then the system (S) is BIBO stable. In the remainder of this paper we exhibit a class of systems which have this property.

First of all, we only consider systems (S) for which f is continuous in (x, u, t) and for which there exists a constant k such that

$$| f(x, u, t) - f(x, 0, t) | \leqq k | u |, \tag{17}$$

for all $u \in R^m$. It is well known [6, p. 71] that if (16) is u.s.a.w., then there exists a Lyapunov function $V(t, x)$ possessing properties A and B and such that $\dot{V}$ is negative definite. Next we show that if V satisfies an additional condition, then (S) is BIBO stable.

THEOREM 4. *For the zero-input system* (16), *if there exists a Lyapunov function* $V(x, t)$ *possessing properties A and B, if* $\dot{V}$ *is negative definite, and in addition, if for every number* $b \geqq 0$, *there exists a number* $M = M(b)$ *such that*

$$| \dot{V}(x, t) | \geqq b | \nabla V | \tag{18}$$

whenever $| x | \geqq M$, *then* (S) *is* BIBO *stable.*

Remark. Theorem 4 imposes conditions only on the zero-input system (16) and not on (S) (except for the condition (17)).

Proof. Let $\dot{V}_u \triangleq \langle \nabla V, f(x, u, t) \rangle + V_t$ and $\dot{V} \triangleq \langle \Delta V, f(x, 0, t) \rangle + V_t$. Then for every function u, $\| u \| \leqq a$, by (17) and (18), we see that

$$\dot{V}_u = \langle \nabla V, f(x, u, t) - f(x, 0, t) \rangle + \dot{V} \leqq ka | \nabla V | + \dot{V} \leqq 0,$$

for all x, $| x | \geqq M(ka)$. Hence, by Theorem 1, (S) is BIBO stable.

COROLLARY 1. *Suppose that* (S) *satisfies* (17) *and that the zero-input system is either homogeneous of rational order* $r > 0$ [6, p. 90] *or has intensive behavior* [6, p. 85]. *Then, if the zero-input system is u.a.s.w., the system* (S) *is* BIBO *stable.*

Proof. If (S) satisfies the hypothesis of the corollary, then the theorems of Zubov [6, p. 91] and of Krasovskii [6, p. 86] assert the existence of a Lyapunov function which satisfies the hypothesis of Theorem 4.

REFERENCES

[1] T. YOSHIZAWA, *Liapunov's function and boundedness of solutions*, Funkcial. Ekvac., 2 (1959), pp. 95–142.

[2] G. SANSONE AND R. CONTI, *Nonlinear Differential Equations*, Macmillan, New York, 1964.
[3] C. CARATHÉODORY, *Vorlesungen über Reele Funktionen*, Chelsea, New York, 1948, p. 665.
[4] J. P. LASALLE, *Stability and control*, this Journal, 1 (1962), pp. 3–15.
[5] V. A. YAKUBOVICH, *The method of matrix inequalities in the stability theory of nonlinear control systems II*. Automat. Remote Control, 26 (1962), pp. 577–592.
[6] W. HAHN, *Theory and Application of Liapunov's Direct Method*, Prentice-Hall Englewood Cliffs, New Jersey, 1963.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

A FIXED-POINT METHOD FOR A MINIMUM-NORM CONTROL PROBLEM*

J. E. RUBIO†

Abstract. A method for obtaining the control which minimizes the terminal value of the norm of the state vector of a linear, time-varying system is presented. It is shown that the control that minimizes the linear functional given by the inner product $(\hat{c}, x(t_f))$, with $\hat{c}$ a fixed point of an operator L_0 and $x(t_f)$ the terminal state vector, also minimizes the norm of $x(t_f)$. The operator L_0 maps vectors c of unity norm into vectors $x(t_f)/\|x(t_f)\|$, with $x(t_f)$ obtained by applying the control that minimizes $(c, x(t_f))$.

This linear minimization problem is studied in detail, and existence conditions for the optimum control are established. An iterative technique is given for the computation of fixed points of L_0 ; it is shown that the procedure converges. The method is characterized by its comparatively modest computational requirements.

1. Introduction. The problem to be treated in this paper is the minimization of the terminal value of the norm of the state vector of a linear, time-varying system with control inputs bounded in magnitude. Let the system be described by the following vector differential equation:

$$\dot{x} = A(t)x + B(t)u. \tag{1}$$

Here x is an n-dimensional state vector, u is an r-dimensional control vector, $A(t)$ is a real $n \times n$ matrix and $B(t)$ is a real $n \times r$ matrix. The control u is to belong to the class U of piecewise continuous controls with components bounded in magnitude by $|u_k| \leqq 1$, $k = 1, \cdots, r$. The system of equations (1) is to satisfy the initial condition $x(t_0) = x_0 \neq 0$, and the problem is to find a control vector $u^* \in U$ such that $\| x(t_f, u)\|$, the state vector norm at terminal time t_f when the control u has been applied, is a minimum; t_f is a given number greater than t_0. The usual Euclidean norm will be used throughout; that is, $\| x \|^2 = (x, x) = \sum_{i=1}^{n} x_i^2$.

This problem has interest for two reasons. Firstly, if $\| x(t_f, u^*)\| = a$, it follows that $|x_i(t_f, u^*)| \leqq a$, $i = 1, \cdots, n$; that is, all components of the state vector at the terminal time are bounded in magnitude by the minimum possible bound. The minimization of the terminal value of the norm of the state vector is then a way to keep a rather efficient control over all of its components by using a scalar performance criterion. Besides, as pointed out by Ho [1], the time-optimal problem can be solved by repeatedly minimizing $\| x(t_f, u)\|$ for different values of the terminal time t_f.

The application of the traditional optimization techniques to this prob-

* Received by the editors February 1, 1966.

† Department of Electrical and Electronic Engineering, University of Leeds, Leeds, England.

lem does not seem conducive to much success. For instance, if Pontryagin's optimization procedure [2] is tried, the following differential equations are readily obtained; here the p_j are a set of auxiliary variables.

$$\dot{x}_i = \sum_{j=1}^{n} a_{ij}x_j + \sum_{k=1}^{r} b_{ik}u_k^*, \qquad x_i(t_0) = x_{i0}, \quad i = 1, \cdots, n,$$

$$\dot{p}_j = -\sum_{i=1}^{n} (p_i - 2x_i)a_{ij}, \qquad p_j(t_f) = 0, \quad j = 1, \cdots, n, \tag{2}$$

$$u_k^* = \operatorname{sgn} \sum_{i=1}^{n} b_{ik}(p_i - 2x_i), \qquad k = 1, \cdots, r.$$

However, this system is not simple to study from a theoretical standpoint. (That is, existence and uniqueness properties of the solutions seem quite difficult to establish.) Nor is it simple to treat as a practical one: estimation of these solutions by means of a computer appears as a cumbersome task, involving a search in n-dimensional space for the initial values $p(t_0)$ that will make the solutions satisfy the terminal boundary conditions.

It is not surprising, then, that several authors have developed other methods to cope with this problem; the most significant contribution is by Ho [1]; he presents an iterative procedure that he applies to time-invariant systems. Recently, Fancher [3] has extended this procedure to normal [4], time-varying systems, and has improved the computational aspects.

The procedure to be presented below is iterative in nature. It will be proved that the optimal control u^* that minimizes $\| x(t_f, u)\|$ can be obtained by means of a process that involves the successive minimizations of linear functionals of the type $S = \sum_{i=1}^{n} c_i x_i(t_f, u)$. This is quite fortunate, because each of these problems can be studied quite thoroughly from a theoretical standpoint. Pontryagin's equations are known to have a unique solution under some conditions on the matrices $A(t)$ and $B(t)$; these equations can also be handled numerically in quite a simple way, no search being needed for the estimation of the solutions.

2. The fundamental theorem. For the system described by (1), with $u \in U$, consider the problem of determining the optimum control u_c^* that minimizes the following functional:

$$S_c(u) = \sum_{i=1}^{n} c_i x_i(t_f, u) = (c, x(t_f, u)).$$

It will be assumed in what follows that $c \in \Omega$; Ω is the set $\Omega = \{x \mid \| x \| = 1\}$; c is the vector with components c_i, $i = 1, \cdots, n$.

This problem has a unique solution under several restrictions on $A(t)$ and $B(t)$; this solution will be studied in the next section. For the moment,

it is sufficient to admit the existence of an optimal control u_c^* for any $c \in \Omega$ by means of which a global minimum is attained as indicated by (3) below.

Consider then the transformation L_0 (the reason for the use of the subscript "0" will be apparent later) that maps the vector c into the vector $x(t_f, u_c^*)/\| x(t_f, u_c^*)\|$. Clearly, L_0 maps Ω into itself and is well-defined unless $\| x(t_f, u_c^*)\|$ equals zero. The following theorem is of great importance.

THEOREM 1. *If $\hat{c}$, a fixed point of the operator L_0, exists, and $\min_{u \in U} \| x(t_f, u)\|$ is not zero, then the control $u_{\hat{c}}^*$ that minimizes $S_{\hat{c}}$ also minimizes $\| x(t_f, u)\|$.*

Proof. Since $u_{\hat{c}}^*$ minimizes $S_{\hat{c}}$,

$$S_{\hat{c}}(u_{\hat{c}}^*) \leqq S_{\hat{c}}(u) \quad \text{for all} \quad u \in U. \tag{3}$$

If $\hat{c}$ is a fixed point of L_0, and $\| x(t_f, u_{\hat{c}}^*)\|$ is not zero (which certainly will be the case, since $\min_{u \in U} \| x(t_f, u)\|$ is not zero), then $\hat{c} = x(t_f, u_{\hat{c}}^*)/\| x(t_f, u_{\hat{c}}^*)\|$; thus

$$S_{\hat{c}}(u_{\hat{c}}^*) = \| x(t_f, u_{\hat{c}}^*)\|. \tag{4}$$

Using now (3), (4) and Schwarz's inequality,

$$\begin{aligned} \| x(t_f, u_{\hat{c}}^*)\| = S_{\hat{c}}(u_{\hat{c}}^*) \leqq S_{\hat{c}}(u) &\leqq \| \hat{c}\| \, \| x(t_f, u)\| \\ &= \| x(t_f, u)\| \quad \text{for all} \quad u \in U. \end{aligned} \tag{5}$$

Then, $u_{\hat{c}}^*$ minimizes $\| x(t_f, u)\|$.

COROLLARY. *If a and b are fixed points of L_0, then $\| x(t_f, u_a^*)\| = \| x(t_f, u_b^*)\|$.*

Proof. By applying (5) with $\hat{c} = a$, it is concluded that

$$\| x(t_f, u_a^*)\| \leqq \| x(t_f, u_b^*)\|.$$

By applying (5) with $\hat{c} = b$, the reverse inequality follows. The corollary is then proved.

Of course, the conclusions of this corollary can be extended to an arbitrary number of fixed points; this means that all the controls $u_{\hat{c}}^*$, with $\hat{c}$ any fixed point of L_0, are equally effective in minimizing $\| x(t_f, u)\|$. Then, questions concerning uniqueness of fixed points of L_0 are uninteresting for the present purpose.

An iterative method will be presented below to obtain these fixed points and at the same time construct approximations to the optimal control u^* that minimizes $\| x(t_f, u)\|$. As a matter of fact, it will be proved that the sequence $c, L_0c, L_0^2c, \cdots, L_0^mc, \cdots$, with $c \in \Omega$ but otherwise arbitrary, converges to a fixed point of L_0, that is, to a solution $\hat{c}$ of the equation $L_0y = y$. Approximations to u^* are also obtained by this process, since the

application of L_0 to $L_0{}^m c$, say, involves the calculation of the optimal control $u_{c'}^*$, with $c' = L_0{}^m c$.

When $\min_{u \in U} \| x(t_f, u) \| = 0$, the conclusions of Theorem 1 cannot be applied; this case is considered in Appendix A.

3. The linear minimization problem. The problem of minimizing S_c will be considered in detail now; expressions will be derived for the optimal control $u_c{}^*$, and existence conditions studied.

To minimize S_c, according to Pontryagin's method [2], a Hamiltonian H is to be constructed first:

$$H = \sum_{j=1}^{n} x_j \left(\sum_{i=1}^{n} a_{ij}(t) p_i \right) + \sum_{k=1}^{r} u_k \left(\sum_{i=1}^{n} b_{ik}(t) p_i \right).$$

The differential equations for p do not depend upon u, and therefore can now be written:

$$\dot{p}_j = -\sum_{i=1}^{n} a_{ij}(t) p_i, \qquad p_j(t_f) = -c_j, \qquad j = 1, \cdots, n. \tag{6}$$

This is of course the adjoint system to system (1). If the transition matrix of (6) is $\Psi(t, \tau)$, then the solution of (6) is

$$p(c, t) = -\Psi(t, t_f) c. \tag{7}$$

No search is necessary to compute the vector p, and only the matrix $\Psi(t, t_f)$ needs to be evaluated.

The control vector $u_c{}^*$ that minimizes S_c maximizes H. Define $\beta_k(c, t) = \sum_{i=1}^{n} b_{ik}(t) p_i$, $k = 1, \cdots, r$. It is clear that if $\beta_k(c, t')$, $t' \in [t_0, t_f]$, is positive (negative) then $u_{ck}^*(t')$ must take the value one (minus one). If $\beta_k(c, t')$ is zero, then H does not depend on u_k, and the problem is one of singular control [5].

Call E_k the set of points $t \in [t_0, t_f]$ such that $\beta_k(c, t) = 0$, $k = 1, \cdots, r$. If the Lebesgue measure $m(E_k)$ is zero, the value which is assigned to u_k at $t \in E_k$ is of no importance. Conditions will be then assumed on $A(t)$ and $B(t)$ such that $m(E_k) = 0$, $k = 1, \cdots, r$. A set of such conditions has been derived by Pontryagin and his co-workers [2, Chap. III, Theorem 15] in connection with the time-optimal problem:

(i) The functions $a_{ij}(t)$ and $b_{ik}(t)$ are defined on an open interval (a, b) that contains $[t_0, t_f]$. They have over this interval respectively $(n - 2)$ and $(n - 1)$ continuous derivatives.

(ii) The general position condition is satisfied [2, pp. 182–183].

These conditions will be assumed to be satisfied in what follows; then, the values of u_{ck}^* can be defined arbitrarily for $t \in E_k$, and will be made equal to zero. Then,

$$u_{ck}^*(t) = \operatorname{sgn} \beta_k(c, t), \qquad k = 1, \cdots, r, \tag{8}$$

with $\operatorname{sgn} \lambda = 0$ if $\lambda = 0$.

Due to the linearity of (1), Pontryagin's maximum principle is known to be a sufficient condition for optimality [5]; then, the control u_c^* defined by (8) satisfies (3).

Once u_c^* is computed by means of (8), $x(t_f, u_c^*)$ can be obtained from the well-known expression [6]:

$$x(t_f, u_c^*) = \Phi(t_f, t_0)x_0 + \int_{t_0}^{t_f} \Phi(t_f, \tau)B(\tau)u_c^*(\tau)\, d\tau. \tag{9}$$

Here $\Phi(t, \tau)$ is the transition matrix of the system $\dot{x} = Ax$. It is well-known [6] that this matrix is related to $\Psi(t, \tau)$ by the expression

$$\Psi^*(t, \tau) = \Phi(\tau, t), \tag{10}$$

where $\Psi^*(t, \tau)$ is the adjoint matrix of $\Psi(t, \tau)$.

The application of the operator L_0 to a vector $c \in \Omega$ can be summarized by the following steps:

(i) The vector $p(c, t)$ is computed by means of (7).

(ii) The optimal control u_c^* is obtained from (8).

(iii) The terminal state vector $x(t_f, u_c^*)$ is computed by means of (9).

(iv) The vector $x(t_f, u_c^*)/\| x(t_f, u_c^*) \| = L_0 c$ is computed if $x(t_f, u_c^*)$ is not zero. Otherwise, L_0 is not defined for that particular value of c.

This procedure is comparatively simple; only one of the matrices Ψ and Φ needs to be computed because of (10). Several efficient numerical methods exist for estimating the solutions of systems like (6); this needs to be done only once during the iterative procedure that was introduced in §2 for obtaining the fixed points of L_0.

4. The sequence $\{L_0^m c\}$. It will be proven now that the sequence $\{L_0^m c\}$ converges to a solution $\hat{c}$ of the equation $L_0 y = y$ for any initial vector $c \in \Omega$. With this purpose in mind, it will prove convenient to introduce the operator L_σ, defined as follows: to apply L_σ to a vector $c \in \Omega$, the sequence of operations (i) to (iv) of §3 is carried out without modification with the exception of step (ii), where instead of using the control u_c^* given by (8), the following control u_c^σ is applied:

$$u_{ck}^\sigma = 2G(\beta_k(c, t), \sigma) - 1, \qquad k = 1, \cdots, r,$$

where $G(y, \sigma)$ is the cumulative Gaussian distribution

$$G(y, \sigma) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y/\sigma} e^{-\frac{1}{2}z^2}\, dz.$$

It is clear that $\lim_{\sigma \to 0} u_c^\sigma = u_c^*$, the optimal control. The operator L_0 will be studied by introducing some properties of the operators L_σ, $\sigma \in [0, \infty)$. It will be assumed again that $\min_{u \in U} \| x(t_f, u) \| = d > 0$; this implies that all terminal state vectors resulting from the application of controls u_c^σ, for all $c \in \Omega$ and all $\sigma \in [0, \infty)$, will satisfy the relation

$$\| x(t_f, u_c^\sigma) \| \geqq d > 0. \tag{11}$$

In Appendix A, a simple procedure to be used when $\min_{u \in U} \| x(t_f, u) \| = 0$ is presented; it is shown there that, by a simple change in the origin of the state space, the general minimization method can be applied also in this case.

It will be proved first that the sequence $\{L_\sigma{}^m c\}$ converges for large enough values of σ and all $c \in \Omega$. This will be achieved by showing that L_σ is a contracting operator [7] for sufficiently large values of σ. Since L_σ maps Ω, a closed set, into itself, it will follow [7] that the sequence $\{L_\sigma{}^m c\}$, $c \in \Omega$, converges for sufficiently large values of σ.

THEOREM 2. *There exists a $\sigma_0 > 0$ such that L_σ is a contracting operator for $\sigma > \sigma_0$.*

Proof. For $c_1, c_2 \in \Omega$, it is to be proved [7] that

$$\| L_\sigma c_1 - L_\sigma c_2 \| \leqq \alpha \| c_1 - c_2 \|,$$

where $\alpha < 1$ should be independent of c_1, c_2.

Assuming $\sigma \neq 0$,

$$(12) \qquad \begin{aligned} | u^\sigma_{c_1 k} - u^\sigma_{c_2 k} | &= \sqrt{\frac{2}{\pi}} \left| \int_{\beta_k(c_2,t)/\sigma}^{\beta_k(c_1,t)/\sigma} e^{-\frac{1}{2}z^2} \, dz \right| \\ &\leqq \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} | \beta_k(c_1, t) - \beta_k(c_2, t) |, \qquad k = 1, \cdots, r, \quad t \in [t_0, t_f]. \end{aligned}$$

Besides,

$$(13) \qquad \begin{aligned} | \beta_k(c_1, t) - \beta_k(c_2, t) | &= \left| \sum_{i=1}^{n} b_{ik}(t)(p_i(c_1, t) - p_i(c_2, t)) \right| \\ &= \left| \sum_{j=1}^{n} (c_{1j} - c_{2j}) \left(\sum_{i=1}^{n} \chi_{ij}(t, t_f) b_{ik}(t) \right) \right|, \\ & \qquad k = 1, \cdots, r, \quad t \in [t_0, t_f], \end{aligned}$$

where the $\chi_{ij}(t, t_f)$ are the entries of the matrix $\Psi(t, t_f)$. Since, under the conditions on $A(t)$, they are bounded over the interval $[t_0, t_f]$, and since the functions $b_{ik}(t)$ are continuous over this interval, a number M can be found such that

$$(14) \qquad \left| \sum_{i=1}^{n} \chi_{ij}(t, t_f) b_{ik}(t) \right| < M \quad \text{for} \quad t \in [t_0, t_f], \qquad k = 1, \cdots, r, \quad j = 1, \cdots, n.$$

From (13) and (14) and Schwarz's inequality it can be inferred that

$$(15) \qquad | \beta_k(c_1, t) - \beta_k(c_2, t) | \leqq \| c_1 - c_2 \| N$$

with $N = \sqrt{n}\, M$, for $t \in [t_0, t_f]$. From (12) and (15) it follows that

$$\| u^\sigma_{c_1} - u^\sigma_{c_2} \| \leqq \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} N \| c_1 - c_2 \|.$$

If $u_{c_1}^{\sigma}$ and $u_{c_2}^{\sigma}$ are applied to the system described by (1), and the final values of the state vectors are called x_{f1} and x_{f2} respectively, then

$$x_{f1} - x_{f2} = \int_{t_0}^{t_f} \Phi(t_f, \tau)B(\tau)(u_{c_1}^{\sigma}(\tau) - u_{c_2}^{\sigma}(\tau))\, d\tau.$$

Since the matrices $\Phi(t_f, \tau)$ and $B(\tau)$ have bounded components, a constant η can be found such that

$$\begin{aligned} \| x_{f1} - x_{f2} \| &\leqq \int_{t_0}^{t_f} \| \Phi(t_f, \tau)B(\tau) \| \, \| u_{c_1}^{\sigma}(\tau) - u_{c_2}^{\sigma}(\tau) \| \, d\tau \\ &\leqq \eta(t_f - t_0) \max_{[t_0, t_f]} \| u_{c_1}^{\sigma}(t) - u_{c_2}^{\sigma}(t) \| . \end{aligned} \tag{16}$$

The $\max_{[t_0, t_f]} \| u_{c_1}^{\sigma}(t) - u_{c_2}^{\sigma}(t) \|$ exists because of the continuity of $u_{c_1}^{\sigma}(t)$ and $u_{c_2}^{\sigma}(t)$ over $[t_0, t_f]$.

Then, since (15) holds for all $t \in [t_0, t_f]$,

$$\| x_{f1} - x_{f2} \| \leqq \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} N\eta \, \| c_1 - c_2 \| .$$

Without loss of generality, assume that $\| x_{f2} \| \geqq \| x_{f1} \|$. Using then (11) and (22) in Appendix B,

$$\begin{aligned} \left\| \frac{x_{f1}}{\| x_{f1} \|} - \frac{x_{f2}}{\| x_{f2} \|} \right\| &\leqq \frac{1}{\| x_{f1} \|} \| x_{f1} - x_{f2} \| \\ &\leqq \frac{1}{d} \sqrt{\frac{2}{\pi}} \frac{1}{\sigma} N\eta \, \| c_1 - c_2 \| . \end{aligned} \tag{17}$$

It is clear at last that $\alpha = (1/\sigma)\sqrt{2/\pi}\,(N\eta/d)$ can be made less than unity by choosing $\sigma > \sigma_0 = \sqrt{\pi/2}\,(d/N\eta)$, and that α does not depend upon c_1 or c_2. Theorem 2 is proved.

The fact that the sequence $\{L^m c\}$ is known to converge for $\sigma > \sigma_0$ will be used now to establish the convergence of this sequence for $\sigma \in [0, \sigma_0]$. The following lemma will be useful.

LEMMA 1. *The vector $L_\sigma c$ is a continuous function of σ for any $\sigma \in [0, \infty)$ and any $c \in \Omega$.*

Proof. (i) The continuity of $L_\sigma c$ will be established first for $\sigma = \sigma_1 \neq 0$. Take $\sigma_2 \neq 0$; in a similar way as before,

$$| u_{ck}^{\sigma_1} - u_{ck}^{\sigma_2} | \leqq \sqrt{\frac{2}{\pi}} \, | \beta_k(c, t) | \left| \frac{1}{\sigma_1} - \frac{1}{\sigma_2} \right| \leqq \sqrt{\frac{2}{\pi}} M \frac{| \sigma_1 - \sigma_2 |}{\sigma_1 \sigma_2} ,$$

for all $t \in [t_0, t_f]$, $k = 1, \cdots, r$, $c \in \Omega$, $\sigma_1 \neq 0$, $\sigma_2 \neq 0$. The last inequality was obtained by using (14) and Schwarz's inequality in the expression for $\beta_k(c, t)$. Then,

$$\| u_c^{\sigma_1} - u_c^{\sigma_2} \| \leqq \sqrt{\frac{2}{\pi}} N \frac{| \sigma_1 - \sigma_2 |}{\sigma_1 \sigma_2} , \qquad \text{for} \quad t \in [t_0, t_f], \quad c \in \Omega,$$

$$\sigma_1 \neq 0, \qquad \sigma_2 \neq 0.$$

Using (16) again, and calling x_{f1} and x_{f2} the values of the terminal state vectors for σ_1 and σ_2, respectively,

$$\| x_{f1} - x_{f2} \| \leqq \sqrt{\frac{2}{\pi}}\, \eta N \frac{| \sigma_1 - \sigma_2 |}{\sigma_1 \sigma_2}.$$

Again, without loss of generality assume that $\| x_{f1} \| \geqq \| x_{f2} \|$. Then, as in (17),

$$\left\| \frac{x_{f1}}{\| x_{f1} \|} - \frac{x_{f2}}{\| x_{f2} \|} \right\| = \| L_{\sigma 1}\, c - L_{\sigma 2}\, c \| \leqq \sqrt{\frac{2}{\pi}}\, \frac{\eta N}{d \sigma_1 \sigma_2} | \sigma_1 - \sigma_2 |.$$

Given an $\epsilon > 0$, a $\delta(\epsilon) < \sqrt{\pi/2}(d\sigma_1\sigma_2/\eta N)\epsilon$ can be found such that, if $| \sigma_1 - \sigma_2 | < \delta$, $\| L_{\sigma 1} c - L_{\sigma 2} c \| < \epsilon$. Note that $\delta(\epsilon)$ does not depend on the vector c.

(ii) Assume now that $\sigma_1 = 0$. It is clear that, for any value of $\beta_k(c, t) \neq 0$, u_{ck}^{σ} can be made as close to $+1$ (or to -1) as desired by making σ sufficiently small, which will make β_k/σ as large (small) as necessary. If $\beta_k(c, t) = 0$, $u_{ck}^{\sigma} = 0$ and $u_{ck}^{*} = 0$; continuity is also preserved in this case.

THEOREM 3. *The sequence* $\{L_{\sigma}{}^{m} c\}$ *converges for* $\sigma \in [0, \infty)$ *for any* $c \in \Omega$.

Proof. (i) It will be proved first that, if $\{L_{\sigma 1}^{m} c\}$ converges, $\{L_{\sigma 2}^{m} c\}$ converges if $| \sigma_1 - \sigma_2 |$ is sufficiently small.

Since $\{L_{\sigma 1}^{m} c\}$ converges, given an $\epsilon > 0$, an $N(\epsilon)$ can be found such that, if $p \geqq N(\epsilon)$ and $q \geqq N(\epsilon)$, then $\| L_{\sigma 2}^{p} c - L_{\sigma 2}^{q} c \| < \epsilon$ for all $c \in \Omega$. Since $L_{\sigma} c$ is continuous with respect to σ, $L_{\sigma}{}^{p} c$ and $L_{\sigma}{}^{q} c$ are also continuous with respect to this parameter; given $\epsilon_1 > 0$ and $\epsilon_2 > 0$, numbers $\delta_1(\epsilon_1)$ and $\delta_2(\epsilon_2)$ can be found such that, if $| \sigma_1 - \sigma_2 | < \delta_1(\epsilon_1)$, $\| L_{\sigma 2}^{p} c - L_{\sigma 1}^{p} c \| < \epsilon_1$, and if $| \sigma_1 - \sigma_2 | < \delta_2(\epsilon_2)$, $\| L_{\sigma 2}^{q} c - L_{\sigma 1}^{q} c \| < \epsilon_2$. If $| \sigma_1 - \sigma_2 | < \min(\delta_1, \delta_2)$, and if $p \geqq N(\epsilon)$, $q \geqq N(\epsilon)$, then

$$\| L_{\sigma 2}^{p} c - L_{\sigma 2}^{q} c \| \leqq \| L_{\sigma 1}^{p} c - L_{\sigma 1}^{q} c \| + \| L_{\sigma 2}^{p} c - L_{\sigma 1}^{p} c \| + \| L_{\sigma 1}^{q} c - L_{\sigma 2}^{q} c \|$$
$$< \epsilon + \epsilon_1 + \epsilon_2.$$

Thus the sequence $\{L_{\sigma 2}^{m} c\}$ converges for all $c \in \Omega$ for sufficiently small values of $| \sigma_1 - \sigma_2 |$, since it is a fundamental sequence and the Euclidean n-space is complete [7].

(ii) The convergence of the sequence $\{L_{\sigma}{}^{m} c\}$ will be established now for $\sigma \in [0, \sigma_0]$ and $c \in \Omega$. Since it converges for $\sigma \in (\sigma_0, \infty)$, according to part (i) of this proof it will converge for $\sigma = \sigma_0$, and for $\sigma = \sigma_1$ less than σ_0 and sufficiently near to it. It will also converge for $\sigma_2 < \sigma_1$, and near enough to it, etc.; in this way, the whole interval $[0, \sigma_0]$ can be covered, and Theorem 3 is proved. In particular, the convergence of the sequence $\{L_0{}^{m} c\}$ for any $c \in \Omega$ is established.

At last, it remains to be proved that the limit of the sequence $\{L_0^m c\}$ is a fixed point of the operator L_0. A lemma will be proved beforehand.

LEMMA 2. *The operator L_σ is continuous at $c = c_1 \in \Omega$ for $\sigma \in [0, \infty)$.*

Proof. If c_1 and c_2 are in Ω, it has to be established that, given an $\epsilon > 0$, a $\delta(\epsilon)$ can be found such that if $\| c_1 - c_2 \| < \delta(\epsilon)$, then $\| L_\sigma c_1 - L_\sigma c_2 \| < \epsilon$. For $\sigma \neq 0$, this follows from (17). For $\sigma = 0$, continuity can be proved in the following way.

Since $L_\sigma c$ is a continuous function of σ for any $c \in \Omega$, given $\epsilon_1 > 0$ ($\epsilon_2 > 0$), a δ_1 (a δ_2) can be found such that if $\sigma < \delta_1$ ($\sigma < \delta_2$), $\| L_0 c_1 - L_\sigma c_1 \| < \epsilon_1$ ($\| L_0 c_2 - L_\sigma c_2 \| < \epsilon_2$). Since L_σ is continuous at $c = c_1$ for $\sigma \neq 0$, given $\epsilon_3 > 0$, a δ_3 can be found such that if $\| c_1 - c_2 \| < \delta_3$, $\| L_\sigma c_1 - L_\sigma c_2 \| < \epsilon_3$. Then

$$\| L_0 c_1 - L_0 c_2 \| \leqq \| L_0 c_1 - L_\sigma c_1 \| + \| L_\sigma c_2 - L_0 c_2 \| + \| L_\sigma c_1 - L_\sigma c_2 \|$$
$$< \epsilon_1 + \epsilon_3 + \epsilon_3 ,$$

if $\sigma < \min(\delta_1, \delta_2)$ and $\| c_1 - c_2 \| < \delta_3$. Then, L_0 is continuous at any $c_1 \in \Omega$.

THEOREM 4. *The limit $\hat{c}$ of the sequence $\{L_0^m c\}$, with $c \in \Omega$, is a fixed point of the operator L_0; that is, it satisfies the equality $L_0 \hat{c} = \hat{c}$.*

Proof. This theorem is a well-known [8] consequence of Lemma 2, and its proof will be omitted.

The claims made in §2 have then been justified. Due to Theorem 1, if a control exists that minimizes $S_{\hat{c}}(u)$, with $\hat{c}$ a fixed point of the operator L_0, it also minimizes $\| x(t_f, u) \|$. As it was indicated in §3, this control exists under the appropriate conditions on $A(t)$ and $B(t)$, and Theorem 4 indicates how to compute a fixed point $\hat{c}$ and then the control that minimizes $\| x(t_f, u) \|$. Existence properties for this control have then been determined quite simply; the method is also characterized by its comparatively modest computational requirements.

Appendix A. A procedure to follow when $\min_{u \in U} \| x(t_f, u) \| = d = 0$. If d is zero, the sequence $\{L_0^m c\}$ might or might not converge. If it does not, and since it converges for $d \neq 0$, this situation can be diagnosed readily. If it does converge to $\hat{c}$, it is not apparent that $u_{\hat{c}}^*$ is the optimum control, and an alternative procedure is required. This situation can be diagnosed by the small values that $\| x(t_f, u) \|$ takes after several iterations; this will also be the case when d is not zero but very small; but it will also be convenient to apply the procedure to follow to avoid dividing very small numbers by very small numbers when computing $x(t_f, u)/\| x(t_f, u) \|$ at an advanced stage of the iteration.

The procedure consists of a change in the origin of the state space; put $z = x + a$ with a a constant vector different from zero. Then, if $\| x(t_f, u^*) \|$

$= 0$ and since the optimal control u^* does not change under the change in the origin of the state space, $\min_{u \in U} \| z(t_f, u) \| = \| z(t_f, u^*) \| = \| a \|$, which can be made of a reasonable size.

The general procedure can be applied in this case, because (1) becomes

$$\dot{z} = A(t)z + B(t)u + f(t), \tag{21}$$

where $f(t) = -A(t)a$; the conclusions of §3 remain unchanged, since (6) and (7) do not vary, and the fact that $m(E_k)$ equals zero, $k = 1, \cdots, r$, remains true; for a system like (21) this will be the case if, besides satisfying conditions (i) and (ii) of §3, the functions $f_i(t)$, $i = 1, \cdots, n$, have one continuous derivative [2], which is true for system (21). The conclusions of §4 are also valid when system (1) is replaced by system (21), since (16) is maintained with a different meaning for the constant η.

Appendix B. A useful inequality. It is clear that, for any two vectors x_{f1} and x_{f2} such that $\| x_{f1} \| \neq 0$, $\| x_{f2} \| \neq 0$,

$$\left\| \frac{x_{f1}}{\| x_{f1} \|} - \frac{x_{f2}}{\| x_{f2} \|} \right\| = \frac{1}{\| x_{f1} \|} \| x_{f1} - \gamma x_{f2} \|,$$

where $\gamma = \| x_{f1} \| / \| x_{f2} \|$. Without loss of generality, assume that $\gamma \leqq 1$. Then,

$$\begin{aligned} \| x_{f1} - x_{f2} \|^2 - \| x_{f1} - \gamma x_{f2} \|^2 &= (1 - \gamma^2) \| x_{f2} \|^2 + 2(\gamma - 1)(x_{f1}, x_{f2}) \\ &\geqq (1 - \gamma^2) \| x_{f2} \|^2 + 2(\gamma - 1) \| x_{f1} \| \| x_{f2} \| \\ &= \| x_{f2} \|^2 (\gamma - 1)^2 \geqq 0. \end{aligned}$$

That is,

$$\| x_{f1} - \gamma x_{f2} \| \leqq \| x_{f1} - x_{f2} \|,$$

or

$$\left\| \frac{x_{f1}}{\| x_{f1} \|} - \frac{x_{f2}}{\| x_{f2} \|} \right\| \leqq \frac{1}{\| x_{f1} \|} \| x_{f1} - x_{f2} \|. \tag{22}$$

REFERENCES

[1] Y. C. Ho, *A successive approximation technique for optimal control systems subject to input saturation*, Trans. ASME Ser. D. J. Basic Engrg., 84D (1962), pp. 33–40.

[2] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[3] P. S. Fancher, *Iterative computation procedures for an optimum control problem*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 346–348.

[4] J. P. La Salle, *The time-optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. 5, Princeton University Press, Princeton, 1960, pp. 1–24.
[5] L. I. Rozonoer, *L. S. Pontryagin maximum principle in the theory of optimum systems. II*, Automat. Remote Control, 20 (1959), pp. 1405–1421.
[6] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
[7] L. V. Kantarovich and G. P. Akilov, *Functional Analysis in Normed Spaces*, Pergamon Press, Oxford, 1964.
[8] B. Z. Vulikh, *Functional Analysis for Scientists and Technologists*, Pergamon Press, Oxford, 1963, §4.3.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

NECESSARY CONDITIONS FOR SINGULAR EXTREMALS INVOLVING MULTIPLE CONTROL VARIABLES*

B. S. GOH†

Abstract. New necessary conditions are obtained from the second variation, via a transformed accessory minimum problem, for an important class of singular Bolza problems, which includes most of the singular optimal control problems that have been studied in recent years. This set of necessary conditions is a generalization of the classical Clebsch (Legendre) necessary condition. It is in a form easily used. For problems with multiple control variables, it is required that a certain matrix be symmetric; and if this symmetry property is satisfied, it then requires another matrix to be positive semidefinite. The positive semidefiniteness of the diagonal terms of the latter matrix imposes the same conditions as those obtained by other authors (Kelley, Kopp and Moyer). Should this generalized Clebsch condition be satisfied only in a semidefinite manner, then another similar set of necessary conditions can deduced, and so on.

Three examples are studied. Firstly, we impose new necessary conditions on the variable thrust arcs of a rocket moving in a resisting medium in a vertical plane of a flat earth with two degrees of freedom, namely, the lift and thrust programs. Secondly, the doubly singular arcs of a problem in interplanetary guidance, formulated by Breakwell, are shown to be nonoptimal. Thirdly, a simple optimality condition is deduced for a class of identically singular optimal control problems, certain members of which were previously studied by Haynes, using an extension of the Green's Theorem approach to higher dimension.

1. Introduction. Recently this author [1] laid down a procedure by which a generalized Clebsch necessary condition can be deduced for singular arcs of the general Bolza problem as formulated by Bliss [2, p. 189]. In this procedure we first eliminate some derivatives of variations from the accessory minimum problem, so that the status of these variations can be, and is, raised to that of derivatives. This is followed by the application of the classical Clebsch condition to the transformed accessory minimum problem, leading to a generalized Clebsch condition for the original Bolza problem.

The repeated indices summation convention will be employed throughout this paper, unless otherwise stated, and the following set of indices will be used:

$$
\begin{aligned}
& i, j = 1, 2, \cdots, n; && r, s = 1, 2, \cdots, m; \\
& \alpha, \beta = 1, 2, \cdots, m^* < m; && \rho, \nu = m^* + 1, m^* + 2, \cdots, m; \\
& \mu = 1, 2, \cdots, N \leqq n + 1. &&
\end{aligned} \tag{1}
$$

* Received by the editors May 16, 1966, and in final form July 8, 1966.

† Department of Mathematics, University of Canterbury, Christchurch, New Zealand.

2. A class of Bolza problems. In current optimal rocket trajectories and optimal control problems the following Bolza problem is of fundamental importance: Find the control functions $u_r(t)$ which minimize the performance index

$$J \equiv g[x(t_1), t_1] + \int_{t_0}^{t_1} f(x, u, t)\, dt, \tag{2}$$

where the functions $x_i(t)$, $u_r(t)$ are subject to the conditions

$$\dot{x}_i = f_i(x, u, t), \tag{3}$$

$$x_i(t_0) = x_{i0} \quad \text{(constants)}, \tag{4}$$

$$\psi^\mu[x(t_1), t_1] = 0, \tag{5}$$

$$a_r \leqq u_r \leqq b_r \quad (a_r, b_r \quad \text{constants}). \tag{6}$$

However, unless otherwise stated we shall assume that the control vector u_r belongs to an open region U, i.e., $a_r < u_r < b_r$. After deriving the main results we shall outline the necessary modifications to include the cases where some of the u_r's are equal to the corresponding a_r's or b_r's.

This problem becomes an equivalent Bolza problem as formulated by Bliss [2] if we introduce m auxiliary variables $z_r(t)$ and transform the problem by eliminating the u_r's using the equations

$$u_r = \dot{z}_r, \qquad z_r(t_0) = 0, \tag{7}$$

and letting $z_r(t_1)$ be arbitrary. The initial values of z_r have been put equal to zero for definiteness and this step is of no consequence. Henceforth it will be assumed that this transformation has been carried out, but whenever convenient we retain the use of the variables u_r.

Assuming that the problem is *normal*, Bliss's first order necessary conditions [2, p. 214] for an arc without corners to be minimizing are:

(i) *The Euler-Lagrange equations.*

$$\dot{\lambda}_i = -\frac{\partial H}{\partial x_i}, \tag{8}$$

$$\frac{\partial H}{\partial u_r} = 0, \tag{9}$$

where

$$H(t, x, u, \lambda) = f(x, u, t) + \lambda_i f_i(x, u, t). \tag{10}$$

(ii) *The transversality condition.*

$$dg + e^\mu\, d\psi^\mu + H_1\, dt_1 - \lambda_{i1}\, dx_{i1} = 0 \tag{11}$$

for all arbitrary values of dx_{i1} and dt_1. The constants e^μ are Lagrange multipliers.

(iii) *The Weierstrass condition.*

$$H(t, x, u^*, \lambda) \geqq H(t, x, u, \lambda) \tag{12}$$

for all admissible u_r^*, i.e., $(u_r^*) \in U$, and for every element (t, x, u, λ) of the extremal under examination.

(iv) *The classical Clebsch condition.*

$$\frac{\partial^2 H}{\partial u_r \, \partial u_s} \pi_r \, \pi_s \geqq 0 \tag{13}$$

for all arbitrary values of π_r.

For this discussion it is important to consider this classical Clebsch condition as a first order condition. It is in fact a direct consequence of the Weierstrass condition [2, p. 224]. Originally this Clebsch condition was obtained via the second variation. It would then be meaningless to apply the Clebsch condition to the accessory minimum problem. The corner (junction) conditions do not arise because of the assumption that the reference arc is smooth.

3. The singular accessory minimum problem. Let the $(n + m)$ functions $y_i(t)$, $v_r(t)$ be the variations of $x_i(t)$ and $z_r(t)$ along the minimizing arc. The second variation of the functional J along the minimizing arc is expressible in the form [2, p. 227]

$$J_2 \equiv 2\gamma[\xi_1 , y(t_1)] + \int_{t_0}^{t_1} 2\omega(t, y, \dot{v}) \, dt, \tag{14}$$

in which $2\gamma[\xi_1 , y(t_1)]$ is a homogenous quadratic form in its arguments and

$$2\omega = \frac{\partial^2 H}{\partial u_r \, \partial u_s} \dot{v}_r \dot{v}_s + 2 \frac{\partial^2 H}{\partial u_r \, \partial x_i} \dot{v}_r y_i + \frac{\partial^2 H}{\partial x_i \, \partial x_j} y_i y_j . \tag{15}$$

The equations of variation are

$$\dot{y}_i - \frac{\partial f_i}{\partial u_r} \dot{v}_r - \frac{\partial f_i}{\partial x_j} y_j = 0, \tag{16}$$

$$y_i(t_0) = 0, \qquad v_r(t_0) = 0, \tag{17}$$

$$\Psi^\mu[\xi_1 , y(t_1)] = 0, \tag{18}$$

where the $v_r(t_1)$ are arbitrary and where the $\Psi^\mu[\xi_1 , y_i(t_1)]$ stand for N sets of linear homogenous forms in its arguments.

In matrix notation, 2ω and the constraints in (16) have the forms

$$2\omega \equiv \dot{\eta}^T R\dot{\eta} + 2\dot{\eta}^T Q\eta + \eta^T P\eta, \tag{19}$$

$$\phi\dot{\eta} + \theta\eta = 0, \tag{20}$$

where

$$\eta_i \equiv y_i\,, \qquad \eta_{n+r} \equiv v_r\,, \tag{21}$$

and where R, Q, P are $(n + m) \times (n + m)$ order matrices, ϕ, θ are $n \times (n + m)$ order matrices and the superscript T denotes matrix transpose. Thus we have

$$R \equiv \begin{bmatrix} 0 & 0 \\ 0 & \dfrac{\partial^2 H}{\partial u_r\, \partial u_s} \end{bmatrix}, \tag{22}$$

$$Q \equiv \begin{bmatrix} 0 & 0 \\ \dfrac{\partial^2 H}{\partial u_r\, \partial x_i} & 0 \end{bmatrix}, \tag{23}$$

$$P \equiv \begin{bmatrix} \dfrac{\partial^2 H}{\partial x_i\, \partial x_j} & 0 \\ 0 & 0 \end{bmatrix}, \tag{24}$$

$$\phi \equiv [\delta_{ij} \qquad -\partial f_i/\partial u_r], \tag{25}$$

$$\theta \equiv [-\partial f_i/\partial x_j \qquad 0], \tag{26}$$

where the δ_{ij} are the Kronecker deltas and where partition lines run between the nth and $(n + 1)$th rows/columns.

DEFINITION. An extremal arc of a Bolza problem is said to be *singular* if the determinant

$$\Delta \equiv \begin{vmatrix} R & \phi^T \\ \phi & 0 \end{vmatrix} \equiv 0 \tag{27}$$

for all elements (t, x, u, λ) belonging to the extremal [2, p. 207].

With the matrices R, ϕ displayed in (22) and (25),

$$\Delta = (-1)^{n^2} \begin{vmatrix} \dfrac{\partial^2 H}{\partial u_r\, \partial u_s} \end{vmatrix}. \tag{28}$$

This is easily seen by evaluating the determinant Δ by the first n columns followed by the first n rows. Hence for this class of Bolza problems we can say that an extremal is singular if

$$\begin{vmatrix} \dfrac{\partial^2 H}{\partial u_r\, \partial u_s} \end{vmatrix} \equiv 0 \tag{29}$$

along the extremal.

An important class of singular Bolza problems consists of problems in which one or more control variables appear linearly in both (2) and (3) and in which the classical Clebsch condition is satisfied. Thus if the control variable u_m, say, appears linearly, we have

$$\frac{\partial^2 H}{\partial u_m{}^2} \equiv 0. \tag{30}$$

The positive semidefiniteness of the matrix $(\partial^2 H/\partial u_r \partial u_s)$, required by the classical Clebsch condition, implies that all the 2×2 order determinants

$$\begin{vmatrix} \dfrac{\partial^2 H}{\partial u_s{}^2} & \dfrac{\partial^2 H}{\partial u_m\, \partial u_s} \\ \dfrac{\partial^2 H}{\partial u_m\, \partial u_s} & 0 \end{vmatrix} \geqq 0 \tag{31}$$

for $s = 1, 2, \cdots, m-1$. Inequality (31) implies

$$-\left(\frac{\partial^2 H}{\partial u_m\, \partial u_s}\right)^2 \geqq 0, \tag{32}$$

i.e.,

$$\frac{\partial^2 H}{\partial u_m\, \partial u_s} \equiv 0. \tag{33}$$

Hence the $m \times m$ order matrix $(\partial^2 H/\partial u_r \partial u_s)$ has the form

$$\left[\frac{\partial^2 H}{\partial u_r\, \partial u_s}\right] \equiv \begin{bmatrix} R_1 & 0 \\ 0 & 0 \end{bmatrix}, \tag{34}$$

where R_1 is an $(m-1) \times (m-1)$ order matrix. Hence the extremal is singular, from (29).

In a similar manner when $(m - m^*)$ control variables appear linearly in both (2) and (3) and when the classical Clebsch condition is satisfied, the $m \times m$ order matrix $(\partial^2 H/\partial u_r \partial u_s)$ is expressible in the form displayed in (34), where R_1 is then an $m^* \times m^*$ order matrix. Most of the singular Bolza problems of optimal rocket trajectories and optimal control belong to this important class.

For other singular extremals of the Bolza problems formulated in the last section we may first have to subject the accessory minimum problem to a certain regular linear transformation [1, (71)] in order to assume that the $m \times m$ order matrix $(\partial^2 \omega/\partial \dot{\eta}_{n+r} \partial \dot{\eta}_{n+s})$ can be partitioned thus:

$$2\left[\frac{\partial^2 \omega}{\partial \dot{\eta}_{n+r}\, \partial \dot{\eta}_{n+s}}\right] = \begin{bmatrix} R_1 & 0 \\ 0 & 0 \end{bmatrix}, \tag{35}$$

where partition lines run between the m^*th and $(m^* + 1)$th rows/columns.

The matrix R_1 should be nonsingular or nonexisting so that we can deduce the most effective generalized Clebsch condition. This can be brought about by the preliminary linear transformation just mentioned.

Assuming that the partition displayed in (35) is valid for a Bolza problem, we have from (22)

$$R \equiv \begin{bmatrix} 0 & 0 & 0 \\ 0 & R_1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{36}$$

where partition lines run between the nth and $(n+1)$th rows/columns and the $(n+m^*)$th and $(n+m^*+1)$th rows/columns. The matrices Q, P, ϕ, θ are then partitioned in a similar manner giving

$$Q \equiv \begin{bmatrix} 0 & 0 & 0 \\ Q_1 & 0 & 0 \\ Q_2 & 0 & 0 \end{bmatrix}, \tag{37}$$

$$P \equiv \begin{bmatrix} P_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{38}$$

$$\phi \equiv [I_n \quad -B_1 \quad -B_2], \tag{39}$$

$$\theta \equiv [-A \quad 0 \quad 0]. \tag{40}$$

We note that the zero submatrices in (37) to (40) occur because of the form in which the Bolza problem is formulated, namely the introduction of control variables.

FUNDAMENTAL THEOREM. *If the matrices R, Q, P, ϕ, θ of an accessory minimum problem can be partitioned in the manner displayed in* (36) *to* (40), *then for each element of a minimizing singular extremal of the Bolza problem, the following conditions are necessary:*

(i) *The $(m - m^*) \times (m - m^*)$ order matrix Q_2B_2 must be identically symmetric.*

(ii) *If Q_2B_2 is identically symmetric, then the matrix*

$$\begin{bmatrix} R_1 & R_2{}^T \\ R_2 & R_3 \end{bmatrix} \equiv R_4, \text{ say,} \tag{41}$$

must be positive semidefinite, where

$$R_2 \equiv B_2{}^T Q_1{}^T - Q_2 B_1, \tag{42}$$

$$R_3 \equiv B_2{}^T P_1 B_2 - \frac{d}{dt} Q_2 B_2 - Q_2 B_3 - B_3{}^T Q_2{}^T, \tag{43}$$

and

$$B_3 \equiv AB_2 - B_2. \tag{44}$$

Proof. We shall first prove condition (ii), that is, we first assume Q_2B_2 is symmetric. As laid down in Goh [1, (75)] the variation vector η is subjected to the regular transformation

$$\eta = V\zeta, \tag{45}$$

where

$$V \equiv \begin{bmatrix} I_n & 0 & B_2 \\ 0 & I_{m^*} & 0 \\ 0 & 0 & I_{m-m^*} \end{bmatrix}. \tag{46}$$

Under such a transformation 2ω and the constraints (16) retain the forms displayed in (19) and (20) but with

$$R \equiv \begin{bmatrix} 0 & 0 & 0 \\ 0 & R_1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{47}$$

$$Q \equiv \begin{bmatrix} 0 & 0 & 0 \\ Q_1 & 0 & Q_1 B_2 \\ Q_2 & 0 & Q_2 B_2 \end{bmatrix}, \tag{48}$$

$$P \equiv \begin{bmatrix} P_1 & 0 & P_1 B_2 \\ 0 & 0 & 0 \\ B_2{}^T P_1 & 0 & B_2{}^T P_1 B_2 \end{bmatrix}, \tag{49}$$

$$\phi \equiv [I_n \quad -B_1 \quad 0], \tag{50}$$

$$\theta \equiv [-A \quad 0 \quad -AB_2 + \dot{B}_2]. \tag{51}$$

Let

$$\zeta^T \equiv [\tau^T \quad \sigma^T \quad \kappa^T], \quad \text{say.} \tag{52}$$

In the transformed accessory minimum problem the derivated function $\dot{\kappa}_\rho(t)$ occurs only in the bilinear forms

$$2\dot{\kappa}^T Q_2 \tau \quad \text{and} \quad 2\dot{\kappa}^T Q_2 B_2 \kappa. \tag{53}$$

The assumption that Q_2B_2 is symmetric leads to

$$\int_{t0}^{t_1} 2\dot{\kappa}^T Q_2 B_2 \kappa \, dt = \kappa^T Q_2 B_2 \kappa \Big|_{t_0}^{t_1} - \int_{t_0}^{t_1} \kappa^T \frac{d}{dt}(Q_2 B_2)\kappa \, dt, \tag{54}$$

and in general,

$$\int_{t_0}^{t_1} 2\dot{\kappa}^T Q_2 \tau \, dt = 2\kappa^T Q_2 \tau \Big|_{t_0}^{t_1} - \int_{t_0}^{t_1} 2\kappa^T \frac{d}{dt}(Q_2 \tau) \, dt. \tag{55}$$

Employing the matrix equation (20) with the appropriate matrices displayed in (50) to (52), we have

$$
\begin{aligned}
-2\kappa^T \frac{d}{dt}(Q_2\tau) &= -2\kappa^T\dot{Q}_2\tau - 2\kappa^T Q_2\dot{\tau} \\
&= -2\kappa^T\dot{Q}_2\tau - 2\kappa^T\dot{Q}_2 B_1\dot{\sigma} \\
&\quad -2\kappa^T Q_2 A\tau - 2\kappa^T Q_2(AB_2 - \dot{B}_2)\kappa.
\end{aligned} \tag{56}
$$

Hence (54) to (56) permit the $\dot{\kappa}_\rho(t)$ terms to be completely eliminated from the second variation. The equation of variation (20) with matrices ϕ, θ displayed in (50) and (51), is also devoid of the terms $\dot{\kappa}_\rho(t)$. Hence the $\dot{\kappa}_\rho(t)$ terms are completely eliminated from the accessory minimum problem. Thus the status of the κ_ρ terms can be raised to that of derivatives; this step is taken by replacing κ_ρ with $\dot{\kappa}_\rho{}^*$, say. Following this it is observed that the terms $\dot{\kappa}_\rho{}^*(t_1)$ appear in

$$
2\gamma[\xi_1, \tau_i(t_1), \dot{\kappa}_\rho{}^*(t_1)] \quad \text{and} \quad \Psi^\mu[\xi_1, \tau_i(t_1), \dot{\kappa}_\rho{}^*(t_1)]. \tag{57}
$$

As shown in [1], the terms $\dot{\kappa}_\rho{}^*(t_1)$ are treated like parameters such as ξ_1, and henceforth we let ξ represent ξ_1 and $\dot{\kappa}_\rho{}^*(t_1)$.

After rearrangements, 2ω and the differential constraints of the accessory minimum problem remain in the forms displayed in (19) and (20) but with

$$
R \equiv \begin{bmatrix} 0 & 0 & 0 \\ 0 & R_1 & R_2{}^T \\ 0 & R_2 & R_3 \end{bmatrix}, \tag{58}
$$

$$
Q \equiv \begin{bmatrix} 0 & 0 & 0 \\ Q_1 & 0 & 0 \\ Q_3 & 0 & 0 \end{bmatrix}, \tag{59}
$$

$$
P \equiv \begin{bmatrix} P_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{60}
$$

$$
\phi \equiv [I_n \quad -B_1 \quad -B_3], \tag{61}
$$

$$
\theta \equiv [-A \quad 0 \quad 0], \tag{62}
$$

where

$$
R_2 \equiv B_2{}^T Q_1{}^T - Q_2 B_1, \tag{63}
$$

$$
R_3 \equiv B_2{}^T P_1 B_2 - \frac{d}{dt} Q_2 B_2 - Q_2 B_3 - B_3{}^T Q_2{}^T, \tag{64}
$$

$$
Q_3 \equiv B_2{}^T P_1 - Q_2 A - \dot{Q}_2, \tag{65}
$$

$$
B_3 \equiv AB_2 - \dot{B}_2. \tag{66}
$$

Applying the classical Clebsch condition to the accessory extremal $\xi = 0, \eta \equiv 0$ of the transformed accessory minimum problem arrived at in the last paragraph, we deduce the following necessary condition for the original Bolza problem: Along a minimizing arc

$$\pi^T R \pi \geqq 0 \tag{67}$$

for all $(n + m) \times 1$ order matrices π satisfying

$$\phi\pi = 0, \tag{68}$$

where the matrices R, ϕ are displayed in (58) and (61). Taking $\pi_{n+1}, \pi_{n+2}, \cdots, \pi_{n+m}$ to be arbitrary, condition (67) subjected to (68) is seen to imply that the $m \times m$ order matrix

$$\begin{bmatrix} R_1 & R_2^T \\ R_2 & R_3 \end{bmatrix} \tag{69}$$

must be positive semidefinite, along a minimizing singular arc. Hence we have proved condition (ii) of the Fundamental Theorem.

We shall now prove condition (i) of the Fundamental Theorem. If Q_2B_2 is not symmetric, we introduce new partition lines running between the $(n + m^{**})$th and $(n + m^{**} + 1)$th rows/columns where $m^* < m^{**} \leqq m - 1$. Then

$$R \equiv \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & R_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{70}$$

$$Q \equiv \begin{bmatrix} 0 & 0 & 0 & 0 \\ Q_1 & 0 & 0 & 0 \\ Q_2^* & 0 & 0 & 0 \\ Q_2^{**} & 0 & 0 & 0 \end{bmatrix}, \text{ say,} \tag{71}$$

$$P \equiv \begin{bmatrix} P_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{72}$$

$$\phi \equiv [I_n \ -B_1 \ -B_2^* \ -B_2^{**}], \text{ say,} \tag{73}$$

$$\theta \equiv [-A \quad 0 \quad 0 \quad 0]. \tag{74}$$

As before partition lines run also between the nth and $(n + 1)$th rows/columns.

Furthermore, m^{**} is chosen such that the $(m - m^{**}) \times (m - m^{**})$ order matrix $Q_2^{**}B_2^{**}$ is symmetric and is of the highest possible order.

With m^{**} chosen in this manner, the nonsymmetry of Q_2B_2 implies that

$$Q_2{}^{**}B_2{}^{*} - B_2{}^{**T}Q_2{}^{*T} \not\equiv 0. \tag{75}$$

This is easily seen on examining

$$Q_2 B_2 \equiv \begin{bmatrix} Q_2{}^{*}B_2{}^{*} & Q_2{}^{*}B_2{}^{**} \\ Q_2{}^{**}B_2{}^{*} & Q_2{}^{**}B_2{}^{**} \end{bmatrix}. \tag{76}$$

If $Q_2{}^{**}B_2{}^{*} - B_2{}^{**T}Q_2{}^{*T} \equiv 0$, then the order of the symmetric matrix $Q_2{}^{**}B_2{}^{**}$ can be increased by at least one. The reason for this is that the matrix consisting of $Q_2{}^{**}B_2{}^{**}$ and any one column of $Q_2{}^{**}B_2{}^{*}$ and the corresponding row of $Q_2{}^{*}B_2{}^{**}$ and the corresponding diagonal element of $Q_2{}^{*}B_2{}^{*}$ is symmetric. Finally we note that if $Q_2{}^{**}B_2{}^{**}$ is a 1×1 order matrix it is trivially symmetric.

As $Q_2{}^{**}B_2{}^{**}$ is symmetric, condition (ii) of the Fundamental Theorem is applicable to the accessory minimum problem with matrices R, Q, P, ϕ, θ partitioned in the manner displayed in (70) to (74). Thus we are led to the condition that the matrix

$$\begin{bmatrix} R_1 & 0 & R_2{}^{*T} \\ 0 & 0 & R_2{}^{**T} \\ R_2{}^{*} & R_2{}^{**} & R_3 \end{bmatrix} \equiv R_4\text{, say,} \tag{77}$$

must be positive semidefinite where

$$\begin{aligned} [R_2{}^{*}R_2{}^{**}] &\equiv B_2{}^{**T}[Q_1{}^{T}Q_2{}^{*T}] - Q_2{}^{**}[B_1B_2{}^{*}] \\ &= (B_2{}^{**T}Q_1{}^{T} - Q_2{}^{**}B_1 \quad B_2{}^{**T}Q_2{}^{*T} - Q_2{}^{**}B_2{}^{*}). \end{aligned}$$

Therefore

$$R_2{}^{*} \equiv B_2{}^{**T}Q_1{}^{T} - Q_2{}^{**}B_1\text{,} \tag{78}$$

and

$$R_2{}^{**} \equiv B_2{}^{**T}Q_2{}^{*T} - Q_2{}^{**}B_2{}^{*}. \tag{79}$$

Now we shall show that $R_2{}^{**}$ must be a zero matrix. The positive semidefiniteness of the matrix R_4 of (77) implies that all the 2×2 order determinants of the form

$$\begin{vmatrix} 0 & (R_2{}^{**})_{pq} \\ (R_2{}^{**})_{pq} & (R_3)_{pp} \end{vmatrix} \tag{80}$$

(p not summed) are greater than or equal to zero. Hence $-(R_2{}^{**})_{pq}^{2} \geqq 0$. Therefore $(R_2{}^{**})_{pq} \equiv 0$, i.e.,

$$R_2{}^{**} \equiv 0. \tag{81}$$

But from (75) and (79), condition (81) is not satisfied. Hence matrix Q_2B_2 must be symmetric.

COROLLARY 1. *If the matrix Q_2B_2 is symmetric, then the positive semidefiniteness of the matrix R_4 of* (41) *implies that the diagonal terms of the matrix R_3 of* (43) *must be greater than or equal to zero.*

It can be shown that these diagonal terms (r not summed) are

$$(82)\qquad \begin{aligned}(R_3)_{rr} \equiv{}& \frac{\partial f_i}{\partial u_r}\frac{\partial^2 H}{\partial x_i\,\partial x_j}\frac{\partial f_j}{\partial u_r} - \frac{d}{dt}\left(\frac{\partial^2 H}{\partial u_r\,\partial x_i}\frac{\partial f_i}{\partial u_r}\right)\\ &- 2\frac{\partial^2 H}{\partial u_r\,\partial x_i}\left(\frac{\partial f_i}{\partial x_j}\frac{\partial f_j}{\partial u_r} - \frac{d}{dt}\frac{\partial f_i}{\partial u_r}\right) \geqq 0\end{aligned}$$

for $r = m^* + 1, m^* + 2, \cdots, m$. This is equivalent to the result first obtained by Kelley [3], [4].

COROLLARY 2. *If Q_2B_2 is identically symmetric and $R_2 \equiv 0$, $R_3 \equiv 0$, then the following conditions are necessary for each element of the singular extremal of the original Bolza problem:*

(i) *The $(m - m^*) \times (m - m^*)$ order matrix Q_3B_3 must be identically symmetric. The matrices Q_3 and B_3 are displayed in* (65) *and* (66).

(ii) *If Q_3B_3 is identically symmetric, the matrix*

$$(83)\qquad \begin{bmatrix} R_1 & R_{2,1}^T \\ R_{2,1} & R_{3,1}\end{bmatrix}$$

must be positive semidefinite, where

$$(84)\qquad R_{2,1} \equiv B_3{}^T Q_1{}^T - Q_3 B_1\,,$$

$$(85)\qquad R_{3,1} \equiv B_3{}^T P_1 B_3 - \frac{d}{dt}(Q_3 B_3) - Q_3 B_4 - B_4{}^T Q_3{}^T,$$

$$(86)\qquad B_4 \equiv AB_3 - \dot{B}_3\,.$$

There exists a series of similar necessary conditions involving the symmetry of Q_kB_k (k not summed, $k > 2$) and the positive semidefiniteness of

$$(87)\qquad \begin{bmatrix} R_1 & R_{2,k-2}^T \\ R_{2,k-2} & R_{3,k-2}\end{bmatrix},$$

where

$$(88)\qquad R_{2,k-2} \equiv B_k{}^T Q_1{}^T - Q_k B_1\,,$$

$$(89)\qquad R_{3,k-2} \equiv B_k{}^T P_1 B_k - \frac{d}{dt}Q_k B_k - Q_k B_{k+1} - B_{k+1}^T Q_k{}^T,$$

$$(90)\qquad Q_k \equiv B_{k-1}^T P_1 - Q_{k-1}A - \dot{Q}_{k-1}\,,$$

$$(91)\qquad B_{k+1} \equiv AB_k - \dot{B}_k\,,$$

assuming that $R_{2,k-3}$, $R_{3,k-3}$ vanish identically and $Q_{k-1}B_{k-1}$ is symmetric.

This is a generalization of the series of necessary conditions obtained by Kopp and Moyer [4], [5].

Proof. The matrices R, Q, P, ϕ, θ displayed in (58) to (62) are in the same form as the corresponding matrices displayed in (36) to (40) because R_2, R_3 vanish identically. Moreover the end variations appear in the transformed accessory minimum problem in the same manner as they appeared in the original accessory minimum problem. Hence Corollary 2 can be proved employing the same arguments as those used to prove the Fundamental Theorem. Repeating in this manner the series of necessary conditions may be obtained.

Remarks.

1. In Corollary 2 we have limited ourselves to the special cases where the matrices R_2 and R_3 vanish identically. The more general case where only certain submatrices of R_2 and R_3 vanish identically could have been considered. However this requires introducing many new matrices.

2. The Fundamental Theorem and corollaries remain valid when the control functions $u_r(t)$ are subjected to the inequalities displayed in (6). Using Valentine's device [6], [7], the new multipliers can be absorbed into the matrix R_1. Alternatively we let the variation of any control function $u_r(t)$, attaining its bounds, be zero.

4. Applications.

4.1. Variable thrust arcs for rocket flight in a resisting medium. Consider the extremal variable thrust arcs of a rocket moving in a resisting medium and in a vertical plane of a flat earth. This problem has been studied extensively by Miele [8], [9]. We shall examine the simplified case where the thrust direction is always tangential to the velocity vector and where there are two degrees of freedom, namely, the lift program and the mass flow program. The case of one further degree of freedom, namely, the thrust direction program, becomes unmanageable analytically.

If X denotes a horizontal coordinate, h a vertical coordinate, V the magnitude of velocity, γ the angle between the velocity vector and horizontal direction, m the mass, g the acceleration of gravity, c the equivalent exit velocity of the rocket engine, β the mass flow, D the drag and L the lift, the equations of motion of the rocket are:

$$\dot{X} = V \cos \gamma, \tag{92}$$

$$\dot{h} = V \sin \gamma, \tag{93}$$

$$\dot{V} = -g \sin \gamma - \frac{D - c\beta}{m}, \tag{94}$$

$$\dot{\gamma} = -\frac{g \cos \gamma}{V} + \frac{L}{mV}, \tag{95}$$

$$\dot{m} = -\beta, \tag{96}$$

$$0 \leqq \beta \leqq \beta_{\max}, \tag{97}$$

where

$$D = D(h, V, L) \quad \text{and} \quad g = \text{const.} \tag{98}$$

The problem is to minimize a certain terminal performance index $G(X, h, V, \gamma, m, t)$ subject to conditions (92) to (97) and prescribed end conditions. As only extremal variable thrust arcs are being examined, the constraint (97) is ignored. The control variables are L and β.

Let $\lambda_1, \lambda_2, \cdots, \lambda_5$ be the Lagrange multipliers defined in (10). Then the Euler-Lagrange equations are

$$\dot{\lambda}_1 = 0, \tag{99}$$

$$\dot{\lambda}_2 = \frac{\lambda_3}{m}\frac{\partial D}{\partial h}, \tag{100}$$

$$\begin{aligned}\dot{\lambda}_3 = &-\lambda_1 \cos\gamma - \lambda_2 \sin\gamma + \frac{\lambda_3}{m}\frac{\partial D}{\partial V} \\ &- \frac{\lambda_4 g}{V^2}\cos\gamma + \frac{\lambda_4 L}{mV^2},\end{aligned} \tag{101}$$

$$\begin{aligned}\dot{\lambda}_4 = &\lambda_1 V \sin\gamma - \lambda_2 V \cos\gamma + \lambda_3 g \cos\gamma \\ &- \frac{\lambda_4 g}{V}\sin\lambda,\end{aligned} \tag{102}$$

$$\dot{\lambda}_5 = -\lambda_3 \frac{D - c\beta}{m^2} + \frac{\lambda_4 L}{m^2 V}, \tag{103}$$

$$-\frac{\lambda_3}{m}\frac{\partial D}{\partial L} + \frac{\lambda_4}{mV} = 0, \tag{104}$$

$$\frac{c\lambda_3}{m} - \lambda_5 = 0. \tag{105}$$

Differentiating (105) with respect to time and using (96), (101) and (102), we are led to

$$\begin{aligned}\lambda_1 \cos\gamma + \lambda_2 \sin\gamma = &\frac{\lambda_3}{m}\left(\frac{D}{c} + \frac{\partial D}{\partial V}\right) \\ &- \frac{\lambda_3}{mV}\frac{\partial D}{\partial L}\left(mg\cos\gamma + \frac{L}{c}(V - c)\right).\end{aligned} \tag{106}$$

The Euler-Lagrange equations and (106) imply that the matrices

R_1, R_2, R_3 of the Fundamental Theorem are given by

$$R_1 = -\frac{\lambda_3}{m}\left[\frac{\partial^2 D}{\partial L^2}\right], \tag{107}$$

$$R_2 = -\frac{c\lambda_3}{m^2}\left[\frac{\partial^2 D}{\partial V \partial L} + \frac{1}{V}\frac{\partial D}{\partial L}\right], \tag{108}$$

$$\begin{aligned} R_3 = &-\frac{c^2\lambda_3}{m^3}\left[\frac{\partial^2 D}{\partial V^2} + \frac{2}{c}\frac{\partial D}{\partial V} + \frac{D}{c^2}\right] \\ &- \frac{\lambda_3}{m^3 V^2}\frac{\partial D}{\partial L}\left[2\,mgc^2\cos\gamma - L(2c^2 - 2cV + V^2)\right]. \end{aligned} \tag{109}$$

The positive semidefiniteness of the 2×2 order matrix R_4 of (41) implies

$$R_1 \geqq 0, \tag{110}$$

$$R_3 \geqq 0, \tag{111}$$

and

$$R_1 R_3 - R_2^2 \geqq 0. \tag{112}$$

The inequality condition (110) is contained in the classical Clebsch necessary condition. The inequality (111) can also be obtained by Kelley's test [3], [4]. Inequality (112) is a new optimality condition. Note that the matrix corresponding to $Q_2 B_2$ of the Fundamental Theorem is trivially symmetric because it is a 1×1 order matrix. In general, if singularity is due to only one control variable appearing linearly, the matrix corresponding to $Q_2 B_2$ is trivially symmetric.

4.2. A problem in optimal interplanetary guidance. We shall examine a singular Bolza problem which was formulated by Breakwell [10]. The statement of the problem is: Find control functions $u(t)$ and $r(t)$ so as to minimize the performance index

$$J \equiv \int_0^T \left[2u\sqrt{p} + \alpha r\right] dt \tag{113}$$

with state variables $p(t)$, $q(t)$ satisfying

$$\dot{p} = -2ru p + raq^2, \tag{114}$$

$$\dot{q} = -raq^2, \tag{115}$$

where $\tau = T - t$ with T predetermined, α is a specified constant and $a(t)$ is a known function. We shall not specify the end conditions of p and q but assume that they are such that singular extremals may exist. We shall also

confine our attention to singular arcs involving intermediate levels of both $r(t)$ and $u(t)$.

As displayed in (10), define

$$H \equiv \lambda(raq^2 - 2\tau up) - \mu raq^2 + 2u\sqrt{p} + \alpha r. \tag{116}$$

The Euler-Lagrange equations lead to

$$\dot{\lambda} = 2\lambda\tau u - u/\sqrt{p}, \tag{117}$$

$$\dot{\mu} = 2(\mu - \lambda)raq, \tag{118}$$

$$\lambda\tau\sqrt{p} = 1, \tag{119}$$

$$\alpha - (\mu - \lambda)aq^2 = 0. \tag{120}$$

Using (119) and (120) the matrix corresponding to Q_2B_2 of the Fundamental Theorem is given by

$$Q_2 B_2 \equiv \begin{bmatrix} 2\tau\sqrt{p} & -aq^2/\sqrt{p} \\ 0 & 2\alpha aq \end{bmatrix}; \tag{121}$$

and, in order that Q_2B_2 is identically symmetric,

$$aq^2/\sqrt{p} \equiv 0 \quad \text{implies} \quad q \equiv 0. \tag{122}$$

This rules out the doubly singular extremals involving intermediate levels of both $r(t)$ and $u(t)$.

4.3. A class of identically singular optimal control problems. We shall examine a class of identically singular optimal control problems, certain members of which have been studied by Haynes [11], using an extension of the Green's Theorem approach to higher dimension. The indices displayed in (1) will be employed.

The statement of the problem is: Find control functions $u_r(t)$ which minimize the performance index

$$J \equiv g\,[x(t_1), t_1] + \int_{t_0}^{t_1} L(x, t)\,dt \tag{123}$$

with the state variables $x_i(t)$ satisfying the following system equations and end conditions:

$$\dot{x}_i = A_i(x, t) + B_{ir}(x, t)u_r\,, \tag{124}$$

$$x_i(t_0) = x_{i0}\,, \quad \text{(constants)}, \tag{125}$$

$$\psi^\mu[x(t_1), t_1] = 0, \tag{126}$$

and the vector $u_r \in U$, an open region. Here L, A_i, B_{ir} are functions of x_j and t. As displayed in (10) define

$$H \equiv L + \lambda_i A_i + \lambda_i B_{ir} u_r\,. \tag{127}$$

The Fundamental Theorem is applicable to the problem. In the notation of the Fundamental Theorem the matrices R_1, R_2 do not occur and

$$Q_2 \equiv \left[\frac{\partial^2 H}{\partial u_r \, \partial x_i}\right] = \left[\lambda_j \frac{\partial B_{jr}}{\partial x_i}\right], \tag{128}$$

$$B_2 \equiv \left[\frac{\partial f_i}{\partial u_s}\right] = [B_{is}]. \tag{129}$$

Therefore

$$Q_2 B_2 = \left[\lambda_j B_{is} \frac{\partial B_{jr}}{\partial x_i}\right] = [\pi_{rs}], \quad \text{say.} \tag{130}$$

Thus condition (i) of the Fundamental Theorem requires that the $m \times m$ order matrix $[\pi_{rs}]$ must be identically symmetric along the minimizing singular extremals. This is a rather stringent condition and should be an effective optimality condition for most problems. However if it is satisfied, condition (ii) of the Fundamental Theorem is still available.

Acknowledgment. This research was undertaken for the Ph.D. degree under the supervision of Professor D. F. Lawden.

The author is indebted to the New Zealand Government for the Colombo Plan scholarship which made this research possible.

REFERENCES

[1] B. S. Goh, *The second variation for the singular Bolza problem*, this Journal, 4 (1966), pp. 309–325.

[2] G. A. Bliss, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, 1946.

[3] H. J. Kelley, *A second variation test for singular extremals*, AIAA J., 2 (1964), pp. 1380–1382.

[4] H. J. Kelley, R. E. Kopp and H. G. Moyer, *Singular extremals in optimal control*, Optimization Techniques, vol. 2, G. Leitmann, ed., Academic Press, New York, to be published.

[5] R. E. Kopp and H. G. Moyer, *Necessary conditions for singular extremals*, AIAA J., 3 (1965), pp. 1439–1444.

[6] D. F. Lawden, *Optimal Trajectories for Space Navigation*, Butterworths, Washington, 1963.

[7] L. D. Berkovitz, *Variational methods of control and programming*, J. Math. Anal. Appl., 3 (1961), pp. 145–169.

[8] A. Miele, *The calculus of variations in applied aerodynamics and flight mechanics*, Optimization Techniques, G. Leitmann, ed., Academic Press, New York, 1962.

[9] ———, *General variational theory of the flight paths of rocket powered aircrafts, missiles and satellite carriers*, Astronaut. Acta, 4 (1958), pp. 264–288.

[10] J. V. Breakwell, *A doubly singular problem in optimal interplanetary guidance*, this Journal, 3 (1965), pp. 71–77.

[11] G. W. Haynes, *On the optimality of a totally singular vector control: An extension of the Green's theorem approach to higher dimension*, Final Report, Contract NAS-2-2351, Martin Company, Denver, 1965.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

ON STATIONARY POINTS OF NONLINEAR MAXIMUM-PROBLEMS IN BANACH SPACES*

K. KIRCHGÄSSNER† AND K. RITTER‡

1. Introduction. In this paper various aspects of the nonlinear maximum-problem in a Banach space B are considered. A nonlinear functional F has to be maximized on a bounded or unbounded convex polyhedral region R in a Banach space.

The various types of "stationary points" are investigated and it will be shown that the character of a stationary point depends only on the local properties of the functional F in the intersection of those hyperplanes in which the point lies. The results obtained are generalizations of theorems proved in [6] for the finite-dimensional case, which could be used to develop a method for the solution of the nonconcave quadratic maximum-problem [7].

Some results have been obtained earlier in other connections. Thus M. Altman [1] has given some properties of stationary points and L. Hurwicz [4] has proved the theorem of Kuhn and Tucker for locally convex linear spaces. In this paper, however, this basic theorem is given in a form which is analogous to the finite-dimensional case.

An example, given at the end of §6, shows that the results of this paper may also be applied to problems of control theory.

2. Statement of the problem. The maximum-problem is considered on a real Banach space B whose elements will be denoted by s, x, y. The adjoint space of the continuous linear functionals over B will be denoted by B^* and its elements by l_ν. Furthermore, F will denote a continuous generally nonlinear functional defined over an open set Ω of B containing the domain R which is defined as follows:

$$R = \{x: \quad x \in B,\ l_\nu x \leqq \beta_\nu,\ \nu \in I\},$$

where l_ν are elements of B^*, β_ν are real numbers and I is a given finite set of indices.

Throughout this paper it is assumed that the functional F is differentiable in the sense of Fréchet and that the linear operator F' defines a continuous mapping from B into B^*.

* Received by the editors November 29, 1965, and in final revised form July 18, 1966.

† Institut für angewandte Mathematik und Mechanik der Deutschen Versuchsanstalt für Luft- und Raumfahrt, Freiburg, Hebelstrasse 27, Germany.

‡ Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706.

Then the statement of the problem is as follows: *An element $x_0 \in R$ has to be determined such that F achieves its absolute maximum over R at x_0, that is, $F(x_0) \geqq F(x)$ for all $x \in R$.*

3. Local maximum conditions. For the purpose of this paper an appropriate generalization of the Kuhn-Tucker theorem [6] will be needed which is proved in this section. The proof is based on the following generalization of Farkas' lemma [3].

LEMMA (Farkas). *Let $l \in B^*$, $l \neq 0$, $l_\nu \in B^*$, $\nu \in I$. Furthermore, let $lx \leqq 0$ for all $x \in \tilde{R}$ where $\tilde{R} = \{x: \ x \in B, l_\nu x \leqq 0, \nu \in I\}$. Then there exist real numbers $\lambda_\nu \geqq 0$ such that*

$$l = \sum_{\nu \in I} \lambda_\nu l_\nu .$$

A proof of this lemma for arbitrary linear spaces has been given by Fan [2].

THEOREM (Kuhn-Tucker). (a) *Let $x_0 \in R$ be a local maximum of F in R. Then there exist real numbers $\lambda_\nu \geqq 0$, $\nu \in I$, with*

$$\begin{aligned} F'(x_0) - \sum_{\nu \in I} \lambda_\nu l_\nu &= 0, \\ (l_\nu x_0 - \beta_\nu)\lambda_\nu &= 0, \quad \nu \in I. \end{aligned} \tag{3.1}$$

(b) *Let F be a concave functional on R and let $x_0 \in R$, $\lambda_\nu \geqq 0$, $\nu \in I$, be a solution of* (3.1). *Then F achieves its absolute maximum over R at x_0.*

Proof.

(a) Let x_0 be a relative maximum. If x_0 is an interior point of R, i.e., if $l_\nu x_0 < \beta_\nu$, $\nu \in I$, it follows easily that $F'(x_0) = 0$. Therefore with $x = x_0$ and $\lambda_\nu = 0$, $\nu \in I$, the conditions (3.1) are satisfied.

Now suppose that

$$\begin{aligned} l_\nu x_0 &= \beta_\nu \quad \text{for} \quad \nu \in I_1 , \\ l_\mu x_0 &< \beta_\mu \quad \text{for} \quad \mu \in I - I_1 , \end{aligned}$$

where $I_1 \neq \varnothing$. Furthermore, let s be an arbitrary element of B with the properties

$$\| s \| = 1, \qquad l_\nu s \leqq 0, \qquad \nu \in I_1 .$$

Then $x = x_0 + \tau s$, $0 \leqq \tau \leqq \tau_1$, is an element of R if $\tau_1 > 0$ is sufficiently small. It follows that

$$F(x) = F(x_0) + \tau F'(x_0)s + Q(x_0 , \tau s)$$

with

$$\lim_{\tau \to 0} \frac{1}{\tau} \| Q \| = 0.$$

Therefore $F'(x_0)s \leqq 0$ is valid and the assertion of the theorem follows from Farkas' lemma.

(b) Let $x_0 \in R$, $\lambda_\nu \geqq 0$, $\nu \in I$, be a solution of (3.1). Assume $F(x)$ to be a concave functional over R. Then for all $s \in B$ with $l_\nu s \leqq 0$, $\nu \in I_1$, we have

$$F(x_0 + s) - F(x_0) \leqq F'(x_0)s.$$

Using (3.1) it follows that

$$F(x_0 + s) - F(x_0) \leqq (\sum_{\nu \in I} \lambda_\nu l_\nu)s \leqq 0.$$

4. Properties of stationary points. Since the conditions (3.1) are necessary for an absolute maximum of $F(x)$ in R, the set of solutions of the maximum-problem reduces from R to the set of elements satisfying (3.1). These elements will be called stationary according to the following definition.

DEFINITION. Let $x_0 \in R$ and

$$\begin{aligned} l_\nu x_0 &= \beta_\nu, \quad \nu \in I_1, \\ l_\mu x_0 &< \beta_\mu, \quad \mu \in I - I_1. \end{aligned} \tag{4.1}$$

Then x_0 is called *stationary* if real numbers $\lambda_\nu \geqq 0$, $\nu \in I_1$, exist such that

$$F'(x_0) - \sum_{\nu \in I_1} \lambda_\nu l_\nu = 0.$$

If I_1 is the empty set, x_0 is called a *free* stationary point.

If, for some $\nu \in I_1$, $\lambda_\nu = 0$ holds, it follows immediately from the above definition that those equations in (4.1) corresponding to these values of ν can be canceled without changing the stationary point x_0, i.e., these hyperplanes, even though they are incident in x_0, do not determine the location of the stationary point.

In order to simplify the following discussion of stationary points we assume that the linear functionals l_ν, $\nu \in I_1$, are linearly independent and that $\lambda_\nu > 0$ for each $\nu \in I_1$. Without these assumptions, similar results to those obtained in the following can be derived; but the presentation would be somewhat tedious.

Generally, there may be several stationary points in the intersection of the hyperplanes corresponding to the set I_1. A condition is easily derived under which those stationary points are equivalent.

(1) *Let L_1 be the subspace of B^* generated by the l_ν's, $\nu \in I_1 \subseteq I$, $x(t)$ a Fréchet-differentiable mapping of the real interval $[a, b]$ into B with $l_\nu x(t) = \beta_\nu$, $\nu \in I_1$. Furthermore let*

$$F'[x(\tau)] \in L_1 \quad \textit{for all} \quad \tau \in [a, b].$$

Then

$$F[x(\tau)] = \text{const.} \quad \textit{for all} \quad \tau \in [a, b].$$

The proof follows immediately with the rule of implicit differentiation. We have

$$\frac{dF[x(\tau)]}{d\tau} = F'[x(\tau)] \frac{dx(\tau)}{d\tau}.$$

On the other hand,

$$l_\nu \frac{dx(\tau)}{d\tau} = 0 \quad \text{for} \quad \nu \in I_1,$$

from which the assertion follows.

The following statement gives a sufficient criterion for a stationary point to be isolated.

(2) *Let $x_0 \in R$ be a stationary point with*

$$l_\nu x_0 = \beta_\nu, \qquad \nu \in I_1, \qquad F'(x_0) \in L_1,$$

where L_1 is as in (1). *If the functional F is twice differentiable in the sense of Fréchet and if the bilinear operator $F''(x_0)$ satisfies the condition*

$$| [F''(x_0)y]y | \geqq \epsilon \| y \|^2, \quad \epsilon > 0, \tag{4.2}$$

then there exists a $\delta > 0$ such that for all

$$y \in R_0^\delta = \{y: \quad y \in B, l_\nu y = 0, \nu \in I_1, \| y \| \leqq \delta\}$$

we have $F'(x_0 + y) \notin L_1$, except for $y = 0$.

For the proof we assume that a sequence $\{x_n\}$ exists converging to x_0 with

$$l_\nu x_n = \beta_\nu, \qquad F'(x_n) \in L_1, \qquad \nu \in I_1.$$

Setting $y_n = x_0 - x_n$ yields $l_\nu y_n = 0$, $\nu \in I_1$. Hence $[F'(x_n) - F'(x_0)]y_n = 0$, and therefore

$$\| [F'(x_n) - F'(x_0)]y_n - [F''(x_0)y_n]y_n \| = \| [F''(x_0)y_n]y_n \| = o(\| y_n \|^2).$$

But this is a contradiction to (4.2).

The section concludes with an illustrative interpretation of the "dual variables" λ_ν.

(3) *Let F be Fréchet-differentiable and x_0 and L_1 as in* (2). *For an arbitrary, but fixed $\kappa \in I$, and any $y \in B$ with*

$$\begin{aligned} l_\nu y &= 0, \qquad \nu \in I_1, \qquad \nu \neq \kappa, \\ l_\kappa y &= 1, \end{aligned} \tag{4.3}$$

we define $x(\tau) = x_0 + (\tau - \beta_\kappa)y$. *Then*

$$\frac{dF[x(\beta_\kappa)]}{d\tau} = \lambda_\kappa .$$

To prove this statement we first remark that the linear independence of the l_ν, $\nu \in I_1$, implies the existence of an element y of B such that (4.3) is satisfied [8, p. 138, Theorem 3.5-C]. Now we have

$$\lim_{\tau \to \beta_\kappa} \frac{F[x(\tau)] - F[x(\beta_\kappa)]}{\tau - \beta_\kappa} = \lim_{\tau \to \beta_\kappa} \frac{F[x_0 + (\tau - \beta_\kappa)y] - F(x_0)}{\tau - \beta_\kappa}$$

$$= F'(x_0)y = (\sum_{\nu \in I_1} \lambda_\nu l_\nu)y = \lambda_\kappa .$$

This completes the proof.

5. Classification of stationary points. The main result of this section is the fact that the character of a stationary point depends only on the local properties of the functional $F(x)$ within the intersection of those hyperplanes in which the stationary point is located. It will always be assumed that $F(x)$ is twice differentiable in the sense of Fréchet. Then the following statements are valid:

(1) *Let the stationary point* $x_0 \in R$ *be a relative minimum of* $F(x)$ *in* R. *Then* x_0 *is a free stationary point and* $[F''(x_0)s]s \geqq 0$ *for all* $s \in B$.

Proof. Let the stationary point x_0 be a relative minimum of $F(x)$ in R with

$$l_\nu x_0 = \beta_\nu , \qquad \nu \in I_1 , \qquad I_1 \neq \varnothing .$$

If x_0 is not a free stationary point there exists a $\lambda_\kappa > 0$, $\kappa \in I_1$. As in §4, statement (3), we set $x = x_0 + (\tau - \beta_\kappa)y$ so that for $\tau = \beta_\kappa$,

$$\frac{dF[x(\tau)]}{d\tau} = \lambda_\kappa > 0.$$

Choosing $x \in R$ with $l_\kappa x < \beta_\kappa$, i.e., $\tau < \beta_\kappa$, and $\beta_\kappa - \tau$ sufficiently small, we have $F(x) < F(x_0)$ in contradiction to the assumption that x_0 is a relative minimum.

Hence, x_0 is a free stationary point and $F'(x_0)s = 0$ for all $s \in B$. Therefore, $[F''(x_0)s]s \geqq 0$ follows, i.e., $F(x)$ is a convex functional in the neighborhood of x_0.

(2) *The stationary point* x_0 *satisfying the conditions*

$$F'(x_0) - \sum_{\nu \in I_1} \lambda_\nu l_\nu = 0,$$

$$l_\nu x_0 = \beta_\nu , \qquad \lambda_\nu > 0, \qquad \nu \in I_1 ,$$

is a relative maximum of $F(x)$ *in* R *if for all* $s \in B$ *with* $l_\nu s = 0$, $\nu \in I_1$, *we have* $[F''(x_0)s]s < 0$.

Proof. It is obvious that x_0 is a relative maximum in the intersection of the hyperplanes $l_\nu x = \beta_\nu$, $\nu \in I_1$. Let $x \in R$ but not in this intersection. Let $s = x - x_0$; then there exists at least one $\kappa \in I_1$ with $l_\kappa s < 0$. Therefore

$$F(x) - F(x_0) = F'(x_0)s + o(\| s \|) = [\sum_{\nu \in I_1} \lambda_\nu l_\nu]s + o(\| s \|),$$

from which the assertion follows.

On the other hand, the fact that a stationary point x_0 is a relative maximum of $F(x)$ in R implies $[F''(x_0)s]s \leqq 0$ for all $s \in B$ satisfying $l_\nu s \leqq 0$, $\nu \in I_1$.

(3) *The stationary point* x_0 *is a saddle-point if an* $s_1 \in B$ *exists such that*

$$l_\nu s_1 = 0, \qquad \nu \in I_1 \quad \textit{and} \quad [F''(x_0)s_1]s_1 > 0,$$

provided x_0 *is not a free stationary point.*

This assertion follows immediately from the statements (1) and (2).

6. Applications to a quadratic maximum-problem. In this section we assume that $F(x)$ has the following special form:

$$F(x) = lx - \tfrac{1}{2}(Ax)x,$$

where $l \in B^*$ and A is a bounded linear operator mapping B into B^* with the property that

$$(Ax_1)x_2 = (Ax_2)x_1 \quad \text{for any pair} \quad x_1, x_2 \in B.$$

For this special case some of the results obtained above can be formulated more completely. Furthermore, this case is of particular interest because of its application to quadratic programming.

Because of statement (1) in §4 all stationary points located in the intersection of the same hyperplanes are equivalent in the sense described there. Let x_0 and x_1 be stationary points satisfying

$$\begin{aligned} l_\nu x_0 = l_\nu x_1 &= \beta_\nu, \quad \nu \in I_1 \subseteq I, \\ l - Ax_0 - \sum_{\nu \in I_1} \lambda_\nu^0 l_\nu &= 0, \quad \lambda_\nu^0 > 0, \\ l - Ax_1 - \sum_{\nu \in I_1} \lambda_\nu^1 l_\nu &= 0, \quad \lambda_\nu^1 > 0. \end{aligned}$$

Forming the convex combination

$$\begin{aligned} x &= (1 - \tau)x_0 + \tau x_1, \quad 0 \leqq \tau \leqq 1, \\ \lambda_\nu &= (1 - \tau)\lambda_\nu^0 + \tau\lambda_\nu^1, \quad \nu \in I_1, \end{aligned}$$

it becomes obvious that every x of this combination is a stationary point so that the statement (1) in §4 yields

$$F[x(\tau)] = \text{const.}$$

Let M be a subspace of B. We say A is *positive* in M if $(Ax)x \geqq 0$ for any $x \in M$. If $(Ax)x \leqq 0$ for any $x \in M$, A is called *negative*. With this notation the stationary points can be completely classified for the special case considered in this section:

(1) *The stationary point $x_0 \in R$ is a relative minimum of $F(x)$ if and only if x_0 is a free stationary point and the operator A is negative.*

(2) *Let $x_0 \in R$ be a stationary point satisfying the conditions*

$$\begin{aligned} &(\alpha) \quad l - Ax_0 = \sum_{\nu \in I_1} \lambda_\nu l_\nu , \quad \lambda_\nu > 0, \\ &(\beta) \qquad l_\nu x_0 = \beta_\nu , \qquad \nu \in I_1 \subseteq I. \end{aligned} \tag{6.1}$$

The stationary point x_0 is a relative maximum of $F(x)$ in R if and only if the operator A is positive in the intersection of the hyperplanes $l_\nu x_0 = 0$, $\nu \in I_1$.

(3a) *The stationary point $x_0 \in R$ satisfying conditions* (6.1) *is a saddle-point of $F(x)$ in R if and only if the operator A is not positive in the intersection of the hyperplanes $l_\nu x_0 = 0$, $\nu \in I_1$, provided $I_1 \neq \varnothing$.*

(3b) *If $I_1 \neq \varnothing$, then x_0 is a saddle-point of F if and only if the operator A is neither positive nor negative.*

The proof of these statements follows immediately from the corresponding proofs in §5 invoking the fact that for all $s \in B$ with

$$(As, s) = 0, \qquad l_\nu s = 0, \qquad \nu \in I_1 \subseteq I,$$
$$F(s_0 + \tau s) = F(x_0)$$

for all real τ.

The following example shows that the results of this paper may also be applied to problems of control theory.

Let X be the space of all n-vector functions $x(t)$ which are continuous on the interval $[0, T]$, and let U be the space of all m-vector functions $u(t)$ which are continuous on $[0, T]$. Consider the problem of maximizing

$$\int_0^T g[x(t), u(t)]\, dt \tag{6.2}$$

subject to

$$\dot{x} = A(t)x(t) + B(t)u(t), \qquad x(0) = 0, \tag{6.3}$$

$$\int_0^T a_\nu' x(t)\, dt \leqq \alpha_\nu , \qquad \nu = 1, \cdots, m_1 , \tag{6.4}$$

$$\int_0^T b_\mu' u(t)\, dt \leqq \beta_\mu , \qquad \mu = 1, \cdots, m_2 , \tag{6.5}$$

where $g[x(t), u(t)]$ is a continuously differentiable function on $E^n \times E^m$, while $A(t)$ and $B(t)$ are $n \times n$ and $n \times m$ matrices whose components are continuous functions of t and a_ν and b_μ are n- and m-vectors respectively.

Let $\Phi(t)$ be a solution of $\dot{x} = A(t)x(t)$ such that $\Phi(0) = I$. Then the system (6.3) of linear differential equations has the solution

$$(6.6) \qquad x(t) = \int_0^t \Phi(t)\Phi(s)^{-1}B(s)u(s)\,ds = \Phi(t)\int_0^t C(s)u(s)\,ds.$$

If we introduce (6.6) into (6.2) and (6.4) and use the abbreviations

$$\tilde{g}[u(t)] = g\left[\Phi(t)\int_0^t C(s)u(s)\,ds, u(t)\right],$$
$$\tilde{a}_\nu'(t) = a_\nu'\Phi(t),$$

we obtain (6.2) and (6.4) in the following form:

$$(6.7) \qquad \int_0^T \tilde{g}[u(t)]\,dt,$$

$$(6.8) \qquad \int_0^T \int_0^t \tilde{a}_\nu'(t)C(s)u(s)\,ds\,dt \leqq \alpha_\nu, \qquad \nu = 1, \cdots, m_1.$$

With the norm

$$\| u \| = \sup_{0 \leqq t \leqq T} \| u(t) \|_m$$

($\|\cdot\|_m$ = Euclidean norm in E^m), U becomes a Banach space in which (6.7) can be written in the form $F(u)$, where F is a functional on U, while each of the inequalities (6.5) and (6.8) is of the form $l_\nu u \leqq \alpha_\nu$, where l_ν is a continuous linear functional on U.

REFERENCES

[1] M. Altman, *Stationary points in nonlinear programming*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 12 (1964), pp. 29–35.

[2] K. Fan, *On systems of linear inequalities*, Linear Inequalities and Related Systems, Annals of Mathematical Studies, Princeton University Press, Princeton, 1956, pp. 99–156.

[3] J. Farkas, *Über die Theorie der einfachen Ungleichungen*, J. Reine Angew. Math., 124 (1902), pp. 1–27.

[4] L. Hurwicz, *Programming in linear spaces*, Studies in Linear and Nonlinear Programming, K. J. Arrow, L. Hurwicz, H. Uzawa, eds., Stanford University Press, Stanford, 1958, pp. 38–102.

[5] H. W. Kuhn and A. W. Tucker, *Nonlinear programming*, Proceedings of the Second Berkeley Symposium on Mathematics Statistics and Probability, University of California Press, Berkeley, 1951, pp. 481–492.

[6] K. Ritter, *Stationary points of quadratic maximum-problems*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 4 (1965), pp. 143–158.

[7] ———, *A method for solving maximum-problems with a nonconcave quadratic objective function*, Ibid., 4 (1966), pp. 340–351.

[8] A. E. Taylor, *Introduction to Functional Analysis*, John Wiley, New York, 1964.

J. SIAM CONTROL
Vol. 4, No. 4, 1966
Printed in U.S.A.

AN EXAMPLE CONCERNING ROCKETS CAPABLE OF IMPULSIVE THRUST*

R. W. RISHEL†

Introduction. It is a widely held intuitive idea that properties of a rocket capable of impulsive thrust are in some sense the limiting case of the properties of rockets with bounded thrust as the magnitude of the bound becomes infinite. The purpose of this paper is to show that if the impulsively thrusting rocket is required to satisfy Newton's laws, this intuitive idea becomes questionable.

A rigorous derivation of the equations of motion for a rocket capable of impulsive thrust which satisfies Newton's laws of motion is given in the proof of Theorem 1. Then it is shown, for a constant gravitational field, that there is a solution of these equations which uses an amount of fuel strictly less than the infimum of the amounts of fuel used by bounded thrust rocket trajectories joining the same points.

Rigorous treatments of impulsive thrust rocket problems with different formulations have been given by Ewig [1], Ewig and Haseltine [2], and Neustadt [3].

Equations of Motion. Let the notation $\int_{\tau}^{t}$ denote the integral over the closed interval defined by the limits of integration and $\int_{(\tau,t]}$ denote the integral over a half-open interval. A rocket will be said to be capable of impulsive thrust if the fuel used (mass expelled) by the rocket on the closed interval $[t_0, t]$ is given by the formula

$$\int_{t_0}^{t} R(ds) \tag{1}$$

in which R is a positive measure.

Let $m(t), x(t), v(t), u(t)$ denote the mass, position, velocity and exhaust velocity of the rocket. Suppose the rocket is moving in a gravitational field which causes an acceleration $G(t, x)$ on the rocket. Consider the rocket as an idealized point mass.

THEOREM 1. *Let the velocity and exhaust velocity of a rocket be given by locally integrable functions. Suppose that the motion of the rocket is such that its velocity has a limit from the left $v^{-}(t)$ at each time t. Suppose the rocket is capable of impulsive thrusting, that is, its mass is given by the formula*

* Received by the editors October 20, 1965, and in revised form May 20, 1966.
† Aero-Space Group, The Boeing Company, Seattle, Washington 98124.

$$m(t) = m_0 - \int_{t_0}^{t} R(ds) \tag{2}$$

for some positive measure R. Then, if Newton's laws are satisfied, there is a measure A such that

$$v(t) = v_0 + \int_{t_0}^{t} A(ds), \tag{3}$$

and the equations

$$\int_{t_0}^{t} m(s)A(ds) = \int_{t_0}^{t} m(s)G(s, x(s))\, ds - \int_{t_0}^{t} u(s)R(ds), \tag{4}$$

$$x(t) = x_0 + \int_{t_0}^{t} v(s)\, ds \tag{5}$$

are satisfied at each time t during the flight of the rocket.

Proof. Let $X(t, \tau, x, v)$ and $V(t, \tau, x, v)$ denote the position and velocity at time t of an object which was at position x with velocity v at time τ and was acted upon by the gravitational field during the time from τ to t. Then the total momentum of the rocket and its expended fuel is given by

$$m(t)v(t) + \int_{t_0}^{t} V(t, \tau, x(\tau), v^{-}(\tau) + u(\tau))R(d\tau). \tag{6}$$

The definition of $V(t, \tau, x, v)$ implies that

$$\begin{aligned} &V(t, \tau, x(\tau), v^{-}(\tau) + u(\tau)) \\ &\quad = v^{-}(\tau) + u(\tau) + \int_{\tau}^{t} G(s, X(x, \tau, x(\tau), v^{-}(\tau) + u(\tau)))\, ds. \end{aligned} \tag{7}$$

Newton's law asserts that the change in total momentum is the integral of the external forces applied. Hence,

$$\begin{aligned} &m(t)v(t) + \int_{t_0}^{t} V(t, \tau, x(\tau), v^{-}(\tau) + u(\tau))R(d\tau) - m_0 v_0 \\ &\quad = \int_{t_0}^{t} m(s)G(s, x(s))\, ds \\ &\quad\quad + \int_{t_0}^{t} \int_{t_0}^{s} G(s, X(s, \tau, x(\tau), v^{-}(\tau) + u(\tau)))R(d\tau)\, ds. \end{aligned} \tag{8}$$

After an interchange of order of integration, the last integral of (8) becomes

$$\int_{t_0}^{t} \int_{\tau}^{t} G(s, X(s, \tau, x(\tau), v^{-}(\tau) + u(\tau)))\, dsR(d\tau). \tag{9}$$

Using (7) gives

$$(10)\quad m(t)v(t) - m_0 v_0 = \int_{t_0}^{t} m(s)G(s, x(s))\, ds - \int_{t_0}^{t} (v^-(\tau) + u(\tau))R(d\tau).$$

Equation (10) implies that $m(t)\, v(t)$ is a function of bounded variation. The mass $m(t)$ is defined by (2) and $m(t) > 0$, hence $m(t)^{-1}$ is of bounded variation. Hence, $v(t)$ is also of bounded variation. Reference [4, p. 54, Theorem 11] implies that there is a measure A such that

$$(11)\qquad v(t) = v_0 + \int_{t_0}^{t} A(ds).$$

Consider the formula

$$(12)\quad \int_{t_0}^{t} R(ds) \int_{t_0}^{t} A(ds) = \int_{t_0}^{t} \int_{t_0}^{\tau} R(ds)A(d\tau) + \int_{t_0}^{t} \int_{(\tau,t]} R(ds)A(d\tau).$$

Interchanging the order of integration in the last integral of (12) and using (2) and (11) gives

$$(13)\quad m(t)v(t) - m_0 v_0 = \int_{t_0}^{t} m(\tau)A(d\tau) - \int_{t_0}^{t} v^-(\tau)R(d\tau).$$

Substituting (13) in (10) gives

$$(14)\quad \int_{t_0}^{t} m(s)A(ds) = \int_{t_0}^{t} m(s)G(s, x(s))\, ds - \int_{t_0}^{t} u(s)R(ds).$$

This completes the proof of Theorem 1.

The quantities m_0, v_0, and x_0 in (2), (3), and (5) are initial conditions for the rocket. The rocket will be said to satisfy terminal conditions m_1, v_1, x_1 at time t_1 if $m(t_1) = m_1$, $v(t_1) = v_1$, $x(t_1) = x_1$.

Example. Consider the one-dimensional system whose equations are

$$(15)\qquad \int_0^t m(s)A(ds) = -g \int_0^t m(s)\, dt + \int_0^t R(ds),$$

$$(16)\qquad x(t) = \int_0^t v(s)\, ds,$$

$$(17)\qquad v(t) = \int_0^t A(ds),$$

$$(18)\qquad m(t) = m_0 - \int_0^t R(ds);$$

that is, a rocket in a field with constant gravity g, with unit exhaust velocity in a direction opposite to the gravitational acceleration. Suppose it is desired to guide this rocket from position and velocity $(0; 0)$ at time zero, to

position and velocity (X, V), in which $X > 0$, using a minimum amount of fuel.

Consider the infimum of the fuel used in guiding the rocket when the class of measures R is restricted to those given by bounded functions $r(s)$ for which the initial and terminal conditions are satisfied. This class of measures satisfies the equation

$$\int_0^t R(ds) = \int_0^t r(s)\, ds \tag{19}$$

in which $r(s)$ is a bounded positive function. This infimum is greater than or equal to

$$m_0[1 - e^{-(V^2+2gX)^{\frac{1}{2}}}]. \tag{20}$$

This statement may be deduced from general theorems of [3]. A short direct proof is given below.

If the measure R is of the form (19), equations (15) through (18) can be differentiated with respect to t to obtain the equations

$$m\dot{v} = -mg + r, \tag{21}$$

$$\dot{m} = -r, \tag{22}$$

$$\dot{x} = v. \tag{23}$$

For a given function $r(s)$, let t_1 be the time at which the terminal conditions are satisfied. Then integrating (21) with the integrating factor e^{v+gt} between 0 and t_1 gives

$$m(t_1) = m_0 e^{-(V+gt_1)}. \tag{24}$$

Since $r(s)$ and $m(s)$ are nonnegative, (21) implies

$$\dot{v} + g \geqq 0. \tag{25}$$

Hence, if $s \leqq t_1$,

$$V + g(t_1 - s) \geqq v(s) \tag{26}$$

and

$$Vs + gt_1 s - \tfrac{1}{2}gs^2 \geqq x(s). \tag{27}$$

Setting $s = t_1$ gives

$$Vt_1 + \tfrac{1}{2}gt_1^2 \geqq X. \tag{28}$$

The smallest value of $t_1 > 0$ for which (28) holds is given by

$$t_1 = -g^{-1}[V - (V^2 + 2gX)^{\frac{1}{2}}]. \tag{29}$$

Hence, (24) and (29) imply

$$m_0 - m(t_1) \geqq m_0[1 - e^{-(V^2+2gX)^{\frac{1}{2}}}], \tag{30}$$

which is the assertion to be proved.

Consider the measure

$$R(ds) = m_0(V^2 + 2gX)^{\frac{1}{2}}[1 + (V^2 + 2gX)^{\frac{1}{2}}]^{-1}\delta(s)\,ds, \tag{31}$$

in which $\delta(s)\,ds$ indicates the measure with mass one concentrated at time zero. Since $\int_0^0$ denotes the integral over the closed interval [0, 0], integrating (15) through (18) over the interval [0, 0] gives

$$m(0) = m_0 - \int_0^0 R(ds) = m_0[1 + (V^2 + 2gX)^{\frac{1}{2}}]^{-1}, \tag{32}$$

$$v(0) = \int_0^0 m(0)^{-1}R(ds) = (V^2 + 2gX)^{\frac{1}{2}}, \tag{33}$$

$$x(0) = 0. \tag{34}$$

Since R assigns measure zero to the open interval $(0, \infty)$, it is seen that the values of $x(t)$ and $v(t)$ on the interval $[0, \infty)$ are given by

$$x(t) = -\tfrac{1}{2}gt^2 + (V^2 + 2gX)^{\frac{1}{2}}t, \tag{35}$$

$$v(t) = -gt + (V^2 + 2gX)^{\frac{1}{2}}. \tag{36}$$

Equations (35) and (36) imply the terminal conditions are satisfied at time

$$t_1 = -g^{-1}[V - (V^2 + 2gX)^{\frac{1}{2}}]. \tag{37}$$

Now

$$\int_0^{t_1} R(ds) = m_0(V^2 + 2gX)^{\frac{1}{2}}[1 + (V^2 + 2gX)^{\frac{1}{2}}]^{-1}. \tag{38}$$

Now, (38) is strictly less than the lower bound (20). Hence, the measure R defined by (31) uses an amount of fuel strictly less than the infimum of the fuel used by measures given by bounded functions.

REFERENCES

[1] G. M. Ewing, *A fundamental problem of navigation in free space*, Quart. Appl. Math., 18 (1961), pp. 355–362.

[2] G. M. Ewing and W. R. Haseltine, *Optimal programs for an ascending missile*, this Journal, 2 (1964), pp. 66–88.

[3] L. W. Neustadt, *A general theory of minimum fuel space trajectories*, Ibid., 3 (1965), pp. 317–356.

[4] L. Schwartz, *Theorie des Distributions*, Hermann, Paris, 1951.