SIAM J. CONTROL
Vol. 9, No. 1, February 1971

CONTROLLABILITY AND POLE ASSIGNMENT FOR DISCRETE TIME LINEAR SYSTEMS DEFINED OVER ARBITRARY FIELDS*

S. K. MITTER† AND R. FOULKES‡

1. Introduction. The theory of controllability and observability due to Kalman is certainly one of the most important conceptual contributions to linear systems theory. An account of the development of the ideas of controllability and observability as well as its implications on feedback control theory and realization theory may be found in the recent book of Kalman, Falb and Arbib [1].

It has been known for some time that for a linear continuous, finite-dimensional autonomous system with a scalar control variable, complete controllability is equivalent to being able to assign arbitrary poles to the closed loop transfer matrix by a suitable choice of state variable feedback gain matrix. This result was generalized to the vector control case by Wonham [2] and Simon and Mitter [3]. In [3] constructive recursive algorithms to achieve pole assignment were also presented. The objective of this note is to generalize this result to cover discrete-time, finite-dimensional, autonomous linear systems defined over arbitrary fields. The result can thus be applied to the feedback control of linear sequential machines [4]. By duality arguments the problem of state determination is also solved.

2. Notation and system definition. Let

T = time set = Z = (ordered Abelian group of) integers;
U = input values = F^m = vector space of m-tuples over the field F;
X = state space = F^n;
Y = output space = F^p;
Ω = input space of functions $t \mapsto u(t)$; that is, arbitrary sequences $u(-1), u(0), u(1), \cdots$, with $u(t) \in U$.

We shall be concerned with the discrete-time, autonomous, linear dynamical system Σ defined over a field F,

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) \end{aligned} \tag{2.1}$$

with $t \in Z$, $x(t) \in F^n$, $u(t) \in F^m$, $y(t) \in F^p$ and where

$$\begin{aligned} A &: F^n \to F^n, \\ B &: F^m \to F^n, \\ C &: F^n \to F^p \end{aligned} \tag{2.2}$$

are F-homomorphisms.

* Received by the editors July 10, 1969, and in revised form February 27, 1970. This work was supported in part by the National Science Foundation under Grant GK-3714 at Case Western Reserve University and the National Aeronautics and Space Administration under Grant NGL-22-009-124 at Electronic Systems Laboratory, Massachusetts Institute of Technology.

† Electronic Systems Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, and Systems Research Center, Case Western Reserve University, Cleveland, Ohio 44106.

‡ Systems Research Center, Case Western Reserve University, Cleveland, Ohio 44106.

We shall usually not make a distinction between (A, B) and (A, C) as a pair of F-homomorphisms or as a pair of matrices representing these homomorphisms with respect to a given basis.

With respect to the system (2.1) we make the assumption:

(i) the pair (A, B) is *completely reachable*, that is, the rank of the $n \times nm$ matrix

$$H(A, B) = [B, AB, \cdots, A^{n-1}B] \tag{2.3}$$

is n.

(ii) the pair (A, C) is *completely observable*, that is, the rank of the $n \times np$ matrix

$$K(A, C) = [C^T, A^TC^T, \cdots, (A^T)^{n-1}C^T] \tag{2.4}$$

is n.

3. Statement of main theorem.[1] The principal result of this paper is the following theorem.

THEOREM 3.1. *For the linear autonomous system* (2.1), (A, B) *is a completely reachable pair if and only if for every monic polynomial* g *of degree* n, *there exists an* $m \times n$ *matrix* K *over* F *such that the characteristic polynomial of* $A + BK$ is *precisely* g (*up to a factor of* ± 1).

The proof of the result proceeds via several propositions and is presented in the next section.

4. Proof of main theorem. The proof will be divided into three parts: necessity for the case when B is a column vector, necessity for a general B and sufficiency.

PROPOSITION 4.1 (case $m = 1$). *In* (2.1) *let* $B = b = n \times 1$ *matrix. If* (A, B) *is a completely reachable pair, then there exists a* $1 \times n$ *matrix* k *such that the characteristic polynomial of* $A + bk$ *has an arbitrary preassigned form* (*of degree* n).

The proof of this proposition essentially consists of transforming A to rational canonical form and is identical to the proof given for the field of real numbers (see, for example, [5, Theorems 7 and 9]).

We now consider the case where B is an $n \times m$ matrix.

PROPOSITION 4.2. *If* (A, B) *is a completely reachable pair, then there exists a matrix* K *and a vector* b *such that* $(A + BK, b)$ *is a completely reachable pair and* b *is in the column space of* B.

Proof. The proof presented is essentially the same as independently given by Heymann [6] and hence only an outline of the proof will be given.

Let b_j be the jth column of B and let E_j be the cyclic subspace of the coordinate space $E = F^n$ generated by b_j. Since (A, B) is a completely reachable pair $E = E_1 + \cdots + E_m$. In general, the E_i are not independent, that is, $E_i \cap E_j \neq \varnothing$ for $i \neq j$. However, it is easy to see that there are subspaces S_i and a finite integer t, $0 < 1 \leqq m$, such that $E = S_1 + \cdots + S_t$ and $S_i \cap S_j = \varnothing$ for $i \neq j$, that is, E is

[1] It was pointed out by the reviewer that a similar result has been obtained by R. E. Kalman in the unpublished notes: *Lectures on Controllability and Observability*, CIME Seminar, Italy, February 1969.

a direct sum of the subspaces S_i. A basis for E can now be obtained by combining the bases for the subspaces.

By rearranging the columns of B (hence the coordinates of the control) it can be assumed that the first t columns of B are used. Hence the basis is

$$b_1, \cdots, A^{k_1-1}b_1, \cdots, b_t, \cdots, A^{k_t-1}b_t \quad \text{and} \quad \sum_{i=1}^{t} k_i = n.$$

Let $R = [b_1, \cdots, A^{k_1-1}b_1, \cdots, b_t, \cdots, A^{k_t-1}b_t]$ be the matrix whose columns are the above basis vectors. Clearly R is invertible.

Define an $m \times n$ matrix $S = [s_1 \cdots s_n]$, where each column is an m-tuple defined as follows:

$$s_{r_j} = \varepsilon_{j+1}^{(m)} \quad \text{if} \quad r_j = \sum_{i=1}^{j} k_i \quad \text{and} \quad j = 1, \cdots, t-1; \quad s_j = 0 \quad \text{otherwise,}$$

where $\varepsilon_i^{(m)}$ is the ith standard basis vector of F^m.

Finally, let $P = SR^{-1}$. Clearly $PA^{k_i-1}b_j = \varepsilon_{j+1}^{(m)}$, $j = 1, \cdots, t-1$, and $PA^ib_j = 0$ for all other powers of A.

Let $\underline{A} = A + BP$. Then the controllability matrix of the pair $(\underline{A}, b_1)$ is $\underline{H} = [b_1 \underline{A}b_1 \cdots \underline{A}^{n-1}b_1]$ and it has rank n. Clearly b_1 is in the column space of $\underline{B}$. The necessity part of the theorem now follows from Proposition 4.1.

We now prove sufficiency.

PROPOSITION 4.3. *Given an arbitrary monic polynomial g of degree n, if there exists an $m \times n$ matrix K such that the characteristic polynomial of $A + BK$ is precisely g, then (A, B) is a completely reachable pair.*

Proof. We first assume that the field F has a sufficient number of scalars $a_1, \cdots, a_n$ such that $\det(A - a_iI) \neq 0$, $i = 1, 2, \cdots, n$. From the above assumption and by hypothesis there is a K such that $(A + BK)v_i = a_iv_i$ and $v_i \neq 0$. Since $a_iI - A$ is invertible, we have

$$(a_iI - A)^{-1}BKv_i = v_i, \qquad i = 1, 2, \cdots, n. \tag{4.1}$$

Now for each a_i there are scalars $b_j(a_i)$ such that

$$(a_iI - A)^{-1} = \sum_{j=1}^{n} b_j(a_i)A^{j-1}, \qquad i = 1, 2, \cdots, n. \tag{4.2}$$

Hence from (4.1) and (4.2), we obtain

$$\sum_{j=1}^{n} A^{j-1}B(b_j(a_i)Kv_i) = v_i, \qquad i = 1, 2, \cdots, n. \tag{4.3}$$

Let $H = [B, AB, \cdots, A^{n-1}B]$ and $y = (y_1 \cdots y_n)^T \in F^{nm}$. Then

$$Hy = \sum_{j=1}^{n} A^{j-1}By_j.$$

If we set $y_j = b_j(a_i)Kv_j$, then (4.3) becomes

$$Hy_i^* = v_i, \qquad i = 1, 2, \cdots, n, \tag{4.4}$$

where $y_i^* = (b_1(a_i)Kv_i \cdots b_n(a_i)Kv_i)^T$.

Since the eigenvalues of $A + BK$ are distinct, the eigenvectors $v_1, \cdots, v_n$ are linearly independent and form a basis for F^n. Hence, by using (4.4), any $v \in F^n$ can be written as $v = H(\sum_i c_i y_i^*)$. Therefore the range of H is F^n and hence (A, B) is a completely reachable pair.

Now if F does not contain enough distinct scalars, apply Proposition A.3 of the Appendix to $f = \det(A - xI)$ and $g = \det(A + BK - xI)$. Then over some extension field $F' \supset F$, g has n distinct roots none of which are roots of f. Now from the proof of Proposition 4.3, H considered as a linear transformation of $(F')^{nm} \to (F')^n$ has rank n. But H is a matrix over $F \subset F'$; hence it has rank n over F also.

For finite fields (containing at least 2 elements) the following stronger result can be proved.

THEOREM 4.4. *The following statements are equivalent*:

(i) (A, B) *is a completely reachable pair*;

(ii) *Given a monic polynomial* g *of degree* n, *there exists a matrix* K *such that the characteristic polynomial of* $A + BK$ *is precisely* g;

(iii) $B \neq 0$, *and given an irreducible polynomial* p *of degree* n, *there exists a matrix* K *such that the characteristic polynomial of* $A + BK$ *is* p.

Proof. The theorem will be proved by showing that the statements (i) and (iii) are equivalent.

(i) $\Rightarrow$ (iii) from Theorem 3.1.

We now prove the reverse implication. For $n = 1$, the result is obvious. For $n > 1$, by Proposition A.1 we can construct an irreducible polynomial of degree n.

Let $\mathscr{R}$ denote the range of $H(A, B)$. Define the map

$$\bar{A}: F^n/\mathscr{R} \to F^n/\mathscr{R}$$

by

$$\bar{A}\bar{x}_i = \overline{(A + BK)x_i}, \qquad i = 1, 2, \cdots, n,$$

where $F^n/\mathscr{R}$ is the quotient space, $\{x_1, \cdots, x_n\}$ is a basis for F^n and $\bar{x}$ denotes the coset of x in the quotient space $F^n/\mathscr{R}$ and K is an $m \times n$ matrix. This is a well-defined map since $\mathscr{R}$ is an A-invariant subspace of F^n.

Let $p(x) = \sum_{i=0}^{n} p_i x^i$ be the characteristic polynomial of $A + BK$. Then by the Cayley–Hamilton theorem $p(A + BK) = 0$. It is easily verified that $\bar{A}$ is an endomorphism of $F^n/\mathscr{R} \to F^n/\mathscr{R}$. Using an induction on k we can show

$$\bar{A}^k\bar{x} = \overline{(A + BC)^k x}$$

and we may verify that $p(\bar{A}) = 0$ (that is, the zero map on $F^n/\mathscr{R}$).

Let m be the minimal polynomial of $\bar{A}$. Then m divides p since $p(\bar{A}) = 0$. Since by hypothesis p is irreducible, either $m = 1$ or $m = \pm p$.

Since $B \neq 0$, $\deg m < n = \deg p$, so $m = 1$. But $m(\bar{A}) = 0$; this means that the identity map on $F^n/\mathscr{R}$ is equal to the zero map and hence $F^n/\mathscr{R} = \mathscr{R}$. Therefore $\mathscr{R} = F^n$ and (A, B) is a completely reachable pair.

5. An example. As an example, consider the following three-state circuit over the field Z_3.

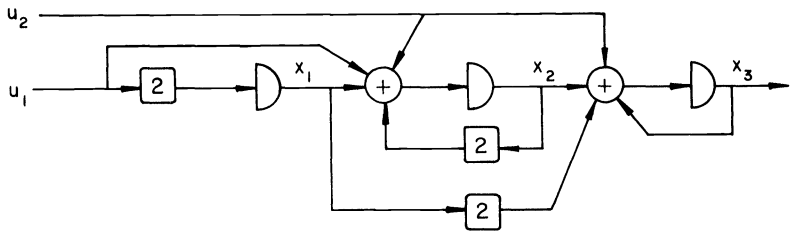


FIG. 1

By inspection of the diagram,

$$x_1(n+1) = 2u_1(n),$$
$$x_2(n+1) = x_1(n) + 2x_2(n) + u_1(n) + u_2(n),$$
$$x_3(n+1) = 2x_1(n) + x_2(n) + x_3(n) + u_2(n).$$

Letting $x^T(n) = [x_1(n) \quad x_2(n) \quad x_3(n)]$ and $u^T(n) = [u_1(n) \quad u_2(n)]$,

$$x(n+1) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 2 & 0 \\ 2 & 1 & 1 \end{pmatrix} x(n) + \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} u(n) = Ax(n) + Bu(n).$$

By direct calculation,

$$H = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 2 & 2 & 1 \\ 0 & 1 & 2 & 2 & 0 & 1 \end{pmatrix},$$

which has rank 3. Following the construction in Proposition 4.2,

$$R = [b_1 \quad Ab_1 \quad A^2b_1] = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 1 & 2 \\ 0 & 2 & 0 \end{pmatrix}.$$

Then $R^{-1} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 2 \\ 2 & 2 & 2 \end{pmatrix}$; also, $S = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Therefore $P = SR^{-1} = \begin{pmatrix} 0 & 0 & 0 \\ 2 & 2 & 2 \end{pmatrix}$.

Again by direct calculation, $\underline{A} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 2 \\ 1 & 0 & 0 \end{pmatrix}$ and $\underline{H} = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 1 & 2 \\ 0 & 2 & 0 \end{pmatrix}$, which is the same as R in this case, so $\underline{H}$ has rank 3. Letting $p_1 = (p_{11}p_{12}p_{13})$, the characteristic polynomial of $\underline{A} + b_1p_1$ is $-x^3 + (2p_{11} + p_{12} + 1)x^2 + (2p_{13} + p_{11})x + (p_{12} + p_{13})$. To see that these coefficients may be chosen arbitrarily, it suffices to note that the following determinant is nonzero:

$$\begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & 2 \\ 0 & 1 & 1 \end{pmatrix} = 1 \neq 0.$$

6. Observability and state reconstruction. Since the pair (A, C) has been assumed to be a completely observable pair, it follows from Theorem 3.1 that for every monic polynomial g of degree n, there exists a $p \times n$ matrix $-D^T$ such that the characteristic polynomial of $A^T - C^T D^T$ is precisely g. Hence the characteristic polynomial of $A - DC$ can be made arbitrary.

Now consider an observer [7]:

$$\hat{x}(t+1) = X\hat{x}(t) + Dy(t) + Bu(t), \qquad t = 0, 1, 2, \cdots, \tag{6.1}$$

where X and D are $n \times n$ and $n \times p$ matrices respectively. Let D be chosen such that $A^T - C^T D^T$ has arbitrary characteristic polynomial and let $X = A - DC$.

Then

$$\begin{aligned} &\hat{x}(t+1) - x(t+1) = (A - DC)[\hat{x}(t) - x(t)], \\ &\hat{x}(0) - x(0) = \text{given}. \end{aligned} \tag{6.2}$$

From (6.2) it follows that

$$\hat{x}(n) - x(n) = (A - DC)^n[\hat{x}(0) - x(0)].$$

Since the characteristic polynomial of $A - DC$ can be made arbitrary, the matrix $A - DC$ can in particular be made nilpotent and hence the observer reconstructs the initial state in at most n steps.

Appendix. In this Appendix some results on finite fields which are used in the proof of Proposition 4.3 are presented.

The following results are needed. The proof of the first two propositions are consequences of well-known results on finite fields (see Lang [8, Chap. VII, § 5]).

PROPOSITION A.1 (see [9, p. 128]). *If F is a finite field consisting of at least two elements, then the polynomial ring $P(F)$ contains irreducible polynomials of every degree $\geqq 2$.*

PROPOSITION A.2. *For every irreducible polynomial over a finite field F, there is an extension field F' such that the given polynomial has n distinct roots in F', where n is the degree of the polynomial.*

PROPOSITION A.3. *Let F be a finite field and f a given polynomial of degree n over F. Then there is a polynomial g of degree n over F and some extension field $F' \supset F$ such that g has n distinct roots in F', none of which are roots of f.*

Proof. First, consider the case when f has at least one root in F. Then $f = f'f''$, where f' is a product of linear factors and f'' has no roots in F. Also, $\deg f' \geqq 1$, so $\deg f'' < n$.

By Proposition A.1, there is an irreducible polynomial g over F of degree n. Then $\gcd(f'', g) = 1$ since $\deg f'' < \deg g$. Clearly $\gcd(f', g) = 1$, so $\gcd(f, g) = 1$.

By Proposition A.2, there is an extension field $F' \supset F$ such that g has n distinct roots in F'. But $\gcd(f, g) = 1$ in F' also, so no root of g is a root of f in F'.

Next, consider the case when f has no roots in F. Now, etiher F has n distinct scalars or not. If it has, let $g = (a_1 - x)(a_2 - x)\cdots(a_n - x)$, where the a_i are distinct scalars in F. Then no root of g is a root of f, and F is the desired extension field.

If F does not have n distinct scalars, consider the prime factorization of f: $f = p_1 p_2 \cdots p_k$, where each p_i is an irreducible polynomial of degree at least 2, say $\deg p_i = m_i$, and $\sum_{i=1}^{k} m_i = n$. (Note that the p_i may not be distinct.)

If $k > 1$, then each $m_i < n$. Pick a polynomial g over F irreducible of degree n. Then $\gcd(f, g) = 1$. Again. let $F' \supset F$ be an extension of F containing n distinct roots of g. Since $\gcd(f, g) = 1$, no root of g is a root of f.

If $k = 1$, then $f = p_1$ and $m_1 = n$. Let F have p distinct scalars, and choose $g' = (a_1 - x)\cdots(a_p - x)$, where the $a_i \in F$ are distinct. Also choose g'' over F irreducible of degree $n - p$. Then $\gcd(f, g'') = 1$ and $\gcd(f, g') = 1$; hence, $\gcd(f, g'g'') = 1$.

Let $F' \supset F$ be an extension field in which g'' has $n - p$ distinct roots. Therefore $g = g'g''$ has n distinct roots in F'. Also, since $\gcd(f, g) = 1$, no root of g is a root of f.

Acknowledgment. The authors are indebted to the reviewer for constructive comments and suggestions. In particular the statement and proof of Theorem 4.4 was obtained as a result of comments by the reviewer.

Note added in proof. It was recently pointed out to me by R. W. Brockett that V. M. Popov proved the result on pole assignment earlier in his paper: *Hyperstability and optimality of automatic systems with several control functions*, Rev. Roumaine Sci. Techn.-Electrotechnct. Energ., 9 (1964), pp. 629–690.

REFERENCES

[1] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1967.

[2] W. M. WONHAM, *On pole-assignment in multi-input controllable linear systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 660–665.

[3] J. D. SIMON AND S. K. MITTER, *A theory of modal control*, Information and Control, 13 (1968), pp. 316–353.

[4] A. GILL, *Linear Sequential Circuits: Analysis, Synthesis and Applications*, McGraw-Hill, New York, 1967.

[5] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[6] M. HEYMANN, *Comments on "Pole Assignment in Multi-Input Controllable Linear Systems,"* IEEE Trans. Automatic Control, AC-13 (1969), pp. 748–749.

[7] D. G. LUENBERGER, *Observers for multivariable systems*, Ibid., AC-11 (1966), pp. 190–197.

[8] S. LANG, *Algebra*, Addison-Wesley, Reading, Mass., 1965.

[9] I. T. ADAMSON, *Introduction to Field Theory*, Oliver and Boyd, Edinburgh and London, 1964.

SIAM J. CONTROL
Vol. 9, No. 1, February 1971

STABILITY OF A NONAUTONOMOUS DIFFERENTIAL EQUATION OF FOURTH ORDER*

A. S. C. SINHA AND R. G. HOFT†

1. Introduction. In this paper, the object is to establish sufficient conditions for stability of (1) in the cases $p(t) \equiv 0$, and $p(t) \not\equiv 0$, respectively, which follows from some fundamental results due to LaSalle [1]. Further, it is worth noting that, in view of LaSalle's results, it is not necessary to show that the Lyapunov function is positive definite (and this would be a tedious computation) in order to conclude global asymptotic stability

$$\dddot{x} + f(\ddot{x})\ddot{x} + \phi(\dot{x}, \ddot{x})\ddot{x} + \psi(\dot{x}) + \theta(x) = p(t). \tag{1}$$

It is easily seen that the results in this analysis become Routh–Hurwitz conditions when functions f, ψ, ϕ and θ are constants in the case $p(t) \equiv 0$. The conditions obtained are generalizations of the work of Ezeilo [2] and Harrow [3] for a class of fourth order systems. The Lyapunov function given here has been constructed by forming linear combinations of line integrals. However, the constants for the linear combinations are obtained by a trial method. The functions f, ϕ, ψ and θ are real-valued with continuous first partials. A smoothness property is also assumed to ensure the existence of integrals appearing in the analysis.

2. Stability. Consider the differential equation (1) and the associated system (*Case* 1. $p(t) \equiv 0$):

$$\begin{aligned} &\dot{x} = y, \quad \dot{y} = z, \quad \dot{z} = w, \\ &\dot{w} = -f(z)w - \phi(y, z)z - \psi(y) - \theta(x), \end{aligned} \tag{2}$$

where $\psi(0) = \theta(0) = 0$, so that the origin is a critical point.

It is assumed that there exist positive constants such that

$$b \geqq f(z) \geqq f^o \geqq a > 0, \tag{3}$$

$$\phi(y, z) \geqq \phi^o > 0, \tag{4}$$

$$\psi(y)/y \geqq \psi^o > 0, \tag{5}$$

$$(\theta^o\delta)^{1/2} \geqq \theta(x)/x \geqq \theta^o > 0, \tag{6}$$

$$\psi^o \leqq \psi'(y) \leqq \delta_1, \qquad \theta^o \leqq \theta'(x) \leqq \theta'(0) \leqq \delta, \tag{7}$$

* Received by the editors June 10, 1969.

† College of Engineering, University of Missouri, Columbia, Missouri 65201. This work was supported in part by the National Science Foundation under Grant NSF GK805.

where the superscript o designates the evaluation of the given function at the origin. Let the following relations between the constants of (3) to (7) hold:

$$a\phi^o\psi^o\theta^o - \delta ab\theta^o - \delta_1^2\delta = \beta > 0, \tag{8}$$

$$\frac{\beta}{(\psi^o)^2(2a\psi^o\theta)^{1/2}} > \left[\frac{\delta_1}{\psi^o} - \frac{\theta^o}{\delta}\right], \tag{9}$$

$$\frac{1}{z}\int_0^z f(s)\,ds - f(z) < \frac{\beta}{a^2\psi^o\theta^o} \qquad \text{for all } z \neq 0. \tag{10}$$

The following V-functions are introduced:

$$V_1(y, z, w) = \frac{z}{F(z)}\left[w + F(z) + \frac{\delta}{\psi^o}\frac{F(z)}{z}y\right]^2 + \left[\frac{1}{a} - \frac{z}{F(z)}\right]w^2, \tag{11}$$

where $\int_0^z f(s)\,ds = F(z)$;

$$V_2(y, z) = \frac{2}{a}\int_0^z s\{\phi(y, s) - \phi^o\}\,ds + \frac{2\delta}{\psi^o}\int_0^y s\{\phi(s, 0) - \phi^o\}\,ds, \tag{12}$$

$$V_3(x, y, z) = \frac{\theta^o\delta}{\psi^o}\left\{x + \frac{\psi^o}{\theta^o\delta}\left[\frac{\theta(x)}{x}\right]y + \frac{\psi^o}{a\theta^o\delta}\left[\frac{\theta(x)}{x}\right]z\right\}^2 + \frac{2\delta}{\psi^o}\int_0^x s\left\{\frac{\theta(s)}{s} - \theta^o\right\}ds, \tag{13}$$

$$\begin{aligned} V_4(x, y, z) = {} & 2\int_0^y \left\{\frac{\psi(s)}{s} - \frac{\psi^o}{\theta^o\delta}\left[\frac{\theta(x)}{x}\right]^2\right\}s\,ds \\ & + \left\{\frac{\phi^o\delta}{\psi^o} - \frac{\delta^2}{(\psi^o)^2}\left[\frac{F(z)}{z}\right] - \frac{\theta'(0)}{a}\right\}y^2 \\ & + \left\{\frac{\phi^o}{a} - \frac{\delta}{\psi^o} - \frac{\psi^o}{\delta\theta^o a^2}\left[\frac{\theta(x)}{x}\right]^2\right\}z^2 + 2\int_0^z sf(s)\,ds - zF(z) \\ & + \frac{2}{a}\left\{\frac{\psi(y)}{y} - \frac{\psi^o}{\theta^o\delta}\left[\frac{\theta(x)}{x}\right]^2\right\}yz. \end{aligned} \tag{14}$$

Consider the following inequalities which are used to show that $V \to \infty$ as $|x| \to \infty$. Now $0 < F(z)/z < b$ and therefore from (7) it follows that

$$\alpha(x, y) = \frac{\psi(y)}{y} - \frac{\psi^o}{\theta^o \delta}\left[\frac{\theta(x)}{x}\right]^2 \geqq \frac{\psi(y)}{y} - \psi^o \geqq 0, \tag{15}$$

$$\begin{aligned} \frac{\phi^o \delta}{\psi^o} - \frac{\delta^2}{(\psi^o)^2}\left[\frac{F(z)}{z}\right] - \frac{\theta'(0)}{a} &\geqq \frac{\delta}{a(\psi^o)^2\theta^o}[a\phi^o\psi^o\theta^o - \delta ab\theta^o - (\psi^o)^2\delta] \\ &\geqq \frac{\delta}{a\psi^{o2}\theta^o}[a\phi^o\psi^o\theta^o - \delta ab\theta^o - \delta_1^2\delta] = \frac{\delta\beta}{a\psi^{o2}\theta^o} > 0. \end{aligned} \tag{16}$$

From the mean value theorem,

$$\psi^o \leqq \frac{\psi(y)}{y} = \int_0^1 \psi'(yt)\, dt \leqq \delta_1, \qquad \theta^o \leqq \frac{\theta(x)}{x} = \int_0^1 \theta'(xt)\, dt \leqq \delta$$

and (3), (6) and (7) give

$$\begin{aligned} \frac{\phi^o}{a} - \frac{\delta}{\psi^o} - \frac{\psi^o}{\delta\theta^o a^2}\left[\frac{\theta(x)}{x}\right]^2 &\geqq \frac{1}{a^2\psi^o\theta^o}[a\phi^o\psi^o\theta^o - \delta a^2\theta^o - (\psi^o)^2\theta^o] \\ &\geqq \frac{1}{a^2\psi^o\theta^o}[a\phi^o\psi^o\theta^o - \delta ab\theta^o - \delta_1^2\delta] = \frac{\beta}{a^2\psi^o\theta^o} > 0, \end{aligned} \tag{17}$$

$$\begin{aligned} 2\int_0^z sf(s)\, ds - zF(z) &= \int_0^z sf(s)\, ds - \int_0^z F(s)\, ds \\ &= -\int_0^z \left\{\frac{F(s)}{s} - f(s)\right\} s\, ds > -\frac{\beta}{2a^2\psi^o\theta^o}z^2. \end{aligned} \tag{18}$$

Therefore, using (15), (16), (17) and (18) in (14) gives

$$\begin{aligned} V_4(x, y, z) &\geqq \int_0^y \alpha(x, s)s\, ds + \frac{\delta\beta}{a(\psi^o)^2\theta^o}y^2 + \frac{\beta}{2a^2\psi^o\theta^o}z^2 + \frac{2}{a}\alpha(x, y)yz \\ &\geqq 2\int_0^y \alpha(x, s)s\, ds + \frac{\beta}{a}\left[\frac{y}{\psi^o} + \frac{\psi^o}{\beta}\alpha(x, y)z\right]^2 + Cz^2, \end{aligned}$$

where

$$C = \frac{\beta}{2a^2\psi^o\theta^o} - \frac{(\psi^o)^2}{a\beta}\alpha^2(x, y) > 0.$$

Now $\alpha(x, y) \leqq \psi^o[\delta_1/\psi^o - \theta^o/\delta]$; therefore, from (9), $\alpha^2(x, y) < \beta^2/[(\psi^o)^2 2a\psi^o\theta^o]$.

Consider the Lyapunov function

$$
\begin{aligned}
2V(x, y, z, w) &= V_1 + V_2 + V_3 + V_4 \\
&\geqq \frac{z}{F(z)}\left[w + F(z) + \frac{\delta}{\psi^o}\frac{F(z)}{z}y\right]^2 + \left[\frac{1}{a} - \frac{z}{F(z)}\right]w^2 \\
&\quad + \frac{\theta^o\delta}{\psi^o}\left\{x + \frac{\psi^o}{\theta^o\delta}\left[\frac{\theta(x)}{x}\right]y + \frac{\psi^o}{a\theta^o\delta}\left[\frac{\theta(x)}{x}\right]z\right\}^2 \\
&\quad + \frac{2\delta}{\psi^o}\int_0^x s\left\{\frac{\theta(s)}{s} - \theta^o\right\}ds + 2\int_0^y \alpha(x, s)s\,ds \\
&\quad + \frac{\beta}{a}\left[\frac{y}{\psi^o} + \frac{\psi^o}{\beta}\alpha(x, y)z\right]^2 + Cz^2.
\end{aligned}
$$

Thus $V \to \infty$ as $|x| \to \infty$. A little computation gives

$$
\begin{aligned}
2V(x, y, z, w) &= \frac{2\delta}{\psi^o}\int_0^x \theta(s)\,ds + \left[\frac{\phi^o\delta}{\psi^o} - \frac{\theta'(0)}{a}\right]y^2 \\
&\quad + \left(\frac{\phi^o}{a} - \frac{\delta}{\psi^o}\right)z^2 + \frac{1}{a}w^2 + 2wz + \frac{2\delta}{\psi^o}yw \\
&\quad + \frac{2\delta}{\psi^o}y\int_0^z f(s)\,ds + 2\theta(x)y + \frac{2}{a}\theta(x)z \\
&\quad + \frac{2}{a}\psi(y)z + 2\int_0^y \psi(s)\,ds + \frac{2}{a}\int_0^z s\{\phi(y, s) - \phi^o\}\,ds \\
&\quad + \frac{2\delta}{\psi^o}\int_0^y s\{\phi(s, 0) - \phi^o\}\,ds + 2\int_0^z sf(s)\,ds.
\end{aligned} \tag{19}
$$

Evaluating the time derivative of (19) along the trajectories of (2) yields

$$
\begin{aligned}
\dot{V}(x, y, z, w) &= \frac{1}{a}[f(z) - a]w^2 - \left[\phi(y, z) - \frac{\psi'(y)}{a} - \frac{\delta}{\psi^o}\frac{F(z)}{z}\right]z^2 \\
&\quad - \left[\frac{\delta}{\psi^o}\frac{\psi(y)}{y} - \theta'(x)\right]y^2 - \frac{1}{a}[\theta'(0) - \theta'(x)]yz \\
&\quad + \frac{1}{a}z\int_0^z s\phi_y(y, s)\,ds - \frac{\delta}{\psi^o}[\phi(y, z) - \phi(y, 0)]yz \\
&\leqq -\frac{1}{a}[f(z) - a]w^2 - \left\{\phi^o - \frac{1}{a}\psi'(y) - \frac{b\delta}{\psi^o} - \frac{[\theta'(0) - \theta'(x)]}{4a^2}\right\}z^2 \\
&\quad - \delta_2 y^2 - [\theta'(0) - \theta'(x)]\left(y + \frac{z}{2a}\right)^2 \\
&\quad + \frac{1}{a}z\int_0^z s\phi_y(y, s)\,ds - \frac{\delta}{\psi^o}\left[y\int_0^1 \frac{\partial\phi}{\partial z}(y, zt)\,dt\right]
\end{aligned} \tag{20}
$$

and with $\delta = \theta'(0) + \delta_2$, $\delta_2 > 0$,

$$\phi(y, z) - \phi(y, 0) = z \int_0^1 \frac{\delta\phi}{\delta z}(y, zt)\, dt.$$

Then (20) is negative semidefinite if the following conditions hold:

$$z\phi_y(y, z) \leqq 0, \tag{21}$$

$$y \int_0^1 \frac{\partial\phi}{\partial z}(y, zt)\, dt \geqq 0, \tag{22}$$

$$\theta'(0) > 0, \qquad \theta'(0) \geqq \theta'(x) \geqq \theta'(0) - \gamma_1, \qquad \gamma_1 \geqq 0, \tag{23}$$

where $\gamma_1 = 4a^2(\beta/a\psi^o\theta^o - \gamma)$ for some $\gamma \leqq \beta/a\psi^o\theta^o$. Using (7), (8) and (23) gives

$$\begin{aligned}\left[\phi^o - \frac{1}{a}\psi'(y) - \frac{b\delta}{\psi^o}\right] - \left[\frac{\theta'(0) - \theta'(x)}{4a^2}\right] &= \frac{1}{a\psi^o\theta^o}[a\phi^o\psi^o\theta^o - \psi^o\theta^o\psi'(y) - ab\delta\theta^o] - \left[\frac{\theta'(0) - \theta'(x)}{4a^2}\right] \\ &\geqq \frac{\beta}{a\psi^o\theta^o} - \frac{\gamma_1}{4a^2} = \gamma > 0.\end{aligned}$$

The conditions of Theorems 2 and 3 of LaSalle's paper [1] are satisfied; therefore global asymptotic stability has been proved. Thus these results are combined into the following theorem.

THEOREM 1. *If there exist positive constants as defined in* (3)–(7) *of the system* (2) *such that the following conditions hold*:

(i) $\psi(0) = \theta(0) = 0$, *and* f, ϕ, ψ, θ *have continuous first partials*;

(ii) *conditions* (8)–(10) *and* (21)–(23) *hold*;

(iii) $\dot{V} \not\equiv 0$ *along every nontrivial solution*;

then the system (2) *is asymptotically stable.*

3. Nonautonomous systems. In this section, the system (1) is analyzed to give sufficient conditions for which every solution of (1) is bounded for all $t \geqq 0$. Here it is assumed (Case 2) that $p(t)$ is an integrable function.

THEOREM 2. *If the conditions* (i) *and* (ii) *of Theorem* 1, *along with the following conditions hold*:

(i) $f(z) \geqq a + \gamma \geqq a + \frac{1}{2}|p(t)|$;

(ii) $\delta_2 \geqq (\delta/2\psi^o)|p(t)|$;

(iii) $\int_0^\infty |p(s)|\, ds \leqq c < \infty$;

then every solution of (1) *is bounded in the future* [1].

Proof. To prove Theorem 2, consider the Lyapunov function (19) which was shown to be such that $V \to \infty$ as $|x| \to \infty$. Evaluating the time derivative along

the trajectories of (1) results in the addition of $p(t)$ terms to (20).

$$
\begin{aligned}
\dot{V}(x, y, z, w) \leqq & -\frac{1}{a}[f(z) - a]w^2 - \left\{\phi^o - \frac{1}{a}\psi'(y) - \frac{b\delta}{\psi^o} - \frac{[\theta'(0) - \theta'(x)]}{4a^2}\right\} z^2 \\
& - \delta_2 y^2 - [\theta'(0) - \theta'(x)]\left(y + \frac{z}{2a}\right)^2 \\
& + \frac{1}{a} z \int_0^s s\phi_y(y, s)\, ds - \frac{\delta}{\psi^o}\left[y \int_0^z \frac{\partial \phi}{\partial z}(y, zt)\, dt \right] z^2 \\
& + \frac{1}{a} p(t)w + \frac{\delta}{\psi^o} p(t)y + p(t)z \qquad (24) \\
\leqq & -\frac{1}{a}\left[f(z) - a - \frac{1}{2}|p(t)| \right] w^2 - \left[\gamma - \frac{1}{2}|p(t)| \right] z^2 \\
& - \left[\delta_2 - \frac{\delta}{2\psi^o}|p(t)| \right] y^2 + \frac{1}{2}\left(\frac{1}{a} + \frac{\delta}{\psi^o} + 1 \right) |p(t)| \\
\leqq & \frac{1}{2}\left(\frac{1}{a} + \frac{\delta}{\psi^o} + 1 \right) |p(t)|
\end{aligned}
$$

if the conditions (i), (ii) and (iii) hold.

Now integrating (24), we have

$$
\begin{aligned}
V(x, y, z, w) & \leqq V(0) + \frac{1}{2}\left(\frac{1}{a} + \frac{\delta}{\psi^o} + 1 \right) \int_0^t |p(s)|\, ds \qquad (25) \\
& \leqq V(0) + \frac{1}{2}\left(\frac{1}{a} + \frac{\delta}{\psi^o} + 1 \right) c \leqq k.
\end{aligned}
$$

Therefore, every solution of (1) is bounded in the future.

4. Example. A very simple case of differential equation (1) is when the functions f and ϕ are constants. In this case the previous results become the Routh–Hurwitz conditions. Consider

$$
\dddot{x} + a_1 \ddot{x} + a_2 \dot{x} + a_3 \dot{x} + a_4 x = p(t). \qquad (26)
$$

Let $p(t) = 0$; then the conditions (8)–(10) result in only one nontrivial relation (27) which is clearly a Routh–Hurwitz condition for the system.

$$
a_1 a_2 a_3 a_4 - a_4^2 a_1^2 - a_3^2 a_4 = \beta > 0. \qquad (27)
$$

Let $p(t) \not\equiv 0$ and define the constants $a = a_1/2$, $b = f^o = a_1$ and $\delta = 2a_4$. Then (26) has bounded solutions in the future if

(i) $\frac{1}{2}a_1 a_2 a_3 a_4 - a_4^2 a_1^2 - 2a_3^2 a_4 = \beta > 0$,

(ii) $a_1 \geqq |p_1(t)|$, $a_3 \geqq |P(t)|$

and the conditions of Theorem 2 are satisfied.

REFERENCES

[1] J. P. LaSalle, *Stability theory for ordinary differential equations*, Differential Equations, 4 (1968), pp. 57–65.

[2] J. O. C. EZEILO, *Further results for the solutions of third order differential equations*, Proc. Cambridge Philos. Soc., 63 (1967), pp. 147–154.

[3] M. HARROW, *Further results for the solution of third order differential equations*. Proc. London Math. Soc., 43 (1968), pp. 587–592.

[4] J. LASALLE AND S. LEFCHETZ, *Stability by Liapunov's Direct Method with Applications*, Academic Press, New York, 1961.

[5] J. O. C. EZEILO, *Further results for the solutions of a third-order differential equation*, Proc. Cambridge Philos. Soc., 59 (1963), pp. 111–116.

[6] N. N. KRASOVSKII, *Stability of Motion*, Stanford University Press, Stanford, California, 1963.

[7] W. LEIGHTON, *On the construction of Liapunov functions for certain autonomous nonlinear equations*, Contributions to Differential Equations, 4 (1963), pp. 367–383.

[8] A. I. OGURTSOV, *The stability of solutions of two nonlinear differential equations of the third and fourth orders*, J. Appl. Math. Mech., 23 (1959), pp. 247–251.

SIAM J. CONTROL
Vol. 9, No. 1, February 1971

ERROR BOUNDS OF HIGH ORDER ACCURACY FOR THE STATE REGULATOR PROBLEM VIA PIECEWISE POLYNOMIAL APPROXIMATIONS*

W. E. BOSARGE, JR. AND O. G. JOHNSON†

Abstract. Consider the linear quadratic cost control problem $\dot{x} = A(t)x + B(t)u$, $x(0) = x_0$, with a cost functional $J[u] = \frac{1}{2}\int_0^T [\langle x, Q(t)x\rangle + \langle u, R(t)u\rangle]\,dt$. Let S be a suitable space of piecewise cubic polynomials on a mesh of norm h on the interval $[0, T]$. Then it is shown that a so-called Ritz–Trefftz method for minimizing $J[\cdot]$ over S leads to an approximation to $J[\cdot]$ of order $O(h^7)$. Further, a computable error bound can be exhibited. It is also shown that the computed pair $(\bar{u}, \bar{x})$ converges to the optimal pair (u^*, x^*) with order $O(h^3)$. Similar statements are made for piecewise polynomial approximation of arbitrary positive order.[1]

1. Introduction. In the past considerable effort has been devoted to the use of direct methods in calculus of variations and control problems (see, for example, [4], [12], [6]). Specifically, the effort was directed towards finding the minimum of a functional $J[u; x]$ (defined for some admissible controls and states) over a suitable finite-dimensional subspace of basis functions $\{u_n(t), x_n(t); n = 0, 1, \cdots, N\}$. In this paper we consider the application of a modified Ritz direct method to the so-called state regulator control problem using finite-dimensional piecewise polynomial bases. We show that, for the "well-behaved" linear quadratic cost control problem, a numerically useful computational algorithm can be obtained for computing very high order approximations to the optimal control, the corresponding optimal state, and the associated cost functional. In particular, for the "smooth" problem, we show that a so-called Ritz–Trefftz method employing α-order splines delivers approximations to the optimal control and optimal state with an error $O(h^\alpha)$, where h represents the mesh size for the spline basis. The corresponding approximate optimal cost estimates the true (optimal) cost to within $O(h^{2\alpha})$. We also exhibit an $O(h^{2\alpha+1})$ error bound using the Hermite space $H^\alpha(\Pi)$. In addition, we represent an explicit derivation of the numerical Ritz–Trefftz algorithm. In a forthcoming paper we will discuss the practical utility of the method including results on its numerical stability. We will also present some comparisons of our method to the classical approach where numerical integration of a nonlinear Ricatti differential equation is required. (Note $O(h^\alpha)$ is a function less than Kh^α in norm for some $K > 0$, where K is independent of h.)

2. Problem description. In this section we present a detailed description of the infinite-dimensional state regulator problem and recall a number of useful analytic results which we shall require in the sequel.

Consider the linear time-varying system

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t), \\ x(0) &= x_0 \end{aligned} \tag{2.1}$$

* Received by the editors October 9, 1969, and in revised form May 26, 1970.

† International Business Machines Scientific Center, Houston, Texas 77025 and Department of Mathematical Sciences, Rice University, Houston, Texas.

[1] A detailed numerical study of the algorithm resulting from this modified Ritz approach will appear in a forthcoming paper.

and the cost functional $J[\cdot]$ defined by

$$J[u] = \frac{1}{2}\int_0^T [\langle x(t), Q(t)x(t)\rangle + \langle u(t), R(t)u(t)\rangle]\, dt. \tag{2.2}$$

Assume that $x(\cdot)$ is an n-dimensional state vector, $u(\cdot)$ an r-dimensional control vector, $A(\cdot)$ an $n \times n$ matrix, $B(\cdot)$ an $n \times r$ matrix, $Q(\cdot)$ an $n \times n$ positive definite matrix, and $R(\cdot)$ an $r \times r$ positive definite matrix. In addition, assume that A, B, Q, R are (elementwise) piecewise $C^{\alpha}[0, T], \alpha \geqq 0$. We require that the terminal time T be fixed, that $0 < r \leqq n$, and that $u(\cdot)$ be unconstrained.

The problem we treat in this paper can be described in the following equivalent ways.

DEFINITION 2.1 (Problem 1). Given the linear system (2.1) and the cost functional (2.2) subject to the assumptions stated above, find the optimal control, i.e., the control which will drive the system (2.1) so as to minimize the cost functional (2.2). Thus we seek a u^* such that $J[u^*] = \min_{u \in \mathscr{A}_u} J[u]$ with $\dot{x}^* = A(t)x^*(t) + B(t)u^*(t)$ and $x^*(0) = x_0$. Here $\mathscr{A}_u$ is some set of admissible vector-valued functions on $[0, T]$.

Since Problem 1 can be viewed as a standard calculus of variations minimization problem subject to nonholonomic constraints, we may introduce Lagrange multipliers $\lambda(\cdot)$ (time-varying n-vector) and γ(n-vector) and define the Lagrangian $L[u, x; \lambda, \gamma]$ by

$$\begin{aligned} L[u, x, \lambda, \gamma] \equiv J[u, x] &+ \int_0^T \langle \lambda(t),\ -\dot{x} + A(t)x + B(t)u\rangle dt \\ &+ \langle \gamma, (x(0) - x_0)\rangle, \end{aligned} \tag{2.3}$$

where $J[u, x]$ is defined in (2.2) but is viewed as a function of two variables. Thus, an alternative way of describing Problem 1 is given by the following definition.

DEFINITION 2.2 (Problem 1′). Given the linear system (2.1) and the cost functional (2.2) subject to the assumptions of Definition 2.1, find the u^*, x^*, λ^* and γ^* such that the Lagrangian is extremized; that is,

$$L[u^*, x^*; \lambda^*, \gamma^*] = \max_{\substack{\lambda \in \mathscr{A}_\lambda \\ \gamma \in R_n}} \min_{\substack{u \in \mathscr{A}_u \\ x \in \mathscr{A}_x}} L[u, x; \lambda, \gamma], \tag{2.4}$$

where R_n is real Euclidean n-space, and $\mathscr{A}_x$, $\mathscr{A}_\lambda$ and $\mathscr{A}_u$ are innerrelated.

One of the essential properties of the multipliers $\lambda(\cdot)$ and γ is that, in the process of extremizing L over u, x, λ and γ, the Lagrangian is maximized over the multipliers λ and γ, and minimized with respect to u and x (see [10]). This fact is often misunderstood (see [12], [9] or [5]) and sometimes permits erroneous conclusions to be drawn in the convergence analysis of direct methods which have been applied to problems of the above type (see, for example, [12]).

We introduce still a third formulation of Problem 1 which is useful in the sequel.

DEFINITION 2.3 (Problem 1″). Under the assumptions of Definition 2.2, find the λ^* and γ^* such that

$$L[u^*, x^*; \lambda^*, \gamma^*] = \max_{\substack{\lambda \in \mathscr{A}_\lambda \\ \gamma \in R_n}} L[u, x; \lambda, \gamma] \tag{2.5}$$

subject to

$$\frac{\partial L}{\partial u}[u, x; \lambda, \gamma] = 0, \qquad \frac{\partial L}{\partial x}[u, x; \lambda, \gamma] = 0, \tag{2.6}$$

where $\partial L/\partial u$ and $\partial L/\partial x$ are partial Fréchet derivatives of the scalar Langrangian.

The equivalence of the three formulations of Problem 1 follows from well-known "principles" of the calculus of variations (see, for example, [4]).

By smoothness of Problem 1 we shall mean the following.

DEFINITION 2.4. We say Problem $1 \in PC^{\alpha}[0, T]$ if and only if A, B, Q, R have (elementwise) piecewise continuous αth order derivatives. Equivalently we say Problem 1 is αth order *smooth* if Problem $1 \in PC^{\alpha}[0, T]$, $\alpha \geqq 0$.

Basically, the solution of the state regulator problem leads to an optimal feedback system with the property that the state vector is "kept near zero" without excessive expenditure of control energy. Due to the linear nature of the system equations and the quadratic nature of the cost functional, the problem can be "handled analytically" in the sense that the optimal control is an explicit linear function of the state, i.e., is of the form

$$u(t) = G(t)x(t), \qquad t \in [0, T], \tag{2.7}$$

with $G(\cdot)$ an $r \times n$ matrix-valued function defined by

$$G(t) = -R^{-1}(t)B^{T}(t)K(t)x(t), \tag{2.8}$$

where $K(t)$ is the $n \times n$ symmetric matrix solution of the well-known Ricatti equation. Therefore, a complete solution is obtained (in general) only after $K(\cdot)$ has been numerically approximated. We pursue this point further in a later section of the paper.

We now present a number of well-known results for Problem 1.

THEOREM 2.1. *Let us assume that an optimal control u^* for Problem 1 exists for the state x_0. Then, in order that u^* be optimal, it is necessary that there exist a vector function $\lambda^*(\cdot)$ (Lagrange multiplier) such that*

(a) *$\lambda^*(\cdot)$ corresponds to $u^*(\cdot)$ and $x^*(\cdot)$ with $\lambda^*(\cdot)$ and $x^*(\cdot)$ solutions of the equation pair*

$$\dot{x}^*(t) = A(t)x^*(t) + B(t)u^*(t) \tag{2.9}$$

and

$$\dot{\lambda}^*(t) = -Q(t)x^*(t) - A^{T}(t)\lambda^*(t), \tag{2.10}$$

subject to the boundary condition $x^*(0) = x_0$ and the transversality condition $\lambda^*(T) = 0$;

(b) *along the optimal trajectory $x^*(\cdot)$ we must have*

$$R(t)u^*(t) + B^{T}(t)\lambda^*(t) = 0. \tag{2.11}$$

Proof. See Athans and Falb [2].

THEOREM 2.2 (Analytic solution of the state regulator problem 1). *Given the linear system (2.1) and the cost functional (2.2), where $u(\cdot)$ is not constrained, T is specified and $Q(\cdot)$ and $R(\cdot)$ are positive definite matrices. Then an optimal control exists, is unique and is defined by (2.7) and (2.8), where the $n \times n$ symmetric*

matrix K is the solution of the Ricatti equation

$$\dot{K}(t) = -K(t)A(t) - A^T(t)K(t) + K(t)B(t)R^{-1}(t)B^T(t)K(t) - Q(t) \tag{2.12}$$

with $K(T) = 0$. The state $x^*(\cdot)$ of the optimal system is then the solution of the linear system

$$\begin{aligned} \dot{x}(t) &= [A(t) - B(t)R^{-1}(t)B^T(t)K(t)]x(t), \\ x(0) &= x_0. \end{aligned} \tag{2.13}$$

Proof. See [2].

The notation x^*, u^*, λ^* and J^* will be used throughout the paper to denote optimal quantities. We now state a well-known theorem which we shall require in the sequel.

THEOREM 2.3. *Given the linear system* (2.1) *and the cost functional J of* (2.2), *let J^* denote the minimum value of J; then J^* is given by*

$$J^*[x(t); t] = \tfrac{1}{2}\langle x^*(t), \lambda^*(t)\rangle, \qquad 0 \leqq t \leqq T, \tag{2.14}$$

where $\lambda^*(t) = K(t)x^*(t)$ with $K(\cdot)$ the solution of (2.12).

Proof. See [2].

We now prove a theorem which tells us that the optimal quantities u^*, x^* and λ^* for Problem 1 are in $PC^\alpha[0, T]$ provided Problem 1 $\in PC^{\alpha-1}[0, T]$.

THEOREM 2.4. *Assume that Problem 1 is $(\alpha - 1)$st order smooth. Then the optimal control $u^*(\cdot)$, the optimal state $x^*(\cdot)$ and the optimal costate $\lambda^*(\cdot)$ are each in $PC^\alpha[0, T]$.*

Proof. From (2.9), (2.10) and (2.11) we see that $x^*(\cdot)$ and $\lambda^*(\cdot)$ satisfy the $2n$-equation system

$$\begin{bmatrix} \dot{x}^*(t) \\ \dot{\lambda}^*(t) \end{bmatrix} = \left[\begin{array}{c|c} A(t) & -S(t) \\ \hline -Q(t) & -A^T(t) \end{array}\right] \begin{bmatrix} x^*(t) \\ \lambda^*(t) \end{bmatrix} \tag{2.15}$$

subject to $x^*(0) = x_0$ and $\lambda^*(T) = 0$. Here $S(t)$ is defined by

$$S(t) = B(t)R^{-1}(t)B^T(t). \tag{2.16}$$

Since $A, B, Q, R \in PC^{\alpha-1}[0, T]$ and R is positive definite, it follows that $R^{-1} \in PC^{\alpha-1}[0, T]$. Thus, $S \in PC^\alpha[0, T]$. From the theory of ordinary differential equations we now conclude that x^* and $\lambda^* \in PC^\alpha[0, T]$. From (2.11) we deduce that $u^* \in PC^\alpha[0, T]$ also and thus the theorem is proved.

3. Finite-dimensional analogue of Problem 1. In this section we present a special finite-dimensional analogue to Problem 1 using certain subspaces of piecewise polynomials. We first describe a number of useful piecewise polynomial spaces.

3.1. Spaces of piecewise polynomials. Let S_m^α ($\alpha \geqq 1; m \geqq 1$, with α, m integers) be an m-dimensional space of piecewise polynomials of fixed order $\alpha - 1$. Then we assume S_m^α is characterized by the following properties:

P1. There exists a linear operator $L_m: PC^\alpha[0, T] \to S_m^\alpha$.

P2. For all $f \in PC^\alpha[0, T]$, $\|L_m f - f\|_2 = O(m^{-\alpha})$.

P3. For all $f \in PC^{\alpha}[0, T]$, $\|(d/dt)(L_m f - f)\|_2 = O(m^{-\alpha+1})$.

P4. $(L_m f)^{(i)}(0) = f^{(i)}(0)$, $(L_m f)^{(i)}(T) = f^{(i)}(T)$, $i = 0, 1$.

P5. $S_m^{\alpha} \subset \mathscr{A}_{\lambda}$.

It is well known that there exist spaces of piecewise polynomials possessing properties P1–P5 provided, of course, $\mathscr{A}_{\lambda}$ is piecewise smooth. As an example, suppose we define π to be a mesh on $[0, T]$ such that $h = |\Pi|$, i.e., $\Pi : 0 = t_0 < t_1 < \cdots < t_v = T$ and

$$h = \max_{i=1,\cdots,v} |t_i - t_{i-1}|. \tag{3.1}$$

Now suppose $f \in PC^{\alpha}[0, T]$ and let Π contain all points of discontinuity of $f^{(\alpha)}$ in $[0, T]$. Then the closure of the graph of $f \in C^{\alpha}[t_{i-1}, t_i]$, $i = 1, \cdots, v$. Thus, $f^{(\alpha-1)}$ is absolutely continuous and $f^{(\alpha)} \in L_{\infty}$, Now, if m_0 is an integer in $[\alpha/2, \alpha + 1]$, then there exists, for example, a Hermite interpolate to f (which we denote by $L_m^H f$) of order $2m_0 - 1$. Further, it is well known in this case that (see, for example, [3])

$$\|L_m^H f - f\|_2 = O(h^{\alpha+1/2}) \tag{3.2}$$

and

$$\left\| \frac{d}{dt}(L_m^H f - f) \right\| = O(h^{\alpha-1/2}). \tag{3.3}$$

We remark that the order of convergence is actually better (by $\frac{1}{2}$) than we require.

Another extremely useful piecewise polynomial space possessing properties P1–P4 is the space of splines of order $\alpha - 1$ (see, for example, [1]). Spline subspaces have been used frequently in recent years in the development of practical and efficient numerical algorithms for attacking wide classes of problems. In fact, for many practical problems, spline subspaces "deliver" the best results for an equivalent amount of computation, compared with an alternative finite dimension space of piecewise polynomials (see [8]). We shall therefore emphasize spline subspaces S_m^{α} for our numerical comparisons which we present in a forthcoming paper.

3.2. Application to state regulator Problem 1. We now consider a special finite-dimensional approximation to Problem 1 (Definition 2.1) over a typical S_m^{α}. The approximation is, in a sense, a Ritz method, but differs from the classical Ritz procedure in a fundamental way.

We begin by stating a special finite-dimensional analogue for Definition 2.1. Instead of requiring that the differential equation side condition (2.1) be satisfied identically, we relax the constraint as indicated in the following definition (see Trefftz [11]).

DEFINITION 3.1 (Problem 2). Given the linear system (2.1) and the cost functional (2.2) subject to the standard assumptions of Definition 2.1, find the optimal control $\bar{u}$ such that

$$J[\bar{u}] = \min_{u \in \mathscr{A}_u} J[u] \tag{3.4}$$

subject to the side conditions

$$\bar{x}(0) = x_0, \tag{3.5a}$$

$$\int_0^T \langle w_j(t), (-\dot{\bar{x}} + A(t)\bar{x} + B(t)\bar{u})_i \rangle \, dt = 0 \tag{3.5b}$$

for each basis function $w_j \in S_m^\alpha$, $i = 1, \cdots, n$, $j = 1, \cdots, m$.

The analogues of Problems 1′ and 1″ are given in Definitions 3.2 and 3.3, respectively.

DEFINITION 3.2 (Problem 2′). Under the same assumptions as in Definition 3.1, find the $\bar{u}, \bar{x}, \bar{\lambda}$ and $\bar{\gamma}$ such that the Lagrangian (2.3) is extremized when λ is restricted to S_m^α ($\lambda \in S_m^\alpha$ will henceforth mean that each $\lambda_i \in S_m^\alpha$). Thus

$$L[\bar{u}, \bar{x}; \bar{\lambda}, \bar{\gamma}] = \max_{\substack{\lambda \in S_m^\alpha \\ \gamma \in R_n}} \min_{\substack{u \in \mathscr{A}_u \\ x \in \mathscr{A}_x}} L[u, x; \lambda, \gamma]. \tag{3.6}$$

DEFINITION 3.3 (Problem 2″). Under the assumptions of Definition 3.1, find the $\bar{\lambda}$ and $\bar{\gamma}$ such that

$$L[\bar{u}, \bar{x}; \bar{\lambda}, \bar{\gamma}] = \max_{\substack{\lambda \in S_m^\alpha \\ \gamma \in R_n}} L[u, x; \lambda, \gamma] \tag{3.7}$$

subject to the constraints

$$\frac{\partial L}{\partial x}[u, x; \lambda, \gamma] = 0, \tag{3.8}$$

$$\frac{\partial L}{\partial u}[u, x; \lambda, \gamma] = 0. \tag{3.9}$$

The equivalence of Problems 2 and 2′ follows immediately from the fundamental theorem of Lagrange multipliers. Further, the equivalence of Problems 2′ and 2″ is guaranteed by fundamental principles of the calculus of variations (see [4]). The three formulations of the finite-dimensional problem will be required in subsequent sections. In particular, the formulation of Problem 2′ will be required to generate certain key inequalities, essential in the convergence analysis of the numerical Ritz–Trefftz procedure. The third formulation (Problem 2″) is required in the development of a useful computational algorithm for the numerical solution of Problem 1.

3.3. Development of numerical algorithm. We consider now the development of a feasible computational algorithm using the problem formulation 2″. We begin by extremizing $L[u, x; \lambda, \gamma]$ for $\lambda_i \in S_m^\alpha$ and $\gamma \in R_n$, subject to the constraints (3.8) and (3.9). Thus, since $\lambda_i \in S_m^\alpha$, we can express λ as a linear combination of basis functions w_i, $i = 1, \cdots, m$. Consequently, we write

$$\lambda(t) = \sum_{j=1}^{m} c_j w_j(t), \qquad 0 \leqq t \leqq T, \tag{3.10}$$

where each c_j, $j = 1, \cdots, m$, is an n-vector. Now, since $x(T)$ is unconstrained, we conclude from (3.8) that

$$\begin{aligned} &\dot{\lambda}(t) + A^T(t)\lambda(t) + Q(t)x(t) = 0, \qquad 0 \leqq t \leqq T, \\ &\lambda(T) = 0, \qquad \gamma = -\lambda(0). \end{aligned} \tag{3.11}$$

Similarly, we deduce from condition (3.9) that

$$B^T(t)\lambda(t) + R(t)u(t) = 0, \qquad 0 \leqq t \leqq T. \tag{3.12}$$

From (3.11) and (3.12) we observe that L can be written as a functional of the Lagrange multiplier $\lambda(\cdot)$ only. Thus the problem is reduced to maximizing L over $\lambda(\cdot)$. We require, therefore, that

$$\frac{\partial L}{\partial \lambda} = 0, \tag{3.13}$$

where λ is defined in (3.10). Thus, we must first express L in terms of the c_j's. We have

$$\begin{aligned} L[u, x; \lambda, \gamma] &= \frac{1}{2}\int_0^T \langle u, Ru\rangle\, dt + \frac{1}{2}\int_0^T \langle x, Qx\rangle\, dt \\ &\quad + \int_0^T \langle \lambda, -\dot{x} + Ax + Bu\rangle\, dt - \langle \lambda(0), x(0) - x_0\rangle \\ &= J[u] - \langle \lambda(t), x(t)\rangle|_0^T + \int_0^T \langle \dot{\lambda} + A^T\lambda, x\rangle\, dt \\ &\quad + \int_0^T \langle B^T\lambda, u\rangle\, dt - \langle \lambda(0), x(0) - x_0\rangle. \end{aligned} \tag{3.14}$$

From conditions (3.11) and (3.12) we have

$$L[u, x; \lambda, \gamma] = -J[u] + \langle \lambda(0), x_0\rangle. \tag{3.15}$$

From (3.10) we obtain

$$\int_0^T \langle u(t), R(t)u(t)\rangle\, dt = \sum_{i=1}^m \sum_{j=1}^m c_i^T \left[\int_0^T E(t)w_i(t)w_j(t)\, dt\right] c_j, \tag{3.16}$$

where E is an $n \times n$ symmetric matrix given by $E(t) = B(t)R^{-1}(t)B^T(t)$. Similarly

$$\int_0^T \langle x(t), Q(t)x(t)\rangle\, dt = \int_0^T [\langle \dot{\lambda}, Q^{-1}(t)\dot{\lambda}\rangle + 2\langle \lambda, F(t)\dot{\lambda}\rangle + \langle \lambda(t), G(t)\lambda\rangle]\, dt \tag{3.17}$$

or

$$\begin{aligned} \int_0^T \langle x(t), Q(t)x(t)\rangle\, dt &= \sum_{i=1}^m \sum_{j=1}^m c_i \left[\int_0^T Q^{-1}(t)\dot{w}_i(t)\dot{w}_j(t)\, dt\right] c_j \\ &\quad + 2\sum_{i=1}^m \sum_{j=1}^m c_i^T \left[\int_0^T F(t)w_i(t)\dot{w}_j(t)\, dt\right] c_j \\ &\quad + \sum_{i=1}^m \sum_{j=1}^m c_i^T \left[\int_0^T G(t)w_i(t)w_j(t)\, dt\right] c_j, \end{aligned}$$

where $F(t) = A(t)Q^{-1}(t)$ and $G(t) = A(t)Q^{-1}(t)A^T(t)$. Finally, we have that

$$\langle \lambda(0), x_0\rangle = \left\langle \sum_{i=1}^m c_i w_i(0), x_0 \right\rangle. \tag{3.18}$$

Maximizing $\tilde{L}[c_1, \cdots, c_m] = L[u(c_1, \cdots, c_m), x(c_1, \cdots, c_m); \lambda(c_1, \cdots, c_m), \gamma(c_1, \cdots, c_m)]$ over c_i $(i = 1, \cdots, m)$, we obtain

$$\begin{aligned}\frac{\partial \tilde{L}}{\partial c_i} = -\sum_{j=1}^{m} \int_0^T & [(E(t) + G(t))w_i(t)w_j(t) + F(t)w_i(t)\dot{w}_j(t) \\ & + F^T(t)\dot{w}_i(t)w_j(t) + Q^{-1}(t)\dot{w}_i(t)\dot{w}_j(t)]\, dt\, c_j \\ & + w_i(0)x_0 = 0\end{aligned} \tag{3.19}$$

for $i = 1, \cdots, m$. We define H_{ij} by

$$H_{ij} = \int_0^T [(E + G)w_i w_j + F w_i \dot{w}_j + F^T \dot{w}_i \dot{w}_j + Q^{-1}\dot{w}_i \dot{w}_j]\, dt. \tag{3.20}$$

It follows that solving (3.19) for c_i, $i = 1, \cdots, m$, is equivalent to solving the linear system

$$Hy = b, \tag{3.21}$$

where the $(m \times n) \times (m \times n)$ symmetric matrix H is defined by

$$H = \begin{bmatrix} H_{11} & H_{12} & \cdots & H_{1m} \\ H_{21} & H_{22} & \cdots & H_{2m} \\ \vdots & & & \vdots \\ H_{m1} & \cdots & & H_{mm} \end{bmatrix}, \qquad H_{ij} = H_{ji}^t \tag{3.22}$$

and y and b are given by

$$y = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} \qquad b = \begin{bmatrix} w_1(0)x_0 \\ w_2(0)x_0 \\ \vdots \\ w_m(0)x_0 \end{bmatrix}. \tag{3.23}$$

We note that H possesses certain desirable numerical properties (such as sparseness, definiteness, etc.), which insures that (3.21) is numerically "well-conditioned". We discuss these aspects of the numerical algorithm described by (3.21) in greater detail in a forthcoming paper. If we now assume that $\bar{y}$, the solution of (3.21), is known exactly, then we write $\bar{y} = [\bar{c}_1^T, \cdots, \bar{c}_m^T]^T$ and set

$$\bar{\lambda}(t) = \sum_{i=1}^{m} \bar{c}_i w_i(t), \qquad \bar{u} = u_{\bar{\lambda}}, \quad \bar{x} = x_{\bar{\lambda}}, \tag{3.24}$$

where $\bar{x}$ and $\bar{u}$ are computed according to (3.11) and (3.12). In the remainder of the paper we shall be concerned with how "well" $J[\bar{u}]$ approximates $J[u^*]$, and, likewise, how "well" the $\bar{u}$ and $\bar{x}$ approximate u^* and x^*, respectively.

4. Cost functional convergence for pair $(\bar{u}, \bar{x})$. In this section we state and prove the key convergence theorems for the cost functional $J[\cdot]$, interpreting the results in a number of useful piecewise polynomial subspaces.

4.1. An order error bound for Ritz–Trefftz method. From the equivalence of Problem 2 with Problem 2″ we can conclude that $\bar{x}(0)$ is indeed x_0. Then it is clear that the final term $\langle \gamma, \bar{x}(0) - x_0 \rangle$ may be omitted and we write $L[\bar{u}, \bar{x}; \bar{\lambda}]$, replacing $L[\bar{u}, \bar{x}; \bar{\lambda}, \bar{\gamma}]$.

THEOREM 4.1. *Assume Problem* $1 \in PC^{\alpha-1}[0, T]$, *and let* $\bar{u}$, $\bar{x}$ *and* $\bar{\lambda}$ *be prescribed by the Ritz–Trefftz algorithm (defined in* (3.21) *to* (3.24)) *over an arbitrary* S_m^α. *Suppose* $\lambda_s \in S_m^\alpha$, *where* λ_s *approximates* λ^* *to order* h^α, $h = 1/m$. *Further, suppose* (u_s, x_s) *is the pair generated by* λ_s *according to* (3.11) *and* (3.12). *Then*

$$\|\varepsilon_u\|_2 = O(h^\alpha), \quad \|\varepsilon_x\|_2 = O(h^{\alpha-1}), \tag{4.1}$$

$$J[u^*] \geqq L[\bar{u}, \bar{x}; \bar{\lambda}] \geqq J[u^*] + L[\varepsilon_u, \varepsilon_x; \varepsilon_\lambda], \tag{4.2}$$

where $\varepsilon_u = u_s - u^*, \varepsilon_x = x_s - x^*$ *and* $\varepsilon_\lambda = \lambda_s - \lambda^*$, *with* $\|\cdot\|_2 = \left[\int [\cdot]^2\right]^{1/2}$.

Proof. Since Problem $1 \in PC^{\alpha-1}[0, T]$, we have $\lambda^* \in PC^\alpha[0, T]$. Hence, from the hypothesis there exists a $\lambda_s \in S_m^\alpha$ such that $\|\lambda_s - \lambda^*\|_2 = O(h^\alpha)$. Thus, from (3.11) and (3.12) we obtain

$$\|u_s - u^*\|_2 = \|R^{-1}B^T(\lambda_s - \lambda^*)\|_2 \leqq O(h^\alpha), \tag{4.3}$$

$$\|x_s - x^*\|_2 = \|Q^{-1}[(\dot{\lambda}_s - \dot{\lambda}^*) + A^T(\lambda_s - \lambda^*)]\|_2 \leqq O(h^{\alpha-1}) \tag{4.4}$$

which proves (4.1).

For the proof of (4.2) we recall formulations 1″ and 2″, whence

$$J[u^*] = L[u^*, x^*; \lambda^*, \gamma^*] \geqq L[\bar{u}, \bar{x}; \bar{\lambda}, \bar{\gamma}], \tag{4.5}$$

Since $S_m^\alpha \subset \mathscr{A}_\lambda$ (in general, $\mathscr{A}_\lambda$ will be a set of continuously differentiable functions on $[0, T]$). And, from (3.7) we conclude that $L[\bar{u}, \bar{x}; \bar{\lambda}] \geqq L[u_s, x_s; \lambda_s]$. But

$$\begin{aligned} L[u_s, x_s; \lambda_s] &= L[u^* + \varepsilon_u, x^* + \varepsilon_x; \lambda^* + \varepsilon_\lambda] \\ &= L[u^*, x^*; \lambda^*] + L[\varepsilon_u, \varepsilon_x; \varepsilon_\lambda] \\ &\quad + \int_0^T \langle u^*(t), R(t)\varepsilon_u(t)\rangle\, dt + \int_0^T \langle x^*(t), Q(t)\varepsilon_x(t)\rangle\, dt \\ &\quad + \int_0^T \langle \lambda^*(t), -\dot{\varepsilon}_x(t) + A(t)\varepsilon_x(t) + B(t)\varepsilon_u(t)\rangle\, dt. \end{aligned} \tag{4.6}$$

From (3.11), (3.12) and integration by parts we conclude that

$$\begin{aligned} &\int_0^T \langle \lambda^*(t), -\dot{\varepsilon}_x(t) + A(t)\varepsilon_x(t) + B(t)\varepsilon_u(t)\rangle\, dt \\ &\quad = -\int_0^T \langle u^*(t), R(t)\varepsilon_u(t)\rangle\, dt - \int_0^T \langle x^*(t), Q(t)\varepsilon_x(t)\rangle\, dt \end{aligned} \tag{4.7}$$

(note that $\varepsilon_x(0) = 0$ since $\lambda_s(0) = \lambda^*(0)$, $\lambda_s'(0) = \lambda^{*\prime}(0)$ by P4) which proves that

$$J[u^*] \geqq L[\bar{u}, \bar{x}; \bar{\lambda}] \geqq J[u^*] + L[\varepsilon_u, \varepsilon_x; \varepsilon_\lambda]. \tag{4.8}$$

We consider now an important corollary to the previous theorem.

COROLLARY 4.2. *Suppose the conditions of Theorem* 4.1 *hold. Then*

$$0 \leqq J[u^*] - L[\bar{u}, \bar{x}; \bar{\lambda}] \leqq O(h^{2(\alpha-1)}). \tag{4.9}$$

Proof. From (4.8) we observe that

$$0 \leqq J[u^*] - L[\bar{u}, \bar{x}; \bar{\lambda}] \leqq - L[\varepsilon_u, \varepsilon_x; \varepsilon_\lambda]. \tag{4.10}$$

But

$$\begin{aligned} L[\varepsilon_u, \varepsilon_x; \varepsilon_\lambda] = \frac{1}{2}\Big[\int_0^T \langle \varepsilon_u, R(t)\varepsilon_u \rangle \, dt + \int_0^T \langle \varepsilon_x, Q(t)\varepsilon_x \rangle \, dt \Big] \\ + \int_0^T \langle \varepsilon_\lambda, -\dot{\varepsilon}_x + A(t)\varepsilon_x + B(t)\varepsilon_u \rangle \, dt. \end{aligned} \tag{4.11}$$

Again integrating by parts and applying the order bounds to $\varepsilon_\lambda, \varepsilon_x$ and ε_u, we obtain

$$\begin{aligned} 0 \leqq J[u^*] - L[\bar{u}, \bar{x}; \bar{\lambda}] &\leqq O(h^{2(\alpha-1)}) + O(h^{2\alpha}) \\ &= O(h^{2(\alpha-1)}), \end{aligned}$$

and the corollary is proved.

Two important piecewise polynomial subspaces which are useful computationally are $S(\Pi)$ (the space of piecewise cubic splines on $[0, T]$) and $H(\Pi)$ (the piecewise cubic Hermite subspace on $[0, T]$). If, for example, Problem $1 \in PC^3[0, T]$, then $\alpha = 4$, and we have

$$0 \leqq J[u^*] - L[\bar{u}, \bar{x}; \bar{\lambda}] \leqq O(h^6), \qquad h = |\Pi|,$$

with $\bar{u}, \bar{x}, \bar{\lambda}$ generated over $S(\Pi)$. Similarly, for $\bar{u}, \bar{x}, \bar{\lambda}$ generated over $H(\Pi)$, we can show that

$$0 \leqq J[u^*] - L[\bar{u}, \bar{x}; \bar{\lambda}] \leqq O(h^6), \qquad h = |\Pi|.$$

(Note that $L[\bar{u}, \bar{x}; \bar{\lambda}]$ is just $\frac{1}{2}\int_0^T \langle \bar{u}, R\bar{u} \rangle \, dt + \frac{1}{2}\int_0^T \langle \bar{x}, Q\bar{x} \rangle \, dt$.)

5. Norm convergence of the pair $(\bar{u}, \bar{x})$. In this section we discuss the convergence of the computed pair $\bar{u}$ and $\bar{x}$ to the optimal pair u^* and x^*. We begin by proving a series of simple lemmas.

LEMMA 5.1. *Assume Problem* $1 \in PC^{\alpha-1}[0, T]$ *and let* $\bar{u}, \bar{x}$ *and* $\bar{\lambda}$ *be prescribed by the Ritz–Trefftz procedure over the space* S_m^α. *Then*

$$\begin{aligned} &\int_0^T \langle \lambda^*, -\dot{x} + A(t)\bar{x} + B(t)\bar{u} \rangle \, dt \\ &\qquad = 2J[u^*] - \int_0^T \langle u^*, R(t)\bar{u} \rangle \, dt - \int_0^T \langle x^*, Q(t)\bar{x} \rangle \, dt. \end{aligned} \tag{5.1}$$

Proof. Integration of the left side of (5.1) by parts yields

$$\int_0^T \langle \lambda^*(t), -\dot{\bar{x}} + A(t)\bar{x} + B(t)\bar{u} \rangle \, dt$$

$$= -\langle \lambda^*, \bar{x} \rangle|_0^T + \int_0^T \langle \dot{\lambda}^* + A^T(t)\lambda^*, \bar{x} \rangle \, dt + \int_0^T B^T\lambda^*, \bar{u} \rangle \, dt.$$

But, by transversality $\lambda^*(T) = 0$, and from (2.14) we conclude that $2J[u^*] = \langle \lambda^*(0), x(0) \rangle$. Since $\bar{x}(0) = x_0$, we have

$$2J[u^*] = -\langle \lambda^*, \bar{x} \rangle|_0^T.$$

The desired result now follows from the standard necessary conditions for Problem 1 (see (2.9) and (2.10)).

LEMMA 5.2. *Assume the hypotheses of Lemma* 5.1 *hold. In addition suppose that* λ_s *is the* S_m^α *approximation to* λ^*, *and suppose* (u_s, x_s) *is the pair generated by* λ_s *according to* (3.11) *and* (3.12). *Then*

$$\int_0^T \langle \lambda^*, -\dot{\bar{x}} + A(t)\bar{x} + B(t)\bar{u} \rangle \, dt = O(h^{\alpha-1}). \tag{5.2}$$

Proof. From the formulation of Problem 2, we deduce that

$$\int_0^T \langle w_j, (-\dot{\bar{x}} + A(t)\bar{x} + B(t)\bar{u})_i \rangle \, dt = 0$$

for each $w_j \in S_m^\alpha$, $i = 1, \cdots, n$, $j = 1, \cdots, m$. But $\lambda_s(t) = \sum_{j=1}^m c_j^s w_j(t)$ for some set c_j^s of n-vectors, $j = 1, \cdots, m$. Hence

$$\int_0^T \langle \lambda_s, -\dot{\bar{x}} + A(t)\bar{x} + B(t)\bar{u} \rangle \, dt = 0. \tag{5.3}$$

Hence

$$\int_0^T \langle \lambda^*, -\dot{\bar{x}} + A(t)\bar{x} + B(t)\bar{u} \rangle \, dt$$

$$= \int_0^T \langle \dot{\varepsilon}_\lambda, \bar{x} \rangle \, dt + \int_0^T \langle \varepsilon_\lambda, A\bar{x} \rangle \, dt + \int_0^T \langle \varepsilon_\lambda, B\bar{u} \rangle \, dt$$

$$\leqq \|\dot{\varepsilon}_\lambda\|_2 \|\bar{x}\|_2 + \|\varepsilon_\lambda\|_2 \|A\|_2 \|\bar{x}\|_2 \, \|\varepsilon_\lambda\|_2 \|B\|_2 \|\bar{u}\|_2.$$

But we know that $\|\varepsilon_\lambda\|_2 = O(h^\alpha)$ and thus $\|\dot{\varepsilon}_\lambda\|_2 = O(h^{\alpha-1})$. Since $\|\bar{x}\|_2$ and $\|\bar{u}\|_2$ are each bounded uniformly in h (Theorem 4.1), (5.2) is established and the proof is complete.

LEMMA 5.3. *Assume the hypotheses of Lemma* 5.1 *hold. Then*

$$\|\bar{u} - u^*\|_R^2 \equiv \int_0^T \langle \bar{u} - u^*, R(t)(\bar{u} - u^*) \rangle \, dt = O(h^{\alpha-1}), \tag{5.4}$$

$$\|\bar{x} - x^*\|_Q^2 \equiv \int_0^T \langle \bar{x} - x^*, Q(t)(\bar{x} - x^*) \rangle \, dt = O(h^{\alpha-1}). \tag{5.5}$$

Proof. We can easily verify the identity

$$\|u^* - \bar{u}\|_R^2 + \|x^* - \bar{x}\|_Q^2 = 2J[u^*] - 2\left[\int_0^T \langle \bar{u}, R(t)u^* \rangle\, dt + \int_0^T \langle \bar{x}, Q(t)x^* \rangle\, dt + 2J[\bar{u}, \bar{x}]\right].$$

Because of the equivalence of Problems 2, 2′ and 2″ we recall that $L[\bar{u}, \bar{x}; \bar{\lambda}, \bar{\gamma}] = J[\bar{u}, \bar{x}]$. Hence

$$\|u^* - \bar{u}\|_R^2 + \|x^* - \bar{x}\|_Q^2 \leqq 4J[u^*] - 2\left[\int_0^T \langle \bar{u}, R(t)u^* \rangle\, dt + \int_0^T \langle \bar{x}, Q(t)x^* \rangle\, dt\right] + O(h^{2(\alpha-1)}).$$

Now, applying the previous two lemmas, we obtain

$$0 \leqq \|u^* - \bar{u}\|_R^2 + \|x^* - \bar{x}\|_Q^2 \leqq 4J[u^*] - 4J[u^*] + O(h^{\alpha-1})$$

or

$$0 \leqq \|u^* - \bar{u}\|_R^2 + \|x^* - \bar{x}\|_Q^2 \leqq O(h^{\alpha-1})$$

from which (5.4) and (5.5) follow.

LEMMA 5.4. *Under the assumptions of Lemma* 5.2 *we assert that*

$$\int_0^T \langle \lambda^*, -\dot{x} + A(t)\bar{x} + B(t)\bar{u} \rangle\, dt \leqq \|\dot{\varepsilon}_\lambda\|_2 \|\bar{x} - x^*\|_2 + \|\varepsilon_\lambda\|_2 \|A\|_2 \|\bar{x} - x^*\|_2 + \|\varepsilon_\lambda\|_2 \|B\|_2 \|\bar{u} - u^*\|. \tag{5.6}$$

Proof. The lemma follows immediately from (5.3) and integration by parts.

THEOREM 5.5. *Under the hypotheses of Lemma* 5.2 *we assert that*

$$\|\bar{u} - u^*\|_R = O(h^{\alpha-1}), \tag{5.7}$$

$$\|\bar{x} - x^*\|_Q = O(h^{\alpha-1}). \tag{5.8}$$

Proof. We observed in proving Lemma 5.3 that

$$0 \leqq \|\bar{u} - u^*\|_R^2 + \|\bar{x} - x^*\|_Q^2 \leqq 4J[u^*] - 2\left[\int_0^T \langle \bar{u}, R(t)u^* \rangle\, dt + \int_0^T \langle \bar{x}, Q(t)x^* \rangle\, dt\right] + O(h^{2(\alpha-1)}).$$

Applying Lemmas 5.1, 5.2 and 5.4, we find that

$$\|\bar{u} - u^*\|_R^2 + \|\bar{x} - x^*\|_Q^2 \leqq 2[\|\dot{\varepsilon}_\lambda\|_2 \|\bar{x} - x^*\|_2 + \|\varepsilon_\lambda\|_2 \|A\|_2 \|\bar{x} - x^*\|_2 + \|\varepsilon_\lambda\|_2 \|B\|_2 \|\bar{u} - u^*\|_2] + O(h^{2(\alpha-1)}).$$

Now, since $R(\cdot)$ and $Q(\cdot)$ are positive definite, we can find positive constants v_1, v_2, q_1, q_2 such that

$$v_1 \|\bar{u} - u^*\|_2^2 \leqq \|\bar{u} - u^*\|_R^2 \leqq v_2 \|\bar{u} - u^*\|_2^2 \tag{5.9}$$

and

$$q_1\|\bar{x} - x^*\|_2^2 \leqq \|\bar{x} - x^*\|_Q^2 \leqq q_2\|\bar{x} - x^*\|_2^2. \tag{5.10}$$

Therefore, we obtain the inequality

$$\begin{aligned}\|\bar{u} - u^*\|_R^2 + \|\bar{x} - x^*\|_Q^2 \\ \leqq O(h^{\alpha-1})\|\bar{u} - u^*\|_R + O(h^{\alpha-1})\|\bar{x} - x^*\|_Q + O(h^{2(\alpha-1)}).\end{aligned} \tag{5.11}$$

Set $\delta_u = \|\bar{u} - u^*\|_R$ and $\delta_x = \|\bar{x} - x^*\|_Q$. Then (5.11) becomes

$$\delta_u^2 + O(h^{\alpha-1})\delta_u + \delta_x^2 + O(h^{\alpha-1})\delta_x \leqq O(h^{2(\alpha-1)})$$

or, equivalently,

$$[\delta_u - O(h^{\alpha-1})]^2 + [\delta_x - O(h^{\alpha-1})]^2 \leqq O(h^{2(\alpha-1)}).$$

Hence

$$\begin{aligned}\delta_u &= \|\bar{u} - u^*\|_R = O(h^{\alpha-1}), \\ \delta_x &= \|\bar{x} - x^*\|_Q = O(h^{\alpha-1})\end{aligned}$$

and the theorem is established.

COROLLARY 5.6. $\|\bar{u} - u^*\|_2 = O(h^{\alpha-1})$, $\|\bar{X} - x^*\|_2 = O(h^{\alpha-1})$.

Proof. The proof is trivial and is therefore omitted.

6. Remarks on the numerical properties of the method. The above theorems establish that the Ritz–Trefftz method has the unique theoretical property that it converges to the optimal value of the cost functional at a rate equal to the square of the order of the approximating subspaces. This advantage is not shared by any other approximate method, direct or indirect. In fact, the standard Ritz direct method is not known to converge for variational problems with nonholonomic constraints.

The numerical implications are obvious. The authors have a working program for the method and have compared it with the usual numerical approximation of the Ricatti equation for numerous examples. The method is significantly faster, except on trivial problems where the two approaches yield approximately the same results. Since the resulting algebraic problem is definite, it can be solved by a Cholesky decomposition without pivoting, hence the band structure is preserved (providing a patch basis is used). Thus the numerical algorithm is extremely stable and fast. Also, reasonable bounds on the condition number can be computed.

In [13] we discuss the numerical utility of the algorithm in considerable detail. Additional convergence results for $(\bar{u}, x_{\bar{u}})$ and $J[\bar{u}; x_{\bar{u}}]$ were obtained and appear in [14].

REFERENCES

[1] J. H. AHLBERG, E. N. NILSON AND J. L. WALSH, *The Theory of Splines and Their Applications*, Academic Press, New York, 1967.

[2] M. ATHANS AND P. L. FALB, *Optimal Control: An Introduction to the Theory and Its Applications*, McGraw-Hill, New York, 1966.

[3] G. BIRKHOFF, M. SCHULTZ AND R. VARGA, *Piecewise Hermite interpolation in one and two variables with applications to partial differential equations*, Numer. Math., 11 (1968), pp. 232–256.

[4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. I, Interscience, New York, 1953.
[5] V. W. EVELEIGH, *Adaptive Control and Optimization Techniques*, McGraw-Hill, New York, 1967.
[6] P. L. FALB, *Direct Methods in Optimal Control*, to appear.
[7] G. HADLEY, *Nonlinear and Dynamic Programming*, Addison-Wesley, Reading, Mass., 1964.
[8] O. G. JOHNSON, *Error bounds for Sturm–Liouville eigenvalues by several piecewise cubic Rayleigh–Ritz methods*, SIAM J. Numer. Anal., 6 (1969), pp. 317–333.
[9] R. C. K. LEE, *Optimal Estimation, Identification and Control*, Research Monograph No. 28, M.I.T. Press, Cambridge, Mass., 1964.
[10] D. G. LUENBENGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
[11] E. TREFFTZ, *Konvergenz und Fehlerschätzung beim Ritzschen Verfahren*, Math. Ann., 100 (1928), pp. 503–521.
[12] A. P. SAGE, *Optimum Systems Control*, Prentice-Hall, Englewood Cliffs, N.J., 1968.
[13] W. E. BOSARGE, JR. AND O. G. JOHNSON, *Numerical properties of the Ritz–Trefftz algorithm for optimal control*, IBM DPD Tech. Rep. 320.2376, 1970; Math. Comp., to appear.
[14] ———, *Direct method approximation to the state regulator control problem using a Ritz–Trefftz suboptimal control*, Proc. Joint Automatic Control Conference, 1970, pp. 51–56.

SIAM J. CONTROL
Vol. 9, No. 1, February 1971

BOUNDARY VALUE CONTROL OF THE HIGHER-DIMENSIONAL WAVE EQUATION*

DAVID L. RUSSELL†

1. Introduction. Many important control processes can be described approximately by means of partial differential equations with control parameters appearing in the boundary conditions. For example, a triangular airplane wing may be equipped with ailerons on the trailing edge. An idealized model for such a plant would involve the partial differential equations which describe the motion of a plate with arbitrary control functions appearing in the boundary conditions along one of the sides of the triangle. Many other examples could be given.

Linear hyperbolic problems in one space dimension have been studied rather extensively; see, e.g., [1], [2], [3], [4], [5]. Here the theory is relatively uncomplicated. One can study questions of controllability using the geometric techniques based on characteristic curves or the more algebraic techniques based on the theory of nonharmonic Fourier series. One obtains not only theorems asserting the existence of controls transferring one state to another within a finite time period but also constructive proofs of these theorems which can be adapted to yield numerical techniques whereby the appropriate control functions can be calculated. The papers of Grainger [4] and Cirina [5] are noteworthy in this respect. Cirina's paper shows that such methods can even be used for quasilinear systems.

The theory is not nearly as complete for problems involving two or more space variables. The reasons why this should be so become apparent when one compares Chaps. V and VI of the treatise [6] of Courant–Hilbert. Some results have been obtained in this area by Fattorini [7] who considers, for the most part, boundary value control problems wherein the controls can be described by finitely many functions of the time t—physically the most realistic situation.

The purpose of the present paper is to study hyperbolic problems in several space dimensions using certain uniqueness theorems due to Holmgren [8] and John [9]. Using these results we can obtain very explicit estimates on the length of time required to transfer a given state into an arbitrarily small neighborhood of any other state using boundary value controls restricted to a subset of the boundary of the region in question. More specifically, we are able to show for the wave equations in 3 or fewer space variables that the system can be controlled in any time T which exceeds twice the wave propagation time from the boundary set where controls are applied to the rest of the physical medium. It should be noted that such a result is in agreement with known results [1], [2] for the case of a single space dimension.

2. The control problem. Let Ω be a bounded, open connected domain in R^n whose boundary is an analytic surface Γ of dimension $n-1$. We indicate

* Received by the editors February 12, 1970.

† Departments of Mathematics and Computer Sciences, University of Wisconsin, Madison, Wisconsin 53706, and Research Consultant, Honeywell Inc., St. Paul, Minnesota 55113. This work was supported in part by the National Science Foundation under Grant GP-11495.

points in Ω by

$$x = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^n \end{pmatrix}.$$

The boundary surface Γ is parametrized by an $(n-1)$-dimensional vector variable s. Integrals over Ω will be denoted by $\int_\Omega (\cdot)\, dx$ while integrals over Γ will be written $\int_\Gamma (\cdot)\, ds$. If we wish to indicate a point in R^n which lies on Γ we will write $x(s)$. The surface Γ, being analytic, has everywhere a unique unit outward normal vector which we will indicate by $\eta(s)$.

We consider a second order linear hyperbolic partial differential equation

$$L(w) = \rho(x)w_{tt} - \sum_{i,j=1}^{n} (\alpha_{ij}(x)w_i)_j = 0. \tag{2.1}$$

The subscripts i, j indicate partial differentiation with respect to x^i, x^j, respectively. The coefficients $\rho(x)$, $\alpha_{ij}(x)$ are real analytic in some open subset of R^n which includes $\Omega \cup \Gamma$. Moreover, if $A(x)$ is the $n \times n$ matrix with entries $\alpha_{ij}(x)$, $A(x)$ is symmetric and there are positive numbers ρ_0 and α_0 such that

$$\rho(x) \geqq \rho_0, \qquad x \in \Omega,$$
$$v'A(x)v \geqq \alpha_0 \|v\|^2, \qquad x \in \Omega, \quad v \in R^n.$$

Let Γ_1 be a relatively open subset of Γ. For $T > 0$, we denote by F the space of all C^∞ functions $f: \Gamma \times [0, T] \to R^1$ with the property that f vanishes outside some compact subset of $\Gamma_1 \times (0, T)$ (this set varies with f). We pose for (2.1) the initial boundary value problem

$$\begin{aligned} &w(x, 0) \equiv w_t(x, 0) \equiv 0, \qquad x \in \Omega, \\ &w_x(x(s), t)A(x(s))\eta(s) \equiv f(s, t), \qquad (x(s), t) \in \Gamma \times [0, T]. \end{aligned} \tag{2.2}$$

The symbol w_x denotes the row vector of spatial partial derivatives of w:

$$w_x = (w_1, w_2, \cdots, w_n).$$

With these assumptions it is known that the initial boundary value problem (2.1), (2.2) has a unique solution $w^f(x, t)$ which lies in the class $C^\infty((\Omega \cup \Gamma) \times [0, T])$. The reader is referred to the papers of Friedrichs, Lax and Duff [10], [11], [12].

If $f(s, t) \equiv 0$ for $t_1 \leqq t \leqq t_2$, then $w^f(\cdot, t)$ can be considered as a vector-valued function with range in $L^2(\Omega)$ which solves the evolution equation

$$d^2w/dt^2 + Bw = 0, \qquad t_1 \leqq t \leqq t_2, \tag{2.3}$$

where B is the unbounded operator on $L^2(\Omega)$ which is the unique self-adjoint extension of the operator $\sum_{i,j=1}^{n} (\alpha_{ij}(x)w_i)_j$ defined on twice continuously

differentiable functions $w(x)$ satisfying the boundary conditions

$$w_x(x(s))A(x(s))\eta(s) = 0, \qquad x(s) \in \Gamma.$$

Let H_1 denote the space of pairs of real-valued functions $w(x)$, $w_t(x)$ defined on Ω with $w_t(x)$ square integrable and $w(x)$ having square integrable derivatives:

$$\int_\Omega \left(w_t(x)^2 + \sum_{i=1}^{n} w_i(x)^2 + w(x)^2 \right) dx < \infty.$$

Let H_E denote the space of equivalence classes of H_1 modulo the zero energy states $w_t(x) \equiv 0$, $w(x) \equiv$ const. The "energy"

$$\mathscr{E}(w, w_t) = \int_\Omega (\rho(x)(w_t(x))^2 + w_x(x)A(x)w_x(x)')\, dx$$

is a constant on each such equivalence class. H_E is a Hilbert space with the inner product

$$\langle (w, w_t); (v, v_t) \rangle_E = \int_\Omega (\rho(x)w_t(x)v_t(x) + w_x(x)A(x)v_x(x))\, dx$$

and resulting norm

$$\|(w, w_t)\|_E = \sqrt{\langle (w, w_t); (w, w_t) \rangle_E} = \sqrt{\mathscr{E}(w, w_t)}.$$

We shall not stress the distinction between H_1 and H_E where unnecessary, and we shall say $(w, w_t) \in H_E$ if the equivalence class of (w, w_t) is a member of H_E.

For each $f \in F$ the corresponding solution $w^f(x, t)$ of (2.1), (2.2) is such that $(w^f(\cdot, T), w_t^f(\cdot, T)) \in H_E$. In fact, if we put

$$R_T = \{(w^f(\cdot, T), w_t^f(\cdot, T)) | f \in F\},$$

then R_T is a subspace of H_E which we will call the *reachable space.* Following others we make the following definition.

DEFINITION. The control system $\{(2.1), (2.2), f \in F\}$ is *approximately controllable* in time $T > 0$ if R_T is dense in H_E relative to the topology induced by the norm $\|\cdot\|_E$.

We shall conclude this section with a theorem which relates approximate controllability to "observability." (Cf. parallel results for ordinary differential equations [13].)

THEOREM 1. *Let $(\hat{v}, \hat{v}_t) \in H_E$ be such that both $\hat{v}$ and $\hat{v}_t$ lie in $C^\infty(\Omega \cup \Gamma)$ and $\hat{v}$ satisfies the consistency conditions*

$$\hat{v}_x(x(s))A(x(s))\eta(x(s)) = 0, \qquad \hat{v}_{tx}(x(s))A(x(s))\eta(x(s)) = 0, \qquad x(s) \in \Gamma.$$

Let $v(x, t)$ be the unique C^∞ solution of $L(v) = 0$ which satisfies the boundary conditions

$$v_x(x(s), t)A(x(s))\eta(x(s)) \equiv 0, \qquad (x(s), t) \in \Gamma \times [0, T] \tag{2.4}$$

and the terminal conditions

$$v(x, T) \equiv \hat{v}(x), \qquad v_t(x, T) \equiv \hat{v}_t(x). \tag{2.5}$$

Then $(\hat{v}, \hat{v}_t) \in (R_T)^\perp$ *in* H_E *if and only if* $v_t(x(s), t) \equiv 0$, $(x(s), t) \in \Gamma_1 \times [0, T]$.

Proof. For $f \in F$ we have

$$0 = \int_{\Omega \times [0,T]} (v_t L(w^f) + w_t^f L(v))\, dx\, dt$$

$$= \int_{\Omega \times [0,T]} \operatorname{div}_{x,t} \begin{pmatrix} -v_t(w_x^f a_1) - w_t^f(v_x a_1) \\ \vdots \\ -v_t(w_x^f a_n) - w_t^f(v_x a_n) \\ \rho w_t^f v_t + w_x^f A v_x' \end{pmatrix} dx\, dt, \tag{2.6}$$

where, for convenience, we have suppressed the arguments in the integrand and the column vectors $a_i(x)$ are the columns of the symmetric matrix $A(x)$:

$$A(x) = (a_1(x), a_2(x), \cdots, a_n(x)).$$

Applying the divergence theorem to the second member of (2.6) we obtain

$$0 = \int_{\Omega \times \{T\}} (\rho w_t^f v_t + w_x^f A v_x')\, dx - \int_{\Omega \times \{0\}} (\rho w_t^f v_t + w_x^f A v_x')\, dx$$

$$- \int_{\Gamma \times [0,T]} (v_t(w_x^f A\eta) + w_t^f(v_x A\eta))\, ds\, dt. \tag{2.7}$$

Using (2.5), (2.2), respectively, in the first two members of (2.7) and (2.2), (2.4) in the third member, we obtain

$$\int_\Omega (\rho(x) w_t^f(x, T)\hat{v}_t(x) + w_x^f(x, T) A(x) \hat{v}_x(x)')\, dx = \int_{\Gamma_1 \times [0,T]} (v_t f)\, ds\, dt. \tag{2.8}$$

From the definition of $\langle \cdot, \cdot \rangle_E$, we see that (2.8) becomes

$$\langle (w^f(\cdot, T), w_t^f(\cdot, T)); (\hat{v}, \hat{v}_t) \rangle_E = \int_{\Gamma_1 \times [0,T]} (v_t f)\, ds\, dt. \tag{2.9}$$

The right-hand side of (2.9) vanishes for all $f \in F$ if and only if $v_t(x(s), t) \equiv 0$, $(x(s), t) \in \Gamma_1 \times [0, T]$, and thus the proof is complete.

Theorem 1 is fundamental in the proofs of the controllability theorems of the subsequent sections.

3. The time T_0. In order to state and prove our theorems on approximate controllability of (2.1), (2.2) we must employ the concept of a *characteristic surface* for (2.1) in R^{n+1}. This concept is treated in detail in [6], for example, but we give a brief description to make our presentation somewhat self-contained.

Let S be a surface in R^{n+1} given by

$$S = \{(x, t) | \Phi(x, t) = 0\},$$

where $\Phi(x, t)$ is a smooth real-valued function of $n + 1$ variables. We define the *characteristic form*

$$\chi(\Phi, x, t) = \rho(x)(\Phi_t(x, t))^2 - \Phi_x(x, t) A(x) \Phi_x'(x, t).$$

The surface S is: characteristic if $\chi(\Phi, x, t) \equiv 0$ for $(x, t) \in S$; uniformly space-like if there exists a $\delta > 0$ such that $\chi(\Phi, x, t) \geqq \delta$ for $(x, t) \in S$; uniformly time-like if

there exists a $\delta > 0$ such that $\chi(\Phi, x, t) \leqq -\delta$ for $(x, t) \in S$. For what is usually called the wave equation $\rho(x) \equiv 1$ and $A(x) \equiv I$, the $n \times n$ identity matrix, a surface is characteristic if and only if it everywhere makes an angle of 45° with any intersecting surface $t = \text{const}$. It is this special case that we shall use in our diagrams since it is less confusing than the general case.

Let $(x_0, t_0) \in (\Omega \cup \Gamma) \times [0, T]$. We define the *forward cone of influence of* (x_0, t_0) to be the subset $K^+(x_0, t_0)$, the largest closed subset of $(\Omega \cup \Gamma) \times [t_0, T]$ which contains (x_0, t_0) and does not meet any uniformly space-like surface passing through (x_0, t_0). Similarly we define $K^-(x_0, t_0)$, the *backward cone of influence of* (x_0, t_0), by replacing $[t_0, T]$ with $[0, t_0]$. It is easy to see that $K^+(x_0, t_0)$ and $K^-(x_0, t_0)$ have characteristic boundary surfaces. When $\rho(x) \equiv 1$, $A(x) \equiv I$, we have $K^+(x_0, 0) = \{(x, t) \in (\Omega \cup \Gamma) \times [0, T] | t^2 - \|x - x_0\|^2 \geqq 0\}$. If G is a subset of $\Omega \cup \Gamma$, we define forward and backward cones of influence of (G, t_0) by

$$K^+(G, t_0) = \overline{\bigcup_{x_0 \in G} K^+(x_0, t_0)}, \qquad K^-(G, t_0) = \overline{\bigcup_{x_0 \in G} K^-(x_0, t_0)}.$$

Let t_0, t_1 lie in $[0, T]$ and let $G \subseteq \Omega \cup \Gamma$. We define

$$K(G, t_0, t_1) = K^+(G, t_0) \cap K^-(G, t_1).$$

Since the coefficients of the operator L do not depend upon t, $K(G, t_0, t_1)$ is symmetric about the plane $t = \frac{1}{2}(t_0 + t_1)$.

The fact that $A(x)$ is uniformly positive definite can be used to prove that there is a least time $T_0 > 0$ such that $K^+(\Gamma_1, 0)$ includes the set $\Omega \times \{T_0\}$. Then $K(\Gamma_1, 0, 2T_0)$ also includes $\Omega \times \{T_0\}$. If $T > 2T_0$, there is an $\alpha > 0$ such that $K(\Gamma_1, 0, T)$ includes $\Omega \times \{t\}$ for $|t - T/2| < \alpha$. If $T < 2T_0$, the set

$$J(T/2) = \Omega \times \{T/2\} - K(\Gamma_1, 0, T)$$

is a nonempty set. For $\rho(x) \equiv 1$, $\Omega =$ unit disc in R^2, Figs. 1–4 illustrate the geometry of the situations described above both when $\Gamma_1 = \Gamma$ and when Γ_1 is a small subarc of Γ.

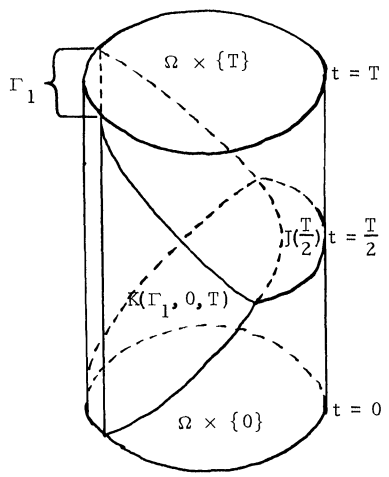


FIG. 1. Γ_1 *is a subarc of* Γ, $T < 2T_0$

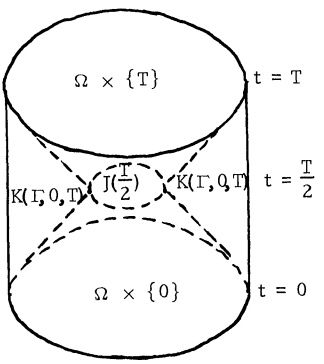


FIG. 2. $\Gamma_1 = \Gamma$, $T < 2T_0$

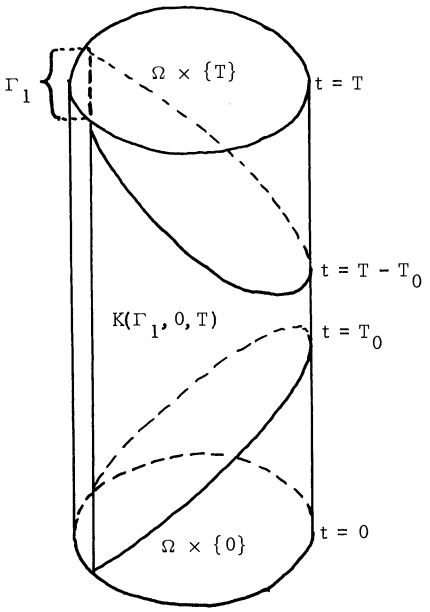


FIG. 3. Γ_1 *is a subarc of* Γ, $T > 2T_0$

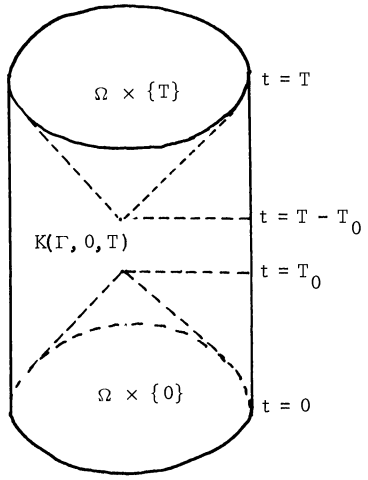


FIG. 4. $\Gamma_1 = \Gamma$, $T > 2T_0$

4. Noncontrollability for $T < 2T_0$. Since the reachable set R_T is a linear subspace of the Hilbert space H_E, R_T fails to be dense in H_E just in case there is a nonzero element of H_E which is orthogonal to all elements of R_T. In view of our definition of H_E in terms of equivalence classes of states in H_1 modulo the zero energy states, we see that R_T is dense in H_E if and only if the equations

$$\int_\Omega (\rho(x)w_t^f(x, T)\hat{v}_t(x) + w_x^f(x, T)A(x)\hat{v}_x(x)')\,dx = 0, \qquad f \in F, \tag{4.1}$$

where $\hat{v}$, $\hat{v}_t$ is a fixed element of H_1, imply that $\hat{v}_t \equiv 0$, $\hat{v} = \text{const.}$

When $T < 2T_0$ the subset J of Ω given by

$$J = \{x | (x, T/2) \in J(T/2)\}$$

is a nonempty open set. Let $\tilde{v}(x)$, $\tilde{v}_t(x)$ be a state in H_1 such that (i) $\tilde{v}(x)$, $\tilde{v}_t(x)$ has nonzero energy norm, (ii) $\tilde{v}(x)$, $\tilde{v}_t(x) \in C^\infty(\Omega)$ and vanish outside a compact subset of the interior of J.

We now permit the state $\tilde{v}(x)$, $\tilde{v}_t(x)$ to evolve, via the partial differential equation with

$$v\left(x, \frac{T}{2}\right) = \tilde{v}(x), \qquad v_t\left(x, \frac{T}{2}\right) = \tilde{v}_t(x).$$

We put

$$\hat{v}(x) = v(x, T), \qquad \hat{v}_t(x) = v_t(x, T) \tag{4.2}$$

and note that (2.1), (2.4) imply conservation of energy so that

$$\|(\hat{v}, \hat{v}_t)\|_E = \|(\tilde{v}, \tilde{v}_t)\|_E \neq 0.$$

Now our requirement (ii) on $\tilde{v}$, $\tilde{v}_t$ guarantees that $v(x, t) \in C^\infty(\Omega \times [0, T])$, as does $w^f(x, t)$ for each $f \in F$. This enables us to apply Theorem 1 to show that (4.1) holds if and only if $v_t(x(s), t) \equiv 0$ for $(x(s), t) \in \Gamma_1 \times [0, T]$.

It is easily seen that the fact that $K(\Gamma_1, 0, T) \cap J = \varnothing$ implies that the interior of $K(\Gamma_1, 0, T)$ does not meet $K^+(J, T/2) \cup K^-(J, T/2)$, the cone of influence of $(J, T/2)$. Well-known results for hyperbolic partial differential equations [6] then show that $v(x, t) \equiv 0$, $v_t(x, t) \equiv 0$ in $K(\Gamma_1, 0, T)$. But $\Gamma_1 \times [0, T] \subseteq K(\Gamma_2, 0, T)$ so we conclude that $v_t(x(s), t) \equiv 0$, $(x(s), t) \in \Gamma_1 \times [0, T]$. Therefore (4.1) must hold for the nonzero energy state $\hat{v}$, $\hat{v}_t$ and we have proved the following theorem.

THEOREM 2. *The system* (2.1), (2.2) *is not approximately controllable in time* T *if* $T < 2T_0$.

This theorem may be compared with comparable results [1], [2] for hyperbolic systems in one space dimension.

5. The Holmgren–Fritz John uniqueness theorem. The uniqueness theorems of Holmgren and Fritz John [8], [9], applied to the case we have in mind, reduce to the following theorem.

THEOREM 3. *Let* $u(x, t)$ *be a twice continuously differentiable solution of* $L(u) = 0$ (*cf.* (2.1)) *in* $K(\Gamma_1, t_0, t_1)$, $[t_0, t_1] \subseteq [0, T]$, *with*

$$u_x(x(s), t)A(x(s))\eta(x(s)) \equiv 0, \qquad u(x(s), t) \equiv 0, \quad (x(s), t) \in \Gamma_1 \times [t_0, t_1]. \tag{5.1}$$

Then

$$u(x, t) \equiv 0 \quad \text{in } K(\Gamma_1, t_0, t_1). \tag{5.2}$$

The proof of this theorem is detailed in the works cited. However, we need to strengthen the theorem somewhat for our needs and this strengthening requires that we have some details of the proof. For this reason we give a short proof of Theorem 3. An important part of the proof is the lemma stated below, which we do not prove. See [9] for details in certain cases.

LEMMA. *If* $(\bar{x}, \bar{t})$ *lies in the interior of* $K(\Gamma_1, t_0, t_1)$, *there is a uniformly time-like family of surfaces* $S(\lambda)$, $0 \leqq \lambda \leqq 1$, *with the following properties*:

(i) $S(\lambda)$ *is a compact subset of a relatively open analytic* $(n-1)$*-dimensional surface*;

(ii) $S(\lambda)$ *varies analytically with respect to* λ, $0 \leqq \lambda \leqq 1$;

(iii) $S(\lambda) \subseteq K(\Gamma_1, t_0, t_1)$, $0 \leqq \lambda \leqq 1$, *and* $S(0)$ *is a subset of the interior of* $\Gamma_1 \times [0, T]$;

(iv) *If* $0 \leqq \lambda \leqq 1$, *then* $S(0) \cup S(\lambda)$ *is the boundary of an open subset* $D(\lambda) \subseteq K(\Gamma_1, t_0, t_1)$ *and* $(\bar{x}, \bar{t}) \in D(1)$.

Assuming this lemma, we proceed.

Proof of Theorem 3. The uniformly time-like character of the surfaces $S(\lambda)$ together with the analyticity of these surfaces enables one to employ the Cauchy–Kowalewski theorem [6] to show that there are n-dimensional neighborhoods $N(\lambda)$ of the surfaces $S(\lambda)$ such that if analytic Cauchy data for z are prescribed on $S(\lambda)$, there will be a corresponding unique analytic solution $z(x, t)$ of $L(z) = 0$ in $N(\lambda)$. Since the equation $L(z) = 0$ is linear, $N(\lambda)$ depends only on $S(\lambda)$, not on the particular Cauchy data. Moreover, $N(\lambda)$ varies continuously with λ, $0 \leqq \lambda \leqq 1$.

Thus for sufficiently small $\lambda > 0$, $S(0) \subseteq N(\lambda)$ and the analytic solution $z(x, t)$ is defined throughout the domain $D(\lambda)$ which is bounded by $S(0)$ and $S(\lambda)$. We consider the identity

$$
\begin{aligned}
0 &= \int_{D(\lambda)} (uL(z) - zL(u))\, dx\, dt \\
&= \int_{D(\lambda)} \operatorname*{div}_{x,t} \begin{pmatrix} -uz_x a_1 + zu_x a_1 \\ \vdots \qquad\qquad \vdots \\ -uz_x a_n + zu_x a_n \\ \rho u z_t - \rho z u_t \end{pmatrix} dx\, dt \qquad (5.3) \\
&= \int_{S(0)\cup S(\lambda)} (-uz_x A\eta + zu_x A\eta + \eta_0 \rho u z_t - \eta_0 \rho z u_t)\, d\sigma,
\end{aligned}
$$

where $d\sigma$ denotes integration with respect to surface area on $S(0) \cup S(\lambda)$ and $\begin{pmatrix} \eta \\ \eta_0 \end{pmatrix}$ is the unit outward normal to $S(0) \cup S(\lambda)$ in R^{n+1}, defined in the relative interiors of $S(0)$ and $S(\lambda)$. We observe that $\eta_0 \equiv 0$ on $S(0)$, and that (5.1) holds on $S(0)$ since $S(0) \subseteq \Gamma_1 \times [0, T]$. Thus (5.3) reduces to

$$
\int_{S(\lambda)} (u(\eta_0 \rho z_t - z_x A\eta) + z(\eta_0 \rho u_t - u_x A\eta))\, d\sigma = 0. \qquad (5.4)
$$

We choose analytic Cauchy data for z on $S(\lambda)$ as follows: we put $z \equiv 0$ on $S(\lambda)$ and we take the normal derivative of z across $S(\lambda)$ to be an arbitrary real-valued analytic function α, i.e.,

$$
z(x(\sigma), t(\sigma)) = 0, \qquad (x(\sigma), t(\sigma)) \in S(\lambda), \qquad (5.5)
$$

$$
z_x(x(\sigma), t(\sigma))\eta(\sigma) + z_t(x(\sigma), t(\sigma))\eta_0(\sigma) = \alpha(\sigma), \quad (x(\sigma), t(\sigma)) \in S(\lambda). \qquad (5.6)
$$

The equations (5.5), (5.6) together imply that

$$
(z_x(x(\sigma), t(\sigma)), z_t(x(\sigma), t(\sigma))) \equiv \alpha(\sigma)(\eta'(\sigma), \eta_0(\sigma)), \qquad (x(\sigma), t(\sigma)) \in S(\lambda). \qquad (5.7)
$$

Substituting (5.7) and (5.5) in (5.4) we obtain

$$
\int_{S(\lambda)} u\alpha(\rho(\eta_0)^2 - \eta' A\eta)\, d\sigma = 0. \qquad (5.8)
$$

Since $S(\lambda)$ is uniformly time-like, we have

$$
\rho(x(\sigma))(\eta_0(\sigma))^2 - \eta'(\sigma)A(x(\sigma))\eta(\sigma) \equiv \beta(\sigma) \leqq -\beta_0 < 0 \qquad (5.9)
$$

for all values of the vector σ parametrizing $S(\lambda)$. The equation (5.8) becomes

$$
\int_{S(\lambda)} u\alpha\beta\, d\sigma = 0.
$$

Since this equation holds for all real analytic functions α, we conclude

$$
u(x(\sigma), t(\sigma))\beta(\sigma) \equiv 0, \qquad (x(\sigma), t(\sigma)) \in S(\lambda),
$$

and, since (5.9) shows that β never vanishes, we have

$$u(x(\sigma), t(\sigma)) \equiv 0, \qquad (x(\sigma), t(\sigma)) \in S(\lambda).$$

Repeating this argument on surfaces $S(\mu)$, $0 < \mu < \lambda$, which sweep out the interior of $D(\lambda)$, we conclude

$$u(x, t) \equiv 0, \qquad (x, t) \in D(\lambda).$$

We now let I denote the largest subinterval of $[0, 1]$ which includes 0 and has the property that $u \equiv 0$ on $S(\lambda)$ if $\lambda \in I$. We have seen above that I is nonempty. Essentially the same technique can be used to show that I is open. But it is obvious that I is closed, and we conclude, from the connectedness of $[0, 1]$, that $I = [0, 1]$. Thus

$$u(x, t) \equiv 0, \qquad (x, t) \in D(1),$$

and, since $(\bar{x}, \bar{t}) \in D(1)$, we have

$$u(\bar{x}, \bar{t}) \equiv 0.$$

Since $(\bar{x}, \bar{t})$ is an arbitrary point in the interior of $K(\Gamma_1, t_0, t_1)$ and since u is continuous in $K(\Gamma_1, t_0, t_1)$, we see that (5.2) follows and the proof is complete.

6. Controllability for $T > 2T_0$, $n \leqq 3$. Let the state $(\hat{v}, \hat{v}_t)$ lie in the finite energy space H_E, and suppose that for all $f \in F$ we have

$$\int_\Omega (\rho(x) w_t^f(x, T)\hat{v}_t(x) + W_x^f(x)A(x)\hat{v}_x'(x))\, dx = 0. \tag{6.1}$$

If this implies $\hat{v}_t(x) \equiv 0$, $\hat{v}(x) \equiv$ const., then R_T is dense in the Hilbert space H_E and we have approximate controllability.

Let $v(x, t)$ be the generalized solution in $\Omega \times [0, T]$ of the partial differential equation $L(v) = 0$ corresponding to homogeneous boundary conditions (2.4). If $v(x, t)$ were smooth, say $v \in C^3((\Omega \cup \Gamma) \times [0, T])$, the proof of our controllability result would not be difficult. Applying Theorem 1 we would get $v_t(x(s), t) \equiv 0$ for $(x(s), t) \in \Gamma_1 \times [0, T]$. Putting $u(x, t) \equiv v_t(x, t)$, we would have a solution of $L(u) = 0$ satisfying the hypotheses (5.1) of Theorem 3 and we could conclude $v_t \equiv u \equiv 0$ in $K(\Gamma_1, 0, T)$. If $T > 2T_0$, the set $K(\Gamma_1, 0, T)$ includes $\Omega \times [T/2 - \varepsilon, T/2 + \varepsilon]$ for some $\varepsilon > 0$. If v_t vanishes in $\Omega \times [T/2 - \varepsilon, T/2 + \varepsilon]$, then $v_{tt}(x, T/2) \equiv 0$ which would imply $\sum_{i,j=1}^n (\alpha_{ij}(x)v_i(x, T/2))_j \equiv 0$. Thus $v(x, T/2)$ would be a solution of the elliptic boundary value problem

$$\sum_{i,j=1}^{n} \left(\alpha_{ij}(x)v_i\left(x, \frac{T}{2}\right)\right)_j = 0, \qquad x \in \Omega, \tag{6.2}$$

$$v_x\left(x(s), \frac{T}{2}\right)A(x(s))\eta(s) = 0, \qquad x(s) \in \Gamma. \tag{6.3}$$

It is clear that the only solutions of (6.2), (6.3) have the form $v \equiv$ const. Thus we would have $v_t(x, T/2) \equiv 0$, $v_x(x, T/2) \equiv 0$ which would show that $\mathscr{E}(v(\cdot, T/2), v_t(\cdot, T/2)) = 0$. Since solutions of $L(v) = 0$ with boundary conditions (2.4) conserve energy, we could then conclude $\mathscr{E}(v(\cdot, T), v_t(\cdot, T)) = \mathscr{E}(\hat{v}, \hat{v}_t) = 0$, so that $\hat{v}_t \equiv 0$,

$\hat{v} \equiv$ const. and the proof would be complete. In fact, under these conditions we could obtain the result for $T = 2T_0$ also and there is nothing special about $n \leqq 3$.

Unfortunately we are not at all justified in assuming such smoothness for $v(x, t)$. A rigorous proof requires that we allow $(\hat{v}, \hat{v}_t)$ to be an arbitrary finite energy state. All this gives us is that $\hat{v}_t \in L^2(\Omega)$ and $\hat{v}_i \in L^2(\Omega)$, $i = 1, 2, \cdots, n$. The generalized solution $v(x, t)$ is no smoother. For this reason it becomes a nontrivial task to justify the argument presented above. In the present paper we will give such justification only for $n \leqq 3$, $T > 2T_0$. The result undoubtedly remains true for larger values of n and for $T = 2T_0$ but rather involved arguments seem to be required. Fortunately, $n \leqq 3$ includes most cases of physical interest.

THEOREM 4. *If* $(\hat{v}, \hat{v}_t) \in H_E$ *is such that* (6.1) *holds for all* $f \in F$ *then* $\hat{v}_t \equiv 0$, $\hat{v} \equiv$ *const. provided* $n \leqq 3$, $T > 2T_0$. *Thus the system* (2.1), (2.2) *is approximately controllable in time* $T > 2T_0$ *when* $n \leqq 3$.

Proof. Let $v(x, t)$ be the generalized solution of $L(v) = 0$ with homogeneous boundary conditions (2.4) and satisfying the terminal conditions $v(x, t) \equiv \hat{v}(x)$, $v_t(x, T) \equiv \hat{v}_t(x)$. Let $(\hat{v}^k, \hat{v}_t^k)$ be a sequence of states in H_E converging to $(\hat{v}, \hat{v}_t)$ in the energy norm as $k \to \infty$. Moreover, $(\hat{v}^k(x), \hat{v}_t^k(x)) \in C^\infty(\Omega \cup \Gamma)$ and satisfy the consistency conditions

$$\hat{v}_x^k(x(s))A(x(s))\eta(s) = 0, \quad \hat{v}_{tx}^k(x(s))A(x(s))\eta(s) = 0, \qquad x(s) \in \Gamma. \tag{6.4}$$

One way in which this could be done is to expand $\hat{v}(x)$, $\hat{v}_t(x)$ in terms of the eigenfunctions $\varphi_j(x)$ of the operator B introduced in (2.3):

$$\hat{v}(x) = \sum_{j=0}^{\infty} \alpha_j \varphi_j(x), \qquad \hat{v}_t(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x)$$

and take

$$\hat{v}^k(x) = \sum_{j=0}^{k} \alpha_j \varphi_j(x), \qquad \hat{v}_t^k(x) = \sum_{j=0}^{k} \beta_j \varphi_j(x).$$

For $k = 0, 1, 2, \cdots$ let $v^k(x, t)$ be the C^∞ solutions of $L(v^k) = 0$ which satisfy the terminal and boundary conditions

$$\begin{aligned} &v^k(x, T) \equiv \hat{v}^k(x), \qquad v_t^k(x, T) \equiv \hat{v}_t^k(x), \\ &v_x^k(x(s), t)A(x(s))\eta(s) \equiv 0, \qquad (x(s), t) \in \Gamma \times [0, T]. \end{aligned} \tag{6.5}$$

It is known [6, Chap. VI] that for each fixed t_1, $0 \leqq t_1 \leqq T$, the states $(v^k(\cdot, t_1), v_t^k(\cdot, t_1))$ converge to $(v(\cdot, t_1), v_t(\cdot, t_1))$ in the space H_E; in fact, this is just a consequence of the energy conservation.

Since

$$\lim_{k \to \infty} \|(\hat{v}, \hat{v}_t) - (\hat{v}^k, \hat{v}_t^k)\|_E = 0 \tag{6.6}$$

and (6.1) is assumed to hold, we have

$$\lim_{k \to \infty} \int_\Omega [\rho w_t^f \hat{v}^k + w_x^f A \hat{v}_x^{k\prime}] \, dx = 0.$$

Performing a computation similar to that in the proof of Theorem 1, we conclude that for each fixed $f \in F$,

$$\lim_{k \to \infty} \int_{\Gamma_1 \times [0,T]} v_t^k f \, ds \, dt = 0. \tag{6.7}$$

Let us now define functions $D^{-j}v$, $j = 1, 2, \cdots$, for $x \in \Omega$ by

$$(D^{-1}v)(x, t) = \int_T^t v(x, \tau) \, d\tau, \qquad (D^{-(j+1)}v)(x, t) = \int_T^t (D^{-j}v)(x, \tau) \, d\tau.$$

We define $D^{-j}v^k$ similarly and verify without difficulty, since the v^k satisfy the evolution equation (2.3), that

$$\begin{aligned} &\frac{d^2}{dt^2}(D^{-j}v^k) + B(D^{-j}v^k) = \hat{v}_t^k \frac{(t-T)^{j-1}}{(j-1)!} + \hat{v}^k \frac{(t-T)^{j-2}}{(j-2)!}, \\ &(D^{-j}v^k)(T) = (D^{-j}v^k)_t(T) = 0, \end{aligned} \tag{6.8}$$

for $j > 2$.

A result proved in [14, Theorem 1.19, p. 486] shows that the inhomogeneous linear initial value problem

$$\frac{d^2r}{dt^2} + Br = g(t), \qquad r(T) = r_t(T) = 0,$$

where $g : [0, T] \to L^2(\Omega)$ is continuously differentiable with respect to t, has a unique solution r such that $r_{tt}(t_1) \in L^2(\Omega)$ and $r(t_1)$ lies in the domain $\Delta(B) \subseteq L^2(\Omega)$ of the unbounded self-adjoint operator B for $0 \leqq t_1 \leqq T$. Moreover it is also shown in the theorem cited that there are constants M_0, M_1 such that

$$\|Br(t_1)\| < M_0 \sup_{0 \leqq \tau \leqq T} \|g(\tau)\| + M_1 \sup_{0 \leqq \tau \leqq T} \|g_t(\tau)\| \tag{6.9}$$

uniformly for all $t_1 \in [0, T]$, where $\|\cdot\|$ denotes the usual norm in $L^2(\Omega)$, and it is also shown that $Br(t)$ is continuous in t with respect to the norm $\|\cdot\|$.

Applying this theorem we see that, for $j \geqq 2$, $D^{-j}v$ lies in $\Delta(B)$ and solves

$$\begin{aligned} &\frac{d^2}{dt^2}(D^{-j}v) + B(D^{-j}v) = \hat{v}_t \frac{(t-T)^{j-1}}{(j-1)!} + \hat{v} \frac{(t-T)^{j-2}}{(j-2)!}, \\ &(D^{-j}v)(T) = (D^{-j}v)_t(T) = 0. \end{aligned} \tag{6.10}$$

Further, (6.9) together with (6.8), (6.10) shows that

$$\lim_{k \to \infty} \|B(D^{-j}v)(\cdot, t) - B(D^{-j}v^k)(\cdot, t)\| = 0 \tag{6.11}$$

uniformly for $0 < t < T$.

Now the operator B is uniformly elliptic and we can apply known results from the theory of elliptic boundary value problems [15, Theorem 9.11, p. 132 and Remarks, p. 148] to show that the fact that $D^{-j}v \in \Delta(B)$ together with (6.6) implies that $D^{-j}v(\cdot, t)$ lies in the space $H_2(\Omega)$ (for definition of $H_m(\Omega)$, see [15]) for $0 \leqq t \leqq T$, $\|(D^{-j}v)(\cdot, t)\|_{2,\Omega}$ is continuous and uniformly bounded for $0 \leqq t \leqq T$, and

$$\lim_{k \to \infty} \|(D^{-j}v)(\cdot, t) - (D^{-j}v^k)(\cdot, t)\|_{2,\Omega} = 0 \tag{6.12}$$

uniformly for $0 \leqq t \leqq T$, where $\|\cdot\|_{2,\Omega}$ is the norm in $H_2(\Omega)$, the sum of the integrals of the squares of the partial derivatives of order $\leqq 2$.

The theorem of Sobolev [15, Theorem 3.9, p. 32] states that if $r \in H_m(\Omega)$, then r can be modified on a set of measure zero so that $r \in C^l(\Omega \cup \Gamma)$, provided l is an integer such that $l < m - n/2$. For $m = 2$, we have $0 < 2 - n/2$ when $n \leqq 3$, so, for such n, $r \in H_2(\Omega)$ implies $r \in C^0(\Omega \cup \Gamma)$. Moreover, if $\|\cdot\|_s$ denotes the usual "sup" norm in $C^0(\Omega \cup \Gamma)$, we have

$$\|r\|_s \leqq \alpha_0 \|r\|_{2,\Omega} + \alpha_1 \|r\|.$$

Applying these results with $r = D^{-j}v$, $j \geqq 2$, the uniform boundedness and continuity of $\|(D^{-j}v)(\cdot, t)\|_{2,\Omega}$ and (6.11), we conclude that $(D^{-j}v)(x, t)$ is continuous for $(x, t) \in (\Omega \cup \Gamma) \times [0, T]$ and

$$\lim_{k \to \infty} (D^{-j}v^k)(x, t) = (D^{-j}v)(x, t) \tag{6.13}$$

uniformly for such (x, t).

Having now obtained the continuity of $(D^{-2}v)(x, t)$ we return to (6.7). Integrating by parts three times, we conclude that for all $f \in F$,

$$\lim_{k \to \infty} \int_{\Gamma_1 \times [0,T]} (D^{-2}v^k) f_{ttt} \, ds \, dt = 0$$

which with (6.13) implies

$$\int_{\Gamma_1 \times [0,T]} (D^{-2}v) f_{ttt} \, ds \, dt = 0, \qquad f \in F. \tag{6.14}$$

Taking account of the fact that f and all its derivatives vanish outside a compact subset of $\Gamma_1 \times (0, T)$, (6.14) implies that $(D^{-2}v)(x(s), t)$ is a polynomial in t of degree at most 2 whose coefficients are continuous functions of $x(s)$, for all $(x(s), t) \in \Gamma_1 \times [0, T]$. Let δ be a small positive number. We define the third order difference

$$\Delta^3 r(x, t) = r(x, t + 3\delta) - 3r(x, t + 2\delta) + 3r(x, t + \delta) - r(x, t)$$

for any function r defined on $\Omega \times [0, T]$. The function $\Delta^3 r$ is defined on $\Omega \times [0, T - 3\delta]$ and possesses all smoothness properties of $r(x, t)$. Applying this difference operator to $D^{-2}v$ and $D^{-2}v^k$ we obtain functions

$$\begin{aligned} \hat{u}(x, t) &= \Delta^3(D^{-2}v)(x, t), \\ u^k(x, t) &= \Delta^3(D^{-2}v^k)(x, t), \qquad k = 0, 1, 2, \cdots. \end{aligned} \tag{6.15}$$

From (6.8), (6.10) we see that

$$L(\hat{u}) = \frac{d^2\hat{u}}{dt^2} + B\hat{u} = 0, \qquad L(u^k) = \frac{d^2u^k}{dt^2} + Bu^k = 0,$$

the u^k satisfy the homogeneous boundary conditions

$$u_x^k(x(s), t)A(x(s))\eta(s) = 0, \qquad (x(s), t) \in \Gamma \times [0, T - 3\delta], \tag{6.16}$$

while $\hat{u}(x, t)$ is continuous, $\hat{u}(\cdot, t)$ lies in $\Delta(B)$ (which means $\hat{u}(x, t)$ satisfies the boundary conditions $\hat{u}_x(x(s), t)A(x(s))\eta(s) = 0$, $(x(s), t) \in \Gamma \times [0, T - 3\delta]$ in some

sense which we need not specify) and, from the fact that $D^{-2}v$ is a polynomial of degree at most 2 on $\Gamma_1 \times [0, T]$, we have

$$\hat{u}(x(s), t) \equiv 0, \qquad (x(s), t) \in \Gamma_1 \times [0, T - 3\delta]. \tag{6.17}$$

Now we refer back to Theorem 3, or, more precisely, its proof. We let $t_0 = 0$, $t_1 = T - 3\delta$ and define the surfaces $S(\lambda)$ as we did there. We define z as we did there and put u^k in place of the function u of Theorem 3. Repeating the calculations following (5.3), we see that

$$\int_{S(0)\cup S(\lambda)} (-u^k z_x A\eta + z u_x^k A\eta + \eta_0 \rho u^k z_t - \eta_0 \rho z u_t^k)\, d\varphi = 0.$$

On $S(0)$ we have $\eta_0(s) \equiv 0$ and $u_x^k A\eta \equiv 0$. Defining α and β as in the proof of Theorem 3 and recalling $z \equiv 0$ on $S(\lambda)$, we have

$$\int_{S(\lambda)} u^k \alpha\beta\, d\varphi = \int_{S(0)} u^k z_x A\eta\, ds\, dt.$$

Now u^k converges uniformly to $\hat{u}$ in $(\Omega \cup \Gamma) \times [0, T - 3\delta]$, so we have

$$\int_{S(\lambda)} \hat{u}\alpha\beta\, d\delta = \int_{S(0)} \hat{u} z_x A\eta\, ds\, dt = 0$$

since $\hat{u}$ obeys (6.17). As in Theorem 3 we conclude that $\hat{u} \equiv 0$ on $S(\lambda)$. A continuation process similar to that described in Theorem 3 can be used to show that $\hat{u} \equiv 0$ on every surface $S(\lambda)$, $0 \leqq \lambda \leqq 1$. Then, just as in Theorem 3, we conclude that

$$\hat{u}(x, t) \equiv 0, \qquad (x, t) \in K(\Gamma_1, 0, T - 3\delta).$$

Now if $T > 2T_0$, we have

$$\Omega \times \left[\frac{T}{2} - \varepsilon, \frac{T}{2} + \varepsilon\right] \subseteq K(\Gamma_1, 0, T - 3\delta)$$

if ε and δ are both chosen sufficiently small. (We need $\varepsilon + 3\delta < T/2 - T_0$.) Thus we have, for small ε and δ,

$$\hat{u}(x, t) \equiv 0, \qquad x \in \Omega, \qquad \frac{T}{2} - \varepsilon \leqq t \leqq \frac{T}{2} + \varepsilon. \tag{6.18}$$

Returning to the definition (6.15) we see that if (6.18) holds for all small δ, then it must be true that for $T/2 - \varepsilon \leqq t \leqq T/2 + \varepsilon$, $(D^{-2}v)(x, t)$ is a polynomial in t of degree not greater than 2 with coefficients which are functions of x lying in $\Delta(B)$ (since $D^{-2}v \in \Delta(B)$). Differentiating $D^{-2}v$ twice with respect to t in $T/2 - \varepsilon < t < T/2 + \varepsilon$, we see that there is a function $\tilde{v}(x)$ with $\tilde{v} \in \Delta(B)$ such that

$$v(x, t) \equiv \tilde{v}(x), \quad x \in \Omega, \quad \frac{T}{2} - \varepsilon \leqq t \leqq \frac{T}{2} + \varepsilon.$$

But if $v(x, t)$ is a generalized solution of $d^2v/dt^2 + Bv = 0$ such that $v \in \Delta(B)$ and $v(x, t)$ is constant with respect to t, then we must have

$$Bv(x, t) \equiv B\tilde{v}(x) \equiv 0.$$

But the only elements $\tilde{v} \in \Delta(B)$ for which $B\tilde{v} = 0$ are of the form $\tilde{v} \equiv$ const. Therefore, for $T/2 - \varepsilon \leqq t \leqq T/2 + \varepsilon$,

$$v(x, t) \equiv \tilde{v}(x) \equiv \text{const.}, \qquad v_t(x, t) \equiv 0.$$

Since the energy associated with the generalized solution $v(x, t)$ is constant, we conclude that $v(x, t) \equiv \hat{v}(x) \equiv$ const., $v_t(x, T) \equiv \hat{v}_t(x) \equiv 0$ and the proof of Theorem 4 is complete.

Acknowledgment. I should like to express my appreciation to Professor J. L. Lions of the University of Paris whose suggestions in a 1966 letter provided the germinal idea for the proofs presented in this paper.

REFERENCES

[1] D. L. RUSSELL, *On boundary-value controllability of linear symmetric hyperbolic systems*, Proc. Conference on Mathematical Theory of Control (Univ. of Southern California, Los Angeles, 1967), Academic Press, New York, 1967.

[2] ———, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Appl., 18 (1967), pp. 542–560.

[3] M. A. CIRINA, *Boundary controllability of nonlinear hyperbolic systems*, this Journal, 7 (1969), pp. 198–212.

[4] J. J. GRAINGER, *Boundary-value control of distributed systems characterized by hyperbolic differential equations*, Doctoral thesis, Electrical Engineering Dept., Univ. of Wisconsin, Madison, 1967.

[5] R. H. THOMAS, *The identification and control of hyperbolic distributed processes*, Doctoral thesis, Electrical Engineering Dept., Univ. of Wisconsin, Madison, 1969.

[6] R. COURANT AND D. HILBERT, *Partial Differential Equations*, Methods of Mathematical Physics, vol. II; Interscience, New York, 1962.

[7] H. O. FATTORINI, *Boundary control systems*, this Journal, 6 (1968), pp. 349–388.

[8] E. HOLMGREN, *Über Systeme von linearen partiellen Differentialgleichungen*, Ofversigt af kongl. Vetenskapsakad. Förk., 58 (1901), pp. 91–103.

[9] F. JOHN, *On linear partial differential equations with analytic coefficients—Unique continuation of data*, Comm. Pure Appl. Math., 2 (1949), pp. 209–253.

[10] K. O. FRIEDRICHS, *Symmetric hyperbolic linear differential equations*, Ibid., 7 (1954), pp. 345–392.

[11] P. D. LAX, *On Cauchy's problem for hyperbolic equations and the differentiability of solutions of elliptic equations*, Ibid., 8 (1955), pp. 615–633.

[12] G. F. D. DUFF, *Mixed problems for hyperbolic equations of general order*, Canad. J. Math., 2 (1959), pp. 195–221.

[13] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, Interscience, New York, 1968.

[14] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[15] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, Princeton, 1965.

SIAM J. CONTROL
Vol. 9, No. 1, February 1971

ON OPTIMAL CONTROLS FOR MEASURE DELAY-DIFFERENTIAL EQUATIONS*

P. C. DAS† AND R. R. SHARMA‡

1. Introduction. In recent years it has been a matter of interest to study the systems described by differential equations containing impulses which cause discontinuous changes in the values of the state variables of the system. Such systems include pulse frequency modulation systems and models for biological neural nets. (See [10], [11] and the references given therein.) Schmaedeke [13] has considered the control system described by the equations

$$\frac{dx^i}{dt} = f^i(t, x^1, x^2, \cdots, x^n, u^1, u^2, \cdots, u^m) + \sum_{j=1}^{m} g_j^i(t)\frac{du^j}{dt}, \qquad i = 1, 2, \cdots, n, \tag{1.1}$$

in which the control inputs $u^j(t)$ may have discontinuities of the first kind giving rise to impulse control input du^j/dt. He developed the theory of the equations (1.1) when the control functions $u^j(t)$ and the solutions $x^i(t)$ are functions of bounded variation and the derivatives dx^i/dt and du^j/dt are taken in the sense of distribution derivatives, to be denoted by Dx^i and Du^j respectively, which can be identified with Stieltjes measures. He then considered the existence of optimal control for the system given by (1.1). But in this optimal control problem the coefficients g_j^i of the impulsive controllers may be expected in many cases to depend not only on t but also on the state variables x^i. Moreover, the system may have hereditary effect. Due to these considerations an attempt has been made in this paper to generalize the results in [13] by considering the control system described by delay-differential equations

$$\begin{aligned} Dx^i(t) &= f^i(t, x_t^1, x_t^2, \cdots, x_t^n, u^1, u^2, \cdots, u^m) \\ &\quad + \sum_{j=1}^{m} g_j^i(t, x_t^1, x_t^1, \cdots, x_t^n, u^1, u^2, \cdots, u^m)Du^j(t), \qquad i = 1, 2, \cdots, n, \end{aligned} \tag{1.2}$$

where x_t^i represents the restriction of the function $x^i(s)$ on the interval $p(t) \leqq s \leqq q(t)$, p and q being real functions with the property $p(t) \leqq q(t) \leqq t$ for each t; for each fixed t, f^i and g_j^i are functionals defined on the space $BV([p(t), q(t)])$; and $u^j(t)$ are right continuous functions of bounded variation. If the g_j^i depend only on t, and $p(t) = q(t) = t$ so that $x_t^i = x^i(t)$, then (1.2) coincides with (1.1).

Throughout this paper, the (Stieltjes) integrals are taken to be Lebesgue (–Stieltjes) integrals. This requires in the proof of Theorem 1 a careful use of the integration by parts formula which does not always hold for Lebesgue–Stieltjes integrals. (In the corresponding theorem in [13] the integral used is actually Riemann–Stieltjes, though it has not been stated explicitly there.)

* Received by the editors May 14, 1969, and in final revised form March 18, 1970.

† Department of Mathematics, Indian Institute of Technology, Kanpur, India.
‡ Department of Mathematics, Regional Institute of Technology, Jamshedpur, India.

2. Preliminaries. Let Ω be a subset of n-dimensional Euclidean space E^n. We denote by $C_c^\infty(\Omega)$ the class of infinitely partially differentiable complex functions, defined on Ω, which have compact support. $C_c^\infty(\Omega)$ is a normed linear space with addition, scalar multiplication and norm defined by

$$(\psi_1 + \psi_2)(x) = \psi_1(x) + \psi_2(x),$$
$$(\alpha\psi)(x) = \alpha\psi(x),$$
$$\|\psi\| = \sup_{x\in\Omega} |\psi(x)|.$$

A continuous linear functional defined on $C_c^\infty(\Omega)$ is called a distribution on Ω. It follows from the Riesz representation theorem that any distribution F on Ω can be identified with a complex Borel measure μ by the relation

$$F(\psi) = \int_\Omega \psi \, d\mu, \qquad \psi \in C_c^\infty(\Omega). \tag{2.1}$$

If a distribution F is given by (2.1) and g is a μ-integrable function, then we define the product gF by

$$(gF)(\psi) = \int_\Omega g\psi \, d\mu, \qquad \psi \in C_c^\infty(\Omega). \tag{2.2}$$

It is easy to see that gF is a distribution on Ω. A distribution F on an interval I is to be identified with the Lebesgue–Stieltjes measure $dh(t)$ if for every closed finite interval J contained in I, $h(t)$ is of bounded variation on J and

$$F(\psi) = \int_J \psi(t)\, dh(t) \tag{2.3}$$

for all $\psi \in C_c^\infty(J)$. A distribution F on an interval I is to be identified with a point function f if for every closed finite interval J contained in I, f is integrable on J and

$$F(\psi) = \int_J f(t)\psi(t)\, dt \tag{2.4}$$

for all $\psi \in C_c^\infty(J)$. The derivative DF of a distribution F on I is a distribution defined by

$$DF(\psi) = -F(\psi'), \qquad ' \equiv d/dt. \tag{2.5}$$

For any vector $x = (x^1, x^2, \cdots, x^n) \in E^n$, the norm will be defined by

$$|x| = \sum_{i=1}^{n} |x^i|. \tag{2.6}$$

The norm of an $n \times m$ matrix $G = (g_j^i)$ will be defined by

$$|G| = \sum_{i=1}^{n} \sum_{j=1}^{m} |g_j^i|. \tag{2.7}$$

The space $BV(I)$ is defined for an interval I and consists of all scalar functions f on I which are of bounded variation. If a is the left endpoint of I, then the norm of f is

$$|f|_I = v(f, I) + |f(a+)|, \tag{2.8}$$

where $v(f, I)$ denotes the total variation of f on I. With this norm the space $BV(I)$ is a Banach space. The space $BV(I)_n$ is defined for an interval I and consists of all vector functions f with values in E^n whose individual components belong to $BV(I)$. The norm of f is

$$\|f\|_I = \sum_{i=1}^{n} |f^i|_I = \sum_{i=1}^{n} \{v(f^i, I) + |f^i(a+)|\}$$
$$= v(f, I) + |f(a+)|.$$

With this norm $BV(I)_n$ is a Banach space.

3. Existence and uniqueness of solutions. Let S be a domain (an open connected set) in E^n. The set of all functions in $BV(I)_n$ with values in S will be denoted by $BV(I, S)$. Let α, β and t_0 be numbers such that

$$-\infty \leqq \alpha < t_0 < \beta \leqq \infty.$$

In what follows, the interval $[\alpha, t_0]$ will be understood to be $(-\infty, t_0]$ in case $\alpha = -\infty$; similarly, if $\beta = \infty$, the interval $[t_0, \beta]$ will mean $[t_0, \infty)$. Let p and q be two real functions defined on $[t_0, \beta]$ and satisfying

$$\alpha \leqq p(t) \leqq q(t) \leqq t$$

for each t. We define the interval

$$I_t = [p(t), q(t)].$$

Let x_t denote the restriction of the function $x(\tau)$ on the interval I_t. For each $t \in [t_0, \beta]$, we define

$$R_t = \{(t, x_t) | x \in BV([\alpha, t], S)\} \tag{3.1}$$

and let R be defined by

$$R = \bigcup_{t \in [t_0, \beta]} R_t. \tag{3.2}$$

Let $f(t, x_t)$ be an n-vector functional and $G(t, x_t)$ an $n \times m$ matrix functional defined on R. Let $u(t)$ be a right continuous m-vector function of bounded variation defined on $[t_0, \beta]$. We assume that for each given $x \in BV([\alpha, \beta], S)$, $f(t, x_t)$ is Lebesgue measurable and $G(t, x_t)$ is integrable with respect to the Lebesgue–Stieltjes measure $du(t)$ on $[t_0, \beta]$

Consider now the delay-differential equation

$$Dx = f(t, x_t) + G(t, x_t)Du, \qquad t > t_0, \tag{3.3}$$

where the operations of differentiation are to be understood in the sense of distribution derivatives with respect to the real variable t. Since the distribution derivative Du of a function u of bounded variation can always be identified with

a Lebesgue–Stieltjes measure, we shall call (3.3) a measure delay-differential equation.

DEFINITION 1. A function $x(t)$ is called a *solution of* (3.3) *on an interval* I, $[\alpha, t_0] \subseteq I \subseteq [\alpha, \beta]$, with the initial function $\varphi \in BV([\alpha, t_0], S)$, if

(i) $x \in BV(I, S)$,

(ii) $x(t) = \varphi(t)$ for $t \in [\alpha, t_0]$,

(iii) $x(t)$ is continuous from the right on $I \cap [t_0, \beta]$,

(iv) $x(t)$ satisfies (3.3) on $I \cap (t_0, \beta]$.

Consider the integral equation

$$x(t) = \begin{cases} \varphi(t) & \text{for } t \in [\alpha, t_0], \\ \varphi(t_0) + \int_{t_0}^{t} f(s, x_s)\, ds + \int_{t_0}^{t} G(s, x_s)\, du(s) & \text{for } t > t_0. \end{cases} \tag{3.4}$$

DEFINITION 2. A function $x \in BV(I, S)$ is called a *solution of* (3.4) *on the interval* I, $[\alpha, t_0] \subseteq I \subseteq [\alpha, \beta]$, if it satisfies this equation for $t \in I$.

We shall now prove the following theorem.

THEOREM 1. *$x(t)$ is a solution of* (3.4) *if and only if it is a solution of* (3.3) *with initial function φ.*

Proof. Let $x(t)$ be a solution of (3.4) on the interval $[\alpha, T]$. Then conditions (i) and (ii) of Definition 1 are satisfied. The right continuity of $u(t)$ implies that the integral $\int_{t_0}^{t} G(s, x_s)\, du(s)$ is also a right continuous function of t. The integral $\int_{t_0}^{t} f(s, x_s)\, ds$ is an absolutely continuous (and hence continuous) function of t. Thus $x(t)$ is right continuous. We shall show that $x(t)$ satisfies (3.3) on $(t_0, T]$. Let $F^i(\psi)$ be the distribution on $[t_0, T]$ to be identified with the ith component $x^i(t)$ of $x(t)$. Then for any closed interval $J = [a, b]$ contained in $[t_0, T]$, we have

$$F^i(\psi) = \int_J \left\{ \varphi^i(t_0) + \int_{t_0}^{t} f^i(s, x_s)\, ds + \int_{t_0}^{t} [G(s, x_s)\, du(s)]^i \right\} \psi(t)\, dt$$

for all $\psi \in C_c^{\infty}(J)$. The derivative-distribution is

$$\begin{aligned} DF^i(\psi) = -F(\psi') = -\int_J \Bigg\{ & \varphi^i(t_0) + \int_{t_0}^{t} f^i(s, x_s)\, ds \\ & + \int_{t_0}^{t} \left(\sum_{j=1}^{m} g_j^i(s, x_s)\, du^j(s) \right) \Bigg\} \psi'(t)\, dt, \end{aligned} \tag{3.5}$$

where $g_j^i(s, x_s)$ is the i, jth element of $G(s, x_s)$ and $u^j(s)$ is the jth component of $u(s)$.

Integration by parts yields

$$\begin{aligned} & \int_J \left\{ \varphi^i(t_0) + \int_{t_0}^{t} f^i(s, x_s)\, ds \right\} \psi'(t)\, dt \\ & = \left[\left(\varphi^i(t_0) + \int_{t_0}^{t} f^i(s, x_s)\, ds \right) \psi(t) \right]_a^b - \int_J \psi(t) f^i(t, x_t)\, dt \\ & = -\int_J \psi(t) f^i(t, x_t)\, dt \qquad \text{(since } \psi(b) = \psi(a) = 0\text{)}. \end{aligned} \tag{3.6}$$

The function $h(t) = \int_{t_0}^{t} g_j^i(s, x_s)\, du^j(s)$ is right continuous, and is of bounded variation on the interval $J = [a, b]$, by [6, § 52.18, p. 275]. We have

$$\int_J h(t)\psi'(t)\, dt = \int_J h(t)\, d\psi(t) = \int_{\{a\}} h(t)\, d\psi(t) + \int_{(a,b]} h(t)\, d\psi(t).$$

But

$$\int_{\{a\}} h(t)\, d\psi(t) = h(a)(\psi(a) - \psi(a-)) \qquad \text{(by [7, Example 5, p. 199])}$$

$$= 0 \qquad \text{(since } \psi \text{ is continuous);}$$

and

$$\int_{(a,b]} h(t)\, d\psi(t) = h(b)\psi(b) - h(a)\psi(a) - \int_{(a,b]} \psi(t)\, dh(t) \quad \text{(by [7, Example n, p. 185])}$$

$$= -\int_{(a,b]} \psi(t)\, dh(t) \qquad \text{(since } \psi(a) = \psi(b) = 0)$$

$$= -\int_{[a,b]} \psi(t)\, dh(t) + \int_{\{a\}} \psi(t)\, dh(t)$$

$$= -\int_{[a,b]} \psi(t)\, dh(t) + \psi(a)(h(a) - h(a-))$$

$$= -\int_{[a,b]} \psi(t)\, dh(t) \qquad \text{(since } \psi(a) = 0).$$

Therefore,

$$\int_J h(t)\psi'(t)\, dt = -\int_J \psi(t)\, dh(t);$$

i.e.,

$$\int_J \left(\int_{t_0}^{t} g_j^i(s, x_s)\, du^j(s) \right) \psi'(t)\, dt$$

$$= \int_J \psi(t)\, d\left\{ \int_{t_0}^{t} g_j^i(s, x_s)\, du^j(s) \right\}$$

$$= \int_J \psi(t) g_j^i(t, x_t)\, du^j(t) \qquad \text{(by [1, Corollary 6, p. 180]);}$$

and, therefore,

$$\int_J \psi(t)\, d\left\{ \sum_{j=1}^{m} \int_{t_0}^{t} g_j^i(s, x_s)\, du^j(s) \right\} = \int_J \psi(t) \left(\sum_{j=1}^{m} g_j^i(t, x_t)\, du^j(t) \right). \tag{3.7}$$

From (3.5), (3.6) and (3.7), we obtain

$$DF'(\psi) = \int_J \psi(t) f^i(t, x_t)\, dt + \int_J \psi(t)[G(t, x_t)\, du(t)]^i.$$

Since $G(t, x_t)$ is integrable with respect to $du(t)$, the last continuous linear functional in the above equation is identified with the measure $[G(t, x_t)\, du(t)]^i$ (see (2.2)) while the first continuous linear functional in this equation is identified with the function $f^i(t, x_t)$. This holds for $i = 1, 2, \cdots, n$, and therefore the derivative distribution $DF(\psi)$ is identified with $f(t, x_t) + G(t, xt)Du$. Hence, $x(t)$ is also a solution of (3.3).

Conversely, let us suppose that $x(t)$ is a solution of (3.3) with initial function $\varphi(t)$. Then for any closed interval J contained in $[t_0, T]$ we have

$$\int_J \psi(t) Dx^i(t) = \int_J \psi(t) f^i(t, x_t)\, dt + \int_J \psi(t)[G(t, x_t)\, du(t)]^i \tag{3.8}$$

for $\psi \in C_c^\infty(J)$. By using [1, Corollary 6, p. 180] again we may write

$$\int_J \psi(t)[G(t, x_t)\, du(t)]^i = \int_J \psi(t)\, d\left(\int_{t_0}^t [G(s, x_s)\, du(s)]^i \right).$$

Integrating the three integrals in (3.8) in the way we have done above, we obtain

$$\int_J \psi'(t)(x^i(t) - \varphi^i(t_0))\, dt = \int_J \psi'(t) \left\{ \int_{t_0}^t f^i(s, x_s)\, ds + \int_{t_0}^t [G(s, x_s)\, du(s)]^i \right\} dt.$$

Therefore,

$$x^i(t) = \varphi^i(t_0) + \int_{t_0}^t f^i(s, x_s)\, ds + \int_{t_0}^t [G(s, x_s)\, du(s)]^i$$

almost everywhere in J. But, since $x^i(t)$ is continuous from the right, being a solution of (3.3), and since the right-hand side of the above equation is a right continuous function of t, equality holds everywhere in J in the above equation. Hence $x(t)$ is a solution of (3.4). The proof of Theorem 1 is thus complete.

We now enumerate the hypotheses which will be used in this section.

H_1. $f(t, x_t)$ is locally Lipschitzian with respect to x; i.e., for any $t_1 > t_0$ and every $k_1 < \infty$, there exists $k_2 = k_2(t_1, k_1)$ such that $\|x\|_{[\alpha, t_1]} \leqq k_1$, $\|y\|_{[\alpha, t_1]} \leqq k_1$ imply

$$|f(t, x_t) - f(t, y_t)| \leqq k_2 \|x - y\|_{I_t}$$

for every $t \in [t_0, t_1]$.

H_2. $f(t, x_t)$ is continuous in t and x; i.e., $\lim_{\nu \to \infty} t^{(\nu)} = t^*$ and $\lim_{\nu \to \infty} x^{(\nu)}(t) = x(t)$ imply

$$\lim_{\nu \to \infty} f(t^{(\nu)}, x_{t^{(\nu)}}^{(\nu)}) = f(t^*, x_{t^*}).$$

H_3. There exists a Lebesgue integrable function $r(t)$ such that

$$|f(t, x_t)| \leqq r(t)$$

uniformly with respect to x.

H_4. $G(t, x_t)$ is locally Lipschitzian with respect to x.

H_5. $G(t, x_t)$ is continuous in t and x.

H_6. There exists a function $w(t)$ integrable with respect to the Lebesgue–Stieltjes measure $d\, V_u(t)$ ($V_u(t) \equiv v(u, [t_0, t])$ denotes the total variation function

of $u(t)$) such that

$$|G(t, x_t)| \leqq w(t)$$

uniformly with respect to x.

THEOREM 2 (Local existence and uniqueness). *Let the hypotheses* H_1, H_3, H_4, H_6 *be satisfied on the interval* $[t_0, T]$. *Then there exists a unique solution of* (3.3) *on an interval* $[\alpha, t_0 + a]$ *with a given initial function* $\varphi \in BV([\alpha, t_0], S)$.

Proof. Denote by Q_t, $t_0 \leqq t \leqq T$, the space of all functions $x(t)$ with the properties

(i) $x \in BV([\alpha, t])_n$,
(ii) $x(s) = \varphi(s)$ for $s \in [\alpha, t_0]$,
(iii) $v(x, [t_0, t]) \leqq b$, $b > 0$.

Suppose that $Q_t \subseteq BV([\alpha, t], S)$. This is always possible if b is suitably chosen.

Choose a, $0 < a \leqq T - t_0$, such that

$$\int_{t_0}^{t_0+a} r(t)\,dt + \int_{t_0}^{t_0+a} w(t)\,dV_u(t) \leqq b. \tag{3.9}$$

Since $\int_{t_0}^{t} r(s)\,ds$ is a continuous function of t, and since the right continuity of $u(t)$ and hence of $dV_u(t)$ imply that $\int_{t_0}^{t} w(s)\,dV_u(s)$ is also a right continuous function t, it is possible to choose such an a.

Now consider Q_{t_0+a} which forms a complete metric space. Let A be the mapping defined on Q_{t_0+a} through the relations

$$(Ax)(t) = \begin{cases} \varphi(t) & \text{for } t \in [\alpha, t_0], \\ \varphi(t_0) + \int_{t_0}^{t} f(s, x_s)\,ds + \int_{t_0}^{t} G(s, x_s)\,du(s) & \text{for } t \in (t_0, t_0 + a]. \end{cases} \tag{3.10}$$

We shall show that A maps Q_{t_0+a} into itself. Since both of the integrals on the right of (3.10) are functions of t which are of bounded variation on $[t_0, t_0 + a]$, it follows that Ax is a function in $BV([\alpha, t_0 + a])_n$. Furthermore,

$$\begin{aligned} v(Ax, [t_0, t_0 + a]) &\leqq v\left(\int_{t_0}^{t} f(s, x_s)\,ds, [t_0, t_0 + a] \right) \\ &\quad + v\left(\int_{t_0}^{t} G(s, x_s)\,du(s), [t_0, t_0 + a] \right) \\ &\leqq \int_{t_0}^{t_0+a} |f(s, x_s)|\,ds + \int_{t_0}^{t_0+a} |G(s, x_s)|\,dV_u(s) \qquad (3.11) \\ &\leqq \int_{t_0}^{t_0+a} r(s)\,ds + \int_{t_0}^{t_0+a} w(s)\,dV_u(s) \qquad \text{(by } H_3 \text{ and } H_6\text{)} \\ &\leqq b \qquad \text{(by (3.9))}. \end{aligned}$$

Thus a maps Q_{t_0+a} into itself.

We shall now show that A is a contraction. We have

$$\begin{aligned}\|Ax - Ay\|_{[\alpha,t_0+a]} &\leqq \left\| \int_{t_0}^{t} [f(s, x_s) - f(s, y_s)]\, ds \right\|_{[t_0,t_0+a]} \\ &\quad + \left\| \int_{t_0}^{t} [G(s, x_s) - G(s, y_s)]\, du(s) \right\|_{[t_0,t_0+a]} \\ &\leqq \int_{t_0}^{t_0+a} |f(s, x_s) - f(s, y_s)|\, ds \\ &\quad + \int_{t_0}^{t_0+a} |G(s, x_s) - G(s, y_s)|\, dV_u(s).\end{aligned} \tag{3.12}$$

We have, for every $x \in Q_{t_0+a}$,

$$\begin{aligned}\|x\|_{[\alpha,t_0+a]} &= \|\varphi\|_{[\alpha,t_0]} + v(x, [t_0, t_0 + a]) \\ &\leqq \|\varphi\|_{[\alpha,t_0]} + b = k_1, \quad \text{say.}\end{aligned}$$

Therefore, by H_1 and H_4, there exist constants $k_2 = k_2(k_1)$, $k_3 = k_3(k_1)$ such that

$$\begin{aligned}|f(s, x_s) - f(s, y_s)| &\leqq k_2 \|x - y\|_{I_s} \leqq k_2 \|x - y\|_{[\alpha,s]}, \\ |G(s, x_s) - G(s, y_s)| &\leqq k_3 \|x - y\|_{I_s} \leqq k_3 \|x - y\|_{[\alpha,s]}.\end{aligned} \tag{3.13}$$

From (3.12) and (3.13), we obtain

$$\begin{aligned}\|Ax - Ay\|_{[\alpha,t_0+a]} &\leqq \int_{t_0}^{t_0+a} k_2 \|x - y\|_{[\alpha,s]}\, ds + \int_{t_0}^{t_0+a} k_3 \|x - y\|_{[\alpha,s]}\, dV_u(s) \\ &\leqq \{ak_2 + k_3 v(u, [t_0, t_0 + a])\} \|x - y\|_{[\alpha,t_0+a]}.\end{aligned} \tag{3.14}$$

Since $u(t)$ is right continuous, a can be chosen such that

$$ak_2 + k_3 v(u, [t_0, t_0 + a]) < 1,$$

and then A is, by (3.14), a contraction. Hence, by the principle of contraction mapping there is a unique fixed point. This completes the proof.

Our next theorem shows that the local existence of a solution can also be proved if the hypotheses H_1 and H_4 in the above theorem are replaced by the hypotheses H_2 and H_5.

THEOREM 3 (Local existence). *Let the hypotheses* H_2, H_3, H_5 *and* H_6 *be satisfied. Then there exists a solution of* (3.3) *on an interval* $[\alpha, t_0 + a]$ *with a given initial function* $\varphi \in BV([\alpha, t_0], S)$.

Proof. Define Q_t as in the proof of Theorem 2 and choose $a > 0$ so as to satisfy (3.9).

Now define

$$\bar{\varphi}(t) = \begin{cases} \varphi(t) & \text{for } t \in [\alpha, t_0], \\ \varphi(t_0) & \text{for } t \in (t_0, t_0 + a] \end{cases} \tag{3.15}$$

and

$$(3.16)\quad x^{(\nu)}(t) = \begin{cases} \bar{\varphi}(t) & \text{for } t \in [\alpha, t_0 + a/\nu], \\ \varphi(t_0) + \displaystyle\int_{t_0 + a/\nu}^{t} f\left(s - \frac{a}{\nu}, x^{(\nu)}_{s - a/\nu}\right) ds & \\ \quad + \displaystyle\int_{t_0 + a/\nu}^{t} G\left(s - \frac{a}{\nu}, x^{(\nu)}_{s - a/\nu}\right) du(s) & \text{for } t \in (t_0 + a/\nu, \cdot] \end{cases}$$

$$\nu = 1, 2, \cdots.$$

For any $\nu \geqq 2$, the first expression in (3.16) defines $x^{(\nu)}(t)$ on $[\alpha, t_0 + a/\nu]$, and then the second expression of (3.16) defines $x^{(\nu)}(t)$ on $(t_0 + a/\nu, t_0 + 2a/\nu]$. Let us assume that $x^{(\nu)}(t)$ is defined on $[\alpha, t_0 + ja/\nu]$ for $1 < j < \nu$. Then the second expression of (3.16) defines $x^{(\nu)}(t)$ on $(t_0 + ja/\nu, t_0 + (j+1)\alpha/\nu]$. $x^{(\nu)}(t)$ is thus defined on $[\alpha, t_0 + a]$.

From H_3, H_6 and (3.9), we shall obtain, by calculations similar to (3.11),

$$(3.17)\qquad v(x^{(\nu)}, [t_0, t_0 + a]) \leqq b, \qquad \nu = 1, 2, \cdots.$$

Thus,

$$x^{(\nu)} \in Q_{t_0 + a}, \qquad \nu = 1, 2, \cdots,$$

and hence $x^{(\nu)}(t)$ are of uniformly bounded variation and are also uniformly bounded. Hence by Helly's selection principle [2, Chap. XII, Theorem 33], there exists a subsequence $x^{(\nu_j)}(t)$ and a function $x^*(t)$ of bounded variation such that

$$(3.18)\qquad \lim_{j \to \infty} x^{(\nu_j)}(t) = x^*(t),$$

and moreover,

$$v(x^*, [\alpha, t_0 + a]) \leqq \liminf_{j \to \infty} v(x^{(\nu_j)}, [\alpha, t_0 + a]) \leqq b$$

by (3.17). Obviously $x^*(t) = \varphi(t)$ for $t \in [\alpha, t]$. Hence $x^* \in Q_{t_0 + a} \subset BV([\alpha, t_0], s)$.

By H_2, H_5 and (3.18) we obtain

$$(3.19)\qquad \lim_{j \to \infty} f(t - a/\nu_j, x^{(\nu_j)}_{t - a/\nu_j}) = f(t, x^*_t),$$

$$(3.20)\qquad \lim_{j \to \infty} G(t - a/\nu_j, x^{(\nu_j)}_{t - a/\nu_j}) = G(t, x^*_t).$$

From H_3 and (3.19), we obtain, by using Lebesgue's dominated convergence theorem,

$$(3.21)\qquad \lim_{j \to \infty} \int_{t_0}^{t} f\left(s - \frac{a}{\nu_0}, x^{(\nu_j)}_{s - a/\nu_j}\right) ds = \int_{t_0}^{t} f(s, x^*_s)\, ds.$$

Let $du^+(t)$ and $du^-(t)$ be upper and lower variations (also called positive and negative variations) of the Lebesgue–Stieltjes measure $du(t)$. Then

$$\left| \int_{t_0}^{t} G(s, x_s)\, du^+(s) \right| \leqq \int_{t_0}^{t} |G(s, x_s)|\, du^+(s)$$

$$\leqq \int_{t_0}^{t} w(s)\, du^+(s) \quad (\text{by } H_6)$$

$$\leqq \int_{t_0}^{t} w(s)\, dV_u(s),$$

since $w(s) \geqq 0$, $dV_u(s) = du^+(s) + du^-(s)$. Since $w(s)$ is integrable with respect to $dV_u(s)$, $G(t, x_t)$ is integrable with respect to $du^+(s)$. Similarly $G(t, x_t)$ is also integrable with respect to $du^-(s)$. Therefore,

$$\int_{t_0}^{t} G\left(s - \frac{a}{v_j}, x^{(v_j)}_{s-a/v_j}\right) du(s) = \int_{t_0}^{t} G\left(s - \frac{a}{v_j}, x^{(v_j)}_{s-a/v_j}\right) du^+(s) - \int_{t_0}^{t} G\left(s - \frac{a}{v_j}, x_{s-a/v_j}\right) du^-(s) \tag{3.22}$$

(see [4, Example (7), § 29]; also see [7, p. 184]). Now take limits as $v_j \to \infty$. Due to H_6 and (3.20), Lebesgue's dominated convergence theorem can be used on the integrals on the right-hand side of (3.22) to obtain

$$\lim_{j\to\infty} \int_{t_0}^{t} G\left(s - \frac{a}{v_j}, x^{(v_j)}_{s-a/v_j}\right) du(s) = \int_{t_0}^{t} G(s, x_s)\, du^+(s) - \int_{t_0}^{t} G(s, x_s)\, du^-(s) = \int_{t_0}^{t} G(s, x_s)\, du(s). \tag{3.23}$$

Also

$$\left| \int_{t_0}^{t_0 + a/v_j} G\left(s - \frac{a}{v_j}, x^{(v_j)}_{s-a/v_j}\right) du(s) \right| \leqq \int_{t_0}^{t_0 + a/v_j} w(s)\, dV_u(s) \qquad \text{(by } H_6\text{)}$$

$$\to 0 \quad \text{as } j \to \infty,$$

since $\int_{t_0}^{t} w(s)\, dV_u(s)$ is a right continuous function of t. So

$$\lim_{j\to\infty} \int_{t_0}^{t_0 + a/v_j} G\left(s - \frac{a}{v_j}, x^{(v_j)}_{s-a/v_j}\right) du(s) = 0. \tag{3.24}$$

Similarly,

$$\lim_{j\to\infty} \int_{t_0}^{t_0 + a/v_j} f\left(s - \frac{a}{v_j}, x^{(v_j)}_{s-a/v_j}\right) ds = 0. \tag{3.25}$$

Now, by (3.16), we have

$$x^{(v_j)}(t) = \begin{cases} \bar{\varphi}(t) \quad \text{for } t \in [\alpha, t_0 + a/v], \\ \varphi(t_0) + \displaystyle\int_{t_0}^{t} f\left(s - \frac{a}{v_j}, x^{(v_j)}_{s-a/v_j}\right) ds - \int_{t_0}^{t_0 + a/v_0} f\left(s - \frac{a}{v_j}, x^{(v_j)}_{s-a/v_j}\right) ds \\ \qquad + \displaystyle\int_{t_0}^{t} G\left(s - \frac{a}{v_j}, x^{(v_j)}_{s-a/v_j}\right) du(s) - \int_{t_0}^{t_0 + a/v_j} G\left(s - \frac{a}{v_j}, x^{(v_j)}_{s-a/v_j}\right) du(s) \\ \qquad\qquad \text{for } t \in (t_0 + a/v, t_0 + a]. \end{cases}$$

Taking limits as $j \to \infty$ and making use of (3.18), (3.19), (3.20), (3.24) and (3.25), we obtain

$$x^*(t) = \begin{cases} \varphi(t) & \text{for } t \in [\alpha, t_0], \\ \varphi(t_0) + \int_{t_0}^{t} f(s, x_s^*)\, ds + \int_{t_0}^{t} G(s, x_s^*)\, du(s) & \text{for } t \in (t_0, t_0 + a]. \end{cases}$$

$x^*(t)$ is thus the solution of (3.3) on the interval $[\alpha, t_0 + a]$ with initial function $\varphi(t)$. This completes the proof.

THEOREM 4 (Extended existence). *Let the hypotheses* H_2, H_3, H_5 *and* H_6 *be satisfied on* $[t_0, \beta]$, *and let* $\varphi \in BV([\alpha, t_0], S)$ *be given. If there exists a solution* $x(t, t_0, \varphi)$ *of* (3.3) *on* $[\alpha, T)$, *where* T *is a point of continuity of* $u(t)$, *and if* $T < \beta$ *and* T *cannot be increased, then for any compact set* $F \subset S$ *there exists a sequence* $t_0 < t_1 < t_2 \cdots < t_k < \cdots \to T$ *such that*

$$x(t_k) \in S - F \quad \text{for } k = 1, 2, \cdots.$$

Proof. Let $x(t, t_0, \varphi)$ be a solution of (3.3) on $[\alpha, T)$ with initial function φ, where $T \in (t_0, \beta)$ is a point of continuity of $u(t)$. Suppose that there exists a compact set $F \subset S$ such that $x(t) \in F$ for $t \in [\alpha, T]$. We have

$$x(t) = \begin{cases} \varphi(t) & \text{for } t \in [\alpha, t_0], \\ \varphi(t_0) + \int_{t_0}^{t} f(s, x_s)\, ds + \int_{t_0}^{t} G(s, x_s)\, du(s) & \text{for } t \in (t_0, T). \end{cases}$$

If $t_0 \leqq t_1 < t_2 < T$, then

$$\begin{aligned} |x(t_1) - x(t_2)| &\leqq \int_{t_1}^{t_2} |f(s, x_s)|\, ds + \int_{t_1}^{t_2} |G(s, x_s)|\, dV_u(s) \\ &\leqq \int_{t_1}^{t_2} r(s)\, ds + \int_{t_1}^{t_2} w(s)\, dV_u(s). \end{aligned} \tag{3.26}$$

Since $u(t)$ is continuous at the point T, the function $\int_{t_0}^{t} w(s)\, dV_u(s)$ is also continuous at T. So if $\varepsilon > 0$, we can choose t_1 so close to T (but $t_1 < T$) that

$$\int_{t_1}^{T} w(s)\, dV_u(s) < \frac{\varepsilon}{2},$$

and then we choose t_2 so close to t_1 that

$$\int_{t_1}^{t_2} r(s)\, ds < \frac{\varepsilon}{2}.$$

Thus, we obtain, from (3.26),

$$|x(t_1) - x(t_2)| < \varepsilon$$

for all $t_0 \leqq t_1 < t_2 < T$ such that t_1 is sufficiently close to T. Therefore, by Cauchy's criterion,

$$x(T-) \equiv \lim_{t \to T-} x(t)$$

exists. Define

$$x(T) = x(T-).$$

Let

$$\varphi^*(t) = x(t) \quad \text{for } t \in [\alpha, T].$$

Obviously, $\varphi^*(t) \in F \subset S$. Replacing t_0 by T, φ by φ^*, we can prove, as in Theorem 3, that there exists a solution $x^*(t, T, \varphi^*)$ of (3.3) on $[\alpha, T + \delta]$, where $\delta > 0$, which is represented by

$$x^*(t) = \begin{cases} \varphi^*(t) & \text{for } t \in [\alpha, T], \\ \varphi^*(T) + \int_T^t f(s, x_s)\, ds + \int_T^t G(s, x_s)\, du(s) & \text{for } t \in (T, T + \delta). \end{cases}$$

Obviously $x^*(t)$ is also a solution of (3.3) with the initial condition $x^*(t) = \varphi(t)$ for $t \in [\alpha, t_0]$; i.e., $x^*(t)$ is the extension of $x(t)$ over the interval $[\alpha, T + \delta]$. Thus, the value of T can be increased. This proves the theorem.

4. Formulation of an optimal control problem. Let $[t_0, \beta]$ be a fixed interval and S a domain of E^n. Let $B = \bigcup_{t \in [t_0, \beta]} BV([\alpha, t], S)$. We shall denote by U the set of all right continuous functions u of bounded variation on $[t_0, \beta]$ into a non-empty compact subset Q of E^m.

Consider a control process governed by the measure delay-differential equation

$$Dx = f(t, x_t, u(t)) + G(t, x_t, u(t))\, Du, \qquad t > t_0, \tag{4.1}$$

satisfying the following assumptions:

A_1. The functional f with range in E^n is defined for all $t \in [t_0, \beta]$, for all $x \in B$ and for all $u \in U$.

A_2. $f(t, x_t, u(t))$ is continuous in t, x and u.

A_3. There exists a Lebesgue integrable real function $r(t)$ for $t \in [t_0, \beta]$ such that

$$|f(t, x_t, u(t))| \leqq r(t)$$

uniformly with respect to $x \in B$, $u \in U$.

A_4. G is an $n \times m$ matrix defined for $t \in [t_0, \beta]$, for all $x \in B$ and for all $u \in U$.

A_5. $G(t, x_t, u(t))$ is continuous in t, x and u.

A_6. There exists a constant K such that

$$|G(t, x_t, u(t))| \leqq K$$

uniformly with respect to $x \in B$, $u \in U$.

Let $\varphi \in BV([\alpha, t_0], S)$ be given. Under the assumptions A_1 to A_6, (4.1) has a bounded variation solution $x(t)$, for each choice of $u \in U$, satisfying the initial condition $x(t) = \varphi(t)$ for $t \in [\alpha, t_0]$. This follows from Theorem 3. If $u \in U$ defined on the interval $[t_0, t_1]$ is such that the corresponding solution $x(t, t_0, \varphi)$ of (4.1) is also defined on $[t_0, t_1]$, then $u(t)$ is called an admissible control and $x(t)$ is called

the corresponding trajectory. We define the cost of the control $u(t)$ by

$$J(u) = \int_{t_0}^{t_1} f^0(t, x(t), u(t))\, dt,$$

where f^0 is a given continuous real function defined on $[t_0, \beta] \times S \times Q$. The functional J is called the cost functional of the system (4.1).

A target set is a family $\mathscr{T}$ of nonempty compact sets $T_t \subset E^n$ defined for $t \in [t_0, \beta]$. We say that an admissible control $u(t)$ defined on $[t_0, t_1]$ transfers the function $\varphi \in BV([\alpha, t_0], S)$ to the target set $\mathscr{T}$, if the trajectory $x(t) = x(t, t_0, \varphi, u)$, corresponding to the initial function $\varphi(t)$ and the control $u(t)$, satisfies the relation $x(t_1) \in T_{t_1}$. Let A be the set of all admissible controls which transfer φ to $\mathscr{T}$. A control $u^* \in A$ is called an optimal (minimal) control if

$$J(u^*) \leqq J(u) \quad \text{for each } u \in A.$$

The optimization problem consists in finding such an optimal control.

5. Existence of an optimal control.

THEOREM 5. *We are given the control problem with the following data*:

(a) $$Dx = f(t, x_t, u(t)) + G(t, x_t, u(t))Du, \qquad t \in [t_0, \beta],$$

with f and G satisfying the assumptions A_1 *to* A_6 *of* § 4;

(b) *the fixed initial function $\varphi \in BV([\alpha, t_0], S)$*;

(c) *a target set $\mathscr{T}$ of nonempty compact sets $T_t \subset E^n$ defined on $[t_0, \beta]$ and upper semicontinuous with respect to inclusion*;

(d) *the set A of admissible controls $u(t)$ defined on subintervals $[t_0, t_1]$ contained in $[t_0, \beta]$ with the same left endpoint (and perhaps different right endpoint $t_1 > t_0$) which transfer φ to $\mathscr{T}$, which is such that for all $u \in A$,*

$$|\Delta u| \leqq \Delta h$$

on each subinterval of $[t_0, t_1]$, where h is a given nondecreasing right continuous function defined on $[t_0, t_1]$; (the symbol Δh on the interval, say, $[t_0, t_1]$ denotes $h(t_1) - h(t_0)$);

(e) *the cost functional*

$$J(u) = \int_{t_0}^{t_1} f^0(t, x(t), u(t))\, dt,$$

where f^0 is a continuous real function defined on $[t_0, \beta] \times S \times Q$.

Then if A is nonempty, there exists an optimal control in A.

Proof. Since for each $u \in A$, $|\Delta u| \leqq \Delta h$ on each subinterval of $[t_0, t_1]$, therefore $u \in A$ are of uniform bounded variation. We shall show that the corresponding trajectories $x(t)$ are also of uniform bounded variation.

We have

$$(5.1) \quad x(t) = \begin{cases} \varphi(t) & \text{for } t \in [\alpha, t_0], \\ \varphi(t_0) + \int_{t_0}^{t} f(s, x_s, u(s))\, ds + \int_{t_0}^{t} G(s, x_s, u(s))\, du(s) & \text{for } t \in (t_0, t_1]. \end{cases}$$

As in (3.11),

$$
\begin{aligned}
v(x, [t_0, t_1]) &\leqq \int_{t_0}^{t_1} |f(s, x_s)|\, ds + \int_{t_0}^{t_1} |G(s, x_s)|\, dV_u(s) \\
&\leqq \int_{t_0}^{t_1} r(s)\, ds + Kv(u, [t_0, t_1]) \qquad (5.2) \\
&\leqq \int_{t_0}^{t_1} r(s)\, ds + K(h(\beta) - h(t_0)),
\end{aligned}
$$

which shows that $v(x, [t_0, t_1])$ are uniformly bounded for all the trajectories $x(t)$ corresponding to controls in A. Hence the trajectories $x(t)$ are also uniformly bounded.

Now, since A is nonempty, and the corresponding trajectories are uniformly bounded, $\inf J(u) = \tilde{m} > -\infty$ for all $u \in A$. Either A is a finite set, in which case the theorem is trivially true, or we can select from A a sequence of controls $u^{(k)}$ on the intervals $[t_0, t_1^{(k)}]$ for which $J(u^{(k)})$ decreases monotonically to $\tilde{m}$. Select a subsequence (which we shall still denote by $u^{(k)}$) such that $t_1^{(k)} \to t_1^*$ monotonically. Consider the case when $\{t_1^{(k)}\}$ is monotonic decreasing (the case when it is increasing will be considered later). Next choose $\hat{t}_1$ such that $\hat{t}_1 = t_1$ if $t_1^{(k)} = t_1$ for all k; otherwise let $t_1^* < t_1^{(k_0+1)} \leqq \hat{t}_1 < t_1^{(k_0)} < \beta$ for some k_0. (From now on all reference to the index k tacitly assumes $k > k_0$.) We define the extended control $\hat{u}^{(k)}$ on $[t_0, \hat{t}_1]$ by

$$
\hat{u}^{(k)}(t) = \begin{cases} u^{(k)}(t) & \text{if } t \in [t_0, t_1^{(k)}], \\ u^{(k)}(t_1^{(k)}) & \text{if } t \in (t_1^{(k)}, \hat{t}_1]. \end{cases} \qquad (5.3)
$$

Since $u^{(k)}(t) \in Q$ for all $t \in [t_0, t_1^{(k)}]$, $\hat{u}^{(k)}(t)$ also belongs to Q for $t \in [t_0, \hat{t}_1]$. Evidently $\hat{u}^{(k)}$ is right continuous on $[t_0, \hat{t}_1)$, and $|\Delta\hat{u}^{(k)}| \leqq \Delta h$ on every subinterval of $[t_0, \hat{t}_1]$. Now, by Helly's principle of choice, there exists a subsequence (for which we shall use the same notation $\hat{u}^{(k)}$) and a function of bounded variation u^* such that

$$
\lim_{k \to \infty} \hat{u}^{(k)}(t) = u^*(t) \qquad (5.4)
$$

everywhere on $[t_0, \hat{t}_1]$. Our aim is to show that $u^* \in A$ and is an optimal control.

We shall first show that u^* is right continuous. Let τ be any point in $[t_0, \hat{t}_1)$. Since $|\Delta\hat{u}^{(k)}| \leqq \Delta h$ on every subinterval of $[t_0, \hat{t}_1]$, we have

$$
|\hat{u}^{(k)}(\tau + s) - \hat{u}^{(k)}(\tau)| \leqq h(\tau + s) - h(\tau), \qquad \tau \leqq \tau + s \leqq \hat{t}_1;
$$

and since h is right continuous at τ, it follows that given any $\varepsilon > 0$, there exists a $\delta > 0$ such that for any k,

$$
|\hat{u}^{(k)}(\tau + s) - \hat{u}^{(k)}(\tau)| < \varepsilon \quad \text{for } 0 \leqq s \leqq \delta.
$$

Therefore,

$$
\lim_{s \to 0+} \hat{u}^{(k)}(\tau + s) = \hat{u}^{(k)}(\tau) \quad \text{uniformly for } k. \qquad (5.5)
$$

Also

$$\lim_{k\to\infty} \hat{u}^{(k)}(\tau + s) = u^*(\tau + s).$$

Hence, applying the Moore theorem on interchange of order of repeated limits, we obtain

$$\lim_{s\to 0+} \lim_{k\to\infty} \hat{u}^{(k)}(\tau + s) = \lim_{k\to\infty} \lim_{s\to 0+} \hat{u}^{(k)}(\tau + s);$$

i.e.,

$$u^*(\tau + 0) = u^*(\tau),$$

Thus, u^* is right continuous at τ. But τ was taken arbitrarily in $[t_0, \hat{t}_1)$. Hence u^* is right continuous at each $t \in [t_0, \hat{t}_1)$.

Since $\hat{u}^{(k)}(t) \in Q$ for $t \in [t_0, \hat{t}_1]$ and Q is compact, it follows from (5.4) that $u^*(t)$ also belongs to Q for each $t \in [t_0, \hat{t}_1]$. Since $\Delta u^* = \lim_{k\to\infty} \Delta\hat{u}^{(k)}$ on every subinterval of $[t_0, \hat{t}_1]$, it follows that, given any $\varepsilon > 0$,

$$\begin{aligned} \Delta u^* &\leqq |\Delta\hat{u}^{(k)}| + \varepsilon \\ &\leqq \Delta h + \varepsilon \quad \text{for sufficiently large } k. \end{aligned}$$

Since ε is arbitrary, we get

$$\Delta u^* \leqq \Delta h \tag{5.6}$$

on every subinterval of $[t_0, t_1]$.

Let $x^{(k)}$ be the trajectory defined on $[t_0, t_1^{(k)}]$ corresponding to the control $u^{(k)}$. The extended control $\hat{u}^{(k)}$ coincides with $u^{(k)}$ on $[t_0, t_1^{(k)}]$ and is constant and, therefore, continuous on $[t_1^{(k)}, \hat{t}_1]$ (see (5.3)). Hence, by Theorem 4, $x^{(k)}$ can be extended to $[t_0, \hat{t}_1]$. The extended trajectory $\hat{x}^{(k)}$ is given by

$$\hat{x}^{(k)}(t) = \begin{cases} \varphi(t) & \text{for } t \in [\alpha, t_0], \\ \varphi(t_0) + \displaystyle\int_{t_0}^{t} f(s, \hat{x}_s^{(k)}, \hat{u}^{(k)}(s))\, ds & \\ \qquad + \displaystyle\int_{t_0}^{t} G(s, \hat{x}_s^{(k)}, \hat{u}^{(k)}(s))\, d\hat{u}^{(k)}(s) & \text{for } t \in (t_0, \hat{t}_1]. \end{cases} \tag{5.7}$$

Since

$$v(\hat{u}^{(k)}, [t_0, \hat{t}_1]) = v(u^{(k)}, [t_0, t_1^{(k)}]),$$

total variations of $\hat{x}^{(k)}$ can be seen, as in (5.2), to be uniformly bounded. The $\hat{x}^{(k)}$ are also uniformly bounded. Hence there exists a subsequence (still labeled $\hat{x}^{(k)}$) and a function x^* such that

$$\lim_{k\to\infty} \hat{x}^{(k)}(t) = x^*(t) \tag{5.8}$$

everywhere on $[t_0, \hat{t}_1]$. By selecting the corresponding subsequence from $\hat{u}^{(k)}$ we do not change any of the preceding limiting operations satisfied by $\hat{u}^{(k)}$.

Since $\hat{u}^{(k)}(t) \to u^*(t)$, $\hat{x}^{(k)}(t) \to x^*(t)$ everywhere on $[t_0, \hat{t}_1]$, and the assumptions A_2 and A_3 of § 4 hold, we obtain, by Lebesgue's dominated convergence theorem,

$$\lim_{k\to\infty} \int_{t_0}^{t} f(s, \hat{x}_s^{(k)}, \hat{u}^{(k)}(s))\, ds = \int_{t_0}^{t} f(s, x_s^*, u^*(s))\, ds \tag{5.9}$$

for all $t \in [t_0, \hat{t}_1]$

Now, $|\Delta\hat{u}^{(k)}| \leqq \Delta h, |\Delta u^*| \leqq \Delta h$ on every subinterval of $[t_0, \hat{t}_1]$; and $\hat{u}^{(k)}(t) \to u^*(t)$ everywhere on $[t_0, \hat{t}_1]$. For each k, $G(t, \hat{x}_t^{(k)}, \hat{u}^{(k)}(t))$ is a continuous function of t (by A_5 of § 4), and therefore each element $g_j^i(t, \hat{x}_t^{(k)}, \hat{u}^{(k)}(t))$ of the matrix function $G(t, \hat{x}_t^{(k)}, \hat{u}^{(k)}(t))$ is integrable with respect to $dh(t)$. Moreover,

$$\begin{aligned}\left|\int_{t_1}^{t_2} g_j^i(t, \hat{x}_t^{(k)}, \hat{u}^{(k)}(t))\, dh(t)\right| &\leqq \int_{t_1}^{t_2} |g_j^i(t, \hat{x}_t^{(k)}, \hat{u}^{(k)}(t))|\, dh(t) \\ &\leqq \int_{t_1}^{t_2} |G(t, \hat{x}_t^{(k)}, \hat{u}^{(k)}(t))|\, dh(t) \leqq \int_{t_1}^{t_2} K\, dh(t) \quad \text{(by } A_6\text{)} \\ &= K(h(t_2) - h(t_1)),\end{aligned}$$

and therefore the integrals $\int g_j^i(t, \hat{x}_t^{(k)}, \hat{u}^{(k)}(t))\, dh(t)$ are absolutely continuous with respect to $h(t)$ uniformly in k, and bounded uniformly. Hence by [2, Theorem 27, p. 285] and since A_5 holds, we obtain

$$\lim_{k\to\infty} \int_{t_0}^{t} G(s, \hat{x}_s^{(k)}, \hat{u}^{(k)}(s))\, d\hat{u}^{(k)}(s) = \int_{t_0}^{t} G(s, x_s^*, u^*(s))\, du^*(s) \tag{5.10}$$

for all $t \in [t_0, \hat{t}_1]$. By (5.7), (5.8), (5.9), (5.10), we obtain

$$x^*(t) = \begin{cases} \varphi(t) & \text{for } t \in [\alpha, t_0], \\ \varphi(t_0) + \displaystyle\int_{t_0}^{t} f(s, x_s^*, u(s))\, ds & \\ \qquad + \displaystyle\int_{t_0}^{t} G(s, x_s^*, u^*(s))\, du^*(s) & \text{for } t \in (t_0, \hat{t}_1]. \end{cases} \tag{5.11}$$

We shall now show that

$$\lim_{k\to\infty} \hat{x}^{(k)}(t_1^{(k)}) = x^*(t_1^*). \tag{5.12}$$

We have

$$|x^{(k)}(t_1^{(k)}) - x^*(t_1^*)| \leqq |x^{(k)}(t_1^{(k)}) - x^*(t_1^{(k)})| + |x^*(t_1^{(k)}) - x^*(t_1^*)|. \tag{5.13}$$

Since x^* is continuous from the right (because u^* is), $x^*(t_1^{(k)}) \to x^*(t_1^*)$; and hence the last term on the right in (5.13) is zero as $k \to \infty$. We have

$$\begin{aligned}x^{(k)}(t_1^{(k)}) - x^*(t_1^{(k)}) &= \int_{t_0}^{t_1^{(k)}} f(s, x_s^{(k)}, u^{(k)}(s))\, ds + \int_{t_0}^{t_1^{(k)}} G(s, x_s^{(k)}, u^{(k)}(s))\, du^{(k)}(s) \\ &\quad - \int_{t_0}^{t_1^{(k)}} f(s, x_s^*, u^*(s))\, ds - \int_{t_0}^{t_1^{(k)}} G(s, x_s^*, u^*(s))\, du^*(s)\end{aligned}$$

$$= \int_{t_0}^{t_1^*} [f(s, x_s^{(k)}, u^{(k)}(s)) - f(s, x_s^*, u^*(s))]\, ds$$

$$+ \left[\int_{t_0}^{t_1^*} G(s, x_s^{(k)}, u^{(k)}(s))\, du^{(k)}(s) - \int_{t_0}^{t_1^*} G(s, x_s^*, u^*(s))\, du^*(s) \right]$$

$$+ \int_{t_1^*}^{t_1^{(k)}} [f(s, x_s^{(k)}, u^{(k)}(s)) - f(s, x_s^*, u^*(s))]\, ds$$

$$+ \left[\int_{t_1^*}^{t_1^{(k)}} G(s, x_s^{(k)}, u^{(k)}(s))\, du^{(k)}(s) - \int_{t_1^*}^{t_1^{(k)}} G(s, x_s^*, u^*(s))\, du^*(s) \right]$$

$$= I_1 + I_2 + I_3 + I_4, \quad \text{say.}$$

By (5.9) and (5.10), we have

$$I_1 \to 0, \quad I_2 \to 0 \quad \text{as } k \to \infty.$$

We have, further,

$$|I_3| \leqq \int_{t_1^*}^{t_1^{(k)}} |f(s, x_s^{(k)}, u^{(k)}(s))|\, ds + \int_{t_1^*}^{t_1^{(k)}} |f(s, x_s^*, u^*(s))|\, ds$$

$$\leqq 2 \int_{t_1^*}^{t_1^{(k)}} r(s)\, ds \to 0 \quad \text{as } k \to \infty,$$

$$|I_4| \leqq \int_{t_1^*}^{t_1^{(k)}} |G(s, x_s^{(k)}, u^{(k)}(s))|\, dh(s) + \int_{t_1^*}^{t_1^{(k)}} |G(s, x_s^*, u^*(s))|\, dh(s)$$

$$\leqq 2K[h(t_1^{(k)}) - h(t_1^*)] \to 0 \quad \text{as } k \to \infty,$$

since h is right continuous. Thus (5.12) is established.

Now, $x^{(k)}(t_1^{(k)}) \in T_{t_1^{(k)}}$ for $k = 1, 2, \cdots$, and

$$x^*(t_1^*) = \lim_{k \to \infty} x^{(k)}(t_1^{(k)}).$$

If $x^*(t_1^*)$ were not in $T_{t_1^*}$, then there would exist a neighborhood N of the compact set $T_{t_1^*}$ so that $x^*(t_1^*)$ is not in the closure $\bar{N}$ of N. But, since T_t is upper semicontinuous, $T_t \subset N$ for t sufficiently near t_1^*. Thus $x^{(k)}(t_1^{(k)}) \in N$ for large k and yet $x^*(t_1^*) \notin \bar{N}$. This contradiction shows that $x^*(t_1^*) \in T_{t_1^*}$. Hence the control u^* on $[t_0, t_1^*]$ belongs to A.

We have

$$J(u^{(k)}) = \int_{t_0}^{t_1^{(k)}} f^0(s, x^{(k)}(s), u^{(k)}(s))\, ds$$

$$= \int_{t_0}^{t_1^*} f^0(s, x^{(k)}(s), u^{(k)}(s))\, ds + \int_{t_1^*}^{t_1^{(k)}} f^0(s, x^{(k)}(s), u^{(k)}(s))\, ds.$$

Since $f^0(s, x^{(k)}(s), u^{(k)}(s))$ is uniformly bounded on $[t_0, t_1^{(k)}]$ and so the last integral in the above equation approaches zero, and since

$$x^{(k)}(t) \to x^*(t), \qquad u^{(k)}(t) \to u^*(t)$$

everywhere on $[t_0, t_1^*]$, therefore, by applying Legesgue's dominated convergence theorem to the first integral, we obtain

$$\lim_{k\to\infty} J(u^{(k)}) = \int_{t_0}^{t_1^*} f^0(s, x^*(s), u^*(s))\, ds = J(u^*).$$

But $\lim_{k\to\infty} J(u^{(k)}) = \tilde{m}$. Therefore, $J(u^*) = \tilde{m}$. Hence u^* on $[t_0, t_1^*]$ is an optimal control.

We shall now consider the case when $\{t_1^{(k)}\}$ is monotonic increasing. We extend all the controls $u^{(k)}$ to the interval $[t_0, \hat{t}_1]$, where $\hat{t}_1 = t_1^* + \delta$ for appropriately small $\delta > 0$, by defining

$$\hat{u}^{(k)}(t) = \begin{cases} u^{(k)}(t) & \text{for } t \in [t_0, t_1^{(k)}], \\ u^{(k)}(t_1^{(k)}) & \text{for } t \in (t_1^{(k)}, \hat{t}_1]. \end{cases} \tag{5.14}$$

As before, there exists a subsequence $\hat{u}^{(k)}$ (without changing the notation) and a right continuous function of bounded variation u^* such that everywhere on $[t_0, \hat{t}_1]$,

$$\lim_{k\to\infty} \hat{u}^{(k)}(t) = u^*(t) \tag{5.15}$$

and

$$\lim_{k\to\infty} \hat{x}^{(k)}(t) = x^*(t), \tag{5.16}$$

where $\hat{x}^{(k)}$ and x^* are the trajectories corresponding to $\hat{u}^{(k)}$ and u^* respectively.

We shall show that

$$\lim_{k\to\infty} x^{(k)}(t_1^{(k)}) = x^*(t_1^*). \tag{5.17}$$

Consider

$$|x^{(k)}(t_1^{(k)}) - x^*(t_1^*)| \leqq |x^{(k)}(t_1^{(k)}) - \hat{x}^{(k)}(t_1^*)| + |\hat{x}^{(k)}(t_1^*) - x^*(t_1^*)|. \tag{5.18}$$

Since $\hat{x}^{(k)}(t_1^*) \to x^*(t_1^*)$, the last term on the right in (5.18) approaches zero as $k \to \infty$. For the first term on the right, we write

$$\begin{aligned} x^{(k)}(t_1^{(k)}) - \hat{x}^{(k)}(t_1^*) &= \int_{t_0}^{t_1^{(k)}} f(s, x_s^{(k)}, \hat{u}^{(k)}(s))\, ds + \int_{t_0}^{t_1^{(k)}} G(s, x_s^{(k)}, \hat{u}^{(k)}(s))\, d\hat{u}^{(k)}(s) \\ &\quad - \int_{t_0}^{t_1^*} f(s, \hat{x}_s^{(k)}, \hat{u}^{(k)}(s))\, ds - \int_{t_0}^{t_1^{(k)}} G(s, \hat{x}_s^{(k)}, \hat{u}^{(k)}(s))\, d\hat{u}^{(k)}(s) \\ &= -\int_{t_1^{(k)}}^{t_1^*} f(s, \hat{x}_s^{(k)}, \hat{u}^{(k)}(s))\, ds - \int_{t_1^{(k)}}^{t_1^*} G(s, \hat{x}_s^{(k)}, \hat{u}^{(k)}(s))\, d\hat{u}^{(k)}(s) \\ &= I_1 + I_2, \quad \text{say.} \end{aligned}$$

Now,

$$|I_1| \leqq \int_{t_1^{(k)}}^{t_1^*} |f(s, \hat{x}_s^{(k)}, \hat{u}^{(k)}(s))|\, ds$$

$$\leqq \int_{t_1^{(k)}}^{t_1^*} r(s)\, ds \to 0 \quad \text{as } k \to \infty$$

and

$$I_2 = -\int_{t_1^{(k)}}^{t_1^*} G(s, \hat{x}_s^{(k)}, \hat{u}^{(k)}(s))\, d\hat{u}^{(k)}(s) = 0$$

for each k, since $\hat{u}^{(k)}(t) = u^{(k)}(t_1^{(k)})$ for $t \in [t_1^{(k)}, t_1^*]$ and is thus constant on $[t_1^{(k)}, t_1^*]$. Thus (5.17) is established. Exactly as before, $x^*(t_1^*) \in T_{t_1^*}$; and therefore u^* on $[t_0, t_1^*]$ belongs to A. It can be seen, as before, that $J(u^*) = \tilde{m}$. Hence u^* is an optimal control. This completes the proof.

Acknowledgment. The authors are very thankful to the referee for his comments which helped them to improve this paper.

REFERENCES

[1] N. DUNFORD AND J. SCHWARTZ, *Linear Operators. Part I*, Interscience, New York, 1964.
[2] L. M. GRAVES, *The Theory of Functions of Real Variables*, McGraw-Hill, New York, 1956.
[3] A. HALANAY, *Differential Equations: Stability, Oscillations, Time Lags*, Academic Press, New York, 1965.
[4] P. R. HALMOS, *Measure Theory*, East-West ed., Van Nostrand, New Delhi, 1964.
[5] I. HALPERIN, *Introduction to the Theory of Distributions*, University of Toronto Press, Toronto, 1952.
[6] E. J. MCSHANE, *Integration*, Princeton University Press, Princeton, 1944.
[7] M. E. MUNROE, *Introduction to Measure and Integration*, Addison-Wesley, Reading, Mass., 1959.
[8] I. P. NATANSON, *Theory of Functions of a Real Variable*, vol. I, Frederick Ungar, New York, 1955.
[9] M. N. OGUZTORELI, *Time-Lag Control Systems*, Academic Press, New York, 1966.
[10] T. PAVLIDIS, *Optimal control of pulse frequency modulated systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 676–684.
[11] ———, *Stability of systems described by differential equations containing impulses*, Ibid., AC-12 (1967), pp. 43–45.
[12] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
[13] W. W. SCHMAEDEKE, *Optimal control theory for nonlinear vector differential equations containing measures*, this Journal, 3 (1965), pp. 331–380.

SIAM J. CONTROL
Vol. 9, No. 1, February 1971

NECESSARY AND SUFFICIENT OPTIMALITY CONDITIONS FOR A CLASS OF DISTRIBUTED PARAMETER CONTROL SYSTEMS*

EARL R. BARNES†

Abstract. A class of optimal control problems arising in systems with distributed parameters is considered. Pontryagin's maximum principle is shown to be a necessary condition for optimality. Under certain convexity assumptions, it is shown that the maximum principle gives a sufficient condition for optimality.

The maximum principle of L. S. Pontryagin [4] gives a necessary condition for optimality for a large class of optimal control problems involving systems with lumped parameters. E. B. Lee has shown (cf. [1]) that the maximum principle is also a sufficient condition for optimality for a certain subclass of such problems. In this paper we shall derive analogous necessary and sufficient conditions for optimality in the form of a maximum principle for certain optimal control problems arising in systems with distributed parameters.

The paper is divided into two sections. The first section is concerned with the optimal control of systems (such as vibrating strings) that can be described by second order linear hyperbolic partial differential equations. In the second section, we discuss the problem of optimally controlling vibrating beams. The results obtained here are analogous to those obtained in Section 1 except now we restrict ourselves to problems in one space dimension.

In all cases we have tried to formulate problems that generalize certain control problems that have been studied for vibrating systems having finitely many degrees of freedom and lumped parameters. Other problems for vibrating strings and beams have been considered by Russell and Komkov in [2] and [3], respectively. These problems involve systems having quadratic cost functionals. A fairly extensive treatment of similar systems has been given by Lions in [11].

Necessary optimality conditions similar to ours have been obtained by Egorov in [5], [6] and [7] for a different class of distributed parameter control problems. However, our approach is different than Egorov's and we obtain certain advantages over his method of deriving necessary conditions. In the first place, our proofs are shorter and simpler. But more importantly, our technique of proof illustrates that Egorov's completeness assumption on the class of admissible controls (cf. [6]) can be dropped. This is desirable since there is apparently no technique available for determining when a given class of admissible controls is complete in the sense of Egorov.

---

* Received by the editors August 12, 1969, and in revised form June 1, 1970.

† IBM Watson Research Center, Yorktown Heights, New York 10598. A portion of this paper was taken from the author's doctoral dissertation, written under the directorship of Professor G. S. Jones at the Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, 1968. The author's graduate research was supported in part by an NDEA Fellowship and in part by a Minta Martin Fellowship.

1. The optimal control of a vibrating string.

1.1. Introduction. The problem of controlling vibrating systems is of considerable importance to engineers. Many vibrating systems can be approximated by lumped parameter systems which can in turn be described by a set of second order linear ordinary differential equations. Such systems have received considerable attention in the literature. See, for example, [8], [9] and [10, pp. 444–456].

The physical model for a vibrating system having lumped parameters is usually taken to be a system consisting of a finite number of springs and masses in combination. The control functions are then taken to represent forces that can be applied to the masses so as to bring the system to an equilibrium position in some optimal fashion. In this section we shall discuss the optimal control of the simplest vibrating systems having distributed parameters. Such systems can be described by one-dimensional wave equations. We assume that a physical model for our system can be taken to be an inhomogeneous string stretched between two fixed points and undergoing small planar vibrations. We also assume that we are able to apply a distributed force which acts in the plane of vibration and normal to the string, along the entire length of the string, in an effort to control the vibrations. Our aim is to determine a force which eliminates the vibrations in some optimal fashion.

Assume that the string is stretched between the points $x = 0$ and $x = l$ on the x-axis and is vibrating in a vertical plane. Denote the displacement at the point x and time t by $u(x, t)$ and denote the externally applied force by f. Then, assuming that the frictional forces acting on the string are negligible, u and f are related by the differential equation

$$\mu(x)\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(p(x)\frac{\partial u}{\partial x}\right) - q(x)u + f(x, t), \tag{1.1}$$

$$0 \leqq x \leqq l, \quad t > 0.$$

μ is the variable density of the string, p is the tension, and q the elastic restoring force. u satisfies boundary and initial conditions of the form

$$\begin{aligned} &u(0, t) = 0, \qquad u(l, t) = 0, \\ &u(x, 0) = \varphi(x), \qquad u_t(x, 0) = \psi(x). \end{aligned} \tag{1.2}$$

The functions μ, p and q in (1.1) are assumed to satisfy μ, $p \in c^3$, $q \in c^1$, and $\mu(x)$, $q(x)$, $p(x) > 0$ for $0 \leqq x \leqq l$. The initial conditions φ and ψ are elements of $L_2[0, l]$. The function f denotes a control parameter. f is assumed to be admissible in the sense of the following definition.

DEFINITION 1.1. Let $F \subset E_1$ be a given constraint set. A square integrable function f defined on $[0, l] \times [0, T]$ and taking values in F is called an *admissible control.*

For a given admissible control f, let $\{f_n\}$ be a sequence of continuous functions converging to f in the mean square. Similarly, let $\{\varphi_n\}$ and $\{\psi_n\}$ be sequences of c^2 and c^3 functions, respectively, converging in the mean square to φ and ψ. Then it is well known that there exist twice continuously differentiable functions u_n satisfying (1.1)–(1.2) with f, φ and ψ replaced by f_n, φ_n and ψ_n, respectively (cf. [12, p. 157]).

Moreover, each u_n satisfies the energy inequality

$$\int_0^l \left[\mu(x)\left(\frac{\partial u_n}{\partial t}(x,t)\right)^2 + p(x)\left(\frac{\partial u_n}{\partial x}(x,t)\right)^2 + q(x)u_n^2(x,t) \right] dx$$

$$\leqq c \int_0^t \int_0^l f_n^2(x,\tau)\, dx\, d\tau \tag{1.3}$$

$$+ c \int_0^l \left[\mu(x)\psi_n^2(x) + p(x)\left(\frac{\partial \varphi_n}{\partial x}(x)\right)^2 + q(x)\varphi_n^2(x) \right] dx.$$

c is a constant, independent of the functions u_n.

If we apply the energy inequality to the differences $u_k - u_n$, we see that the limits

$$\lim_{k\to\infty} \frac{\partial u_k}{\partial t}(x,t), \quad \lim_{k\to\infty} \frac{\partial u_k}{\partial x}(x,t), \quad \lim_{k\to\infty} u_k(x,t)$$

exist in $L_2[0, l]$ uniformly in t for $0 \leqq t \leqq T$. We shall denote these limits by

$$\frac{\partial u}{\partial t}(x,t), \quad \frac{\partial u}{\partial x}(x,t) \quad \text{and} \quad u(x,t),$$

respectively, and shall refer to this convergence as convergence in the energy norm.

The function u will be called a solution of (1.1)–(1.2) having finite energy. $\partial u/\partial t$ and $\partial u/\partial x$ will be called generalized derivatives of u.

In the sequel we shall obtain certain results by formally integrating by parts. The integrations will be valid for classical solutions, and hence, the results obtained will be valid for solutions with finite energy, by passing to the limit in the result for classical solutions.

1.2. Formulation of the control problem. The system (1.1)–(1.2) will be at rest at time $T > 0$ if the condition $u(x, T) \equiv 0$ and $u_t(x, T) \equiv 0$ are satisfied. These conditions are equivalent to

$$\int_0^l u^2(x,T)\, dx = 0 \quad \text{and} \quad \int_0^l u_t^2(x,T)\, dx = 0. \tag{1.4}$$

The constraints (1.4) are special cases of more general terminal constraints assumed in the discussion to follow.

Let $h_1(x,u)$, $h_2(x,u)$, $g_1(x,u)$, $g_2(x,u)$, $F_{-2}(x,t,u,f)$, $F_{-1}(x,t,u,f)$, $F_0(x,t,u,f)$, $\cdots$, $F_n(x,t,u,f)$ be continuous functions, each having continuous first and second derivatives with respect to u. Consider the following optimal control problem.

Find an admissible control f^0 which minimizes the functional

$$J_0(f) = \int_0^l [g_1(x,u(x,T)) + g_2(x,u_t(x,T))]\, dx$$

$$+ \int_0^T \int_0^l F_0(x,t,u(x,t),f(x,t))\, dx\, dt \tag{1.5}$$

subject to the differential equation (1.1)–(1.2) and the constraints

$$
\begin{aligned}
&\int_0^l h_2(x, u_t(x, T))\,dx + \int_0^T \int_0^l F_{-2}(x, t, u(x, t), f(x, t))\,dx\,dt = c_{-2},\\
&\int_0^l h_1(x, u(x, T))\,dx + \int_0^T \int_0^l F_{-1}(x, t, u(x, t), f(x, t))\,dx\,dt = c_{-1},\\
&\int_0^T \int_0^l F_i(x, t, u(x, t), f(x, t))\,dx\,dt \leqq c_i, \qquad 1 \leqq i \leqq n',\\
&\int_0^T \int_0^l F_i(x, t, u(x, t), f(x, t))\,dx\,dt = c_i, \qquad n' < i \leqq n.
\end{aligned} \tag{1.6}
$$

The discussion to follow is still applicable if all these constraints are of one type, either equality or inequality.

DEFINITION 1.2. An admissible control which solves this control problem is said to be an *optimal control.*

1.3. Necessary and sufficient conditions for optimality. A set of necessary conditions for optimality is contained in the following theorem. We shall show that the conditions stated in the theorem are also sufficient for a certain class of problems.

THEOREM 1.1 (Maximum principle). *In order that an admissible control* f^0, *with the response* u^0, *be optimal, it is necessary that there exist constants* $\lambda_{-2}, \lambda_{-1}, \lambda_0, \cdots, \lambda_n$ *and a solution* $v(x, t)$ *of*

$$
\begin{aligned}
&\mu(x)\frac{\partial^2 v}{\partial t^2} = \frac{\partial}{\partial x}\left(p(x)\frac{\partial v}{\partial x}\right) - q(x)v\\
&\qquad + \sum_{i=-2}^{n} \lambda_i \frac{\partial F_i}{\partial u}(x, t, u^0(x, t), f^0(x, t)),\\
&v(0, t) = 0, \qquad v(l, t) = 0,\\
&v(x, T) = \frac{1}{\mu(x)}\lambda_0 \frac{\partial g_2}{\partial u}(x, u_t^0(x, T)) + \frac{1}{\mu(x)}\lambda_{-2}\frac{\partial h_2}{\partial u}(x, u_t^0(x, T)),\\
&\frac{\partial v}{\partial t}(x, T) = -\frac{1}{\mu(x)}\lambda_{-1}\frac{\partial h_1}{\partial u}(x, u^0(x, T)) - \frac{1}{\mu(x)}\lambda_0 \frac{\partial g_1}{\partial u}(x, u^0(x, T))
\end{aligned} \tag{1.7}
$$

such that

$$
\begin{aligned}
&\max_{f \in F}\left[v(x, t)f + \sum_{i=-2}^{n} \lambda_i F_i(x, t, u^0(x, t), f)\right]\\
&\qquad = v(x, t)f^0(x, t) + \sum_{i=-2}^{n} \lambda_i F_i(x, t, u^0(x, t), f^0(x, t))
\end{aligned}
$$

for almost all points $(x, t) \in (0, l) \times (0, T)$. *The* λ_i *can be chosen such that* $\lambda_i \leqq 0$ *for* $0 \leqq i \leqq n'$ *and some* $\lambda_i \neq 0$.

Remark. The conclusion of this theorem will be deduced from three lemmas. Before stating and proving the lemmas, we introduce a special perturbation of an optimal control.

Suppose that f^0 is an optimal control with the corresponding response u^0. Let $(x_1, t_1), \cdots, (x_N, t_N)$ be N arbitrary points in the open region $(0, l) \times (0, T)$ and let $f_1, \cdots, f_N$ be N arbitrary points in F. We may assume that $x_1 \leqq x_2 \leqq \cdots \leqq x_N$. Choose $\delta > 0$ such that $x_i + N\sqrt{\delta} < x_j$ if $x_i < x_j$, and such that $x_N + N\sqrt{\delta} < l$ and $t_i + \sqrt{\delta} < T$ for each i. Let $\varepsilon_1, \cdots, \varepsilon_N$ be real parameters satisfying $0 \leqq \varepsilon_j \leqq \delta$ $(1 \leqq j \leqq N)$. Let

$$X_1 = x_1 \quad \text{and} \quad X_j = x_j + \sqrt{\varepsilon_1} + \cdots + \sqrt{\varepsilon_{j-1}}$$

for $1 < j \leqq N$. Clearly, the intervals $X_j \leqq x \leqq X_j + \sqrt{\varepsilon_j}$ are nonoverlapping. Therefore, the rectangles

$$R_j : [X_j, X_j + \sqrt{\varepsilon_j}] \times [t_j, t_j + \sqrt{\varepsilon_j}]$$

are nonoverlapping.

Define the admissible control f_ε on $[0, l] \times [0, T]$ by

$$f_\varepsilon(x, t) = \begin{cases} f^0(x, t) & \text{if } (x, t) \notin \bigcup_{j=1}^{N} R_j, \\ f_j & \text{if } (x, t) \in R_j, \quad j = 1, \cdots, N. \end{cases} \tag{1.8}$$

Here ε denotes the vector $(\varepsilon_1, \cdots, \varepsilon_N) \in E_N$. The norm of ε is defined by

$$|\varepsilon| = \varepsilon_1 + \cdots + \varepsilon_N.$$

LEMMA 1.1. *Let u_ε denote the response of the system* (1.1)–(1.2) *due to f_ε. Let $\Delta u = u_\varepsilon - u^0$ and let $\Delta f = f_\varepsilon - f^0$. Then*

$$\int_0^l \Delta u^2(x, T)\, dx = o(\varepsilon), \qquad \int_0^l \Delta u_t^2(x, T)\, dx = o(\varepsilon)$$

and

$$\int_0^T \int_0^l \Delta u^2(x, t)\, dx\, dt = o(\varepsilon).$$

$o(\varepsilon)$ is a quantity such that

$$\lim_{\varepsilon \to 0} (o(\varepsilon)/|\varepsilon|) = 0.$$

Proof. Let

$$E(t) = \int_0^l \left\{ \mu(x) \left(\frac{\partial \Delta u}{\partial t}(x, t) \right)^2 + p(x) \left(\frac{\partial \Delta u}{\partial x}(x, t) \right)^2 + q(x) \Delta u^2(x, t) \right\} dx.$$

Proceeding formally as in [14, p. 145], we obtain

$$E(t) = \int_0^t \frac{dE(\tau)}{d\tau} d\tau = 2 \int_0^t \int_0^l \frac{\partial \Delta u}{\partial t}(x, \tau) \Delta f(x, \tau)\, dx\, d\tau. \tag{1.9}$$

This equality is valid for weak solutions since the latter is the limit of classical solutions in the energy norm.

Applying the Cauchy–Schwarz inequality to the space integral, we obtain

$$E(t) \leqq 2\int_0^t \left(\int_0^l \left(\frac{\partial \Delta u}{\partial t}\right)^2 dx\right)^{1/2}\left(\int_0^l (\Delta f)^2\, dx\right)^{1/2} d\tau$$

$$\leqq \int_0^T 2E^{1/2}(\tau)\left(\int_0^l (\Delta f)^2\, dx\right)^{1/2} d\tau.$$

It follows that

$$\sup_{0\leqq t\leqq T} E(t) \leqq 2 \sup_{0\leqq t\leqq T} E^{1/2}(t) \int_0^T \left(\int_0^l (\Delta f)^2\, dx\right)^{1/2} d\tau \tag{1.10}$$

$$= 2 \sup_{0\leqq t\leqq T} E^{1/2}(t) \sum_{i=1}^{N} O(\varepsilon_i^{3/4}).$$

$O(r)$ is a quantity such that

$$\lim_{r\to 0^+} (O(r)/r) = \text{const.}$$

Inequality (1.10) shows that, for each $t \in [0, T]$, we have

$$0 \leqq E^{1/2}(t) \leqq 2 \sum_{i=1}^{N} O(\varepsilon_i^{3/4})$$

or

$$E(t) = o(\varepsilon). \tag{1.11}$$

The conclusion of the lemma now follows immediately from (1.11), since the coefficients q and μ are bounded away from zero.

Let L denote the differential operator defined by

$$Lu \equiv \mu(x)\frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x}\left(p(x)\frac{\partial u}{\partial x}\right) + q(x)u.$$

The formal adjoint of L has the same definition. Thus we shall write

$$Mv = \mu(x)\frac{\partial^2 v}{\partial t^2} - \frac{\partial}{\partial x}\left(p(x)\frac{\partial v}{\partial x}\right) + q(x)v$$

to denote the adjoint of L. We use M instead of L to distinguish the roles of u and v in the discussion to follow.

LEMMA 1.2. *Let u and v be two functions for which the expressions Lu and Mv are square integrable. Thus u and v are weak solutions in the sense mentioned above. If u and v satisfy the boundary conditions* (1.2), *then*

$$\int_0^T \int_0^l [vLu - uMv]\, dx\, dt = \int_0^l \Big\{\mu(x)v(x, T)\frac{\partial u}{\partial t}(x, T) - \mu(x)u(x, T)\frac{\partial v}{\partial t}(x, T)$$

$$+ \mu(x)u(x, 0)\frac{\partial v}{\partial t}(x, 0) - \mu(x)v(x, 0)\frac{\partial u}{\partial t}(x, 0)\Big\} dx. \tag{1.12}$$

Proof. Formally we have

$$vLu - uMv = -\frac{\partial}{\partial x}\left(vp(x)\frac{\partial u}{\partial x} - up(x)\frac{\partial v}{\partial x}\right) + \frac{\partial}{\partial t}\left(-\mu(x)u\frac{\partial v}{\partial t} + \mu(x)v\frac{\partial u}{\partial t}\right).$$

Therefore, the conclusion of the lemma can be obtained for c^2 functions by integrating both sides of this identity over the region $[0, l] \times [0, T]$ and making use of the boundary conditions in (1.2). The lemma holds for weak solutions u and v by passing to the limit in (1.12) for c^2 functions.

DEFINITION 1.3. Let f^0 be any admissible control with the response u^0. Let v be the solution of system (1.9) for arbitrary constants $\lambda_{-2}, \lambda_{-1}, \lambda_0, \cdots, \lambda_n$. Let g be any one of the following functions:

$$g(x, t) = F_i(x, t, u^0(x, t), f), \quad f \in F \text{ is fixed},$$
$$g(x, t) = F_i(x, t, u^0(x, t), f^0(x, t)),$$
$$g(x, t) = v(x, t)f, \quad f \in F \text{ is fixed},$$
$$g(x, t) = v(x, t)f^0(x, t).$$

We shall call $(\bar{x}, \bar{t})$ a *regular point* of f^0 if

$$\int_{\bar{t}}^{\bar{t}+\sqrt{\varepsilon}} \int_{\bar{x}}^{\bar{x}+\sqrt{\varepsilon}} g(x, t)\, dx\, dt = \varepsilon g(\bar{x}, \bar{t}) + o(\varepsilon)$$

for any $\varepsilon > 0$ and sufficiently small.

It follows from [15, Theorem 6.3, p. 118], that almost all points of $[0, l] \times [0, T]$ are regular for each admissible control f.

LEMMA 1.3. *Let the functions f^0 and f_ε be defined as in Lemma 1.1 where we assume now that the points (x_i, t_i) are regular, $i = 1, \cdots, N$. If $N = 1$, then there exist constants $\lambda_{-2}, \lambda_{-1}, \lambda_0, \lambda_1, \cdots, \lambda_n$ (not all zero) such that*

(a) $$\lambda_i \leqq 0 \quad \text{for } 0 \leqq i \leqq n',$$

(b) $$\lim_{\varepsilon \to 0^+} \frac{\bar{J}(f_\varepsilon) - \bar{J}(f^0)}{\varepsilon} \leqq 0,$$

where $\bar{J}(f)$, for any admissible control f, is defined by

$$\bar{J}(f) = \lambda_0 \int_0^l [g_1(x, u(x, T)) + g_2(x, u_t(x, T))]\, dx$$
$$+ \int_0^l [\lambda_{-2}h_2(x, u_t(x, T)) + \lambda_{-1}h_1(x, u(x, T))]\, dx$$
$$+ \int_0^T \int_0^l \sum_{i=-2}^{n} \lambda_i F_i(x, t, u(x, t), f(x, t))\, dx\, dt.$$

Proof. Define the functionals $J_{-2}, J_{-1}, J_1, \cdots, J_n$ on the class of admissible controls by

$$J_{-1}(f) = \int_0^l h_1(x, u(x, T))\,dx + \int_0^T \int_0^l F_{-1}(x, t, u(x, t), f(x, t))\,dx\,dt,$$

$$J_{-2}(f) = \int_0^l h_2(x, u_t(x, T))\,dx + \int_0^T \int_0^l F_{-2}(x, t, u(x, t), f(x, t))\,dx\,dt,$$

$$J_i(f) = \int_0^T \int_0^l F_i(x, t, u(x, t), f(x, t))\,dx\,dt,$$

$i = 1, \cdots, n$. J_0 is defined by (1.5). Let J denote the vector-valued functional $(J_{-2}, J_{-1}, J_0, \cdots, J_n)$. Let Y denote the set

$$Y = \{J(f) : f \text{ admissible}\} \subset E_{n+3}.$$

We assume the reader is familiar with the terminology and the generalized multiplier rule introduced by Hestenes in [13, Chap. IV]. We shall construct a derived set K for the set Y at $J(f^0)$.

Define functions $v_{-2}, v_{-1}, v_0, \cdots, v_n$ by requiring the following conditions to be satisfied:

(1.13a) $$\mu(x)\frac{\partial^2 v_j}{\partial t^2} = \frac{\partial}{\partial x}\left(p(x)\frac{\partial v_j}{\partial x}\right) - q(x)v_j + \frac{\partial F_j}{\partial u}(x, t, u^0(x, t), f^0(x, t))$$
$$0 \leqq x \leqq l, \quad 0 \leqq t \leqq T, \quad -2 \leqq j \leqq n,$$

(1.13b) $$v_j(0, t) = 0, \quad v_j(l, t) = 0, \qquad -2 \leqq j \leqq n,$$

(1.13c) $$v_{-2}(x, T) = \frac{1}{\mu(x)}\frac{\partial h_2}{\partial u}(x, u_t^0(x, T)), \qquad \frac{\partial v_{-2}}{\partial t}(x, T) = 0,$$

(1.13d) $$v_{-1}(x, T) \equiv 0, \qquad \frac{\partial v_{-1}}{\partial t}(x, T) = -\frac{1}{\mu(x)}\frac{\partial h_1}{\partial u}(x, u^0(x, T)),$$

(1.13e) $$v_0(x, T) = \frac{1}{\mu(x)}\frac{\partial g_2}{\partial u}(x, u_t^0(x, T)),$$
$$\frac{\partial v_0}{\partial t}(x, T) = -\frac{1}{\mu(x)}\frac{\partial g_1}{\partial u}(x, u^0(x, T)),$$

(1.13f) $$v_j(x, T) = 0, \qquad \frac{\partial v_j}{\partial t}(x, T) = 0, \qquad j = 1, \cdots, n.$$

Thus, each v_j is defined to be a solution of a boundary value problem satisfying certain terminal conditions. The transformation $x \to x$, $t \to T - t$ transforms each of these problems into an initial-boundary value problem which has a solution for the same reason that the system (1.1)–(1.2) has a solution for each admissible control.

For each point $(x, t) \in (0, l) \times (0, T)$ and for each point $\bar{f} \in F$, define the scalars

$$k^i(x, t, \bar{f}) = v_i(x, t)(\bar{f} - f^0(x, t)) + F_i(x, t, u^0(x, t), \bar{f}) - F_i(x, t, u^0(x, t), f^0(x, t)) \qquad i = -2, -1, 0, \cdots, n. \tag{1.14}$$

We shall show that the set

$$K = \{k | k = (k^{-2}(x, t, \bar{f}), \cdots, k^n(x, t, \bar{f})), (x, t) \text{ a regular point of } f^0, \bar{f} \in F\}$$

is a derived set for Y at $J(f^0)$.

Let $k_1, \cdots, k_N$ be an arbitrary finite collection of vectors from K. We must show that there exist points $J_\varepsilon \in Y$ depending continuously on the vector parameter $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_N)$, for all sufficiently small positive values of ε, such that

$$J_\varepsilon = J(f^0) + \sum_{j=1}^{N} k_j \varepsilon_j + o(\varepsilon).$$

Since $k_j \in K$, $j = 1, \cdots, N$, there exist points $(x_1, t_1), \cdots, (x_N, t_N)$ of regularity of f^0 and points $f_1, \cdots, f_N \in F$ such that

$$k_j = (k^{-2}(x_j, t_j, f_j), \cdots, k^n(x_j, t_j, f_j)),$$

$j = 1, \cdots, N$. We shall show that J_ε can be defined by $J_\varepsilon = J(f_\varepsilon)$, where f_ε is the admissible control defined in Lemma 1.1.

We have, for $i = 1, \cdots, n$,

$$\begin{aligned} J_i(f_\varepsilon) - J_i(f^0) &= \int_0^T \int_0^l [F_i(x, t, u_\varepsilon(x, t), f_\varepsilon(x, t)) \\ &\qquad - F_i(x, t, u^0(x, t), f^0(x, t))]\, dx\, dt \\ &= \int_0^T \int_0^l [F_i(x, t, u_\varepsilon(x, t), f_\varepsilon(x, t)) \\ &\qquad - F_i(x, t, u^0(x, t), f_\varepsilon(x, t))]\, dx\, dt \\ &\quad + \int_0^T \int_0^l [F_i(x, t, u^0(x, t), f_\varepsilon(x, t)) \\ &\qquad - F_i(x, t, u^0(x, t), f^0(x, t))]\, dx\, dt \\ &= \int_0^T \int_0^l \frac{\partial F_i}{\partial u}(x, t, u^0(x, t), f^0(x, t)) \Delta u(x, t)\, dx\, dt \\ &\quad + \sum_{j=1}^{N} \varepsilon_j [F_i(x_j, t_j, u^0(x_j, t_j), f_j) \\ &\qquad - F_i(x_j, t_j, u^0(x_j, t_j), f^0(x_j, t_j))] + \sum_{j=1}^{N} o(\varepsilon_j). \end{aligned} \tag{1.15}$$

In (1.15) we used the fact that f^0 is regular at each of the points (x_j, t_j) and the fact that

$$\int_0^T \int_0^l \Delta u^2(x, t)\, dx\, dt = o(\varepsilon).$$

If now we use the fact that

$$Mv_i = \frac{\partial F_i}{\partial u}(x, t, u^0(x, t), f^0(x, t)), \qquad i = 1, \cdots, n,$$

and substitute in (1.12), we obtain

$$\begin{aligned}\int_0^T \int_0^l \Delta u(x, t) Mv_i\, dx\, dt &= \int_0^T \int_0^l v_i(x, t)(f_\varepsilon(x, t) - f^0(x, t)) \\ &= \sum_{j=1}^N \varepsilon_j v_i(x_j, t_j)(f_j - f^0(x_j, t_j)) + o(\varepsilon).\end{aligned}$$

It now follows from (1.14) and (1.15) that

$$J_i(f_\varepsilon) - J_i(f^0) = \sum_{j=1}^N \varepsilon_j k_j^i + o(\varepsilon), \qquad i = 1, \cdots, n, \tag{1.16}$$

where k_j^i denotes the ith component of k_j.

For $i = 0$, we have

$$\begin{aligned}J_0(f_\varepsilon) - J_0(f^0) = \int_0^l \left[\frac{\partial g_1}{\partial u}(x, u^0(x, T))\Delta u(x, T) + \frac{\partial g_2}{\partial u}(x, u_t^0(x, T))\Delta u_t(x, T)\right] dx \\ + \sum_{j=1}^N \varepsilon_j [F_0(x_j, t_j, u^0(x_j, t_j), f_j) \\ - F_0(x_j, t_j, u^0(x_j, t_j), f^0(x_j, f_j))] \\ + \int_0^T \int_0^l (Mv_0)\Delta u(x, t)\, dx\, dt + o(\varepsilon).\end{aligned}$$

If now we observe that by (1.12) and (1.13e)

$$\begin{aligned}\int_0^T \int_0^l \Delta u(x, t) Mv_0\, dx\, dt = \int_0^T \int_0^l v_0(x, t)(f_\varepsilon(x, t) - f^0(x, t))\, dx\, dt \\ - \int_0^l \left[\frac{\partial g_1}{\partial u}(x, u^0(x, T))\Delta u(x, T) \right. \\ \left. + \frac{\partial g_2}{\partial u}(x, u_t(x, T))\Delta u_t(x, T)\right] dx,\end{aligned}$$

we obtain (1.16) for $i = 0$.

Equations (1.12) and (1.13c)–(1.13d) can be used in a similar fashion to prove (1.16) for $i = -1$ and -2. Thus we can write

$$J(f_\varepsilon) - J(f^0) = \sum_{j=1}^{N} \varepsilon_j k_j(x_j, t_j, f_j) + o(\varepsilon).$$

Define $J_\varepsilon = J(f_\varepsilon)$. The proof that K is a derived set for Y at $J(f^0)$ is now complete.

It now follows from Theorem 3.1 in [13, p. 177] that there exist multipliers $\lambda_{-2}, \lambda_{-1}, \lambda_0, \cdots, \lambda_n$ with $\lambda_i \leqq 0$ for $0 \leqq i \leqq n'$ and some $\lambda_i \neq 0$, such that

$$\sum_{i=-2}^{n} \lambda_i k^i \leqq 0 \tag{1.17}$$

for any vector $k = (k^{-2}, k^{-1}, k^0, \cdots, k^n)$ in K.

To obtain the conclusion of Lemma 1.3, we take $N = 1$ in the foregoing discussion and put

$$\bar{J} = \sum_{i=-2}^{n} \lambda_i J_i.$$

We then have, by (1.16),

$$\bar{J}(f_\varepsilon) - \bar{J}(f^0) = \varepsilon \sum_{i=-2}^{n} \lambda_i k^i + o(\varepsilon)$$

for any vector $k = (k^{-2}, k^{-1}, k^0, \cdots, k^n)$ in K. The proof of Lemma 1.3 now follows from the fact that

$$\lim_{\varepsilon \to 0^+} \frac{\bar{J}(f_\varepsilon) - \bar{J}(f^0)}{\varepsilon} = \sum_{i=-2}^{n} \lambda_i k^i(x, t, f) \leqq 0.$$

Proof of Theorem 1.1. Assume that the admissible control f^0 is optimal. Let (x, t) be a regular point of f^0. Then, by Lemma 1.3, there exist constants $\lambda_{-2}, \lambda_{-1}, \lambda_0, \cdots, \lambda_n$ independent of (x, t), with $\lambda_i \leqq 0$ for $0 \leqq i \leqq n'$ and some $\lambda_i \neq 0$, such that

$$\sum_{i=-2}^{n} \lambda_i [v_i(x, t)(f - f^0(x, t)) + F_i(x, t, u^0(x, t), f) - F_i(x, t, u^0(x, t), f^0(x, t))] \leqq 0$$

for any point $f \in F$. The v_i's are defined in (1.13). This says that the function

$$\sum_{i=-2}^{n} \lambda_i [v_i(x, t) f + F_i(x, t, u^0(x, t), f)]$$

attains its maximum value on F at $f = f^0(x, t)$. If we put $v(x, t) = \sum_{i=-2}^{n} \lambda_i v_i(x, t)$, we obtain the conclusion of the maximum principle. This completes the proof of Theorem 1.1.

THEOREM 1.2. *Consider the control system* (1.1), (1.2), (1.5), (1.6). *Assume*

(a) $g_1, g_2, F_0, \cdots, F_{n'}$, *are convex functions of* u;

(b) *the functions* F_i *are of the form*

$$F_i(x, t, u, f) = F^i(x, t, u) + H^i(x, t, f), \qquad i = -2, \cdots, n.$$

Suppose there exist an admissible control f^0 *and constants* $\lambda_{-2}, \lambda_{-1}, \lambda_0, \cdots, \lambda_n$ *such that*

$$\max_{f \in F} \left[v(x,t)f + \sum_{i=-2}^{n} \lambda_i H^i(x,t,f) \right] = v(x,t)f^0(x,t) + \sum_{i=-2}^{n} \lambda_i H^i(x,t,f^0(x,t)),$$

where v *is a nonzero solution of*

$$Mv = \sum_{i=-2}^{n} \lambda_i \frac{\partial F^i}{\partial u}(x,t,u^0(x,t))$$

with the terminal and boundary conditions (1.9). *Assume further that*

(c) $\lambda_0 < 0, \lambda_i \leqq 0$ *for* $i = 1, \cdots, n'$;

(d) *the constraints* (1.6) *are satisfied by* f^0;

(e) *if strict inequality holds in* (1.6), *the corresponding multiplier* λ_i *is zero*;

(f) $-\lambda_i F^i, -\lambda_{-2} h_2, -\lambda_{-1} h_1$ *are convex functions of* u *for* $n' < i \leqq n$.

Then f^0 *is optimal.*

Remarks. Condition (e) is a necessary assumption since, as Hestenes shows in the proof of his multiplier rule [13], if an optimal control satisfies a strict inequality in (1.6), the corresponding multiplier must be zero. (f) is satisfied when the functions h_1, h_2, F_i, $n' < i \leqq n$, are linear in u.

Proof of Theorem 1.2. If f and u satisfy (1.6), then by (e),

$$\int_0^T \int_0^l \lambda_i [F^i(x,t,u(x,t)) - F^i(x,t,u^0(x,t))]\,dx\,dt$$

$$+ \int_0^T \int_0^l \lambda_i [H^i(x,t,f(x,t)) - H^i(x,t,f^0(x,t))]\,dx\,dt \geqq 0$$

for $i = 1, \cdots, n$. A similar result holds for $i = -2$ and -1.

It follows that

$$\begin{aligned} -\lambda_0[J_0(f) - J_0(f^0)] \geqq & -\lambda_0 \int_0^l [g_2(x,u_t(x,T)) - g_2(x,u_t^0(x,T)) + g_1(x,u(x,T)) \\ & - g_1(x,u^0(x,T))]\,dx \\ & - \sum_{i=-2}^{n} \int_0^T \int_0^l \lambda_i [F^i(x,t,u(x,t)) - F^i(x,t,u^0(x,t))]\,dx\,dt \\ & - \sum_{i=-2}^{n} \int_0^T \int_0^l \lambda_i [H^i(x,t,f(x,t)) - H^i(x,t,f^0(x,t))]\,dx\,dt \\ & - \lambda_{-1} \int_0^l [h_1(x,u(x,T)) - h_1(x,u^0(x,T))]\,dx \\ & - \lambda_{-2} \int_0^l [h_2(x,u_t(x,T)) - h_2(x,u_t^0(x,T))]\,dx. \end{aligned} \tag{1.18}$$

Therefore, by the convexity assumption (f),

$$
\begin{aligned}
-\lambda_0[J_0(f) - J_0(f^0)] \geqq & -\lambda_0 \int_0^l \left[\frac{\partial g_1}{\partial u}(x, u^0(x, T))\Delta u(x, T) \right. \\
& \left. + \frac{\partial g_2}{du}(x, u_t(x, T))\, \Delta u_t(x, T) \right] dx \\
& - \int_0^T \int_0^l \sum_{i=-2}^{n} \lambda_i \frac{\partial F^i}{\partial u}(x, t, u^0(x, t))\Delta u(x, t)\, dx\, dt \\
& + \int_0^T \int_0^l \sum_{i=-2}^{n} \lambda_i [H^i(x, t, f^0(x, t)) - H^i(x, t, f(x, t))]\, dx\, dt \\
& - \lambda_{-1} \int_0^l \frac{\partial h_1}{\partial u}(x, u^0(x, T))\Delta u(x, T)\, dx \\
& - \lambda_{-2} \int_0^l \frac{\partial h_2}{\partial u}(x, u_t^0(x, T))\Delta u_t(x, T)\, dx \\
= & -\lambda_0 \int_0^l \left[\frac{\partial g_1}{\partial u}(x, u^0(x, T))\Delta u(x, T) \right. \\
& \left. + \frac{\partial g_2}{\partial u}(x, u_t^0(x, T))\Delta u_t(x, T) \right] dx \\
& - \int_0^T \int_0^l (Mv)\Delta u(x, t)\, dx\, dt \\
& + \int_0^T \int_0^l \sum_{i=-2}^{n} \lambda_i [H^i(x, t, f^0(x, t)) - H^i(x, t, f(x, t))]\, dx\, dt \\
& - \int_0^l \left[\lambda_{-1} \frac{\partial h_1}{\partial u}(x, u^0(x, T))\Delta u(x, T) \right. \\
& \left. + \lambda_{-2} \frac{\partial h_2}{\partial u}(x, u_t^0(x, T))\Delta u_t(x, T) \right] dx.
\end{aligned}
$$

If we now apply Lemma 1.2 and the conditions (1.7), we obtain

$$
\begin{aligned}
-\lambda_0[J_0(f) - J_0(f^0)] \geqq & \int_0^T \int_0^l \{v(x, t)[f^0(x, t) - f(x, t)] \\
& + \sum_{i=-2}^{n} \lambda_i [H^i(x, t, f^0(x, t)) - H^i(x, t, f(x, t))]\}\, dx\, dt.
\end{aligned} \tag{1.19}
$$

The right-hand side of (1.19) is nonnegative. Hence, $J_0(f) - J_0(f^0) \geqq 0$.

1.4. Generalizations. We wish to point out briefly how the results of the previous section can be extended to more general systems. Thus, let us consider a

system governed by the general linear hyperbolic partial differential equation

$$\frac{\partial^2 u}{\partial t^2} = \sum_{i,j=1}^{m} \frac{\partial}{\partial x_i}\left(p_{ij}(x)\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{m} r_i(x)\frac{\partial u}{\partial x_i} + q(x)u + f(x,t), \qquad (x,t) \in \Omega \times [0,T] = Q, \tag{1.20}$$

together with boundary and initial conditions of the form

$$u(x,0) = \varphi(x), \qquad \frac{\partial u}{\partial t}(x,0) = \psi(x) \quad \text{in } \Omega, \qquad u(x,t) = 0 \quad \text{on } \partial\Omega \times [0,T]. \tag{1.21}$$

Here Ω represents an open connected set in E_m and $x = (x_1, \cdots, x_m)$ is a point in Ω. $\partial\Omega$ denotes the boundary of Ω.

The functions p_{ij}, $\partial p_{ij}/\partial x_k$, r_i, $\partial r_i/\partial x$ and q are assumed to be bounded and measurable in Ω. Moreover, we assume that $p_{ij} = p_{ji}$ and that there exist c and C such that

$$c\sum_{i=1}^{m} \xi_i^2 \leqq \sum_{i,j=1}^{m} p_{ij}(x)\xi_i\xi_j \leqq C\sum_{i=1}^{m} \xi_i^2$$

for all $x \in \Omega$ and all $\xi = (\xi_1, \cdots, \xi_m) \in E_m$. c and C are constants satisfying $0 < c \leqq C < \infty$.

With these assumptions holding, the system (1.20)–(1.21) has a weak solution having finite energy for every choice of $f \in L_2(Q)$ and $\varphi, \psi \in L_2(\Omega)$. This follows from Theorem 5.1 in [17].

Now let us imagine that in § 1.2 we have replaced the interval $(0, l)$ by the set Ω in E_m and that x is now an m-dimensional vector. The optimal control problem formulated in § 1.2 now makes sense for the system (1.20)–(1.21). Moreover, the statement of the necessary and sufficient optimality conditions remain the same except that now in place of the differential equations (1.1) and (1.7) we must substitute the differential equations

$$Lu = f$$

and

$$Mv = \sum_{i=1}^{n} \lambda_i \frac{\partial F_i}{\partial u}(x,t,u^0(x,t),f^0(x,t)),$$

where L and M are defined by (1.22) and (1.23) below. We must also take $v(x,t) = 0$, $x \in \partial\Omega$, and $\mu(x) \equiv 1$ in (1.7).

The reader who has understood the proofs in § 1.2 and who is familiar with the elementary theory of second order linear hyperbolic partial differential equations as discussed, for example, in [17] will have no trouble constructing the proofs of these facts for himself. Nevertheless, we shall state and indicate the proofs of the needed results. In the first place, we recall that Theorem 1.1 follows from Lemmas 1.1, 1.2 and 1.3. Lemma 1.3 is a consequence of Lemmas 1.1, 1.2 and Theorem 3.1 in [13, p. 177]. To obtain Lemma 1.2 for the system (1.20)–(1.21),

we must define

$$Lu \equiv \frac{\partial^2 u}{\partial t^2} - \sum_{i,j=1}^{m} \frac{\partial}{\partial x_i}\left(p_{ij}(x)\frac{\partial u}{\partial x_j}\right) - \sum_{i=1}^{m} r_i(x)\frac{\partial u}{\partial x_i} - q(x)u \tag{1.22}$$

and

$$Mv \equiv \frac{\partial^2 v}{\partial t^2} - \sum_{i,j=1}^{m} \frac{\partial}{\partial x_i}\left(p_{ij}(x)\frac{\partial v}{\partial x_j}\right) + \sum_{i=1}^{m} \frac{\partial}{\partial x_i}\left(r_i\frac{\partial v}{\partial x_i}\right) - q(x)v. \tag{1.23}$$

Lemma 1.2 now follows immediately from the identity (which is given by (4.10) in [17])

$$\int_0^T \int_\Omega vLu\,dx\,dt = \int_\Omega \left\{\frac{\partial u}{\partial t}(x, T)v(x, T) - \frac{\partial u}{\partial t}(x, 0)v(x, 0)\right\} dx$$
$$- \int_0^T \int_\Omega \left\{\frac{\partial u}{\partial t}\frac{\partial v}{\partial t} - \sum_{i,j=1}^{m} p_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j} + \sum_{i=1}^{m} r_i\frac{\partial u}{\partial x_i} + quv\right\} dx\,dt$$

and its analogue for

$$\int_0^T \int_\Omega uMv\,dx\,dt.$$

We turn now to a discussion of the m-dimensional version of Lemma 1.1. To this end, let $(x^1, t^1), \cdots, (x^N, t^N)$ be N arbitrary points in $\Omega \times (0, T)$ and let $f_1, \cdots, f_N$ be N arbitrary points in F. Assume that the (x^i, t^i) are labeled so that $x_1^1 \leqq x_1^2 \leqq \cdots \leqq x_1^N$, where x_i^j denotes the ith component of x^j. Choose $\delta > 0$ such that $x_1^i + N\delta < x_1^j$ if $x_1^i < x_1^j$ and such that the points $(y^j, \tau^j) \in E_{m+1}$ defined by

$$y_i^j = x_i^j + N\delta^{1/(m+1)}, \qquad \tau^j = t^j + \delta^{1/(m+1)},$$

$i = 1, \cdots, m$, $j = 1, \cdots, N$, belong to $\Omega \times (0, T)$. Let $\varepsilon_1, \cdots, \varepsilon_N$ be N real parameters satisfying $0 \leqq \varepsilon_j \leqq \delta\,(1 \leqq j \leqq N)$, and let $\hat{m} = m + 1$. Let

$$X_1 = x_1^i \quad \text{and} \quad X_j = x_1^j + \sqrt[\hat{m}]{\varepsilon_1} + \cdots + \sqrt[\hat{m}]{\varepsilon_{j-1}}$$

for $1 < j \leqq N$. The intervals $X_j \leqq x \leqq X_j + \sqrt[\hat{m}]{\varepsilon_j}$ are nonoverlapping, and therefore, the $(m + 1)$-dimensional cubes

$$R_j{:}[X_j, X_j + \sqrt[\hat{m}]{\varepsilon_j}] \times [x_2^j, x_2^j + \sqrt[\hat{m}]{\varepsilon_j}] \times \cdots \times [x_m^j, x_m^j + \sqrt[\hat{m}]{\varepsilon_j}] \times [t^j, t^j + \sqrt[\hat{m}]{\varepsilon_j}],$$

$j = 1, \cdots, N$, are nonintersecting.

Let f_ε be defined on $\Omega \times [0, T]$ according to definition (1.8). To obtain the analogue of Lemma 1.1 for the system (1.20)–(1.21), we assert the inequality

$$\int_\Omega \left\{\left(\frac{\partial \Delta u}{\partial t}(x, t)\right)^2 + \sum_{i,j=1}^{m} p_{ij}\frac{\partial \Delta u}{\partial x_i}(x, t)\frac{\partial \Delta u}{\partial x_j}(x, t) + \Delta u^2(x, t)\right\} dx$$
$$\leqq 2e^{Kt}\int_0^t \int_\Omega e^{-K\tau}\frac{\partial \Delta u}{\partial t}(x, \tau)\Delta f(x, \tau)\,dx\,d\tau \tag{1.24}$$

which is obtained from the energy inequality (8.1) in [17]. K is a constant depending only on m, c and the bounds for the coefficients of L. Δu and Δf are defined as in Lemma 1.1.

Denoting the left-hand side of (1.24) by $E(t)$, we obtain from the Cauchy–Schwarz inequality,

$$E(t) \leqq 2e^{Kt} \int_0^t e^{-K\tau} E^{1/2}(\tau) \left(\int_\Omega (\Delta f)^2 \, dx \right)^{1/2}.$$

From this, it follows that

$$E^{1/2}(t) \leqq 2 \int_0^t \left(\int_\Omega (\Delta f)^2 \, dx \right)^{1/2}$$
$$= \sum_{i=1}^{N} \varepsilon_i^{1/(m+1)} O(\varepsilon_i^{m/(2(m+1))}) = \sum_{i=1}^{N} O(\varepsilon_i^{(m+2)/(2(m+1))}).$$

Now since $(m + 2)/(2(m + 1)) > 1/2$, it follows that $E(t) = o(\varepsilon)$ uniformly in t for $0 \leqq t \leqq T$. This completes the development necessary for the extension of the results in § 1.2 to the system (1.20)–(1.21).

Example. Let the controlled process be described by the differential equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \qquad 0 \leqq x \leqq \pi, \quad 0 \leqq t \leqq \pi, \tag{1.25}$$

with the initial and boundary conditions

$$u(x, 0) = \sin x, \quad \partial u(x, 0)/\partial t = -\sin x, \quad u(0, t) = u(\pi, t) = 0. \tag{1.26}$$

Find an admissible control f^0 which minimizes the functional

$$J_0(f) = \int_0^\pi [u^2(x, \pi) + u_t^2(x, \pi)] \, dx + \int_0^\pi \int_0^\pi f^2(x, t) \, dx \, dt,$$

subject to the differential equation (1.25)–(1.26). For this example, we shall assume that the set F appearing in the definition of an admissible control is the whole real line.

To find an admissible control f^0 satisfying the condition of the maximum principle (Theorem 1.1), we look for a function which maximizes an expression of the form

$$vf + \lambda_0 f^2, \tag{1.27}$$

where $\lambda_0 \leqq 0$. We assume $\lambda_0 < 0$. Then we may assume without loss of generality that $\lambda_0 = -1$. Since f^0 maximizes (1.27), we must have

$$f^0 = -\tfrac{1}{2} v.$$

Putting this in (1.25) we are faced with the problem of solving the system

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} - \frac{1}{2} v, \qquad \frac{\partial^2 v}{\partial t^2} = \frac{\partial^2 v}{\partial x^2},$$

$$u(0, t) = u(\pi, t) = v(0, t) = v(\pi, t) = 0,$$

$$u(x, 0) = \sin x, \qquad \partial u(x, 0)/\partial t = -\sin x,$$

$$v(x, \pi) = -2u_t(x, \pi), \qquad v_t(x, \pi) = 2u(x, \pi).$$

The solution of this system is

$$v(x, t) = \frac{2}{1 + \pi/2}(\sin t + \cos t)\sin x,$$

$$u(x, t) = (\cos t - \sin t)\sin x - \frac{\sin x}{1 + \pi/2}\int_0^t \sin(t - \tau)(\sin \tau + \cos \tau)\, d\tau.$$

Hence, an optimal control is

$$f^0(x, t) = -\frac{1}{2}v(x, t) = -\frac{1}{1 + \pi/2}(\sin t + \cos t)\sin x,$$

since the maximum principle is a sufficient condition for optimality for this example.

The minimum cost is $4\pi/(2 + \pi)^2 + 2\pi^2/(2 + \pi)^2$.

2. The optimal control of a vibrating beam. In this section we discuss the optimal control of the system studied by Komkov in [3]. Thus we let u denote the transverse deflection of an inhomogeneous beam subject to an external force f. Then u and f are related by the differential equation

$$\rho(x)A(x)\frac{\partial^2 u}{\partial t^2} + \frac{\partial^2}{\partial x^2}\left(E(x)I(x)\frac{\partial^2 u}{\partial x^2}\right) = f(x, t), \qquad 0 \leqq x \leqq l, \quad 0 \leqq t. \tag{2.1}$$

l denotes the length of the beam. $\rho(x)$ is the material density, $A(x)$ the cross-sectional area, $E(x)$ Young's modulus and $I(x)$ is the moment of inertia of the cross-sectional area about the neutral axis.

We assume the boundary conditions for (2.1) to be given by

(i) $$u(0, t) = u(l, t) = 0, \qquad \frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(l, t) = 0,$$

or by

(ii) $$u(0, t) = u(l, t) = 0, \qquad EI\frac{\partial^2 u}{\partial x^2}(0, t) = EI\frac{\partial^2 u}{\partial x^2}(l, t) = 0,$$

or by

(iii) $$EI\frac{\partial^2 u}{\partial x^2}(0, t) = EI\frac{\partial^2 u}{\partial x^2}(l, t) = 0,$$

$$\frac{\partial}{\partial x}\left[EI\frac{\partial^2 u}{\partial x^2}(0, t)\right] = \frac{\partial}{\partial x}\left[EI\frac{\partial^2 u}{\partial x^2}(l, t)\right] = 0.$$

Conditions (i) correspond to a beam with built-in ends. Conditions (ii) correspond to a beam with freely supported ends. Conditions (iii) correspond to a beam with free ends.

The initial condition for our system is of the form

$$u(x,0) = \varphi(x), \qquad \frac{\partial u}{\partial t}(x,0) = \psi(x), \tag{2.2}$$

where φ and ψ are continuously differentiable functions on $[0, l]$. As in [3], we assume the coefficients ρ, A, E and I to be twice continuously differentiable and positively bounded away from zero on $[0, l]$.

2.1. Formulation of the control problem. We shall formulate the control problem for the vibrating beam analogously to the formulation given above for the problem of controlling the vibrating string.

Let functions $h_1, h_2, g_1, g_2, F_{-2}, \cdots, F_n$ be given as in § 1.2.

We wish to find an admissible control f^0 which minimizes the functional

$$\begin{aligned} J_0(f) = &\int_0^l [g_1(x, u(x,T)) + g_2(x, u_t(x,T))]\,dx \\ &+ \int_0^T \int_0^l F_0(x,t,u(x,t),f(x,t))\,dx\,dt \end{aligned} \tag{2.3}$$

subject to the differential equation (2.1) with initial conditions (2.2) and subject to any set of boundary conditions (i), (ii) or (iii) and the side constraints (1.6). $T > 0$ is fixed.

An admissible control for this problem is defined by Definition 1.1 in § 1.2.

For each admissible control f the system (2.1)–(2.2) has a weak solution u which is continuously differentiable on $[0, l] \times [0, T]$. Moreover, the energy function

$$E(t) = \frac{1}{2}\int_0^l \left\{\rho(x)A(x)\left(\frac{\partial u(x,t)}{\partial t}\right)^2 + E(x)I(x)\left(\frac{\partial^2 u}{\partial x^2}(x,t)\right)^2\right\} dx \tag{2.4}$$

is continuous in t and uniformly bounded for $0 \leqq t \leqq T$. It is easy to show that each weak solution is the limit of a sequence of classical solutions in the energy norm (2.4). This justifies certain results obtained below by formally integrating by parts.

DEFINITION 2.1. An admissible control which minimizes the functional (2.3) subject to the above constraints is said to be an *optimal control.*

THEOREM 2.1 (Maximum principle). *In order that an admissible control f^0 with response u^0 be optimal, it is necessary that there exist constants $\lambda_{-2}, \lambda_{-1}, \lambda_0, \cdots, \lambda_n$ and a solution $v(x,t)$ of*

$$\rho(x)A(x)\frac{\partial^2 v}{\partial t^2} + \frac{\partial^2}{\partial x^2}\left(E(x)I(x)\frac{\partial^2 v}{\partial x^2}\right) = \sum_{i=-2}^{n} \lambda_i \frac{\partial F_i}{\partial u}(x,t,u^0(x,t),f^0(x,t)),$$

$$v(x,T) = \frac{\lambda_0}{\rho(x)A(x)}\frac{\partial g_2}{\partial u}(x, u_t^0(x,T)) + \frac{\lambda_{-2}}{\rho(x)A(x)}\frac{\partial h_2}{\partial u}(x, u_t^0(x,T)), \tag{2.5}$$

$$\frac{\partial v}{\partial t}(x,T) = -\frac{\lambda_0}{\rho(x)A(x)}\frac{\partial g_1}{\partial u}(x, u^0(x,T)) - \frac{\lambda_{-1}}{\rho(x)A(x)}\frac{\partial h_1}{\partial u}(x, u^0(x,T))$$

such that

$$\max_{f\in F}\left[v(x,t)f+\sum_{i=-2}^{n}\lambda_i F_i(x,t,u^0(x,t),f)\right]$$
$$=v(x,t)f^0(x,t)+\sum_{i=-2}^{n}\lambda_i F_i(x,t,u^0(x,t),f^0(x,t)).$$

v satisfies the same boundary conditions as u. Not all of the λ_i are zero, and $\lambda_i \leqq 0$ for $0 \leqq i \leqq n'$.

The proof of this theorem closely parallels the proof of Theorem 1.1. We shall merely state and indicate the proofs of the necessary lemmas. The other details of the proof are left to the reader.

LEMMA 2.1. *Let the admissible control f^0 be optimal for the system* (2.1)–(2.2). *Let u^0 denote the optimal response corresponding to f^0. Let the functions f_ε and u_ε be defined as in Lemma* 1.1. *Then*

$$\int_0^l \Delta u^2(x,T)\,dx = o(\varepsilon), \qquad \int_0^l \Delta u_t^2(x,T)\,dx = o(\varepsilon)$$

and

$$\int_0^T\int_0^l \Delta u^2(x,t)\,dx\,dt = o(\varepsilon),$$

where $\Delta u = u_\varepsilon - u_0$.

Proof. Equation (14) in [3] shows that

$$\frac{1}{2}\int_0^l\left\{\rho(x)A(x)\left(\frac{\partial \Delta u}{\partial t}(x,t)\right)^2+E(x)I(x)\left(\frac{\partial^2\Delta u}{\partial x^2}(x,t)\right)^2\right\}dx$$
$$=\int_0^t\int_0^l \frac{\partial \Delta u}{\partial t}(x,t)\Delta f(x,t)\,dx\,dt.^1$$

The portion of the lemma concerning Δu_t now follows as in the proof of Lemma 1.1. To obtain the remaining portion of the lemma we observe that

$$|\Delta u(x,t)| = \left|\int_0^t \Delta u_t(x,\tau)\,d\tau\right| \leqq T^{1/2}\left(\int_0^T (\Delta u_t)^2\,d\tau\right)^{1/2}$$

and so

$$\int_0^l \Delta u^2(x,t)\,dx = o(\varepsilon)$$

uniformly in t for $0 \leqq t \leqq T$. This completes the proof of the lemma.

Let L denote the differential operator defined by

$$Lu \equiv \rho(x)A(x)\frac{\partial^2 u}{\partial t^2}+\frac{\partial^2}{\partial x^2}\left(E(x)I(x)\frac{\partial^2 u}{\partial x^2}\right).$$

[1] This derivation in [3] is purely formal, but it can be justified by using the approximating series technique that by now has become standard in this paper.

The formal adjoint of L is defined by

$$Mv \equiv \rho(x)A(x)\frac{\partial^2 v}{\partial t^2} + \frac{\partial^2}{\partial x^2}\left(E(x)I(x)\frac{\partial^2 v}{\partial x^2}\right).$$

Again we are dealing with a self-adjoint system. As before, we shall use L and M to distinguish the roles of u and v in what follows.

LEMMA 2.2. *Let u and v be two functions for which Lu and Mv are defined and are square integrable. Thus u and v are weak solutions of certain differential equations. If u and v satisfy any set of the boundary conditions* (i), (ii) *or* (iii), *then*

$$\int_0^T \int_0^l [vLu - uMv]\,dx\,dt = \int_0^l \rho(x)A(x)\left[v(x,T)\frac{\partial u}{\partial t}(x,T) - u(x,T)\frac{\partial v}{\partial t}(x,T)\right]dx + \int_0^l \rho(x)A(x)\left[u(x,0)\frac{\partial v}{\partial t}(x,0) - v(x,0)\frac{\partial u}{\partial t}(x,0)\right]dx. \tag{2.6}$$

Proof. For sufficiently smooth functions u and v, the lemma is proved by integrating the left-hand side of (2.6) by parts and making use of the boundary conditions (i), (ii) or (iii). For weak solutions u and v, the lemma is obtained by applying (2.6) to a sequence of smooth functions converging to u and v in the energy norm (2.4).

LEMMA 2.3. *Lemma* 1.3 *is valid for the problem of this section.*

There is no essential change in the proof given for Lemma 1.3. We shall, therefore, leave the proof to the reader. It is probably worthwhile to point out here that the functions v_j defined in (1.13a)–(1.13f) are now defined as follows:

$$\rho(x)A(x)\frac{\partial^2 v_j}{\partial t^2} + \frac{\partial^2}{\partial x^2}\left(E(x)I(x)\frac{\partial^2 v_j}{\partial x^2}\right) = \frac{\partial F_j}{\partial u}(x,t,u^0(x,t),f^0(x,t)), \tag{2.7a}$$

$$0 \leqq x \leqq l, \quad 0 \leqq t \leqq T, \quad -2 \leqq j \leqq n.$$

These v_j satisfy the same set of boundary conditions that have been prescribed for the function u in (2.1). The terminal conditions on the v_j are given by

$$v_{-2}(x,T) = \frac{1}{\rho(x)A(x)}\frac{\partial h_2}{\partial u}(x,u_t(x,T)), \qquad \frac{\partial v_{-2}}{\partial t}(x,T) = 0, \tag{2.7b}$$

$$v_{-1}(x,T) = 0, \qquad \frac{\partial v_{-1}}{\partial t}(x,T) = -\frac{1}{\rho(x)A(x)}\frac{\partial h_1}{\partial u}(x,u_t^0(x,T)), \tag{2.7c}$$

$$v_0(x,T) = \frac{1}{\rho(x)A(x)}\frac{\partial g_2}{\partial u}(x,u_t^0(x,T)),$$

$$\frac{\partial v_0}{\partial t}(x,T) = -\frac{1}{\rho(x)A(x)}\frac{\partial g_1}{\partial u}(x,u^0(x,T)), \tag{2.7d}$$

$$v_j(x,T) = 0, \qquad \frac{\partial v_j}{\partial t}(x,T) = 0, \qquad j = 1,\cdots,n. \tag{2.7e}$$

The proof of Theorem 2.1 can now be carried out following the pattern set forth in the proof of Theorem 1.1.

Similarly, one can prove the analogue of Theorem 1.2 for the system (2.1)–(2.2). If this is done, one should apply the remarks following Theorem 1.2 to the new theorem.

REFERENCES

[1] E. B. LEE, *A sufficient condition in the theory of optimal control*, this Journal, 1 (1963), pp. 241–245.
[2] D. L. RUSSELL, *Optimal regulation of linear symmetric hyperbolic systems with finite dimensional controls*, this Journal, 4 (1966), pp. 276–295.
[3] V. KOMKOV, *The optimal control of a transverse vibration of a beam*, this Journal, 6 (1968), pp. 401–421.
[4] L. S. PONTRYAGIN, V. BOLTYANSKII, R. GAMKRELIDZE AND E. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, L. W. Neustadt, ed., Interscience, New York, 1962.
[5] A. I. EGOROV, *On optimal control of processes in distributed objects*, J. Appl. Math. Mech., 27 (1963), pp. 1045–1058.
[6] ———, *Necessary optimality conditions for distributed parameter systems*, this Journal, 5 (1967), pp. 352–408.
[7] ———, *Optimal processes in distributed parameter systems and certain problems in invariance theory*, this Journal, 4 (1966), pp. 601–661.
[8] G. S. JONES AND A. STRAUSS, *An example of optimal control*, SIAM Rev., 10 (1968), pp. 25–55.
[9] D. W. BUSHAW, *Optimal discontinuous forcing terms*, Contributions to the Theory of Nonlinear Oscillations, vol. IV, Annals of Mathematics Studies, No. 41, Princeton Univ. Press, Princeton, 1958, pp. 29–52.
[10] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[11] J. L. LIONS, *Contrôle optimal de systems gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.
[12] G. HELLWIG, *Partial Differential Equations*, Blaisdell, New York, 1964.
[13] M. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
[14] N. S. KOSHLYAKOV, M. M. SMIRNOV AND E. B. GLIBER, *Differential Equations of Mathematical Physics*, North-Holland, Amsterdam, 1964.
[15] S. SAKS, *Theory of the Integral*, Hafner, New York, 1937.
[16] E. R. BARNES, *The optimal control of systems with distributed parameters*, Doctoral thesis, Mathematics Department, University of Maryland, College Park, 1968.
[17] C. WILCOX, *Initial-boundary value problems for linear hyperbolic partial differential equations of the second order*, Arch. Rational Mech. Anal., 10 (1962), pp. 361–400.

SIAM J. CONTROL
Vol. 9, No. 1, February 1971

CONVERGENCE OF A DISCRETIZATION FOR CONSTRAINED SPLINE FUNCTION PROBLEMS*

JAMES W. DANIEL†

1. Introduction. In [10], some generalizations of the basic ideas of spline functions were developed by considering certain minimization problems under a mixture of discrete and continuous inequality constraints, extending concepts in [1], [8], [9] and [13]. Sufficient and sometimes necessary conditions for a function to solve the minimization problem were presented via optimal control techniques, but no computational methods were discussed. In the present paper, we shall analyze the convergence of simple discretizations of the problem, such discretizations in many cases being finitely solvable by standard quadratic programming methods. Let us first define the problem.

Let m be a positive integer and let $1 < p < \infty$. For $i = 1, 2, \cdots, k$, let M_i be not identically zero linear differential operators on $[0, 1]$ of degree less than m, and similary for N_i, $i = 1, 2, \cdots, n$; we write

$$M_i x = \sum_{j=0}^{m-1} b_{ij}(t)x^{(j)}(t), \qquad N_i x = \sum_{j=0}^{m-1} c_{ij}(t)x^{(j)}(t).$$

We allow $k = 0$ and $n = 0$. Let L be a linear differential operator on $[0, 1]$ of exact degree m,

$$Lx = \sum_{j=0}^{m} a_j(t)x^{(j)}(t), \qquad a_m(t) \neq 0 \quad \text{in } [0, 1].$$

Let $W^{m,p}$ be the (Sobolev) space of real-valued runctions x on $[0, 1]$ such that $x^{(m-1)}$ is absolutely continuous and $x^{(m)} \in L^p(0, 1)$. Then our minimization problem is to

$$\text{minimize} \quad f(x) \equiv \int_0^1 |Lx(t)|^p \, dt \quad \text{over}$$

$$C = \{x; x \in W^{m,p}, \alpha_i(t) \leqq M_i x(t) \leqq \beta_i(t) \text{ for } 0 \leqq t \leqq 1, \tag{1.1}$$
$$i = 1, \cdots, k, \gamma_i \leqq N_i x(\xi_i) \leqq \delta_i \text{ for } i = 1, \cdots, n\},$$

where α_i and β_i are given functions, the γ_i and δ_i are given scalars, and the ξ_i are points in $[0, 1]$. Some simple generalizations are possible by allowing one-sided constraints or by allowing the N_i to be difference operators, but we shall not consider these here. It is shown in [10] that, if C is nonempty, if $a_j \in C^j[0, 1]$ for $j = 0, \cdots, m$, if $b_{ij} \in C[0, 1]$ for $i = 1, \cdots, k$ and $j = 0, \cdots, m - 1$, if $c_{ij} \in C[0, 1]$ for $i = 1, \cdots, n$ and $j = 0, \cdots, m - 1$, and if α_i and β_i lie in $M_i W^{m,p}$ for $i = 1, \cdots, k$, then there exists a solution x^* to the problem in (1.1). *We shall assume the above hypotheses to hold throughout the ensuing discussion.* Since the solution x^*

* Received by the editors March 20, 1970, and in revised form June 26, 1970.

† Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706. This work was supported by the Office of Naval Research under Contract N00014-67-A-0128-0004.

may hit the boundary of C at unknown points, perhaps countably many times, the computation of x^* is difficult. One obvious way to handle this is as follows.

Let $h > 0$ be some mesh size, say $h = 1/Q$, and let $[0, 1]$ be partitioned by $t_i = ih$, $i = 0, \cdots, 1/h = Q$; we suppose that all the points ξ_i lie on this mesh for all h to be used, that is, ξ_i/h is an integer (this assumption is not necessary but simplifies the notation). Our first discretization consists in merely replacing the continuous constraints by discrete ones, that is, we

$$\text{minimize} \quad f(x) = \int_0^1 |Lx(t)|^p \, dt \quad \text{over}$$

$$C_1(h) \equiv \{x; x \in W^{m,p}, \alpha_i(t_j) \leqq M_i x(t_j) \leqq \beta_i(t_j) \text{ for } j = 0, \cdots, Q, i = 1, \cdots, k, \gamma_i \leqq N_i x(\xi_i) \leqq \delta_i \text{ for } i = 1, \cdots, n\}. \tag{1.2}$$

As analyzed in [13], this problem can be solved in the common case of $p = 2$ in finitely many steps by minimizing a quadratic function of $2k(Q + 1) + m + 2n$ variables subject to $2k(Q + 1) + m + 2n$ linear inequality constraints. We shall prove the following (§ 2, Theorem 2.1): *All weak limit points (in the $W^{m,p}$ sense), at least one of which exists, of a sequence of solutions to the first discretization in* (1.2) *must solve the original problem of* (1.1); *if the solution to the original problem is unique, the approximating solutions converge to it $W^{m,p}$-weakly, and in particular the function and the first $m - 1$ derivatives converge uniformly, that is, in $C[0, 1]$.*

If one must take h very small to obtain a reasonable approximation to x^*, one might well be satisfied to have only approximate values of x^* at the grid points t_i rather than throughout $[0, 1]$; if $x^*(t)$ were desired and $x^*(t_i)$ was accurately known, unconstrained interpolation could be used to generate a reasonable approximation to x^*. Thus we are led to a second, more complete, discretization. If z is a function defined at least on the mesh points $t_i = ih$, $i = 0, \cdots, 1/h = Q$, let $D = D_h$ be the operator such that $Dz(t_i) = (z(t_{i+1}) - z(t_i))/h$, $i = 0, 1, \cdots, Q - 1$. We then have

$$D^l z(t_i) = \frac{1}{h^l} \sum_{j=0}^{l} (-1)^{l-j} \binom{l}{j} z(t_{i+j})$$

as a natural analogue of the lth derivative of z. We therefore define

$$M_{i,h} z(t_j) = \sum_{l=0}^{m-1} b_{il}(t_j) D^l z(t_j), \qquad 0 \leqq j \leqq Q - m + 1,$$

$$N_{i,h} z(t_j) = \sum_{l=0}^{m-1} c_{il}(t_j) D^l z(t_j), \qquad 0 \leqq j \leqq Q - m + 1,$$

$$L_h z(t_j) = \sum_{l=0}^{m} a_l(t_j) D^l z(t_j), \qquad 0 \leqq j \leqq Q - m.$$

Our second discretized problem is now to

$$
\begin{aligned}
\text{minimize} \quad & f_h(x_h) = h \sum_{j=0}^{Q-m} |L_h x_h(t_j)|^p \quad \text{over} \\
& C_2(h) \equiv \{x_h; x_h \equiv (x_h(0), x_h(t_1), \cdots, x_h(1))^T \in R^{Q+1}, \\
& \quad -\varepsilon_h + \alpha_i(t_j) \leqq M_{i,h} x_h(t_j) \leqq \beta_i(t_j) + \varepsilon_h \\
& \quad \text{for } j = 0, \cdots, Q - m + 1, \; i = 1, \cdots, k, \\
& \quad -\varepsilon_h + \gamma_i \leqq N_{i,h} x_h(\xi_i) \leqq \delta_i + \varepsilon_h \text{ for } i = 1, \cdots, n\}.
\end{aligned} \tag{1.3}
$$

Here ε_h, which tends to zero, gives a small expansion of the constraint set as $h \to 0$. If some such expansion is not allowed, the set $C_2(h)$ can be empty [6], [7]; as will be clear from the use made of the expansion by ε_h, constraints of the form $\alpha(t) \leqq x(t) \leqq \beta(t)$ or $\alpha \leqq x(\xi) \leqq \beta$ *need not be expanded.* This problem in the common case of $p = 2$ can be solved in finitely many steps since it involves a quadratic function of $Q + 1$ variables subject to $2k(Q - m + 2) + 2n$ linear inequalities. Since in general $Q = 1/h$ is large, seeking only approximations to $x^*(t_i)$ reduces the difficulty considerably. Under slight additional hypotheses on the problem, we shall prove the following (roughly stated) (§ 3, Theorem 3.1): *All $W^{m,p}$ weak limit points (of certain "interpolations"), at least one of which exists, of a sequence of solutions to the discretization in* (1.3) *must solve the original problem of* (1.1); *if the solution to the original problem is unique, the ("interpolations" of the) approximate solutions converge to it $W^{m,p}$-weakly, and in particular the function values and the first $m - 1$ divided differences converge uniformly, that is, in $C[0, 1]$, to x^* and its first $m - 1$ derivatives.*

In our analysis of these discretizations, we shall follow the general approach of [6] and [7] for studying discretized optimization problems; in particular, we use the approach of [6] and [7] for studying discretized optimal control problems, since our problem can easily be stated in that formalism, as in our (2.6). Once we use the control theory formalism, however, other approaches for analyzing discretizations of control problems can be used; for example, the ideas of Cullum [2] and [3]. Our present problem has such special structure however (for example, the control variable is unconstrained but the state variable is subject to constraints) that the more general results give us too little information; we must make use of the problem's structure.

2. Analysis of the first, simpler, discretization. We have yet to define a norm on the space $W^{m,p}$; two common norms are

$$
\|x\|_0 \equiv \left\{ \sum_{i=0}^{m} \int_0^1 |x^{(i)}(t)|^p \, dt \right\}^{1/p},
$$

$$
\|x\|_1 \equiv \left\{ \sum_{i=1}^{m} |x(\theta_i)|^p + \int_0^1 |x^{(m)}(t)|^p \, dt \right\}^{1/p}
$$

for $0 \leqq \theta_1 < \theta_2 < \cdots < \theta_m \leqq 1$. It is well known [14], [12] that these norms are equivalent; that is, there exist positive constants a, A such that $a\|x\|_0 \leqq \|x\|_1 \leqq A\|x\|_0$ for all x in $W^{m,p}$. This equivalence is related to the famous Sobolev

inequalities which say that, for $0 \leqq i \leqq m-1$, there exist constants A_i such that $|x^{(i)}(t)| \leqq A_i \|x\|_1$ for all t in $[0,1]$ and x in $W^{m,p}$. If a sequence $\{x_n\}$ converges weakly in $(W^{m,p}, \|\cdot\|_1)$ to x, then $\|x_n\|_1$ is uniformly bounded and therefore, by the Sobolev inequalities, for $0 \leqq i \leqq m-1$, the functions $x_n^{(i)}$ are uniformly bounded in $C[0,1]$. Hence the $\{x_n^{(i)}\}$ are equicontinuous for $0 \leqq i \leqq m-2$ which implies that $\{x_n^{(i)}\}$ converges uniformly, that is, in $C[0,1]$, to $x^{(i)}$ for $0 \leqq i \leqq m-2$. A slightly more subtle argument, using

$$|x^{(m-1)}(t) - x^{(m-1)}(t')| \leqq \left| \int_t^{t'} |x^{(m)}(\tau)|\, d\tau \right| \leqq \left[\int_0^1 |x^{(m)}(\tau)|^p\, d\tau \right]^{1/p} |t' - t|^{1-1/p},$$

shows as well that $\{x_n^{(m-1)}\}$ converges uniformly to $x^{(m-1)}$. For $1 < p < \infty$, $W^{m,p}$ is reflexive and hence the unit sphere is weakly compact and weakly sequentially compact.

From the computational standpoint, serious difficulties arise if the original problem (1.1) admits solutions of arbitrarily large norm. For example, the functions $x_n(t) = n$ form a minimizing sequence (in fact, they are all solutions) for $\int_0^1 |x^{(2)}(t)|^2\, dt$ over the set of x satisfying $0 \leqq x^{(1)}(t) \leqq 1$ but have no convergent subsequence. In this situation, our analysis to follow could not guarantee that the approximate solutions have limit points; to avoid this we must eliminate problems admitting solutions of arbitrarily large norm. We pause to see what this means. For $0 \leqq l \leqq k$, let $S_l \equiv \{x; x \in W^{m,p}, Lx \equiv 0, M_i x \equiv 0 \text{ for } 1 \leqq i \leqq l, \text{ and } N_i x(\xi_i) = 0 \text{ for } 1 \leqq i \leqq n\}$. It is shown in [10] that, if d_{k+1} is the dimension of S_k and if d_{k+1-j} is the dimension of $S_{k+1-j} - S_{k+2-j}$ for $1 \leqq j \leqq k$, then there exist points $\theta_{l,i}$ with $0 \leqq \theta_{l,1} < \theta_{l,2} < \cdots < \theta_{l,d_l} \leqq 1$ for $1 \leqq l \leqq k+1$, with the points $\{\theta_{k+1,i}\}_{i=1}^{d_{k+1}}$ being completely arbitrary in $[0,1]$, such that

$$\|x\| \equiv \left\{ \sum_{l=1}^{k} \sum_{j=1}^{d_l} |M_l x(\theta_{l,j})|^p + \sum_{j=1}^{d_{k+1}} |x(\theta_{k+1,j})|^p + \sum_{i=1}^{n} |N_i x(\xi_i)|^p + \int_0^1 |Lx(t)|^p\, dt \right\}^{1/p} \tag{2.1}$$

defines a norm on $W^{m,p}$. We remark that if for some l_0 one has $M_{l_0} x \equiv x$, then one may take $d_{l_0} = m$, all other $d_l = 0$, eliminate the sum in N_i from (2.1), and take arbitrary distinct points for $\theta_{l_0,i}$ to define the norm. By using the Sobolev inequalities, it is simple to show that this norm is in fact equivalent to $\|\cdot\|_0$ and $\|\cdot\|_1$.

LEMMA 2.1. *$\|\cdot\|$ is equivalent to $\|\cdot\|_0$ and $\|\cdot\|_1$.*

Proof. The existence of an A' such that $\|x\| \leqq A' \|x\|_1$ for all x in $W^{m,p}$ is clear from the Sobolev inequalities; we ask whether or not an $a' > 0$ exists such that $\|x\| \geqq a' \|x\|_1$. If not, we can find $x_n \in W^{m,p}$ such that $\|x_n\| \to 0$ but $\|x_n\|_1 = 1$; by the weak sequential compactness of the sphere in $(W^{m,p}, \|\cdot\|_1)$, we may assume x_n converges $(W^{m,p}, \|\cdot\|_1)$-weakly to some x in $W^{m,p}$. Since L is bounded from $(W^{m,p}, \|\cdot\|_1)$ into $L^p(0,1)$, since M_i and N_i are bounded from $(W^{m,p}, \|\cdot\|_1)$ into $C[0,1]$, and since $\int_0^1 |Lx_n(t)|^p\, dt$ tends to zero, we have $\|x\| = 0$ and therefore $x \equiv 0$. Since x_n converges uniformly, that is, in $C[0,1]$, to x and $\|x_n\|_1 = 1$, we have

$\int_0^1 |x_n^{(m)}(t)|^p\,dt$ converging to 1. Then

$$\left\{\int_0^1 |Lx_n(t)|^p\,dt\right\}^{1/p} \geqq \left\{\int_0^1 |a_m(t)x_n^{(m)}(t)|^p\,dt\right\}^{1/p} - \left\{\int_0^1 \left|\sum_{i=0}^{m-1} a_i(t)x_n^{(i)}(t)\right|^p dt\right\}^{1/p},$$

which then is bounded away from zero since $|a_m(t)| \geqq \varepsilon > 0$ for some ε, since $\int_0^1 |x_n^{(m)}(t)|^p\,dt \to 1$, and since $x_n^{(i)}$ converges uniformly to $x^{(i)} \equiv 0$ for $0 \leqq i \leqq m-1$. This contradicts $\|x_n\| \to 0$. This completes the proof.

Now for any x in C, the values $|M_l x(\theta_{l,i})|$ and $|N_i x(\xi_i)|$ are bounded, uniformly in i, l and x. Given any $\bar{x}$ in C, we would seek x^* only from among those x in C satisfying $f(x) \leqq f(\bar{x})$, which in turn gives an upper bound for $\int_0^1 |Lx(t)|^p\,dt$. From this then, we can deduce an a priori bound on $\|x^*\|$ if and only if the $\{|x(\theta_{k+1,j})|\}$ are uniformly bounded for $1 \leqq j \leqq d_{k+1}$. If $d_{k+1} = 0$, this is certainly true; if $d_{k+1} \neq 0$, there exists a nonzero function z in S_k and thus $x^* + \alpha z \in C$ for all scalars α and $f(x^* + \alpha z) = f(x^*)$. Therefore the original problem admits of an a priori bound on its solutions if and only if $d_{k+1} = 0$. *We hereafter assume that* $d_{k+1} = 0$. Computationally this may be accomplished simply by adding one continuous constraint $|x(t)| \leqq E$ for some large E or m discrete constraints $|x(\theta_i)| \leqq E$ and thus this is not a computationally significant restriction. Adding $|x(t)| \leqq E$ means that we may use as our norm the simple expression

$$\left\{\sum_{i=1}^{m} |x(\theta_i)|^p + \int_0^1 |Lx(t)|^p\,dt\right\}^{1/p}. \tag{2.2}$$

We hereafter assume that the points $\theta_{l,i}$ *of* (2.1) *are mesh points for all h used. If some* $M_{l,0}x \equiv x$, *then we need only assume that the m points* θ_i *of* (2.2) *are mesh points for all h used.*

Now let, for $Q = 1/h$, x_Q^* be a solution to the simple discretized problem of (1.2); $C_1(h)$ is not empty since, in particular, $x^* \in C_1(h)$, where x^* solves the original problem. Since $f(x_Q^*) \leqq f(x^*)$, since $|M_l x_Q^*(\theta_{l,j})| \leqq \max\{|\alpha_l(\theta_{l,j})|, |\beta_l(\theta_{l,j})|\}$ and since $|N_i x_Q^*(\xi_i)| \leqq \max\{|\delta_i|, |\gamma_i|\}$, we conclude that $\|x_Q^*\|$ is uniformly bounded. We note that x_Q^* is not necessarily in C; it is however "near" to C as the following more general lemma demonstrates.

LEMMA 2.2. *If* $x \in W^{m,p}$ *and* $-\varepsilon + \alpha_i(t_j) \leqq M_i x(t_j) \leqq \beta_i(t_j) + \varepsilon$ *for* $i = 1, \cdots, k$ *and* $j = 0, \cdots, Q$, *then* $-(\varepsilon + \eta_i(h)) + \alpha_i(t) \leqq M_i x(t) \leqq \beta_i(t) + (\varepsilon + \eta_i(h))$ *for* $0 \leqq t \leqq 1$, *where*

$$\eta_i(h) = \max\left[\left\{\int_0^1 |\beta_i^{(1)}(t)|^p\,dt\right\}^{1/p}, \left\{\int_0^1 |\alpha_i^{(1)}(t)|^p\,dt\right\}^{1/p}\right] h^{1/q} + B_i\|x\|_0 h^{1/q} + \|x\|_0 w_i(h),$$

where $1/p + 1/q = 1$, $|b_{il}(t)| \leqq B_i$, $|b_{il}(t_1) - b_{il}(t_2)| \leqq w_i(h)$ *if* $|t_1 - t_2| \leqq h$, $l = 0, \cdots, m-1$.

Proof. For the upper bound,

$$M_i x(t) - \beta_i(t) = M_i x(t) - M_i x(t_j) + M_i x(t_j) - \beta_i(t_j) + \beta_i(t_j) - \beta_i(t).$$

Since

$$\beta_i \in M_i W^{m,p} \subset W^{1,p},$$

$$|\beta_i(t) - \beta_i(t_j)| \leqq \int_{t_j}^{t} |\beta_i^{(1)}(t)|\, dt \leqq \left\{ \int_0^1 |\beta_i^{(1)}(t)|^p\, dt \right\}^{1/p} |t - t_j|^{1/q}.$$

Also

$$|M_i x(t) - M_i x(t_j)| \leqq \sum_{l=0}^{m-1} |b_{il}(t)x^{(l)}(t) - b_{il}(t_j)x^{(l)}(t_j)|$$

$$\leqq B_i \|x\|_0 |t - t_j|^{1/q} + \|x\|_0 w_i(|t - t_j|)$$

arguing as for β_i. Thus, letting t_j be such that $|t - t_j| \leqq h$, we have

$$M_i x(t) - \beta_i(t) \leqq \varepsilon + \left[\left\{ \int_0^1 |\beta_i^{(1)}(t)|^p\, dt \right\}^{1/p} + B_i \|x\|_0 \right] h^{1/q} + \|x\|_0 w_i(h).$$

Similarly for the lower bound. This completes the proof.

We note that if the b_{il} are Hölder continuous with exponent greater than or equal to $1/q$ (that is, if $|b_{il}(t_1) - b_{il}(t_2)| \leqq c|t_1 - t_2|^r$ for some constant $c \geqq 0$ and $r \geqq 1/q$) then $|\eta_i(h)| \leqq Fh^q$ for a constant F uniformly bounded whenever $\|x\|$ is bounded.

LEMMA 2.3. *If functions $x_Q \in W^{m,p}$ satisfy $-\varepsilon_Q + \alpha_i(t) \leqq M_i x_Q(t) \leqq \beta_i(t) + \varepsilon_Q$ for $i = 1, \cdots, k$ and $-\varepsilon_Q + \gamma_i \leqq N_i x_Q(\xi_i) \leqq \delta_i + \varepsilon_Q$ for $i = 1, \cdots, n$, if $\lim_{Q\to\infty} \varepsilon_Q = 0$ and if x_Q converges $W^{m,p}$-weakly to x, then $x \in C$.*

Proof. This is obvious since $x_Q^{(j)}(t)$ converges to $x^{(j)}(t)$ for each t in $[0, 1]$ for $0 \leqq j \leqq m - 1$, and the constraints involve derivatives of order at most $m - 1$.

LEMMA 2.4. *As $\varepsilon_Q \geqq 0$ tends to zero, the minimum of f over the set $C_{\varepsilon_Q} = \{x;$ $x \in W^{m,p}$, $-\varepsilon_Q + \alpha_i(t) \leqq M_i x(t) \leqq \beta_i(t) + \varepsilon_Q$ for $1 \leqq i \leqq k$ and $0 \leqq t \leqq 1$, $-\varepsilon_Q + \gamma_i \leqq N_i x(\xi_i) \leqq \delta_i + \varepsilon_Q$ for $1 \leqq i \leqq n\}$ converges to the minimum of f over C.*

Proof. Clearly each set C_{ε_Q} is weakly closed, and for the minimization problem over C_{ε_Q} we may restrict ourselves to those x satisfying $f(x) \leqq f(x^*)$, since $x^* \in C_{\varepsilon_Q}$, where x^* minimizes f over C. Since, for all x in this set, $\|x\|$ is uniformly bounded, the weakly lower semicontinuous functional f attains its minimum over the weakly compact set C_{ε_Q} at some point x_Q. Since $\|x_Q\|$ is uniformly bounded, we may assume that x_Q converges weakly to some x, which must be in C by Lemma 2. Thus $f(x) \leqq \liminf_{Q\to\infty} f(x_Q) \leqq \limsup f(x_Q) \leqq f(x^*)$ since $f(x_Q) \leqq f(x^*)$ for all Q. Thus $\lim_{Q\to\infty} f(x_Q) = f(x^*)$.

We can now prove our discretization result for the simpler discretization.

THEOREM 2.1. *Let the general assumptions of § 1 hold and let the fixed points $\theta_{l,i}$ defining the norm $\|\cdot\|$ in (2.1) (or θ_i in (2.2)) be mesh points in our discretization for all h. Let $x_Q^* \in W^{m,p}$ solve the problem of (1.2), that is, minimize $f(x)$ over $C_1(h)$. Then $f(x_Q^*)$ converges to $f(x^*)$ and all $W^{m,p}$ weak limit points, at least one of which exists, of $\{x_Q^*\}$ minimize f over C; if x^* minimizing f over C is unique, then x_Q^* converges weakly to x^*, that is, $x_Q^{*(i)}$ converges uniformly to $x^{*(i)}$ for $0 \leqq i \leqq m - 1$ and $x_Q^{*(m)}$ converges L^p-weakly to $x^{*(m)}$.*

Proof. Arguing as in Lemma 2.4, we see that x_Q^* always exists, $f(x_Q^*) \leqq f(x^*)$ and there exists a constant E such that $\|x_Q^*\| \leqq E$. Thus, by Lemma 2.2, there exist functions $\eta_i(h)$ for $1 \leqq i \leqq k$ tending to zero with h, and such that $x_Q^* \in C_Q \equiv \{x;\ x \in W^{m,p},\ -\eta_i(h) + \alpha_i(t) \leqq M_i x(t) \leqq \beta_i(t) + \eta_i(h)$ for $1 \leqq i \leqq k,\ \gamma_i \leqq N_i x(\xi_i) \leqq \delta_i$ for $1 \leqq i \leqq n\}$. By Lemma 2.4, $\zeta_Q = \min_C f - \min_{C_Q} f$ tends to zero. We write

$$f(x^*) = \min_C f = \min_{C_Q} f + \zeta_Q \leqq f(x_Q^*) + \zeta_Q \leqq f(x^*) + \zeta_Q. \tag{2.3}$$

This implies, since $\zeta_Q \to 0$, that $f(x_Q^*) \to f(x^*)$. Since $\{x_Q^*\}$ is bounded, it has weak limit points. For any such weak limit point x' with $x_{Q'}^*$ weakly converging to x', we have $x' \in C$ by Lemma 2.3 and thus

$$f(x^*) \leqq f(x') \leqq \liminf_{Q' \to \infty} f(x_{Q'}^*) = \lim_{Q \to \infty} f(x_Q^*) = f(x^*).$$

This says that $f(x') = f(x^*)$, that is, x' minimizes f over C. The remainder follows from the definition of convergence in $W^{m,p}$.

We have not been able to estimate the rate of convergence as a function of h.

3. Analysis of the more complete discretization. We shall here use the norm $\|\cdot\|$ defined in (2.1) via points $\theta_{l,i}$ (or θ_i in (2.2)) and we shall assume that the $\theta_{l,i}$ are mesh points for all h used. We shall analyze our complete discretization, the relationship between the problems in (1.1) and (1.3), by the general discretization analysis of [5], [6] and [7]; for completeness our arguments are self-contained.

We wish to use roughly the arguments of Theorem 2.1 in this case also. If x_h^* minimizes f_h over $C_2(h)$, we unfortunately cannot talk about $f(x_h^*)$ or $f_h(x^*)$ as in (2.3) in the proof of Theorem 2.1, since these make no sense in our new situation. Instead, with x^* we shall associate a point $y_h = r_h x^* \in C_2(h)$ by a "discretization" or "restriction" mapping r_h such that $|f_h(r_h x^*) - f(x^*)|$ converges to zero with h. Similarly, with x_h^* we shall associate a $z_h = p_h x_h^* \in W^{m,p}$ "converging into C" by an "interpolation" or "prolongation" mapping p_h and such that $|f(p_h x_h^*) - f_h(x_h^*)|$ converges to zero with h. We can then imitate the proof of Theorem 2.1 by replacing (2.3) by roughly

$$\begin{aligned} f(x^*) &= \min_{C_Q} f + \zeta_Q \leqq f(p_h x_h^*) + \zeta_Q \\ &= f_h(x_h^*) + \zeta_Q + [f(p_h x_h^*) - f_h(x_h^*)] \\ &\leqq f_h(r_h x^*) + \zeta_Q + [f(p_h x_h^*) - f_h(x_h^*)] \\ &= f(x^*) + \zeta_Q + [f(p_h x_h^*) - f_h(x_h^*)] + [f_h(r_h x^*) - f(x^*)]. \end{aligned}$$

Having outlined our approach and the reasons for constructing certain mappings p_h and r_h, we now proceed with the technical details. We define the restriction r_h in the obvious manner. Let $y_h = r_h x^*$ be the discrete mesh function (that is, defined at points $t_i = ih$ only) defined by $y_h(t_i) \equiv x^*(t_i)$.

We need to develop some tools for using divided differences. For any l with $0 \leqq l \leqq m$, by Peano's theorem we can write

$$D^l x^*(t) = \frac{1}{(l-1)!} \int_0^1 D_t^l (t - \tau)_+^{l-1} x^{*(l)}(\tau)\, d\tau,$$

where the D_t indicates differences with respect to t, that is, $D_t g(t, \tau) = (g(t + h, \tau) - g(t, \tau))/h$, and where

$$(t - \tau)_+^{l-1} = \begin{cases} (t - \tau)^{l-1} & \text{for } t - \tau \geqq 0, \\ 0 & \text{for } t - \tau \leqq 0 \end{cases}$$

as usual. $K_l(\tau - t) \equiv (1/(l-1)!)D_t^l(t - \tau)_+^{l-1}$ is a "basic spline" [4], that is,

$$K_l(s) \equiv \frac{1}{h^l(l-1)!} \sum_{i=0}^{l} (-1)^{l-i} \binom{l}{i} (ih - s)_+^{l-1}$$

vanishes identically for $s \geqq lh$ and $s \leqq 0$, is strictly positive for s in $(0, lh)$ and lies in $C^{l-2}(-\infty, \infty)$. Thus $D^l t^l \equiv 1! = \int_0^{lh} K_l(s)\, ds$, and we find

$$\int_0^{lh} K_l(s)\, ds = 1.$$

Also

$$K_l(s) = \frac{1}{h^l(l-1)!} \sum_{i=0}^{l} (-1)^{l-i} \binom{l}{i} \left(i - \frac{s}{h}\right)_+^{l-1} h^{l-1} \leqq \frac{G}{h}$$

for some fixed G, since we have $0 \leqq s/h \leqq l$.

We now show that $r_h x^* \in C_2(h)$ for large enough ε_h.

LEMMA 3.1. *$-\varepsilon_h + \alpha_i(t_j) \leqq M_{i,h} y_h(t_j) \leqq \beta_i(t_j) + \varepsilon_h$ for $0 \leqq j \leqq Q - m + 1$ and $1 \leqq i \leqq k$, and $-\varepsilon_h + \gamma_i \leqq N_{i,h} y_h(\xi_i) \leqq \delta_i + \varepsilon_h$ for $1 \leqq i \leqq n$, where $\varepsilon_h = Gh^{1/q}$ and $G \geqq \|x^*\|_0 \sum_{l=0}^{m-1} \|b_{il}\|_\infty l^{1/q}$ for $1 \leqq i \leqq k$, and $G \geqq \|x^*\|_0 \sum_{l=0}^{m-1} \|c_i l\|_\infty l^{1/q}$ for $1 \leqq i \leqq n$.*

Proof. For $0 \leqq l \leqq m - 1$,

$$\begin{aligned} |D^l y_h(t_j) - x^{*(l)}(t_j)| &= |D^l x^*(t_j) - x^{*(l)}(t_j)| \\ &= \left| \int_{t_j}^{t_j + lh} K_l(\tau - t_j)[x^{*(l)}(\tau) - x^{*(l)}(t_j)]\, dt \right| \\ &\leqq \sup_{t_j \leqq \tau \leqq t_j + lh} |x^{*(l)}(\tau) - x^{*(l)}(t_j)|. \end{aligned}$$

Now

$$|x^{*(l)}(\tau) - x^{*(l)}(t_j)| \leqq \int_{t_j}^{\tau} |x^{*(l+1)}(s)|\, ds \leqq \|x^*\|_0 |\tau - t_j|^{1/q} \leqq \|x^*\|_0 |lh|^{1/q}.$$

More generally,

$$\begin{aligned} |M_{i,h} y_h(t_j) - M_i x^*(t_j)| &= |M_{i,h} x^*(t_j) - M_i x^*(t_j)| \\ &\leqq \sum_{l=0}^{m-1} |b_{il}(t_j)|\, |D^l x^*(t_j) - x^{*(l)}(t_j)| \\ &\leqq \sum_{l=0}^{m-1} |b_{il}(t_j)|\, \|x^*\|_0 |lh|^{1/q} \end{aligned}$$

and similarly for N_i and $N_{i,h}$. The lemma follows since $\alpha_i(t_j) \leqq M_i x^*(t_j) \leqq \beta_i(t_j)$ and similarly for N_i.

As our last step in treating r_h, we show that $|f_h(r_h x^*) - f(x^*)|$ converges to zero.

LEMMA 3.2. $\lim_{h\to 0} |f_h(r_h x^*) - f(x^*)| = 0$.

Proof. Since the m-times continuously differentiable functions are dense in $W^{m,p}$, we can find such a function z arbitrarily near x^* and such that $|f(x^*) - f(z)|$ is arbitrarily small. Since, for $0 \leqq l \leqq m$,

$$\begin{aligned}|D^l x^*(t_j) - D^l z(t_j)| &\leqq \int_{t_j}^{t_j+lh} K_l(\tau - t_j)|x^{*(l)}(\tau) - z^{(l)}(\tau)|\,d\tau \\ &\leqq \left\{\int_{t_j}^{t_j+lh} K_l(\tau - t_j)\,d\tau\right\}^{1/q} \\ &\qquad \cdot \left\{\int_{t_j}^{t_j+lh} K_l(\tau - t_j)|x^{*(l)}(\tau) - z^{(l)}(\tau)|^p d\tau\right\}^{1/p} \\ &\leqq \left\{Gl\int_0^1 |x^{*(l)}(\tau) - z^{(l)}(\tau)|^p\,d\tau\right\}^{1/p}\end{aligned}$$

which is arbitrarily small, and since the functions a_l are bounded, it is also clear that $|f_h(r_h x^*) - f_h(r_h z)|$ can be made arbitrarily small independent of h by choosing z near x^*. Thus we are through if, after fixing z, we can show that $|f(z) - f_h(r_h z)|$ tends to zero. By using the triangle inequality, we immediately find

$$\begin{aligned}|f_h(r_h z)^{1/p} - f(z)^{1/p}|^p \leqq &\sum_{i=0}^{Q-m}\int_{ih}^{ih+h}\left|\sum_{l=0}^{m-1}[a_l(t_i)D^l z(t_i) - a_l(t)z^{(l)}(t)]\right|^p dt \\ &+ \int_{1-mh}^{1}\left|\sum_{l=0}^{m-1} a_l(t)z^{(l)}(t)\right|^p dt,\end{aligned}$$

the latter term of which clearly tends to zero with h for fixed z. For the former term, since $z^{(l)}$ is continuous, we have $D^l z(t_i) = z^{(l)}(\lambda_i)$ for some λ_i in $(t_i, t_i + lh)$. Then the former term equals

$$\sum_{i=0}^{Q-m}\int_{ih}^{ih+h}\left|\sum_{l=0}^{m-1}[a_l(t_i)z^{(l)}(\lambda_i) - a_l(t)z^{(l)}(t)]\right|^p dt.$$

Since a_l and $z^{(l)}$ are both continuous, the term under the integral sign for this fixed z is bounded by some function $w(h)$ tending to zero with h and thus the whole expression is bounded by $\sum_{i=0}^{Q-m}\int_{ih}^{ih+h} w(h)\,dt \leqq w(h)$.

Remark. The preceding lemma is of some independent interest. As a special case, it says that $h\sum_{i=0}^{Q-m}|D^m x(t_i)|^p$ converges to $\int_0^1 |x^{(m)}(t)|^p\,dt$ for all x in $W^{m,p}$; since the sum looks something like a Riemann sum for the integral, it is interesting that convergence can be proved. This is vital for the work in this paper since [10] did not give broad necessary continuity conditions for x^*; we know of examples in which $x^{*(m)}$ has countably infinitely many finite jumps although the constraining functions are very smooth.

Next, we must consider a mapping p_h of x_h^* into $p_h x_h^* = z_h \in W^{m,p}$ near C with $|f(p_h x_h^*) - f_h(x_h^*)|$ converging to zero. Let v_h be an m–vector function on the mesh

points, $v_h = (v_{h,0}, \cdots, v_{h,m-1})^T$, solving

$$\begin{aligned} v_{h,0}(t_{j+1}) &= v_{h,0}(t_j) + hv_{h,1}(t_j), \\ &\vdots \\ v_{h,m-2}(t_{j+1}) &= v_{h,m-2}(t_j) + hv_{h,m-1}(t_j), \\ v_{h,m-1}(t_{j+1}) &= v_{h,m-1}(t_j) + h\left[-\sum_{i=0}^{m-1}(a_i(t_j)v_{h,i}(t_j)) + L_h x_h^*(t_j)\right] \end{aligned} \tag{3.1}$$

for $0 \leqq j \leqq Q - m$, with $v_{h,i}(0) = D^i x_h^*(0)$ for $0 \leqq i \leqq m - 1$. For convenience we have assumed $a_m(t) \equiv 1$ without loss of generality. Clearly then, we have that

$$v_{h,i}(t_j) = D^i x_h^*(t_j) \quad \text{for } 0 \leqq i \leqq m - 1 \quad \text{and} \quad 0 \leqq j \leqq Q - m.$$

Consider the m–vector function V_h on $[0, 1]$, $V_h = (V_{h,0}, \cdots, V_{h,m-1})^T$, solving the system of differential equations

$$\begin{aligned} V_{h,0}^{(1)} &= V_{h,1}, \\ &\vdots \\ V_{h,m-2}^{(1)} &= V_{h,m-1}, \\ V_{h,m-1}^{(1)} &= -\sum_{i=0}^{m-1} a_i V_{h,i} + u_h \end{aligned} \tag{3.2}$$

with $V_{h,i}(0) = v_{h,i}(0)$ for $i = 0, \cdots, m - 1$, and $u_h(t) = L_h x_h^*(t_i)$ for $t_i \leqq t < t_{i+1}$ and $0 \leqq i \leqq Q - m$, $u_h(t) = 0$ for $t \geqq 1 - (m - 1)h$. We see immediately that the v_h is obtained by applying Euler's method to solve the system in (3.2) which, for convenience, we write as

$$V_h^{(1)} = AV_h + eu_h, \qquad V_h(0) = v_h(0), \tag{3.3}$$

where A is the obvious matrix and $e = (0, 0, \cdots, 0, 1)^T$.

We now define $z_h = p_h x_h^* \equiv V_{h,0}$. We notice that $z_h = V_{h,0}$ solves the equation $Lz_h = u_h$ and thus

$$\int_0^1 |Lz_h(t)|^p \, dt = \int_0^1 |u_h(t)|^p \, dt = h \sum_{i=0}^{Q-m} |L_h x_h^*(t_i)|^p, \tag{3.4}$$

that is, $f(p_h x_h^*) = f_h(x_h^*)$. Thus we have accomplished the goal of making $|f(p_h x_h^*) - f_h(x_h^*)|$ tend to zero; we now check to see if $z_h = p_h x_h^*$ is "near" C by relating V_h to v_h.

Now we write

$$V_h(t_{j+1}) = V_h(t_j) + \int_{t_j}^{t_{j+1}} [A(t)V(t) + eu_h(t)] \, dt$$

and

$$v_h(t_{j+1}) = v_h(t_j) + \int_{t_j}^{t_{j+1}} [A(t_j)v_h(t_j) + eu_h(t)] \, dt.$$

Letting $e_h(t_j) = V_h(t_j) - v_h(t_j)$ and arguing in the usual way, we find, writing

$$\|e_h(t_j)\|_\infty = \max_{0 \leqq i \leqq m-1} |e_{h,i}(t_j)| \quad \text{and} \quad F = \max_{0 \leqq t \leqq 1} \|A(t)\|_\infty,$$

$$\|e_h(t_j)\|_\infty \leqq \frac{\exp(F) - 1}{(1 + 1/q)F} h^{1/q} \int_0^1 |(A(t)V_h(t))^{(1)}|^p \, dt.$$

Thus, v_h and V_h will be uniformly close if $AV_h \in W^{1,p}$ and is uniformly bounded in $W^{1,p}$. If $A^{(1)} \in C[0, 1]$, that is $A \in C^1[0, 1]$, then since $(AV_h)^{(1)} = A^{(1)}V_h + A^2V_h + Aeu_h$ and $\|A^{(1)}\|_\infty$, $\|A\|_\infty$ and $\int_0^1 |u_h(t)|^p \, dt$ are uniformly bounded, AV_h will be uniformly bounded in $W^{1,p}$ if V_h is uniformly bounded in $L^p(0, 1)$; this finally is clearly true if $V_h(0)$ is uniformly bounded in $\mathscr{R}^m$.

LEMMA 3.3. *If the coefficients a_i defining the operator L are in $C^1[0, 1]$, then there exists a constant K such that $\|v_h(t_j) - V_h(t_j)\|_\infty \leqq Kh^{1/q}$ for $0 \leqq j \leqq Q - m$, $1/p + 1/q = 1$. Also $\|V_{h,0}\| = \|p_h x_h^*\|$ is uniformly bounded.*

Proof. Because of the preceding arguments, we need only show that $V_h(0) = v_h(0)$ is uniformly bounded in $\mathscr{R}^m$. Because of (3.1) we can write $v_h(t_j)$ via

$$\begin{aligned} v_h(t_j) = h \sum_{i=0}^{j-1} &[I + hA(t_{j-1})] \cdots [I + hA(t_{i+1})]u_h(t_i) \\ &+ [I + hA(t_{j-1})] \cdots [I + hA(t_0)]v_h(0). \end{aligned} \tag{3.5}$$

Consider the term in this expression involving the sum; this, call it w_h, solves (3.1) with $w_h(0) = 0$ and is therefore within $O(h^{1/q})$ of the solution W_h to (3.3) with $W_h(0) = 0$ by our preceding arguments. Note that, by applying the operators $M_{l,h}$ at $\theta_{l,r}$ and $N_{r,h}$ at ξ_r to the first components of the vectors on both sides of (3.5), we immediately see that $v_h(0)$ solves a certain system of linear equations. Applying one of these operators to the first component of the left-hand side yields merely that operator applied to x_h^*, and these values are uniformly bounded at the $\theta_{l,r}$ and ξ_r. Applying an operator D^l for $0 \leqq l \leqq m - 1$ to $w_{h,0}$ merely gives $w_{h,l}$, which is uniformly close to $W_{h,l} = W_{h,0}^{(l)}$, which is uniformly bounded; it then follows that application of one of the operators $M_{l,h}$ or $N_{r,h}$ to $w_{h,0}$ gives uniformly bounded values.

Thus we have found that $v_h(0)$ solves a linear system with right-hand side uniformly bounded in $\mathscr{R}^m$. A typical row in the matrix B_h of this system consists of, say, $M_{l,h}$ applied at $\theta_{l,r}$ to the components of the first row of the matrix function whose value at t_j is

$$[I + hA(t_{j-1})] \cdots [I + hA(t_0)].$$

Arguing as we have done above, it is easy to show that such an expression converges uniformly to the row (the collection of which forms a matrix B) consisting of the application of M_l at $\theta_{l,r}$ to the components of the first row of the matrix function whose value at t is

$$\exp\left[\int_0^t A(\tau) \, d\tau\right].$$

A matrix B of such rows, however, must be of full rank, since by assumption there are no nonzero functions $x \in W^{m,p}$ such that $\|x\| = 0$. If we only apply those operators at those points which in the limit give an $m \times m$ nonsingular matrix, as we can always do since rank $(B) = m$, then for small h, the matrices multiplying $v_h(0)$ are uniformly nonsingular, and therefore the $v_h(0) = V_h(0)$ are uniformly bounded in $\mathscr{R}^m$. Since $V_h(0)$ is uniformly bounded, it follows that $\|V_{h,0}\|$ is also.

We can now prove convergence for the more complex discretization.

THEOREM 3.1. *Let the general assumptions of* § 1 *hold and let the fixed points* $\theta_{l,i}$ *defining the norm* $\|\cdot\|$ *in* (2.1) (*or the* θ_i *in* (2.2)) *be mesh points in our discretization for all h. Let the problem in* (1.1) *not admit solutions of arbitrarily large* $W^{m,p}$*-norm, for example, some* $M_l x \equiv x$. *Let* $x_h^* \in \mathscr{R}^{Q+1}$ *solve the problem in* (1.3), *that is, minimize* $f_h(x_h)$ *over* $C_2(h)$, *where* $\varepsilon_h \geqq Gh^{1/q}$ *and* G *is defined in Lemma* 3.1; *one may thus take* $h^{1/q} = o(\varepsilon_h)$ *for small h. Suppose the functions* a_i *defining L lie in* $C^1[0, 1]$. *Let* $p_h x_h^* = z_h$ *solve*

$$Lz_h = \begin{cases} L_h x_h^*(t_i) & \text{for } t_i \leqq t < t_{i+1}, 0 \leqq i \leqq Q - m, \\ 0 & \text{for } t \geqq 1 - (m-1)h. \end{cases}$$

Then $f_h(x_h^*)$ *converges to* $f(x^*)$ *and all* $W^{m,p}$ *weak limit points, at least one of which exists, of* $\{p_h x_h^*\}$, *minimize f over C; if* x^* *minimizing f over C is unique, then* $p_h x_h^*$ *converges weakly to* x^*. *If* (*some subsequence of*) $p_h x_h^*$ *converges weakly to a point x, then* x_h^* *and its first* $m - 1$ *difference approximations* $D^l x_h^*$ *evaluated at the points* $t_i = ih, 0 \leqq i \leqq Q - l$, *converge uniformly, that is, in* $C[0, 1]$, *to x and its first* $m - 1$ *derivatives at the points* t_i.

Proof. By Lemma 3.1 and the hypothesis on ε_h, $r_h x^* \in C_2(h)$, so $C_2(h)$ is not empty. Since $C_2(h)$ is not empty, x_h^* exists. By Lemmas 3.3 and 2.2 and the facts that $v_{h,0} = x_h^* \in C_2(h)$ and $\|p_h x_h^*\|$ is uniformly bounded, there exist functions $\eta_i(h)$ tending to zero with h and such that $p_h x_h^* \in C_h \equiv \{x; x \in W^{m,p}, \ -\eta_i(h) - \varepsilon_h - Kh^{1/q} + \alpha_i(t) \leqq M_i x(t) \leqq \beta_i(t) + \eta_i(h) + \varepsilon_h + Kh^{1/q}$ for $1 \leqq i \leqq k, \ -\varepsilon_h - Kh^{1/q} + \gamma_i \leqq N_i x(\xi_i) \leqq \delta_i + \varepsilon_h + Kh^{1/q}$ for $1 \leqq i \leqq n\}$. By Lemma 2.4, $\zeta_h \equiv \min_C f - \min_{C_h} f$ tends to zero. We write

$$f(x^*) = \min_C f = \min_{C_h} f + \zeta_h \leqq f(p_h x_h^*) + \zeta_h = f_h(x_h^*) + \zeta_h,$$

the last equality following from the construction of p_h. Thus, we have

$$\begin{aligned} f(x^*) &\leqq f(p_h x_h^*) + \zeta_h = f_h(x_h^*) + \zeta_h \leqq f_h(r_h x^*) + \zeta_h \\ &\leqq f(x^*) + \zeta_h + [f_h(r_h x^*) - f(x^*)]. \end{aligned}$$

From Lemma 3.2 and this inequality, we conclude that $f(x^*) = \lim_{h \to 0} f(p_h x_h^*) = \lim_{h \to 0} f_h(x_h^*)$. Since, from Lemma 3.3, $p_h x_h^*$ is bounded, it has weak limit points; for any such weak limit point x' we have $x' \in C$ by Lemma 2.3 and thus

$$f(x^*) \leqq f(x') \leqq \liminf_{h' \to 0} f(p_h x_h^*) = f(x^*),$$

which says that x' minimizes f over C. If $p_h x_h^*$ converges to some x weakly in $W^{m,p}$, then $p_h x_h^*$ and its first $m - 1$ derivatives converge uniformly to x and its first $m - 1$ derivatives. By Lemma 3.3, the numbers $v_{h,l}(t_j) = D^l v_{h,0}(t_j) = D^l x_h^*(t_j)$ are uniformly close to $V_{h,l}(t_j) = V_{h,0}^{(l)}(t_j) = p_h x_h^{*(l)}(t_j)$ for $0 \leqq l \leqq m - 1$.

We have not been able to estimate the rate of convergence as a function of h.

4. An elementary example. Consider the example in [10] with $m = 1$, $Lx \equiv x^{(1)}$, $k = 1$, $M_1 x \equiv x$, $\alpha_1(t) = t - t^2$, $\beta_1(t) = t$, $n = 1$, $N_1 x(\xi_1) = x(1)$, $\delta_1 = \gamma_1 = c \in [0, 1]$, that is,

$$\text{minimize} \quad \int_0^1 |x^{(1)}(t)|^2 \, dt \quad \text{over}$$

$$C = \{x; x \in W^{1,2}, t - t^2 \leqq x(t) \leqq t, x(1) = c\}.$$

The unique solution to this problem is

$$x^*(t) = \begin{cases} t - t^2 & \text{for } 0 \leqq t \leqq 1 - \sqrt{c}, \\ (2\sqrt{c} - 1)t + (1 + c - 2\sqrt{c}) & \text{for } 1 - \sqrt{c} \leqq t \leqq 1, \end{cases}$$

as pictured in Fig. 4.1.

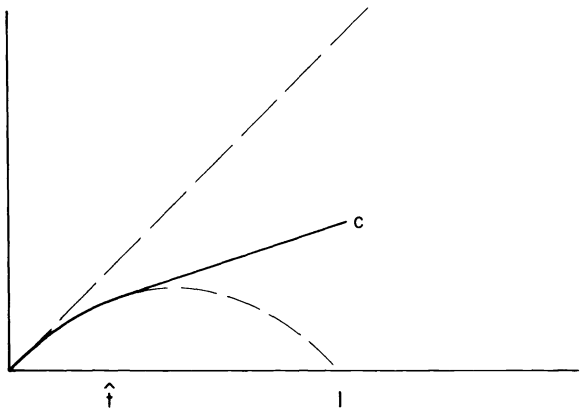


FIG. 4.1

Let $\hat{t} = 1 - \sqrt{c}$, the point at which x^* leaves the lower curve.

If we use the simple discretization and merely discretize the constraints at $t_i = ih$, $0 \leqq i \leqq Q = 1/h$, then if $\hat{t}_i = \max\{t_i; t_i \leqq \hat{t}\}$, then x_Q^* is just the piecewise linear interpolant of $t - t^2$ at t_i for $t \leqq \hat{t}_i$ and is the linear interpolant between $\hat{t}_i - \hat{t}_i^2$ and c for $\hat{t}_i \leqq t \leqq 1$. For small enough h, the solution for the complete discretization is also unique; such discrete variational splines are studied in [11].

If we define, for $c > 1/2$, the numbers

$$\alpha_h = \min\{t_i; t_i \geqq \sqrt{2\varepsilon_h}\}, \qquad \beta_h = \max\{t_i; t_i \leqq 1 - \sqrt{c}\},$$

the unique solution x_h^* for the complete discretization with $\varepsilon_h = h^{1/2-\delta}$,

$\delta \in (0, 1/2)$, is

$$x_h^*(t_i) = \begin{cases} \varepsilon_h + t_i\left[\dfrac{\alpha_h - \alpha_h^2 - 2\varepsilon_h}{\alpha_h}\right] & \text{if } 0 \leqq t_i \leqq \alpha_h, \\ t_i - t_i^2 - \varepsilon_h & \text{if } \alpha_h \leqq t_i \leqq \beta_h, \\ \beta_h - \beta_h^2 - \varepsilon_h + (t_i - \beta_h)\left[\dfrac{c - \beta_h + \beta_h^2}{1 - \beta_h}\right] & \text{if } \beta_h \leqq t_i \leqq 1. \end{cases}$$

Acknowledgment. The author thanks O. L. Mangasarian and L. L. Schumaker for their suggestion of the problem and for their continuing interest.

REFERENCES

[1] M. Atteia, *Fonctions "spline" définés sur un ensemble convexe*, Numer. Math., 12 (1968), pp. 192–210.

[2] J. Cullum, *Discrete approximations to continuous optimal control problems*, this Journal, 7 (1969), pp. 32–49.

[3] ———, *An explicit procedure for discretizing continuous optimal control problems*, J. Optimization Theory and Applications, to appear.

[4] H. B. Curry and I. J. Schoenberg, *On Polya frequency functions. IV: The fundamental spline functions and their limits*, J. Analyse Math., 17 (1966), pp. 71–107.

[5] J. W. Daniel, *On the approximate minimization of functionals*, Math. Comp., 23 (1969), pp. 573–582.

[6] ———, *On the convergence of a numerical method for optimal control problems*, J. Optimization Theory and Applications, 4 (1969), pp. 330–342.

[7] ———, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, N.J., to appear.

[8] M. Golumb and J. W. Jerome, *Ordinary differential equations with boundary conditions on arbitrary sets*, to appear.

[9] J. W. Jerome and L. L. Schumaker, *On Lg-splines*, J. Approx. Theor., 2 (1969), pp. 29–49.

[10] O. L. Mangasarian and L. L. Schumaker, *Splines via optimal control*, Approximations with Special Emphasis on Spline Functions, I. J. Schoenberg, ed., Academic Press, New York, 1969.

[11] ———, *Discrete splines via mathematical programming*, this Journal, to appear.

[12] C. B. Morrey, Jr., *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, 1966.

[13] K. Ritter, *Generalized spline interpolation and nonlinear programming*, Approximations with Special Emphasis on Spline Functions, I. J. Schoenberg, ed., Academic Press, New York, 1969.

[14] S. L. Sobolev, *Applications of Functional Analysis in Mathematical Physics*, vol. 7, Trans. Math. Monographs, American Mathematical Society, Providence, R.I., 1963.

SIAM J. CONTROL
Vol. 9, No. 1, February 1971

INVARIANCE PROPERTIES IN THE THEORY OF ORDINARY DIFFERENTIAL EQUATIONS WITH APPLICATIONS TO STABILITY PROBLEMS*

NELSON ONUCHIC†

1. Introduction. The main purpose of this paper is to find conditions under which every solution of a second order scalar differential equation tends to zero with its derivative as $t \to \infty$. The basic tool to carry out this objective is provided by Theorems 1, 2 and 3 given below.

Theorem 1 is a modified but closely related version of Yoshizawa's theorem 5 in [6]. Theorem 2 is essentially a special case of Miller's theorem 1 in [3]. Some remarks relating Theorems 1 and 2 and those ones by Yoshizawa and Miller will be done following each theorem. Theorem 3 is a particular case of La Salle's Theorem 1 in [1].

Theorem 4 is a result concerning the above posed problem obtained as an application of Theorems 1 and 2. This result is more general than the one by Yoshizawa in [6, Example 2]. See also [2] and [7, Example 14.1, p. 63]. Theorem 5, previously obtained by us [4], is here stated because it gives additional information on the problem under consideration which is intimately related to Theorem 4. Theorem 6 also gives a contribution to our problem and it is strongly dependent on Theorem 3.

2. Invariance properties. Consider a system of differential equations defined on an open set $Q \subset R^n$:

$$\dot{x} = H(x), \tag{1}$$

where $H(x)$ is continuous on Q.

If M is a subset of Q, then M is called quasi-invariant with respect to (1) if and only if for each $x_0 \in M$ there is a solution $x(t)$ of (1) with $x(0) = x_0$, such that $x(t)$ exists and remains in M for all real t. Let $x(t)$ be a continuous function defined in the future, that is, for all $t \geqq$ some real t_0. A point p of R^n is said to be a limit point of $x(t)$ if there exists a sequence $\{t_m\}$, $t_m \to \infty$ as $m \to \infty$, such that $x(t_m) \to p$ as $m \to \infty$. The set of all limit points of $x(t)$ is denoted by Ω and is called the ω-limit set of $x(t)$. If $x(t)$ is bounded in the future, that is, $x(t)$ is bounded on some interval $[a, \infty)$, $a > -\infty$, it is easily seen that Ω is a nonempty, connected and compact set with $x(t) \to \Omega$ as $t \to \infty$, that is, $\operatorname{dist}(x(t), \Omega) \to 0$ as $t \to \infty$.

Consider the differential system defined on $[0, \infty) \times Q$, Q being an open set of R^n,

$$\dot{x} = f(t, x), \tag{2}$$

where f is assumed to be continuous for $t \geqq 0$, $x \in Q$.

Let $V(t, x)$ be a real-valued C^1 function defined for $t \geqq 0$, $x \in Q$. Define

$$\dot{V}_{(2)}(t, x) = \sum_{j=1}^{n} \frac{\partial V(t, x)}{\partial x_j} f_j(t, x) + \frac{\partial V(t, x)}{\partial t}.$$

* Received by the editors June 17, 1970.

† Escola de Engenharia de São Carlos, São Carlos, SP Brazil. This research was supported in part by the Conselho Nacional de Pesquisas, Brazil.

Consider also the differential system defined on $[0, \infty) \times Q$:

$$\dot{x} = F(t, x) + G(t, x), \tag{3}$$

where $F(t, x)$ and $G(t, x)$ are continuous for $t \geqq 0$, $x \in Q$.

THEOREM 1. *Suppose that the following hypotheses hold with respect to system* (3):

(i) *$F(t, x)$ is bounded for all $t \geqq 0$ when x belongs to an arbitrary compact subset of Q.*

(ii) *For every compact set $B \subset Q$ and every continuous function $z(t) \in B$, defined on $[0, \infty)$, it follows that*

$$\int_s^{s+t} G(v, z(v))\, dv \to 0 \quad as \quad s \to \infty, \tag{4}$$

uniformly on $t \in [0, 1]$.

(iii) *There are a real-valued nonnegative C^1 function $V(t, x)$ and a real-valued nonnegative continuous function $W(x)$ such that*

$$\dot{V}_{(3)}(t, x) \leqq - W(x), \quad t \geqq 0 \quad and \quad x \in Q.$$

Let $x(t)$ be a solution of (3), *defined in the future, with $x(t) \in K$ for $t \geqq$ some t_0, where K is a compact subset of Q.*

Then $\Omega \subset E = \{x \in Q | W(x) = 0\}$, where Ω is the ω-limit set of $x(t)$.

Note. The main difference between the above theorem and Theorem 5 in [6] is that Yoshizawa assumes the condition that $\int^{\infty} \|G(v, z(v))\|\, dv < \infty$, which is stronger than the one given by (4).

A sufficient condition for (4) is given as follows:

(a) For every compact set $B \subset Q$ there corresponds a scalar function $\sigma_B(t)$ defined for $t \geqq 0$ such that

$$\int_t^{t+1} \sigma_B(s)\, ds \to 0 \quad \text{as} \quad t \to \infty,$$

and $\|G(t, x)\| \leqq \sigma_B(t)$ for all $t \geqq 0$, $x \in B$.

An example considered in [5] shows that condition (4) is not implied by (a).

The proof of Theorem 1 can be done by following essentially the same ideas contained in the proof of Theorem 5 in [6].

Consider the differential system defined on $[0, \infty) \times Q$:

$$\dot{x} = H(x) + S(t, x) + G(t, x), \tag{5}$$

where $H(x)$, $S(t, x)$ and $G(t, x)$ are assumed to be continuous on $[0, \infty) \times Q$.

Let A be a fixed subset of Q. Assume that $S(t, x)$ satisfies the following property with respect to the set A:

(b) For each $\varepsilon > 0$ and each compact subset K or Q there corresponds $\delta = \delta(\varepsilon, K) > 0$ and $T_0 = T_0(\varepsilon, K) > 0$ such that $t \geqq T_0$, $x \in K$ and dist $(x, A) < \delta$ imply $\|S(t, x)\| < \varepsilon$.

THEOREM 2. *Let hypotheses* (4) *and* (b) *hold. Let* $x(t)$ *be a solution of* (5) *such that* $x(t) \in K$, $T \leqq t < \infty$, *where* $T \geqq 0$ *and* K *is a compact subset of* Q *and* $x(t) \to A$ *as* $t \to \infty$. *Then the* ω*-limit set* Ω *of* $x(t)$ *is a nonempty, connected, compact and quasi-invariant set with respect to* (1).

Note. As observed, this theorem is essentially a particular case of Miller's Theorem 1 in [3]. Miller deals with delay equations where the unperturbed system is uniformly almost periodic. However, condition (4) considered in Theorem 2 is weaker than the one in Miller's theorem. But Miller's proof also works well with condition (4).

Consider the system

$$\dot{x} = F(t, x), \tag{6}$$

where F is continuous for $t \geqq 0$, $x \in R^n$ and satisfies any one of the conditions guaranteeing uniqueness of solutions.

THEOREM 3. *Let* $V(t, x)$ *and* $W(x)$ *be real-valued nonnegative* C^1 *functions on* $[0, \infty) \times R^n$ *such that* $\dot{V}_{(8)}(t, x) \leqq -W(x)$ *for all* $t > 0$ *and* $x \in G$, *where* G *is a compact set of* R^n. *Let* $x(t)$ *be a solution of* (6) *such that* $x(t) \in G$ *for all* $t > t_0$ *and let* $\dot{W}$ *be bounded from above or below along the solution* $x(t)$. *Then* $\Omega \subset E = \{x \in G | W(x) = 0\}$, *where* Ω *is the* ω*-limit set of* $x(t)$.

3. Applications. The objective of this section is to study the problem posed in § 1 by using the results discussed in § 2. Specifically, our purpose is to find conditions under which, for all solutions $x(t)$ of the second order scalar differential equation

$$\ddot{x} + h(t, x, \dot{x})\dot{x} + f(x) + g(t, x, \dot{x}) + p(t, x, \dot{x}) = 0, \tag{7}$$

we can guarantee that $(x(t), \dot{x}(t)) \to (0, 0)$ as $t \to \infty$. To this end, consider the equivalent equation

$$\begin{aligned} &\dot{x} = y, \\ &\dot{y} + h(t, x, y)y + f(x) + g(t, x, y) + p(t, x, y) = 0 \end{aligned} \tag{7'}$$

and the following set of assumptions with respect to the functions in (7').

(H_1) $h(t, x, y)$ is continuous in $[0, \infty) \times R^2$ and $h(t, x, y)$ is bounded when $x^2 + y^2$ is bounded, and, moreover, $h(t, x, y) \geqq k(x, y) \geqq 0$, where $k(x, y)$ is a continuous function in R^2.

(H_2) The sets $R^+ = \{(x, 0) | x > 0\}$ and $R^- = \{(x, 0) | x < 0\}$ are connected components with respect to the topological space $\Gamma - \{(0, 0)\}$, where $\Gamma = \{(x, y) \in R^2 | yk(x, y) = 0\}$.

(H_3) $f(x)$ is continuous, $xf(x) > 0$ for all $x \neq 0$, and

$$\int_0^x f(s)\, ds \to \infty \quad \text{as } |x| \to \infty.$$

(H_3') $f(x)$ is continuous in R and there is a positive ρ such that $\int_0^x f(s)\, ds > 0$ for $0 < |x| \leqq \rho$.

(H_3'') $f(x)$ is continuous in R and $\int_0^x f(s)\,ds \to \infty$ as $|x| \to \infty$.

We see that (H_3) implies (H_3') and (H_3) implies (H_3'').

(H_4) $p(t, x, y)$ is continuous and $|p(t, x, y)| \leqq \beta(t)$ for all $t \geqq 0$, x, y in R, where $\beta(t)$ is continuous with $\int_0^\infty \beta(t)\,dt < \infty$.

(H_5) $g(t, x, y)$ is continuous and $yg(t, x, y) \geqq 0$ for all $t \geqq 0$, x, y in R.

(H_6) For every compact subset B of R and for all continuous functions $x(t)$ and $y(t)$, defined on $[0, \infty)$ with values in B, it follows that

$$\int_s^{s+t} g(v, x(v), y(v))\,dv \to 0 \quad \text{as} \quad s \to \infty,$$

uniformly on $t \in [0, 1]$.

A sufficient condition for (H_6) is given as follows:

$(\tilde{H}_6)$ For every compact subset B of R there corresponds a real function $\sigma_B(t)$, defined for $t \geqq 0$, such that

$$\int_t^{t+1} \sigma_B(s)\,ds \to 0 \quad \text{as} \quad t \to \infty,$$

and $|g(t, x, y)| \leqq \sigma_B(t)$ for all $t \geqq 0$, x, y in B.

LEMMA 1. *Let hypotheses (H_3'), (H_4) and (H_5) hold. Let $h(t, x, y)$ be non-negative and continuous for $t \geqq 0$, x, y in R. Then for each $\varepsilon > 0$ there are positive $T(\varepsilon)$ and $\delta(\varepsilon)$ such that $t \geqq t_0 \geqq T(\varepsilon)$ and $(x_0, y_0) \in R^2$, with $|x_0| + |y_0| < \delta$, imply $|x(t)| + |y(t)| < \varepsilon$ for every solution $(x(t), y(t))$ of (7') satisfying $x(t_0) = x_0$ and $y(t_0) = y_0$.*

Proof. Consider the function

$$V(t, x, y) = \left[y^2 + 2\int_0^x f(s)\,ds\right]^{1/2} + \int_t^\infty \beta(s)\,ds$$

defined for $t \geqq 0$ and $|x| + |y| < \rho$.

Given $0 < \varepsilon < \rho$, define

$$2m = \min_{|x|+|y|=\varepsilon} \left[y^2 + 2\int_0^x f(s)\,ds\right]^{1/2} > 0.$$

Then condition (H_3') implies

$$\inf_{\substack{t \geqq 0 \\ |x|+|y|=\varepsilon}} V(t, x, y) \geqq 2m > 0.$$

Let $T(\varepsilon) > 0$ and $0 < \delta(\varepsilon) < \varepsilon$ be chosen such that

$$\int_{T(\varepsilon)}^\infty \beta(t)\,dt > m \quad \text{and} \quad \left[y^2 + 2\int_0^x f(s)\,ds\right]^{1/2} < m$$

for $|x| + |y| \leqq \delta(\varepsilon)$. Then

$$\max_{\substack{t \geqq T(\varepsilon) \\ |x|+|y|=\delta(\varepsilon)}} V(t, x, y) < 2m.$$

An easy computation shows that

$$\dot{V}_{(7')}(t, x, y) \leqq \frac{-y^2 h(t, x, y) - y g(t, x, y)}{\left[y^2 + 2\int_0^x f(s)\,ds\right]^{1/2}} + |p(t, x, y)| - \beta(t) \leqq 0$$

for all $t \geqq 0$, $0 < |x| + |y| < \rho$.

Suppose that there are $t_0 \geqq T(\varepsilon)$ and $(x_0, y_0) \in R^2$ with $|x_0| + |y_0| < \delta(\varepsilon)$, such that for some solution $(x(t), y(t))$ of (7′) satisfying $(x(t_0), y(t_0)) = (x_0, y_0)$ and for some $\tilde{t} > t_0$ we have $|x(\tilde{t})| + |y(\tilde{t})| \geqq \varepsilon$. Then there are real numbers t_1 and t_2, $t_0 < t_1 < t_2$, such that $|x(t_1)| + |y(t_1)| = \delta(\varepsilon)$, $|x(t_2)| + |y(t_2)| = \varepsilon$ and $\delta(\varepsilon) \leqq |x(t)| + |y(t)| \leqq \varepsilon$ for $t_1 \leqq t \leqq t_2$. But this implies $V(t_1, x(t_1), y(t_1)) < V(t_2, x(t_2), y(t_2))$ and, since $(x(t), y(t)) \neq (0, 0)$ on $[t_1, t_2]$, we have that $V(t, x(t), y(t))$ is differentiable in $[t_1, t_2]$. Hence there is a real number s, $t_1 < s < t_2$, satisfying $\dot{V}(s, x(s), y(s)) > 0$, leading to a contradiction. The proof is complete.

COROLLARY 1. *Suppose that some uniqueness condition with respect to the initial value problem holds for* (7′). *Suppose that all hypotheses of Lemma* 1 *are satisfied with* $p(t, 0, 0) = 0$ *for* $t \geqq 0$. *Then the null solution of* (7′) *is uniformly stable.*

Proof. Corollary 1 follows from Lemma 1, taking into account that the origin is an equilibrium point of (7′) and considering the usual continuous dependence argument.

We observe that $g(t, 0, 0) = 0$ is a consequence of hypothesis (H_5).

LEMMA 2. *Let hypotheses* (H_3'), (H_4) *and* (H_5) *hold. Let* $h(t, x, y)$ *be nonnegative and continuous for* $t \geqq 0$, x, y *in* R. *Then every solution of* (7′) *is bounded in the future.*

Proof. Let $V(t, x, y) = \left[y^2 + 2\int_0^x f(s)\,ds + M\right]^{1/2} + \int_t^\infty \beta(s)\,ds$, where M is chosen so that $2\int_0^x f(s)\,ds + M > 0$ for all x in R. It is easy to see that $\dot{V}_{(7')}(t, x, y) \leqq 0$ for all $t \geqq 0$ and x, y in R. Then as $W(x, y) = \left[y^2 + 2\int_0^x f(s)\,ds + M\right]^{1/2} \leqq V(t, x, y)$ and $W(x, y) \to \infty$ as $x^2 + y^2 \to \infty$ it follows that every solution of (7′) is bounded in the future.

The proof is complete.

THEOREM 4. *Let hypotheses* (H_1) *through* (H_6) *hold. Then for every solution* $x(t)$ *of* (7) *we have that* $x(t) \to 0$ *and* $\dot{x}(t) \to 0$ *as* $t \to \infty$.

Proof. Let $(x(t), y(t))$ be any solution of (7′). Lemma 2 implies that this solution is bounded in the future. Then we know that the ω-limit set Ω of $(x(t), y(t))$ is a nonempty, connected and compact set with $(x(t), y(t)) \to \Omega$ at $t \to \infty$. We must have $\Omega \cap R_x \neq \varnothing$, where R_x is the x-axis, because, if this were not true, it would follow that $|x(t)| \to \infty$ as $t \to \infty$, a contradiction.

Defining $V(t, x, y) = \left[y^2 + 2\int_0^x f(s)\,ds + 1\right]^{1/2} + \int_t^\infty \beta(s)\,ds$ and

$$W(x, y) = \frac{y^2 k(x, y)}{\left[y^2 + 2\int_0^x f(s)\,ds + 1\right]^{1/2}},$$

a computation shows, by taking into account hypothesis (H_1), that

$$\dot{V}_{(7')}(t, x, y) \leqq -W(x, y) \leqq 0$$

for $t \geqq 0$, $(x, y) \in Q = R^2$. Then it follows from Theorem 1 that

$$\Omega \subset \Gamma = \{(x, y) \in R^2 | yk(x, y) = 0\}.$$

We claim that $(0, 0) \in \Omega$. Indeed, otherwise, since $\Omega \cap R_x \neq \varnothing$, it would follow that $\Omega \subset \Gamma - \{(0, 0)\}$ and $\Omega \cap (R^+ \cup R^-) \neq \varnothing$. But, since Ω is a connected set and (H_2) holds, it would follow that $\Omega \subset R^+ \cup R^-$. Then, by applying Theorem 2 with $A = R_x$, $H = (y, -f(x))$, $S = (0, -h(t, x, y)y)$, $G = (0, -g(t, x, y) - p(t, x, y))$ and $Q = R^2$, it would follow that Ω is quasi-invariant with respect to the system

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} + f(x) &= 0. \end{aligned} \tag{8}$$

But condition (H_3) implies that the unique quasi-invariant set for (8) contained in R_x is $(0, 0)$, leading to a contradiction. Thus $(0, 0) \in \Omega$, and hence there exists a sequence $\{t_m\}$, $t_m \to \infty$ as $m \to \infty$, such that $(x(t_m), y(t_m)) \to (0, 0)$ as $m \to \infty$. Then it follows from Lemma 1 that $(x(t), y(t)) \to (0, 0)$ as $t \to \infty$, completing the proof.

COROLLARY 2. *Suppose that some uniqueness condition with respect to the initial value problem holds for* (7'). *Let hypotheses* (H_1) *through* (H_6) *hold and* $p(t, 0, 0) = 0$. *Then the origin is globally asymptotically stable for* (7').

Proof. This corollary follows from Corollary 1 and Theorem 4.

The next theorem, which is a result previously obtained by us [4, Theorem], gives sufficient conditions to guarantee that for every solution $x(t)$ of the scalar equation

$$\ddot{x} + h(x, \dot{x}) + f(x) + g(t, x, \dot{x}) + p(t, x, \dot{x}) = 0, \tag{9}$$

we have $x(t) \to 0$ and $\dot{x}(t) \to 0$ as $t \to \infty$. Most of the ideas used in the proof of Theorem 5 are closely related to the ones in the proof of Theorem 4, but Theorem 5 depends also on the well-known Poincaré–Bendixon theorem for two-dimensional autonomous systems.

THEOREM 5. *Suppose that the following hypotheses hold with respect to* (9):

(i) $h(x, y)$ *is continuous and* $yh(x, y) \geqq 0$ *for all real numbers* x *and* y.

(ii) *For every Jordan curve* γ *in* R^2 *containing the origin in its interior, there exists at least one point* $(x, y) \in \gamma$ *such that* $y \neq 0$ *and* $h(x, y) \neq 0$.

(iii) $g(t, x, y)$ *is continuous with* $y[h(x, y) + g(t, x, y)] \geqq 0$ *for all* $t \geqq 0$ *and all real numbers* x *and* y.

(iv) *Hypotheses* (H_3), (H_4) *and* (H_6) *hold.*

Then for every solution $x(t)$ *of* (9) *we have that* $x(t) \to 0$ *and* $\dot{x}(t) \to 0$ *as* $t \to \infty$.

COROLLARY 3. *Suppose that some uniqueness condition with respect to the initial value problem holds for*

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} + h(x, y) + f(x) + g(t, x, y) + p(t, x, y) &= 0. \end{aligned} \tag{9'}$$

Suppose that all hypotheses of Theorem 5 are satisfied, with $p(t, 0, 0) = 0$ *for all* $t \geqq 0$. *Then the origin is globally asymptotically stable for* (9′).

THEOREM 6. *Let* $h(t, x, y)$ *be continuous for* $t \geqq 0$, x, y *in* R *and moreover,* $h(t, x, y) \geqq k(x, y) \geqq 0$, *where* $k(x, y)$ *is a* C^1 *function with* $yk_y(x, y) \geqq 0$ *for all* x, y *in* R. (*Here* k_y *denotes the partial derivative with respect to* y.) *Let hypotheses* (H_2), (H_3'), (H_3'') *and* (H_5) *hold. Let hypothesis* (H_4) *be satisfied with* $\beta(t)$ *bounded on* $(0, \infty)$. *Suppose that any one of the conditions guaranteeing uniqueness of solutions is satisfied with respect to* (7′).

Then every solution $x(t)$ *of* (7) *is bounded in the future and* $\dot{x}(t) \to 0$ *as* $t \to \infty$. *Furthermore if* $x(t)$ *is a solution of* (7) *such that* $\liminf_{t\to\infty} |x(t)| = 0$, *then also* $x(t) \to 0$ *as* $t \to \infty$.

Proof. Let $x(t)$ be any solution of (7). Then it follows from Lemma 2 that $(x(t), y(t))$ is bounded in the future. Hence, there are $t_0 \geqq 0$ and a compact set G of R^2 such that $(x(t), y(t)) \in G$ for all $t > t_0$.

Define $V(t, x, y)$ as in Lemma 2, that is,

$$V(t, x, y) = \left[y^2 + 2\int_0^x f(s)\,ds + M \right]^{1/2} + \int_t^\infty \beta(s)\,ds.$$

It is easily seen that

$$\dot{V}_{(7')}(t, x, y) \leqq -\frac{y^2 k(x, y)}{\left[y^2 + 2\int_0^x f(s)\,ds + M \right]^{1/2}}.$$

By taking

$$\gamma = \inf_{(x,y)\in G} \frac{1}{\left[y^2 + 2\int_0^x f(s)\,ds + M \right]^{1/2}},$$

it follows that

$$\frac{y^2 k(x, y)}{\left[y^2 + 2\int_0^x f(s)\,ds + M \right]^{1/2}} \geqq \gamma y^2 k(x, y)$$

for all $(x, y) \in G$. Therefore, defining $W(x, y) = \gamma y^2 k(x, y)$, we have that $\dot{V}_{(7')}(t, x, y) \leqq -W(x, y)$ for all $t > 0$, $(x, y) \in G$.

To apply Theorem 3, let us show that $\dot{W}$ is bounded from above along each solution which remains in G for all $t > t_0 \geqq 0$.

Let $z(t)$ be a solution of (7) such that $(z(t), \dot{z}(t)) \in G$ for all $t > t_0$. Then

$$\begin{aligned}
\dot{W}(z(t), \dot{z}(t)) &= \gamma \frac{d}{dt}\{[\dot{z}(t)]^2 k(z(t), \dot{z}(t))\} \\
&= 2\gamma \dot{z}(t)\ddot{z}(t)k(z(t), \dot{z}(t)) + \gamma[\dot{z}(t)]^3 k_x(z(t), \dot{z}(t)) + \gamma[\dot{z}(t)]^2 k_y(z(t), \dot{z}(t))\ddot{z}(t) \\
&= \gamma[\dot{z}(t)]^3 k_x(z(t), \dot{z}(t)) + 2\gamma \dot{z}(t)k(z(t), \dot{z}(t))[-h(t, z(t), \dot{z}(t))\dot{z}(t) - f(z(t)) \\
&\qquad - g(t, z(t), \dot{z}(t)) - p(t, z(t), \dot{z}(t))]
\end{aligned}$$

$$+ \gamma[\dot{z}(t)]^2 k_y(z(t), \dot{z}(t))[-h(t, z(t)), \dot{z}(t))\dot{z}(t)$$
$$- f(z(t)) - g(t, z(t), \dot{z}(t)) - p(t, z(t), \dot{z}(t))]$$
$$\leqq \gamma[\dot{z}(t)]^3 k_x(z(t), \dot{z}(t)) - 2\gamma\dot{z}(t)k(z(t), \dot{z}(t))f(z(t))$$
$$+ 2\gamma|\dot{z}(t)|k(z(t), \dot{z}(t))\beta_0 - \gamma[\dot{z}(t)]^2 k_y(z(t), \dot{z}(t))f(z(t))$$
$$+ \gamma[\dot{z}(t)]^2|k_y(z(t), \dot{z}(t))|\beta_0,$$

where $\beta_0 = \sup_{0 \leqq t < \infty} \beta(t)$.

Then as G is compact and $(z(t), \dot{z}(t)) \in G$, there is a positive number $C = C(G)$ such that $\dot{W}(z(t), \dot{z}(t)) \leqq C$ for $t > t_0$.

Therefore Theorem 3 implies $\Omega \subset \Gamma = \{(x, y) \in R^2 | yk(x, y) = 0\}$, where Ω is the ω-limit set of $(x(t), y(t))$. We must have $\Omega \cap R_x \neq \varnothing$, because otherwise it would follow that $|x(t)| \to \infty$ as $t \to \infty$, a contradiction.

Consider, then, the two possibilities:

(a) $(0, 0) \in \Omega$, and

(b) $\Omega \subset \Gamma - \{(0, 0)\}$ and hence $\Omega \cap (R^+ \cup R^-) \neq \varnothing$.

Case (a) implies, taking into account Lemma 1, that $(x(t), \dot{x}(t)) \to (0, 0)$ as $t \to \infty$.

Case (b) implies, by using (H_2), that $\Omega \subset R^+ \cup R^-$ and consequently $\dot{x}(t) \to 0$ as $t \to \infty$.

In case $\liminf_{t \to \infty} |x(t)| = 0$, we have that $(0, 0) \in \Omega$ and, consequently, $(x(t), \dot{x}(t)) \to (0, 0)$ as $t \to \infty$. The proof is complete.

Acknowledgment. The author is indebted to Professor Taro Yoshizawa for his invaluable comments on the original manuscript.

REFERENCES

[1] J. P. La Salle, *An invariance principle in the theory of stability*, Differential Equations and Dynamical Systems, Academic Press, New York, 1967, pp. 277–286.

[2] J. J. Levin, *On the global asymptotic behavior of nonlinear systems of differential equations*, Arch. Rational Mech. Anal., 6 (1965), pp. 65–74.

[3] R. K. Miller, *Asymptotic behavior of nonlinear delay-differential equations*, J. Differential Equations, 1 (1965), pp. 293–305.

[4] N. Onuchic, *Stability properties of a second order differential equation*, Acta Mexicana de Ciencia y Tecnologia, 3 (1969), pp. 6–11.

[5] A. Strauss and J. A. Yorke, *On asymptotically autonomous differential equations*, Math. Systems Theor., 1, 2 (1967), pp. 175–182.

[6] T. Yoshizawa, *Asymptotic behavior of solutions of a system of differential equations*, Contributions to Differential Equations, 1 (1963), pp. 371–387.

[7] ———, *Stability theory by Lyapunov's second method*, J. Math. Soc. Japan, 1966.

SIAM J. CONTROL
Vol. 9, No. 1, February 1971

THE GENERATION OF LYAPUNOV FUNCTIONS FOR INPUT-OUTPUT STABLE SYSTEMS*

JAN C. WILLEMS †

Abstract. This paper discusses the relationship between properties of input-output descriptions and state space models for dynamical systems. It is shown that a state space realization of an input-output stable dynamical system is globally asymptotically stable in the sense of Lyapunov if it is uniformly observable and if every state is reachable. This result is proved in the context of abstract dynamical systems and leads to the equivalence of input-output stability and asymptotic stability for uniformly controllable and uniformly observable linear finite-dimensional systems. The generation of Lyapunov functions is subsequently considered, and variational techniques for the construction of Lyapunov functions are presented. Passivity and related energy concepts are particularly exploited in this context. These results yield the Lyapunov functions used in the proofs of the circle criterion and the Popov criterion as particular cases. The generality of the approach, however, makes these ideas applicable to much more general situations. Examples illustrating the results and the unifying point of view are included.

1. Introduction. "Dynamical systems" as they are studied and defined[1] in modern system theory distinguish themselves from arbitrary operators in mathematics by one basic property: they are causal, i.e., nonanticipatory: future values of the input do not influence past values of the output. This basic realizability property of physical systems may be incorporated in the mathematical model in two ways: either by appropriately restricting the operator defining the input-output relationship, or by working with a state space description which will then automatically ensure this causality. This last approach has proved particularly useful in optimal control theory due to the fact that any deterministic optimal controller can always be implemented with a memoryless function of the state in the feedback. It is therefore very advantageous to work with a state space model from the very start.

This duality in the possible description of systems has reflected itself in other areas of system theory and is particularly prevalent in stability theory. The input-output approach leads to the concept of input-output stability and has been developed mainly in the last decade, especially following the work of Sandberg [4] and Zames [5]. The state space description leads to concepts such as global asymptotic stability in the sense of Lyapunov and poses the stability problem in a setting which does not involve inputs, thus making use of the theory of classical dynamical systems. Which of the two approaches is to be preferred depends on

* Received by the editors January 20, 1970.

† Electronic Systems Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. Now at Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, England. This research was supported by the National Aeronautics and Space Administration under Contract NGL-22-009-124 and by the National Science Foundation under Grant GK-14152.

[1] There appears to be no agreement as to the use of the term "dynamical system." For the purpose of this paper any causal input-output relation will be termed a dynamical system. It will be shown that this is equivalent to the existence of a state. Zadeh [1] and Balakrishnan [2] appear to reserve the term for systems in which the state evolution is governed by a differential equation. The dynamical systems studied in classical mechanics [3] correspond to the state evolution equations in the absence of inputs.

the particular application. (For a discussion of these issues, see the survey paper by the author [6].) It has become clear, however, that the input-output approach leads to more powerful and general results. There are, in fact, a number of interesting stability criteria available which have been obtained in an input-output stability setting, but for which no proofs using Lyapunov methods exist.

Notwithstanding this success of input-output stability, there are certain aspects of Lyapunov stability theory which make the study and development of this "internal" approach to stability theory both useful and important. Not the least of these advantages is the possibility of obtaining estimates on the domain of attraction of an equilibrium in the case of nonglobal stability, a concept which has not even been satisfactorily formulated, let alone developed in the context of input-output stability. The main stumbling block in applying Lyapunov methods to the stability analysis particularly of nonlinear systems remains the absence of general methods for the construction of Lyapunov functions. This paper in part addresses itself to this problem.

The paper is concerned with the implications of input-output stability to global stability of dynamical systems, and with the construction of Lyapunov functions for input-output stable systems. The converse question, i.e., the implications of global stability to input-output stability, will not be considered, in view of space limitations and in view of the fact that such implications are much easier to obtain.

The first part of the paper introduces the concept of a dynamical system, which is defined as a causal operator between signal spaces, and the concept of a realization in which state space concepts become relevant. Some important properties of dynamical systems and realizations are then introduced: they are those of stability, controllability, observability, reachability, connectedness, and irreducibility. These notions play an important role in the sequel.

The second part of the paper discusses the generation of Lyapunov functions for input-output stable systems. Particular emphasis is placed on passive systems and on concepts such as available energy, required energy, and cyclic energy, the latter of which is very reminiscent of certain notions in thermodynamics.

The third part of the paper is concerned with feedback systems. Feedback systems are very important in control, and their stability is, of course, the main qualification on the performance of a feedback structure as a controller. Moreover, for design purposes, it is extremely desirable that properties of feedback systems be concluded from considerations of the open-loop elements. This aspect makes the results of the previous sections not easily applicable to the analysis of feedback systems, and a somewhat different approach is thus required. The ensuing Lyapunov functions are defined in terms of variational problems.

The paper ends with a list of examples. They illustrate the viewpoint adopted here and lead to the Lyapunov functions used to prove the circle criterion and the Popov criterion.

The work reported here has been directly inspired by a very interesting paper by Baker and Bergen [7] which appeared recently. They indeed posed the problem of constructing Lyapunov functions as a variational problem, an approach which has been fully exploited in the context presented here. Some of these ideas already appeared in the work of Popov [8], Kalman [9], and Anderson [10], [11]. The

author obtained a great deal of insight from the work of Brockett on passivity and stability [12]. It is interesting to note that the ingenious independence of path argument as exploited by the latter author in his construction of Lyapunov functions follows here as a rather logical consequence of the variational problems which lead to the desired Lyapunov functions.

The paper also indicates what may be the basic reason why stability conditions appear to be easier to obtain using input-output methods than through the construction of suitable Lyapunov functions: both input-output stability and Lyapunov stability can be posed as variational minimization problems, and whereas Lyapunov methods need the explicit solution of these variational problems (thus the boundedness *and* the value of an infimum), input-output stability only requires the boundedness of this infimum. This observation is due to Zames (private communication).

2. Dynamical systems. A dynamical system is usually defined on a subset of the real line as a mapping between function spaces satisfying an appropriate set of axioms. This paper will be concerned with continuous time systems only. Moreover, it will be assumed that the inputs and the outputs take their values in appropriate inner product spaces and that their norm is a locally square integrable function of time. This restriction precludes a certain amount of generality and is made mainly for expository purposes since the results of the paper generalize to much more general situations. In particular, the assumption that the input and output spaces are inner product spaces is of no consequence to many of the results in the paper. One of the reasons for treating systems in this setting is the possibility of introducing and exploiting concepts related to energy and passivity of systems. Indeed, these have far-reaching implications in stability theory.

There are two main avenues for obtaining mathematical models of systems: the first one starts with an internal model in which physical laws and interconnections are used to describe the dynamics and which then yield the relation between the influence variables (the inputs) and the variables of interest (the outputs). The second approach starts with an input-output relation as the basic mathematical model to be used. Such a model is usually the logical consequence from identification experiments at the input-output terminals.

Besides inputs and outputs there is an additional set of variables which is of fundamental importance in the description of dynamical systems. These are the so-called states which summarize the effect of past inputs. The internal modeling approach, in fact, usually displays a state explicitly. More often than not the state has no immediate physical significance and there is never any uniqueness as to its choice. Although the basic mechanism of interest in system theory is the generation of outputs from inputs, it is very often advantageous, however, to view this process as taking place through this intermediate variable, the state. This point of view has been particularly useful in such fields as dynamic optimization theory and the study of Markov processes.

These concepts are formally introduced in the present section, and it is shown that input-output descriptions and state space descriptions of dynamical systems are essentially equivalent.

The first notion is that of signal spaces which will be the input and output function spaces.

Let V be an inner product space and let R denote the real line. Let f be a V-valued function defined on R. Then the *causal truncation* of f at T is defined to be the result of the projection operator P_T defined by

$$(P_T f)(t) \triangleq \begin{cases} f(t) & \text{for } t \leqq T, \\ 0 & \text{otherwise.} \end{cases}$$

The *anticausal truncation* of f at T is defined as $Q_T f = f - P_T f$. Consider now the vector space of V-valued functions on R with $\int_{-\infty}^{+\infty} \|f(t)\|_v^2 \, dt < \infty$. This vector space is itself an inner product space with

$$\langle f_1, f_2 \rangle = \int_{-\infty}^{+\infty} \langle f_1(t), f_2(t) \rangle_V \, dt$$

as the inner product. It will be denoted by $L_2(V)$ and is complete if V is. As usual, no attention is paid to the fact that a function in $L_2(V)$ actually represents the equivalence class of functions which are equal to it almost everywhere with respect to Lebesgue measure.

As Wiener remarked when extending Fourier transforms, $L_2(V)$ is not a very interesting class of functions since it consists of functions which were small in the remote past and are destined to become small in the remote future. This last aspect, in particular, makes this function space of very limited use in stability studies which precisely refer to this remote future, and any a priori limitations on the future would therefore be very inappropriate.

A useful extension of $L_2(V)$ is its so-called *causal extension* denoted by $L_{2e}(V)$, which consists of all V-valued functions on R whose causal truncations belong to $L_2(V)$, i.e.,

$$L_{2e}(V) \triangleq \{f : R \to V | P_T f \in L_2(V), \textit{ all } T \in R\}.$$

The *anticausal extension* of $L_{2e}(V)$ is similarly defined as

$$\{f : R \to V | Q_T f \in L_2(V), \text{ all } T \in R\}.$$

Since all time functions considered in this paper will be assumed to start at some finite time, very little use of this anticausal extension will be made.

DEFINITION 1. Let $L_{2e}(V)$ denote the causal extension of $L_2(V)$. Then the subspace of $L_{2e}(V)$ defined by

$$S(V) \triangleq \{f \in L_{2e}(V) | Q_T f = 0 \text{ for } \textit{some } T \in R\}$$

will be called a *signal space*. Elements of $S(V)$ will be called *signals*, and elements of $L_2(V) \cap S(V)$ will be called *small signals*.

Thus, signal spaces consist of functions which vanish in the remote past and which have, in a sense, no finite escape, but are otherwise quite arbitrary. As is customary in the related literature, it will be assumed that inputs are applied to systems starting at some finite time in the past. This time need not be a priori fixed and will, in general, be different for each experiment. Note that signal spaces are closed under concatenation and that any "reasonable" physical signal belongs to a signal space. For the purposes of this paper, signal spaces represent a very convenient abstraction of reality. The fact that signals are required to have their

support on a half-line is very important and results in mild conditions for the well-posedness of mathematical models. In other words, given a mathematical model for a system (e.g., an integral equation or a differential equation), it will, in general, be relatively easy to establish that inputs in a given class generate well-defined outputs, provided, however, that these inputs have their support on a half-line. For inputs defined on the whole real line $(-\infty, +\infty)$, establishing this existence and uniqueness of outputs usually leads to stringent conditions and requires typically input-output continuity of the system. This case is moreover of dubious physical significance. For details see [13], [14, § 4.6], [15].

Let $U = S(V_u)$ and $Y = S(V_y)$ be signal spaces. U will be called the *input space*, and Y will be called the *output space*. Elements of U and Y will be called respectively *input signals* and *output signals*. A mapping F from U into Y is said to be *causal* (or *nonanticipatory*) if for all $T \in R$ and all $u_1, u_2 \in U$ with $P_T u_1 = P_T u_2$, the equality $P_T F u_1 = P_T F u_2$ holds.[2] This condition is equivalent to requiring that $P_T F P_T = P_T F$ on U.

Note that the signal spaces as introduced above could have been called somewhat more consistently causal signal spaces. The analogous concepts of anticausal signal spaces and anticausal operators thus become straightforward. No use will be made of these concepts, however. An additional notion which is of some importance is that of a *memoryless* operator. This would most logically be defined as an operator which is *both* causal *and* anticausal but is easiest (although equivalently) defined as an operator F, defined by an element $r \in Y$ and an instantaneous map, f, from $V_u \times R$ into V_y with $f(0, t) = 0$ for all $t \in T$ and $Fu \triangleq r + Nu$, where $(Nu)(t) \triangleq f(u(t), t)$ is such that any function $u \in L_2(V_u)$ with compact support yields $Nu \in L_2(V_y)$ (consequently also with compact support).

DEFINITION 2. A *dynamical system* is defined as a causal mapping from the input signal space U into the output signal space Y. If this mapping is memoryless, then the dynamical system will similarly be called memoryless.

The above setting for the study of input-output relations is similar to the one employed by Balakrishnan in [16]. The definition eliminates the possibility of studying differentiators, for instance, but for the purposes of this paper (stability) such a restriction is not very disturbing. In the study of networks, however, one clearly wants a more general definition which admits singularity functions in the impulse response. Zemanian [17] and Balakrishnan [18] have studied systems in which the inputs are assumed to be infinitely smooth functions and the outputs are distributions. Extended spaces appeared first in the context of stability theory as a result of the work of Sandberg [4] and Zames [5].

For many purposes, it is convenient to impose some smoothness conditions on the operators in question. Note that U and Y have, as signal spaces, no topology since they are, although derived from normed spaces, not normed themselves. However, causality enables one nevertheless to make a suitable definition of local continuity. Although simple continuity is the most logical smoothness condition to impose, it is very often advantageous to require somewhat stronger conditions, more specifically Lipschitz continuity. Recall that a (in general nonlinear) map F between normed spaces is said to be *Lipschitz continuous* if there

[2] Note the abuse of notation in the fact that the symbol P_T is used to denote an operator on U *and* an operator on Y. This ambiguity, however, causes no difficulty.

exists a real constant $K < \infty$ such that for all x_1, x_2 in the domain of F, $\|Fx_1 - Fx_2\| \leqq K\|x_1 - x_2\|$. Let F be a causal map from the input space U into the output space Y. Then F is said to be *locally Lipschitz continuous* if for all $t_0, t_1 \in R$, $P_{t_1}FQ_{t_0}$ is Lipschitz continuous as a map from $L_2(V_u)$ into $L_2(V_y)$.

DEFINITION 3. A dynamical system is said to be *smooth* if the defining map G is locally Lipschitz continuous. It is said to be *uniformly smooth* if for any given $T > 0$, $P_{t+T}GQ_t$ is Lipschitz continuous uniformly in t.

Convention. For convenience it will be assumed that all dynamical systems under consideration are *unbiased*, i.e., that they map the zero input into the zero output; hence $G0 = 0$. This absence of a bias term can always be obtained by a trivial redefinition of G and assumes, for instance, that for memoryless operators the element $r \in Y$ appearing in the definition is the zero element.

Now that the definition of input-output models of dynamical systems has been introduced, attention is focused on the formalism for the state space description of dynamical systems. Let R_2^+ denote the *causal sector* of R_2 defined as $R_2^+ \triangleq \{(t_2, t_1) | t_2, t_1 \in R, t_2 \geqq t_1\}$.

DEFINITION 4. A (mathematical model of a) dynamical system is said to be in *state space form* if it is determined by an abstract set X (the *state space*) and two maps, ϕ, the *state transition map*, and y, the *output reading map*, satisfying the following axioms:

(i) ϕ maps $R_2^+ \times X \times U$ into X;

(ii) (*Causality*): $\phi(t, t_0, x_0, u) = \phi(t, t_0, x_0, P_tQ_{t_0}u)$ for all $(t, t_0) \in R_2^+$, $x_0 \in X$, and $u \in U$;

(iii) (*Consistency*): $\phi(t_0, t_0, x_0, u) = x_0$ for all $t_0 \in R$, $x_0 \in X$, and $u \in U$;

(iv) (*Composition law or semi-group property*): $\phi(t_2, t_0, x_0, u) = \phi(t_2, t_1, \phi(t_1, t_0, x_0, u), u)$ for all $(t_1, t_0), (t_2, t_1) \in R_2^+$, $x_0 \in X$, and $u \in U$;

(v) y maps $R \times X \times V_u$ into V_y and the value of the output at time t is given by $y(t, x(t), u(t))$;

(vi) X is a subset of an inner product space V_x;

(vii) (*Unbiasedness*): $\phi(t, t_0, 0, 0) = 0$ for all $(t, t_0) \in R_2^+$, and $y(t, 0, 0) = 0$ for all $t \in R$;

(viii) Let $\tilde{X}$ denote the signal space induced by V_x (i.e., $\tilde{X} = S(V_x)$); it is then assumed that the functions

$$x(t) = Q_{t_0}\phi(t, t_0, x, u) = \begin{cases} \phi(t, t_0, x, u) & \text{for } t \geqq t_0, \\ 0 & \text{otherwise,} \end{cases}$$

$$Q_{t_0}y(t, x(t), u(t)) = \begin{cases} y(t, x(t), u(t)) & \text{for } t \geqq t_0, \\ 0 & \text{otherwise,} \end{cases}$$

belong to $\tilde{X}$ and Y respectively for all $t_0 \in R$, $x_0 \in X$, and $u \in U$.

Axioms (i)–(v) are the usual axioms involved in describing dynamical systems from a state space point of view. Axiom (vi) induces a topology on the state space and will be needed in the definitions of Lyapunov stability, for instance.[3] Axiom

[3] The assumption that X is an inner product space is restrictive and inconvenient for many applications, more so than would appear at first sight. For a study of dynamical systems whose state space is a group manifold see [26].

(vii) is in keeping with the unbiasedness convention introduced above, and axiom (viii) guarantees the absence of a finite escape and results in the fact that locally square integrable inputs produce locally square integrable outputs.

A dynamical system in state space form thus views the generation of outputs from inputs as occurring through the composition of two maps, G_x and G_y, with $G_x\colon U \to X$ and $G_y\colon X \times U \to Y$. The map G_x is a dynamical system in its own right (with X viewed as the output space) but satisfies a richer set of axioms than merely those implied by the dynamical system axioms: in addition, it is required that this map have the Markov property which results in a decoupling of the past from the future in the sense that the present value of the state has sufficient information in it so as to summarize the effect of past inputs. The state space thus represents an adequate memory-bank. The map G_y is memoryless (with input space $X \times U$) and the dependence of $y(t)$ on past values of u is obtained through the dependence on $x(t)$. It is a simple matter to verify that the composite system $y = G_y(G_x u, u)$ is indeed a dynamical system in the input-output sense. Note also that $G_x 0$ satisfies the axioms for dynamical systems without inputs as studied in classical mechanics and its extensions.

The next definitions refer to the smoothness of the state transition map and the output reading map. These smoothness conditions are generally quite important aspects of a particular dynamical system in state space form. For instance, it can be shown that otherwise any finite-dimensional dynamical system can be realized by a one-dimensional dynamical system if this latter is not required to have any smoothness. It suffices therefore to consider a one-to-one map from R^n into R and appropriately modify the state space and the maps defining the dynamical system.

DEFINITION 5. A dynamical system in state space form is said to be *smooth* if for any $(t_1, t_0) \in R_2^+$ there exist $K_1, K_2, K_3, K_4 < \infty$ such that

$$\|P_{t_1}Q_{t_0}(\phi(t, t_0, x_1, u_1) - \phi(t, t_0, x_2, u_2))\| \leqq K_1\|x_1 - x_2\| + K_2\|P_{t_1}Q_{t_0}(u_1 - u_2)\|$$

for all $x_1, x_2 \in X$ and $u_1, u_2 \in U$, and

$$\|P_{t_1}Q_{t_0}(y(t, x_1(t), u_1(t)) - y(t, x_2(t), u_2(t)))\| \leqq K_3\|P_{t_1}Q_{t_0}(x_1 - x_2)\| + K_4\|P_{t_1}Q_{t_0}(y_1 - u_2)\|$$

for all $x_1, x_2 \in \tilde{X}$ and $u_1, u_2 \in U$.

It is said to be *uniformly smooth* if for any $T \geqq 0$ and $t_1 = t_0 + T$ in the above inequalities, the constants K_1, K_2, K_3, K_4 may be taken independent of t_0.

These definitions are entirely analogous to those imposed for input-output systems. It is a simple matter to verify that (uniformly) smooth dynamical systems in state space form define (uniformly) smooth input-output dynamical systems. It is also clear that uniform smoothness and smoothness are equivalent for time-invariant systems.

The final discussion of this section involves the relationship between the above definitions of dynamical systems. As might be expected they are indeed equivalent.

DEFINITION 6. Consider the dynamical system G and a dynamical system in state space form with defining maps ϕ and y. Then the dynamical system in state space form is said to be a *realization* of G if any $u \in U$ with $t_0 \in R$ such that $P_{t_0}u = 0$, yields $(Gu)(t) = y(t, \phi(t, t_0, 0, u), u(t))$ for all $t \in R$. The dynamical system in state space form thus defines the same input-output relation as G.

THEOREM 1. *Every (smooth, uniformly smooth) dynamical system has a (smooth, uniformly smooth) realization in state space form.*

Proof. The proof proceeds by construction. The state space $X = V_x$ will be taken to be the collection of all functions in $L_2(V_u)$ with compact support, and the state at time t will be taken to be $P_t u$, where $u \in U$ is the input to the system. Thus, for instance,

$$\phi(t, x_0, t_0, u) = S_t P_t u \stackrel{\Delta}{=} x_0 + S_t P_t Q_{t_0} u,$$

where S_T denotes the shift operator $(S_T z)(t) = z(t - T)$, and

$$y(t, x, u) = (Gu)(t) = (GP_t u)(t) = (GS_{-t}S_t P_t u)(t) = \stackrel{\Delta}{=} (GS_{-t}x(t))(t).$$

It is left to the reader to verify that these maps indeed satisfy the axioms of dynamical systems in state space form. The smoothness claims are also easily verified directly in view of the simplicity of the state transition map and the fact that the output map and the original dynamical system are essentially identical.

The above theorem, although too trivial and general to be of significance in specific instances, yields a rather interesting canonical decomposition of nonlinear dynamical systems into a linear, time-invariant, reachable dynamical part followed by a memoryless nonlinear part.[4] Note also that the dynamical part in this decomposition may be described by the partial differential equation

$$\partial x(z, t)/\partial t = \partial x(z, t)/\partial z, \qquad z \leqq 0,$$

with the boundary control $x(0, t) = u(t)$ and with solutions defined in an appropriate sense. The function $x(z, t)$ for $z \geqq 0$ then plays the role of the state at time t, and the partial differential equation describes the evolution of the initial state $x(z, t_0)$ resulting from the input $u(t)$. Notice also that in the above realization the map y inherits linearity and time-invariance of G.

It should be noted that the equivalence of a dynamical system and a state space realization of a dynamical system might nevertheless lead the latter model to produce an output which could not be the result of any input to the former model. Such outputs result from initial states which are somewhat artificial in the sense that they cannot be produced by past inputs. The equivalence of a dynamical system and one of its realizations is thus really a zero initial state equivalence.

Notation. Let G denote a dynamical system in state space form, $x_0 \in X$, $t_0 \in R$, and $u \in U$. Then the function defined by

$$\begin{cases} y(t, \phi(t, t_0, x_0, u), u(t)) & \text{for } t \geqq T \geqq t_0, \\ 0 & \text{otherwise,} \end{cases}$$

[4] It was pointed out to the author that similar decomposition due to Wiener [19] and Balakrishnan [16] have appeared in the literature.

will be denoted by $Q_T G(t_0, x_0, u)$ (or $Q_T y(t_0, x_0, u)$ when there is no danger of confusion).

3. Fundamental properties of dynamical systems. A number of fundamental concepts related to dynamical systems and their state space realizations are introduced and discussed in this section: they relate to the influence of the control on the state (reachability, controllability, and connectedness), of the state on the output (observability and irreducibility), and of the input on the output (stability and continuity).

DEFINITION 7. The state space of a dynamical system in state space form is said to be *reachable* if given any $x \in X$ and $t \in R$, there exists a $t_0 \in R$, $t_0 \leqq t$, and a $u \in U$ such that $\phi(t, t_0, 0, u) = x$. A dynamical system in state space form is said to be *controllable* if given any $x_0 \in X$ and $t_0 \in R$, there exists a $t \in T$, $t \geqq t_0$, and a $u \in U$ such that $\phi(t, t_0, x_0, u) = 0$. The state space of a dynamical system in state space form is said to be *connected* if given any $x_0, x_1 \in X$ there exists an element $(t_1, t_0) \in R_2^+$ and a $u \in U$ such that $\phi(t_1, t_0, x_0, u) = x_1$.

Reachability thus requires the map $\phi(t, \cdot, 0, \cdot)$ to be onto X, whereas controllability requires that 0 be in the range space of $\phi(\cdot, t_0, x_0, \cdot)$. Note that reachability, controllability, and time-invariance imply connectedness.

DEFINITION 8. A dynamical system in state space form is said to be *observable* if for any $t_0 \in R$, knowledge of $Q_{t_0} y(t_0, x_0, 0)$ (uniquely) determines $x_0 \in X$. The state space of a dynamical system in state space form is said to be *irreducible* if for any given $t_0 \in R$ and $x_0 \in X$ there exists a $Q_{t_0} u \in U$, such that knowledge of $Q_{t_0} y(t_0, x_0, u)$ (uniquely) determines $x_0 \in X$.

Observability thus requires the map $y(t_0, \cdot, 0)$ to be one-to-one on X, whereas irreducibility requires the map $y(t_0, \cdot, u)$ to be one-to-one on X by choosing $(Q_{t_0} u)(x)$. It is clear that observability implies irreducibility and that the nomenclature "irreducible" is quite appropriate since if the state space is not irreducible, then there exist at least two initial states which will be completely indistinguishable under experimentation: these two states are thus entirely equivalent, and nothing will be lost by eliminating one of them from the state space.

The above nomenclature is common (although far from standard) in the related literature with the possible exception of irreducibility which is often taken to indicate the set-theoretic minimality of the state space. Observability and irreducibility are equivalent for linear systems with a finite-dimensional smooth state space realization. The simplest example of systems in which these concepts are different are systems with multiplicative control described, e.g., by $\dot{x} = uAx$. It should also be remarked that the above definitions, although natural, are not the most convenient ones for certain applications. Although it can be shown that every dynamical system has a realization[5] with a reachable state space, it is sometimes very difficult to discover exactly what states are reachable (and to define X then appropriately). For instance, in systems described by partial differential equations these reachable states have certain smoothness properties which are not a priori known; therefore, in certain applications it is much more convenient to adopt an "almost" reachability requirement. The same remark holds for the following

[5] See, for example, the proof of Theorem 1.

definitions which, in addition, require an appropriate choice of the topology on the state space.

DEFINITION 9. The state space of a dynamical system in state space form is said to be *uniformly reachable* if there exist a continuous function $\alpha: R^+ \to R^+$ (R^+ denotes the nonnegative real numbers) with $\alpha(0) = 0$ and a constant $T \geqq 0$, such that for any $x \in X$ and $t \in R$, there exists a $u \in U$ with $\|P_t u\|^2 \leqq \alpha(\|x\|)$ such that $\phi(t, t - T, 0, u) = x$. *Uniform controllability* and *uniform connectedness* are similarly defined. A dynamical system is said to be *uniformly observable* if there exist a strictly monotone increasing continuous function $\beta: R^+ \to R$ with $\beta(0) = 0$ and $\lim_{\sigma \to +\infty} \beta(\sigma) = +\infty$ and a constant $T \in R$, $T \geqq 0$, such that for any $x \in X$ and $t_0 \in R$,

$$\|P_{t_0+T} Q_{t_0} y(t_0, x, 0)\|^2 \geqq \beta(\|x\|).$$

The state space of a dynamical system in state space form is said to be *uniformly irreducible* if with β and T as before, the inequality

$$\|P_{t_0+T} Q_{t_0}(y(t_0, x_1, u) - y(t_0, x_2, u))\|^2 \geqq \beta(\|x_1 - x_2\|)$$

holds for all $x_1, x_2 \in X$, $t_0 \in R$, and some $u \in U$.

The above definitions differ somewhat from those in the literature. Most of the papers concerned with uniform controllability for linear systems follow Kalman's [20] original definition, which imposes many more restrictions than the definitions used here. In particular, it requires any control which makes the transfer from state 0 at time $t - T$ to state x at time t to be such that $\|P_t Q_{t-T} u\|^2 \geqq \alpha(\|x\|) > 0$.

The most efficient realization of a dynamical system is one in which the state space is reachable *and* irreducible. This indeed guarantees that every output which can be observed as a result of initial conditions and inputs could have been observed by properly choosing the past input and that two different initial conditions will lead to different outputs by properly choosing the input. Two realizations which are both reachable *and* irreducible are thus isomorphic. They differ in the sense that their state spaces are labeled differently. The one-to-one onto map between these state spaces may, in general, be a function of time, however. A realization of a dynamical system in which the state space is reachable and irreducible can thus properly be called *minimal*, a notion which has many more substantive implications for linear systems. In looking for reachable *and* irreducible realizations it is natural to consider as the candidate for the state space the equivalence classes of those inputs up to time t which yield the same output after time t, regardless of the input after time t; more precisely, by considering the equivalence class $\{P_t u |\ Q_t y$ is fixed for all $Q_t u\}$ as a typical element of the state space. The difficulty with this representation is that, in general, the state space itself then becomes a function of time. There are two methods of getting around this difficulty: one is to modify the original axioms and definitions so as to allow for a state space which is itself a function of time; the other is to define a dynamical system as a causal *and* a noncausal map depending on whether one considers time moving forward or backward from the initial time. The state is thus alternatively required to summarize past and future, and the state space thus has many more invariant properties with respect to time. This device has been used successfully by Kalman

[21] and others in their study of systems described by the Volterra integral equation

$$y(t) = \int_{t_0}^{t} w(t, \tau)u(\tau)\, d\tau$$

with separable kernel w. This principle rests on dubious physical grounds, however, and leads to technical difficulties for infinite-dimensional systems. The above problems do not occur in stationary systems.

Recall that a mapping between normed spaces is said to be *bounded* if it maps bounded sets into bounded sets. It is said to have a *finite gain* if there exists a $K < \infty$ such that for any $\rho \geqq 0$ the ball with radius ρ gets mapped into the ball with radius $K\rho$. The infimum of all real numbers K achieving the above inequality is usually called the *gain* of the operator.

DEFINITION 10. A dynamical system, G, is said to be *input–output stable* if it maps bounded sets of small signals in U into bounded sets of small signals in Y. It is said to be *finite-gain input-output stable* if it is stable and if there exists a $K < \infty$ such that for any small signal $u \in U$, $\|Gu\| \leqq K\|u\|$. The infimum of all such real numbers K will be denoted by $\|G\|$. A dynamical system G is said to be *input-output continuous* if it is stable and if the map G is continuous (in the topology induced by $L_2(V_u)$ and $L_2(V_y)$) as a map from $U \cap L_2(V_u)$ into $Y \cap L_2(V_y)$. It is said to be *input-output Lipschitz continuous* if G is actually Lipschitz continuous.

It can be shown [14, § 2.4] that a dynamical system is finite-gain stable if and only if the gain of $P_{t_1}GQ_{t_0}$ is bounded for all $t_0, t_1 \in R$, uniformly in t_0 and t_1. In fact,

$$\|G\| = \lim_{-t_0, t_1 \to \infty} \|P_{t_1}GQ_{t_0}\|,$$

and this limit is approached monotonically. A similar relationship holds for Lipschitz continuity.

Related, but not identical, are the following more familiar Lyapunov stability concepts for dynamical systems in state space form.

DEFINITION 11. The equilibrium state of a dynamical system in state space form is said to be *globally attractive* if for any $x_0 \in X$ and $t_0 \in R$,

$$\lim_{T \to \infty} \phi(t_0 + T, t_0, x_0, 0) = 0.$$

It is said to be *uniformly globally attractive* if this limit is uniform in t_0. It is said to be *stable* if for any $\varepsilon > 0$ and $t_0 \in R$ there exists a $\delta(\varepsilon, t_0)$ such that $\|\phi(t_0 + T, t_0, x_0, 0)\| \leqq \varepsilon$ for all $T \geqq 0$ whenever $\|x_0\| \leqq \delta$. It is said to be *uniformly stable* if $\delta(\varepsilon, t_0)$ may be chosen independent of t_0. A dynamical system in state space form is said to be *bounded* if for any $x_0 \in X$ and $t_0 \in R$, $\phi(t_0 + T, t_0, x_0, 0)$ is bounded on the half-line $T \geqq 0$. It is said to be *uniformly bounded* if this bound may be chosen independent of t_0. A dynamical system in state space form is said to be *globally asymptotically stable* if the equilibrium state is globally attractive and stable. It is said to be *uniformly globally asymptotically stable* if the equilibrium state is uniformly globally attractive, uniformly stable, and uniformly bounded.

The usual method of proving stability of systems in state space form is to consider an appropriate Lyapunov function. The following definition of a Lyapunov function is a convenient one for the present discussion.

DEFINITION 12. Let V be a mapping from $X \times R$ into R^+, with $V(x, t) = 0$ if and only if $x = 0$. Then V is said to be a *Lyapunov function* for a dynamical system in state space form if for any $x_0 \in X$ and $t_0 \in R$,

(i) $V(\phi(t, t_0, x_0, 0), t)$ is a monotone nonincreasing function of t for $t \geqq t_0$;

(ii) $\lim_{T \to \infty} V(\phi(t_0 + T, t_0, x_0, t), t_0 + T) = 0$.

The function V will be called a *uniform Lyapunov function* if, in addition, the limit in (ii) is uniform in t_0 and if $V(x, t)$ is bounded in t for all $x \in X$. The function V is said to be *decrescent* if there exists a continuous function $\alpha: R^+ \to R^+$ with $\alpha(0) = 0$ such that $V(x, t) \leqq \alpha(\|x\|)$ for all $x \in X$ and $t \in R$. It is said to be *positive definite* if there exists a monotone increasing continuous function $\beta: R^+ \to R^+$ with $\beta(0) = 0$ such that $V(x, t) \geqq \beta(\|x\|)$ for all $x \in X$ and $t \in R$. It is said to be *radially unbounded* if there exists a continuous function $\gamma: R^+ \to R^+$ with $\lim_{\sigma \to +\infty} \gamma(\sigma) = +\infty$ such that $V(x, t) \geqq \gamma(\|x\|)$ for all $x \in X$ and $t \in R$. If V is a Lyapunov function for a dynamical system in state space form, then the equilibrium state is globally asymptotically stable if V is positive definite, and the dynamical system is bounded if V is radially unbounded. If V is a uniform Lyapunov function, then the equilibrium state is uniformly globally attractive if V is positive definite, and uniformly stable if V is positive definite and decrescent; the system is uniformly bounded if V is radially unbounded, and uniformly globally asymptotically stable if V is radially unbounded, positive definite, and decrescent. Notice also that decrescency implies the last condition in the definition of a uniform Lyapunov function, and thus a decrescent Lyapunov function for a uniformly globally asymptotically stable system is a uniform Lyapunov function.

The main purpose of this paper is to study the relations between input-output stability and global stability. It seems reasonable to expect that an input-output stable system will be globally stable if inputs sufficiently influence states and if states sufficiently influence outputs. Then internal instability should reflect into external instability. That this can be made precise is shown in the next section.

4. Input-output stability and global stability. This section establishes the fundamental relationship between input-output stability and global stability. In trying to obtain these internal stability implications from external data, one defines certain functions which depend on the external variables only. In order for these functions to be well-defined and to qualify as suitable Lyapunov functions, a number of additional assumptions have to be made, and it is at this point that reachability, controllability, observability, and input-output stability become relevant. There are two natural functions to consider for this purpose:

(i) $V_r(x, t) \triangleq \inf \|P_t Q_{t_0} u\|^2$, where the infimum is to be taken over all $t_0 \leqq t$ and $u \in U$ with $\phi(t, t_0, 0, u) = x$ (the infimum (supremum) over the void set is by assumption $+\infty (-\infty)$), and

(ii) $V_0(x, t) \triangleq \|Q_t y(t, x, 0)\|^2$.

The symbolism is clear: the first function is inspired by reachability, and the second by observability. V_r is well-defined if the state space is reachable, and V_0 is well-defined if the state space is reachable and if the dynamical system is input-output stable. Indeed, let $u \in U$ be such that $\phi(t, t_0, 0, u) = x$. Then

$$\|Q_t y(t, x, 0)\|^2 \leqq \|G P_t u\|^2 \leqq \|G\|^2 \|P_t u\|^2.$$

If $V_0(x, t)$ is well-defined, then it is clearly monotone nonincreasing along undriven solutions and approaches zero when $t \to \infty$ since

$$V_0(\phi(t, t_0, x_0, 0), t) = \int_t^\infty \|y(\tau, \phi(\tau, t_0, x_0, 0), 0)\|_{V_y}^2 \, d\tau.$$

The function V_r is also monotone nonincreasing along undriven solutions, although this is not as immediate. The argument used in the demonstration of this fact will repeatedly be used in the sequel and will therefore only here be done explicitly. Thus, consider $V_r(\phi(t_1, x, t, 0), t_1)$ with $t_1 \geqq t$, and denote $\phi(t_1, x, t, 0)$ by x_1. Then $V_r(x_1, t_1) = \inf\|P_{t_1} Q_{t_0} u\|^2$. The state of the dynamical system can be driven to x_1 at time t_1 by first driving it from 0 at time t_0 to x at time t and then applying zero control until time t_1. This is, in general, a suboptimal control for reaching x_1 at time t_1, even when it is driven to x at time t in an optimal fashion. This suboptimal strategy thus shows that $V_r(x_1, t_1) \leqq V_r(x, t)$.

The basic relationships between input-output stability and global stability are stated in the following theorems.

THEOREM 2. *A uniformly observable realization of an input-output stable dynamical system with a reachable state space is globally asymptotically stable and bounded, and V_0 is a positive definite radially unbounded Lyapunov function for it.*

Proof. It suffices to show that V_0 is a positive definite radially unbounded Lyapunov function. By reachability and input-output stability, V_0 is well-defined. By observability, $V_0 = 0$ if and only if $x = 0$; it is monotone nonincreasing and approaches zero along undriven solutions since

$$\int_t^\infty \|y(\tau, \phi(\tau, t, x, 0), 0)\|_{V_y}^2 \, d\tau < \infty.$$

It remains to be shown that V is positive definite and radially unbounded. This, however, is an immediate consequence of uniform observability.

Note that in the above theorem the uniform observability condition cannot simply be relaxed to only observability, even if at the same time one assumes uniform reachability rather than merely reachability.

THEOREM 3. *A uniformly observable realization of a finite-gain input-output stable dynamical system with a uniformly reachable state space is uniformly globally asymptotically stable, and V_r and V_0 are positive definite radially unbounded decrescent uniform Lyapunov functions for it.*

Proof. From finite-gain input-output stability it follows that

$$\|P_t y(t, 0, u)\|^2 \leqq \|G\|^2 \|P_t u\|^2$$

and thus that

$$V_0(x, t) \leqq \|G\|^2 V_r(x, t).$$

V_r is well-defined and decrescent by uniform reachability, and V_0 is well-defined, positive definite, and radially unbounded by uniform observability. Thus both V_0 and V_r are positive definite, radially unbounded, and decrescent, and the theorem follows if it can be shown that $V_0(\phi(t_0 + T, t_0, x_0, 0), t_0 + T)$ approaches zero as $T \to \infty$, uniformly in t_0. Notice that with T as in the definition of uniform

observability it follows that

$$V_0(\phi(t_0 + T, t_0, x_0, 0), t + T) | V_0(x_0, t_0) \leqq 1 - \frac{\beta(\|x_0\|)}{\|G\|^2 \alpha(\|x_0\|)},$$

where α and β are as in the definitions of uniform reachability and uniform observability. This implies uniform convergence of the Lyapunov function V_0 since for any $\varepsilon > 0$ and $M < \infty$ there exists a $\delta > 0$ such that $\beta(\sigma)|\alpha(\sigma) \geqq \delta > 0$ for all $0 < \varepsilon \leqq \sigma \leqq M < \infty$. Hence the system is uniformly globally asymptotically stable, which in turn implies that V_r is a uniform Lyapunov function.

Note that finite-gain stability (or uniform smoothness) and uniform observability yield that every control u transferring state 0 at time t_0 to state x at time t requires

$$\|P_t Q_{t_0} u\|^2 \geqq \|G\|^{-2} \beta(\|x\|),$$

a condition which is usually part of the definition of uniform controllability [20]. As a final remark in this section, note that the fact that the inputs and outputs take their values in inner product spaces is inessential and that the results hold, mutatis mutandis, if these spaces are merely normed spaces. The inner product structure becomes very important in the next section, which is concerned with passivity.

5. Lyapunov functions for passive systems. The notions which will be introduced in this section are those of passivity and certain concepts related to energy. It will be assumed in this section that the inner product spaces under consideration are real.[6]

DEFINITION 13. Let $U = Y$ and let G be a dynamical system from U into Y. Then G is said to be *passive* if for all $u \in U$ and $t \in R$, $\langle P_t u, P_t G u \rangle \geqq 0$. It is said to be *strictly passive* if $G - \varepsilon I$ is passive for some $\varepsilon > 0$.

This terminology is to be interpreted as follows: $\langle u(t), y(t) \rangle_{V_u}$ represents the instantaneous power delivered to the system from the outside. Thus $\langle P_t u, P_t y \rangle$ represents the total energy at time t delivered to the system from the outside. If regardless of the termination and in the absence of initial excitations this energy is nonnegative, then the system is passive viewed from its input-output terminals.

It can be shown (see [14, § 2.17]) that if G is input-output stable, then it is passive if and only if $\langle u, Gu \rangle \geqq 0$ for all small input signals u.

DEFINITION 14. Let G be a dynamical system in state space form. Then the *required energy*, E_r, is defined on $X \times R$ as

$$E_r(x, t) \triangleq \inf \langle P_t u, P_t G u \rangle,$$

where the infimum is to be taken over all $t_0 \leqq t$ and $u \in U$ with $P_{t_0} u = 0$ which yield $\phi(t, t_0, 0, u) = x$. The *available energy*, E_a, is defined on $X \times R$ as

$$E_a(x, t) \triangleq \sup_{\substack{u \in U \\ t_1 \geqq t}} - \langle P_{t_1} Q_t u, P_{t_1} Q_t y(t, x, u) \rangle.$$

The *cycle energy*, E_c, is defined on $X \times R$ as $E_r - E_a$. Thus

$$E_c(x, t) = \inf \langle P_{t_1} u, P_{t_1} G u \rangle,$$

[6] Complex inner product spaces can be treated equally well by considering the real part of the inner product in the definitions of passivity and energy.

where the infimum is to be taken over all t_0 and t_1 with $t_0 \leqq t \leqq t_1$ and $u \in U$ with $P_{t_0}u = 0$, which yields $\phi(t, t_0, 0, u) = x$.

The available energy is thus the maximum energy which can be extracted from a system, whereas the required energy is the energy needed to excite a system to a given set of initial conditions. The cycle energy is the minimum energy it takes to cycle a system between the equilibrium and a given state. Note that all of the above energies are defined in terms of input-output relations.

LEMMA 1. *Consider a realization of a passive dynamical system and assume that the state space is reachable. Then E_a, E_r and E_c exist (i.e., E_a, E_r, $E_c < \infty$) and are nonnegative. Moreover, $0 \leqq E_a, E_c \leqq E_r$.*

Proof. That E_r and E_c are finite and nonnegative follows immediately from passivity and reachability. Hence, since $E_a + E_c = E_r$, $E_a \leqq E_r$.

It remains to be shown that E_a is nonnegative. This follows by considering $Q_t u = 0$, which shows that the supremum in the definition of E_c is taken over a set which contains zero. This completes the proof of the lemma.

The inequality $E_a \leqq E_r$ formalizes the intuitive notion that passive systems cannot supply more energy to the outside than has previously been supplied to them from the outside. Note that none of the above notions satisfactorily defines the *stored energy*, $E_s(x, t)$, which is an internal property of a dynamical system and thus usually a function of the realization. The passivity definition employed here is purely input–output. Similar definitions of internal passivity can be made, and the theory for linear time-invariant dissipative systems [22] is available. One can then pose the question of whether or not every input-output passive system has a passive realization. These ramifications fall beyond the scope of the present paper. It would be interesting to verify that the stored energy in a passive realization of a passive system satisfies the inequality $E_a \leqq E_s \leqq E_r$, as it should.

The cycle energy E_c is a measure of the degree of irreversibility of a system. This is the intuitive basis for the following definitions.

DEFINITION 15. A passive dynamical system in state space form is said to be *irreversible* if $E_c(x, t) = 0$ only if $x = 0$. It is said to be *uniformly irreversible* if there exists a monotone increasing function $\gamma : R^+ \to R^+$ with $\gamma(0) = 0$ and $\lim_{\sigma \to +\infty} \gamma(\sigma) = +\infty$ such that for all $x \in X$ and $t \in R$, $E_c(x, t) \geqq \gamma(\|x\|)$. It is said to be *reversible* if $E_c = 0$, i.e., if $E_r = E_a$.

THEOREM 4. *The available energy, E_a, and the required energy, E_r, are decrescent uniform Lyapunov functions for a uniformly observable realization of a passive finite-gain input-output stable dynamical system with a uniformly reachable state space.*

Proof. It will first be shown that E_r is decrescent. By the Schwarz inequality,

$$|\langle P_t u, P_t G u \rangle| \leqq \|P_t u\| \, \|P_t G u\| .$$

It thus follows from finite-gain stability that there exists a constant $K < \infty$ such that $E_r(x, t) \leqq K \inf \|P_t u\|^2$, where the infimum is to be taken over all $t_0 \leqq t$ and $u \in U$ with $P_{t_0}u = 0$, $\phi(t, t_0, 0, u) = x$. By uniform reachability, $E_r(x, t)$ is thus decrescent. It will now be shown that E_r is a Lyapunov function. That $E_r(x, t)$ is nonincreasing along undriven solutions follows from its definition and by letting $u = 0$ from t_0 until t_1, using an analogous argument to the one used in § 4 in

showing that V_r is monotone nonincreasing. By Theorem 3,

$$\phi(t_0 + T, t_0, x_0, 0) \to 0 \quad \text{as } T \to \infty,$$

uniformly in t_0, which by decrescence indeed implies that E_r is a uniform Lyapunov function. Since $E_a \leqq E_r$, it remains to be shown that E_a is monotone nonincreasing along undriven solutions. This follows from an analogous argument to the one used to show that E_r is nonincreasing. This completes the proof.

Theorem 4 is not as convincing as one might like it to be since it does not make any claims about the positive definiteness of the Lyapunov functions. This positive definiteness can be obtained using somewhat stronger hypotheses. The available energy E_a will be positive definite if the feedback system with the dynamical system G in the forward and some constant gain $k > 0$ in the feedback loop remains a well-defined dynamical system. This is the case under weak additional assumptions on G. The resulting control to be used to show definiteness of E_a is the solution e of the feedback equation $Q_t(e + kG(t, x, e)) = 0$. This corresponds to the input which results from a termination of the system with a positive resistor. The required energy E_r will be positive definite if E_a is or if the system is strictly passive (rather than merely passive). A third possibility is to require uniform irreversibility, since $E_c \leqq E_r$.

6. Feedback systems. One of the main reasons for being interested in stability stems from its importance in feedback control. The canonical form of the feedback system considered in this paper is shown in Fig. 1, and the closed loop system is thus described by the implicit equations

$$\text{(FE)} \qquad (I + G)e = u, \qquad y = Ge.$$

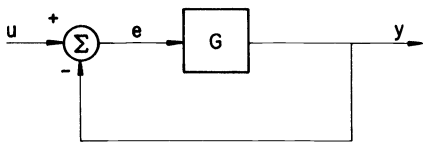


FIG. 1. *The feedback system under consideration*

It will be assumed that the input signal space U and the output signal space Y are the same and that G is a dynamical system from U into Y. Two questions related to the well-posedness of this feedback system arise: first, whether or not the closed loop feedback system still represents a well-defined dynamical system in its own right and, second, whether or not the state space induced by a realization of G will also qualify as the state space for the closed loop system. These issues fall beyond the scope of this paper, and it will be explicitly assumed rather that these well-posedness conditions are satisfied. It is thus assumed that:

(i) $(I + G)^{-1}$ exists (as a map from U into itself) and is causal. This implies that the closed loop system $G(I + G)^{-1} = I - (I + G)^{-1}$ is itself a dynamical system from the input space U into the output space Y.

(ii) If G is described in state space form with state space X, then X also qualifies as the state space for the closed loop dynamical system $G(I + G)^{-1}$, and a unique solution to the feedback equations exists for any initial condition $x_0 \in X$, initial time $t_0 \in R$, and input $u \in U$.

These well-posedness questions have been investigated in the literature [14, § 4.2], [15] and the simplest sufficient conditions essentially impose a restriction on the feedthrough in G, in addition to some smoothness conditions on the open loop system. They are satisfied in most models and, in particular, whenever G contains a pure or generalized delay.

In the study of feedback systems it is important to establish conditions on the open loop operator in order to draw conclusions about the closed loop system. The first questions thus answered are those related to controllability and irreducibility.

THEOREM 5. *Consider the feedback system described by equations* (FE). *Then reachability* (*controllability, connectedness, irreducibility*) *of the state space realization of the open loop system implies reachability* (*controllability, connectedness, irreducibility*) *of the associated state space realization of the closed loop system. Uniform reachability* (*controllability, connectedness, irreducibility*) *of the state space realization of the open loop system implies uniform reachability* (*controllability, connectedness, irreducibility*) *of the associated state space realization of the closed loop system provided the open loop system is in addition uniformly smooth.*

Proof. Preservation of reachability, controllability, or connectedness is essentially obvious. Indeed, let u_1 be a control which results in the desired transfer for the open loop system. Then the control $u = u_1 + Gu_1$ will clearly result in the same transfer for the closed loop system. Irreducibility of the closed loop system will be established by contradiction. Assume therefore that there exist $x_1, x_2 \in X$, $x_1 \neq x_2$, and $t_0 \in R$ such that

$$Q_{t_0}G(I + G)^{-1}(t_0, x_1, u) = Q_{t_0}G(I + G)^{-1}(t_0, x_2, u) \qquad \text{for all } u \in U.$$

This implies that

$$Q_{t_0}G(t_0, x_1, u_1) = Q_{t_0}G(t_0, x_2, u_1)$$

for all u_1 which can be written as

$$u_1 = u - Q_{t_0}G(I + G)^{-1}(t_0, x_1, u), \qquad u \in U.$$

Since u_1 can thus be taken to be any element of U by choosing $u = u_1 + Q_{t_0}G(t_0, x_1, u_1)$, this shows that

$$Q_{t_0}G(t_0, x_1, u_1) = Q_{t_0}G(t_0, x_2, u_1) \qquad \text{for all } u_1 \in U.$$

Hence, the open loop system is not irreducible whenever the closed loop system is not irreducible. To show uniform reachability, let u be a control such that $\phi(t, t_0, 0, u) = x$ with $t_0 = t - T$ and T as in the definition of uniform reachability. The control $u + Gu$ then transfers the closed loop system from state 0 at $t_0 = t - T$ to state x at t. Since

$$\|P_tQ_{t-T}u + P_tGQ_{t-T}u\| \leqq (1 + \|P_tGQ_{t-T}\|)\|P_tQ_{t-T}u\|,$$

uniform reachability of the closed loop system thus becomes a consequence of uniform reachability of the open loop system if $\|P_t G Q_{t-T}\|$ is uniformly bounded, which in turn is a consequence of the uniform smoothness assumption. Preservation of uniform irreducibility is shown in a similar way. This completes the proof.

The above theorem contains no surprises with the possible exception that it does not state the preservation of observability under feedback. This is in fact untrue, and it is necessary to consider nonlinear systems to obtain satisfactory counterexamples. The system $\dot{x} = uAx$, $y = Cx$ will lead to a counterexample for the contention that closed loop observability follows from open loop observability. This phenomenon is by and large a consequence of the definition of observability which is really observability under zero input. If one were to modify this definition and require that for all given inputs $u \in U$ the response to different initial states should be different, then this observability under arbitrary inputs would indeed by preserved under feedback. These two types of observability are equivalent for linear systems. Note that the above theorem states the equivalence of reachability (controllability, connectedness, irreducibility) of the open loop and the closed loop system since putting positive unit feedback around the closed loop system gives back the open loop system. The fact that only feedback systems with unit feedback are being considered is also inessential to the basic result.

7. Lyapunov functions for feedback systems. In trying to define Lyapunov functions for input-output stable feedback systems, one can of course apply the techniques developed by § 4 and § 5. Such an approach is not very promising since the computation of some of the Lyapunov functions defined there requires detailed knowledge of the closed loop system, whereas it is desirable to pose the calculations and variational questions entirely in terms of the open loop dynamical system. This holds, in particular, for the function V_0 defined in § 4.

It can be shown [13] that a feedback system is finite-gain input-output stable if and only if there exists a constant $\varepsilon > 0$ such that the inequality $\|P_t(I + G)u\| \geqq \varepsilon \|P_t u\|$ holds for all $u \in U$ and $t \in R$. In fact, ε^{-1} may be taken as any real number larger than the gain of $(I + G)^{-1}$. It should also be remarked that for linear feedback systems stability, continuity and finite-gain stability are equivalent.

Now consider the following two functions which are defined as variational problems and will lead to Lyapunov functions for feedback systems:

(i) $V_r(x, t) = \inf \|P_t Q_{t_0}(u + y(t_0, 0, u))\|^2$ when the infimum is to be taken over all $u \in U$ and $t_0 \in R$ such that $\phi(t, t_0, 0, u) = x$ (ϕ and y denote the state transition and output reading map of the open loop dynamical system).

(ii) $V_\varepsilon(x, t) = -\inf(\|P_{t_1} Q_t(u + y(t, x, u))\|^2 - \varepsilon^2 \|P_{t_1} Q_t u\|^2)$, where the infimum is to be taken over all $u \in U$ and $t_1 \geqq t$.

THEOREM 6. *Assume that the feedback system described by equation* (FE) *is uniformly observable, finite-gain input-output stable, and that the state space is uniformly reachable. Let K denote the gain of $(I + G)^{-1}$ and let $0 < \varepsilon \leqq K^{-1}$. Then the feedback system is uniformly globally asymptotically stable and V_r and V_ε are positive definite radially unbounded decrescent uniform Lyapunov functions for it.*

Proof. Note that uniform global asymptotic stability and the claims about V_r follow from Theorem 3. It will now be shown that V_ε is finite.

Let $u_1 \in U$ and $t_0 \leqq t$ be such that $P_{t_0}u_1 = 0$ and $\phi(t, t_0, 0, u_1) = x$. Then

$$P_{t_1}Q_t(u + y(t, x, u)) = P_{t_1}(Q_t u + P_t u_1 + y(t_0, 0, Q_t u + P_t u_1)) - P_t(u_1 + y(t_0, 0, P_t u_1)).$$

Thus,

$$\|P_{t_1}Q_t(u + y(t, x, u))\|^2 \geqq \varepsilon^2 \|P_{t_1}(Q_t u + P_t u_1)\|^2 - \|P_t(u_1 + y(t_0, 0, P_t U_1))\|^2$$
$$= \varepsilon^2 \|P_{t_1}Q_t u\|^2 + \varepsilon^2 \|P_t u_1\|^2 - \|P_t(u_1 + y(t_0, 0, P_t u_1))\|^2$$

and

$$-V_\varepsilon(x, t) \geqq \varepsilon^2 \|P_t u_1\|^2 - \|P_t(u_1 + y(t_0, 0, P_t u_1))\|^2.$$

Since the right-hand side of this inequality depends on $P_t u_1$ only, the result follows. The fact that V_ε is nonnegative follows from taking $u = G(I + G)^{-1}(t, x, 0)$. That V_ε is monotone nonincreasing along solutions follows from the usual argument explained earlier. To show that

$$V_\varepsilon(\phi(t_0 + T, t_0, x_0, 0), t_0 + T) \to 0 \quad \text{as } T \to \infty$$

uniformly in t_0, it suffices to show that V_ε is decrescent since $\phi(t_0 + T, t_0, x_0, 0) \to 0$ as $T \to \infty$ uniformly in t_0. It follows from the above inequality that

$$V_\varepsilon(x, t) \leqq \inf(\|P_t(u_1 + y(t_0, 0, P_t u_1))\|^2 - \varepsilon^2 \|P_t u_1\|^2),$$

where the infimum is to be taken over all $t_0 \leqq t_1$ and $u_1 \in U$ with $P_{t_0}u_1 = 0$ and $\phi(t, t_0, 0, u_1) = x$. Decrescence thus follows from uniform reachability and finite-gain stability. Positive definiteness and radial unboundedness follows by considering $u = e = -G(I + G)^{-1}(t, x, 0)$ which yields $V_\varepsilon(x, t) \geqq \beta(\|x\|)$ by uniform observability. This completes the proof.

The standard methods for proving stability of feedback systems is to show that the open loop gain is less than unity (small loop gain theorem) or to show that the open loop dynamical system may be viewed as the cascade of two passive systems (positive operator conditions). These cases admit special consideration and are treated in the remainder of this section.

Consider therefore the following two functions:

(i) $V_1(x, t) = -\inf(\|Q_t u\|^2 - \|Q_t y(t, x, u)\|^2)$, where the infimum is to be taken over all $u \in U$ with $\|u\| < \infty$, and

(ii) $V_2(x, t) = \inf(\|P_t u\|^2 - \|P_t G u\|^2)$, where the infimum is to be taken over all $t_0 \in R$ and $u \in U$ with $P_{t_0}u = 0$ and $\phi(t, t_0, 0, u) = x$.

THEOREM 7. *Assume that the feedback system described by equations* (FE) *is uniformly observable, that the open loop dynamical system has gain less than unity and that the state space is uniformly reachable. Then the feedback system is finite-gain input-output stable and uniformly globally asymptotically stable and V_1 and V_2 are positive definite radially unbounded decrescent uniform Lyapunov functions for it.*

The case in which the open loop consists of the composition of a dynamical system G_1 followed by a memoryless dynamical system G_2 whose gain product is less than unity has received a great deal of attention and leads to a Lyapunov

function which only depends on G_1. Consider therefore the following functions:

(i) $V_1^*(x, t) = -\inf(\|Q_t u\|^2 - \|G_1\|^{-2}\|Q_t G_1(t, x, u)\|^2)$, where the infimum is to be taken over all $u \in U$ with $\|u\| < \infty$, and

(ii) $V_2^*(x, t) = \inf(\|P_t u\|^2 - \|G_1\|^{-2}\|P_t G u\|^2)$, where the infimum is to be taken over all $t_0 \in R$ and $u \in U$ with $P_{t_0} u = 0$ and $\phi(t, t_0, 0, u) = x$.

THEOREM 8. *Assume that the feedback system described by equations* (FE) *is uniformly observable and that the state space is uniformly reachable. Let* $G = G_2 G_1$, *where* G_1 *is a uniformly controllable, uniformly observable dynamical system and* G_2 *is a memoryless dynamical system. Assume that the product of the gains of* G_1 *and* G_2, $\|G_1\|\,\|G_2\|$, *is less than unity. Then the feedback system is finite-gain input-output stable and uniformly globally asymptotically stable and* V_1^* *and* V_2^* *are positive definite radially unbounded decrescent uniform Lyapunov functions for it.*

Proof. The proofs of Theorems 7 and 8 offer no surprises considering the previous theorems, and the details will be omitted. The stability claims follow from the so-called small-gain theorem [5] for input-output stability and Theorem 2. Existence of V_1, V_2, V_1^* and V_2^* follows from the small gain condition, decrescence from uniform reachability, and positive definiteness by taking $u = 0$. Monotonicity along undriven solutions requires a minor modification of the usual argument. Consider, for instance, the function V_1^* of Theorem 8. Let $t_1 \geqq t_0$ and $x_1 = \psi(t_1, t_0, x_0, 0)$, with ψ the state transition map of the closed loop feedback system. Choose u on (t_0, t_1) to equal $e = P_{t_1} Q_{t_0}(I + G)^{-1}(t_0, x_0, 0)$. Thus,

$$V_1^*(x_0, t_0) \geqq -\|e\|^2 + \|G_1\|^{-2}\|G_1(t_0, x_0, e)\|^2 + V_1^*(x_1, t_1).$$

Since, however,

$$\begin{aligned} -\|e\|^2 + \|G_1\|^{-2}\|P_{t_1} Q_{t_0} G_1(t_0, x_0, e)\|^2 &= -\|G_2 P_{t_1} Q_{t_0} G_1(t_0, x_0, e)\|^2 \\ &\quad + \|G_1\|^{-2}\|P_{t_1} Q_{t_0} G_1(t_0, x_0, e)\|^2 \\ &\geqq (-\|G_2\|^2 + \|G_1\|^{-2})\|P_{t_1} Q_{t_0} G_1(t_0, x_0, e)\|^2 \end{aligned}$$

and $\|G_1\|^{-2} - \|G_2\|^2 > 0$, $V_1^*(x_0, t_0) \geqq V_1^*(x_1, t_1)$ as desired. This completes the proof.

Theorem 8 is particularly useful, for instance, when G_1 is linear and G_2 is nonlinear or when G_1 is linear and time invariant and G_2 is time-varying. The variational problems which then result are indeed much simpler if one applies Theorem 8 than those needed in Theorem 7.

The two theorems which follow are the counterparts of the preceding ones, but with passivity conditions replacing the small gain condition. The stability theorem which lies at the basis of these results states that a feedback system is finite-gain input-output stable if the open loop dynamical system G is the composition of a passive system, G_1, and a strictly passive finite-gain input-output stable system, G_2. This decomposition is usually not the result of physical considerations, but rather a mathematical device which allows one to prove stability of the closed loop system.

Assume thus that G, G_1 and G_2 with $G = G_2 G_1$ are dynamical systems in state space form with state spaces X, X_1 and X_2 respectively. The space $X_1 \times X_2$ certainly qualifies as another state space for G but will, in general, be much larger than X, particularly if the latter is minimal (i.e., reachable and irreducible). In

general, this is true in stability applications since the factors G_1 and G_2 are usually not natural decompositions of G but are constructed with the aid of so-called "multipliers," which usually results in this inflation of the state space. Assume now that the dynamical system G in state space form has a reachable and irreducible state space X. Consider the dynamical system $G = G_2G_1$ with state space $X_1 \times X_2$ and assume that the state $(x_1, x_2) \in X_1 \times X_2$ is reachable at $t \in R$, i.e., that there exist a $t_0 \leqq t$ and $u \in U$ with $P_{t_0}u = 0$ such that $\phi_1(t, t_0, 0, u) = x_1$ and $\phi_2(t, t_0, 0, G_1u) = x_2$ (ϕ_1 and ϕ_2 denote the state transition maps of G_1 and G_2 respectively). Consider now on this subset of reachable states at time t the equivalence classes of those which yield the same output after time t for all $u \in U$, i.e., the reachable states (x_1', x_2') and (x_1'', x_2'') will be considered equivalent if

$$Q_tG_2(t, x_2', G_1(t, x_1', u)) = Q_tG_2(t, x_2'', G_1(t, x_1'', u)) \quad \text{for all } u \in U.$$

There is (by minimality) a one-to-one and onto correspondence between these equivalence classes and the space X. Denote by $X_t(x_1, x_2)$ the element of X corresponding in this sense to the equivalence class derived from the reachable state (x_1, x_2). The map X_t may in general depend explicitly on t. Assume furthermore that there exist constants k and K such that

$$k(\|x_1\|^2 + \|x_2\|^2) \leqq \|X_t(x_1, x_2)\|^2 \leqq K(\|x_1\|^2 + \|x_2\|^2) \quad \text{for all } t \in R$$

and reachable states $(x_1, x_2) \in X_1 \times X_2$. The decomposition of G into $G = G_2G_1$ will then be called a *compatible factorization* of the dynamical system G.

The statement of the theorem which follows involves Lyapunov functions defined on $X_1 \times X_2$, but these can, by the above remarks, also be considered as Lyapunov functions on the state space X provided one only considers pairs (x_1, x_2) which are reachable. The following theorem statement then becomes clear.

THEOREM 9. *Assume that the feedback system described by equations* (FE) *is uniformly observable and that the state space is uniformly reachable. Assume also that the open loop dynamical system G has a uniformly reachable and uniformly irreducible state space and that it admits a compatible factorization* $G = G_2G_1$ *into the uniformly observable dynamical systems* G_1 *and* G_2 *with uniformly reachable state spaces* X_1 *and* X_2 *respectively. Assume that one of these factors is passive and uniformly smooth and that the other is strictly passive and finite-gain input-output stable. Then the closed loop feedback system is finite-gain input-output stable and uniformly globally asymptotically stable, and the total available energy,* $E_a = E_a^{(1)} + E_a^{(2)}$, *and the total required energy,* $E_r = E_r^{(1)} + E_r^{(2)}$, *are decrescent uniform Lyapunov functions for it.* (The superscripts refer to the dynamical systems composing G.)

Proof. Decrescence of E_r on $X_1 \times X_2$ follows from Theorem 4 with an appropriate modification in the proof in order to replace finite-gain stability by the uniform smoothness condition. Decrescence on $X_1 \times X_2$ then implies decrescence on X by the inequality in the definition of a compatible factorization. Since $E_a \leqq E_r$, E_a is also decrescent. The stability claims about the feedback system are well known [5], and it remains to be shown that the energy functions are monotone nonincreasing along undriven solutions. This will only be shown for the required energy. The proof for the available energy is similar.

Let $(x_1', x_2') \in X_1 \times X_2$ and $t_0 \in R$ be given, and let $(x_1'', x_2'') \in X_1 \times X_2$ denote the state of the dynamical systems G_2G_1 at time $t_1 \geqq t_0$ resulting from the transfer along solutions of the undriven feedback system. Then

$$\inf_{\to x_1'', t_1} \langle P_{t_1}u_1, P_{t_1}G_1u_1\rangle + \inf_{\to x_2'', t_2} \langle P_{t_1}u_2, P_{t_1}G_2u_2\rangle$$
$$\leqq \inf_{\to x_1', t_0} \langle P_{t_0}u_1, P_{t_0}G_1u_1\rangle + \inf_{\to x_2', t_0} \langle P_{t_0}u_2, P_{t_0}G_2u_2\rangle$$
$$+ \langle P_{t_1}Q_{t_0}e_1, P_{t_1}Q_{t_0}G_1(t_0, x_1', e_1)\rangle + \langle P_{t_1}Q_{t_0}e_2, P_{t_1}Q_{t_0}G_2(t_0, x_2', e_2)\rangle,$$

where the notation $\inf_{\to x_1'', t_1}$, for instance, denotes the infimum over all $t \leqq t_1$ and $u_1 \in U$ with $P_t u_1 = 0$ and $\phi_1(t_1, t, 0, u_1) = x_1''$. The other symbolism is to be interpreted in an analogous way. The inputs e_1 and e_2 denote respectively $(I + G)^{-1}(t_0, (x_1', x_2'), 0)$ and $G_1(I + G)^{-1}(t_0, (x_1', x_2'), 0)$. The desired result then follows if one notices that $Q_{t_0}e_2 = Q_{t_0}G_1(t_0, x_1', e_1)$ and that $Q_{t_0}e_1 = -Q_{t_0}G_2(t_0, x_2', G_1(t_0, x_1', e_1))$ since this shows that the contributions of the last two terms in the above inequality cancel. This completes the proof.

The reader is referred to the remark following Theorem 4 for conditions to ensure positive definiteness and radial unboundedness. Notice again that positive definiteness on $X_1 \times X_2$ suffices for positive definiteness on X by the definition of a compatible factorization. The case in which the operator G_2 is memoryless leads, as in the small gain case, to a simplification. This is stated in the following final theorem.

THEOREM 10. *Assume that the feedback system described by equations* (FE) *is uniformly observable and that the state space is uniformly reachable. Assume that the open loop dynamical system, $G = G_2G_1$, consists of the composition of a uniformly observable, finite-gain input-output stable, strictly passive dynamical system, G_1, with a uniformly reachable state space, followed by a memoryless passive dynamical system, G_2. Then the closed loop feedback system is finite-gain input–output stable and uniformly globally asymptotically stable, and the available energy, $E_a^{(1)}$, and the required energy, $E_r^{(1)}$, are decrescent uniform Lyapunov functions for it.*

Proof. The proof combines the ideas in the proofs of Theorems 8 and 9 and is left to the reader.

The theorems developed here treat the small gain stability conditions and the passive operator stability conditions. The methods can, however, easily be extended to treat conic operators as well.

8. Examples.[7]

Example 1. Let $G(s)$ be a $p \times m$ matrix of rational functions of s with $\lim_{s\to\infty} G(s) = 0$, and assume that $\{A, B, C\}$ is a minimal[8] realization of $G(s)$. Assume that

[7] The norms and inner products involved in these examples are the usual norms and inner products of Euclidean spaces. Prime denotes transposition. For the calculations involved in the solution of the variational problems in this section, see [23, §§ 21, 22, 23 and 25]. Although some of the problems are not treated explicitly there, the modifications merely require algebraic manipulation and no new methodology.

[8] Algebraically this means that A is an $n \times n$ matrix, that the $n \times nm$ and $n \times np$ matrices $(B, AB, \cdots, A^{n-1}B)$ and $(C', A'C', \cdots, (A')^{n-1}C')$ are of full rank n, and that $G(s) = C(Is - A)^{-1}B$. The full

the poles of $G(s)$, which by minimality equal the eigenvalues of A, are in $\operatorname{Re} s < 0$. The system $\dot{x} = Ax + Bu$; $y = Cx$ is thus finite-gain input–output stable and globally asymptotically stable, and Theorem 3 yields as positive definite decrescent radially unbounded Lyapunov functions the quadratic forms $x'K_1x$ and $x'K_2x$, where K_1 is the (unique, positive definite) solution of the linear matrix equation $A'X + XA = -C'C$, and K_2^{-1} is the (unique, positive definite) solution of the linear matrix equation $AX + XA' = -BB'$. Clearly, these are only two of many possible Lyapunov functions for this asymptotically stable dynamical system.

Example 2. Let $G(s)$ be an $m \times m$ matrix of rational functions of s with $\lim_{s\to\infty} G(s) < \infty$, and assume that $\{A, B, C, D\}$ is a minimal realization of $G(s)$. Assume that the poles of $G(s)$ are in $\operatorname{Re} s < 0$, that $G(j\omega) + G'(-j\omega)$ is Hermitian positive definite for all $\omega \in R$, and that $D + D'$ is positive definite. The n-dimensional system $\dot{x} = Ax + Bu$; $y = Cx + Du$ and thus strictly passive, finite-gain input-output stable, and globally asymptotically stable. The available energy, $E_a(x_0, t_0)$, is given by

$$-\inf_{u\in L_2(0,\infty)} \eta, \qquad \text{where } \eta = \int_0^\infty u'(t)y(t)\,dt,$$

subject to the constraint $\dot{x} = Ax + Bu$; $y = Cx + Du$, $x(0) = x_0$, and is independent of t_0. This variational problem is a least squares problem and, by Lemma 1, an infimum exists. This infimum is, in fact, attained by the feedback control

$$u = -(D + D')^{-1}(C + B'K)x$$

and

$$\min_{u\in L_2(0,\infty)} \eta = \eta^* = -x_0'Kx_0/2,$$

where $K = K'$ is the (unique) negative definite solution of the algebraic Riccati equation

$$0 = -A'X - XA + (C + B'X)'(D + D')^{-1}(C + B'X).$$

Note [8], [9], [10] that this implies the existence of an $n \times n$ positive definite matrix $P = P'$ ($P = -K$), and $n \times m$ matrix L, and an $m \times m$ matrix W_0 such that (Kalman–Yakubovich–Popov):

$$A'P + PA = -LL',$$

$$PB = C' - LW_0,$$

$$W_0'W_0 = D + D'.$$

---

rank condition on $(B, AB, \cdots, A^{n-1}B)$ is equivalent to controllability, reachability, and connectedness, and the full rank condition on $(C', A'C', \cdots, (A')^{n-1}C')$ is equivalent to observability and irreducibility, where these notions refer to the linear time-invariant finite-dimensional system $\dot{x} = Ax + Bu$; $y = Cx$. For these systems, the observability considered here is equivalent to observability under arbitrary inputs, and all of these properties hold uniformly whenever they hold. Global asymptotic stability requires all the eigenvalues of A to be in $\operatorname{Re} \lambda < 0$ and is equivalent to input-output continuity if the system is minimal.

Since, moreover, in this (linear) case the passivity is necessary and sufficient for the existence of the required infimum, and since passivity is equivalent to positive realness of $G(s)$, the above conditions (existence of K or P) are also necessary for positive realness. Thus the available energy $E_a(x, t) = x'Px/2$ is a positive definite radially unbounded decrescent Lyapunov function. The required energy, $E_r(x_0, t_0)$, is somewhat more involved to calculate and is defined by

$$\inf_{T \geqq 0} \inf_{u \in L_2(-T,0)} \eta,$$

where

$$\eta = \int_{-T}^{0} u'(t)y(t)\,dt,$$

subject to the constraint[9] $\dot{x} = Ax + Bu$; $y = Cx + Du$, $x(-T) = 0$, $x(0) = x_0$, and is independent of t_0. The above variational problem is again a least squares problem, and by Lemma 1, an infimum exists. This infimum can be characterized as follows:

$$\eta^* = \inf_{T \geqq 0} \min_{u \in L_2(-T,0)} \eta = -x_0'\Sigma x_0/2,$$

where $\Sigma = P + W^{-1}$ and $P = P'$ is the (unique) positive definite solution of the algebraic Riccati equation

$$0 = A'X + XA + (C - B'X)'(D + D')^{-1}(C - B'X).$$

In fact, this matrix is the same as the one appearing in the calculation of the available energy and is such that $A_1 = A - B(D + D')^{-1}(C - B'P)$ is an asymptotically stable matrix. W is the (unique) solution of the linear matrix equation $A_1X + XA_1' = -B(D + D')^{-1}B'$, and is symmetric positive definite. The required energy $E_r(x, t) = \frac{1}{2}x'Px + \frac{1}{2}x'W^{-1}x$ is also a positive definite radially unbounded decrescent Lyapunov function. The cycle energy $E_c = E_r - E_a$ is given by $E_c(x, t) = \frac{1}{2}x'W^{-1}x$. The system is thus lossy.

Example 3. Let $g(s)$ be a rational function of s with $\lim_{s\to\infty} g(s) = 0$ and assume that $\{A, b, c'\}$ is a minimal realization of $g(s)$. Let k be a scalar. Assume that the Nyquist locus of $g(s)$ does not intersect but encircles the $-1/k$ point in the complex plane $-\rho$ times in the clockwise direction, where ρ is the number of poles of g in $\operatorname{Re} s \geqq 0$. Then the closed loop system $\dot{x} = (A - kbc')x$ is globally asymptotically stable, and Theorem 6 yields as a positive definite radially unbounded decrescent Lyapunov function $-x'Rx$, where $R = R'$ is the (unique) negative definite solution of the algebraic matrix Riccati equation

$$0 = -A'X - XA + \frac{(kc + Xb)(kc + Xb)'}{1 - \varepsilon^2} - cc'$$

with $\varepsilon > 0$ such that $|1 + kg(j\omega)| \geqq \varepsilon_1 > \varepsilon$ for all $\omega \in R$.

---

[9] It is important to realize that this variational problem is *not* equivalent to the simpler one which asks to evaluate $\inf_{u \in L_2(-\infty,0)} \int_{-\infty}^{0} u'(t)y(t)\,dt$ subject to $\dot{x} = Ax + Bu$; $y = Cx + Du$, $x(0) = x_0$. (This latter variational problem leads again to the available energy.)

Example 4. Let $G(s)$ be a $p \times m$ matrix of rational functions of s with $\lim_{s\to\infty} G(s) = 0$ and assume that $\{A, B, C\}$ is a minimal realization of $G(s)$. Assume that the poles of $G(s)$ are in $\operatorname{Re} s < 0$, and that for all $\omega \in R$ the eigenvalues of the matrix $G'(-j\omega)G(j\omega)$ are inside the open ball with radius ρ^{-2} in the complex plane. Let $f(\sigma, t)$ be a R^m-valued function defined on $R^p \times R$, Lipschitz continuous on R^p, uniformly in t, and satisfying, for some $\alpha < \rho$, the inequality $\|f(\sigma, t)\| \leqq \alpha\|\sigma\|$ for all $(\sigma, t) \in R^p \times R$. Consider now the nonlinear differential equation

$$\dot{x}(t) = Ax(t) - Bf(Cx(t), t).$$

This differential equation may be viewed as the mathematical model of the undriven feedback system studied in § 6 with the open loop dynamical system described by the equations

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = f(Cx(t), t),$$

and the closed loop dynamical system determined by the equations

$$\dot{x}(t) = Ax(t) - Bf(Cx(t), t) + Bu(t), \qquad y(t) = f(Cx(t), t).$$

This system satisfies all the assumptions for Theorem 8 to be applicable, and the nonlinear differential equation is thus uniformly globally asymptotically stable by the small gain theorem. Consider now

$$\inf_{u\in L_2(0,\infty)} \int_0^\infty [u'(t)u(t) - \rho^2 y'(t)y(t)]\, dt,$$

subject to the constraint $\dot{x} = Ax + Bu$; $y = Cx$, $x(0) = x_0$. This infimum exists and is given by $x_0'Kx_0$ when K is the (unique) negative definite solution of the matrix Riccati equation

$$0 = -A'X - XA + XBB'X + \rho^2 C'C.$$

Theorem 8 thus states that $-x'Kx$ is a positive definite radially unbounded decrescent uniform Lyapunov function for this nonlinear differential equation. Theorem 8 yields as another positive definite radially unbounded decrescent uniform Lyapunov function $x'(K + W^{-1})x$, where K is as defined above and W is the (unique, positive definite) solution of the linear matrix equation

$$(A - BB'K)X + X(A - BB'K)' = -BB'.$$

Example 5. Let $g(s)$ be a rational function of s with $\lim_{s\to\infty} g(s) = 0$, and assume that $\{A, b, c'\}$ is a minimal realization of $g(s)$. Assume that the poles of $g(s)$ are in $\operatorname{Re} s < 0$ and that there exists a real number $\alpha \geqq 0$ such that for some constant $\varepsilon > 0$, $\operatorname{Re}(\alpha + j\omega)g(j\omega) \geqq \varepsilon > 0$ for all $\omega \geqq 0$. Let $f(\sigma)$ be a real-valued function defined on R, Lipschitz continuous on R, and satisfying for some $\delta > 0$ the inequality $f(\sigma)/\sigma > \delta > 0$ for all $\sigma \in R$, $\sigma \neq 0$. Consider now the nonlinear differential equation $\dot{x}(t) = Ax(t) - bf(c'x(t))$. This differential equation may be viewed as the mathematical model of the undriven feedback system studied in § 6 with the open loop dynamical system described by the equation

$$\dot{x} = Ax + bu; \qquad y = f(c'x),$$

and the closed loop dynamical system described by the equations

$$\dot{x} = Ax - bf(c'x) + bu, \qquad y = f(c'x).$$

The open loop dynamical system can be viewed as the cascade of the two systems: a linear time-invariant system with transfer function $(s + \alpha)g(s)$ followed by a nonlinear system which is the cascade of a linear time-invariant system with transfer function $1/(s + \alpha)$ followed by the memoryless nonlinearity $f(\cdot)$. The state equations of the first system are

$$\dot{x} = Ax + bu, \qquad y = \alpha c'x + c'Ax + c'bu.$$

This system satisfies the assumptions of the system of Example 3 which thus yields expressions for the available energy, the required energy, and the cycle energy. The state equations for the second system are

$$\dot{z} = -\alpha z + u, \qquad y = f(z).$$

The available energy $E_a(z, t)$ for this system is independent of t_0 and is defined by

$$E_a(z_0) = -\inf_{T \geqq 0} \inf_{u \in L_2(0,T)} \eta,$$

where

$$\eta = \int_0^T u(t)y(t)\,dt,$$

subject to $\dot{z} = -\alpha z + u$, $y = f(z)$, $z(0) = z_0$. Thus

$$\eta = F(z(T)) + \alpha \int_0^T z(t)f(z(t))\,dt - F(z_0)$$

with

$$F(z) = \int_0^z \sigma f(\sigma)\,d\sigma.$$

Since $F(\sigma) \geqq 0$ and $\sigma f(\sigma) \geqq 0$ for all $\sigma \in R$, and since the value of

$$\alpha \int_0^T z(t)f(z(t))\,dt + F(z(T))$$

can be made arbitrarily small by proper choice of u, it follows that

$$E_a(z_0) = F(z_0) = \int_0^{z_0} \sigma f(\sigma)\,d\sigma.$$

Similarly, the required energy

$$E_r(z_0) = F(z_0) = \int_0^{z_0} \sigma f(\sigma)\,d\sigma.$$

The cycle energy E_c for this first order nonlinear system is thus zero, and the system is reversible. It is a simple matter to show that the above factorization of the system $\dot{x} = Ax + bu$, $y = f(c'x)$ is a compatible factorization as defined in § 7. Notice that reachable states satisfy the condition $z_0 = c'x_0$, which defines a hyperplane in the space $R^n \times R$. Theorem 9 thus yields as positive definite radially unbounded

decrescent Lyapunov functions for the nonlinear differential equation $\dot{x} = Ax - bf(c'x)$ satisfying the conditions enumerated earlier (i.e., the conditions of the Popov criterion):

(i) $V(x) = \frac{1}{2}x'Px + \int_0^{c'x} f(\sigma)\,d\sigma$, where $P = P'$ is the (unique) positive definite solution of the algebraic Riccati equation[10]

$$0 = A'X + XA + \frac{1}{2c'b}(\alpha c' + c'A - b'X)'(\alpha c' + c'A - b'X).$$

(ii) $V(x) + \frac{1}{2}x'W^{-1}x$, where W is the (unique) solution of the linear matrix equation

$$A_1X + XA_1' = -\frac{bb'}{2c'b} \quad \text{with} \quad A_1 = A - \frac{b}{2c'b}(\alpha c' + c'A - b'P),$$

and is symmetric positive definite.

Example 6. *Path integrals* [23, § 26], [25].

LEMMA 2 [23, p. 170]. *Assume that $x(t)$ is an n times differentiable function of t and that α_{ij}, $i, j = 0, 1, \cdots, n$, are constants. Then the*

$$\eta = \int_{t_0}^{t_1} \sum_{i,j=0}^{n} \alpha_{ij} \frac{d^ix(t)}{dt^i} \frac{d^jx(t)}{dt^j}\,dt$$

is independent of path (i.e., it depends only on the values of $x(t)$ and its derivatives at $t = t_0$ and $t = t_1$) if and only if the polynomial

$$h(s) = \sum_{i,j=0}^{n} \alpha_{ij}(s^i(-s)^j + (-s)^is^j)$$

vanishes identically.

This lemma leads to rather specific formulas for the Lyapunov functions described in this paper. For instance, Theorem 3 thus yields as a Lyapunov function for the differential equation $p(D)x(t) = 0$ with $D = d/dt$ and $p(s)$ a polynomial with all its roots in $\operatorname{Re} s < 0$,

$$V(x, x^{(1)}, \cdots, x^{(n-1)}) = \inf_{T \geq 0} \inf_{\substack{x(t) \mid x(-T) = \cdots = x^{(n-1)}(-T) = 0 \\ x(0) = x, \cdots, x^{(n-1)}(0) = x^{(n-1)}}} \eta,$$

where $\eta = \int_{-T}^{0} (p(D)x(t))^2\,dt$. Let $r(s)$ be a solution of the polynomial equation $p(s)p(-s) = r(s)r(-s)$, and let $(p\bar{p})^+(s) = p(-s)$ and $(p\bar{p})^-(s) = p(s)$ denote the solutions with poles respectively in $\operatorname{Re} s > 0$ and $\operatorname{Re} s < 0$. Now rewriting η as

$$\eta = \int_{-T}^{0} [(p(D)x(t))^2 - (r(D)x(t))^2]\,dt + \int_{-T}^{0} (r(D)x(t))^2\,dt,$$

one observes that by Lemma 2 the first integral is independent of path and thus depends on the values of $x, x^{(1)}, \cdots, x^{(n-1)}$ only. The integrand in the second integral is nonnegative and should hence be made as small as possible. By choosing

[10] Compare with the results of [24].

$r(s) = (p\bar{p})^+(s) = p(-s)$ and letting $T \to +\infty$, the contribution of this second integral can indeed be made arbitrarily small and yields a positive definite radially unbounded decrescent Lyapunov function for $p(D)x(t) = 0$, the quadratic form

$$V(x, \dot{x}, \cdots, x^{(n-1)}) = \int_{-\infty}^{0} [(p(D)x(t))^2 - (p(-D)x(t))^2]\, dt,$$

with $x(t)$ any n times differentiable function such that

$$x(0) = x, \cdots, x^{(n-1)}(0) = x^{(n-1)}$$

and

$$\lim_{t \to -\infty} x(t) = \cdots = x^{(n-1)}(t) = 0.$$

If $g(s)$ is chosen such that

$$\lim_{s \to \infty} \frac{q(s)}{p(s)} < \infty \quad \text{and} \quad \operatorname{Re} \frac{q(j\omega)}{p(j\omega)} \geqq 0 \quad \text{for all } \omega,$$

then Theorem 4 yields as Lyapunov functions

$$E_r(x, \dot{x}, \cdots, x^{(n-1)}) = \int_{-\infty}^{0} [p(D)x(t)q(D)x(t) - \frac{1}{2}((p\bar{q} + \bar{p}q)^+(D)x(t))^2]\, dt$$

and

$$E_a(x, \dot{x}, \cdots, x^{(n-1)}) = \int_{0}^{\infty} [p(D)x(t)q(D)x(t) - \frac{1}{2}((p\bar{q} + \bar{p}q)^-(D)x(t))^2]\, dt$$

with $x(t)$ any n times differentiable function such that

$$x(0) = x, \cdots, x^{(n-1)}(0) = x^{(n-1)}$$

and

$$\lim_{t \to \pm\infty} x(t) = \cdots = x^{(n-1)}(t) = 0.$$

9. Conclusions. The development of the results and the techniques described in this paper evolves in three stages: the first one introduces and compares the input-output description with the state space description of dynamical systems and shows their equivalence. The second part in the development leads to the equivalence of input-output stability and global stability under appropriate controllability and observability conditions; the third issue is the construction of Lyapunov functions.

The methods for constructing Lyapunov functions involve, for the most part, variational problems and are posed in the framework of systems with inputs and outputs; this notwithstanding the fact that the system for which global asymptotic stability (in the sense of Lyapunov) is to be shown is an autonomous (undriven) system. The results thus obtained serve as a further relationship between the areas of dynamic optimization and stability theory and focus interest on a class of optimization problems, some of which will, in fact, lead to singular controls.

It is felt that the importance of this paper lies in its theoretical contribution in demonstrating the equivalence between global asymptotic stability and input-output stability, which, as expected, merely requires appropriate controllability and observability (more precisely: reachability and uniform observability). It also serves to unify the two main approaches to stability theory: input-output stability and Lyapunov stability. In this latter class it unifies and generalizes the various available results by posing the construction of these Lyapunov functions as variational problems.

The results of the paper could also serve as a starting point to develop techniques which will lead to suitable Lyapunov functions for estimating the domain of attraction for nonglobal stable systems. This is a problem of great practical importance, and the methods of the paper lead to tractable variational problems which could be used in such an analysis.

Acknowledgment. The author acknowledges Professor R. W. Brockett of Harvard University and Dr. J. L. Willems of the University of Ghent, Belgium (on leave at Harvard University) for some helpful discussions.

REFERENCES

[1] L. ZADEH AND C. A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.

[2] A. V. BALAKRISHNAN, *On the state space theory of nonlinear systems*, Functional Analysis and Optimization, E. R. Caianiello, ed., Academic Press, New York, 1966, pp. 15–36.

[3] V. V. NEMYTSKII AND V. V. STEPANOV, *Qualitative Theory of Differential Equations*, Princeton University Press, Princeton, N.J., 1960.

[4] I. W. SANDBERG, *Some results on the theory of physical systems governed by nonlinear functional equations*, Bell System Tech. J., 44 (1965), pp. 871–898.

[5] G. ZAMES, *On the input-output stability of time-varying nonlinear feedback systems. Part I: Conditions derived using concepts of loop gain, conicity, and positivity; Part II: Conditions involving circles in the frequency plane and sector nonlinearities*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238, 465–476.

[6] J. C. WILLEMS, *A survey of stability of distributed parameter systems*, Control of Distributed Parameter Systems, 1969 Joint Automatic Control Conference, Boulder, Colo., ASME Publ., 1969, pp. 63–102.

[7] R. A. BAKER AND A. R. BERGEN, *Lyapunov stability and Lyapunov functions of infinite dimensional systems*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 325–334.

[8] V. M. POPOV, *Hyperstability and optimality of automatic systems with several control functions*, Rev. Roumaine Sci. Tech. Electrotechn. et Energ., 9 (1964), no. 4, pp. 629–690.

[9] R. E. KALMAN, *Lyapunov functions for the problem of Lur'e in automatic control*, Proc. Nat. Acad. Sci. U.S.A., 49 (1963), pp. 201–205.

[10] B. D. O. ANDERSON, *A system theory criterion for positive real matrices*, this Journal, 5 (1967), pp. 171–182.

[11] B. D. O. ANDERSON AND J. B. MOORE, *New results in linear system stability*, this Journal, 7 (1969), pp. 398–414.

[12] R. W. BROCKETT, *Path integrals, Liapunov functions, and quadratic minimization*, Proc. 4th Annual Allerton Conf. on Circuit and System Theory, Monticello, Ill., 1966, pp. 685–698.

[13] J. C. WILLEMS, *Stability, instability, invertibility and causality*, this Journal, 7 (1969), pp. 645–671.

[14] ———, *The Analysis of Feedback Systems*, MIT Press, Cambridge, Mass., 1970.

[15] G. ZAMES, *Realizability conditions for feedback systems*, IEEE Trans. Circuit Theory, CT-11 (1964), pp. 186–194.

[16] A. V. BALAKRISHNAN, *On the controllability of a nonlinear system*, Proc. Nat. Acad. Sci. U.S.A., 55 (1966), pp. 465–468.

[17] A. H. ZEMANIAN, *The Hilbert port*, SIAM J. Appl. Math., 18 (1970), pp. 98–138.

[18] A. V. BALAKRISHNAN, *Foundations of the state-space theory of continuous systems. I*, J. Computer and System Sci., 1 (1967), pp. 91–116.
[19] N. WIENER, *Nonlinear Problems in Random Theory*, MIT Press, Cambridge, Mass., 1958.
[20] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.
[21] ———, *Canonical structure of linear dynamical systems*, Proc. Nat. Acad. Sci. U.S.A., 48 (1962), pp. 596–600.
[22] G. LUMER AND R. S. PHILLIPS, *Dissipative operators in a Banach space*, Pacific J. Math., 11 (1961), pp. 679–698.
[23] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
[24] K. R. MEYER, *On the existence of Lyapunov functions for the problem of Lur'e*, this Journal, 3 (1966), pp. 373–383.
[25] R. W. BROCKETT AND J. L. WILLEMS, *Frequency domain stability criteria. Parts I and II*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 255–261, 401–413.
[26] R. W. BROCKETT, *System theory on group manifolds and coset spaces*, submitted to this Journal.

SIAM J. CONTROL
Vol. 9, No. 1, February 1971

THE DISTRIBUTION OF 2 × n GAME VALUES AND PROGRAM OPTIMA*

DONALD J. SOULTS† AND H. T. DAVID‡

1. Introduction. In this paper we derive an expression (2.14) for the distribution of the value V of a $2 \times n$ game when the n columns of the game matrix are distributed independently, each according to a bivariate distribution satisfying certain continuity conditions.

The sort of expression obtained is best illustrated under the further assumption that the columns are identically distributed. Suppose F is that common distribution, and let $\alpha(p, v, F)$ be the probability under F of the half-plane $px_1 + (1 - p)x_2 > v$. Then, if F satisfies certain continuity conditions, $P\{V > v\}$ is representable as an integral of $\alpha^{n-1}(p, v, F)$:

$$P\{V > v\} = \int_0^1 \alpha^{n-1}(p, v, F)\, dW(p). \tag{1.1}$$

Either of two functions $W(p)$ may be used in (1.1). Of these, one typically has a discontinuity at 0 and the other typically has a discontinuity at 1.

The expectation of the distribution (2.14) is the value of a two-stage game in which Nature, at the first stage, randomly selects the game to be played at the second stage. The random selection is according to $\prod_j F_j$, where F_j is the distribution of the jth column. Hence the expectation of (2.14) is a possible valuation of a statistical population of $2 \times n$ games of which a randomly selected one eventually is to be played.

By a familiar correspondence between games and programs, we obtain as well the distribution (3.2) of the optimum of the linear program:

$$\begin{gathered} \min \{b_1 y_1 + b_2 y_2\}, \\ a_{1j} y_1 + a_{2j} y_2 \geqq c_j, \quad j = 1, 2, \cdots, n, \quad y_1, y_2 \geqq 0, \end{gathered} \tag{1.2}$$

where the b_i and the c_j are nonrandom and positive, and where the n vectors (a_{1j}, a_{2j}) are random and independent. Because of the possibility that (1.2) is infeasible, this distribution typically is improper, with the probability deficit equal to the probability of infeasibility (cf.(3.3)).

The distribution (3.2) and probability (3.3) provide in effect closed expressions for the propagation of variability from the technological matrix to the optimum; hence, they seem of possible interest for sensitivity analysis.

As indicated below, the early parts of our argument apply to the general $m \times n$ case. However, we have not managed to arrive at representations similar to (1.1), (2.14) or (3.2) for $m \geqq 3$.

There is a vast literature in stochastic programming, indicated, for example, in [5]. We dwell solely on the distributional problem, to the exclusion of decision-

* Received by the editors February 17, 1970, and in revised form June 12, 1970.
† Boeing-Aircraft Company, Renton, Washington 98055.
‡ Department of Statistics, Iowa State University, Ames, Iowa 50010.

making and economic considerations. References [1], [2], [4], [6], [7], [8] and the references cited therein have similar emphasis.

2. The value distribution. Let σ_m be the m-simplex, i.e., the set of points $p:(p^{(1)} \cdots, p^{(m)})$ with $p^{(i)} \geqq 0$ and $\sum p^{(i)} = 1$. Consider an $m \times n$ game matrix $\|X_{ij}\|$, $i = 1, \cdots, m; j = 1, 2, \cdots, n$. The value of the corresponding game is

$$V = \max_{p \in \sigma_m} f(p),$$

where $f(p)$, continuous and concave on σ_m, is given by

$$f(p) = \min_{j=1,2,\cdots,n} \sum_{i=1}^{m} p^{(i)} X_{ij}.$$

Hence the event $V > v$ is equivalent to the event:

$$E: \text{for at least one } p \text{ in } \sigma_m, f(p) > v.$$

Let $p_1, p_2, \cdots, p_\nu$ be the points of a suitable grid G_ν in σ_m. Define the approximating events

$$E^\nu = \bigcup_{k=0,1,\cdots,\nu} E_k^\nu,$$

where $E_k^\nu = f(p_k) > v$.

It is readily verified using the continuity of $f(p)$ that E is the set-theoretic limit of E^ν if $\|G_\nu\| \to 0$, in which case [3, p. 40]

$$\lim_{\nu \to \infty} P\{E^\nu\} = P\{E\}, \tag{2.1}$$

and inclusion-exclusion gives

$$\begin{aligned} P\{E^\nu\} = {} & \sum_{k=0}^{\nu} \langle k \rangle - \sum_{k_1 < k_2} \sum \langle k_1, k_2 \rangle \\ & + \sum_{k_1 < k_2 < k_3} \sum \langle k_1, k_2, k_3 \rangle \\ & \cdots\cdots \\ & + (-1)^\nu \langle 0, 1, 2, \cdots, \nu \rangle, \end{aligned} \tag{2.2}$$

where, for example,

$$\langle k_1, k_2, k_3 \rangle = P\{E_{k_1}^\nu \cap E_{k_2}^\nu \cap E_{k_3}^\nu\}.$$

We now discuss a reduction of (2.2), leading to (2.5) when $m = 2$. Consider any index set $(k_1, \cdots, k_h)$, to be denoted by $\mathscr{K}$. Denote the corresponding probability $\langle k_1, \cdots, k_h \rangle$ by $\langle \mathscr{K} \rangle$ and the corresponding subset $(p_{k_1}, \cdots, p_{k_h})$ of G_ν by $\mathscr{P}(\mathscr{K})$. Let $\mathscr{C}(\mathscr{K})$ be the convex hull of $\mathscr{P}(\mathscr{K})$ and let $\mathscr{S}(\mathscr{K})$ be the set of extreme points of $\mathscr{C}(\mathscr{K})$.

It is useful to single out sets $\mathscr{K}$ with the property that

$$\mathscr{C}(\mathscr{K}) \cap [G_\nu - \mathscr{S}(\mathscr{K})] \neq \varnothing. \tag{2.3}$$

Probabilities $\langle \mathscr{K} \rangle$ whose corresponding index sets $\mathscr{K}$ satisfy (2.3) cancel in (2.2);

hence they can be excluded from the summations in (2.2). In other words, the summations in (2.2) need be extended only over index sets $\mathcal{K}$ satisfying

$$\mathcal{C}(\mathcal{K}) \cap [G_v - \mathcal{S}(\mathcal{K})] = \varnothing. \tag{2.4}$$

The cancellation of terms $\langle \mathcal{K} \rangle$ with $\mathcal{K}$ satisfying (2.3) is due to two facts:

(i) In view of the concavity of f over σ_m, $[\mathcal{S}(\mathcal{K}_1) = \mathcal{S}(\mathcal{K}_2)]$ implies $[\langle \mathcal{K}_1 \rangle = \langle \mathcal{K}_2 \rangle]$. In other words, *given* $\mathcal{S}$, $\langle \mathcal{K} \rangle$ *is the same for all index sets* $\mathcal{K}$ *for which* $\mathcal{S}(\mathcal{K}) = \mathcal{S}$.

(ii) For *given* $\mathcal{S}$, *consider all index sets* $\mathcal{K}$ *for which* $\mathcal{S}(\mathcal{K}) = \mathcal{S}$ *and for which* (2.3) *is satisfied. Of the probabilities* $\langle \mathcal{K} \rangle$ *in* (2.2) *corresponding to these index sets, half appear in subtracted sums, and half in added sums.* For example, if $G_v - \mathcal{S} = (p, p')$, then the index sets $\mathcal{K}$ with $\mathcal{S}(\mathcal{K}) = \mathcal{S}$ and $\mathcal{S}$, $\mathcal{S} \cup (p)$, $\mathcal{S} \cup (p')$, and $\mathcal{S} \cup (p, p')$, and it is clear that, if the term $\langle \cdots \rangle$ of (2.2) corresponding to $\mathcal{S}$ appears in a subtracted sum, then so must the term $\langle \cdots \rangle$ corresponding to $\mathcal{S} \cup (p, p')$, while the terms $\langle \cdots \rangle$ corresponding to $\mathcal{S} \cup (p)$ and $\mathcal{S} \cup (p')$ must appear in added sums.

We next consider expression (2.2), and its reduction by cancellation, for the special case $m = 2$, with G_v the grid composed of the evenly spaced points $p_k = (k/v, 1 - k/v)$. For this case, the preceding considerations specialize as follows. Assuming first that $\mathcal{K}$ is not a singleton, we have that $\mathcal{C}(\mathcal{K})$ is the closed line segment $\mathcal{I}(\mathcal{K})$ just containing all the points of $\mathcal{P}(\mathcal{K})$, $\mathcal{S}(\mathcal{K})$ consists of the endpoints of $\mathcal{I}(\mathcal{K})$ and condition (2.4) specifies that, excepting its endpoints, $\mathcal{I}(\mathcal{K})$ contains no points of G_v. Again, if $\mathcal{K}$ contains only k, then $\mathcal{C}(\mathcal{K}) = \mathcal{S}(\mathcal{K}) = (p_k)$, and (2.4) is met. In other words, (2.4) is satisfied either when $\mathcal{K}$ is a singleton or when $\mathcal{P}(\mathcal{K})$ consists of two adjacent points of the grid, and the cancellation of terms $\langle \cdots \rangle$ corresponding to all other index sets $\mathcal{K}$ reduces (2.2) to the more manageable expression

$$\sum_{k=0}^{v} \langle k \rangle - \sum_{k=0}^{v-1} \langle k, k+1 \rangle. \tag{2.5}$$

Further, the summands of the second sum in (2.5) may be written

$$\begin{aligned}\langle k, k+1 \rangle = P\Bigg\{ & \left(\frac{k}{v}\right) X_{11} + \left(1 - \frac{k}{v}\right) X_{21} > v, \left(\frac{k+1}{v}\right) X_{11} \\ & + \left(1 - \left(\frac{k+1}{v}\right)\right) X_{21} > v; \left(\frac{k}{v}\right) X_{12} + \left(1 - \frac{k}{v}\right) X_{22} > v, \\ & \left(\frac{k+1}{v}\right) X_{12} + \left(1 - \left(\frac{k+1}{v}\right)\right) X_{22} > v; \cdots; \left(\frac{k}{v}\right) X_{1n} \\ & + \left(1 - \frac{k}{v}\right) X_{2n} > v, \left(\frac{k+1}{v}\right) X_{1n} + \left(1 - \left(\frac{k+1}{v}\right)\right) X_{2n} > v \Bigg\},\end{aligned}$$

and similarly for the summands $\langle k \rangle$.

Now define:

F: A probability distribution on the plane,

$A(p, v)$: the region $px_1 + (1 - p)x_2 > v$,

$B(p, v)$: the region $px_1 + (1 - p)x_2 \leqq v, x_2 \geqq v$,

$C(p_1, p_2, v)$: the region $p_1x_1 + (1 - p_1)x_2 > v, p_2x_1 + (1 - p_2)x_2 > v$,

$$\alpha(p, v, F): P_F\{A(p, v)\},$$
$$\beta(p, v, F): P_F\{B(p, v)\},$$
$$\gamma(p_1, p_2, v, F): P_F\{C(p_1, p_2, v)\}.$$

Then, assuming independence of columns, i.e., among pairs (X_{1j}, X_{2j}), and supposing that (X_{1j}, X_{2j}) is distributed according to F_j, expression (2.5), and hence $P\{E^v\}$, may be written

$$\sum_{k=0}^{v-1}\left[\prod_{j=1}^{n}\alpha\left(\frac{k}{v}, v, F_j\right) - \prod_{j=1}^{n}\gamma\left(\frac{k}{v}, \frac{k+1}{v}, v, F_j\right)\right] + \prod_{j=1}^{n}\alpha(1, v, F_j).$$

In accordance with (2.1), we must now take this to the limit with v. To this end we need the following lemma.

LEMMA 1. *Suppose that $\{\Phi_v(p)\}$ is a sequence of continuous functions on $[0, 1]$, converging uniformly to $\Phi(p)$. Let $\{F_j; j = 1, \cdots, n\}$ be any set of bivariate distributions such that the n functions $\alpha(p, v, F_j)$ are continuous in p for $p \in [0, 1]$ and the n functions $\gamma(p_1, p_2, v, F)$ are uniformly (with respect to $p_1 \in [0, 1]$) continuous in p_2 at $p_2 = p_1$. Then*

$$\lim_{v\to\infty}\sum_{k=0}^{v-1}\Phi_v\left(\frac{k}{v}\right)\left[\prod_{j=1}^{n}\alpha\left(\frac{k}{v}, v, F_j\right) - \prod_{j=1}^{n}\gamma\left(\frac{k}{v}, \frac{k+1}{v}, v, F_j\right)\right]$$
$$= \sum_{j=1}^{n}\int_0^1 \Phi(p)\prod_{i\neq j}\alpha(p, v, F_i)\, d\beta(p, v, F_j). \tag{2.6}$$

Before proceeding to the induction on n that verifies the lemma note that the introduction of the sequence $\{\Phi_v\}$ is not an idle complication beyond what we actually require; rather, the Φ_v seem required in the induction argument. Also, to abbreviate, set

$$\prod_{i\neq j}\alpha(p, v, F_i) \equiv \partial(p, j, n). \tag{2.7}$$

Proof. To begin with, the right-hand side of (2.6) is well-defined, since $\Phi(z) \cdot \partial(z, j, n)$ is continuous and $\beta(z, v, F_j)$ is nondecreasing.

For $n = 1$, we must show that the limit with v of

$$\left|\sum_{k=0}^{v-1}\Phi_v\left(\frac{k}{v}\right)\left[\alpha\left(\frac{k}{v}, v, F\right) - \gamma\left(\frac{k}{v}, \frac{k+1}{v}, v, F\right)\right] - \int_0^1 \Phi(p)\, d\beta(p, v, F)\right| \tag{2.8}$$

is zero. But

$$\alpha\left(\frac{k}{v}, v, F\right) - \gamma\left(\frac{k}{v}, \frac{k+1}{v}, v, F\right) = \beta\left(\frac{k+1}{v}, v, F\right) - \beta\left(\frac{k}{v}, v, F\right),$$

so that (2.8) is bounded by

$$\left|\sum_{k=0}^{v-1}\Phi\left(\frac{k}{v}\right)\left[\beta\left(\frac{k+1}{v}, v, F\right) - \beta\left(\frac{k}{v}, v, F\right)\right] - \int_0^1 \Phi(p)\, d\beta(p, v, F)\right|$$
$$+\left[\sup_{0\leq k\leq v-1}\left|\Phi_v\left(\frac{k}{v}\right) - \Phi\left(\frac{k}{v}\right)\right|\right]\cdot\beta(1, v, F),$$

of which the first term tends to zero by the continuity of Φ and monotonicity of β, and the last by the uniformity of the convergence of the Φ_v.

To complete the induction using evident abbreviations we write the expression on the left-hand side of (2.6):

$$\sum_{k=0}^{v-1} \Phi_v\left(\frac{k}{v}\right)\left[\prod_{j=1}^{n} \alpha_j - \prod_{j=1}^{n} \gamma_j\right] = \sum_{k=0}^{v-1} \left[\Phi_v\left(\frac{k}{v}\right)\right]\left[\frac{\alpha_n + \gamma_n}{2}\right]\left[\prod_{j=1}^{n-1} \alpha_j - \prod_{j=1}^{n-1} \gamma_j\right]$$

$$+ \sum_{k=0}^{v-1} \left[\Phi_v\left(\frac{k}{v}\right)\right]\left[\left(\prod_{j=1}^{n-1} \alpha_j + \prod_{j=1}^{n-1} \gamma_j\right)\Big/2\right]\left[\alpha_n - \gamma_n\right]. \tag{2.9}$$

Regarding the first term on the right-hand side of (2.9), set

$$\Phi_v^*(p) \equiv \Phi_v(p) \cdot [(\alpha(p, v, F_n) + \gamma(p, p + 1/v, v, F_n))/2],$$

so that product of the first two braces of the summands of that term may be written $\Phi_v^*(k/v)$; and $\Phi_v^*(p)$, in view of the assumed uniform continuity, tends uniformly with v to

$$\Phi^*(p) \equiv \Phi(p)\alpha(p, v, F_n). \tag{2.10}$$

Hence the induction hypothesis for $n - 1$ implies that the first term of (2.9) tends with v to

$$\sum_{j=1}^{n-1} \int_0^1 \Phi^*(p)\partial(p, j, n - 1)\, d\beta(p, v, F_j). \tag{2.11}$$

Similarly, by using the induction hypothesis for $n = 1$ and defining

$$\Phi_v^{**}(p) \equiv \Phi_v(p) \cdot \left[\left(\prod_{j=1}^{n-1} \alpha(p, v, F_j) + \prod_{j=1}^{n-1} \gamma\left(p, p + \frac{1}{v}, v, F_j\right)\right)\Big/2\right],$$

$\Phi_v^{**}(p)$ tends uniformly to

$$\Phi^{**}(p) \equiv \Phi(p) \prod_{j=1}^{n-1} \alpha(p, v, F_j), \tag{2.12}$$

and the second term on the right-hand side of (2.9) tends to

$$\int_0^1 \Phi^{**}(p)\, d\beta(p, v, F_n). \tag{2.13}$$

By using (2.7), (2.10) and (2.12) it is clear that (2.11) and (2.13) add to the desired right-hand side of (2.6).

Substituting $\Phi_v(p) \equiv 1$ in Lemma 1 now gives the limit required in (2.1) and the conclusion that

$$P\{V > v\} = \sum_{j=1}^{n} \int_0^1 \prod_{i \neq j} \alpha(p, v, F_i)\, d\beta(p, v, F_j) + \prod_{i=1}^{n} \alpha(1, v, F_i) \tag{2.14}$$

when the columns of $\| X_{ij} \|$ are independently distributed according to distributions F_j satisfying the continuity conditions of Lemma 1.

Defining

$$W_j(p, v) = \begin{cases} \beta(p, v, F_j) & \text{for } p \in [0, 1), \\ \alpha(1, v, F_j)/n & \text{for } p = 1, \end{cases}$$

we can abbreviate (2.14) to

$$P\{V > v\} = \sum_{j=1}^{n} \int_0^1 \prod_{i \neq j} \alpha(p, v, F_i)\, dW_j(p, v). \tag{2.15}$$

Finally, define

$B^*(p, v)$: the region $px_1 + (1 - p)x_2 \geqq v, x_2 \leqq v$,
$\beta^*(p, v, F) : P_F\{B^*(p, v)\}$.

Since

$$\int_0^1 \partial(p, j, n)\, d\beta^*(p, v, F_j) - \int_0^1 \partial(p, j, n)\, d\beta(p, v, F_j)$$

$$= \int_0^1 \partial(p, j, n)\, d\alpha(p, v, F_j)$$

$$= \prod_{i=1}^{n} \alpha(1, v, F_i) - \prod_{i=1}^{n} \alpha(0, v, F_i),$$

an expression alternate to (2.15) is

$$P\{V > v\} = \sum_{j=1}^{n} \int_0^1 \prod_{i \neq j} \alpha(p, v, F_i)\, dW_j^*(p, v), \tag{2.16}$$

where

$$W_j^*(p, v) = \begin{cases} \beta^*(p, v, F_j) & \text{for } p \in (0, 1], \\ \alpha(0, v, F_j)/n & \text{for } p = 0. \end{cases}$$

We denote the right-hand side of (2.14), (2.15) or (2.16) by $J(v; F_1, \cdots, F_n)$.

3. Implications. Consider the linear program (LP):

$$\min \{u_1 + u_2\},$$
$$X_{1j}u_1 + X_{2j}u_2 \geqq 1, \quad j = 1, \cdots, n, \quad u_1, u_2 \geqq 0. \tag{3.1}$$

It is well known that, for $yz = 1$, LP has the optimum $z > 0$ if and only if the game G with matrix $\|X_{ij}\|$ has the value $y > 0$. Hence assuming independence among pairs (X_{1j}, X_{2j}) with (X_{1j}, X_{2j}) distributed according to F_j satisfying the continuity conditions of Lemma 1, we have, for $t > 0$,

$$P\{\text{optimum of LP} \in (0, t) | F_1, \cdots, F_n\} = P\{V > 1/t | F_1, \cdots, F_n\}$$
$$\equiv J(1/t; F_1, \cdots, F_n). \tag{3.2}$$

Consider now the program (1.2). With the change of variables $u_i = b_i y_i$, that program may be written:

$$\min \{u_1 + u_2\}$$
$$(a_{1j}/d_{1j})u_1 + (a_{2j}/d_{2j})u_2 \geqq 1, \quad j = 1, \cdots, n, \quad u_1, u_2 \geqq 0,$$

where $d_{ij} = b_i c_j$ and so is, in fact, essentially of form (3.1). Hence suppose independence among the pairs of coefficients (a_{1j}, a_{2j}) for the program (1.2), and denote the distribution of $(a_{1j}/d_{1j}, a_{2j}/d_{2j})$ by F_j. Then, if the F_j satisfy the continuity conditions of Lemma 1, the probability that the optimum M of (1.2) is in $(0, t)$ is given by $J(1/t; F_1, \cdots, F_n)$. Now the optimum M must be positive, and the alternative to positive M is that (1.2) is infeasible. Hence our stochastic analysis of (1.2) is completed by noting that

$$P\{(1.2) \text{ infeasible}\} = 1 - \lim_{s \to 0+} J(s; F_1, \cdots, F_n). \tag{3.3}$$

Section 2 also provides some information about the geometric behavior of $P\{V > v\}$ or $P\{M \in (0, t)\}$. For example, concerning the former, suppose that all F_j are the same, say F, and that F, besides satisfying the continuity conditions of Lemma 1, also is such that $\beta(p, v, F)$ is strictly increasing for $p \in [0, 1]$. Then by defining

$$\rho = \max_{0 \leqq p \leqq 1} \alpha(p, v, F),$$

it is clear that $P\{V > v\}$ is geometric in ρ, in the sense that, provided $\rho > 0$,

$$\lim_{n \to \infty} \frac{P\{V > v\}}{(\rho + \varepsilon)^n} = \lim_{n \to \infty} \frac{(\rho - \varepsilon)^n}{P\{V > v\}} = 0.$$

The values of the two $1 \times n$ games corresponding respectively to the two rows of $\| X_{ij} \|$ are of course geometric in $\alpha(0, v, F)$ and $\alpha(1, v, F)$ respectively.

Finally, it is of interest to verify the second formula of [2] for $m = 2$. This is readily done by choosing the F_j circular and centered at (v, v), in which case (2.14) yields

$$P\{V > v\} = (n)(\tfrac{1}{2})^{n-1}(\tfrac{1}{4}) + (\tfrac{1}{2})^n,$$

as required.

REFERENCES

[1] BERNARD BEREANU, *On stochastic linear programming; The Laplace transform of the distribution of the optimum and applications*, J. Math. Anal. Appl., 15 (1966), pp. 280–294.

[2] THOMAS M. COVER, *The probability that a random game is unfair*, Ann. Math. Statist., 37 (1966), pp. 1796–1799.

[3] PAUL R. HALMOS, *Measure Theory*, Van Nostrand, New York, 1950.

[4] ANDRÁS PRÉKOPA, *On the probability distribution of the optimum of a random linear program*, this Journal, 4 (1966), pp. 211–222.

[5] JATI K. SENGUPTA AND GERHARD TINTNER, *The approach of stochastic linear programming: Methods and computation*, submitted to Management Sci.

[6] DONALD J. SOULTS, *Asymptotic value distributions for matrix games*, Doctoral thesis, Iowa State University, Ames, 1968.

[7] DAVID R. THOMAS AND H. T. DAVID, *Game value distributions. I*, Ann. Math. Statist., 38 (1967), pp. 242–250.

[8] DAVID R. THOMAS, *Game value distributions. II*, Ibid., 38 (1967), pp. 251–260.

SIAM J. CONTROL
Vol. 9, No. 1, February 1971

ERRATUM: OPTIMAL CONTROLS FOR SYSTEMS WITH TIME LAG*

A. HALANAY

In § 4, p. 231 the inequality expressing the maximum principle should read

$$
\begin{aligned}
&H(t, \tilde{x}(t), \tilde{x}(\beta_1(t)), \cdots, \tilde{x}(\beta_k(t)), u, \tilde{u}(\beta_1(t)), \cdots, \tilde{u}(\beta_k(t)), \tilde{b}) \\
&\quad + H(\gamma_1(t), \tilde{x}(\gamma_1(t)), \tilde{x}(t), \cdots, \tilde{x}(\beta_k(\gamma_1(t))), \tilde{u}(\gamma_1(t)), u, \cdots, \\
&\qquad\qquad \tilde{u}(\beta_k(\gamma_1(t))), \tilde{b})\dot{\gamma}_1(t) \\
&\quad + \cdots \\
&\quad + H(\gamma_k(t), \tilde{x}(\gamma_k(t)), \tilde{x}(\beta_1(\gamma_k(t))), \cdots, \tilde{x}(t), \tilde{u}(\gamma_k(t)), \tilde{u}(\beta_1(\gamma_k(t))), \cdots, u, \tilde{b})\dot{\gamma}_k(t) \\
&\quad \leqq \text{the same sum where } u \text{ is replaced by } \tilde{u}(t).
\end{aligned}
$$

On p. 233 the coordinate k^ρ should be

$$
\begin{aligned}
k^\rho &= F_\rho(t, \tilde{x}(t), \cdots, \tilde{x}(\beta_k(t)), u, \tilde{u}(\beta_1(t)), \cdots, \tilde{u}(\beta_k(t)), \tilde{b}) \\
&\quad - F_\rho(t, \tilde{x}(t), \cdots, \tilde{x}(\beta_k(t)), \tilde{u}(t), \tilde{u}(\beta_1(t)), \cdots, \tilde{u}(\beta_k(t)), \tilde{b}) \\
&\quad + [F_\rho(\gamma_1(t), \tilde{x}(\gamma_1(t)), \tilde{x}(t), \cdots, \tilde{x}(\beta_k(\gamma_1(t))), \tilde{u}(\gamma_1(t)), u, \cdots, \tilde{u}(\beta_k(\gamma_1(t))), \tilde{b}) \\
&\qquad - F_\rho(\gamma_1(t), \tilde{x}(\gamma_1(t)), \tilde{x}(t), \cdots, \tilde{x}(\beta_k(\gamma_1(t))), \tilde{u}(\gamma_1(t)), \tilde{u}(t), \cdots, \\
&\qquad\qquad \tilde{u}(\beta_k(\gamma_1(t))), \tilde{b})]\dot{\gamma}_1(t) \\
&\quad + \cdots \\
&\quad + [F_\rho(\gamma_k(t), \tilde{x}(\gamma_k(t)), \cdots, \tilde{x}(t), \tilde{u}(\gamma_k(t)), \cdots, u, \tilde{b}) \\
&\qquad - F_\rho(\gamma_k(t), \tilde{x}(\gamma_k(t)), \cdots, \tilde{x}(t), \tilde{u}(\gamma_k(t)), \cdots, \tilde{u}(t), \tilde{b})]\dot{\gamma}_k(t).
\end{aligned}
$$

* This Journal, 6 (1968), pp. 215–234.

SIAM J. CONTROL
Vol. 9, No. 2, May 1971

OUTPUT CONTROLLABILITY AND SYSTEM SYNTHESIS*

A. S. MORSE†

Abstract. The geometric theory of linear multivariable systems is extended by introducing the concept of a controllable output subspace. Necessary and sufficient conditions for an output subspace to be controllable are given. As an example application, controllable output subspaces are used to solve a generalized state-feedback decoupling problem.

1. Introduction. It has recently been shown that the geometric theory of [1] can be successfully applied to a large class of problems in linear systems synthesis. In this article, the theory is extended to encompass systems of the form

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{1}$$

$$y(t) = Cx(t) + Du(t) \tag{2}$$

which differ from those previously considered by the presence of D in (2). The extended theory is based on the concept of a controllable output subspace which is discussed in the next section. This concept is closely related to the idea of a controllability subspace which plays a key role in a number of problems in linear systems synthesis. As an application of the extended theory, conditions under which the above system can be decoupled with state feedback are found.

With regard to (1) and (2), x is an n-vector, u an m-vector and y a p-vector. All matrices above and below are constant, real-valued and of appropriate size. Script letters are used for real vector spaces; $\mathscr{X}$, $\mathscr{U}$, $\mathscr{Y}$ are the state, control and output spaces respectively for (1) and (2). Matrices and their maps are denoted by the same symbol. For a linear map $M:\mathscr{W} \to \mathscr{Z}$, $\{M\}$ or $\mathscr{M}$ denotes the range of M. The notation $\{A|\mathscr{B}\}$ is defined as

$$\{A|\mathscr{B}\} = \mathscr{B} + A\mathscr{B} + \cdots + A^{n-1}\mathscr{B}.$$

It will be convenient to regard $\mathscr{X}$ and $\mathscr{Y}$ as subspaces of $\mathscr{E} \equiv \mathscr{X} \oplus \mathscr{Y}$. With $\mathscr{E}$ so defined, P (respectively Q) denotes the projection from $\mathscr{E}$ onto $\mathscr{X}$ (respectively $\mathscr{Y}$) along $\mathscr{Y}$ (respectively $\mathscr{X}$). Use will be made of the extended maps

$$\bar{A}:\mathscr{E} \to \mathscr{E}, \quad e \mapsto APe;$$

$$\bar{B}:\mathscr{U} \to \mathscr{E}, \quad u \mapsto Bu;$$

$$\bar{C}:\mathscr{E} \to \mathscr{E}, \quad e \mapsto CPe;$$

$$\bar{D}:\mathscr{U} \to \mathscr{E}, \quad u \mapsto Du.$$

Note, in particular, that $\bar{A} = \bar{A}P = P\bar{A}$, $\bar{C} = Q\bar{C} = \bar{C}P$, $\bar{B} = P\bar{B}$ and $\bar{D} = Q\bar{D}$. Below A (respectively B, C, D) and its extension $\bar{A}$ (respectively $\bar{B}$, $\bar{C}$, $\bar{D}$) will be denoted by the same symbol A (respectively B, C, D). The interpretation, in each instance, will be clear from the context.

* Received by the editors May 26, 1970.

† Office of Control Theory and Application, NASA Electronics Research Center, Cambridge, Massachusetts. Now at Department of Engineering and Applied Science, Yale University, New Haven, Connecticut 06520.

The following development is based on the results of [1] which are assumed to be known.

2. Controllable output subspaces. With regard to (1), consider the control

$$u(t) = Fx(t) + Gv(t), \tag{3}$$

where $v \in \mathscr{V}$ (q-space), $G : \mathscr{V} \to \mathscr{U}$ and

$$F : \mathscr{E} \to \mathscr{U}. \tag{4}$$

For the extended map F in (4) to be consistent with its associated matrix in (3), the former must satisfy

$$F = FP. \tag{5}$$

With (3) applied to (1), the largest subspace $\mathscr{R} \subset \mathscr{X}$ which v can completely control is given by

$$\mathscr{R} = \{A + BF|\{BG\}\}. \tag{6}$$

If for fixed $A, B, \mathscr{R} \subset \mathscr{X}$, a pair (F, G) exists for which (6) is true, then $\mathscr{R}$ is called a controllability subspace of (A, B). Existence conditions providing a characterization for controllability subspaces are given in [1].

In [1]–[3], the fundamental role played by controllability subspaces in linear systems synthesis is discussed. However, to study systems with output relations of the form in (2), an additional concept is required. Observe that application of (3) to (1) and (2) affects not only state controllability, but output controllability as well. Thus if $\mathscr{S}$ is the largest subspace of $\mathscr{Y}$ which v can completely control, then clearly

$$\mathscr{S} = (C + DF)\mathscr{R} + \{DG\} \tag{7}$$

with $\mathscr{R}$ given by (6). This suggests a problem similar to the controllability subspace problem of [1]. Namely, for fixed $\mathscr{S}, \mathscr{Y}$, find conditions for the existence of a pair (F, G) and a subspace $\mathscr{R} \subset \mathscr{X}$ for which both (6) and (7) are satisfied.

Conditions (6) and (7) may be simplified.

LEMMA 1. *Let the maps $\hat{A} : \mathscr{E} \to \mathscr{E}$, $\hat{B} : \mathscr{U} \to \mathscr{E}$ be fixed and suppose $\hat{A} = \hat{A}P$. If for fixed G,*

$$\mathscr{R} \equiv \{P\hat{A}|P\{\hat{B}G\}\}, \qquad \mathscr{S} \equiv Q\hat{A}\mathscr{R} + Q\{\hat{B}G\}, \tag{8}$$

then

$$\mathscr{R} = P\{\hat{A}|\hat{\mathscr{B}} \cap (\mathscr{R} + \mathscr{S})\}, \qquad \mathscr{S} = Q\{\hat{A}|\hat{\mathscr{B}} \cap (\mathscr{R} + \mathscr{S})\}. \tag{9}$$

Conversely, if $\mathscr{R}$ and $\mathscr{S}$ are fixed subspaces satisfying (9), *there exists a G for which* (8) *is true.*

Proof. From the relation $\hat{A} = \hat{A}P$ follow the identities

$$\begin{aligned} P\{\hat{A}|\{\hat{B}G\}\} &\equiv \{P\hat{A}|P\{\hat{B}G\}\}, \\ Q\{\hat{A}|\{\hat{B}G\}\} &\equiv Q\hat{A}\{P\hat{A}|P\{\hat{B}G\}\} + Q\{\hat{B}G\}. \end{aligned} \tag{10}$$

Write $\mathscr{W} \equiv \{\hat{A}|\{\hat{B}G\}\}$ and $\mathscr{T} \equiv \{\hat{A}|\hat{\mathscr{B}} \cap (\mathscr{R} + \mathscr{S})\}$.

If (8) is true, (10) provides $\mathscr{R} = P\mathscr{W}$ and $\mathscr{S} = Q\mathscr{W}$. Thus $\mathscr{W} \subset \mathscr{R} + \mathscr{S}$. But $\{\hat{B}G\} \subset \hat{\mathscr{B}} \cap \mathscr{W} \subset \hat{\mathscr{B}} \cap (\mathscr{R} + \mathscr{S})$, so $\mathscr{W} \subset \{\hat{A}|\hat{\mathscr{B}} \cap (\mathscr{R} + \mathscr{S})\} = \mathscr{T}$. Now $\hat{\mathscr{B}} \cap (\mathscr{R} + \mathscr{S}) \subset \mathscr{R} + \mathscr{S}$ and if $\mathscr{V} \subset \mathscr{R} + \mathscr{S}$, $\hat{A}\mathscr{V} \subset \mathscr{R} + \mathscr{S}$. Thus $\mathscr{T} \subset \mathscr{R} + \mathscr{S}$. There follows $\mathscr{R} = P\mathscr{W} \subset P\mathscr{T} \subset \mathscr{R}$ and $\mathscr{S} \subset Q\mathscr{W} \subset Q\mathscr{T} \subset \mathscr{S}$. Therefore $\mathscr{R} = P\mathscr{T}$, $\mathscr{S} = Q\mathscr{T}$, and (9) is true.

Conversely, if (9) holds, $\mathscr{R} = P\mathscr{T}$ and $\mathscr{S} = Q\mathscr{T}$. By Lemma 4.1 of [1], there exists a G such that $\mathscr{T} = \{\hat{A}|\{\hat{B}G\}\}$. It follows from this and (10), that (8) is true. This completes the proof.

Recalling constraint (5) and applying Lemma 1 with

$$\hat{A} \equiv (A + BF) + (C + DF),$$

$$\hat{B} \equiv B + D$$

we can state the existence question as follows.

Let $A, B, C, D, \mathscr{S} \subset \mathscr{Y}$ be fixed. Find conditions for the existence of a map $F: \mathscr{U} \to \mathscr{E}$ and a subspace $\mathscr{T} \subset \mathscr{E}$ such that

$$F = FP, \tag{11}$$

$$\mathscr{S} = Q\mathscr{T}, \tag{12}$$

$$\mathscr{T} = \{A + C + (B + D)F|\{B + D\} \cap (P\mathscr{T} + \mathscr{S})\}. \tag{13}$$

If F and $\mathscr{T}$ satisfying (11)–(13) exist, $\mathscr{S}$ will be called a *controllable output subspace* of (A, B, C, D). The subspace $\mathscr{R} \equiv P\mathscr{T}$ will be called a *generator* of $\mathscr{S}$.

It is possible to solve this problem and thus to characterize controllable output subspaces. Write $\overline{\mathscr{T}}(\mathscr{W})$ for the maximal controllability subspace of $(A + C, B + D)$ contained in a fixed space $\mathscr{W} \subset \mathscr{E}$. A procedure for finding $\overline{\mathscr{T}}(\mathscr{W})$ is given in [1]. Our main result is now presented.

THEOREM 1. *Let (A, B, C, D), $\mathscr{R} \subset \mathscr{X}$, $\mathscr{S} \subset \mathscr{Y}$ be fixed. Then $\mathscr{S}$ is a controllable output subspace with generator $\mathscr{R}$ if and only if*

$$\mathscr{S} = Q\overline{\mathscr{T}}(\mathscr{R} + \mathscr{S}), \qquad \mathscr{R} = P\overline{\mathscr{T}}(\mathscr{R} + \mathscr{S}). \tag{14}$$

Proof. If $\mathscr{S}$ is a controllable output subspace with generator $\mathscr{R}$, then there exist F and $\mathscr{T}$ satisfying (11)–(13) with $\mathscr{R} = P\mathscr{T}$. Thus,

$$\begin{aligned}(A + C + (B + D)F)(\mathscr{R} + \mathscr{S}) &= (A + C + (B + D)F)P(\mathscr{R} + \mathscr{S}) \\ &= (A + C + (B + D)F)\mathscr{T} \\ &\subset \mathscr{T} \subset \mathscr{R} + \mathscr{S}.\end{aligned}$$

By Theorem 4.3 of [1], the controllability subspace $\mathscr{T}$ is maximal relative to $\mathscr{R} + \mathscr{S}$. That is, $\mathscr{T} = \overline{\mathscr{T}}(\mathscr{R} + \mathscr{S})$ and (14) is true.

Conversely, if (14) holds, $\mathscr{S}$ will be a controllable output subspace with $\mathscr{R}$ as a generator if there exists an F satisfying (11) for which

$$\overline{\mathscr{T}}(\mathscr{R} + \mathscr{S}) = \{A + C + (B + D)F|\{B + D\} \cap (\mathscr{R} + \mathscr{S})\}. \tag{15}$$

Since $\overline{\mathscr{T}} \equiv \overline{\mathscr{T}}(\mathscr{R} + \mathscr{S})$ is a controllability subspace, by Theorem 4.1 of [1], $(A + C)\overline{\mathscr{T}} \subset \{B + D\} + \overline{\mathscr{T}}$. This and (14) provide

$$(A + C)\mathscr{R} = (A + C)\overline{\mathscr{T}} \subset \{B + D\} + \overline{\mathscr{T}} \subset \{B + D\} + \mathscr{R} + \mathscr{S}.$$

By Lemma 3.2 of [1], there exists an $\bar{F}$ such that $(A + C + (B + D)\bar{F})\mathscr{R} \subset \mathscr{R} + \mathscr{S}$. If $F \equiv \bar{F}P$, then

$$(A + C + (B + D)F)(\mathscr{R} + \mathscr{S}) = (A + C + (B + D)\bar{F})\mathscr{R} \subset \mathscr{R} + \mathscr{S}.$$

By Theorem 4.3 of [1], (15) holds for this choice of F. In addition, F satisfies (11) as required. This completes the proof.

In general, a fixed controllable output subspace may have many generators. It is useful, therefore, to characterize controllable output subspaces without a hypothesis involving generators. This is easily accomplished.

COROLLARY 1. *Let (A, B, C, D), $\mathscr{S} \subset \mathscr{Y}$ be fixed. Then $\mathscr{S}$ is a controllable output subspace if and only if there exists a controllability subspace $\mathscr{T}$ of $(A + C, B + D)$ such that*

$$\mathscr{S} = Q\mathscr{T}. \tag{16}$$

If (16) *holds, $\mathscr{R} \equiv P\mathscr{T}$ is a generator of $\mathscr{S}$.*

Proof. Necessity is obvious. Take $\mathscr{T} = \bar{\mathscr{T}}(\mathscr{R} + \mathscr{S})$ and apply Theorem 1.

For sufficiency, note that $\mathscr{T} \subset \mathscr{R} + \mathscr{S}$. Thus $\mathscr{T} \subset \bar{\mathscr{T}}(\mathscr{R} + \mathscr{S})$, so $\mathscr{S} = Q\mathscr{T} \subset Q\bar{\mathscr{T}}(\mathscr{R} + \mathscr{S}) \subset \mathscr{S}$ and $\mathscr{R} = P\mathscr{T} \subset P\bar{\mathscr{T}}(\mathscr{R} + \mathscr{S}) \subset \mathscr{R}$. It follows that $\mathscr{R}$ and $\mathscr{S}$ satisfy Theorem 1 which, in turn, provides the desired result.

Remark 1. It is clear from Corollary 1 that $\mathscr{S}$ is a controllable output subspace if and only if $\mathscr{S} = Q\bar{\mathscr{T}}(\mathscr{X} + \mathscr{S})$.

Remark 2. For fixed $\mathscr{R}$ and $\mathscr{S}$ satisfying (14), let $\mathbf{F}$ be the class of F for which (11) and (13) hold with $\mathscr{T} = \bar{\mathscr{T}}(\mathscr{R} + \mathscr{S})$. By Theorem 1, this class is nonempty. Starting with the Theorem 4.2 of [1], it is a simple matter to show that the spectrum of $A + BF$ restricted to $\mathscr{R}$ can be freely assigned by suitable choice of $F \in \mathbf{F}$.

3. Application to state-feedback decoupling. Suppose in (2), the system output y consists of k subvectors

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix}$$

which are to be independently controlled. Algebraically, the structure of y is equivalent to the assumption

$$\mathscr{Y} = \mathscr{Y}_1 \oplus \cdots \oplus \mathscr{Y}_k, \tag{17}$$

where $\mathscr{Y}_i$ is the output subspace associated with y_i. Consider a control of the form

$$u(t) = Fx(t) + \sum_{i=1}^{k} G_i v_i(t). \tag{18}$$

The object of decoupling is to choose F and the G_i so that v_i can control y_i without influencing y_j for $j \neq i$. If $\mathscr{R}_i$ is the controllability subspace of (A, B) associated with v_i, these requirements dictate that

$$\begin{aligned} \mathscr{R}_i &= \{A + BF|\{BG_i\}\}, && i = 1, \cdots, k, \\ \mathscr{Y}_i &= (C + DF)\mathscr{R}_i + \{DG_i\}, && i = 1, \cdots, k. \end{aligned} \tag{19}$$

Utilizing the results of the last section, we may formulate the decoupling problem as follows.

Find conditions for the existence of a map $F:\mathscr{E} \to \mathscr{U}$ *and subspaces* $\mathscr{T}_i$, $i = 1, \cdots, k$, *such that*

$$F = FP, \tag{20}$$

$$\mathscr{T}_i = \{A + C + (B + D)F | \{B + D\} \cap \mathscr{T}_i\}, \qquad i = 1, \cdots, k, \tag{21}$$

$$\mathscr{Y}_i = Q\mathscr{T}_i, \qquad i = 1, \cdots, k. \tag{22}$$

In short, the $\mathscr{Y}_i$ must be controllable output subspaces and they must all be constructed with the same F.

To study this problem, use will be made of the following lemma.

LEMMA 2. *Let* $\mathscr{T}_i$, $i = 1, \cdots, k$, *be fixed subspaces satisfying* (22). *Write* $\overline{\mathbf{F}}$ *for the class of* F *for which* (21) *holds and assume* $\overline{\mathbf{F}}$ *is nonempty. Then there exists an* $F \in \mathbf{F}$ *for which* (20) *is true.*

Proof. Let $F_0 \in \overline{\mathbf{F}}$. If $F_0 = F_0P$, set $F = F_0$. Suppose $F_0 \neq F_0P$. Write $\mathscr{T}_i = \hat{\mathscr{T}}_i \oplus \mathscr{T}_i \cap \mathscr{Y}$, $i = 1, \cdots, k$, where $\hat{\mathscr{T}}_i$ is an arbitrary completion. Since $\mathscr{Y}_i = Q\mathscr{T}_i$, $\mathscr{T}_i \subset \mathscr{X} + \mathscr{Y}_i$. Thus,

$$\mathscr{Y} \cap \sum_{i=1}^{k} \hat{\mathscr{T}}_i = \mathscr{Y} \cap \sum_{i=1}^{k} (\mathscr{Y}_i + \mathscr{X}) \cap \hat{\mathscr{T}}_i$$

$$= \sum_{i=1}^{k} \hat{\mathscr{T}}_i \cap \mathscr{Y} = 0.$$

Therefore the space $\sum_{i=1}^{k} \hat{\mathscr{T}}_i$ is isomorphic to its projection on $\mathscr{X}$ and there exists an F_1 such that

$$F_0 \sum_{i=1}^{k} \hat{\mathscr{T}}_i = F_1P \sum_{i=1}^{k} \hat{\mathscr{T}}_i.$$

Write $F = F_1P$. Then for $i = 1, \cdots, k$,

$$(A + C + (B + D)F)\mathscr{T}_i = (A + C + (B + D)F_0)\hat{\mathscr{T}}_i \subset \mathscr{T}_i.$$

By Lemma 4.3 of [1], $F \in \overline{\mathbf{F}}$. In addition, $F = FP$ as required. This completes the proof.

Using Lemma 2, we can now translate the decoupling problem (20)–(22) into a formulation for which the results of [1] are directly applicable. Define

$$\mathscr{N}_i \equiv \mathscr{X} + \sum_{j \neq i} \mathscr{Y}_j, \qquad i = 1, \cdots, k. \tag{23}$$

The translated formulation is as follows.

Find conditions for the existence of a map $F:\mathscr{E} \to \mathscr{U}$ *and subspaces* $\mathscr{T}_i$, $i = 1, \cdots, k$, *such that*

$$\mathscr{T}_i = \{A + C + (B + D)F | \{B + D\} \cap \mathscr{T}_i\}, \qquad i = 1, \cdots, k, \tag{24}$$

$$\mathscr{T}_i + \mathscr{N}_i = \mathscr{E}, \qquad i = 1, \cdots, k, \tag{25}$$

$$\mathscr{T}_i \subset \bigcap_{j \neq i} \mathscr{N}_j, \qquad i = 1, \cdots, k. \tag{26}$$

It may easily be checked that (25) and (26) are equivalent to (22). If this problem has a solution, Lemma 2 insures the existence of an $F \in \bar{\mathbf{F}}$ for which (20) is true.

Theorem 7.1 of [1] provides the solution to this problem for the case when the number of independent open-loop system inputs equals the number of output subvectors to be controlled. This is equivalent to the assumption

$$\dim(\{B + D\}) = k. \tag{27}$$

We now give the solution to the decoupling problem, stated in notation consistent with (20)–(22).

THEOREM 2. *Let* (27) *hold. The state-feedback decoupling problem has a solution if and only if*

$$Q\bar{\mathscr{T}}_i(\mathscr{Y}_i + \mathscr{X}) = \mathscr{Y}_i, \qquad i = 1, \cdots, k,$$

and

$$\{B + D\} = \sum_{i=1}^{k} (\{B + D\} \cap \bar{\mathscr{T}}_i(\mathscr{Y}_i + \mathscr{X})).$$

Furthermore, if $F, \mathscr{T}_i \cdots \mathscr{T}_k$ *is any solution,*

$$\mathscr{T}_i = \bar{\mathscr{T}}_i(\mathscr{Y}_i + \mathscr{X}), \qquad i = 1, \cdots, k.$$

Remark 3. Theorem 2 asserts that if a solution exists the $\mathscr{T}_i$ are unique. It follows that the generators $\mathscr{R}_i = P\mathscr{T}_i$ are also unique.

4. Conclusion. The state-feedback decoupling problem is by no means the only problem to which the above ideas may be applied. With some effort, controllable output subspaces may be used to extend the general decoupling problem of [2], the triangular decoupling problem of [3] and perhaps the problem of system inversion [4].

REFERENCES

[1] W. M. WONHAM AND A. S. MORSE, *Decoupling and pole assignment in linear multivariable systems: A geometric approach*, this Journal, 8 (1970), pp. 1–18.

[2] A. S. MORSE AND W. M. WONHAM, *Decoupling and pole assignment by dynamic compensation*, this Journal, 8 (1970), pp. 317–337.

[3] ———, *Triangular decoupling of linear multivariable systems*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 447–449.

[4] L. M. SILVERMAN, *Inversion of multivariable linear systems*, Ibid., AC-14 (1969), pp. 270–277.

SIAM J. CONTROL
Vol. 9, No. 2, May 1971

ON INFORMATION STRUCTURES, FEEDBACK AND CAUSALITY*

H. S. WITSENHAUSEN†

Abstract. A finite number of decisions, indexed by $\alpha \in A$, are to be taken. Each decision amounts to selecting a point in a measurable space $(U_\alpha, \mathscr{F}_\alpha)$. Each decision is based on some information fed back from the system and characterized by a subfield $\mathscr{I}_\alpha$ of the product space $(\prod_\alpha U_\alpha, \prod_\alpha \mathscr{F}_\alpha)$. The decision function for each α can be any function γ_α measurable from $\mathscr{I}_\alpha$ to $\mathscr{F}_\alpha$.

A property of the $\{\mathscr{I}_\alpha\}_{\alpha \in A}$ is defined which assures that the setup has a causal interpretation. This property implies that for any combination of choices of the γ_α, the closed loop equations have a unique solution.

The converse implication is false, when card $A > 2$.

1. Introduction. In control-oriented works on dynamic games (in particular, stochastic control problems) one usually finds a "dynamic equation" describing the evolution of a "state" in response to decision (control) variables of the players and to random variables. One also finds "output equations" which define output variables for a player as functions of the state, decision and random variables. Then the information structure is defined by allowing each decision variable to be any desired (measurable) function of the output variables generated for that player up to that time.

Such a setup assumes that the time order in which the various decision variables are selected is fixed in advance. It assumes that each player acts as if he had responsibility only for one station. It assumes that this station has perfect memory.

For large complex systems these tacit assumptions are unlikely to hold. In modeling more general situations it is often natural to consider the number of decisions to be programmed as finite, even though the alternatives to be considered can be infinite in number at each decision. But in the larger systems each "player" is in fact some form of organization unified only by its aims and central preplanning of its policy. Implementation of policy is left for execution by "agents" (which, of course, may be devices). The order in which the various agents of the various organizations will have to act cannot always be predicted, and the information available to different agents, even of the same organization, may be noncomparable in the sense that, of two agents, neither one knows everything his colleague knows.

These difficulties in specifying the information structure of a game were faced and overcome in the early days of game theory [1], [2].

In this paper a different approach is proposed. The decision process is considered as a feedback loop and the game is characterized by its interaction with the policies of the agents, without prejudging questions of chronological order. In this way the relation between causality and the existence of a unique solution for any combination of policies can be brought into focus.

2. The classical theory. Von Neumann and Morgenstern [1] introduced a general description of games in extensive form for the case where the number of

* Received by the editors January 15, 1970.

† Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07974.

decisions is finite and all variables range over finite sets. Their description is in terms of the agents constituting a player's organization. When the umpire calls one of these agents in the course of the play, he provides him with certain data about the situation and the agent must, on this basis, make a decision among a set of alternatives. The agent's decision function is decided in advance by the player. The set of these functions for all his agents constitutes a player's pure strategy.

One weakness of this setup was the assumption that an agent always knows the value of the time variable (move sequence number). Soon afterwards Kuhn [2] introduced a more general description which removes this restriction. A key device of Kuhn is that one separate agent is provided for each possible set of data. Such an agent will select an alternative, according to prior instructions from a player, if and when he is called by the umpire. No data is furnished to the agent. In other words, the decision functions are defined pointwise, with a separate agent for each point. Most information structure theory [3]–[5], [10], [11] is based on Kuhn's model.

The "atomization" of agents in this model is unaccetable when continuous variables are considered for the choice variables. To define the expected payoff for given pure strategies in a stochastic game, measurability on σ-fields smaller than the power sets is essential for the decision functions. For this reason a new formulation of Von Neumann type, with the additional feature of measurability, was constructed by Aumann [6].

For the purposes of the present paper, the Von Neumann and Aumann formulations are inadequate because they assume a fixed known sequencing of the decisions. The Kuhn formulation is inadequate for two reasons. First, the tree describing the game is an expression of the general solution of the closed loop relations. (These relations map information into decisions by the policies, and decisions into information by the rules of the game.) For any combination of policies one can find the corresponding outcome by following the tree along selected branches, and this is an explicit procedure. Thus the difficulties that might arise in solving the loop have been eliminated by defining the game in terms of a general unique solution which must be found before the model can be set up.

Second, as Aumann has pointed out, once decisions are selected from a continuum, measurability conditions must be imposed and this is not conveniently feasible with a tree type model.

The "team theory" of Radner [9] is less general than the Von Neumann description for a single player. Radner assumes that there is an initial chance move and that the players' agents receive information only about this chance move, not about each other's actions. This absence of feedback would be inappropriate in most dynamic situations.

3. Games as multiple feedback loops. As is customary, random effects are modeled as due to the mixed (i.e., randomized) strategy of a player called "Chance," who may have several agents. Once this is done the system upon which the agents act is completely deterministic.

Let $n > 0$ be the finite cardinality of the set A of all agents in the game. Assume, without loss of generality, that each agent α in A will take one decision during the game. If there are circumstances where an agent is never called to action, then one can still let him make a decision which will have no effect. If an agent

were to act more than once, then either the information available on all these occasions is the same, and then these decisions are considered as a single one picked from a product set, or else the information is not the same, and then one provides a separate agent for each occasion.

Agent α must select his decision from a nonempty set U_α on which is given a σ-field $\mathscr{F}_\alpha$. The policy of the agent will be specified as the dependence of the decision upon information available to him and upon the outputs of random devices that may be used to mix a player's strategy.

The crucial observation is that all variables in the game will be determined by the decisions actually taken, that is, by the element u of the Cartesian product $U = \prod_{\alpha \in A} U_\alpha$. Note that this Cartesian product consists of those maps u of A into the union of the U_α which satisfy $u(\alpha) \equiv u_\alpha \in U_\alpha$ for all α in A. This definition does not require any ordering of the set A.

For a nonvoid subset B of A, denote by P_B the projection of the product $\prod_{\alpha \in A} U_\alpha$ upon the product $\prod_{\alpha \in B} U_\alpha$. The projection $P_B(u)$ is just the restriction of u to the domain B. Let $P_\alpha(u) \equiv P_{\{\alpha\}}(u) \equiv u_\alpha$. Let $\prod_{\alpha \in B} \mathscr{F}_\alpha$ be the product σ-field on $\prod_{\alpha \in B} U_\alpha$ and let $\mathscr{F}(B)$ be the smallest σ-field on U such that P_B is measurable. Note that $B_1 \subset B_2$ implies $\mathscr{F}(B_1) \subset \mathscr{F}(B_2)$. The σ-field $\mathscr{F}(A) = \prod_{\alpha \in A} \mathscr{F}_\alpha$, also denoted by $\mathscr{F}$, is the finest field to be considered on U.

The outcome of the game, that is, the list of payoffs to the various players, is determined by u. These payoffs are functions from U into real numbers, measurable on $\mathscr{F}$. Likewise, any data, about the situation, received by agent α, is information about u and will determine a σ-field $\mathscr{I}_\alpha \subset \mathscr{F}$, the *information field* of agent α.

For a nonstochastic game and if only pure strategies are considered, the policy that agent α is assigned by his organization can be any function $\gamma_\alpha : U \to U_\alpha$ measurable as a map of the measurable space $(U, \mathscr{I}_\alpha)$ into $(U_\alpha, \mathscr{F}_\alpha)$. Denoting by Γ_α this set of functions, let $\Gamma = \prod_{\alpha \in A} \Gamma_\alpha$.

Then for the combination of policies $\gamma \in \Gamma$ one has the problem of solving the closed loop equations

$$P_\alpha(u) = \gamma_\alpha(u) \quad \text{for all } \alpha \in A. \tag{1}$$

When random effects are present, a sample space $(\Omega, \mathscr{B}, P)$ for the joint behavior of all random devices is introduced. $\mathscr{B}$ is a σ-field on the set Ω and P a probability measure on $(\Omega, \mathscr{B})$. For each agent a σ-field $\mathscr{B}_\alpha \subset \mathscr{B}$ is given. The possible policies of agent α are the functions

$$\gamma_\alpha : U \times \Omega \to U_\alpha$$

such that $\gamma_\alpha^{-1}(\mathscr{F}_\alpha) \subset \mathscr{I}_\alpha \times \mathscr{B}_\alpha$. Again Γ_α will denote this set of functions and $\Gamma = \prod_{\alpha \in A} \Gamma_\alpha$.

Then for a given combination of policies $\gamma \in \Gamma$ the problem is to find, for each ω in Ω, solutions u (dependent on ω) to the closed loop equations

$$P_\alpha(u) = \gamma_\alpha(u, \omega) \quad \text{for all } \alpha \in A. \tag{2}$$

One would like that there be one and only one solution for each ω, with $\mathscr{B}$-measurable ω-dependence. One also would like this solution to be obtained by actual play of the extensive form of the game in a way compatible with causality.

None of these properties will hold in general for an arbitrary *information structure* characterized by nonvoid sets A, U_α and σ-fields $\mathscr{F}_\alpha$, $\mathscr{I}_\alpha$, with the possible additional specification of Ω, $\mathscr{B}$ and $\mathscr{B}_\alpha$.

For such a structure to become a game one would have to specify a partition of A into players, a probability measure P on $(\Omega, \mathscr{B})$, the mixed strategy of the chance player's agents and the payoff function for each other player. The selection of $(\Omega, \mathscr{B}, \mathscr{P})$ and $\mathscr{B}_\alpha$ must take into account that each player or coalition of players uses an independent random device and that the chance player must be considered as a separate coalition in that respect.

Only the information structure is of concern in this paper.

4. Some basic questions. What condition on the fields $\mathscr{I}_\alpha$ expresses the causality requirement? Since the requirement is a heuristic one, a formal condition (property C below) will be stated and strong arguments given to relate it to the idea of causality.

Define *property* S (solvability) as follows: For each $\gamma \in \Gamma$ and $\omega \in \Omega$ there exists one and only one $u \in U$ satisfying the closed loop equations.

If property S holds, then for each $\gamma \in \Gamma$ the solution defines a map M^γ from Ω into U. *Property* SM holds when this map is always measurable from $(\Omega, \mathscr{B})$ into $(U, \mathscr{F})$. The importance of this property is that it makes the payoff of each player measurable on the sample space, so that the expected payoff can be defined.

Another important consequence of property SM is that the induced fields $\mathscr{I}_\alpha^\gamma$, inverse images under M^γ of $\mathscr{I}_\alpha$, are contained in $\mathscr{B}$. Thus conditional expectations can be defined with respect to $\mathscr{I}_\alpha^\gamma$. Conditional expectations with respect to $\mathscr{I}_\alpha$ are meaningless because this field is not on the probability space Ω. On the other hand, the meaningful conditioning on $\mathscr{I}_\alpha^\gamma$ yields results that may actually depend on γ. Casual treatment of this point has been a source of confusion in stochastic control theory.

The following relations hold among these properties: C implies SM but, surprisingly, SM does not imply C when more than two agents are involved.

5. The causality property. The main difficulty is one of notation. Let n be the cardinality of the set A of agents, and for $1 \leqq k \leqq n$ denote by S_k the set of all injections of $\{1, 2, \cdots, k\}$ into A, i.e., the arrangements without repetition of k elements from A. Let $S^+ = \bigcup_{k=1}^n S_k$, $S_0 = \{\varnothing\}$ and $S = S_0 \cup S^+$.

For $0 \leqq i \leqq j \leqq n$ let $T_i^j \colon S_j \to S_i$ be the truncation map which restricts $s \in S_j$ to the domain $\{1, \cdots, i\}$ or to $\varnothing$ for $i = 0$.

For $s \in S$ denote by $|s|$ the range of s and by $\|s\|$ the cardinality of $|s|$. Note that $s \in S_{\|s\|}$.

It will be assumed that *for all α in A, and x in U_α, the singleton set $\{x\}$ belongs to $\mathscr{F}_\alpha$*. As a consequence, the singletons of $\prod_{\alpha \in B} U_\alpha$ belong to $\prod_{\alpha \in B} \mathscr{F}_\alpha$. In particular, the singletons of U belong to $\mathscr{F}$.

Physically, decisions can be considered as events in space-time. Then the causality condition means that, for any play of the game, the actions of the agents can be ordered and that the information available to an agent may depend on decisions of agents acting earlier but cannot depend upon the decisions of agents acting at concurrent or later times. In the language of relativity theory two agents act at concurrent times if the interval separating their actions is of spatial, as

opposed to temporal, character. For different plays of the game the order may turn out to be different because the decisions made were different and these decisions affect the space-time positions of other decisions.

Thus the causality condition requires that the field $\mathscr{I}_\alpha$, subject to the condition that the set of agents acting before α is B, is contained in the field $\mathscr{F}(B)$ determined by the decisions of the earlier agents.

A way of stating this requirement precisely is the following.

Property C. An information structure $\{U_\alpha, \mathscr{F}_\alpha, \mathscr{I}_\alpha\}_{\alpha \in A}$ is said to possess property C when there exists a map $\varphi : U \to S_n$ such that for all k, s, E satisfying $1 \leqq k \leqq n$, $s = (\alpha_1, \cdots, \alpha_k) \in S_k$, $E \in \mathscr{I}_{\alpha_k}$, one has

$$E \cap (T_k^n \circ \varphi)^{-1}(s) \in \mathscr{F}(|T_{k-1}^k(s)|). \tag{3}$$

Note that in general a set E in $\mathscr{I}_\alpha$ need not belong to $\mathscr{F}(B)$ for any proper subset B of $A - \{\alpha\}$. Neither is it necessary that the set $(T_k^n \circ \varphi)^{-1}(s)$ belong to $\mathscr{I}_\alpha$ for any α. Also note that for $k = 1$ one has $T_0^1(s) = \varnothing$ and $\mathscr{F}(\varnothing) = \{\varnothing, U\}$.

In abbreviated notation the property can be stated as the existence of φ such that for $1 \leqq k \leqq n$ and $s = (\alpha_1, \cdots, \alpha_k)$ in S_k, one has

$$\mathscr{I}_{\alpha_k} \cap (T_k^n \circ \varphi)^{-1}(s) \subset \mathscr{F}(\alpha_1, \cdots, \alpha_{k-1}).$$

The function φ need not be unique, if only because a transposition of concurrently acting agents cannot have any effect.

LEMMA 1. *If property* C *holds with* φ, *then for all* $s \in S^+$,

$$(T_{\|s\|}^n \circ \varphi)^{-1}(s) \in \mathscr{F}(|T_{\|s\|-1}^{\|s\|}(s)|). \tag{4}$$

Proof. For $s \in S_k$ one has $\|s\| = k$ and since U belongs to $\mathscr{I}_\alpha$ for all α, (3) with $E = U$ yields (4).

This lemma says that after $k - 1$ agents have taken their decisions, the selection of the next agent to act is thereby determined mathematically. None of the agents, though, need to know this selection. Even the selected agent need not know that he is the kth one to act nor which $k - 1$ agents have already acted.

LEMMA 2. *The function* $T_1^n \circ \varphi$ *has a constant value* (α^*) *and* $\mathscr{I}_{\alpha^*} = \{\varnothing, U\}$.

Proof. Lemma 1, with $\|s\| = 1$, shows that $T_1^n \circ \varphi$ is measurable on $\mathscr{F}(\varnothing) = \{\varnothing, U\}$ and is therefore a constant. Denoting this constant by (α^*), property C with $k = 1$ and $s = (\alpha^*)$ gives for any set E in $\mathscr{I}_{\alpha^*}$,

$$E \in \mathscr{F}(\varnothing) = \{\varnothing, U\}$$

which shows that $\mathscr{I}_{\alpha^*} = \{\varnothing, U\}$, completing the proof.

The interpretation of Lemma 2 is that there has to be at least one "starting agent" and that a starting agent cannot have any information about decisions taken since none have been taken yet. For instance, many games begin with either one of two agents depending on "the toss." In that case the starting agent is a chance agent who decides according to a fixed mixed strategy the outcome of the toss. The selection of the next agent then depends on this outcome that is on the first agent's decision.

The same argument as in Lemma 2 shows that the function $T_2^n \circ \varphi$ is measurable on $\mathscr{F}(\{\alpha^*\})$.

6. The umpire's information fields. Continuing with the assumption that property C holds with function φ, the possibility for the umpire to organize the play in an orderly manner is to be demonstrated. To this end, introduce the *umpire fields* $\mathscr{F}^{(k)}$ defined as follows:

$$\begin{aligned} &\mathscr{F}^{(0)} \equiv \{\varnothing, U\}, \\ &\mathscr{F}^{(k)} \equiv \left\{ \bigcup_{s\in S_k} [(T_k^n \circ \varphi)^{-1}(s) \cap F_s] \middle| \text{for all } s \in S_k, F_s \in \mathscr{F}(|s|) \right\} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } 1 \leqq k \leqq n. \end{aligned} \tag{5}$$

Heuristically $\mathscr{F}^{(k)}$ is the field characterizing the information available to the umpire after he has seen k agents take their decisions.

LEMMA 3. *The collections of sets* $\mathscr{F}^{(k)}$ *are* σ*-fields on* U. $\mathscr{F}^{(1)} = \mathscr{F}(\{\alpha^*\})$, $\mathscr{F}^{(k)} \subset \mathscr{F}$ *and* $\mathscr{F}^{(n)} = \mathscr{F}$.

Proof. That the $\mathscr{F}^{(k)}$ are indeed σ-fields follows from Auxiliary Lemma 1 (see Appendix) by specialization, noting that $\{(T_k^n \circ \varphi)^{-1}(s) | s \in S_k\}$ is a collection of pairwise disjoint sets covering U. For $k = \|s\| = 1$ this collection has only one nonempty member $(T_1^n \circ \varphi)^{-1}(\alpha^*) = U$. Thus

$$\mathscr{F}^{(1)} = \{F_{(\alpha^*)} | F_{(\alpha^*)} \in \mathscr{F}(\{\alpha^*\})\} = \mathscr{F}(\{\alpha^*\}).$$

Each set F_s in (5) belongs to a subfield of $\mathscr{F}$. The same is true for each set $(T_k^n \circ \varphi)^{-1}(s)$ by virtue of Lemma 1. Thus each set in $\mathscr{F}^{(k)}$ belongs to $\mathscr{F}$.

For $k = \|s\| = n$ one has $|s| = A$ and $\mathscr{F}(|s|) = \mathscr{F}$. Therefore, given a set F in $\mathscr{F}$, one may take $F_s = F$ for all $s \in S_n$ in (5). This represents F as a set in $\mathscr{F}^{(n)}$. Hence $\mathscr{F} = \mathscr{F}^{(n)}$.

A most important fact is that the umpire fields are nested.

LEMMA 4. $\mathscr{F}^{(k)} \subset \mathscr{F}^{(k+1)}$ *for* $0 \leqq k < n$.

Proof. The assertion is trivial for $k = 0$. For $0 < k < n$ and $E \in \mathscr{F}^{(k)}$, one has

$$E = \bigcup_{s\in S_k} [(T_k^n \circ \varphi)^{-1}(s) \cap F_s]$$

with $F_s \in \mathscr{F}(|s|)$ for all $s \in S_k$.

Now since $T_k^n = T_k^{k+1} \circ T_{k+1}^n$, one has

$$(T_k^n \circ \varphi)^{-1} = (T_k^{k+1} \circ T_{k+1}^n \circ \varphi)^{-1} = (T_{k+1}^n \circ \varphi)^{-1} \circ (T_k^{k+1})^{-1}$$

for the inverse set mappings. Thus

$$(T_k^n \circ \varphi)^{-1}(s) = \bigcup_{s'} \{(T_{k+1}^n \circ \varphi)^{-1}(s') | s' \in S_{k+1}, s = T_k^{k+1}(s')\}.$$

Then

$$\begin{aligned} E &= \bigcup_{s\in S_k} [F_s \cap \bigcup_{s'} \{(T_{k+1}^n \circ \varphi)^{-1}(s') | s' \in S_{k+1}, s = T_k^{k+1}(s')\}] \\ &= \bigcup_{s} \bigcup_{s'} \{F_s \cap (T_{k+1}^n \circ \varphi)^{-1}(s') | s \in S_k, s' \in S_{k+1}, s \in T_k^{k+1}(s')\} \\ &= \bigcup_{s'\in S_{k+1}} [F_{T_k^{k+1}(s')} \cap (T_{k+1}^n \circ \varphi)^{-1}(s')]. \end{aligned}$$

But if $s = T_k^{k+1}(s')$ then, as sets, $|s| \subset |s'|$, and therefore $\mathscr{F}(|s|) \subset \mathscr{F}(|s'|)$ so that

$$F_{T_k^{k+1}(s')} \in \mathscr{F}(|s'|).$$

Comparison with (5) shows that $E \in \mathscr{F}^{(k+1)}$ which proves the lemma.

Heuristically, Lemma 4 says that the umpire enjoys perfect recall. The next lemma says that when the umpire has obtained decisions of $k-1$ agents he is able to tell which agent comes into play next.

LEMMA 5. *For* $0 < k \leqq n$, $T_k^n \circ \varphi$ *is measurable on* $\mathscr{F}^{(k-1)}$.

Proof. It suffices to show that for any s' in S_k the set $(T_k^n \circ \varphi)^{-1}(s')$ is of the form

$$\bigcup_{s \in S_{k-1}} [(T_{k-1}^n \circ \varphi)^{-1}(s) \cap F_s] \tag{6}$$

with F_s in $\mathscr{F}(|s|)$ for all s in S_{k-1}.

Let $s^* = T_{k-1}^k(s') \in S_{k-1}$ and let $F_s = (T_k^n \circ \varphi)^{-1}(s')$ for $s = s^*$, noting that, by Lemma 1, this set is in $\mathscr{F}(|s^*|)$, and $F_s = \varnothing$ for $s \in S_{k-1} - \{s^*\}$.

Then (6) becomes

$$(T_{k-1}^n \circ \varphi)^{-1}(s^*) \cap (T_k^n \circ \varphi)^{-1}(s').$$

Since $s^* = T_{k-1}^k(s')$ the first of these sets contains the second and the intersection is just the second set, establishing the claim.

7. The causal solution process. Assuming that property C holds with function φ, it will now be shown that for $\gamma \in \Gamma$ the closed loop equations

$$P_\alpha(u) = \gamma_\alpha(u, \omega) \quad \text{for all } \alpha \in A \tag{7}$$

admit, for each ω, a unique solution and the dependence of this solution upon ω is measurable. Furthermore the solution process can be organized in a recursive way, corresponding to causal play of the game under the direction of the umpire.

LEMMA 6. *Equations* (7) *cannot admit two distinct solutions* u, $\bar{u}$ *for the same decision functions* γ *and the same value of* ω.

Proof. Since ω is fixed, one need only consider the section $\gamma_\alpha(\cdot, \omega)$ of the functions γ_α. These sections are $\mathscr{I}_\alpha$-measurable and will be denoted by $\gamma_\alpha(\cdot)$ for brevity.

Assume then that for some u, $\bar{u} \in U$ and for all α in A, one has both $P_\alpha(u) = \gamma_\alpha(u)$ and $P_\alpha(\bar{u}) = \gamma_\alpha(\bar{u})$.

By Lemma 2, the relations

$$T_k^n(\varphi(u)) = T_k^n(\varphi(\bar{u})) = s = (\alpha_1, \cdots, \alpha_k), \tag{8}$$

$$P_{\alpha_i}(u) = P_{\alpha_i}(\bar{u}), \qquad i = 1, \cdots, k-1, \tag{9}$$

hold trivially for $k = 1$. But when (8), (9) hold for some k one has

$$P_{\alpha_k}(u) = P_{\alpha_k}(\bar{u}). \tag{10}$$

Indeed, otherwise the sets

$$E = \gamma_{\alpha_k}^{-1}(P_{\alpha_k}(u)) \quad \text{and} \quad \bar{E} = \gamma_{\alpha_k}^{-1}(P_{\alpha_k}(\bar{u}))$$

would be disjoint sets in the field $\mathscr{I}_{\alpha_k}$. This is so because of the assumption that the singleton sets $\{P_{\alpha_k}(u)\}$ and $\{P_{\alpha_k}(\bar{u})\}$ in U_{α_k} belong to $\mathscr{F}_{\alpha_k}$. Then by property C, the set

$$F = E \cap (T_k^n \circ \varphi)^{-1}(s),$$

which contains u and the set

$$\bar{F} = \bar{E} \cap (T_k^n \circ \varphi)^{-1}(s),$$

which contains $\bar{u}$, would be disjoint sets in the field $\mathcal{F}(\{\alpha_1, \cdots, \alpha_{k-1}\})$. But since that field is generated by the function $P_{\{\alpha_1,\cdots,\alpha_{k-1}\}}$ which takes the same value at u and $\bar{u}$, a contradiction of Auxiliary Lemma 2 (see Appendix) is obtained.

If $k = n$, then (8), (9), (10) state that $u = \bar{u}$; if $k < n$, then one will have

$$s' \equiv T_{k+1}^n(\varphi(u)) = T_{k+1}^n(\varphi(\bar{u})) = \bar{s'}. \tag{11}$$

Indeed otherwise u and $\bar{u}$ would belong respectively to the disjoint sets $(T_{k+1}^n \circ \varphi)^{-1}(s')$ and $(T_{k+1}^n \circ \varphi)^{-1}(\bar{s'})$. By Lemma 1 these two sets both belong to the field $\mathcal{F}(\{\alpha_1, \cdots, \alpha_k\})$. This field is determined by the function $P_{\{\alpha_1,\cdots,\alpha_k\}}$, which by (9), (10) takes the same value at u and $\bar{u}$. Thus auxiliary Lemma 2 would again be contradicted.

Now with the adjunction of (10) and (11), the relations (8), (9) hold with k increased by unity. Thus Lemma 6 follows by induction.

The successive steps of the recursive solution process, given a set of maps $\gamma_\alpha : U \times \Omega \to U_\alpha$, are described by maps

$$M_k^\gamma : U \times \Omega \to U,$$

where the superscript γ refers to the system of decision functions and $k = 1, \cdots, n$. These maps are defined with the help of an arbitrary reference element $r \in U$.

The action of M_k^γ is defined as follows:

$$P_\alpha(M_k^\gamma(u, \omega)) = \begin{cases} P_\alpha(u) & \text{for } \alpha \in |T_{k-1}^n(\varphi(u))|, \\ \gamma_\alpha(u, \omega) & \text{for } \alpha = (\varphi(u))_k, \\ P_\alpha(r) & \text{otherwise.} \end{cases} \tag{12}$$

LEMMA 7. *The function M_k^γ is measurable, as a map of the measurable space $(U \times \Omega, \mathcal{F}^{(k-1)} \times \mathcal{B})$ into the measurable space $(U, \mathcal{F}^{(k)})$.*

Proof. The set U can be partitioned into the sets $(T_k^n \circ \varphi)^{-1}(s)$ with s ranging over S_k. By Lemma 5 these sets belong to $\mathcal{F}^{(k-1)}$ and, a fortiori, to $\mathcal{F}^{(k)}$ (Lemma 4).

The restriction of M_k^γ to each set in the partition leaves the sequence $s = (\alpha_1, \cdots, \alpha_k)$ unchanged because it preserves the values $u_{\alpha_1}, \cdots, u_{\alpha_{k-1}}$ upon which the restriction of $T_k^n \circ \varphi$ to the same domain depends.

On each set in the partition, the division of the agents into the three groups of (12) is fixed. By the definition of the umpire fields it thus suffices to check that for each $s = (\alpha_1, \cdots, \alpha_k)$ on S_k the restriction of M_k^γ to $(T_k^n \circ \varphi)^{-1}(s) \times \Omega$ is measurable from the field $\mathcal{F}(\{\alpha_1, \cdots, \alpha_{k-1}\}) \times \mathcal{B}$ to the field $\mathcal{F}(\{\alpha_1, \cdots, \alpha_k\})$. By definition of the latter field, it suffices to check the measurability of $P_{\alpha_i} \circ M_k^\gamma$, restricted to the same domain, for $i = 1, \cdots, k$. Now for $i = 1, \cdots, k-1$ the restriction is the projection P_{α_i} defining $\mathcal{F}(\{\alpha_1, \cdots, \alpha_{k-1}\})$. For the index k the function γ_{α_k} is obtained. This function is measurable on $\mathcal{I}_{\alpha_k} \times \mathcal{B}$ and the intersections of sets in $\mathcal{I}_{\alpha_k}$ with $(T_k^n \circ \varphi)^{-1}(s)$ are in $\mathcal{F}(\{\alpha_1, \cdots, \alpha_{k-1}\})$ by virtue of property C. This establishes Lemma 7.

Heuristically, the maps M_k^γ correspond to the idea of state transition equations. From that point of view it is the field $\mathscr{F}^{(k)}$ that represents the state (of knowledge of the umpire).

THEOREM 1. *Property* C *implies property* SM.

Proof. Since M_1^γ is measurable on $\mathscr{F}^{(0)} \times \mathscr{B}$ and $\mathscr{F}^{(0)} = \{\varnothing, U\}$, this function may be considered as a measurable map $\overline{M}_1^\gamma$ of $(\Omega, \mathscr{B})$ into $(U, \mathscr{F}^{(1)})$. For $1 < k \leqq n$ define, recursively,

$$\overline{M}_k^\gamma : (\Omega, \mathscr{B}) \to (U, \mathscr{F}^{(k)})$$

by

$$\overline{M}_k^\gamma(\omega) = M_k^\gamma(\overline{M}_{k-1}^\gamma(\omega), \omega). \tag{13}$$

Then $\overline{M}_n^\gamma$ is the desired solution map. Indeed, the procedure (13) establishes in succession and leaves thereafter invariant the following equations:

$$\begin{aligned} \alpha_1 &= (\varphi(u))_1 = \alpha^*, \\ u_{\alpha_1} &= \gamma_{\alpha_1}(u, \omega), \\ &\cdots \\ \alpha_n &= (\varphi(u))_n, \\ u_{\alpha_n} &= \gamma_{\alpha_n}(u, \omega). \end{aligned} \tag{14}$$

Therefore, the closed loop equations (7) will be satisfied.

The solution is unique for each ω by virtue of Lemma 6.

Finally, by auxiliary Lemma 3, the composition of measurable functions yields measurable functions

$$\overline{M}_k^\gamma : (\Omega, \mathscr{B}) \to (U, \mathscr{F}^{(k)}).$$

For $k = n$ one has $\mathscr{F}^{(k)} = \mathscr{F}$ (Lemma 3), so that the solution map is measurable as claimed.

8. Noncausal information structures. In exploring the possible validity of the implication S $\Rightarrow$ C, only sections $\gamma_\alpha(\cdot, \omega)$ for fixed $\omega \in \Omega$ are involved. Therefore the ω-dependence of γ will be suppressed.

The precise situation is described by the following theorem.

THEOREM 2. *Property* S *implies property* C *when* $n = 1$ *or* 2. *The implication is false for* $n > 2$.

Proof. (i) Suppose $n = 1$, and property S holds. Then $A = \{\alpha\}$, $U_\alpha = U$, $\mathscr{I}_\alpha \subset \mathscr{F}_\alpha = \mathscr{F}$. If a set E in $\mathscr{I}_\alpha$ is nonempty and has nonempty complement E^c, then choose $u_0 \in E$, $\bar{u}_0 \in E^c$ and $\gamma_\alpha(u) = u_0$ for $u \in E$, $\bar{u}_0$ otherwise. Then γ_α is measurable on $\mathscr{I}_\alpha$ and the closed loop equation $u = u_\alpha = \gamma_\alpha(u)$ has two distinct solutions u_0 and $\bar{u}_0$. Since this contradicts property S no such set E can exist, which means that $\mathscr{I}_\alpha = \{\varnothing, U\}$. Then property C holds as claimed. By interchanging $u_0, \bar{u}_0$ in the definition of γ_α a contradiction to existence, instead of to uniqueness, is obtained. The two types of contradictions correspond to the two forms of the classical liars paradox.

(ii) Suppose $n = 2$ and S holds. Then $A = \{\alpha, \beta\}$, $U = U_\alpha \times U_\beta$, $\mathscr{F} = \mathscr{F}_\alpha \times \mathscr{F}_\beta$, $\mathscr{I}_\alpha \subset \mathscr{F}$, $\mathscr{I}_\beta \subset \mathscr{F}$.

If for some fixed $u_\beta^* \in U_\beta$ one had a set E in $\mathscr{I}_\alpha$ such that both E and its complement E^c have a nonempty intersection with the set $\{u|P_\beta(u) = u_\beta^*\}$, then a contradiction to S is obtained by taking $\gamma_\beta(u) = u_\beta^*$ and constructing γ_α as in (i). Hence all sets E in $\mathscr{I}_\alpha$ are cylindrical, in the sense that $(u_\alpha, u_\beta) \in E$ implies $(\bar{u}_\alpha, u_\beta) \in E$ for all $\bar{u}_\alpha$ in U_α. This means that $\mathscr{I}_\alpha \subset \mathscr{F}(\{\beta\})$. Likewise $\mathscr{I}_\beta \subset \mathscr{F}(\{\alpha\})$. Therefore property C will hold if and only if at least one of $\mathscr{I}_\alpha$, $\mathscr{I}_\beta$ is the trivial field $\{\varnothing, U\}$. Suppose this is not the case. Then there are sets $E \in \mathscr{F}_\alpha$, $F \in \mathscr{F}_\beta$ such that $U_\alpha = E + E^c$, $U_\beta = F + F^c$ are proper partitions and $E \times U_\beta \in \mathscr{I}_\beta$, $F \times U_\alpha \in \mathscr{I}_\alpha$. Select $u_\alpha^0 \in E, \bar{u}_\alpha^0 \in E^c, u_\beta^0 \in F, \bar{u}_\beta^0 \in F^c$. Define $\gamma_\alpha(u) = u_\alpha^0$ for $u \in F \times U_\alpha$, $\bar{u}_\alpha^0$ otherwise, $\gamma_\beta(u) = u_\beta^0$ for $u \in E \times U_\beta$, $\bar{u}_\beta^0$ otherwise. These functions are measurable on the respective information fields and the closed loop equations admit the two distinct solutions $u^0 = (u_\alpha^0, u_\beta^0)$ and $\bar{u}^0 = (\bar{u}_\alpha^0, \bar{u}_\beta^0)$. This contradiction proves the claim.

(iii) That the implication does not hold for $n = 3$ (and, a fortiori, for larger n) is shown by the following counterexample which is adapted from a counterexample of R. L. Graham [7] to the corresponding conjecture in combinatorial set theory.

With $A = \{\alpha, \beta, \gamma\}$ let $U_\alpha = U_\beta = U_\gamma = \{\text{true, false}\}$ so that card $U = 8$ and $\mathscr{F} = 2^U$. The notation $x\bar{y}$ will denote the Boolean expression "x and (not y)." Let $\mathscr{I}_\alpha$ be the field generated by $u_\beta\bar{u}_\gamma$. It consists of $\varnothing$, U, $\{u|u_\beta\bar{u}_\gamma = \text{true}\}$, and $\{u|u_\beta\bar{u}_\gamma = \text{false}\}$. Likewise let $\mathscr{I}_\beta$ be generated by $u_\gamma\bar{u}_\alpha$ and $\mathscr{I}_\gamma$ by $u_\alpha\bar{u}_\beta$ Then each agent has a choice of four possible decision functions giving a total of 64 possible combinations. It is easily verified that for each of these combinations the closed loop equations are satisfied by exactly one of the eight elements of U. Since none of the information fields is trivial, property C does not hold while property S does.

Thus the normal form of a game with such a structure is perfectly well-defined. The strategies determine the outcome, but the extensive form of the game cannot be played without some form of precognition.

Since the counterexample used in Theorem 2, part (iii), is nonrandom it trivially satisfies property SM. This shows that property SM does not imply property C.

The last question is a purely technical one: when C is not assumed, does property S imply SM? In most special cases of interest the answer is yes, but in the generality of the present paper this assertion appears to be false, though a counterexample is not on paper.

9. Conclusions. The facts brought to light above are contributions towards a systematic information structure theory. They are entirely compatible with the more usual state equation–output equation formulation, as long as it is realized that (i) the classical information structure is a very special one, and (ii) output equations matter only by virtue of the fields they determine: different outputs determining the same fields are equivalent.

The information fields $\mathscr{I}_\alpha$ belong to the lattice of all subfields of $\mathscr{F}$. When games differ only as to information, relations between values and information can be derived accordingly.

Various important notions can easily be made precise within the given framework. For example, define a *station* as a set of agents, belonging to the same organization, such that the set of information fields of these agents is totally

ordered under inclusion (i.e., nested). A player (or coalition) enjoys *perfect recall* when his whole organization forms a single station. Then it is clear that the derivation of Aumann's theorem on behavior strategies in [6] holds for games satisfying property C when the player under consideration enjoys perfect recall as just defined.

A physical interpretation of the noncausal information structures with property SM, which exist by Theorem 2, appears rather difficult [12].

Appendix. A measurable space $(X, \mathscr{B})$ consists of a set X and a σ-field $\mathscr{B}$ of subsets of X. A function f from measurable space $(X, \mathscr{B})$ into measurable space $(Y, \mathscr{C})$ is called measurable when the inverse image under f of any set in $\mathscr{C}$ belongs to $\mathscr{B}$.

The following facts are simple exercises in Boolean algebra.

AUXILIARY LEMMA 1. *Let S be a nonempty set equal to the disjoint union of the sets $\{E_i | i \in I\}$, where I is a nonempty index set.*[1] *For each i in I let $\mathscr{F}_i$ be a σ-field on S (more generally an $\mathscr{S}$-class in the sense of Loève* [8, p. 59]*). Then the collection of sets*

$$\mathscr{C} = \left\{ \bigcup_{i \in I} (E_i \cap F_i) \middle| \text{ for all } i \in I, F_i \in \mathscr{F}_i \right\}$$

is a σ-field (respectively, $\mathscr{S}$-class) on S.

AUXILIARY LEMMA 2. *Let f be a function from set X into measurable space $(Y, \mathscr{C})$ and let $\mathscr{B} = f^{-1}(\mathscr{C})$ be the σ-field generated on X by f. Assume that for some x_1, x_2 in X, one has $f(x_1) = f(x_2)$. Then there is no set in $\mathscr{B}$ that contains x_1 but not x_2.*

AUXILIARY LEMMA 3. *Let $(X, \mathscr{B})$, $(Y_1, \mathscr{C}_1)$, $(Y_2, \mathscr{C}_2)$, $(Z, \mathscr{D})$ be measurable spaces and let $(Y_1 \times Y_2, \mathscr{C}_1 \times \mathscr{C}_2)$ denote the product space with the product field, the smallest field such that projections are measurable. Suppose the functions f_1: $(X, \mathscr{B}) \to (Y_1, \mathscr{C}_1)$, $f_2 : (X, \mathscr{B}) \to (Y_2, \mathscr{C}_2)$ and $g : (Y_1 \times Y_2, \mathscr{C}_1 \times \mathscr{C}_2) \to (Z, \mathscr{D})$ are measurable. Then the function $h : (X, \mathscr{B}) \to (Z, \mathscr{D})$ defined by $h(x) = g(f_1(x), f_2(x))$ is measurable.*

Acknowledgment. This paper was based on an invited contribution to the first International Conference on the Theory and Application of Differential Games, Amherst, Massachusetts, September 29, 1969.

REFERENCES

[1] J. VON NEUMANN AND O. MORGENSTERN, *The Theory of Games and Economic Behavior*, Princeton Univ. Press, Princeton, N.J., 1944, Chap. II.

[2] H. W. KUHN, *Extensive games and the problem of information*, Contributions to the Theory of Games, vol. 2, Annals of Mathematics Studies, no. 28, Princeton Univ. Press, Princeton, N.J., 1953, pp. 193–216.

[3] N. DALKEY, *Equivalence of information patterns and essentially determinate games*, Ibid., pp. 217–243.

[4] G. L. THOMPSON, *Signaling strategies in n-person games*, Ibid., pp. 267–277.

[5] B. J. BIRCH, *On games with almost complete information*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 275–287.

[1] Some of the E_i may be empty.

[6] R. J. AUMANN, *Mixed and behavior strategies in infinite extensive games*, Advances in Game Theory, Princeton Univ. Press, Princeton, N.J., 1964, pp. 627–650.
[7] R. L. GRAHAM, Private communication, 1969.
[8] M. LOEVE, *Probability Theory*, 3rd ed., Van Nostrand, Princeton, N.J., 1963.
[9] R. RADNER, *Team decision problems*, Ann. Math. Statist., 33 (1962), pp. 857–881.
[10] W. H. CLINGMAN, *Permanently optimal strategies in extensive games*, Proc. Amer. Math. Soc., 17 (1966), pp. 156–161.
[11] E. MARCHI, *A note on temporal games*, Z. Wahrscheinlichkeitstheorie Verw. Gebiete, 12 (1969), pp. 98–106.
[12] R. G. NEWTON, *Particles that travel faster than light?*, Science, 167 (1970), pp. 1569–1574.

SIAM J. CONTROL
Vol. 9, No. 2, May 1971

NECESSARY CONDITIONS FOR JOINING OPTIMAL SINGULAR AND NONSINGULAR SUBARCS*

J. P. McDANELL† AND W. F. POWERS‡

Abstract. Necessary conditions for the optimality of junctions between singular and nonsingular subarcs are developed for singular optimal control problems. Previously known necessary conditions concerning the continuity and smoothness of a piecewise analytic optimal control at a junction are clarified and extended. The main result is that the sum of the order of the singular arc and the lowest order time derivative of the control which is discontinuous at the junction must be an odd integer when the strengthened generalized Legendre-Clebsch condition is satisfied. Also, new necessary conditions which do not require an analyticity assumption are developed. These aid in characterizing problems which may possess nonanalytic junctions.

1. Introduction. Optimal control problems in which the control variables appear only linearly admit the possibility of the occurrence of singular extremals. The analysis of such problems is complicated by the fact that the solution, in general, consists of some combination of singular and nonsingular subarcs, the number and sequence of which are not known a priori. If the solution is totally singular, recent results [1], [2] are available to prove optimality in a large number of cases. If the solution is totally nonsingular, it is the familiar bang-bang control generated by a switching function with isolated zeros, as determined by the minimum principle. However, the mathematical characterization of optimal controls which contain both singular and nonsingular subarcs is far from complete.

This paper is concerned with the problem of characterizing the continuity and smoothness properties of the optimal control at a junction between singular and nonsingular subarcs. The analysis was motivated by the preliminary results obtained in this direction by Kelley, Kopp and Moyer [3] and Johnson [4]. We shall comment on their results in a later section.

2. Problem statement. The class of problems to which this analysis applies is the following. Determine the scalar control $u^*(t)$, $t \in [t_0, t_f]$, which minimizes the functional

$$J(u) = G(t_f, x(t_f)) + \int_{t_0}^{t_f} [L_0(t, x) + L_u(t, x)u]\, dt, \tag{2.1}$$

where the system equation is

$$\dot{x} = f_0(t, x) + f_u(t, x)u \tag{2.2}$$

subject to the constraints

$$|u(t)| \leqq K(t), \qquad t \in [t_0, t_f], \tag{2.3}$$

$$\{t_0, x(t_0), t_f, x(t_f)\} \in S. \tag{2.4}$$

* Received by the editors May 26, 1970, and in revised form August 26, 1970. This research was supported by the National Science Foundation under Grant GK-4990 and by the National Aeronautics and Space Administration under Grant NGR-23-005-329.

† Computer, Information, and Control Engineering Program, University of Michigan, Ann Arbor, Michigan 48104.

‡ Department of Aerospace Engineering, University of Michigan, Ann Arbor, Michigan 48104.

Here x is an n-vector and S is a closed subset of R^{2n+2}. The functions f_0, f_u, L_0, L_u are assumed to be analytic in both arguments in a suitable domain; $K(t)$ is assumed to be analytic in a neighborhood of each junction and $|u(t)| < K(t)$ almost everywhere on the singular subarcs. Of course, the usual case $|u| \leqq K$ with $K = \text{const.}$ is included as a special case. We restrict attention to a scalar control in order to simplify notation. A similar analysis holds for each component of a vector control.

Clearly, the Hamiltonian for this problem is linear in the control, i.e.,

$$H(t, x, \lambda, u) = \lambda^T f_0(t, x) + L_0(t, x) + [\lambda^T f_u(t, x) + L_u(t, x)]u. \tag{2.5}$$

The multiplier equations are then given by

$$\dot{\lambda} = -H_x(t, x, \lambda, u), \tag{2.6}$$

where H_x is also linear in u. The coefficient of u in (2.5) is called the *switching function*, which we shall designate as $\phi(t)$, i.e.,

$$\phi(t) \equiv H_u(t, x(t), \lambda(t)). \tag{2.7}$$

The minimum principle (i.e., Pontryagin's maximum principle in a minimum form) states that for almost every $t \in [t_0, t_f]$ and each u satisfying $|u| \leqq K(t)$, the optimal control $u^*(t)$ must satisfy

$$H(t, x(t), \lambda(t), u^*(t)) \leqq H(t, x(t), \lambda(t), u). \tag{2.8}$$

Therefore, as is well known, on each open subinterval of $[t_0, t_f]$ there are two distinct possibilities for u^*. Either

$$u^*(t) = -K(t) \operatorname{sgn} \phi(t) \tag{2.9}$$

or

$$\phi(t) \equiv 0. \tag{2.10}$$

Equations (2.9) and (2.10) define, respectively, the nonsingular and singular subarcs of the optimal control.

The class of problems defined above will be called *singular control problems*, even though only a portion of the total trajectory may be singular.

3. Notation and definitions. The following definitions will clarify the terminology used in this paper.

DEFINITION 1. A real-valued function g is said to be *piecewise analytic* on an interval (a, b) if for each $t_c \in (a, b)$ there exist $t_1 \in (a, t_c)$ and $t_2 \in (t_c, b)$ such that g is analytic on the open subintervals (t_1, t_c) and (t_c, t_2).

DEFINITION 2. A junction between singular and nonsingular subarcs of the control is said to be a *nonanalytic junction* if the control is not piecewise analytic in any neighborhood of the junction.

DEFINITION 3. Let u be an optimal singular control on the interval $[t_1, t_2]$, and let $(d^{2q}/dt^{2q})[H_u(t, x, \lambda)]$ be the lowest order total derivative of H_u in which u appears explicitly with a coefficient which is not identically zero on $[t_1, t_2]$. Then the integer q is called the *order of the singular arc.*

Implicit in Definition 3 is the property that u first appears explicitly in an even order derivative of H_u; i.e., it is correct to refer to q as an integer. For a proof of this property see Robbins [5].

We also need the well-known generalized Legendre-Clebsch necessary condition for optimality of singular subarcs [3].

THEOREM (Generalized Legendre-Clebsch condition). *On an optimal singular subarc of order q, it is necessary that*

$$(-1)^q \frac{\partial}{\partial u}\left[\frac{d^{2q}}{dt^{2q}} H_u\right] \geqq 0. \tag{3.1}$$

Condition (3.1) hereafter will be called the *GLC condition.* By the strengthened GLC condition we mean that strict inequality holds in (3.1).

In this paper it will be convenient to consider the lowest order derivative of a function to be its zeroth derivative, by which we mean the function itself. We shall use the notation

$$g^{(0)} \equiv g, \qquad g^{(i)} \equiv d^i g/dt^i, \qquad i = 1, 2, \cdots.$$

Also, where the context makes the meaning clear, we shall use u to designate the optimal control instead of u^*.

4. The junction theorems. As indicated in § 1, the theory for totally singular and totally nonsingular optimal controls is rather well developed. The main difficulty with singular control problems occurs when both singular and nonsingular subarcs are present. Since a useful sufficient condition for such problems is not available, one is led naturally to the study of necessary conditions which are valid in the neighborhood of a junction between singular and nonsingular subarcs. It is expected that such conditions can be used to eliminate candidate extremals and/or predict beforehand the way in which singular and nonsingular subarcs must be joined, e.g., whether the optimal control is continuous or discontinuous at a junction.

If the optimal control is well-behaved in a neighborhood of a junction, then the following property must hold.

THEOREM 1. *Let t_c be a point at which singular and nonsingular subarcs of an optimal control u are joined, and let q be the order of the singular arc. Suppose the strengthened GLC condition is satisfied at t_c, i.e., $(-1)^q(\partial/\partial u)H_u^{(2q)} > 0$, and assume that the control is piecewise analytic in a neighborhood of t_c. Let $u^{(r)}$ $(r \geqq 0)$ be the lowest order derivative of u which is discontinuous at t_c. Then $q + r$ is an odd integer.*

Proof. Since $H_u^{(2q)}$ is the lowest order time derivative of H_u which contains u explicitly, from (2.2), (2.5) and (2.6) we see that it must have the form

$$H_u^{(2q)}(t, x, \lambda, u) \equiv A(t, x, \lambda) + B(t, x, \lambda)u. \tag{4.1}$$

Define the functions α and β as follows:

$$\alpha(t) \equiv A(t, x(t), \lambda(t)), \tag{4.2}$$

$$\beta(t) \equiv B(t, x(t), \lambda(t)). \tag{4.3}$$

From the hypotheses it is clear that α and β are continuous and have at least r continuous derivatives at t_c. The switching function ϕ as defined by (2.7) has exactly $2q + r - 1$ continuous derivatives at t_c.

Let ε be a nonzero real number of arbitrarily small magnitude such that $t_c + \varepsilon$ is a point on the nonsingular side of t_c and $t_c - \varepsilon$ is a point on the singular

side. Let u_n and u_s designate the control u on the nonsingular and singular sides of t_c, respectively. By $u_n^{(i)}(t_c)$ and $u_s^{(i)}(t_c)$ we mean the limit as $\varepsilon \to 0$ of $u^{(i)}(t_c + \varepsilon)$ and $u^{(i)}(t_c - \varepsilon)$, respectively.

We wish to expand $\phi(t_c + \varepsilon)$ in a Taylor series about t_c. Let $k = 2q + r$. Then $\phi^{(k)}$ will be the lowest order derivative of the switching function ϕ which is discontinuous at t_c, and since $\phi \equiv 0$ on the singular side of t_c, the first nonzero term of the Taylor series will be the term containing $\phi^{(k)}$. Noting that

$$\phi^{(k)} \equiv \frac{d^r}{dt^r}[\alpha + \beta u] \tag{4.4}$$

we can write

$$\phi(t_c + \varepsilon) = \frac{\varepsilon^k}{k!}\left[\alpha^{(r)}(t_c) + \sum_{i=0}^{r}\binom{r}{i}\beta^{(r-i)}(t_c)u_n^{(i)}(t_c)\right] + o(\varepsilon^k), \tag{4.5}$$

where Leibniz's formula [6, p. 184] for differentiation of a product has been used to differentiate βu_n.

On the singular arc,

$$\phi^{(2q)} = \alpha + \beta u_s \equiv 0. \tag{4.6}$$

Therefore, $\alpha \equiv -\beta u_s$, and

$$\alpha^{(r)} = \frac{d^r}{dt^r}[-\beta u_s] = -\sum_{i=0}^{r}\binom{r}{i}\beta^{(r-i)}u_s^{(i)}. \tag{4.7}$$

Substituting from (4.7) into (4.5), we have

$$\phi(t_c + \varepsilon) = \frac{\varepsilon^k}{k!}\sum_{i=0}^{r}\binom{r}{i}\beta^{(r-i)}(t_c)[u_n^{(i)}(t_c) - u_s^{(i)}(t_c)] + o(\varepsilon^k). \tag{4.8}$$

If $r > 0$,

$$u_n^{(i)}(t_c) = u_s^{(i)}(t_c), \qquad i = 0, \cdots, r-1. \tag{4.9}$$

Therefore, (4.8) becomes

$$\phi(t_c + \varepsilon) = \frac{\varepsilon^k}{k!}\beta(t_c)[u_n^{(r)}(t_c) - u_s^{(r)}(t_c)] + o(\varepsilon^k). \tag{4.10}$$

Let $\sigma = -\operatorname{sgn}\phi(t_c + \varepsilon)$ so that $u_n(t) = \sigma K(t)$. Then recalling that $u_n^{(i)}(t_c) \equiv \lim_{\varepsilon \to 0} u_n^{(i)}(t_c + \varepsilon)$ we have

$$u_n^{(i)}(t_c) \equiv \sigma K^{(i)}(t_c), \qquad i = 0, \cdots, r. \tag{4.11}$$

Now consider the following series expansion on the singular arc:

$$\sigma K(t_c - \varepsilon) - u(t_c - \varepsilon) = \sum_{i=0}^{r}\frac{(-\varepsilon)^i}{i!}[\sigma K^{(i)}(t_c) - u_s^{(i)}(t_c)] + o(\varepsilon^r). \tag{4.12}$$

The right-hand side of (4.12) can be simplified using (4.9) and (4.11) to obtain

$$\sigma K(t_c - \varepsilon) - u(t_c - \varepsilon) = \frac{(-1)^r\varepsilon^r}{r!}[u_n^{(r)}(t_c) - u_s^{(r)}(t_c)] + o(\varepsilon^r). \tag{4.13}$$

Substituting from (4.13) into (4.10) and recalling that $k = 2q + r$, we obtain

$$\phi(t_c + \varepsilon) = (-1)^r \frac{\varepsilon^{2q} r!}{k!} \beta(t_c)[\sigma K(t_c - \varepsilon) - u(t_c - \varepsilon)] + o(\varepsilon^k). \tag{4.14}$$

From the application of the minimum principle on the nonsingular subarc (see (2.9)) we have $\sigma = 1$ if $\phi(t_c + \varepsilon) < 0$ and $\sigma = -1$ if $\phi(t_c + \varepsilon) > 0$. Therefore, the following inequality must hold:

$$(-1)^r \varepsilon^{2q} \beta(t_c)[K(t_c - \varepsilon) \pm u(t_c - \varepsilon)] < 0. \tag{4.15}$$

From the GLC condition we have

$$(-1)^q \beta(t_c) > 0. \tag{4.16}$$

Multiplying the left-hand side of (4.15) by the positive quantity in (4.16) we obtain

$$(-1)^{q+r} \varepsilon^{2q} \beta^2(t_c)[K(t_c - \varepsilon) \pm u(t_c - \varepsilon)] < 0. \tag{4.17}$$

Since $|u(t)| \leqq K(t)$ for all $t \in [t_0, t_f]$, and the singular arc is assumed to be interior almost everywhere, the bracketed quantity in (4.17) is strictly positive, regardless of the choice of sign on $\pm u(t_c - \varepsilon)$. Also $\varepsilon^{2q} > 0$ regardless of the sign of ε. Therefore, condition (4.17) reduces to

$$(-1)^{q+r} < 0 \tag{4.18}$$

from which it is clear that $q + r$ is an odd integer. This completes the proof.

Theorem 1 implies the following important corollaries.

COROLLARY 1. *In q-even problems, assuming u is piecewise analytic and the strengthened GLC condition is satisfied, the optimal control is continuous at each junction.*

COROLLARY 2. *In q-odd problems, assuming u is piecewise analytic and the strengthened GLC condition is satisfied, the optimal control either has a jump discontinuity at each junction, or else the singular control joins the boundary smoothly, i.e., with continuous first derivative.*

In the corollaries above, especially Corollary 1, which applies to the q-even case, the assumption that u is piecewise analytic is not to be taken lightly. In fact, the authors have not seen or been able to produce a q-even example with a continuous junction, i.e., the junction is usually nonanalytic if q is even.

The conclusions reached by Kelley, Kopp and Moyer [3] are consistent with those stated in Corollaries 1 and 2, with one important exception—they ruled out the possibility of a continuous junction for q-odd problems. This erroneous conclusion resulted from the claim that continuity of u implies $(\partial/\partial u)H_u^{(2q)} > 0$ which is not true in general, as can be seen from (4.15) (in which $\beta \equiv (\partial/\partial u)H_u^{(2q)}$). That such a junction is realizable will be demonstrated by means of a simple example in a later section.

Theorem 1 requires that the *strengthened* GLC condition be satisfied at the junction point. While this is the usual and most important case, the possibility exists that the GLC condition is satisfied with equality. To treat this case, note from Definition 3 that for a qth order singular arc the GLC expression $(\partial/\partial u)H_u^{(2q)}$ (i.e., β) cannot be identically zero on the singular subarc. Therefore, in view of our

analyticity assumptions, a derivative of some order must be nonzero at the junction point t_c even if $\beta(t_c) = 0$. This leads to the following theorem, which is a generalization of Theorem 1, but is stated separately to avoid obscuring the result for the important case covered by Theorem 1.

THEOREM 2. *Let t_c be a point at which singular and nonsingular subarcs of an optimal control u are joined, and let q be the order of the singular arc. Assume that the control is piecewise analytic in a neighborhood of t_c. Let $u^{(r)}$ $(r \geqq 0)$ be the lowest order derivative of u which is discontinuous at t_c, and let $\beta^{(m)}$ $(m \geqq 0)$ be the lowest order derivative of the GLC expression $(\partial/\partial u)H_u^{(2q)} \equiv \beta$ which is nonzero at t_c. Then,* (i) *if $m \leqq r$, $q + r + m$ is an odd integer;* (ii) *if $m > r$, $-\operatorname{sgn}[\beta^{(m)}(t_c^+)\beta^{(m)}(t_c^-)] = (-1)^{q+r+m}$.*

Proof outline. The proof is similar to that for Theorem 1; however, in order to obtain a nontrivial term in the Taylor series expansion for $\phi(t_c + \varepsilon)$, one must consider higher order terms with the result that (4.15) is replaced by

$$(-1)^r \varepsilon^{2q+m} \beta_n^{(m)}(t_c)[K(t_c - \varepsilon) \pm u_s(t_c - \varepsilon)] < 0. \tag{4.19}$$

A Taylor series expansion for $\beta(t)$ on the singular arc yields

$$\beta(t_c - \varepsilon) = \frac{(-\varepsilon)^m}{m!}\beta_s^{(m)}(t_c) + o(\varepsilon^m), \tag{4.20}$$

where the subscripts s and n on $\beta^{(m)}(t_c)$ indicate the limit at t_c on the singular and nonsingular sides, respectively. Since $\beta_s^{(m)}(t_c) \neq 0$, from the GLC condition and (4.20) we have

$$(-1)^{q+m} \varepsilon^m \beta_s^{(m)}(t_c) > 0. \tag{4.21}$$

From (4.19) and (4.21) it follows that

$$(-1)^{q+r+m} \beta^{(m)}(t_c^+)\beta^{(m)}(t_c^-) < 0. \tag{4.22}$$

If $m \leqq r$, $\beta^{(m)}$ is continuous at t_c, in which case (4.22) implies that $q + r + m$ is an odd integer. If $m > r$, $\beta^{(m)}$ may not be continuous at t_c, and the conclusion of Theorem 2 for this case follows.

The main restriction in Theorems 1 and 2 is the assumption that the control is piecewise analytic in a neighborhood of the junction. This hypothesis is usually satisfied on the singular subarc, but not always on the nonsingular subarc. Thus, we are led to consider properties which do not require the assumption of analyticity, as stated in the following theorem. The functions A and B in this theorem are those defined by the identity (4.1).

THEOREM 3. *Let u be an optimal control which contains both nonsingular subarcs and piecewise continuous, q-th order singular subarcs.*

(i) *If $H_u^{(2q)} \neq 0$ on the nonsingular side of a junction, then the control is discontinuous.*

(ii) *If $A = 0$, $B \neq 0$, and $K \neq 0$ at a junction, then the control is discontinuous.*

(iii) *If u is piecewise continuous on the nonsingular subarc, $H_u^{(2q)} = 0$ on the nonsingular side of a junction, and $B \neq 0$ at the junction, then the control is continuous.*

Proof. Using the same notation as in the proof of Theorem 4.1, and recalling that $H_u^{(2q)} \equiv 0$ on the singular subarc, we have for case (i),

$$\alpha(t_c) \pm \beta(t_c)K(t_c) \neq 0 = \alpha(t_c) + \beta(t_c)u_s(t_c) \tag{4.23}$$

from which we obtain $|u_s(t_c)| \neq K(t_c)$. Therefore, u is discontinuous.

For case (ii), $\alpha(t_c) = 0$ and $\beta(t_c) \neq 0$ imply $u_s(t_c) = 0$, and since $K(t_c) \neq 0$, the control is discontinuous.

For case (iii) we have

$$\alpha(t_c) + \beta(t_c)u_n(t_c) = 0 = \alpha(t_c) + \beta(t_c)u_s(t_c). \tag{4.24}$$

Since $\beta(t_c) \neq 0$, we must have $u_n(t_c) = u_s(t_c)$.

Case (ii) of Theorem 4.3 may appear to be a rather special case, but it occurs frequently enough to be of interest. Note that this result is independent of an even or odd assumption. Because of this, we can couple (ii) with the previous result for q-even problems to obtain the following interesting property.

COROLLARY 3. *If q is even, $A(t, x, \lambda) \equiv 0$, $K(t_c) \neq 0$, and $B(t_c, x(t_c), \lambda(t_c)) \neq 0$, where t_c is a junction point between optimal singular and nonsingular subarcs, then the junction is nonanalytic.*

Proof. Assume the contrary, i.e., that the optimal control is piecewise analytic in a neighborhood of t_c. Then by Corollary 1 the control is continuous at t_c, but by Theorem 3 (ii) the control is discontinuous, which supplies the necessary contradiction.

In the next section this corollary will be used to predict the nonanalytic junction in the well-known Fuller problem.

5. Example of a nonanalytic junction. Consider the Fuller problem [7], which is to minimize

$$J = \frac{1}{2}\int_0^T x_1^2\, dt \tag{5.1}$$

subject to

$$\begin{aligned} \dot{x}_1 &= x_2, & x_1(0) &= \xi_1 \neq 0, \\ \dot{x}_2 &= u, & x_2(0) &= \xi_2, \end{aligned} \tag{5.2}$$

$$|u| \leqq 1, \tag{5.3}$$

where T is fixed. The Hamiltonian, the multiplier equations and the switching function are given by

$$H = \lambda_1 x_2 + \lambda_2 u + \tfrac{1}{2}x_1^2, \tag{5.4}$$

$$\begin{aligned} \dot{\lambda}_1 &= -x_1, & \lambda_1(T) &= 0, \\ \dot{\lambda}_2 &= -\lambda_1, & \lambda_2(T) &= 0, \end{aligned} \tag{5.5}$$

$$\phi \equiv H_u = \lambda_2. \tag{5.6}$$

The lowest order derivative of H_u which contains u explicitly is

$$H_u^{(4)} = u \tag{5.7}$$

from which we see that the order of the singular arc is even, namely $q = 2$. Also, the strengthened GLC condition holds, and $A(t, x, \lambda) \equiv 0$. Thus, we have precisely the conditions of Corollary 3, indicating that any junctions which occur must be nonanalytic junctions.

This problem has been studied thoroughly by Fuller [7] and Johansen [8], and the result is well known. The singular arc is given by

$$u_s = x_1 = x_2 = 0. \tag{5.8}$$

Since $\xi_1 \neq 0$, the initial control must be nonsingular. The nonsingular arc is characterized by the nonlinear differential equation

$$\phi^{(4)} = -\operatorname{sgn} \phi. \tag{5.9}$$

The solution of (5.9) yields a switching function with an infinite number of zeros such that the ratio of the lengths of successive intervals between zeros is a constant. If T is sufficiently large, the resulting nonsingular (bang-bang) control drives the state to the origin in a finite time t_c, with an infinite number of switches occurring in a neighborhood of t_c, at which point the optimal control becomes singular. The control is clearly discontinuous at the junction point t_c, as it must be according to (iii) of Theorem 3. Even though q is even, Corollary 1 is not violated because the control is not piecewise analytic in a neighborhood of the junction.

The predicted behavior at the junction is useful knowledge for numerical computational schemes; e.g., Jacobson [9] was able to successfully compute bang-bang solutions for this problem with T sufficiently small so that the singular arc did not occur. However, for large T the nonanalytic junction came into play. After computing about ten switches, the method became unstable [10].

Note that the optimal control for this "innocent looking," second order example is measurable but not piecewise continuous. Aside from its physical applicability, the existence of such examples is useful for motivating the assumption of measurable controls in the proof of the minimum principle.

6. Example of a smooth junction with q odd. This example demonstrates not only the realizability of a smooth junction with q odd, but also the dependence of junction phenomena upon boundary conditions. For this case we consider the performance index

$$J = \frac{1}{2}\int_0^T (x_2^2 - x_1^2)\, dt, \tag{6.1}$$

where $T = 2.985$. The equations of motion and constraints are given by

$$\begin{aligned} \dot{x}_1 &= x_2, \quad x_1(0) = 0, \quad x_1(T) = \sigma_1, \\ \dot{x}_2 &= u, \quad x_2(0) = 1, \quad x_2(T) = \sigma_2, \end{aligned} \tag{6.2.}$$

$$|u| \leqq 1. \tag{6.3}$$

The Hamiltonian and the multiplier equations are given by

$$H = \lambda_1 x_2 + \lambda_2 u + \tfrac{1}{2}x_2^2 - \tfrac{1}{2}x_1^2, \tag{6.4}$$

$$\dot{\lambda}_1 = x_1, \qquad \dot{\lambda}_2 = -\lambda_1 - x_2. \tag{6.5}$$

The switching function and its second derivative are

$$\phi \equiv H_u = \lambda_2, \tag{6.6}$$

$$\ddot{H}_u = -x_1 - u \tag{6.7}$$

so the singular arc is first order; i.e., q is odd. From (6.7) we see that the strengthened GLC condition is satisfied. Setting the right-hand side of (6.7) equal to zero and substituting in the equations of motion (6.2), we can readily verify that a singular arc emanating from the initial state (0, 1) is given by

$$u_s = -\sin t, \quad x_1 = \sin t, \quad x_2 = \cos t. \tag{6.8}$$

If the terminal state (σ_1, σ_2) is chosen to lie on the trajectory (6.8), the solution is totally singular, as can be shown by the sufficient conditions in [1] and [2]. However, in this paper we are concerned with junctions. Consider the case where $\sigma_1 = 0$, $\sigma_2 = -\sqrt{2}$. For this case we propose as a candidate for the optimal control

$$u = \begin{cases} -\sin t, & t \in [0, \pi/2), \\ -1 \quad , & t \in [\pi/2, T]. \end{cases} \tag{6.9}$$

This control is admissible and satisfies all the necessary conditions for optimality, including that of Theorem 1. There is a junction at $t_c = \pi/2$. The control and its first derivative are continuous at t_c, but the second derivative is discontinuous, so we have $r = 2$, $q = 1$, and $q + r$ is an odd integer.

The authors are unaware of any workable sufficient conditions in the literature which are applicable to this particular type of problem, i.e., nonconvex and containing both singular and nonsingular subarcs. Consequently, we employed a gradient-type numerical method to justify that the candidate control is indeed optimal, within the bounds of a numerical justification. The modified conjugate gradient method of Pagurek and Woodside [11] was used with penalty functions to enforce the terminal constraints. The result is shown in Fig. 1. The control is continuous and smooth at the junction as expected.

It is apparent that the fortuitous occurrence of this smooth junction is a direct result of our judicious choice of the terminal boundary conditions. In fact, to generate this phenomenon, the form of the candidate control was first selected on the basis of intuition; then a convenient point on the resulting trajectory was selected as the fixed terminal state, and finally the corresponding time was taken to be the explicit final time.

By changing the terminal state, we were able to generate discontinuous controls, which undoubtedly are the usual case. These are shown in Figs. 2 and 3. For these cases $r = 0$, and the condition of Theorem 1 is satisfied again. To further emphasize the special character of the smooth junction, the phase plane trajectories for the controls in Figs. 1–2 are given in Fig. 4.

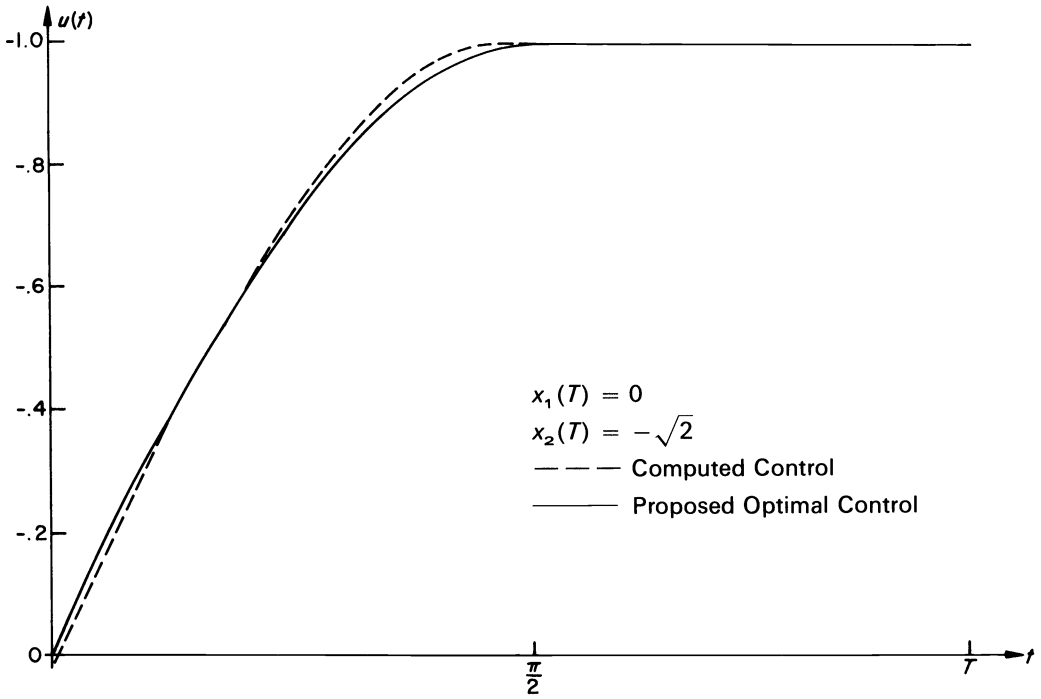


FIG. 1. *Example of an optimal control which is continuous at a singular-to-nonsingular junction with q odd*

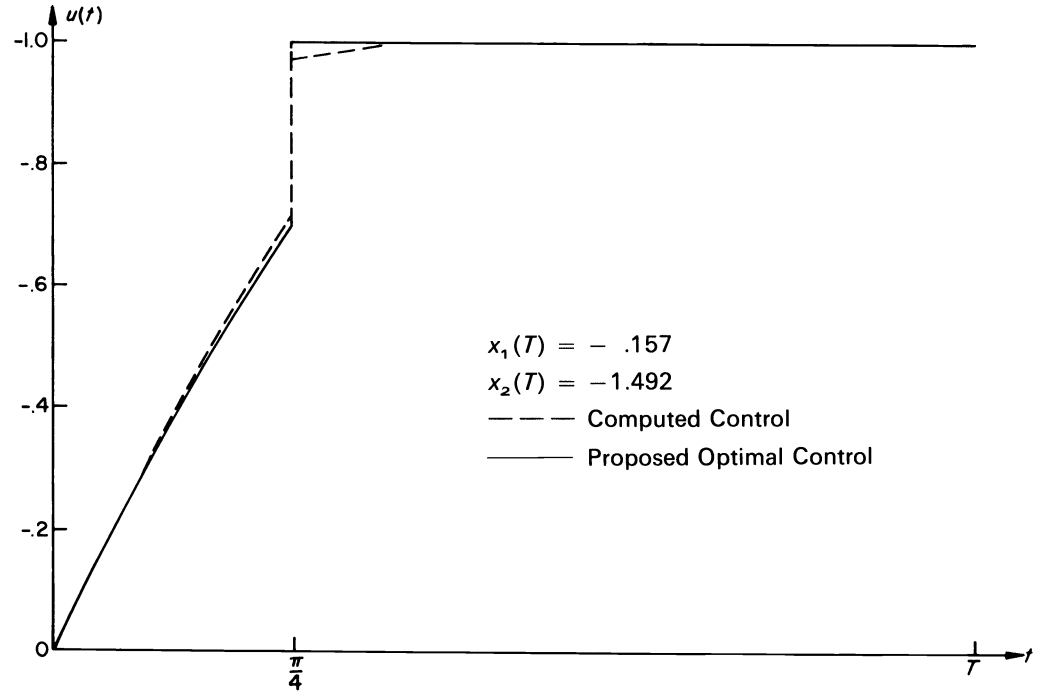


FIG. 2. *A typical discontinuous junction*

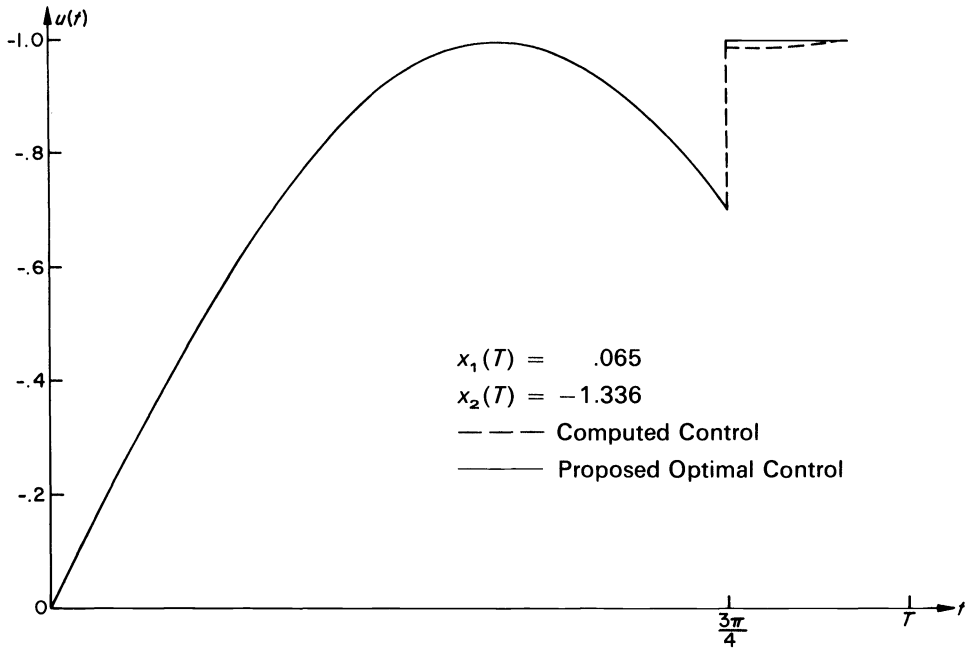


FIG. 3. *A predominantly singular control with discontinuous junction to a short nonsingular subarc*

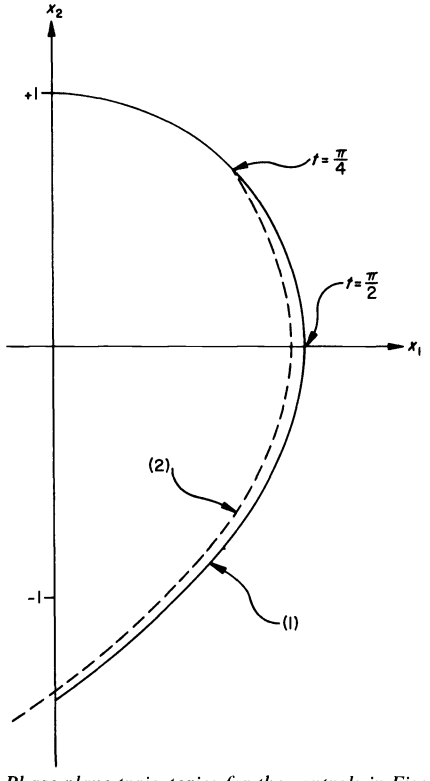


FIG. 4. *Phase plane trajectories for the controls in Figs.* 1 *and* 2

7. Conclusion. Necessary conditions for the optimality of junctions between singular and nonsingular subarcs in singular optimal control problems have been developed. Necessary conditions developed previously by Kelley, Kopp and Moyer [3], which involve an analyticity assumption, have been clarified and extended. The main result in this direction is that the sum of the order of the singular arc and the order of the lowest time derivative of the control which is discontinuous at the junction must be an odd integer when the strengthened generalized Legendre-Clebsch condition is satisfied. Also, new necessary conditions which do not involve an analyticity assumption have been developed. These conditions aid mainly in characterizing problems which may possess nonanalytic junctions.

It should be emphasized that these are *local* necessary conditions for optimality. Yet, as indicated by the example in § 6, the point at which a junction occurs is determined mainly by initial and terminal boundary conditions, i.e., by essentially nonlocal information. This means that any junction theory which, for example, might be used to establish criteria for switching between singular and nonsingular arcs in an indirect computational scheme will have to take such nonlocal information into account.

It is becoming increasingly apparent that a close relationship exists between singular problems and bounded state problems [12], [13]. In this regard it is interesting to note that the result of Theorem 1 bears some similarity to a result of Jacobson, Lele and Speyer [13] which identifies certain properties of optimal trajectories associated with odd order state space constraints. Such similarities suggest the possibility of a duality between these two classes of problems which might be profitably exploited.

REFERENCES

[1] J. P. McDanell and W. F. Powers, *New Jacobi-type necessary and sufficient conditions for singular optimization problems*, AIAA J., 8 (1970), pp. 1416–1420.

[2] J. L. Speyer and D. H. Jacobson, *Necessary and sufficient conditions for optimality for singular control problems: A transformation approach*, Rep. 69-24, Analytical Mechanics Associates, Inc., Cambridge, Mass., 1969; J. Math. Anal. Appl., to appear.

[3] H. J. Kelley, R. E. Kopp and H. G. Moyer, *Singular extremals*, Topics in Optimization, G. Leitmann, ed., Academic Press, New York, 1967.

[4] C. D. Johnson, *Singular Solutions in Problems of Optimal Control*, Advances in Control Systems, vol. 2, C. T. Leondes, ed., Academic Press, New York, 1965.

[5] H. M. Robbins, *A generalized Legendre-Clebsch condition for the singular cases of optimal control*, IBM J. Res. Develop., 11 (1967), pp. 361–372.

[6] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[7] A. T. Fuller, *Study of an optimum nonlinear control system*, J. Electronics Control, 15 (1963), pp. 63–71.

[8] D. E. Johansen, *Solution of a linear mean square estimation problem when process statistics are undefined*, Joint Automatic Control Conference, Troy, New York, 1965.

[9] D. H. Jacobson, *Differential dynamic programming methods for solving bang-bang control problems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 661–675.

[10] ———, personal communication.

[11] B. Pagurek and C. M. Woodside, *The conjugate gradient method for optimal control problems with bounded control variables*, Automatica, 4 (1968), pp. 337–349.

[12] D. H. JACOBSON AND M. M. LELE, *A transformation technique for optimal control problems with a state variable inequality constraint*, Tech. Rep. 574, Harvard Univ., Cambridge, Mass., 1968.

[13] D. H. JACOBSON, M. M. LELE AND J. L. SPEYER, *New necessary conditions of optimality for control problems with state variable inequality constraints*, Tech. Rep. 597, Harvard Univ., Cambridge, Mass., 1969.

SIAM J. CONTROL
Vol. 9, No. 2, May 1971

DISCRETE SPLINES VIA MATHEMATICAL PROGRAMMING*

O. L. MANGASARIAN† AND L. L. SCHUMAKER‡

Abstract. Existence, uniqueness and characterizing properties are given for a class of constrained minimization problems in real Euclidean space. These problems are the discrete analogues of minimization problems in Banach space whose solutions are generalized splines. Solutions of these discrete problems, which are called discrete splines, can be obtained by algorithms of mathematical programming.

1. Introduction. The purpose of this paper is to investigate some constrained minimization problems in real Euclidean space which are suggested by certain minimization problems in Banach space whose solutions are generalized splines (see [5], [9]). Because of the analogy with classical splines, and because the solutions of the discrete problems exhibit a spline-like structure, we call them discrete splines. Our aim here is to obtain existence, uniqueness and characterizing properties of the discrete splines. We include two detailed examples and a section with remarks indicating application to computing continuously constrained spline interpolants as well as some problems for further study.

2. Definition of discrete splines. Let L be a general forward difference operator of order m of the form

$$(Ly)_j = \sum_{v=0}^{m} a_v y_{v+j}, \qquad a_m \neq 0, \qquad j = 0, 1, \cdots, N - m. \tag{2.1}$$

L maps a point $y = (y_0, y_1, \cdots, y_N)$ in R^{N+1} into a point in R^{N-m+1}. Denote by N_0 the null space of L.

For $i = 1, 2, \cdots, k$, let $\mathcal{M}_i$ be forward difference operators of order m_i mapping points of R^{N+1} into R^{N-m_i+1}, where $0 \leqq m_i \leqq m$ and

$$(\mathcal{M}_i y)_j = \sum_{v=0}^{m_i} a_{iv} y_{v+j}, \qquad a_{im_i} \neq 0, \qquad j = 0, 1, \cdots, N - m_i. \tag{2.2}$$

Let α_i, β_i, $i = 1, 2, \cdots, k$, be points in R^{N-m_i+1} such that $\alpha_i \leqq \beta_i$; that is, each component of α_i is less than or equal to the corresponding component of β_i:

$$\alpha_{ij} \leqq \beta_{ij}, \qquad j = 0, 1, \cdots, N - m_i.$$

Define

$$U = \{y \in R^{N+1} : \alpha_i \leqq \mathcal{M}_i y \leqq \beta_i, \ i = 1, 2, \cdots, k\}. \tag{2.3}$$

* Received by the editors March 19, 1970, and in revised form August 7, 1970.

† Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706. The work of this author was supported by the National Science Foundation under Grant GJ-362 and the United States Army under Contract DA-31-124-ARO-D-462.

‡ Department of Mathematics, University of Texas, Austin, Texas 78712. The work of this author was supported by USAFOSR-69-1812.

Finally, let $1 \leqq p \leqq \infty$ and let $g \in R^{N-m+1}$ be prescribed. Throughout this paper we shall be concerned with the discrete minimization problems

$$\begin{aligned} &\underset{y \in U}{\text{minimize}} \, (\|Ly - g\|_p)^p, \qquad 1 \leqq p < \infty, \\ &\underset{y \in U}{\text{minimize}} \, \|Ly - g\|_\infty, \qquad p = \infty, \end{aligned} \tag{2.4}$$

where

$$\|x\|_p = \left(\sum_{j=0}^{N-m} |x_j|^p \right)^{1/p}, \qquad 1 \leqq p < \infty,$$

and

$$\|x\|_\infty = \max_{0 \leqq j \leqq N-m} |x_j|.$$

We shall call the solutions of (2.4) *discrete splines*. The reason for this terminology will become more apparent in later sections.

3. Existence and uniqueness of discrete splines.

THEOREM 3.1.

(i) *A solution of* (2.4) *exists if* U *is nonempty.*

(ii) *For* $1 < p < \infty$, *any two solutions of* (2.4) *differ by an element of* N_0.

(iii) *If* $1 < p < \infty$ *and* $s \in U$ *is a solution of* (2.4), *then it is unique if and only if* $N_0 \cap U(s) = \{0\}$, *where* $U(s) = \{f - s : f \in U\}$.

Proof of Theorem 3.1(i). This theorem is a consequence of the following more general theorem. (For $p = 1$ or ∞, (2.4) is a bounded linear program and hence has a solution [3, Theorem 2, p. 134].)

THEOREM 3.2. *Let* A *be an* $l \times n$ *matrix,* L *a* $k \times n$ *matrix,* b *an* l*-vector,* g *a* k*-vector,*

$$X = \{x : x \in R^n, Ax \leqq b\},$$

$$S = \{LX - g\} = \{z : z \in R^k, z = Lx - g, x \in X\},$$

and let φ *be a lower semicontinuous function on* S *satisfying some growth condition such as this: There exist a* $\hat{z} \in S$ *and* $\rho \geqq 0$ *such that for any* $z \in S$,

$$\|z\|_\infty > \rho \Rightarrow \varphi(z) > \varphi(\hat{z}). \tag{3.1}$$

Then the minimization problem

$$\min \{\varphi(Lx - g) : x \in R^n, Ax \leqq b\} \tag{3.2}$$

has a solution provided X *is nonempty.*

Proof of Theorem 3.2. Because of the growth condition (3.1) the minimization problem (3.2) is equivalent to

$$\min \{\varphi(z) : z \in S, \|z\|_\infty \leqq \rho\}.$$

Because φ is lower semicontinuous on S, and because $\{z : \|z\|_\infty \leqq \rho\}$ is compact, this problem has a solution provided that S is closed. We show now that S is closed.[1]

[1] A linear map of a closed convex set is *not* in general closed. For example the linear map $f(X)$ of the closed convex set $X = \{x : x \in R^2, x_2 \geqq e^{-x_1}\}$, where $f(x) = x_2$, is the open interval $(0, \infty)$. What we show above is that a linear map of a closed *polyhedral* set is closed.

Let $\bar{z}$ be a point of closure of S. Then

$$\inf\{\varepsilon : \varepsilon \in R,\ -e\varepsilon \leqq z - \bar{z} \leqq e\varepsilon,\ z \in S\} = 0, \tag{3.3}$$

where e is a k-vector of ones. But since (3.3) is a linear program, the infimum 0 is attained at some $z^* \in S$ (see [3], [8]). Hence $z^* - \bar{z} = 0$, $\bar{z} = z^* \in S$, and S is closed. (This proof is similar to that of Laurent [7] except that we use linear programming to establish the fact that S is closed, whereas Laurent uses a theorem of Dieudonné concerning the closure of the algebraic difference of two sets in a Hausdorff topological vector space.)

Theorem 3.1(i) follows from Theorem 3.2 by making the identifications: $x = y$, $z = Lx - g$, $k = N - m + 1$, $n = N + 1$, $l = 2\sum_{i=1}^{k}(N - m_i + 1)$, $\varphi(z) = (\|z\|_p)^p$, $1 \leqq p < \infty$, and

$$A = \begin{bmatrix} \tilde{\mathscr{M}}_1 \\ \vdots \\ \tilde{\mathscr{M}}_k \\ -\tilde{\mathscr{M}}_1 \\ \vdots \\ -\tilde{\mathscr{M}}_k \end{bmatrix}_{l \times N+1}, \qquad b = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \\ -\alpha_1 \\ \vdots \\ -\alpha_k \end{bmatrix}_{l \times 1},$$

$$\tilde{\mathscr{M}}_i = \begin{bmatrix} a_{i0} & \cdots & a_{im_i} & 0 & \cdots & & 0 \\ 0 & a_{i0} & \cdots & a_{im_i} & 0 & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & & & & 0 & a_{i0} & \cdots a_{im_i} \end{bmatrix}_{N-m_i+1 \times N+1},$$

observing that all vector norms are continuous functions [4], and that the growth condition (3.1) is satisfied as follows by $\varphi(z) = (\|z\|_p)^p$, $1 \leqq p < \infty$. Let $\hat{z}$ be any point in S. Then for any $z \in S$,

$$\begin{aligned} \varphi(z) \leqq \varphi(\hat{z}) &\Rightarrow \|z\|_p \leqq \|\hat{z}\|_p \\ &\Rightarrow \alpha_p \|z\|_\infty \leqq \|\hat{z}\|_p \quad \text{for some } \alpha_p > 0 \\ &\qquad\qquad (\text{since } \alpha_p\|z\|_\infty \leqq \|z\|_p,\quad \alpha_p > 0 \text{ (see [4])}) \\ &\Rightarrow \|z\|_\infty \leqq \rho = \|\hat{z}\|_p/\alpha_p. \end{aligned}$$

Proof of Theorem 3.1(ii). Since for $1 < p < \infty$, $(\|z - g\|_p)^p$ is a strictly convex function of z, it follows for any minimum solution y, $z = Ly$ is unique. Hence for any two solutions y^1 and y^2 we have that $Ly^1 = Ly^2$ or $L(y^1 - y^2) = 0$ and hence differ by an element of the null space N_0 of L.

Proof of Theorem 3.1(iii). Let s be a solution of (2.4). If $N_0 \cap U(s) \neq \{0\}$, then we may add any nontrivial element $\hat{s} \in N_0 \cap U(s)$ to s and obtain $s + \hat{s} \in U$ and $L(s + \hat{s}) = Ls + L\hat{s} = Ls$, and hence $s + \hat{s}$ is another distinct solution.

Conversely, if s_1 and s_2 are two solutions of (2.4), then by part (ii) of this theorem $L(s_1 - s_2) = 0$, and hence $s_1 - s_2 \in N_0$. But $s_1 - s_2 \in U(s_2)$. Hence

$s_1 - s_2 \in N_0 \cap U(s_2)$, which implies that $s_1 = s_2$ whenever this set consists of the zero element only.

The existence assertion (i) of Theorem 3.1 could also be established without the use of programming theory by following the lines of the proof of a similar theorem in an appropriate Sobolev space for the continuous analogue of (2.4) (see [9]).

4. Necessary and sufficient characterization of discrete splines.

THEOREM 4.1. *For $1 < p < \infty$ a point $y \in U$ is a solution of* (2.4) *if and only if for $i = 1, 2, \cdots, k$ there exist vectors $\lambda_i \geqq 0$, $\mu_i \geqq 0$ in R^{N-m_i+1} such that*

$$H_1 = \sum_{i=1}^{k} [\lambda_i(-\mathcal{M}_i y + \alpha_i) + \mu_i(\mathcal{M}_i y - \beta_i)] = 0 \tag{4.1}$$

and

$$H = (\|Ly - g\|_p)^p + H_1 \tag{4.2}$$

satisfy

$$\nabla_j H = \partial H / \partial y_j = 0, \qquad j = 0, 1, \cdots, N, \tag{4.3}$$

where

$$\lambda_i(-\mathcal{M}_i y + \alpha_i) = \sum_{j=0}^{N-m_i} \lambda_{ij}[-(\mathcal{M}_i y)_j + \alpha_{ij}].$$

Proof. The conditions (4.1)–(4.3) follow by a direct application of the Kuhn–Tucker optimality theorem of nonlinear programming [6], [8]. That they are necessary follows from the fact that the set U is polyhedral, and hence no constraint qualification is needed. That the conditions are sufficient follows from the convexity of U and the objective function $(\|Ly - g\|_p)^p$.

These conditions completely characterize the discrete splines, although in this generality they do not provide a very good idea of the structure of such splines. In the next section we shall specialize problem (2.4) and thereby define what we will call discrete polynomial splines, whose structure can then be more explicitly delineated.

For convenience we note that for $p = 2$,

$$\nabla_j(\|Ly\|_2)^2 = \nabla_j \sum_{i=0}^{N-m} \left(\sum_{v=0}^{m} a_v y_{v+i} \right)^2$$

$$= \begin{cases} 2 \sum_{l=0}^{j} a_{j-l} \left(\sum_{v=0}^{m} a_v y_{v+l} \right), & j = 0, 1, \cdots, m-1, \quad (4.4) \\ 2 \sum_{l=0}^{m} a_{m-l} \left(\sum_{v=0}^{m} a_v y_{v+l+j-m} \right), & j = m, \cdots, N-m, \quad (4.5) \\ 2 \sum_{l=0}^{N-j} a_{m-l} \left(\sum_{v=0}^{m} a_v y_{v+l+j-m} \right), & j = N-m+1, \cdots, N. \quad (4.6) \end{cases}$$

5. Discrete natural polynomial splines. Given integers N, m, a set of integers $\Lambda = \{0 \leqq i_1 < i_2 < \cdots < i_k \leqq N\}$, and real numbers $\gamma = \{\gamma_i\}_1^k$, we define a *discrete natural polynomial spline of degree* $2m-1$ *with knots* $i_1 < \cdots < i_k$ as a vector $y = (y_0, \cdots, y_N)$ which solves the minimization problem

$$\underset{y \in U_N(\gamma)}{\text{minimize}} \frac{1}{h^{2m-1}} \sum_{i=0}^{N-m} (\Delta^m y_i)^2, \tag{5.1}$$

where

$$U_N(\gamma) = \{y = (y_0, \cdots, y_N) : y_{i_j} = \gamma_j, j = 1, 2, \cdots, k\} \tag{5.2}$$

and

$$\Delta^m y_i = \sum_{v=0}^{m} (-1)^{m-v} \binom{m}{m-v} y_{i+v}$$

is the mth forward difference with $h = 1/N$.

The characterization of discrete natural polynomial splines is given by the following theorem.

THEOREM 5.1. *A vector* $y \in U_N(\gamma)$ *solves* (5.1) *if and only if, for* $j \notin \Lambda$,

$$\Delta^{2m} y_{j-m} = 0, \qquad j = m, \cdots, N-m, \tag{5.3a}$$

$$\Delta^m y_j = 0, \qquad j = 0, 1, \cdots, i_1 - 1, \tag{5.3b}$$

$$\Delta^m y_j = 0, \qquad j = i_k + 1, \cdots, N-m. \tag{5.3c}$$

Proof. Let $\Lambda_1 = \{m, \cdots, N-m\}$, $\Lambda_2 = \{0, \cdots, m-1\}$ and $\Lambda_3 = \{N-m+1, \cdots, N\}$. By setting $a_v = h^{-(2m-1)}(-1)^{m-v}\binom{m}{m-v}$, we can write the following with the help of (4.4)–(4.6):

$$\frac{\partial}{\partial y_j}\left(\frac{1}{h^{2m-1}} \sum_{i=0}^{N-m} (\Delta^m y_i)^2\right)$$

$$= \begin{cases} \dfrac{2}{h^{4m-2}} \displaystyle\sum_{l=0}^{m} (-1)^l \binom{m}{l} \sum_{v=0}^{m} (-1)^{m-v} \binom{m}{m-v} y_{v+l+j-m}, & j \in \Lambda_1, \\ \dfrac{2}{h^{4m-2}} \displaystyle\sum_{l=0}^{j} (-1)^{m-j+l} \binom{m}{m-j+l} \sum_{v=0}^{m} (-1)^{m-v} \binom{m}{m-v} y_{v+l}, & \\ & j \in \Lambda_2, \\ \dfrac{2}{h^{4m-2}} \displaystyle\sum_{l=0}^{N-j} (-1)^l \binom{m}{l} \sum_{v=0}^{m} (-1)^{m-v} \binom{m}{m-v} y_{v+l+j-m}, & j \in \Lambda_3, \end{cases}$$

$$= \begin{cases} \dfrac{2}{h^{4m-2}} \displaystyle\sum_{l=0}^{m} (-1)^l \binom{m}{l} \Delta^m y_{l+j-m} = \dfrac{2}{h^{4m-2}} \Delta^{2m} y_{j-m}, & j \in \Lambda_1, \\ \dfrac{2}{h^{4m-2}} \displaystyle\sum_{l=0}^{j} (-1)^{m-j+l} \binom{m}{j-l} \Delta^m y_l, & j \in \Lambda_2, \\ \dfrac{2}{h^{4m-2}} \displaystyle\sum_{l=0}^{N-j} (-1)^l \binom{m}{l} \Delta^m y_{l+j-m}, & j \in \Lambda_3. \end{cases}$$

Now by Theorem 4.1 we have for $j \notin \Lambda$,

$$\Delta^{2m} y_{j-m} = 0, \qquad j \in \Lambda_1, \tag{5.4a}$$

$$(\varphi y)_j = \sum_{l=0}^{j} (-1)^{m+j+l} \binom{m}{j-l} \Delta^m y_l = 0, \qquad j \in \Lambda_2, \tag{5.4b}$$

$$(\psi y)_j = \sum_{l=0}^{N-j} (-1)^l \binom{m}{l} \Delta^m y_{l+j-m} = 0, \qquad j \in \Lambda_3. \tag{5.4c}$$

By coupling (5.4a) for $j < i_1$ with (5.4b) we obtain (5.3b). Similarly, (5.3c) follows from (5.4c) and (5.4a) for $j > i_k$.

Property (5.3a) asserts that the values y_i lie on polynomials of degree $m-1$ (since $\Delta^m y_i = 0$) in the intervals $[0, i_1 - 1]$ and $[i_k + 1, N]$, while between the knots the y_i lie on polynomials of degree $2m-1$ ($\Delta^{2m} y_i = 0$). It is not difficult to determine how these polynomials tie together at the knots. For example, let j be a knot of y separated from the other knots by at least 1; i.e., if j_1, j_2 are the knots preceding and following j, respectively, then $j_1 < j-1$ and $j+1 < j_2$. Let $P_L(t)$ and $P_R(t)$ be the polynomials of degree $2m-1$ on which the y_i lie to the left and right of j. Since $\Delta^{2m} y_{j-m-1} = 0$, the points $\{y_i\}_{j-m-1}^{j+m-1}$ lie on P_L. Similarly, since $\Delta^{2m} y_{j-m+1} = 0$, the points $\{y_i\}_{j-m+1}^{j+m+1}$ lie on P_R. Thus we conclude that the values $\{y_i\}_{j-m+1}^{j+m-1}$ are common to both P_L and P_R. Consequently, $\Delta^\nu P_L(j-m+1) = \Delta^\nu P_R(j-m+1) = \Delta^\nu y_{j-m+1}$ for $\nu = 0, 1, \cdots, 2m-2$. If these differences divided by h^ν converge as $N \to \infty$, then they approach the corresponding derivatives of P_L and P_R, and it follows that P_L and P_R are tied together of class C^{2m-2} in the limit.

It is interesting to compare the discrete natural polynomial spline with the classical natural polynomial spline. We recall that a natural polynomial spline of degree $2m-1$ with knots $x_1 < x_2 < \cdots < x_k$ is a function $s(x) \in C^{2m-2}$ such that

$$s^{(2m)}(x) = 0 \quad \text{for } x \in (x_i, x_{i+1}), \qquad i = 1, 2, \cdots, k-1, \tag{5.5a}$$

$$s^{(m)}(x) = 0 \quad \text{for } x < x_1, \quad x > x_k. \tag{5.5b}$$

These splines are the solutions of

$$\underset{f \in U(\gamma)}{\text{minimize}} \, \|f^{(m)}\|_2, \tag{5.6}$$

where $U(\gamma) = \{f \in H_2^m : f(x_i) = \gamma_i, \ i = 1, 2, \cdots, k\}$ and $H_2^m = \{f \in C^{(m-1)} : f^{(m)}$ absolutely continuous, $f^{(m)} \in L_2\}$.

It is now clear how the discrete polynomial splines resemble their classical analogues in the continuous version (5.6), while (2.4) was obtained as the dis-

cretization of the minimization problem defining very general splines in [9]. While in general the discrete natural splines are not simply the discretized values of their continuous counterparts, it should not be surprising that they are closely connected, and in § 7 we shall make some remarks about convergence in an appropriate sense.

6. Examples. To further motivate the study of the minimization problem (2.4) and to further illustrate the nature of discrete splines as well as their relationship with the well-known continuous splines, we devote this section to some rather simple examples.

Example 6.1. The sequence $\{s_i = s(ih)\}_{i=-2N}^{i=2N}$, where $h = 1/N$ and

$$s(t) = \begin{cases} 1, & -2 \leqq t < -1, \\ -t, & -1 \leqq t < 0. \\ 2t, & 0 \leqq t < 1, \\ 2, & 1 \leqq t \leqq 2, \end{cases} \tag{6.1}$$

is the unique solution of the problem:

$$\text{minimize } \frac{1}{h} \sum_{-2N}^{2N-1} (\Delta y_i)^2$$

$$\text{subject to } \quad 1 \leqq y_{-N} \leqq \tfrac{3}{2}, \quad -\tfrac{1}{4} \leqq y_0 \leqq 0, \quad 2 \leqq y_N \leqq \tfrac{5}{2},$$

where $\Delta y_i = -y_i + y_{i+1}$. We note that the function $s(t)$ above uniquely minimizes $\int_{-2}^{2} (f')^2\, dt$ subject to $1 \leqq f(-1) \leqq \frac{3}{2}$, $-\frac{1}{4} \leqq f(0) \leqq 0$, and $2 \leqq f(1) \leqq \frac{5}{2}$. Hence the sequence solving the discretized problem is obtained by discretizing the continuous solution of the continuous problem. (As noted above, this does not hold in general.)

Discussion of Example 6.1. By Theorem 4.1, $y = (y_{-2N}, y_{-2N+1}, \cdots, y_{2N})$ satisfying the constraints is a solution of Example 6.1 if and only if

$$\Delta y_{-2N} = 0,$$

$$\Delta^2 y_i = 0, \quad i = -2N, -2N+1, \cdots, 2N-2; \quad i \neq -N-1, -1, N-1,$$

$$(2/h)\Delta^2 y_{i-1} = \mu_i - \lambda_i, \quad i = -N, 0, N,$$

$$\Delta y_{2N-1} = 0, \qquad \lambda_{-N}, \mu_{-N}, \lambda_0, \mu_0, \lambda_N, \mu_N \geqq 0,$$

$$(1 - y_{-N})\lambda_{-N} + (y_{-N} - \tfrac{3}{2})\mu_{-N} = 0, \qquad (-\tfrac{1}{4} - y_0)\lambda_0 + y_0\mu_0 = 0,$$

$$(2 - y_N)\lambda_N + (y_N - \tfrac{5}{2})\mu_N = 0.$$

It is easily checked that the above conditions are satisfied by the sequence $\{s_i\}_{-2N}^{2N}$ defined in Example 6.1. Indeed, since s_i hits the lower, upper, and lower constraints respectively at $-N, 0, N$, we have $\mu_{-N} = \lambda_0 = \mu_N = 0$ and

$$(2/h)\Delta^2 s_{-N-1} = -2 = -\lambda_{-N} \leqq 0,$$

$$(2/h)\Delta^2 s_{-1} = 4 = \mu_0 \geqq 0,$$

$$(2/h)\Delta^2 s_{N-1} = -4 = -\lambda_N \leqq 0.$$

The uniqueness of the solution $\{s_i\}_{-2N}^{2N}$ is established by Theorem 3.1(iii). We have here

$$N_0 = \{y : y_i = \text{const.}, i = -2N, \cdots, 2N\}.$$

We also have that

$$y = (y_{-2N}, \cdots, y_{2N}) \in U(s) = \{f - s : f \in U\}$$

implies that $0 \leqq y_{-N} \leqq \frac{1}{2}$, $-\frac{1}{4} \leqq y_0 \leqq 0, 0 \leqq y_N \leqq \frac{1}{2}$. Thus $N_0 \cap U(s) = \{y : y_i = 0, i = -2N, \cdots, 2N\}$ and hence $\{s_i\}_{-2N}^{2N}$ is unique.

Example 6.2. Let $0 < \gamma \leqq 1$. The sequence

$$s_i = \begin{cases} (i/N) - (i/N)^2, & i = 0, 1, \cdots, \bar{\imath}, \\ \alpha(i - N) + \gamma, & i = \bar{\imath} + 1, \cdots, N, \end{cases}$$

where $\bar{\imath}$ is the integer satisfying

$$1 + \frac{1}{N} - \frac{2(\bar{\imath} + 1)}{N} + \frac{(\bar{\imath} + 1)^2 - (\bar{\imath} + 1)}{N^2} < \gamma \leqq 1 + \frac{1}{N} - \frac{2\bar{\imath}}{N} + \frac{\bar{\imath}^2 - \bar{\imath}}{N^2},$$

and

$$\alpha = \frac{1}{N - \bar{\imath}}\left(\gamma - \frac{\bar{\imath}}{N} + \frac{\bar{\imath}^2}{N^2}\right),$$

is the unique solution of the problem:

$$\text{minimize} \quad \frac{1}{h}\sum_{i=0}^{N-1} (\Delta y_i)^2$$

subject to

$$i/N - (i/N)^2 \leqq y_i \leqq i/N, \qquad i = 0, 1, \cdots, N - 1,$$

$$y_N = \gamma.$$

As $N \to \infty$, the point $\bar{\imath}/N$, which is the last point at which s_i is on the curve $t - t^2$, converges to the value $1 - \sqrt{\gamma}$; and the s_i become the discretized values of the function

$$s(t) = \begin{cases} t - t^2, & 0 \leqq t < 1 - \sqrt{\gamma}, \\ (2\sqrt{\gamma} - 1)t + (1 + \gamma - 2\sqrt{\gamma}), & 1 - \sqrt{\gamma} \leqq t \leqq 1. \end{cases}$$

The function $s(t)$ above is known (see [9, Example 2.2]) to be the unique solution of the constrained variational problem:

minimize

$$\int_0^1 (f')^2 \, dt \quad \text{subject to} \quad t - t^2 \leqq f(t) \leqq t \quad \text{and} \quad f(1) = \gamma. \tag{6.2}$$

Discussion of Example 6.2. By Theorem 4.1, a sequence $\{s_i\}_0^N$ satisfying the constraints is a solution of Example 6.2 if and only if

$$(2/h)\Delta s_0 = \mu_0 - \lambda_0,$$

$$(2/h)\Delta^2 s_{i-1} = \mu_i - \lambda_i, \qquad i = 1, \cdots, N-1,$$

$$(2/h)\Delta s_{N-1} = -\mu_N + \lambda_N,$$

$$\left(\frac{i}{N} - \left(\frac{i}{N}\right)^2 - s_i\right)\lambda_i + \left(s_i - \frac{i}{N}\right)\mu_i = 0, \qquad i = 0, \cdots, N-1,$$

$$(-s_N + \gamma)\lambda_N + (s_N - \gamma)\mu_N = 0.$$

It is easily checked that these conditions are satisfied by the sequence $\{s_i\}_0^N$ given in Example 6.2. Since for $i = 1, \cdots, \bar{\imath}$, $s_i = i/N - (i/N)^2$ (that is, s_i is on the lower constraint), we have $\mu_i = 0$, $i = 1, \cdots, \bar{\imath}$. For $i = \bar{\imath}+1, \cdots, N-1$, s_i is not on any constraint and so $\lambda_i = \mu_i = 0$ for these i. We have then

$$\frac{2}{h}\Delta^2 s_{i-1} = \frac{2}{h}\left[\frac{N^2\gamma - \bar{\imath}^2 - N^2 - N + \bar{\imath} + 2N\bar{\imath}}{N^2(N-\bar{\imath})}\right] \leqq 0,$$

$$\frac{2}{h}\Delta s_0 = 2(1-h) = \mu_0 - \lambda_0,$$

$$\frac{2}{h}\Delta^2 s_{i-1} = -\frac{4}{N} = -\lambda_i \leqq 0, \qquad i = 1, \cdots, \bar{\imath}-1,$$

$$\frac{2}{h}\Delta^2 s_{i-1} = 0, \qquad i = \bar{\imath}+1, \cdots, N-1.$$

This example was also solved numerically on a Borroughs 5500 computer for $\gamma = \frac{1}{8}$ using Rosen's gradient projection algorithm [10] of nonlinear programming. Computer time for the case of $N = 10$ was 10 seconds and for $N = 20$ was 40 seconds. The numerical results agreed with the known solution.

It is interesting to note that if we use the 1-norm instead of the 2-norm in this example, that is,

$$\text{minimize } \frac{1}{h}\sum_{i=0}^{N-1} |\Delta y_i|,$$

then for $\frac{1}{4} \leqq \gamma \leqq 1$, any nondecreasing sequence $\{y_i\}_0^N$ satisfying the constraints solves the problem, and the minimum value is γ. For $0 \leqq \gamma \leqq \frac{1}{4}$, any sequence $\{y_i\}_0^N$ satisfying the constraints such that $\{y_i\}_{i \leqq N/2}$ is nondecreasing with $y_{(N-1)/2}$ or $y_{N/2}$ on the lower constraint and $\{y_i\}_{i \geqq N/2}$ is nonincreasing solves the problem. If we use the ∞-norm, that is,

$$\text{minimize } \max_{0 \leqq i \leqq N-1} (1/h)|\Delta y_i|,$$

then for $\gamma \leqq 1 - h$, any sequence $\{y_i\}_0^N$ satisfying the constraints with $|\Delta y_i| \leqq |\Delta y_0| = h - h^2$ is a solution. If $1 - h < \gamma \leqq 1$, then $y_i = i\gamma/N$ uniquely solves the problem. As $N \to \infty$, $\Delta y_0/h \to 1$ and $|\Delta y_i| \leqq h$ for $i = 1, \cdots, N-1$. Both of these problems using the 1-norm and ∞-norm can be formulated as linear programs. Numerical experiments verifying the results in these cases were also performed.

7. Applications and remarks. (a) At present we see the principal application of discrete splines as a means to obtaining reasonable approximations to the interpolation problems with side constraints discussed in [9]. Example 6.2 discussed above illustrates how a rather difficult constrained minimization problem (6.2) for determining a certain spline interpolant can be approximately solved by computing a discrete spline. The computation of discrete splines is not difficult in view of the many available nonlinear programming algorithms (cf. [10], [12]). By using the artifice of an enlarged constraint set, Daniel [2] has shown that the discrete splines converge in an appropriate sense to the spline solutions of the continuous problems. Although rates of convergence are lacking, our experience with some simple examples indicates the practicality of this computational approach.

(b) As a possible further application, we mention the theory of best quadrature formulas in the sense of Sard and its close connection with natural polynomial splines (see, e.g., [5], [11]). Certain summation formulas may have similar connections with discrete splines.

(c) The theory of optimal control was convenient in the study of the continuous analogue of (2.4) in [9]. Another promising direction of investigation would be to examine the use of discrete optimal control for (2.4).

(d) Atteia [1] and Laurent [7] have studied rather general splines in Hilbert space settings which can be considered to subsume problem (2.4) in the special case of $p = 2$. The present investigation differs from theirs in that we consider some very specific problems in a Euclidean space, and thus can get more explicit characterization results.

Acknowledgments. We wish to acknowledge useful discussions with J. W. Daniel and K. Gehner, suggestions by M. Golomb and P. Laurent, and the computational programming of D. Kuba.

REFERENCES

[1] M. ATTEIA, *Fonctions spline définies sur un ensemble convexe*, Numer. Math., 12 (1968), pp. 192–210.

[2] J. W. DANIEL, *Convergence of a discretization for constrained spline function problems*, this Journal, 9 (1971).

[3] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton Univ. Press, Princeton, N.J., 1963.

[4] E. ISAACSON AND H. B. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1966.

[5] J. JEROME AND L. L. SCHUMAKER, *On Lg-splines*, J. Approx. Theory, 2 (1969), pp. 29–49.

[6] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, Proc. Second Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., Univ. of California Press, Berkeley, 1951, pp. 481–492.

[7] P. J. LAURENT, *Construction of spline functions in a convex set*, Approximations with Special Emphasis on Spline Functions, I. J. Schoenberg, ed., Academic Press, New York, 1969, pp. 415–446.

[8] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[9] O. L. MANGASARIAN AND L. L. SCHUMAKER, *Splines via optimal control*, Approximations with Special Emphasis on Spline Functions, I. J. Schoenberg, ed., Academic Press, New York, 1969, pp. 119–156.

[10] J. B. ROSEN, *The gradient projection method for nonlinear programming*, SIAM J. Appl. Math., 8 (1960), pp. 181–217.

[11] I. J. SCHOENBERG, *On best approximations of linear operators*, Nederl. Akad. Wetensch. Proc. Ser. A, 67 (1964), pp. 155–163.

[12] W. I. ZANGWILL, *Nonlinear Programming, A Unified Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1969.

SIAM J. CONTROL
Vol. 9, No. 2, May 1971

OPTIMAL STATIONARY CONTROL WITH STATE AND CONTROL DEPENDENT NOISE*

U. G. HAUSSMANN†

Abstract. The steady state optimal linear regulator with state and control dependent noise is analyzed in a manner similar to that developed by Wonham [1]. By state dependent noise we mean Gaussian white noise with coefficient linear in the state variable, and similarly for control dependent noise. Using a Lyapunov criterion for the existence of stationary probability distributions due to Zakai, it is possible to treat equations leading to diffusion processes with degenerate differential generators. It is found that if the noise is sufficiently small, then an optimal control exists. Further analysis, again using Lyapunov methods, yields conditions under which an optimal control exists no matter how large the noise is.

1. Introduction. Consider the linear control system described formally by

$$\dot{x} = Ax - Bu - C(u)\dot{\omega}_1 + D(x)\dot{\omega}_2 + E\dot{\omega}_3. \tag{1.1}$$

Here u is the control, $\dot{\omega}_1$, $\dot{\omega}_2$ and $\dot{\omega}_3$ are independent Gaussian white noise disturbances, and C and D are linear in their arguments. $D(x)\dot{\omega}_2$ and $C(u)\dot{\omega}_1$ can represent wideband random perturbations of the system matrix A and of the input matrix B respectively.

The problem of interest is to choose a control $u = \varphi(x)$ which minimizes, in the steady state, the expected quadratic cost

$$\mathscr{E}\{x'Mx + u'Nu\}.^1 \tag{1.2}$$

In [1] and [2] it is shown that if EE' is positive definite then an optimal control exists, provided the control and state dependent noise is sufficiently small. Using a recent result due to Zakai [3], we are able to remove the restriction on E. Furthermore, we show that if the control dependent noise as well as the state dependent noise due to the stable modes affects only the stable modes, then an optimal control exists no matter how large the control dependent noise $C(u)\dot{\omega}_1$ is. Further conditions which imply the existence of an optimal control independently of the size of the state dependent noise are also given.

We state the problem precisely in § 2. In § 3 we remove the restriction on E, and § 4 contains the proof that existence is independent of the noise level under certain conditions. In § 5 we apply the method to a stability problem and in § 6 we conclude with an example.

2. The problem. We consider the stochastic differential equation

$$dx = Ax\,dt - Bu\,dt - C(u)\,d\omega_1 + D(x)\,d\omega_2 + E\,d\omega_3, \qquad t \geqq 0, \tag{2.1}$$

where[2] $x \in R^n$, $u \in R^m$, and ω_1, ω_2 and ω_3 are independent Wiener processes of

* Received by the editors March 31, 1970.

† Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada. This work was supported by the National Research Council of Canada under Contracts 67-4058 and 67-3105.

[1] $'$ denotes the transpose.

[2] R^n denotes real Euclidean n-space with norm given by $\|x\| = \{\sum_{i=1}^n |x_i|^2\}^{1/2}$.

dimension d_1, d_2 and d_3, respectively. C and D are given by

$$C(u) = \sum_{i=1}^{m} C^i u_i, \tag{2.2}$$

$$D(x) = \sum_{i=1}^{n} D^i x_i, \tag{2.3}$$

respectively, and A, B, C^i, D^i, E are real constant matrices of corresponding dimensions. The pair (A, B) is assumed to be stabilizable.

If u in (2.1) has the form $u = \varphi(x)$ where

$$\|\varphi(x) - \varphi(y)\| \leqq k\|x - y\|, \qquad x, y \in R^n, \tag{2.4}$$

then (2.1) is an equation of Itô's type. Hence if the random variable $x(0)$ is independent of ω_i, $i = 1, 2, 3$, then (2.1) determines a diffusion process

$$X_\varphi = \{x(t) : t \geqq 0\}$$

(see [2], [4]). Of interest to us is the case where X_φ has an invariant probability measure μ_φ defined on the Borel sets of R^n; i.e., if $x(0)$ has the probability distribution μ_φ, then so does $x(t)$, $t > 0$. As has been shown in [5] and [6], if $EE' > 0$ (i.e., positive definite) then suitable sufficient conditions can be given so that an invariant μ_φ exists. In this case μ_φ is unique. Unfortunately for most problems in control theory one does not have $EE' > 0$. For example if the system is given (formally) by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ a_1 & a_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} (\dot{\omega}_3 + u) \tag{2.5}$$

so that it corresponds to

$$\frac{d^2y}{dt^2} = a_2 \frac{dy}{dt} + a_1 y + \frac{d\omega_3}{dt} + u,$$

then EE' is not positive definite (so that the differential generator of the associated diffusion process is not fully elliptic) even though the system is controllable. Recently Zakai [3] has given alternate conditions which guarantee that an invariant μ_φ exists. However it is in general not unique. (For systems of the form (2.5) it is unique [3].)

We define Φ, the class of admissible control laws, to be the set of functions $\varphi(\cdot)$ such that:

(i) φ satisfies (2.4) for some constant k;

(ii) an invariant probability measure μ_φ exists;

(iii) for any invariant probability measure μ,

$$\mathcal{E}_\mu\{\|x\|^2\} = \int_{R^n} \|x\|^2 \mu(dx) < \infty. \tag{2.6}$$

Let $L(x, u)$ be given by

$$L(x, u) = x'Mx + u'Nu, \tag{2.7}$$

where M and N are constant symmetric positive definite matrices of dimension

$n \times n$, $m \times m$, respectively. For φ in Φ let $\mathscr{E}_\varphi\{L(x,\varphi)\}$ denote the infimum of $\mathscr{E}_\mu\{L(x,\varphi)\}$ taken over all probability measures invariant with respect to X_φ. In [1], [2] Wonham considered the problem (with the extra restriction on E) of finding conditions on C and D such that a control φ° exists, optimal in the sense that

$$\mathscr{E}_{\varphi^\circ}\{L(x,\varphi^\circ)\} = \inf[\mathscr{E}_\varphi\{L(x,\varphi)\} : \varphi \in \Phi].$$

We shall refine these conditions.

Let us point out here that taking the infimum of $\mathscr{E}_\mu$ in the definition of $\mathscr{E}_\varphi$ is irrelevant, for under certain assumptions which we shall always make, it follows that for the optimal control φ°, all invariant probability measures yield the same expected cost, i.e.,

$$\mathscr{E}_{\varphi^\circ}\{L(x,\varphi^\circ)\} = \mathscr{E}_\mu\{L(x,\varphi^\circ)\} \tag{2.8}$$

for all invariant probability measures μ (see Theorem 3.2).

3. Admissible and optimal controls. First we must show that Φ is nonvoid. Let $\mathscr{L}_u$ be the differential operator given by

$$\mathscr{L}_u V(x) = \tfrac{1}{2}\operatorname{tr}\{C(u)'V_{xx}C(u) + D(x)'V_{xx}D(x) + E'V_{xx}E\} + (Ax - Bu)'V_x, \tag{3.1}$$

where V_x denotes the vector $\partial V/\partial x_i$ and V_{xx} denotes the matrix $\partial^2 V/\partial x_i \partial x_j$. We assume from now on that $u = \varphi(x)$ in (2.1) with φ satisfying (2.4). Then $\mathscr{L}_\varphi$ is the differential generator of X_φ, also written simply as $\mathscr{L}$.

Let us consider the following control law φ and function V:

$$\varphi(x) = Kx, \qquad V(x) = x'Px, \tag{3.2}$$

where K and P are two constant matrices, $P \geqq 0$ (i.e., $x'Px \geqq 0, x \in R^n$), to be determined such that

$$\mathscr{L}_\varphi V(x) = \lambda - L(x,\varphi(x)) \tag{3.3}$$

for some $\lambda > 0$. Unfortunately such a V does not satisfy certain conditions imposed in [3][3] and so results of [3] do not apply directly.

THEOREM 3.1. *Assume P and K exist ($P \geqq 0$) such that* (3.2) *and* (3.3) *are satisfied. Then φ is an admissible control.*

Proof. For $n = 1, 2, \cdots$ define

$$\begin{aligned} D_n(x) &= \begin{cases} D(x) & \text{if } \|x\| \leqq n, \\ D[x(n/\|x\|)] & \text{if } \|x\| > n, \end{cases} \\ \tilde{C}_n(x) &= \begin{cases} C(Kx) & \text{if } \|x\| \leqq n, \\ C[(n/\|x\|)Kx] & \text{if } \|x\| > n. \end{cases} \end{aligned} \tag{3.4}$$

[3] Zakai calls this condition (B). It requires that $\mathscr{E}\{f(t)|x(0) = a\}$ be bounded in t on any finite t interval, for any $a \in R^n$, where

$$f(t) = \|V_x(x(t))'[C(K(x(t))), D(x(t)), E]\|^2.$$

We can only conclude that $f(t) \leqq k[1 + \|x(t)\|^4]$ for some $k > 0$.

Then[4]

$$\begin{aligned} \|D(x) - D_n(x)\| &\leqq \begin{cases} 0 & \text{if } \|x\| \leqq n, \\ k\|x\| & \text{if } \|x\| > n, \end{cases} \\ \|C(Kx) - \tilde{C}_n(x)\| &\leqq \begin{cases} 0 & \text{if } \|x\| \leqq n, \\ k\|x\| & \text{if } \|x\| > n, \end{cases} \end{aligned} \tag{3.5}$$

where

$$k = \max\left\{\|K\| \sum_{i=1}^{n} \|C^i\|, \sum_{i=1}^{n} \|D^i\|, \|A - BK\|\right\}.$$

Also $D_n(\cdot)$ and $\tilde{C}_n(\cdot)$ are Lipschitz continuous with constant k (as are $D(\cdot)$ and $C(\varphi(\cdot))$). Now let x^n be the unique solution of

$$dx = (A - BK)x\,dt - \tilde{C}_n(x)\,d\omega_1 + D_n(x)\,d\omega_2 + E\,d\omega_3, \qquad x(0) = \xi. \tag{3.6}$$

We let $\mathscr{L}_n$ be the corresponding differential generator so that

$$\mathscr{L}_n V(x) = \begin{cases} x'\{\Delta(P) + K'\Gamma(P)K + \hat{A}'P + P\hat{A}\}x + \operatorname{tr}\{E'PE\}, & \|x\| \leqq n, \\ x'\left\{\dfrac{n^2}{\|x\|^2}[\Delta(P) + K'\Gamma(P)K] + \hat{A}'P + \hat{P}A\right\}x + \operatorname{tr}\{E'PE\}, & \|x\| > n, \end{cases} \tag{3.7}$$

where

$$(\Delta(P))_{ij} = \operatorname{tr}((D^i)'PD^j), \quad (\Gamma(P))_{ij} = \operatorname{tr}((C^i)'PC^j), \quad \hat{A} = A - BK.$$

Then $\mathscr{L}_n V(x) = \mathscr{L}V(x)$ if $\|x\| \leqq n$, and $\mathscr{L}_n V(x) \leqq \mathscr{L}V(x)$ if $\|x\| > n$, because $\Delta(P) + K'\Gamma(P)K \geqq 0$ as $P \geqq 0$.

We shall require the following three properties of V:

$$\mathscr{E}_x\{V(x(t))\} = \mathscr{E}\{V(x(t))|x(0) = x\} \leqq c\mathscr{E}_x\{\|x(t)\|^2\}, \qquad c < \infty, \tag{3.8a}$$

$$\mathscr{E}_x\{|\mathscr{L}_n V(x(t))|\} \leqq c\mathscr{E}_x\{(1 + \|x(t)\|^2)\}, \qquad n = 1, 2, \cdots, \tag{3.8b}$$

$$\mathscr{E}_x\{\|V_x(x(t))'[\tilde{C}_n(x(t)), D_n(x(t)), E]\|^2\} \leqq c\mathscr{E}_x\{\|x(t)\|^2\}. \tag{3.8c}$$

Thus each of the three left-hand expressions is bounded in t over finite t intervals, for each x in R^n.

As $L(x, u) = x'Mx + u'Nu$ then there is an $R_0 < \infty$ such that $\lambda - L(x, Kx) < -\eta$ for some $\eta > 0$ if $\|x\| > R_0$. Now the proof of [3, Theorem 1] shows that

$$-\frac{1}{t}\mathscr{E}\{V(\xi)\} + \eta \leqq (\rho + \eta)\frac{1}{t}\int_0^t \Pr\{\|x^n(s)\| \leqq R_0\}\,ds \tag{3.9}$$

for any n if $x^n(0) = \xi$ has the distribution μ^+, where μ^+ is any Borel probability measure on R^n with compact support. Here ρ is the maximum of $\mathscr{L}V(x)$ over $\|x\| \leqq R_0$, so that $\mathscr{L}_n V(x) \leqq \rho$ for all n.

[4] If F is a matrix mapping R^n into R^m, then $\|F\|$ denotes the usual operator norm.

In the Appendix we show that $\mathscr{E}\{\|x_t^n - x_t\|^2\} \to 0$. Hence a subsequence, which we call x_t^n again, converges to x_t with probability one, i.e.,

$$\Pr\{\lim_{n\to\infty} \|x_t^n - x_t\| = 0\} = 1.$$

By Egoroff's theorem $x_t^n(\omega)$ converges to $x_t(\omega)$ almost uniformly (ω). Hence for any $\delta > 0$ there is an $n(s)$ such that for $n > n(s)$,

$$\Pr\{\|x_s\| \leqq R_0 + 1\} \geqq \Pr\{\|x_s^n\| \leqq R_0\} - \delta$$

and so[5]

$$\Pr\{\|x_s\| \leqq R_0 + 1\} \geqq \lim_{n\to\infty} \Pr\{\|x_s^n\| \leqq R_0\}.$$

By (3.9) and the dominated convergence theorem ($t < \infty$) we have

$$-\frac{1}{t}\mathscr{E}\{V(\xi)\} + \eta \leqq (\rho + \eta)\frac{1}{t}\int_0^t \Pr\{\|x(s)\| \leqq R_0 + 1\}\, ds,$$

and so (see [3]) the process X_φ has an invariant measure.

Now $|L(x_s^n) - L(x_s)| \leqq k_1\|x_s^n - x_s\|^2$, so that

$$\left|\mathscr{E}_x\left\{\int_0^t L(x_s^n)\, ds - \int_0^t L(x_s)\, ds\right\}\right| \leqq k_1\mathscr{E}_x\left\{\int_0^t \|x_s^n - x_s\|^2\, ds\right\}.$$

As $t < \infty$, $\mathscr{E}_x\{\|x_s^n\|^2\} < \infty$, $\mathscr{E}_x\{\|x_s\|^2\} < \infty$ and as $x(s, \omega)$ is measurable in (s, ω), then

$$\lim_{n\to\infty} \mathscr{E}_x\left\{\int_0^t \|x_s^n - x_s\|^2\, ds\right\} = \lim_{n\to\infty} \int_0^t \mathscr{E}_x\{\|x_s^n - x_s\|^2\}\, ds$$

$$= \int_0^t \lim_{n\to\infty} \mathscr{E}_x\{\|x_s^n - x_s\|^2\}\, ds.$$

By Itô's formula and (3.8),

$$0 \leqq \mathscr{E}_x\{V(x_t^n)\} = V(x) + \mathscr{E}_x\left\{\int_0^t \mathscr{L}_n V(x_s^n)\, ds\right\}$$

$$\leqq V(x) + \lambda t - \mathscr{E}_x\left\{\int_0^t L(x_s^n)\, ds\right\}.$$

It follows that $\mathscr{E}_x\{\int_0^t L(x_s)\, ds\} \leqq V(x) + \lambda t$. Now the argument given in [3, Theorem 2] shows that

$$\int_{R^n} L(x)\mu(dx) \leqq \lambda$$

for any invariant probability measure μ. Hence $\mathscr{E}_\mu\{\|x\|^2\} < \infty$ and $\varphi(x) = Kx$ is admissible. The theorem is established.

[5] This limit can be shown to exist.

COROLLARY. *Under the condition of the above theorem it follows that* X_φ *is stable insofar as*

$$\mathscr{E}_x\{\tau\} \leqq V(x)/\eta, \qquad \|x\| > R_0, \tag{3.10}$$

where τ *is the first passage time of* x_t *to the set* $\{x : \|x\| \leqq R_0\}$.

This follows because if τ_n denotes the corresponding random variable for x_t^n, then

$$0 \leqq \mathscr{E}_x\{V(x^n(t \wedge \tau_n))\} = V(x) + \mathscr{E}_x\left\{\int_0^{t \wedge \tau_n} \mathscr{L}_n V(x^n(s))\, ds\right\},$$

i.e.,

$$0 \leqq V(x) - \eta \mathscr{E}_x\{t \wedge \tau_n\}.$$

But $t \wedge \tau_n(\omega) \leqq t$, and so

$$0 \leqq V(x) - \eta \mathscr{E}_x\{t \wedge \lim_{n \to \infty} \tau_n\} = V(x) - \eta \mathscr{E}_x\{t \wedge \tau\}.$$

Now (3.10) follows by monotone convergence. Relation (3.10) implies that the process is recurrent (and positive if $\mathscr{L}$ is fully elliptic) [5].

Evidently (3.2) yields a solution of (3.3) if and only if

$$\lambda = \operatorname{tr}\{E'PE\} \tag{3.11}$$

and

$$\Delta(P) + K'\Gamma(P)K + (A - BK)'P + P(A - BK) + M + K'NK = 0. \tag{3.12}$$

If (A, B) is stabilizable, then K can be chosen so that $A - BK$ is stable, i.e., all the eigenvalues have negative real parts. The following lemma shows that if C^i, D^i are sufficiently small then (3.12) has a unique positive definite solution. This and Theorem 3.1 imply that $\varphi \in \Phi$.

LEMMA 3.1. *If* $Q > 0$ *and* A *is stable, then*

$$\Delta(P) + K'\Gamma(P)K + A'P + PA + Q = 0 \tag{3.13}$$

has a unique solution $P > 0$, *provided*

$$\left\| \int_0^\infty e^{tA'}[K'\Gamma(I)K + \Delta(I)]e^{tA}\, dt \right\| < 1. \tag{3.14}$$

For the proof see [1], [2].

Thus an admissible control exists if

$$\inf_K \left\| \int_0^\infty e^{t(A-BK)'}[K'\Gamma(I)K + \Delta(I)]e^{t(A-BK)}\, dt \right\| < 1, \tag{3.15}$$

where the infimum is taken only over K such that $A - BK$ is stable. In [1], [2] the following theorem is proved.

THEOREM 3.2. *If* (A, B) *is stabilizable, if* $M > 0$, $N > 0$, *and if* Γ, Δ *satisfy* (3.15), *then an optimal control exists, of the form*

$$\varphi^\circ(x) = K^\circ x. \tag{3.16}$$

Furthermore,

$$K^\circ = [\Gamma(P^\circ) + N]^{-1}B'P^\circ, \tag{3.17}$$

where P° is the unique solution, in the class of positive semidefinite matrices, of the equation

$$M + A'P + PA + \Delta(P) - PB[\Gamma(P) + N]^{-1}B'P = 0. \tag{3.18}$$

The minimum cost is

$$\mathscr{E}_{\varphi^\circ}\{L(x, \varphi^\circ(x))\} = \operatorname{tr}\{E'P^\circ E\}. \tag{3.19}$$

From the same proof it follows that (2.8) holds.

4. Dependence on noise. Theorem 3.2 shows that the existence of an optimal control depends on C and D only through (3.15), although of course the actual value of the control always depends on C and D. We now investigate (3.15) in an effort to determine if this condition can be relaxed. Let $\alpha(\lambda)$ be the minimal polynomial of A. It factors as follows:

$$\alpha(\lambda) = \alpha^+(\lambda)\alpha^-(\lambda), \tag{4.1}$$

where all the zeros of α^+ lie in the closed right half-plane and those of α^- in the open left half-plane. Also define

$$R^n_+ = \{x \in R^n : \alpha^+(A)x = 0\}, \qquad R^n_- = \{x \in R^n : \alpha^-(A)x = 0\}.$$

As α^+ and α^- are coprime it follows that $R^n = R^n_+ \oplus R^n_-$, although the direct sum is not necessarily orthogonal. Let P^+ and P^- be the projections of R^n onto R^n_+ and R^n_- along R^n_- and R^n_+, respectively.

The next two lemmas are useful for computation. The proofs are straightforward and so are omitted. However first we need some notation. Let $g_1, \cdots, g_r$, $r \leqq d_2$, be a basis of the range space of the $n^2 \times d_2$ matrix

$$\begin{bmatrix} D^1 \\ D^2 \\ \vdots \\ D^n \end{bmatrix}, \tag{4.2}$$

and let G^i be the matrix composed of n columns whose direct sum is g_i; i.e., if G^i_j is the jth column of G^i, then

$$g_i = \begin{bmatrix} G^i_1 \\ \vdots \\ G^i_n \end{bmatrix}, \qquad i = 1, 2, \cdots, r. \tag{4.3}$$

LEMMA 4.1. *Let $H(\omega) = [h_1, h_2, \cdots, h_n]$, where $h_i(\omega) = D^i\omega$. Then R^n_- is invariant with respect to $H(\omega)$ for all $\omega \in R^{d_2}$ if and only if R^n_- is invariant with respect to G^i, $i = 1, \cdots, r$.*

LEMMA 4.2. *Assume $\{e_i\}_1^n$ is the usual basis in R^n. If $\{e_i\}_{i \in S}$ spans R^n_-, then R^n_- is invariant with respect to $H(\omega)$ for all ω if and only if $\{D^i\} \subseteq R^n_-$, $i \in S$, where $\{D^i\}$ denotes the range space of D^i.*

The following lemma expresses an algebraic condition as an analytic one and is useful for computation. Krasovskii stated a related sufficient condition [7]. Let s be an eigenvector of A, i.e., for some λ, possibly complex,

$$As = \lambda s.$$

Let s^+ denote any eigenvector s for which the corresponding eigenvalue has non-negative real part. If C is an $n \times m$ matrix, then define $R_C[A]$ by

$$R_C[A] = \{[C, AC, A^2C, \cdots, A^{n-1}C]\}. \tag{4.4}$$

Recall that $\{\cdot\}$ denotes the range space. Suppose the dimension of this space is q. Then C_A denotes an $n \times q$ matrix whose columns are a basis of $R_C[A]$, $n \geqq q$.

LEMMA 4.3. *A necessary and sufficient condition that* $\{C\} \subseteq R^n_-$ *is that*

$$\operatorname{rank}(A - \lambda I)C_A = q \tag{4.5}$$

for each λ *with* $\operatorname{Re}\lambda \geqq 0$.

Proof. We first note that $\{C\} \subseteq R^n_-$ if and only if no s^+ lies in $R_C[A]$. This follows readily by considering A in Jordan form and noting that in that case R^n_+ and R^n_- are orthogonal. Also note that $\{C\} \subseteq R^n_-$ if and only if $c_i \in R^n_-$, $i = 1, \cdots, m$, where c_i is the ith column of C, and that $R_C[A] = \operatorname{span}\{R_{c_i}[A] : i = 1, \cdots, m\}$.

If c^i_A is the ith column of C_A, then any $s \in R_C[A]$ is of the form

$$s = \sum_{i=1}^{q} \sigma_i c^i_A.$$

It follows that some s^+ is in $R_c[A]$ if and only if

$$(A - \lambda I)C_A\sigma = 0$$

has a nontrivial solution σ for some λ with $\operatorname{Re}\lambda \geqq 0$. A simple argument now completes the proof.

We now wish to split our system (2.1) into two parts, one on R^n_+ and one on R^n_-. If R^n_+ has dimension l_1, we can choose a map T_+ of R^n onto R^{l_1} so that T_+ restricted to R^n_+ is one-to-one and onto, and $T_+R^n_- = \{0\}$. Then T_+ is an $l_1 \times n$ matrix. We denote its generalized inverse by $T^\dagger_+$ so that

$$T_+T^\dagger_+ = I, \quad T^\dagger_+T_+ = P^+, \quad P^+T^\dagger_+ = T^\dagger_+, \quad T_+P^+ = T_+.$$

Similarly define T_-, an $l_2 \times n$ matrix corresponding to R^n_-. Also define

$$\tilde{D}^j_+ = T_+ \sum_{i=1}^{n} D^i(T^\dagger_+)_{ij}, \qquad j = 1, 2, \cdots, l_1, \tag{4.6}$$

and

$$\Delta_+(R)_{ij} = \operatorname{tr}\{(\tilde{D}^i_+)'R(\tilde{D}^j_+)\}, \qquad i, j = 1, \cdots, l_1. \tag{4.7}$$

If we let $A_+ = T_+AT^\dagger_+$ and define A_-, Δ_- and $\tilde{D}^j_-$ likewise, then A_- is a stable $l_2 \times l_2$ matrix.

THEOREM 4.1. *If* (A, B) *is stabilizable (i.e.,* $R^n_+ \subseteq R_B[A]$, [8]), *and if* R^n_- *is invariant with respect to* $H(\omega)$ *for all* ω, *then the system* (2.1) *has an admissible control*

φ, provided that

$$\left\| \int_0^\infty e^{tA'_-} \Delta_-(I) e^{tA_-} \, dt \right\|$$

$$= \sup \left\{ \int_0^\infty x' e^{tA'} \Delta(T'_- T_-) e^{tA} x \, dt : x \in R^n_- , \| T_- x \| \leqq 1 \right\} < 1, \tag{4.8}$$

and

$$\inf_{K_+} \left\| \int_0^\infty e^{t(A_+ - T_+ BK_+)'} [\Delta_+(I) + K'_+ \Gamma(T'_+ T_+) K_+] e^{t(A_+ - T_+ BK_+)} \, dt \right\|$$

$$= \inf_K \sup \left\{ \int_0^\infty x' e^{t(A - P^+ BK)'} [\Delta(T'_+ T_+) + K' \Gamma(T'_+ T_+) K] \right. \tag{4.9}$$

$$\left. \cdot e^{t(A - P^+ BK)} x \, dt : x \in R^n_+ , \| T_+ x \| \leqq 1 \right\} < 1.$$

Moreover φ can be chosen linear in x, independent of $P^- x$. *Also* (4.8) *is satisfied if*

$$\sum_{ijk} |(T_- D^k)_{ij}|^2 < \left[\sup \left\{ \int_0^\infty \| e^{tA} x \|^2 \, dt : x \in R^n_- , \| T_- x \| \leqq 1 \right\} \right]^{-1}. \tag{4.10}$$

Proof. As R^n_- is invariant under $H(\omega)$ then

$$T_+ D(x)\omega = T_+ P^+ D(P^+ x)\omega = T_+ D(T^\dagger_+ T_+ x)\omega = \tilde{D}_+(T_+ x)\omega, \tag{4.11}$$

where $\tilde{D}_+(y) = \sum_{i=1}^{l_1} \tilde{D}^i_+ y_i$. We let $T_+ x = y$, $T_- x = z$. Then (2.1) becomes

$$dy = (A_+ y - T_+ Bu)\, dt - T_+ C(u)\, d\omega_1 + \tilde{D}_+(y)\, d\omega_2 + T_+ E\, d\omega_3, \tag{4.12}$$

$$\begin{aligned} dz = {} & (A_- z - T_- Bu)\, dt - T_- C(u)\, d\omega_1 + \tilde{D}_-(z)\, d\omega_2 + T_- E\, d\omega_3 \\ & + T_- D(T^\dagger_+ y)\, d\omega_2, \end{aligned} \tag{4.13}$$

where $\tilde{D}_-(z) = \sum_{i=1}^{l_2} \tilde{D}^i_- z_i = T_- D(T^\dagger_- z)$. Consider (4.13) with $y = 0$, $u = 0$, i.e.,

$$dz = A_- z\, dt + \tilde{D}_-(z)\, d\omega_2 + T_- E\, d\omega_3. \tag{4.14}$$

As (4.14) has a stable matrix, then by (4.8) and Lemma 3.1 there is a symmetric positive definite $l_2 \times l_2$ matrix P_2 such that

$$\Delta_-(P_2) + A'_- P_2 + P_2 A_- + I = 0. \tag{4.15}$$

As $R^n_+ \subseteq R_B[A]$ then $R^{l_1} = T_+ R^n_+ \subseteq R_{T_+ B}[A_+]$ so that $(A_+, T_+ B)$ is controllable [8]. Now (4.9) and Lemma 3.1 show that there are an $l_1 \times m$ matrix K_+ and an $l_1 \times l_1$ matrix $P_1 > 0$, such that

$$\begin{aligned} & K'_+ \Gamma(T'_+ P_1 T_+) K_+ + \Delta_+(P_1) + (A_+ - T_+ BK_+)' P_1 \\ & \qquad + P_1 (A_+ - T_+ BK_+) + I = 0. \end{aligned} \tag{4.16}$$

Let $\lambda_1 = \operatorname{tr} \{(T_+ E)' P_1 (T_+ E)\}$, $\lambda_2 = \operatorname{tr} \{(T_- E)' P_2 (T_- E)\}$, and let $K = K_+ T_+$.

To show that $\varphi(x) = Kx$ is an admissible control we shall use Theorem 3.1. Consider $V(x) = x'Px$, where $P = T'_+ P_1 T_+ + \beta T'_- P_2 T_- > 0$ and $\beta > 0$ is to be

chosen so that for some $\lambda > 0, \alpha > 0$,

$$\mathscr{L}_\varphi V(x) \leqq \lambda - \alpha\|x\|^2. \tag{4.17}$$

This condition, rather than (3.3), suffices in Theorem 3.1. Using (4.11), (4.15) and (4.16) we have

$$\begin{aligned}
\mathscr{L}_\varphi V(x) &= y'K_+'\Gamma(T_+'P_1T_+)K_+y + \beta y'K_+'\Gamma(T_-'P_2T_-)K_+y \\
&\quad + y'\Delta_+(P_1)y + \beta z'\Delta_-(P_2)z + \beta\{y'(T_+^\dagger)'\Delta(T_-'P_2T_-)T_-^\dagger z \\
&\quad + z'(T_-^\dagger)'\Delta(T_-'P_2T_-)T_+^\dagger y + y'(T_+^\dagger)'\Delta(T_-'P_2T_-)T_+^\dagger y\} \\
&\quad + \lambda_1 + \beta\lambda_2 + y'(A_+ - T_+BK_+)'P_1y + y'P_1(A_+ - T_+BK_+)y \\
&\quad + \beta z'A_-'P_2z + \beta z'P_2A_-z - 2\beta y'(T_-BK_+)'P_2z \\
&= \lambda_1 + \beta\lambda_2 - \|y\|^2 - \beta\|z\|^2 + \beta\{y'K_+'\Gamma(T_-'P_2T_-)K_+y \\
&\quad + 2z'(T_-^\dagger)'\Delta(T_-'P_2T_-)T_+^\dagger y + y'(T_+^\dagger)'\Delta(T_-'P_2T_-)T_+^\dagger y \\
&\quad - 2z'P_2T_-BK_+y\} \\
&\leqq \lambda_1 + \beta\lambda_2 - (\beta/(4\gamma))\|x\|^2
\end{aligned} \tag{4.18}$$

if $0 < \beta$ is sufficiently small and $\gamma = \max\{\|T_+^\dagger\|^2, \|T_-^\dagger\|^2\}$. This last assertion is proved as follows. Let $f(y, z)$ be given by

$$\begin{aligned}
f(y,z) &= y'\{K_+'\Gamma(T_-'P_2T_-)K_+ + (T_+^\dagger)'\Delta(T_-'P_2T_-)T_+^\dagger\}y \\
&\quad + 2z'\{(T_-^\dagger)'\Delta(T_-'P_2T_-)T_+^\dagger - P_2T_-BK_+\}y.
\end{aligned}$$

Consider

$$\|y\|^2(1 - \tfrac{1}{2}\beta) + \tfrac{1}{2}\beta\|z\|^2 \geqq \beta c_1\|y\|^2 + 2\beta c_2\|y\|\|z\|, \tag{4.19}$$

where $c_1 \geqq 0, c_2 \geqq 0$. If $\|z\| = 0$, then (4.19) holds for all y if $0 < \beta \leqq (c_1 + \frac{1}{2})^{-1}$. Otherwise let $\|y\| = k\|z\|, k$ variable in $[0, \infty)$. Then (4.19) holds for all k if

$$g(k) = k^2(1 - \tfrac{1}{2}\beta) + \tfrac{1}{2}\beta - \beta c_1k^2 - 2\beta c_2k \geqq 0 \tag{4.20}$$

for all $k \in [0, \infty)$. But if $0 < \beta < (c_1 + \frac{1}{2})^{-1}$ and if $\beta < (2c_2^2 + c_1 + \frac{1}{2})^{-1}$, then (4.20) holds for all $k \in [0, \infty)$. If

$$\begin{aligned}
c_1 &= \|K_+'\Gamma(T_-'P_2T_-)K_+ + (T_+^\dagger)'\Delta(T_-'P_2T_-)T_+^\dagger\|, \\
c_2 &= \|(T_-^\dagger)'\Delta(T_-'P_2T_-)T_+^\dagger - P_2T_-BK_+\|,
\end{aligned}$$

then

$$f(y, z) \leqq c_1\|y\|^2 + 2c_2\|y\|\,\|z\|. \tag{4.21}$$

As $\|x\|^2 \leqq 2\max\{\|T_+^\dagger\|^2, \|T_-^\dagger\|^2\}[\|y\|^2 + \|z\|^2]$, then

$$\begin{aligned}
\lambda_1 + \beta\lambda_2 - \beta/(4\gamma)\|x\|^2 &\geqq \lambda_1 + \beta\lambda_2 - \tfrac{1}{2}\beta[\|y\|^2 + \|z\|^2] \\
&\geqq \lambda_1 + \beta\lambda_2 - \|y\|^2 - \beta\|z\|^2 + \beta f(y, z)
\end{aligned}$$

by (4.21) and (4.19), and (4.18) is established. Now Theorem 3.1 shows that $\varphi(x) = Kx$ is in Φ.

COROLLARY 1. *If* $\{C^i\} \subseteq R^n_-$ *for* $i \in S \subseteq \{1, 2, \cdots, m\}$ *in addition to the assumptions of the theorem, then an admissible control exists no matter how large* C^i, $i \in S$, *is.*

Proof. Existence depends on C^i only through Γ in (4.9). As $\Gamma(T'_+ T_+)_{ij} = \operatorname{tr}\{(T_+ C^i)'(T_+ C^j)\}$ and as $P^+ C^i = 0$ for $i \in S$, then $\Gamma(T'_+ T_+)$ is independent of C^i, $i \in S$.

We recall that if $\{e_i\}_{i \in S}$ spans R^n_-, then by hypothesis and Lemma 4.2, $\{D^i\} \subseteq R^n_-$, $i \in S$, so that $\Delta_+(I)$ is independent of D^i, $i \in S$.

COROLLARY 2. *If* $R^n_+ \subseteq R_B[A]$, $\{C^i\} \subseteq R^n_-$, $i = 1, \cdots, m$, *and if* $R^n_- \subseteq \bigcap_{i=1}^r N[G^i]$ *and* $\{D^i\} \subseteq R^n_-$, $i = 1, \cdots, n$, *then an admissible control exists.*[6]

Proof. As $R^n_- \subseteq N[H(\omega)]$ by hypothesis then $D(P^- x) = 0$ so that $\Delta_-(I) = 0$. Moreover $\Gamma(T'_+ T_+) = 0$ and $\Delta(T'_+ T_+) = 0$ so the conclusion follows.

Let us sum up what this last corollary says. We assume that at least the unstable modes of A are controllable, that the state and control dependent noise affects only the stable modes and that the state dependent noise is due only to the unstable modes; then an admissible control exists.

Let us now assume that the conditions of Theorem 4.1 are satisfied so that we have an admissible control $\varphi(x) = Kx$ and a function $V(x) = x'Px$ such that

$$\mathscr{L}_\varphi V(x) \leqq \lambda - \alpha \|x\|^2, \qquad \lambda > 0, \quad \alpha > 0.$$

In fact $\mathscr{L}_\varphi V(x)$ is of the form

$$\mathscr{L}_\varphi V(x) = l(P) + f_\varphi(x, P),$$

where $l(\cdot)$ and $f_\varphi(x, \cdot)$ are linear and $f_\varphi(x, P) \leqq -\alpha \|x\|^2$, $\alpha > 0$. We can choose $\eta < \infty$ such that

$$\alpha\eta \|x\|^2 \geqq L(x, Kx),$$

and so

$$\mathscr{L}_\varphi(x'Qx) \leqq l(Q) - L(x, \varphi(x))$$

if $Q = \eta P$. Hence

$$0 \geqq \Delta(Q) + K'\Gamma(Q)K + (A - BK)'Q + Q(A - BK) + M + K'NK.$$

This condition suffices to make Wonham's argument [1] proving the existence of an optimal control valid.

THEOREM 4.2. *Under the assumptions of Theorem* 4.1 *or of one of its corollaries, an optimal control exists and* (3.16), (3.17), (3.18) *and* (3.19) *are valid.*

We point out that the following may be useful with regard to (4.9). Consider the system

$$dx/dt = Ax - Bu, \qquad x(0) = x \tag{4.22}$$

and

$$W(x) = \int_0^\infty x(t)'\Delta x(t) + \alpha u(t)'u(t)\, dt,$$

[6] $N[A]$ denotes the null space of A.

where (A, B) are controllable, A, B, Δ are real $n \times n$, $n \times m$ and $n \times n$ matrices, respectively, $\Delta > 0$, and α is a positive scalar.

From the theory of the linear regulator it is known that $\inf_{u=Kx} W(x) = x'P(\alpha)x$, where $P(\alpha)$ is the unique positive solution of

$$A'P + PA - PBB'P/\alpha + \Delta = 0. \tag{4.23}$$

It can be shown that $P(\alpha) - P(\beta) \geqq 0$ if $\alpha > \beta > 0$, and hence that $\lim_{\alpha \to 0} P(\alpha) = P_0$ exists. Then

$$\|P_0\| = \inf_K \left\| \int_0^\infty e^{t(A-BK)'} \Delta e^{t(A-BK)}\, dt \right\|.$$

LEMMA 4.4. $\|P_0\| = 0$ *if and only if* $\operatorname{rank} B = n$.

Proof. By (4.23),

$$A'P_0 + P_0A - \lim_{\alpha \to 0} P(\alpha)\frac{BB'}{\alpha}P(\alpha) + \Delta = 0,$$

and so $P_0BB'P_0 = 0$. If rank $B = n$, then $P_0 = 0$. Conversely if $P_0 = 0$, then

$$\lim_{\alpha \to 0} \frac{P(\alpha)}{\sqrt{\alpha}} BB' \frac{P(\alpha)}{\sqrt{\alpha}} = \Delta.$$

Let p_α^i be the ith column vector of $P(\alpha)$ and let $q_\alpha^i = (1/\sqrt{\alpha})B'p_\alpha^i$. Thus $q_\alpha^i \to q^i$ and $(q^i)'q^j = \Delta_{ij}$. As $\Delta > 0$ then the q^i are linearly independent. Hence $\operatorname{rank} B = n$.

Now we can replace the condition $\{D^i\} \subseteq R_-^n$, $i = 1, \cdots, n$, in Corollary 2 of Theorem 4.1 by $R_+^n \subseteq \{B\}$.

The method of this section yields the following alternate form of Corollary 2. Suppose $C = 0$ and W is a subspace such that $\{D^i\} \subseteq W \subseteq \bigcup_{j=1}^r N[G^j]$, $i = 1, \cdots, n$. Suppose there is a K such that $\hat{A} = A - BK$ has stable modes $\hat{R}_-^n$ and unstable modes $\hat{R}_+^n$, where $W \subseteq \hat{R}_-^n$. We also assume that there is a space W^c complementary to W such that $W^c \subseteq R_B[A]$, and that W and W^c completely reduce $\hat{A}$. Then an admissible control exists independently of D. If $W^c \subseteq \{B\}$, then $\{D^i\} \subseteq W$ is not required.

5. An application to stability. If we set $E = 0$ in (2.1) and if $u = \varphi(x) = Kx$, then $x(t) \equiv 0$ is the unique solution of (2.1) with $x(0) = 0$; i.e., 0 is an equilibrium point. In this case the degenerate probability measure μ with support at 0 is invariant and $\mathscr{E}_\mu\{x'x\} = 0$.

Let us consider the problem of stability of the second mean. We have

$$\begin{aligned} \frac{d}{dt}\mathscr{E}_x\{x_t'Px_t\} &= \mathscr{E}_x\{\mathscr{L}_\varphi(x_t'Px_t)\} \\ &= \mathscr{E}_x\{x_t'[\Delta(P) + K'\Gamma(P) + A'P + PA]x_t\} \\ &\leqq \alpha\lambda(P)\mathscr{E}_x\{x_t'Px_t\} \end{aligned} \tag{5.1}$$

if $P > 0$, where $\alpha > 0$ and $\lambda(P)$ is the largest eigenvalue of $\Delta(P) + K'\Gamma(P)K + A'P + PA$. If $\lambda(P) < 0$, then there exist $\beta, \gamma > 0$ such that

$$\mathscr{E}_x\{x_t'Px_t\} \leqq \beta(x'Px)\exp\{-\gamma t\}, \tag{5.2}$$

and so $\mathscr{E}\{x_t'Px_t\}$ is globally exponentially stable. It follows that if there is a $P > 0$ such that $\lambda(P) < 0$, then the second mean is exponentially stable; i.e., (5.2) holds with $P = I$. On the other hand, Theorem 4.1 and its corollaries give conditions such that $P > 0$ exists and

$$\mathscr{L}_\varphi(x'Px) \leqq -\eta x'Px$$

with $\eta > 0$. Hence if the conditions of the theorem or one of its corollaries are satisfied, then we have exponential stability of the second mean. This implies [9] that the zero solution is asymptotically stable with probability one. More complete results have been obtained in [10] for systems generated by linear nth order differential equations whose constant coefficients are distorted by white noise.

As we mentioned, the invariant measure is degenerate, and so the stationary optimal control is of no interest. However, now the problem of minimizing

$$\mathscr{E}_x\left\{\int_0^\infty [x_t'Mx_t + u_t'Nu_t]\,dt\right\} \tag{5.3}$$

has a solution under exactly the conditions of Theorem 4.1 (see [1]).

6. Example. We consider only one elementary system which does, however, show that the results of § 4 are an improvement over Theorem 3.2. Let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad C(u) = \begin{bmatrix} 0 \\ \alpha \end{bmatrix} u, \quad D(x) = \begin{bmatrix} \beta \\ \gamma \end{bmatrix} x_1.$$

Then

$$R_-^2 = \left\{\begin{bmatrix} 0 \\ \lambda \end{bmatrix} : \lambda \in R^1\right\} \quad \text{and} \quad \{C^1\} \subseteq R_-^2, \quad D(R_-^2) = 0.$$

Also (A, B) is stabilizable and $R_+^2 \subseteq \{B\}$. Hence an optimal control exists for all values of α, β, γ.

On the other hand,

$$\Delta(I) = (\beta^2 + \gamma^2)\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \Gamma(I) = \alpha^2.$$

Then

$$\begin{aligned} \alpha^2 \inf_K Q(K) &\equiv \alpha^2 \inf_K \left\| \int_0^\infty e^{t(A-BK)'} K'K e^{t(A-BK)}\,dt \right\| \\ &\leqq \inf_K \left\| \int_0^\infty e^{t(A-BK)'} [\Delta(I) + \alpha^2 K'K] e^{t(A-BK)}\,dt \right\|. \end{aligned} \tag{6.1}$$

Computation shows that $K = [2, 0]$ minimizes $Q(K)$ in the class of K's such that $A - BK$ is stable. Also $Q([2, 0]) = 2$ so that Theorem 3.2 requires at least $\alpha^2 < \frac{1}{2}$.

Appendix. We shall prove the following lemma.

LEMMA. *If* $x(0) = x^n(0) = \xi$ *with* $\mathscr{E}\{\|\xi\|^2\} < \infty$, *and if* $T < \infty$, *then*

$$\lim_{n\to\infty} \mathscr{E}\{\|x^n(t) - x(t)\|^2\} = 0 \tag{A.1}$$

uniformly in t for $0 \leqq t < T$. *Here* $x(t)$ *satisfies* (2.1) *and* $x^n(t)$ *satisfies* (3.6).

Proof. We shall write $x(t)$ as x_t. Let

$$y_t = \int_0^t D(x_s)\, d\omega_s, \qquad y_t^n = \int_0^t D_n(x_s)\, d\omega_s.$$

As $\mathscr{E}\{\|\xi\|^2\} < \infty$, then $\mathscr{E}\{\max_{0 \leqq t \leqq T} \|x_t\|^2\} < \infty$ so

$$\mathscr{E}\{\|y_t - y_t^n\|^2\} = \mathscr{E}\left\{\int_0^t \|D(x_s) - D_n(x_s^n)\|^2\, ds\right\}.$$

Now

$$\|D(x_s) - D_n(x_s^n)\|^2 \leqq 2\|D(x_s) - D_n(x_s)\|^2 + 2\|D_n(x_s) - D_n(x_s^n)\|^2.$$

If we let $\mathscr{M}_t^n = [0, t] \cap \{s : \|x_s\| > n\}$, then

$$\begin{aligned}\mathscr{E}\{\|y_t - y_t^n\|^2\} &\leqq 2k^2\mathscr{E}\left\{\int_{\mathscr{M}_t^n} \|x_s\|^2\, ds\right\} + 2k^2\mathscr{E}\left\{\int_0^t \|x_s - x_s^n\|^2\, ds\right\} \\ &= 2k^2(\alpha_1 + \alpha_2)\end{aligned}$$

if we set $\alpha_1 = \mathscr{E}\{\int_{\mathscr{M}_t^n} \|x_s\|^2\, ds\}$ and $\alpha_2 = \mathscr{E}\{\int_0^t \|x_s^n - x_s\|^2\, ds\}$. The same result holds for $y_t = \int_0^t C(Kx_s)\, d\omega_s$ and $y_t^n = \int_0^t \tilde{C}_n(x_s)\, d\omega_s$. Moreover,

$$\mathscr{E}\left\{\left\|\int_0^t (A - BK)(x_s - x_s^n)\, ds\right\|^2\right\} \leqq 2tk^2\alpha_2.$$

It follows that

$$M_t^n \equiv \mathscr{E}\{\|x_t^n - x_t\|^2\} \leqq 6tk^2\alpha_2 + 12k^2(\alpha_1 + \alpha_2),$$

i.e.,

$$M_t^n \leqq 6k^2(2 + t)\int_0^t M_s^n\, ds + 12k^2\alpha_1,$$

i.e.,

$$M_t^n \leqq 12k^2\alpha_1 \exp\{6k^2(2 + t)t\} \leqq c_1(T)\alpha_1.$$

However,

$$\begin{aligned}\alpha_1 &= \mathscr{E}\left\{\int_{\mathscr{M}_t^n} \|x_s\|^2\, ds\right\} = \int_0^t \mathscr{E}\{\|x_s\|^2 1_{\|x_s\| > n}(s)\}\, ds \\ &= \int_0^t \int_{\|y\| > n} \|y\|^2 P_s(dy)\, ds,\end{aligned} \tag{A.2}$$

where $P_s(G) = \Pr\{x_s \in G\}$. As $\mathscr{E}\{\max_{0 \leqq s \leqq T} \|x_s\|^2\} < \infty$ then the inner integral of the last term in (A.2) converges to zero uniformly in s. This establishes (A.1).

Acknowledgment. I should like to thank Dr. W. M. Wonham for his suggestions during the course of this work.

REFERENCES

[1] W. M. WONHAM, *Optimal stationary control of a linear system with state-dependent noise*, this Journal, 5 (1967), pp. 486–500.

[2] ———, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. II, A. T. Bharucha-Reid, ed., Academic Press, New York, 1969.

[3] M. ZAKAI, *A Liapunov criterion for existence of stationary probability distributions for systems with noise*, this Journal, 7 (1969), pp. 390–397.

[4] E. B. DYNKIN, *Markov Processes*, vol. I, Academic Press, New York, 1965.

[5] W. M. WONHAM, *Liapunov criteria for weak stochastic stability*, J. Differential Equations, 2 (1966), pp. 195–207.

[6] ———, *A Liapunov method for the estimation of statistical averages*, Ibid., 2 (1966), pp. 365–377.

[7] N. N. KRASOVSKII, *Stabilization of systems in which noise is dependent on the value of the control signal*, Engrg. Cybernetics, 2 (1965), pp. 94–102.

[8] W. M. WONHAM, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 660–665.

[9] F. KOZIN, *On almost sure sample asymptotic properties of diffusion processes defined by stochastic differential equations*, J. Math. Kyoto Univ., 4 (1965), pp. 515–528.

[10] M. B. NEVEL'SON AND R. Z. KHAS'MINSKII, *Stability of a linear system with random disturbances of its parameters*, J. Appl. Math. Mech., 30 (1966), pp. 487–493.

SIAM J. CONTROL
Vol. 9, No. 2, May 1971

INPUT–OUTPUT STRUCTURE OF LINEAR SYSTEMS WITH APPLICATION TO THE DECOUPLING PROBLEM*

L. M. SILVERMAN AND H. J. PAYNE†

1. Introduction. Let $\mathcal{S} = (A, B, C, D)$ be a linear time-invariant system representation described by the equations

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{1.1}$$

$$y(t) = Cx(t) + Du(t), \tag{1.2}$$

where $x(t) \in \Sigma = R^n$, $u(t) \in R^r$, $y(t) \in R^m$ and the constant matrices A, B, C and D are $n \times n$, $n \times r$, $m \times n$ and $m \times r$, respectively. For each initial state $x_0 \in \Sigma$, $\mathcal{S}$ defines a mapping $\mathcal{S}_{x_0}: \mathcal{U} \to \mathcal{Y}$, where $\mathcal{U}$ denotes the input function space over $[0, \infty)$ and $\mathcal{Y}$ the corresponding output space. For simplicity, $\mathcal{U}$ will be taken to be $\mathcal{L}_r^0$, the space of r-vector functions continuous on $[0, \infty)$.

In the first half of this paper (§§ 2–5) two basic problems related to the input–output structure of the system $\mathcal{S}$ are examined. The first is that of explicitly characterizing the functional range of $\mathcal{S}$, while the second is that of describing the set of inputs which generate given elements of the range. Of particular interest for the second problem is the set of inputs which will generate the zero output. These problems have been considered by several authors in the single-input, single-output case [1]–[4] with a complete solution being given in [4]. In the multivariable case attention, for the most part, has centered on situations in which $r = m$ and $\mathcal{S}$ is one-to-one [5]–[8], with a complete solution being given in [8]. However, for many applications, particularly the decoupling problem, this restriction is quite unrealistic. Hence we consider the general problem here. The basis of our approach to the problem is the structure algorithm developed in [8]. In § 2 this algorithm is described and generalized to the case $m \neq r$, and in § 3 several basic properties of the algorithm pertinent to the problems considered are derived. A complete solution to the range and input characterization problems is given in § 4. The main results of this section are Theorems 4.1 and 4.2. Theorem 4.1 gives the necessary and sufficient conditions that a function must satisfy in order that it be in the range of $\mathcal{S}_{x_0}$ while Theorem 4.2 specifies an "inverse system" representation which can be utilized to generate the set of all inputs u such that $w = \mathcal{S}_{x_0}[u]$ for any w in the range of $\mathcal{S}_{x_0}$. Corollaries of these theorems include necessary and sufficient conditions for $\mathcal{S}_{x_0}$ to be left or right invertible and a measure of the rank of the transfer function matrix of $\mathcal{S}$ in terms of quantities defined by the structure algorithm. In § 5 the results are specialized to the problem of generating the zero output. Two equivalent characterizations of the set of "zeroing" inputs are given, the first being an open loop specification as the set of outputs of a fixed dynamic system, and the second, a feedback law of the type

* Received by the editors April 2, 1970.

† Department of Electrical Engineering, University of Southern California, Los Angeles, California 90007. This work was supported by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under Grant AFOSR-69-1622A.

$u = Fx + Gv$. A complete characterization of the class of all pairs (F, G) which can be used for this purpose is also given. Application is also made in this section to the problem of input isolation discussed by Wonham and Morse [9]. It is shown that the structure algorithm very simply classifies the set of inputs which can be isolated from the output by state feedback.

In the second part of the paper (§§ 6–9) application of the structure theory developed in the first part is made to the problem of decoupling $\mathscr{S}$ with respect to a specified partition of the outputs into p subsets either by static feedback (of the type $u = Fx + Gv$ for F and G constant matrices) or dynamic compensation. The static decoupling problem for the case $p = m = r$ and $D = 0$ was first defined by Morgan [10], [11], with a complete solution given by Falb and Wolovich [12], [13], [14] and Gilbert [15]. Partial results were also obtained by Rekasius [16]. The more general static decoupling problem (with $D = 0$) was first defined by Wonham and Morse [9] who also provided solutions for several special cases. The approach of Wonham and Morse differs considerably from that of previous work on the problem with their emphasis being on geometric characterizations. As indicated in their introduction, although they solve two large classes of static decoupling problems their results do not lend themselves directly to computer implementation. The approach taken here is a completely general but algebraic one and is well-suited to computer implementations. In this respect it is similar to Gilbert's [15] procedure when specialized to the case he considers.

The case $D \neq 0$ is handled with no additional complication and leads to results more general than those of [12] and [15] even for the case $p = m = r$.[1] Moreover, for the case $D = 0$, a broader class of decoupling problems is solved than previously reported [9]. An outline of §§ 6–8 follows.

The general static decoupling problem is defined in § 6. The definition given is essentially the same as that given in [9] for the case $D = 0$. In § 7 an explicit algebraic characterization of the class of all feedback pairs (F, G) which decouple $\mathscr{S}$ preserving output controllability is given in terms of matrices defined by the structure algorithm. This characterization consists of a set of (nonlinear) algebraic equations, the set of whose solutions coincides with the set of all decoupling pairs, together with a matrix rank condition which insures output controllability. As yet, a finite algorithm for determining existence of a solution to these equations has not been found. For many important special cases, however, solutions have been found and a number of these are given in § 8. Included in the results of this section are the cases solved (by quite different methods) by Wonham and Morse [9]. A complete solution to the static decoupling problem when G is constrained to be nonsingular is also given. This case includes the now standard result of Falb and Wolovich [12] and Gilbert [15]. Emphasis throughout this section is on constructive proofs which can be directly implemented on a digital computer.

The dynamic decoupling problem is treated in § 9. This problem was first considered in full generality (for $D = 0$) by Morse and Wonham [18], and a complete solution was given for the case in which state feedback is allowed. A simple constructive solution to this problem is also given here. Moreover, the method used here leads quite directly to a complete solution of the more practical

[1] Recently, Morse [17] extended some, but not all, of the results of [9] for the case $D \neq 0$.

problem—that of decoupling with dynamic compensation and *output* feedback alone, while preserving output controllability and stability of the closed loop system (see Theorem 9.3). This result completely solves the outstanding "classical" problem of decoupling a system specified by a transfer function matrix which originally motivated Morgan's state variable approach to the problem (see [11] for references to the early literature on the problem).

2. The structure algorithm. Let $q_0 = \operatorname{rank} D$ and let $\bar{D}_0$ be the submatrix formed from the first q_0 independent rows of D. Then there exists an $m \times m$ nonsingular matrix S_0 such that

$$S_0 D = \begin{bmatrix} \bar{D}_0 \\ 0 \end{bmatrix}.$$

Using S_0 as an output transformation yields a new system representation $\mathscr{T}_0$ defined by the equations

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{2.1}$$

$$y_0(t) = C_0 x(t) + D_0 u(t), \tag{2.2}$$

where $y_0(t) = S_0 y(t)$, $C_0 = S_0 C$ and $D_0 = S_0 D$. It will prove convenient to partition y_0 and C_0 comformably with D_0 as

$$y_0 = \begin{bmatrix} \bar{y}_0 \\ \tilde{y}_0 \end{bmatrix}, \qquad C_0 = \begin{bmatrix} \bar{C}_0 \\ \tilde{C}_0 \end{bmatrix}$$

so that $\bar{y}_0 = \bar{C}_0 x + \bar{D}_0 u$ and $\tilde{y}_0 = \tilde{C}_0 x$. As for the case $r = m$ (see [8]), the remainder of the sequence $\mathscr{T}_i$ can be defined inductively. We assume that $\mathscr{T}_k$ has the form

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{2.3}$$

$$y_k(t) = C_k x(t) + D_k u(t), \tag{2.4}$$

with the partitioning

$$y_k(t) = \begin{bmatrix} \bar{y}_k(t) \\ \tilde{y}_k(t) \end{bmatrix}, \qquad C_k = \begin{bmatrix} \bar{C}_k \\ \tilde{C}_k \end{bmatrix}, \qquad D_k = \begin{bmatrix} \bar{D}_k \\ 0 \end{bmatrix},$$

where $\bar{D}_k$ has q_k rows and rank q_k, $\bar{y}_k$ and $\bar{C}_k$ have q_k rows, and $\tilde{y}_k$ and $\tilde{C}_k$ have $m - q_k$ rows. Observe that if

$$M_k = \begin{bmatrix} I_{q_k} & 0 \\ 0 & I_{m-q_k}(d/dt) \end{bmatrix},$$

then

$$M_k y_k(t) = \begin{bmatrix} \bar{C}_k \\ \tilde{C}_k A \end{bmatrix} x(t) + \begin{bmatrix} \bar{D}_k \\ \tilde{C}_k B \end{bmatrix} u(t).$$

Note that $\tilde{y}_k = \tilde{C}_k x(t)$, hence $\tilde{y}_k$ is continuously differentiable so that $M_k y_k(t)$ always exists. In the expression for $M_k y_k(t)$, let H_{k+1} and J_{k+1} denote the matrices multiplying $x(t)$ and $u(t)$, respectively, and let $q_{k+1} = \operatorname{rank} J_{k+1}$. If $\bar{D}_{k+1}$ is the

matrix formed from the first q_{k+1} independent rows of J_{k+1}, then there exists an $m \times m$ nonsingular matrix S_{k+1} such that

$$S_{k+1}J_{k+1} = \begin{bmatrix} \bar{D}_{k+1} \\ 0 \end{bmatrix}.$$

$\mathscr{T}_{k+1}$ is then defined by the equations

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{2.5}$$

$$y_{k+1}(t) = C_{k+1}x(t) + D_{k+1}u(t), \tag{2.6}$$

where $y_{k+1} = S_{k+1}M_k y_k(t)$, $C_{k+1} = S_{k+1}H_{k+1}$ and $D_{k+1} = S_{k+1}J_{k+1}$. It follows from the above that $y_k(t) = N_k y(t)$, where[2]

$$N_k = \prod_{i=0}^{k} S_{k-i}M_{k-i-1}, \qquad k = 0, 1, \cdots \quad (M_{-1} \equiv I), \tag{2.7}$$

is a sequence of nonsingular matrix differential operators. Furthermore, $\bar{N}_k$ and $\tilde{N}_k$ are defined by $\bar{y}_k = \bar{N}_k y(t)$ and $\tilde{y}_k = \tilde{N}_k y(t)$.

It is clear that the matrices S_i defined above are not unique in general. The following method of constructing the matrices proves to be most convenient and will be utilized in the remainder of the paper.

Let S^*_{k+1} be the (unique) permutation matrix such that $S^*_{k+1}J_{k+1}$ has the following structure. Its first q_{k+1} rows are the first q_{k+1} independent rows of J_{k+1} with the relative order maintained, and its last $m - q_{k+1}$ rows are the remaining rows of J_{k+1}, also with the relative order maintained. It is clear from the form of J_{k+1} that S^*_{k+1} has the form

$$S^*_{k+1} = \begin{bmatrix} I_{q_k} & 0 \\ 0 & \bar{R}_{k+1} \\ 0 & \tilde{R}_{k+1} \end{bmatrix}, \tag{2.8}$$

where $\bar{R}_{k+1}$ has $q_{k+1} - q_k$ rows and $\tilde{R}_{k+1}$ has $m - q_{k+1}$ rows. Furthermore,

$$R_{k+1} = \begin{bmatrix} \bar{R}_{k+1} \\ \tilde{R}_{k+1} \end{bmatrix}$$

is a permutation matrix. Setting

$$\bar{D}_{k+1} = \begin{bmatrix} \bar{D}_k \\ \bar{R}_{k+1}\tilde{C}_k B \end{bmatrix} \tag{2.9}$$

and

$$K^*_{k+1} = \tilde{R}_{k+1}\tilde{C}_k B\bar{D}^{\dagger}_{k+1}, \tag{2.10}$$

where

$$\bar{D}^{\dagger}_{k+1} = \bar{D}'_{k+1}(\bar{D}_{k+1}\bar{D}'_{k+1})^{-1},$$

[2] $\prod_{i=j}^{k} A_i \equiv A_k A_{k-1} \cdots A_j$. The empty product is taken to be the identity.

we have that S_{k+1} can be chosen as

$$S_{k+1} = \begin{bmatrix} I_{q_{k+1}} & 0 \\ -K^*_{k+1} & I_{m-q_{k+1}} \end{bmatrix} S^*_{k+1} \tag{2.11}$$

so that

$$\bar{D}_{k+1} = \begin{bmatrix} \bar{D}_k \\ \bar{R}_{k+1}\tilde{C}_k B \end{bmatrix}, \qquad \bar{C}_{k+1} = \begin{bmatrix} \bar{C}_k \\ \bar{R}_{k+1}\tilde{C}_k A \end{bmatrix} \tag{2.12}$$

and

$$\tilde{C}_{k+1} = \tilde{R}_{k+1}\tilde{C}_k A - K^*_{k+1}\bar{C}_{k+1}. \tag{2.13}$$

This unique explicit representation of the algorithm will be assumed without further comment in the remainder of the paper.

Note that if a is the first integer such that $q_a = q_n$ and

$$K_{k+1} = \tilde{R}_{k+1}\tilde{C}_k B\bar{D}_a^\dagger, \qquad 0 \leqq k \leqq n-1, \tag{2.14}$$

then (2.13) can be expressed as

$$\tilde{C}_{k+1} = \tilde{R}_{k+1}\tilde{C}_k A - K_{k+1}\bar{C}_a, \qquad 0 \leqq k \leqq n-1, \tag{2.15}$$

or

$$\tilde{C}_{k+1} = \tilde{R}_{k+1}\tilde{C}_k(A - B\bar{D}_a^\dagger \bar{C}_a), \qquad 0 \leqq k \leqq n-1. \tag{2.16}$$

This follows from the obvious nesting properties of the sequences $\bar{C}_k$ and $\bar{D}_k$, $k = 0, 1, \cdots, n$ (see [8]).

3. Properties of the structure algorithm. We shall first consider the effect of state feedback on the structure algorithm. Let u be a feedback law of the form

$$u = Fx + v. \tag{3.1}$$

Then the closed loop system is represented by the quadruple $(A + BF, B, C + DF, D)$. If P is any operator associated with $\mathscr{S}$ and Q is the corresponding operator of $(A + BF, B, C + DF, D)$, we shall say that $P \xrightarrow{F} Q$. For example, it is clear that $A \xrightarrow{F} A + BF$, $C \xrightarrow{F} C + DF$ while B and D are invariant under this type of feedback.

LEMMA 3.1. *The operators N_i and the matrices D_i and $\tilde{C}_i$ are invariant under feedback of the type* (3.1), *and*

$$\bar{C}_i \xrightarrow{F} \bar{C}_i + \bar{D}_i F$$

for $i = 0, 1, \cdots$.

Proof. The proof is by induction on the steps of the algorithm. For $i = 0$, the result is obvious. Assuming that the result holds for $i = k$ then we clearly have that

$$H_{k+1} \xrightarrow{F} \begin{bmatrix} \bar{C}_k + \bar{D}_k F \\ \tilde{C}_k A + \tilde{C}_k BF \end{bmatrix} = H_{k+1} + J_{k+1}F$$

and

$$J_{k+1} \xrightarrow{F} \begin{bmatrix} \bar{D}_k \\ \tilde{C}_k B \end{bmatrix} = J_{k+1}.$$

Hence, $S_{k+1} \xrightarrow{F} S_{k+1}$ which in turn implies that $N_{k+1} \xrightarrow{F} N_{k+1}$, $D_{k+1} \xrightarrow{F} D_{k+1}$ and

$$C_{k+1} \xrightarrow{F} S_{k+1}(H_{k+1} + J_{k+1}F) = C_{k+1} + D_{k+1}F,$$

which completes the proof.

Remark. Lemma 3.1 yields as a special case the F-invariants observed by Gilbert [15]. His results obtain by setting $D = 0$ and $m = 1$.

Let the matrix L_k be defined as

$$L_k' = [\tilde{C}_0' : \tilde{C}_1' : \cdots : \tilde{C}_{k-1}']. \tag{3.2}$$

The matrices L_k have many properties in common with observability matrices. In fact, L_k identifies a subspace onto which the projection of the state of $\mathscr{S}$ is determined by the output alone. This follows from the relationship

$$\tilde{Y}_k(t) = L_k x(t), \tag{3.3}$$

where

$$\tilde{Y}_k'(t) = [\tilde{y}_0' \vdots \tilde{y}_1' \vdots \cdots \vdots \tilde{y}_{k-1}']. \tag{3.4}$$

LEMMA 3.2. *If* rank L_k = rank L_{k+1}, *then* rank L_i = rank L_k *for all* $i > k$.

Proof. Let a be the first integer such that $q_a = q_n$. Then by construction, $\bar{D}_i = \bar{D}_a$ and $\bar{C}_i = \bar{C}_a$ for all $i \geqq a$ and $\bar{D}_i$ and $\bar{C}_i$ are leading submatrices of $\bar{D}_a$ and $\bar{C}_a$, respectively, for $i < a$. Hence, by setting $F = -\bar{D}_a^{\dagger}\bar{C}_a$, where $\bar{D}_a^{\dagger} = \bar{D}_a'(\bar{D}_a\bar{D}_a')^{-1}$, it follows from Lemma 3.1 that $\bar{C}_i \xrightarrow{F} 0$ for all i. Also by Lemma 3.1, $L_i \xrightarrow{F} L_i$ so that there is no loss of generality for purposes of the proof in assuming that $\bar{C}_i = 0$ for all i. It follows from (2.12) and (2.13) that with $\bar{C}_k = 0$,

$$\tilde{C}_k A = R_{k+1}^{-1} \begin{bmatrix} 0 \\ \tilde{C}_{k+1} \end{bmatrix}. \tag{3.5}$$

Suppose now that rank L_k = rank L_{k+1}. This implies that there exist matrices P_i such that

$$\tilde{C}_k = \sum_{i=0}^{k-1} P_i \tilde{C}_i. \tag{3.6}$$

By (2.13) with $\bar{C}_k = 0$ we have

$$\tilde{C}_{k+1} = \tilde{R}_{k+1}\tilde{C}_k A = \sum_{i=0}^{k-1} \tilde{R}_{k+1} P_i \tilde{C}_i A. \tag{3.7}$$

Therefore, by (3.5),

$$\tilde{C}_{k+1} = \sum_{i=0}^{k-1} P_i^1 \tilde{C}_{i+1} = \sum_{i=1}^{k} P_{i-1}^1 \tilde{C}_i, \tag{3.8}$$

where P_i^1 is an appropriately defined matrix. Substituting (3.6) into (3.8), therefore,

determines matrices P_i^2 such that

$$\tilde{C}_{k+1} = \sum_{i=0}^{k-1} P_i^2 \tilde{C}_i$$

so that rank L_{k+2} = rank L_k. The result for $i > k+2$ follows by an obvious induction argument.

THEOREM 3.1. *Let a be the first integer such that $q_a = q_n$, let β be the first integer such that* rank L_β = rank $L_{\beta+1}$ *and let* γ = rank $\tilde{C}_0$. *Then*

$$a \leqq \beta \leqq n - \gamma + 1. \tag{3.9}$$

Proof. The upper bound on β follows from the observation that L_1 has rank γ and the fact that

$$\text{rank } L_i - \text{rank } L_{i-1} \geqq 1 \quad \text{for } i \leqq \beta$$

by Lemma 3.2.

The lower bound on β is clear since rank D_i = rank D_β for $i > \beta$, again by Lemma 3.2.

Remark. Theorem 3.1 provides a stopping rule for the algorithm.

LEMMA 3.3. *There exist an integer δ, $\beta \leqq \delta \leqq n$, and matrices P_i, $i = 0, \cdots, \delta - 1$, such that*

$$\tilde{C}_\delta = \sum_{i=0}^{\delta-1} P_i \left(\prod_{j=i+1}^{\delta} \tilde{R}_j \right) \tilde{C}_i. \tag{3.10}$$

Proof. As in the proof of Lemma 3.2, it may be assumed without loss of generality that $\bar{C}_i = 0$ for all i. With this assumption $\tilde{C}_{k+1} = \tilde{R}_{k+1}\tilde{C}_k A$ so that

$$\tilde{C}_k = \left(\prod_{i=1}^{k} \tilde{R}_i \right) \tilde{C}_0 A^k \tag{3.11}$$

and

$$\left(\prod_{i=k+1}^{\delta} \tilde{R}_i \right) \tilde{C}_k = \left(\prod_{i=1}^{\delta} \tilde{R}_i \right) \tilde{C}_0 A^k, \tag{3.12}$$

where δ is as yet undetermined. Let $W_\delta = (\prod_{i=1}^{\delta} \tilde{R}_i)\tilde{C}_0$. Then for a fixed δ, it follows by standard observability matrix arguments (of which the proof of Lemma 3.2 is a generalization) that there exists an integer $\gamma \leqq n$ such that $(W_\delta A^\gamma)'$ is in the range of

$$[W_\delta' \vdots \cdots \vdots (W_\delta A^{\gamma-1})'].$$

Hence, there exists a first integer $\delta \leqq n$ for which $\delta = \gamma$ so that

$$W_\delta A^\delta = \sum_{i=0}^{\delta-1} P_i W_\delta A^i$$

for appropriate matrices P_i. Consequently, since by (3.11), $\tilde{C}_\delta = W_\delta A^\delta$, equation (3.10) follows.

The lower bound on δ derives from the observation that (3.10) implies that rank $L_{\delta+1}$ = rank L_δ.

Remark. The integer δ is not as simply defined as a and β. However, only the form of (3.10) is of importance below and δ need not be calculated for the main results.

For the case $q_a = m$ the results of this section can be considerably strengthened.

THEOREM 3.2. *If* $q_a = m$, *then*

(i) *the rows of* L_a *are linearly independent, and*

(ii) $a = \beta = \delta$.

Proof. The proof of (i) is essentially the same as that of Theorem 3 in [8] and need not be repeated here. That $\beta = a$ follows from (i) and the fact that if $q_a = m$, then $L_j = L_a$ for $j \geqq a$. That $\delta = a$ follows from the fact that $\tilde{C}_a = 0$ so that (3.10) is satisfied trivially with $a = \delta$ and $P_i = 0$, $i = 0, \cdots, a - 1$.

A final property of the structure algorithm that will be used in the sequel is the following lemma.

LEMMA 3.4. *Let* $\lambda \in R^r$. *Then* $\bar{D}_a\lambda = 0$ *if and only if* $\begin{bmatrix} D \\ L_\beta B \end{bmatrix}\lambda = 0$.

Proof. From the structure algorithm,

$$\bar{R}_{k+1}\tilde{C}_k B = [0:I:0]\bar{D}_a$$

and

$$\tilde{R}_{k+1}\tilde{C}_k B = K_{k+1}\bar{D}_a$$

so that

$$\tilde{C}_k B = R_{k+1}^{-1}\begin{bmatrix} 0:I:0 \\ K_{k+1} \end{bmatrix}\bar{D}_a.$$

Hence $D_a\lambda = 0$ implies $L_\beta B\lambda = 0$. By the nesting property of $\bar{D}_k$ and the definition of D_0, one also has $D\lambda = 0$.

For sufficiency, if $\bar{D}_0\lambda = 0$ and $\tilde{C}_0 B\lambda = 0$, then it is clear that $\bar{D}_1\lambda = 0$. Proceeding by induction, suppose $\bar{D}_k\lambda = 0$. Then if $\tilde{C}_k B\lambda = 0$, it follows that

$$\bar{D}_{k+1}\lambda = \begin{bmatrix} \bar{D}_k \\ \bar{R}_{k+1}\tilde{C}_k B \end{bmatrix}\lambda = 0.$$

This completes the proof.

4. System range. In this section it will be shown that the structure algorithm completely determines the functional range of $\mathscr{S}$ and provides a method for generating elements of the range. Two operators derived from the structure algorithm will be utilized for this purpose:

$$M_1 = \left(\frac{d^\delta}{dt^\delta} - \sum_{i=0}^{\delta-1} P_i \frac{d^i}{dt^i}\right)\prod_{j=0}^{a}\tilde{R}_j, \tag{4.1}$$

$$M_2 = \sum_{j=0}^{\delta}\left(\prod_{l=j+1}^{a}\tilde{R}_l\right)K_j\frac{d^{\delta-j}}{dt^{\delta-j}} - \sum_{j=0}^{\delta-1}P_j\sum_{k=0}^{j}\left(\prod_{l=k+1}^{a}\tilde{R}_l\right)K_k\frac{d^{j-k}}{dt^{j-k}}. \tag{4.2}$$

Here and below, δ and P_i, $i = 0, \cdots, \delta - 1$, are the quantities defined in Lemma 3.3; a and β are as defined in Theorem 3.1. The matrices defined in § 2 will be utilized without further comment.

Let $\mathscr{Y}_k$ be the space of m-vector functions for which $N_k y$ is defined and continuous. Then we have the following preliminary result.

LEMMA 4.1. *Let* $y \in \mathscr{Y}_\delta$ *and let* $y_j = N_j y$, $j = 0, \cdots, \delta - 1$. *If* $M_1 y(t) \equiv 0$ *on* $[0, \infty)$, $\bar{y}_a(t) \equiv 0$ *on* $[0, \infty)$ *and* $\tilde{y}_j(0) = 0, j = 0, \cdots, \beta - 1$, *then* $y(t) \equiv 0$ *on* $[0, \infty)$.

Proof. First we show that $\bar{y}_a(t) \equiv 0$ on the interval implies $\tilde{y}_j(t) = (\prod_{i=0}^{j} \tilde{R}_i) y^{(j)}(t)$. To see this, we proceed by induction. By the algorithm,

$$\bar{y}_0(t) = \bar{R}_0 y(t)$$

and

$$\tilde{y}_0(t) = \tilde{R}_0 y(t) - K_0^* \bar{y}_0(t).$$

But $\bar{y}_a(t) \equiv 0$ implies that $\bar{y}_j(t) \equiv 0, j = 0, 1, \cdots, a$. Therefore, $\tilde{y}_0(t) = \tilde{R}_0 y(t)$. Let the desired relation be true for $j = k - 1$. Then by the algorithm,

$$\bar{y}_k(t) = \begin{bmatrix} \bar{y}_{k-1}(t) \\ \bar{R}_k \tilde{y}_{k-1}^{(1)}(t) \end{bmatrix}$$

and

$$\tilde{y}_k(t) = \tilde{R}_k \tilde{y}_{k-1}^{(1)}(t) - K_k^* \bar{y}_k(t).$$

Since $\bar{y}_k(t) \equiv 0$, $\tilde{y}_k(t) = \tilde{R}_k \tilde{y}_{k-1}^{(1)}(t)$, which establishes the suggested relation.

From the differential equation $M_1 y(t) = 0$, the initial conditions $\tilde{y}_j(0) = 0$, $j = 0, 1, \cdots, \beta - 1$, and the result established above, it follows that $(\prod_{j=0}^{a} \tilde{R}_j) y(t) \equiv 0$, hence $\tilde{y}_a(t) \equiv 0$. We proceed by induction to show that $\tilde{y}_k(t) \equiv 0$ for $k < a$. Suppose $\tilde{y}_k(t) \equiv 0$. Then $\bar{y}_k(t) = 0$ implies $\tilde{R}_k \tilde{y}_{k-1}^{(1)}(t) \equiv 0$, and $\bar{R}_k \tilde{y}_{k-1}^{(1)}(t) \equiv 0$. Integrating and employing the initial condition, we have $\tilde{R}_k \tilde{y}_{k-1}(t) \equiv 0$ and $\bar{R}_k \tilde{y}_{k-1}(t) \equiv 0$. Then, since $[\bar{R}_k' \vdots \tilde{R}_k']$ is nonsingular, $\tilde{y}_{k-1}(t) \equiv 0$. By induction, $\tilde{y}_0(t) \equiv 0$. Finally, since $[\bar{R}_0' \vdots \tilde{R}_0]' y(t) = [\bar{y}_0' \vdots \tilde{y}_0]'$, $y(t) \equiv 0$.

The following theorem gives an explicit characterization of the range of a linear system under continuous inputs. A partial result of this nature was first given in [8, Theorem 5], where the history of the problem is also detailed.

THEOREM 4.1 (Range theorem). *An* m*-vector function* w *defined on* $[0, \infty)$ *is in the range of* $\mathscr{S}_{x_0}$ *if and only if*

(i) $w \in \mathscr{Y}_\delta$,

(ii) $\tilde{N}_i w(t)|_{t=0} = \tilde{C}_i x_0$, $i = 0, \cdots, \beta - 1$, *and*

(iii) $(M_1 - M_2 \bar{N}_a) w(t) = 0$ *for all* $t \in [0, \infty)$.

Proof. For necessity, observe that if w is the output of $\mathscr{S}$ for an initial state x_0 and continuous input v then

$$\bar{N}_k w(t) = \bar{C}_k x(t) + \bar{D}_k v(t), \tag{4.3}$$

$$\tilde{N}_k w(t) = \tilde{C}_k x(t). \tag{4.4}$$

Conditions (i) and (ii) follow immediately from (4.3) and (4.4). In fact, it is clear that the two conditions are satisfied for all $i = 0, 1, \cdots$. Setting $\bar{w}_k(t) = \bar{N}_k w(t)$ and $\tilde{w}_k(t) = \tilde{N}_k w(t)$, it follows from the structure algorithm that

$$\tilde{w}_k(t) = \tilde{R}_k \tilde{w}_{k-1}^{(1)}(t) - K_k \bar{w}_a(t). \tag{4.5}$$

Then, iterating, one finds

$$\tilde{w}_k(t) = \left(\sum_{j=0}^{k} \tilde{R}_j\right) w^{(k)}(t) - \sum_{j=0}^{k} \left(\prod_{l=j+1}^{k} \tilde{R}_l\right) K_j \bar{w}_a^{(k-j)}(t). \tag{4.6}$$

Employing Lemma 3.3, and noting that $\tilde{R}_j = I$ for $j > a$, we clearly see that

$$\tilde{w}_\delta(t) = \sum_{i=0}^{\delta-1} P_i \left(\prod_{j=i+1}^{a} \tilde{R}_j\right) \tilde{w}_i(t). \tag{4.7}$$

Substituting for $\tilde{w}_\delta(t)$ and $\tilde{w}_i(t)$ from (4.6), one finds that w must satisfy condition (iii).

For sufficiency, assume that w is such that conditions (i)–(iii) hold. Let K be a matrix whose columns form a basis for the null space of $\bar{D}_a$. Then, since $\bar{D}_a$ has full row rank,

$$T = [\bar{D}_a^\dagger \vdots K] \tag{4.8}$$

is nonsingular. Now let

$$u(t) = \bar{D}_a^\dagger v_1(t) + K v_2(t)$$

and set

$$v_1(t) = \bar{w}_a(t) - \bar{C}_a z(t), \tag{4.9}$$

where z is the solution of

$$\dot{z}(t) = (A - B\bar{D}_a^\dagger \bar{C}_a) z(t) + B\bar{D}_a^\dagger \bar{w}_a(t) + BKv_2(t) \tag{4.10}$$

with initial condition $z(0) = x_0$. It is clear that, for any v_2, the output of $\mathscr{S}_{x_0}$ with this input has the property that $\bar{y}_a(t) \equiv \bar{w}_a(t)$ and y satisfies properties (i)–(iii). Defining $r(t) = y(t) - w(t)$ and $r_j(t) = N_j r(t)$ we then have that $r(t)$ satisfies (iii), and $\bar{r}_a(t) \equiv 0$. Hence, $M_1 r(t) \equiv 0$ and since $\tilde{r}_j(0) = 0, j = 0, \cdots, \beta - 1$, Lemma 4.1 implies $r(t) \equiv 0$ which completes the proof.

Remark. In the case $q_a = m$, it is easily seen that $\tilde{R}_a = K_a = 0$ so that $M_1 = M_2 = 0$. Thus only conditions (i) and (ii) need be satisfied to insure that w is in the range of $\mathscr{S}_{x_0}$. For the subcase $q_a = m = r$, therefore, the range theorem specializes to Theorem 5 of [8]. In general, condition (iii) can be expressed in the form

$$M_1 w(t) = M_2 \bar{w}_a(t).$$

This equation shows that $w(t)$ in the range of $\mathscr{S}_{x_0}$ is uniquely determined by $\bar{w}_a(t)$ and the initial conditions.

COROLLARY 4.1. *The range of $\mathscr{S}$ is invariant under feedback of the type* (3.1).

Proof. By Lemma 3.1, $\mathscr{Y}_i$, $\tilde{N}_i$, $\tilde{C}_i$ and $(M_1 - M_2\bar{N}_a)$ are all invariant under this type of feedback. Hence, the result is immediate from Theorem 4.1.

The following theorem gives a complete characterization of the class of inputs which will generate a particular function w in the range of $\mathscr{S}_{x_0}$.

THEOREM 4.2. *Let w be in the range of $\mathscr{S}_{x_0}$. Then $\mathscr{S}_{x_0}[u] = w$ if and only if u is the output of the system*

$$\dot{z}(t) = (A - B\bar{D}_a^\dagger \bar{C}_a) z(t) + B\bar{D}_a^\dagger \bar{N}_a w(t) + BKv(t), \tag{4.11}$$

$$u(t) = -\bar{D}_a^\dagger \bar{C}_a z(t) + \bar{D}_a^\dagger \bar{N}_a w(t) + Kv(t) \tag{4.12}$$

with $z(0) = x_0$ for some v.

Proof. Sufficiency was established in the proof of Theorem 4.1. To establish necessity, let $u(t) = \bar{D}_a^{\dagger} v_1(t) + K v_2(t)$ be any input such that $\mathcal{S}_{x_0}[u] = w$. Then from the structure algorithm we have

$$\dot{x}(t) = Ax(t) + B\bar{D}_a^{\dagger} v_1(t) + BKv_2(t)$$

and

$$\bar{N}_a w(t) = \bar{C}_a x(t) + v_1(t)$$

with $x(0) = x_0$. Solving the second of these equations for v_1 and substituting into the first yields the pair of equations (4.11), (4.12) with $z = x$ and $v = v_2$.

Remark. The system representation (4.11)–(4.12) is a generalization of the inverse system representation discussed in [8] for the case $r = m = q_a$. For this case observe that $K = 0$ so that a unique u exists satisfying $\mathcal{S}_{x_0}[u] = w$ for a given x_0 and w. This is true more generally when a *left inverse* for $\mathcal{S}_{x_0}$ exists. The conditions for left invertibility are given in the following corollary.

COROLLARY 4.2. *There exists a left inverse* $\mathcal{S}_{x_0}^{L} : \mathcal{Y} \to \mathcal{L}_r^0$ *such that* $\mathcal{S}_{x_0}^{L} \mathcal{S}_{x_0} = I$, *the identity operator, if and only if* $q_a = r$.

Proof. When $q_a = r$, $K = 0$ so that from (4.11)–(4.12) it is clear that the system

$$\dot{z}(t) = (A - B\bar{D}_a^{\dagger}\bar{C}_a)z(t) + B\bar{D}_a^{\dagger}\bar{N}_a w(t), \tag{4.13}$$

$$u(t) = -\bar{D}_a^{\dagger}\bar{C}_a z(t) + \bar{D}_a^{\dagger}\bar{N}_a w(t) \tag{4.14}$$

with $z(0) = x_0$ is a left inverse for $\mathcal{S}_{x_0}$.

Let $\mathcal{Y}_{a,x_0} = \{w \in \mathcal{Y}_a : \tilde{N}_i w(t)|_{t=0} = \tilde{C}_i x_0,\ i = 0, \cdots, \beta - 1\}$. Then from the remarks following Theorem 4.1, $\mathcal{Y}_{a,x_0}$ is precisely the range of $\mathcal{S}_{x_0}$ when $q_a = m$. Moreover, it is clear by Theorem 4.2 that the system representation (4.13)–(4.14) is a right inverse for $\mathcal{S}_{x_0}$ (there may be others). Hence, we have the following corollary.

COROLLARY 4.3. *There exists a right inverse* $\mathcal{S}_{x_0}^{R} : \mathcal{Y}_{a,x_0} \to \mathcal{L}_r^0$ *such that* $\mathcal{S}_{x_0} \mathcal{S}_{x_0}^{R} = I$ *if and only if* $q_a = m$.

Remark. Both the left and right inverse of $\mathcal{S}_{x_0}$, when they exist, may be represented in precisely the same way—as a dynamical system, $(A - B\bar{D}_a^{\dagger}\bar{C}_a,\ B\bar{D}_a^{\dagger},\ -\bar{D}_a^{\dagger}\bar{C}_a, \bar{D}_a^{\dagger})$ following a bank of differentiators realizing N_a. When $q_a = m = r$ this is seen to be the inverse system representation defined in [8].

It should also be noted that in general a representation of the "inverse system" of lower dynamic order than that of (4.11)–(4.12) can be found. Using the procedure outlined in the proof of Theorem 4 in [8] we can show that there exists a representation equivalent to (4.11)–(4.12) whose dynamic portion has dimension $n - \operatorname{rank} L_\beta$.

Let $H(s)$ be the transfer function matrix of (A, B, C, D). Then another generalization of the results of [8] is the following theorem.

THEOREM 4.3. $\operatorname{rank} H(s) = q_a$ *for almost all* s.

Proof. Let $\mathcal{L}$ denote the Laplace transform operator. Then it is clear that for zero initial conditions,

$$\mathcal{L}[N_a y(t)] = [C_a(sI - A)^{-1}B + D_a]\hat{u}(s) = \hat{N}_a(s)\hat{y}(s),$$

where $\hat{N}_a(s)$ results from replacing the operator d/dt by s in N_a, $\hat{u}(s) = \mathscr{L}[u]$, and $\hat{y}(s) = \mathscr{L}[y]$. Since N_a is a nonsingular operator,

$$H(s) = \hat{N}_a^{-1}(s)[C_a(sI - A)^{-1}B + D_a]. \tag{4.15}$$

It follows immediately from (4.15) that rank $H(s) \geqq q_a = \text{rank } \bar{D}_a$ for almost all s.

Also note that rank $H(s) = \text{rank } H(s)T$, where T is defined by (4.8) in the proof of Theorem 4.1. It is clear from that proof that there exists v_1 such that for arbitrary v_2,

$$H(s)T\begin{bmatrix}\hat{v}_1(s)\\ \hat{v}_2(s)\end{bmatrix} = 0.$$

Hence, for all s the null space of $H(s)T$ has dimension at least $r - q_a$. Hence

$$\text{rank } H(s) = \text{rank } H(s)T \leqq r - (r - q_a) = q_a$$

for all s, which completes the proof.

5. Zeroing the output and input isolation. A problem which is important in many areas of control [2]–[4] and one which is central to the decoupling problem is that of "zeroing the output" of a system either by feedback or open loop control. A complete solution to the open loop problem follows directly from the range theorem and its corollaries.

THEOREM 5.1.

(i) *There exists an input $u \in \mathscr{U}$ such that $\mathscr{S}_{x_0}[u]$ is identically zero on $[0, \infty)$ if and only if x_0 is in the null space of L_β.*

(ii) *Let x_0 be in the null space of L_β. Then the output y of $\mathscr{S}_{x_0}$ is identically zero on $[0, \infty)$ if and only if $\bar{y}_a(t)$ is identically zero on $[0, \infty)$.*

(iii) *Let x_0 be in the null space of L_β. Then $\mathscr{S}_{x_0}[u]$ is identically zero on $[0, \infty)$ if and only if u can be expressed as the output of the system representation*

$$\dot{z}(t) = (A - B\bar{D}_a^{\dagger}\bar{C}_a)z(t) + BKv(t), \tag{5.1}$$

$$u(t) = -\bar{D}_a^{\dagger}\bar{C}_a z(t) + Kv(t) \tag{5.2}$$

for some v and $z(0) = x_0$.

Proof. Part (i) follows immediately from Theorem 4.1. Part (ii) follows from the sufficiency proof of Theorem 4.1. Part (iii) follows from Theorem 4.2.

Remark. Let $\mathscr{Z} = ((A - B\bar{D}_a^{\dagger}\bar{C}_a), BK, -\bar{D}_a^{\dagger}\bar{C}_a, K)$. It is seen from part (iii) of Theorem 5.1 that the system representation $\mathscr{Z}$ completely characterizes the set of inputs which zero the output of $\mathscr{S}_{x_0}$ in the sense that any such input must be in the range of $\mathscr{Z}_{x_0}$. Clearly then, any system representation having the same range as $\mathscr{Z}_{x_0}$ will generate the same set of inputs. In particular, for the case $x_0 = 0$, $\mathscr{Z}$ can be replaced by any system which is zero state equivalent to it. Hence, by constructing a minimal such equivalent $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ we can always find a completely controllable and observable system representation whose range for the zero initial state is the set of all inputs which will zero the output of $\mathscr{S}_0$. Moreover, since system range is invariant under state feedback, by Corollary 4.1, we can replace $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ by $(\hat{A} + \hat{B}F, \hat{B}, \hat{C} + \hat{D}F, \hat{D})$ for any F. Since $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ is controllable by construction, $\hat{A} + \hat{B}F$ can have any distribution of eigenvalues by appropriate choice of F [19]. We summarize these remarks in the following corollary.

COROLLARY 5.1. *There exists a system representation* $\hat{\mathscr{Z}} = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ *having any desired set of eigenvalues with the following property*:

$$\mathscr{S}_0[u] = 0 \quad \textit{if and only if} \quad u = \hat{\mathscr{Z}}_0[v] \quad \textit{for some } v.$$

The set of inputs which zero the output of $\mathscr{S}_{x_0}$ also has a feedback representation. It is clear from the form of (5.1)–(5.2) that the same set of inputs can be generated by setting

$$u(t) = -\bar{D}_a^{\dagger}\bar{C}_a x(t) + Kv(t). \tag{5.3}$$

Hence we have the next corollary.

COROLLARY 5.2. *Let* x_0 *be in the null space of* L_β. *Then* $\mathscr{S}_{x_0}[u]$ *is identically zero on* $[0, \infty)$ *if and only if* u *can be expressed in the feedback form* (5.3) *for some* v.

In the remainder of this section we shall be concerned with zero state properties of $\mathscr{S}$. The notation $(A, B, C, D) \sim (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ will be used if $\mathscr{S}_0[u] = \hat{\mathscr{S}}_0[u]$ for all u. If $\mathscr{S}_0[u] = 0$ for all u, we say that $(A, B, C, D) \sim 0$.

Closely related to the problem of zeroing the output of a system is that of isolating the output from a subset of the inputs via a control law of the type (3.1). More precisely, suppose

$$\dot{x} = Ax + Bu + Ew, \tag{5.4}$$

$$y = Cx + Du, \tag{5.5}$$

where w is a "disturbance input" to the original system representation (2.1)–(2.2). When does there exist a control law of the type (3.1) such that in the closed loop system the output is unaffected by variations of w? Equivalently, when does there exist a matrix F such that $(A + BF, E, C + DF) \sim 0$?

This problem was first stated in essentially this form (with $D = 0$) by Wonham and Morse [9] who also provided a solution in terms of a recursively defined subspace. It is shown below that the solution of this problem follows quite easily from the range theorem and that the relevant subspace ($\bar{\mathscr{V}}$ of [9]) determined by Wonham and Morse is actually the null space of L_β. Moreover, use of the structure algorithm gives an explicit characterization of the entire family of feedback matrices F which will isolate a given input set.

THEOREM 5.2. *There exists a matrix* F *such that* $(A + BF, E, C + DF) \sim 0$ *if and only if* $L_\beta E = 0$.

Proof. To establish sufficiency, it is first noted that the inversion algorithm is invariant with respect to all E in the null space of L_β. The proof of this fact follows by a simple induction argument based on the observation that

$$J_{i+1} = \begin{bmatrix} \bar{D}_i \\ \tilde{C}_i(B + E) \end{bmatrix} = \begin{bmatrix} \bar{D}_i \\ \tilde{C}_i B \end{bmatrix}.$$

Since $\bar{C}_a \xrightarrow{F} \bar{C}_a + D_a F$ it follows by Corollary 5.2 that if $F = -\bar{D}_a^{\dagger}\bar{C}_a$ and $u = Fx$, the closed loop system $(A + BF, E, C + DF) \sim 0$ which completes the sufficiency proof.

Suppose now that the zero state response of

$$\dot{x}(t) = (A + BF)x(t) + Ew(t),$$

$$y(t) = (C + DF)x(t)$$

is identically zero for all w. Then certainly, $N_i y(t) \equiv 0$ for all w and $x(0) = 0$. Observe that $\tilde{N}_0 y(t) = \tilde{C}_0 x(t)$ so that

$$\frac{d}{dt}\tilde{N}_0 y(t) = \tilde{C}_0(A + BF)x(t) + \tilde{C}_0 E w(t) \equiv 0$$

which for $t = 0$ and $x(0) = 0$ implies $\tilde{C}_0 E = 0$, since $w(0)$ is unconstrained. By Lemma 3.1 it follows that $\tilde{N}_1 y(t) = C_1 x(t)$. Suppose now that $\tilde{C}_j E = 0$ and $\tilde{N}_{j+1} y(t) = \tilde{C}_{j+1} x(t)$. Then repeating the preceding arguments leads to $\tilde{C}_{j+1} E = 0$ and $\tilde{N}_{j+2} y(t) = \tilde{C}_{j+2} x(t)$. Consequently it follows by induction that $\tilde{C}_j E = 0$ for all $j = 0, 1, \cdots$, which completes the proof.

Since the matrix $[\bar{D}_a^\dagger \vdots K]$ is nonsingular, any feedback matrix F can be expressed in the form

$$F = \bar{D}_a^\dagger F_1 + KF_2, \tag{5.6}$$

where F_1 is a $q_a \times n$ matrix and F_2 is an $(r - q_a) \times n$ matrix. With this notation, a useful characterization of the class of matrices which will isolate a given disturbance matrix E can be given.

THEOREM 5.3. *Let E be such that $L_\beta E = 0$ and let F be expanded in the form* (5.6). *Then $(A + BF, E, C + DF) \sim 0$ if and only if*

$$(\bar{C}_a + F_1)(A - B\bar{D}_a^\dagger \bar{C}_a + BKF_2)^j E = 0 \tag{5.7}$$

for $j = 0, 1, \cdots, n - 1$.

Proof. By Theorem 5.1, part (ii), if $x_0 = 0$, the output y of the system $(A + BF, E, C + DF)$ is identically zero on $[0, \infty)$ if and only if $\bar{y}_a(t)$ is identically zero on the interval. Now

$$\bar{y}_a(t) = (\bar{C}_a + \bar{D}_a F)x(t),$$

where

$$\dot{x}(t) = (A + BF)x(t) + Ew(t).$$

Hence $\bar{y}_a(t) \equiv 0$ for all w if and only if

$$(\bar{C}_a + \bar{D}_a F)(A + BF)^j E = 0, \qquad j = 0, \cdots, n - 1. \tag{5.8}$$

With F represented by (5.6) this equation becomes

$$(\bar{C}_a + F_1)(A + BF)^j E = 0, \qquad j = 0, \cdots, n - 1. \tag{5.9}$$

It remains to show that (5.7) and (5.9) are equivalent. It is clear they are equivalent for $j = 0$.

Suppose (5.9) holds for $j = 0, 1, \cdots, n - 1$ and (5.7) holds for $j = 0, 1, \cdots, k$, and consider

$$(\bar{C}_a + F_1)(A + BF)^{k+1} E = (\bar{C}_a + F_1)(A + BF)(A + BF)^k E. \tag{5.10}$$

By (5.6),

$$BF = B\bar{D}_a^\dagger F_1 + BKF_2, \tag{5.11}$$

and using (5.9) with $j = k$ we have

$$(\bar{C}_a + F_1)(A + BF)^{k+1} E = (\bar{C}_a + F_1)(A - B\bar{D}_a^\dagger \bar{C}_a + BKF_2)(A + BF)^k E.$$

Continuing in this way by using (5.9) and (5.11) with $j = k - 1, k - 2, \cdots, 1$, one finds that

$$(\bar{C}_a + F_1)(A + BF)^{k+1}E = (\bar{C}_a + F_1)(A - B\bar{D}_a^{\dagger}\bar{C}_a + BKF_2)^{k+1}E.$$

Hence by induction, (5.9) implies (5.7). That (5.9) follows from (5.7) is easily established by reversing the steps of the proof.

When E is contained in the column range of B, i.e., $E = BG$ for some matrix G, it follows from Lemma 3.4 that if $\bar{D}_aG = 0$, then $L_\beta E = 0$ (the reverse implication is also true if $D = 0$). This observation together with Theorems 5.2 and 5.3 yields a characterization of the class of *all* feedback laws of the type

$$u = Fx + Gv \tag{5.12}$$

which will zero the output of $\mathcal{S}_0$.

THEOREM 5.4. *Let F be expanded in the form* (5.6). *Then*

$$(A + BF, BG, C + DF, DG) \sim 0 \tag{5.13}$$

if and only if

$$\bar{D}_aG = 0 \tag{5.14}$$

and

$$(\bar{C}_a + F_1)(A - B\bar{D}_a^{\dagger}\bar{C}_a + BKF_2)^jBG = 0 \tag{5.15}$$

for $j = 0, 1, \cdots, n - 1$.

Remark. The class of all pairs (F, G) which zero the output of $\mathcal{S}_0$ can be determined explicitly from (5.14) and (5.15); each G satisfying (5.14) and each F_2 yield a set of matrices F_1 which satisfy the linear equations

$$(\bar{C}_a + F_1)Q(F_2, G) = 0, \tag{5.16}$$

where

$$Q(F_2, G) = [BG \vdots \cdots \vdots (A - B\bar{D}_a^{\dagger}\bar{C}_a + BKF_2)^{n-1}BG].$$

Note that if $G = K$, the range of $Q(F_2, K)$ is the same as that of $Q(0, K)$. Hence, we have the following special case of Theorem 5.4.

COROLLARY 5.3.

$$(A + BF, GK, C + DF, DK) \sim 0 \tag{5.17}$$

if and only if

$$(\bar{C}_a + \bar{D}_aF)Q(0, K) = 0. \tag{5.18}$$

Note that in contrast to (5.16), equation (5.18) is linear in F.

Since left invertibility of a system is of particular importance in many applications, we shall conclude our general study of input-output properties with a summary of several equivalent characterizations of left invertibility in Theorem 5.5 below. Verification is straightforward and will be left to the reader.[3]

[3] $\mathcal{R}(B)$ denotes the column range of B and $\eta(L_\beta)$ denotes the null space of L_β.

THEOREM 5.5. *The following statements are equivalent*:

(i) $\mathscr{S}_{x_0}$ *is left invertible for all* x_0;

(ii) rank $H(s) = r$ *for almost all* s;

(iii) rank $\bar{D}_a = r$;

(iv) rank $\begin{bmatrix} D \\ L_\beta B \end{bmatrix} = r$.

If $D = 0$, *then the following statements are also equivalent to the above*:

(v) $\eta(L_\beta) \cap \mathscr{R}(B) = 0$;

(vi) *if* $E \neq 0$ *is contained in* $\mathscr{R}(B)$, *then there exists no matrix* F *such that* $(A + BF, E, C) \sim 0$;

(vii) *if* $x_0 \neq 0$ *is contained in* $\mathscr{R}(B)$, *then there exists no input* $u \in \mathscr{U}$ *such that* $\mathscr{S}_{x_0}[u] = 0$ *on* $[0, \infty)$.

6. State feedback decoupling. Consider a partition of the output of (A, B, C, D) into p nonempty subsets of components

$$y(t) = \begin{bmatrix} y^1(t) \\ \vdots \\ y^p(t) \end{bmatrix}, \tag{6.1}$$

and let m_i denote the size of the subvector y^i ($0 < m_i$ and $\sum_{i=1}^{p} m_i = m$). This partition induces a corresponding partition of C and D:

$$C = \begin{bmatrix} C^1 \\ \vdots \\ C^p \end{bmatrix}, \qquad D = \begin{bmatrix} D^1 \\ \vdots \\ D^p \end{bmatrix}, \tag{6.2}$$

where C^i and D^i each have m_i rows.

The basic decoupling problem we shall consider is that of finding a control law of the type

$$u(t) = Fx(t) + Gv(t)$$

(denoted as (F, G)), where

$$v(t) = \begin{bmatrix} v^1(t) \\ \vdots \\ v^p(t) \end{bmatrix}$$

is such that the ith input set v^i affects only the ith output set y^i. Let G be partitioned comformably with v as

$$G = [G^1 \vdots \cdots \vdots G^p]. \tag{6.3}$$

Then a more formal definition is the following.

DEFINITION 6.1. The feedback pair (F, G) *decouples* (A, B, C, D) (relative to the output partition (6.2)) if and only if

$$(A + BF, BG^i, C^j + D^jF, D^jG^i) \sim 0, \qquad j \neq i. \tag{6.4}$$

The state feedback decoupling problem then consists of finding a pair (F, G) satisfying (6.4) for a specified output partition. It is clear that the pair $(F, 0)$ will always decouple (A, B, C, D) so that without any further constraints the state feedback decoupling problem always has a solution. However, if the ultimate purpose of decoupling is to control the output of the original system in some non-trivial way, additional constraints are necessary. One such constraint is output controllability [20] of the closed loop system.

Defining

$$Q(F, G) = [BG \vdots \cdots \vdots (A + BF)^{n-1}BG] \tag{6.5}$$

and

$$P(C, D, F, G) = [(C + DF)Q(F, G) \vdots DG], \tag{6.6}$$

we have that the well-known necessary and sufficient condition for output controllability of $(A + BF, BG, C + DF, DG)$ is (see [20])

$$\operatorname{rank} P(C, D, F, G) = m. \tag{6.7}$$

If the closed loop system is decoupled, then this condition is clearly equivalent to saying that

$$\operatorname{rank} P(C^i, D^i, F, G^i) = m_i, \qquad i = 1, \cdots, p. \tag{6.8}$$

DEFINITION 6.2. The feedback pair (F, G) output controllably decouples (A, B, C, D) if and only if (6.4) and (6.8) hold.

For the case $D = 0$ the definition of the decoupling problem is equivalent to that of Wonham and Morse [9]. Other constraints on the decoupled system may also be desired such as the state controllability, observability, stability, etc., but the output controllability constraint appears to be the weakest meaningful one.

In the following sections we shall develop criteria and algorithms for decoupling based on the results of §§ 4 and 5. To do this, however, some new notation is required. Let C and D be partitioned as in (6.2). Then Γ^i is defined to be the matrix formed by deleting C^i from C and Δ^i the matrix formed by deleting D^i from D. Also with G partitioned as in (6.3), define Ω^i to be the matrix formed by deleting G^i from G.

We shall have occasion to perform the structure algorithm described in § 2 on the subsystems (A, B, C^i, D^i) and $(A, B, \Gamma^i, \Delta^i)$. The matrices appearing at the jth stage of the algorithm for (A, B, C^i, D^i) will be denoted as

$$C_j^i = \begin{bmatrix} \bar{C}_j^i \\ \cdots \\ \tilde{C}_j^i \end{bmatrix}, \qquad D_j^i = \begin{bmatrix} \bar{D}_j^i \\ \cdots \\ 0 \end{bmatrix}, \tag{6.9}$$

where $q_j^i = \operatorname{rank} \bar{D}_j^i$ is equal to the number of rows in $\bar{D}_j^i$ and $\bar{C}_j^i$. Also, define a_i to be the first integer such that $\operatorname{rank} \bar{D}_{a_i}^i = \operatorname{rank} \bar{D}_n^i$. For notational simplicity, we write $\bar{D}_{a_i}^i = \bar{D}_{a_i}$ and $C_{a_i}^i = C_{a_i}$.

Similarly, for $(A, B, \Gamma^i, \Delta^i)$ the appropriate matrices appearing at the jth step of the algorithm will be denoted as

$$\Gamma_j^i = \begin{bmatrix} \bar{\Gamma}_j^i \\ \cdots \\ \tilde{\Gamma}_j^i \end{bmatrix}, \qquad \Delta_j^i = \begin{bmatrix} \bar{\Delta}_j^i \\ \cdots \\ 0 \end{bmatrix}, \tag{6.10}$$

where $\xi_j^i = \operatorname{rank} \bar{\Delta}_j^i$ is equal to the number of rows in $\bar{\Delta}_j^i$ and $\bar{\Gamma}_j^i$. Furthermore, α_i is defined to be the first integer such that $\operatorname{rank} \Delta_{\alpha_i}^i = \operatorname{rank} \Delta_n^i$. Again for simplicity, write $\bar{\Delta}_{\alpha_i}^i = \bar{\Delta}_{\alpha_i}$ and $\Gamma_{\alpha_i}^i = \Gamma_{\alpha_i}$.

With the above notation, observe that condition (6.4) can be restated in two equivalent ways:

$$(A + BF, BG^i, \Gamma^i + \Delta^i F, \Delta^i G^i) \sim 0, \qquad i = 1, \cdots, p, \tag{6.11}$$

$$(A + BF, B\Omega^i, C^i + D^i F, D^i \Omega^i) \sim 0, \qquad i = 1, \cdots, p. \tag{6.12}$$

Furthermore, the output controllability constraint (6.8) has the equivalent formulation

$$\operatorname{rank} P(\Gamma^i, \Delta^i, F, \Omega^i) = m - m_i, \qquad i = 1, \cdots, p. \tag{6.13}$$

7. Decoupling characterizations. In this section we shall provide two alternate characterizations of the class of state feedback decoupling pairs. It will be shown in the following section that both characterizations are useful in that one leads to necessary conditions and the other to sufficient conditions for decoupling.

First, define $\mathscr{F}(G^i)$ to be the family of feedback matrices F such that the pair (F, G^i) zeros all but the ith set of outputs, i.e.,

$$\mathscr{F}(G^i) = \{F : (A + BF, BG^i, \Gamma^i + \Delta^i F, \Delta^i G^i) \sim 0\}. \tag{7.1}$$

A complete characterization of $\mathscr{F}(G^i)$ is provided by Theorem 5.4.

Following the development of § 5 we first note that any feedback matrix F can be uniquely represented in the form

$$F = \bar{\Delta}_{\alpha_i}^{\dagger} F_{i1} + K_i F_{i2}, \tag{7.2}$$

where K_i is a fixed matrix whose columns form a basis for the right null space of Δ_{α_i}. Further, define

$$Q_i(F_{i2}, G^i) = [BG^i \vdots \cdots \vdots (A - B\bar{\Delta}_{\alpha_i}^{\dagger} \bar{\Gamma}_{\alpha_i} + BK_i F_{i2})^{n-1} BG^i]. \tag{7.3}$$

Then the following is an immediate corollary of Theorem 5.4.

COROLLARY 7.1. *$F \in \mathscr{F}(G^i)$ if and only if*

$$(\bar{\Gamma}_{\alpha_i} + F_{i1}) Q_i(F_{i2}, G^i) = 0. \tag{7.4}$$

Similarly, denoting by $\mathscr{F}^*(\Omega^i)$ the family of all matrices F for which

$$(A + BF, \Omega^i, C^i + D^i F, D^i \Omega^i) \sim 0$$

and representing F as

$$F = \bar{D}_{a_i}^{\dagger} F_{i1}^* + K_i^* F_{i2}^* \tag{7.5}$$

where K_i^* is a fixed matrix whose columns form a basis for the right null space of $\bar{D}_{a_i}$, we may define

$$Q_i^*(F_{i2}^*, \Omega^i) = [B\Omega^i \vdots \cdots \vdots (A - B\bar{D}_{a_i}^{\dagger} \bar{C}_{a_i} + BK_i^* F_{i2}^*)^{n-1} B\Omega^i]. \tag{7.6}$$

Again by Theorem 5.4 we have the following corollary.

COROLLARY 7.2. *$F \in \mathscr{F}^*(\Omega^i)$ if and only if*

$$(\bar{C}_{a_i} + F_{i1}^*) Q_i^*(F_{i2}^*, \Omega^i) = 0. \tag{7.7}$$

The output controllability criterion (6.8) admits of considerable simplification when F is restricted to $\mathscr{F}(G^i)$. Let

$$P_i(F_{i2}, G^i) = [(C^i - D^i\bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Gamma}_{\alpha_i} + D^iK_iF_{i2})Q_i(F_{i2}, G^i) \vdots D^iG^i]. \tag{7.8}$$

LEMMA 7.1. *If* $F \in \mathscr{F}(G^i)$, *then*

$$Q(F, G^i) = Q_i(F_{i2}, G^i), \tag{7.9}$$

$$P(C^i, D^i, F, G^i) = P_i(F_{i2}, G^i). \tag{7.10}$$

Proof. Let $F_{i0} = \bar{\Gamma}_{\alpha_i} + F_{i1}$. Then by direct substitution,

$$(A + BF)^jBG^i = (A - B\bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Gamma}_{\alpha_i} + BK_iF_{i2} + B\bar{\Delta}_{\alpha i}^{\dagger}\bar{\Gamma}_{\alpha_i}F_{i0})^jBG^i.$$

Hence, by (7.4) and a straightforward induction argument it follows that

$$(A + BF)^jBG^i = (A - B\bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Gamma}_{\alpha_i} + BK_iF_{i2})^jBG^i, \qquad j = 0, 1, \cdots,$$

which establishes (7.9). Equation (7.10) follows by a similar argument.

Similarly, we define

$$P_i^*(F_{i2}^*, \Omega^i) = [(\Gamma^i - \Delta^i\bar{D}_{a_i}^{\dagger}\bar{C}_{a_i} + \Delta^iK_i^*F_{i2}^*, \Omega^i) \vdots \Delta^i\Omega^i]. \tag{7.11}$$

A parallel proof to the above establishes the next lemma.

LEMMA 7.2. *If* $F \in \mathscr{F}^*(\Omega^i)$, *then*

$$Q^*(F, \Omega^i) = Q_i^*(F_{i2}^*, \Omega^i), \tag{7.12}$$

$$P(\Gamma^i, \Delta^i, F, \Omega^i) = P_i^*(F_{i2}^*, \Omega^i). \tag{7.13}$$

The first major characterization of decoupling pairs is given by the following theorem.

THEOREM 7.1. *The pair* (F, G) *output controllably decouples* (A, B, C, D) *if and only if*

$$\bar{\Delta}_{\alpha_i}G^i = 0, \qquad i = 1, \cdots, p, \tag{7.14}$$

$$(\bar{\Gamma}_{\alpha_i} + F_{i1})Q_i(F_{i2}, G^i) = 0, \qquad i = 1, \cdots, p, \tag{7.15}$$

$$\text{rank } P_i(F_{i2}, G^i) = m_i, \qquad i = 1, \cdots, p. \tag{7.16}$$

Proof. It follows by Theorem 5.4 and Corollary 7.1 that (7.14) and (7.15) are necessary and sufficient for (F, G) to decouple (A, B, C, D). In light of Lemma 7.1, equation (7.16) is just a restatement of the output controllability criterion.

By a parallel proof, we also have the following characterization of decoupling pairs.

THEOREM 7.2. *The pair* (F, G) *output controllably decouples* (A, B, C, D) *if and only if*

$$\bar{D}_{a_i}\Omega^i = 0, \qquad i = 1, \cdots, p, \tag{7.17}$$

$$(\bar{C}_{a_i} + F_{i1}^*)Q_i^*(F_{i2}^*, \Omega^i) = 0, \qquad i = 1, \cdots, p, \tag{7.18}$$

$$\text{rank } P_i^*(F_{i2}^*, \Omega^i) = m - m_i, \qquad i = 1, \cdots, p. \tag{7.19}$$

A considerable simplification of the decoupling characterization results if attention is restricted to the class of decoupling pairs for which G is square and nonsingular. Such decoupling pairs which preserve the original number of independent inputs to (A, B, C, D) can be justified in many problems either from physical or mathematical constraints. An example of the latter is the important special case in which the number of inputs to (A, B, C, D) equals the number of output partitions (in particular, when $D = 0$, $m_i = 1$ and $m = r$ we have the problem of Falb and Wolovich). The simplification obtained when G is nonsingular is that output controllability is always preserved independently of F.

LEMMA 7.3. *If G is a square nonsingular matrix, then (A, B, C, D) is output controllable if and only if $(A + BF, BF, C + DF, DG)$ is output controllable for any F.*

The proof of this lemma is essentially the same as for the well-known results for state controllability [2], and will be omitted.

Following immediately from Lemma 7.3 and Theorems 7.1 and 7.2, respectively, are the two simplified characterizations.

THEOREM 7.3. *The pair (F, G), with G square and nonsingular, output controllably decouples (A, B, C, D) if and only if* (7.14) *and* (7.15) *hold and (A, B, C, D) is output controllable.*

THEOREM 7.4. *The pair (F, G), with G square and nonsingular, output controllably decouples (A, B, C, D) if and only if* (7.17) *and* (7.18) *hold and (A, B, C, D) is output controllable.*

8. Decoupling criteria. In this section, the decoupling characterizations as expressed in Theorems 7.1–7.4 of the previous section will be employed to develop explicit criteria for the existence of decoupling pairs (F, G). The results presented, while not representing a complete resolution of this question, do include within them all known previous results [9]–[15], as well as several important extensions.

In general, the determination of decoupling pairs (F, G) is difficult since F and G are interrelated in a highly complicated way through the equations (e.g., (7.14)–(7.16)) characterizing output controllable decoupling. With additional constraints on G, however, the problem becomes quite tractable.

The first group of results in this section derive from the choice $G = K$, where

$$K = [K_1 \vdots \cdots \vdots K_p] \tag{8.1}$$

with the K_i as defined in the previous section. Theorems 8.3–8.5 present sufficient conditions for the existence of a matrix F to complete a decoupling pair of the form (F, K).

The remaining results of this section pertain to those situations in which G is chosen to be (or must be) nonsingular. Necessary and sufficient conditions are derived for a decoupling pair of this type to exist.

The column space of a matrix A will be denoted as $\mathscr{R}[A]$. Let $\bar{Q}_i$ be a matrix whose columns form a basis for $\mathscr{R}[Q_i(0, K_i)]$ and let $\bar{Q}_i^*$ be a matrix whose columns form a basis for $\mathscr{R}[Q_i^*(0, K_i^*)]$. $\bar{P}_i$ and $\bar{P}_i^*$ are defined similarly.

The following two lemmas are basic to the development of the necessary condition expressed in Theorems 8.1 and 8.2 and the sufficient conditions expressed in Theorems 8.3–8.5.

LEMMA 8.1. *If* $\bar{\Delta}_{\alpha_i} G^i = 0$, *then*

$$\mathscr{R}[Q_i(F_{i2}, G^i)] \subseteq \mathscr{R}[\bar{Q}_i], \tag{8.2}$$

$$\mathscr{R}[P_i(F_{i2}, G^i)] \subseteq \mathscr{R}[\bar{P}_i]. \tag{8.3}$$

These relationships hold with equality if $G^i = K_i$.

Proof. To establish (8.2), it need only be shown that there exists a matrix L such that $Q_i(F_{i2}, G^i) = Q_i(0, K_i)L$. Since $\bar{\Delta}_{\alpha_i} G^i = 0$, $G^i = K_i M$ for some matrix M. Then $Q_i(F_{i2}, G^i)$ consists of blocks of the form $(A - B\bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Gamma}_{\alpha_i} + BK_iF_{i2})^k BK_iM$, $k = 0, 1, \cdots, n-1$, so that it is sufficient to establish the existence of a matrix L (not the same for each block) satisfying a similar relationship for each block. Now, if each such block is expanded, each resultant term is of the form BK_iN or $(A - B\bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Gamma}_{\alpha_i})^j N$, $j \leqq k$, for some matrix N (not the same for each term). Hence each term is expressible in the form $Q_i(0, K_i)P$ for some matrix P (not the same for each term), which establishes (8.2). That equality holds when $G^i = K_i$ follows from the invariance of the controllability space of the pair $(A - B\bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Gamma}_{\alpha_i}, BK_i)$ under feedback.

To establish (8.3), define matrices M and N by $G^i = K_iM$, $Q_i(F_{i2}, G^i) = Q_i(0, K_i)N$. Then, employing an obvious partitioning,

$$P_i(F_{i2}, G^i) = P_i(0, K_i)\begin{bmatrix} N & 0 \\ F_{i2}Q_i(F_{i2}, G^i) & M \end{bmatrix}.$$

The condition of equality when $G^i = K_i$ follows from invariance of controllability matrices under feedback.

A parallel form of this lemma is established in an identical fashion.

LEMMA 8.2. *If* $\bar{D}_{a_i}\Omega^i = 0$, *then*

$$\mathscr{R}[Q_i^*(F_{i2}^*, \Omega^i)] \subseteq \mathscr{R}[\bar{Q}_i^*], \tag{8.4}$$

$$\mathscr{R}[P_i^*(F_{i2}^*, \Omega^i)] \subseteq \mathscr{R}[\bar{P}_i^*]. \tag{8.5}$$

These relationships hold with equality if $\Omega^i = K_i^*$.

A parallel can be drawn here with the work of Wonham and Morse [9]. It is clear from the various definitions that for the case $D = 0$, $\mathscr{R}[\bar{Q}_i]$ is the maximum dimension controllability subspace which is simultaneously in the null spaces of C^j, $j \neq i$. For the more general situation it is clear that $\mathscr{R}[\bar{P}_i]$ is the largest output controllability subspace in R^{m_i} which can be generated while simultaneously zeroing $y^j(t)$, $j \neq i$. These remarks lead to the following necessary condition.

THEOREM 8.1. *There exists a pair* (F, G) *which output controllably decouples* (A, B, C, D) *only if*

$$\operatorname{rank} \bar{P}_i = m_i, \qquad i = 1, \cdots, p. \tag{8.6}$$

A similar argument establishes the following theorem.

THEOREM 8.2. *There exists a pair* (F, G) *which output controllably decouples* (A, B, C, D) *only if*

$$\operatorname{rank} \bar{P}_i^* = m - m_i, \qquad i = 1, \cdots, p. \tag{8.7}$$

Theorem 8.1 provides a useful point of departure for the development of sufficient conditions for decoupling. For if (8.6) is satisfied, output controllability

is guaranteed by the choice $G = K$ and, as will be seen, the further conditions to be imposed on F are considerably simplified. On the other hand, a parallel approach starting with the choice $\Omega^i = K_i^*$ is not in general possible since there may not exist a G to simultaneously realize all the K_i^*.

As suggested, considerable simplification is achieved if attention is restricted to the choice $G = K$.

It follows from Theorem 7.1 and Lemma 8.1 that if condition (8.6) is satisfied, the pair (F, K) output controllably decouples $\mathscr{S}$ if and only if the equations

$$(\bar{\Gamma}_{\alpha_i} + \bar{\Delta}_{\alpha_i} F)\bar{Q}_i = 0, \qquad i = 1, \cdots, p, \tag{8.8}$$

have a solution for F. (We have used the fact that $F_{i1} = \bar{\Delta}_{\alpha_i} F$.) Since $\bar{Q}_i$ does not depend on F, this set of equations is linear in the elements of F. Hence standard techniques can be utilized to determine if a solution exists. If no solution exists, however, it cannot be concluded that a solution to the original decoupling problem does not exist, since a solution may still be possible with matrices G_i whose columns do not span the null space of $\bar{\Delta}_{\alpha_i}$.

Let

$$\hat{\Gamma} = \begin{bmatrix} \bar{\Gamma}_{\alpha_i} \\ \vdots \\ \bar{\Gamma}_{\alpha_p} \end{bmatrix}, \qquad \hat{\Delta} = \begin{bmatrix} \bar{\Delta}_{\alpha_i} \\ \vdots \\ \bar{\Delta}_{\alpha_p} \end{bmatrix}. \tag{8.9}$$

Then a special case for which a solution to (8.8) exists is given in the next theorem.

THEOREM 8.3. *If* rank $P_i = m_i$, $i = 1, \cdots, p$, *and* $\mathscr{R}[\hat{\Gamma}] \subset \mathscr{R}[\hat{\Delta}]$, *then there exists a pair* (F, K) *which output controllably decouples* (A, B, C, D).

Proof. Since $\mathscr{R}[\hat{\Gamma}] \subset \mathscr{R}[\hat{\Delta}]$, there exists a matrix F such that $\hat{\Delta}F = -\hat{\Gamma}$. This choice satisfies (8.8) which completes the proof.

Although it is not possible to develop useful sufficient conditions along similar lines by considering the choice $\Omega^i = K_i^*$, the following lemma provides a link so that the characterization of Theorem 7.2 may be employed for a useful parallel result.

For a fixed partition of G in the form (6.3), we have the next lemma.

LEMMA 8.3. *If* $\bar{\Delta}_{\alpha_i} G^i = 0$, $i = 1, \cdots, p$, *then* $\bar{D}_{a_i} \Omega^i = 0$, $i = 1, \cdots, p$.

Proof. From the structure algorithm developed in § 2, it is clear that for any i, $i = 1, \cdots, p$, the rows of $\bar{D}_{a_i}$ can be expressed as a linear combination of the rows of $\bar{\Delta}_{\alpha_j}$ for any $j \neq i$. Hence, if $\bar{\Delta}_{\alpha_j} G^j = 0$, $j \neq i$, $\bar{D}_{a_i} G^j = 0$, $j \neq i$, i.e., $\bar{D}_{a_i} \Omega^i = 0$.

Lemma 8.3 allows one to consider the choice $G = K$ in connection with the first two conditions, (7.17) and (7.18), of Theorem 7.2. The satisfaction of these two guarantee decoupling, while (8.6) guarantees output controllability. Using Lemmas 8.2 and 8.3, one now sees that it is sufficient for decoupling to require that a solution for F exist for the set of equations

$$(\bar{C}_{a_i} + \bar{D}_{a_i} F)\bar{Q}_i^* = 0, \qquad i = 1, \cdots, p, \tag{8.10}$$

with F given by (7.5).

Let

$$\hat{C} = \begin{bmatrix} \bar{C}_{a_1} \\ \vdots \\ \bar{C}_{a_p} \end{bmatrix}, \qquad \hat{D} = \begin{bmatrix} \bar{D}_{a_1} \\ \vdots \\ \bar{D}_{a_p} \end{bmatrix}. \tag{8.11}$$

Then, by a proof analogous to that for Theorem 8.3, we have the following theorem.

THEOREM 8.4. *If* rank $\bar{P}_i = m_i$, $i = 1, \cdots, p$, *and* $\mathscr{R}[\hat{C}] \subset \mathscr{R}[\hat{D}]$, *then there exists a pair* (F, G) *which output controllably decouples* (A, B, C, D).

A special case of this result is the following.

COROLLARY 8.1. *If* rank $\bar{P}_i = m_i$, $i = 1, \cdots, p$, *and the rows of* $\hat{D}$ *are linearly independent, then there exists a pair* (F, G) *which output controllably decouples* (A, B, C, D).

The sufficiency conditions expressed in this corollary, while only a specialization of the conditions of Theorem 8.3, will be seen to be necessary as well if it is required that G be nonsingular.

Under the conditions of Corollary 8.1, a class of decoupling pairs can be explicitly displayed. The independence of the rows of $\hat{D}$ implies the existence of a right inverse, $\hat{D}^\dagger$. Hence, the pair $(-\hat{D}^\dagger\hat{C}, K)$ output controllably decouples (A, B, C, D).

Let

$$\bar{Q} = [\bar{Q}_1 \vdots \cdots \vdots \bar{Q}_p]. \tag{8.12}$$

Then another special case for which a solution to (8.8) is guaranteed is given in the next theorem.

THEOREM 8.5. *If* rank $\bar{P}_i = m_i$, $i = 1, \cdots, p$, *and the columns of* $\bar{Q}$ *are linearly independent, then there exists a pair* (F, G) *which output controllably decouples* (A, B, C, D). *Moreover, under these conditions, the pair* (F_1, K), *with*

$$F_1 = -[\bar{\Delta}_{a_1}^\dagger \bar{\Gamma}_{\alpha_1} \bar{Q}_1 \vdots \cdots \vdots \bar{\Delta}_{\alpha_p}^\dagger \bar{\Gamma}_{\alpha_p} \bar{Q}_p]\bar{Q}^\dagger, \tag{8.13}$$

output controllably decouples (A, B, C, D).

Proof. Equation (8.8) can be rewritten as

$$\bar{\Delta}_{\alpha_i}(F\bar{Q}_i) = -\bar{\Gamma}_{\alpha_i}\bar{Q}_i, \qquad i = 1, \cdots, p.$$

Since the rows of $\bar{\Delta}_{\alpha_i}$ are linearly independent, a solution for $F\bar{Q}_i$ for each of these equations exists. Specifically,

$$F\bar{Q}_i = -\bar{\Delta}_{\alpha_i}^\dagger \bar{\Gamma}_{\alpha_i}\bar{Q}_i, \qquad i = 1, \cdots, p.$$

Collecting these equations, we have

$$F\bar{Q} = -[\bar{\Delta}_{\alpha_1}^\dagger \bar{\Gamma}_{\alpha_1}\bar{Q}_1 \vdots \cdots \vdots \bar{\Delta}_{\alpha_p}^\dagger \bar{\Gamma}_{\alpha_p}\bar{Q}_p].$$

Since the columns of $\bar{Q}$ are linearly independent, $F = F_1$ is a solution of (8.8) which, together with $G = K$, output controllably decouples (A, B, C, D).

Remark. The general solution of (8.8) is obtained by noting that the complete solution for $F\bar{Q}$ is

$$F\bar{Q} = F_1\bar{Q} + KF_2,$$

where $F_2 = \operatorname{diag}(F_{i2})$ and F_{i2}, $i = 1, \cdots, p$, are arbitrary. (F_{i2} has the same number

of rows as K_i has columns, and the same number of columns as $\bar{Q}_i$ has columns.) This in turn yields

$$F = F_1 + KF_2\bar{Q}^{\dagger}. \tag{8.14}$$

This general form for F will now be utilized to show that when the conditions of Theorem 8.5 are satisfied and rank $\bar{Q} = n$, decoupling can be achieved with $A + BF$ strictly stable (all eigenvalues having negative real part).

To see this, first note that by Lemma 8.1 the columns of $\bar{Q}_i$ form a basis for $\mathscr{R}[Q(F, K_i)]$. Performing the coordinate transformation $z = \bar{Q}^{-1}x$, where x is the state of the decoupled system, one can then readily verify that

$$\hat{A} = \bar{Q}^{-1}(A + BF_1)\bar{Q} + \bar{Q}^{-1}(BKF_2\bar{Q}^{-1})\bar{Q}$$

has the block diagonal form

$$\hat{A} = \begin{bmatrix} \hat{A}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{A}_p \end{bmatrix} + \begin{bmatrix} \hat{B}_1F_{12} & & 0 \\ & \ddots & \\ 0 & & \hat{B}_pF_{p2} \end{bmatrix}$$

and

$$\hat{B} = \bar{Q}^{-1}BK = \begin{bmatrix} \hat{B}_1 & & 0 \\ & \ddots & \\ 0 & & \hat{B}_p \end{bmatrix}.$$

The pairs $(\hat{A}_i, \hat{B}_i)$ are clearly controllable, so that $(\hat{A}_i + \hat{B}_iF_{i2})$ can be assigned any desired spectrum by choice of F_{i2} (see [18]). Hence, $\hat{A}$ and therefore $(A + BF)$ can be assigned any spectrum with an F of the form (8.14).

A corollary of Theorem 8.5 is Theorem 4.1 of Wonham and Morse [9] (stated somewhat differently).

COROLLARY 8.2. *Suppose* $D = 0$ *and* $m = n$, *with* rank $C = n$. *Then there exists a pair* (F, G) *which output controllably decouples* (A, B, C, D) *if and only if*

$$\text{rank } \bar{P}_i = m_i, \qquad i = 1, \cdots, p. \tag{8.15}$$

Proof. Since the number of rows of $\bar{P}_i$ is m_i, necessity follows from Theorem 7.1. To prove sufficiency, consider the matrix $C\bar{Q}$. By construction, $C^i\bar{Q}_j = 0, i \neq j$, and $C^j\bar{Q}_j = \bar{P}_j$, i.e., $C\bar{Q} = \text{diag}\,(\bar{P}_i)$, so that rank $C\bar{Q} = \sum_{i=1}^{p} \text{rank } \bar{P}_i = n$. Since rank $C = n$ it follows that rank $\bar{Q} = n$. Also, since $\Gamma^j\bar{Q}_j = 0$, $j = 1, \cdots, p$, and rank $\Gamma^j = n - m_j$, rank $\bar{Q}_j \leqq m_j$. Since each $\bar{Q}_j$ has full column rank, it follows that $\bar{Q}$ has n linearly independent columns. Then the desired result follows immediately from Theorem 8.5.

If the necessary condition of Theorem 8.1 is satisfied but no solution to (8.8) exists, the next logical step is to choose a G of smaller dimension than K and re-examine the decoupling conditions. A guideline for a second choice of G is available from the characterization theorems, Theorems 7.1 and 7.2. From the definition of K_i, it is evident that G must be chosen in the form $G^i = K_iN_i$ in order to satisfy (7.14). Hence, a reasonable approach is to select a set N_i, $i = 1, \cdots, p$, then examine (7.15) and (7.17) for the possibility of a solution for F. Equation (7.17)

should be examined first to determine if a set F_{i2}, $i = 1, \cdots, p$, exists to achieve output controllability. Then if this proves successful, (7.15) can be examined by standard techniques, since for F_{i2} fixed, this equation is linear in the elements of F_{i1}. As yet, no way of generally terminating this algorithm after a finite number of steps has been found.

At this point, we shall return to the condition utilized in Corollary 8.1—that $\hat{D}$ have linearly independent rows. It will be shown that for a large class of systems this condition is both necessary and sufficient for decoupling.

LEMMA 8.4. *If there exists a nonsingular matrix G such that for some F the pair (F, G) decouples (A, B, C, D), then the rows of $\hat{D}$ are linearly independent.*

Proof. If (F, G) decouples (A, B, C, D), then from (7.17), $\bar{D}_{a_i} G^j = 0$, $i \neq j$; hence $\hat{D}G$ must be block diagonal. Also, since the rows of $\bar{D}_{a_i}$ are linearly independent and G is nonsingular, the rows of $\bar{D}_{a_i} G$ are linearly independent. Due to its block diagonal structure, it follows that the rows of $\hat{D}G$ are linearly independent. Then since G is nonsingular, the result follows.

THEOREM 8.6. *There exists a pair (F, G) with G nonsingular which output controllably decouples (A, B, C, D) if and only if*

(i) *(A, B, C, D) is output controllable, and*

(ii) *the rows of $\hat{D}$ are linearly independent.*

Moreover, if conditions (i) *and* (ii) *are satisfied, the pair*

$$\hat{F} = -\hat{D}^{\dagger}\hat{C}, \qquad \hat{G} = [\hat{D}^{\dagger} \vdots K^*], \tag{8.16}$$

where the columns of K^* form a basis for the null space of $\hat{D}$, output controllably decouples (A, B, C, D).

Proof. Necessity follows by the preceding lemma and Theorem 7.4. Sufficiency follows by observing that $\hat{G}$ and $\hat{F}$ satisfy conditions (7.17) and (7.18), respectively.

The following corollary covers a case considered by Wonham and Morse [9, Theorem 5.1].

COROLLARY 8.3. *If $r = p$, then there exists a pair (F, G) which output controllably decouples (A, B, C, D) if and only if*

(i) *(A, B, C, D) is output controllable, and*

(ii) *the rows of $\hat{D}$ are linearly independent.*

Proof. It is clear that if $r = p$, the G matrix of any pair which output controllably decouples must have rank r so that the result follows by the preceding theorem.

Remark. If (ii) holds and $r = p$, then it is clear that $\bar{D}_{a_i}$ is a row vector and rank $\hat{D} = r$. Hence, if (i) and (ii) hold, $(-\hat{D}^{-1}\hat{C}, \hat{D}^{-1})$ output controllably decouples (A, B, C, D). For the case $D = 0$, Wonham and Morse [9] also give a necessary and sufficient condition for the existence of a decoupling pair when $r = p$. Their condition is quite different in form from that given above. One advantage of the present formulation is that it reduces to the "standard" result given by Falb and Wolovich [12] for the subcase $r = p = m$. It is clear from the structure algorithm and the above remark that with $D = 0$ and $r = p$ that if (ii) holds, there is a row r^i of C^i such that

$$\bar{D}_{a_i} = r^i A^{a_i - 1} B$$

and

$$\bar{C}_{a_i} = r^i A^{a_i},$$

where a_i is the first positive integer j such that $r^i A^{j-1} B \neq 0$. Hence, if C^i contains only one row for all i, $\hat{C} = A^*$ and $\hat{D} = B^*$ in the notation of Falb and Wolovich [12], which makes the equivalence with their results apparent.

In our notation, therefore, the Falb-Wolovich result takes the following form.

COROLLARY 8.4. *If $m = p = r$, there exists a pair (F, G) which output controllably decouples (A, B, C, D) if and only if*

$$\text{rank } \hat{D} = m. \tag{8.17}$$

Moreover, if (8.15) is satisfied, $(-\hat{D}^{-1}\hat{C}, \hat{D}^{-1})$ is a decoupling pair.

Proof. Condition (i) of Corollary 8.3 is redundant in this case since rank $\hat{D} = m$ implies the decoupled system is invertible, a stronger condition than output controllability.

When the conditions of Corollary 8.3 are satisfied, it is also possible to give a characterization of the complete class of decoupling pairs.

LEMMA 8.5. *Suppose $p = r$, (A, B, C, D) is output controllable, and $\hat{D}$ has linearly independent rows. Then there exists a matrix F such that the pair (F, G) output controllably decouples (A, B, C, D) if and only if G has the form*

$$G = \hat{D}^{-1}\Lambda, \tag{8.18}$$

where $\Lambda = \text{diag}(\lambda_i)$ is any nonsingular diagonal matrix.

Proof. Necessity follows immediately from (7.17). For sufficiency, note that G of the form (8.18) satisfies (7.17). By noting that $F = -\hat{D}^{-1}\hat{C}$ satisfies (7.18) and applying Theorem 7.4, the proof is completed.

In order to delineate the class of matrices F which together with a G of the form (8.18) output controllably decouple (A, B, C, D), some additional notation will be needed. Let Σ_i be the identity matrix with the ith column deleted, let $\rho_i = n - \text{rank } Q_i^*(0, \hat{D}^{-1}\Sigma_i)$ and let H_i be any $\rho_i \times n$ matrix having rank ρ_i such that

$$H_i Q_i^*(0, \hat{D}^{-1}\Sigma_i) = 0, \qquad i = 1, \cdots, p. \tag{8.19}$$

Also, let

$$H = \begin{bmatrix} H_1 \\ \vdots \\ H_p \end{bmatrix}. \tag{8.20}$$

Then we have the following theorem.

THEOREM 8.7. *Suppose $p = r$, (A, B, C, D) is output controllable, and $\hat{D}$ has linearly independent rows. Then (F, G) output controllably decouples (A, B, C, D) if and only if G has the form* (8.18) *and F has the form*

$$F = \hat{D}^{-1}(-\hat{C} + \hat{F}H), \tag{8.21}$$

where $\hat{F} = \text{diag}(f_i)$ and f_i, $i = 1, \cdots, p$, is an arbitrary $1 \times \rho_i$ matrix.

Proof. Lemma 8.5 established the form of G. The necessary and sufficient conditions for (F, G) to output controllably decouple (A, B, C, D) then reduces to the requirement that F satisfy (7.15) here rewritten in the form

$$(\bar{C}_{a_i} + \bar{D}_{a_i} F) Q_i^*(F_{i2}^*, \Omega^i) = 0, \qquad i = 1, \cdots, p. \tag{8.22}$$

Note that $\Omega^i = \hat{D}^{-1}\Lambda\Sigma_i$, hence $\bar{D}_{a_i}\Omega^i = 0$. Therefore, since rank $\Omega^i = r - 1$, one can take $K_i^* = \Omega^i$, and, by applying Lemma 8.2, $\mathscr{R}[Q_i^*(F_{i2}^*, \Omega^i)] = \mathscr{R}[Q_i^*(0, \hat{D}^{-1}\Lambda\Sigma_i)]$. As Λ is a nonsingular, diagonal matrix, it can be seen that H_i satisfies (8.19) if and only if $H_iQ_i^*(F_{i2}^*, \Omega^i) = 0$. Then F in the form (8.21) is seen to satisfy (8.22). For necessity, note that any F can be expressed as

$$F = -\hat{D}^{-1}\hat{C} + \hat{D}^{-1}M \tag{8.23}$$

for some M. Substitution of (8.23) into (8.22) yields the requirement that

$$M_iQ_i^*(0, \hat{D}^{-1}\Sigma_i) = 0, \qquad i = 1, \cdots, p,$$

where

$$M = \begin{bmatrix} M_1 \\ \vdots \\ M_p \end{bmatrix}.$$

The form (8.21) is then seen to include all solutions F.

This characterization of decoupling pairs appears to be new even for the subcase $r = p = m$. It is, however, quite similar to Gilbert's characterization [15, Theorem 5]. Moreover, the above characterization can be utilized to very simply resolve the question of eigenvalue placement in the decoupled system. This method is also similar to, but somewhat more direct than, Gilbert's procedure.

The following lemma is equivalent to the related result of Gilbert [15, Lemma 1] so that the proof can be omitted.

LEMMA 8.6. *Let H be the matrix defined by* (8.19) *and* (8.20). *Then*

(i) *the rows of H are linearly independent,*

(ii) *there exists a matrix A_i such that*

$$H_i(A - B\hat{D}^{-1}\hat{C}) = A_iH_i$$

and

(iii) $\tilde{C}_j^i$ *for $j = 0, 1, \cdots$ is in the row range of H_i.*

Remark. As a consequence of (iii) above, it can be assumed without loss of generality that if

$$L_j^i = \begin{bmatrix} \tilde{C}_0^i \\ \vdots \\ \tilde{C}_{j-1}^i \end{bmatrix} \tag{8.24}$$

and β_i is the first integer such that rank $L_{\beta_i}^i$ = rank $L_{\beta_{i+1}}^i = \gamma_i$, then the first γ_i rows of H_i can be chosen to be the first γ_i independent rows of L_{β_i}.

Following from Lemma 8.6 and Corollary 8.3 is a simple representation of the class of all decoupled systems for which the conditions of Theorem 8.7 hold. First note that (8.19) implies

$$HB\hat{D}^{-1} = \begin{bmatrix} b_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_p \end{bmatrix}, \tag{8.25}$$

where b_i is a $\rho_i \times 1$ matrix. It is clear from Lemma 8.6 and (8.25) that

$$H(A + BF)H^{\dagger} = \begin{bmatrix} (A_1 - b_1 f_1) & \cdots & 0 \\ \vdots & \ddots & \\ 0 & \cdots & (A_p - b_p f_p) \end{bmatrix},$$

where F is any matrix of the form (8.22) and $H^{\dagger} = H'(HH')^{-1}$.

If J is any matrix whose columns form a basis for the right null space of H and

$$T = \begin{bmatrix} H \\ J^{\dagger} \end{bmatrix}, \qquad J^{\dagger} = (J'J)^{-1}J',$$

is used as a coordinate transformation, it can then be verified that the class of decoupled systems $(T(A + BF)T^{-1}, TBG, (C + DF)T^{-1}, DG)$ has the following explicit form (assume without loss of generality that $D^i = D_0^i$ and $C^i = C_0^i$):

$$T(A + BF)T^{-1} = \left[\begin{array}{ccc|c} (A_1 - b_1 f_1) & \cdots & 0 & \\ \vdots & & \vdots & 0 \\ 0 & \cdots & (A_p - b_p f_p) & \\ \hline J^{\dagger}(A - B\hat{D}^{-1}\hat{C} + B\hat{D}^{-1}\hat{F}H)H^{\dagger} & & & J^{\dagger}(A - B\hat{D}^{-1}\hat{C})J \end{array}\right],$$

$$TBG = \begin{bmatrix} b_1\lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & b_p\lambda_p \\ J^{\dagger}B\hat{D}^{-1}\Lambda & & \end{bmatrix}, \qquad DG = \begin{bmatrix} D^1G \\ \vdots \\ D^pG \end{bmatrix},$$

where

$$D^iG = \begin{cases} \begin{bmatrix} e_i\lambda_i \\ 0 \end{bmatrix} & \text{if } D^i \neq 0, \\ 0 & \text{if } D^i = 0, \end{cases}$$

with e_i the ith row of I_r, and

$$(C + DF)T^{-1} = \begin{bmatrix} (C^1 + D^1F)T^{-1} \\ \vdots \\ (C^p + D^pF)T^{-1} \end{bmatrix},$$

where

$$(C^i + D^iF)T^{-1} = \begin{cases} \begin{bmatrix} \bar{f}_i \\ E^i_{m_i - 1} \end{bmatrix} & \text{if } D^i \neq 0, \\ E^i_{m_i} & \text{if } D^i = 0, \end{cases}$$

and E_l^i is an $l \times n$ matrix whose kth row is the $(\sum_{j=1}^{i-1}\rho_j + k)$th row of I_n and f_i is

the row vector $\vec{f}_i$ with all entries replaced by zero except those indexed $\sum_{j=1}^{i-1} \rho_j + k, k = 1, \cdots, \rho_i$.

Several results follow directly from this characterization. The most important is the following which generalizes and reformulates the related results of Gilbert [15] and Wonham and Morse [9].

THEOREM 8.8. *Suppose* $p = r$, (A, B, C, D) *is state controllable and the rows of* $\hat{D}$ *are linearly independent. Then there exists a pair* (F, G) *such that* $(A + BF, BG, C + DF, DG)$ *is output controllably decoupled and strictly stable if and only if* $J^{\dagger}(A - B\hat{D}^{-1}\hat{C})J$ *is strictly stable.*

Proof. Controllability of (A, B, C, D) implies that the pairs (A_i, b_i) are controllable so that $(A_i + b_i f_i)$ can be assigned any desired spectrum by choice of f_i. It is also clear that the eigenvalues of $J^{\dagger}(A - B\hat{D}^{-1}\hat{C})J$ are invariant under decoupling feedback.

Remark. Under the hypothesis of Theorem 8.8, (A, B, C, D) is left invertible and the dynamic part of the left inverse has the representation (see (4.13)–(4.14))

$$(A - B\hat{D}^{-1}\hat{C}, B\hat{D}^{-1}, -\hat{D}^{-1}\hat{C}, \hat{D}^{-1}).$$

It is clear, therefore, that a sufficient condition for stable decoupling when $\hat{D}$ is nonsingular is that (A, B, C, D) be "minimum phase" in the sense that the eigenvalues of the inverse system have negative real part.

It is also clear from the characterization of decoupled systems that the transfer function matrix $Z(s, F, G)$ has the form

$$Z(s, F, G) = \operatorname{diag}(Z_i(s, F, G)),$$

where

$$Z_i(s, F, G) = \begin{cases} \begin{bmatrix} \vec{f}_i \\ E^i_{m_i - 1} \end{bmatrix} (sI - A_i + b_i f_i)^{-1} b_i \lambda_i + e_i \lambda_i & \text{if } D^i \neq 0, \\ E^i_{m_i} (sI - A_i + b_i f_i)^{-1} b_i \lambda_i & \text{if } D^i = 0. \end{cases}$$

Hence, the transfer function matrix of the decoupled system can be assigned any desired pole pattern.

9. Open loop and dynamic decoupling. As seen in the previous section, a complete solution to the state feedback decoupling problem is not yet possible. Moreover, even for the class of systems in which the decoupling problem is resolved, stable decoupling by state feedback is not always possible. This has motivated the examination of more general decoupling laws [15], [18], [20]. Morse and Wonham [18] have shown that if the class of decoupling laws is sufficiently enlarged, necessary and sufficient conditions can be given both for decoupling and stable decoupling (the conditions turn out to be identical). This problem will also be considered here. It is shown below that the open loop characterization of zeroing inputs given by Corollary 5.1 leads to a very simple proof of the general decoupling criterion as well as an explicit realization of a dynamic decoupling law. Moreover, this approach leads to a considerably stronger result—a necessary and sufficient condition for stable, dynamic decoupling with *output* feedback only.

We shall first consider the open loop decoupling problem.

Let $\mathcal{U}_i$ denote a subset of the input space $\mathcal{U}$ of $\mathcal{S}$ and let $\mathcal{S}^j_{x_0}$ denote the system defined by the quadruple (A, B, C^j, D^j) with initial state x_0. Then the following definition of an open loop decoupling set of inputs will be adopted. It is essentially equivalent to the open loop decoupling definition given in [18].

DEFINITION 9.1. A family of input sets $\{\mathcal{U}_i\}_{i=1}^p$ *output controllably decouples* $\mathcal{S}$ if and only if:

(i) $\mathcal{S}^j_0[u] = 0$ for $j \neq i$ and $u \in \mathcal{U}_i$; and

(ii) for every $\lambda \in R^{m_i}$ there is a $\hat{u} \in \mathcal{U}_i$ such that $\hat{y}^i(1) = \lambda$, where $\hat{y}^i(1)$ is the response of $\mathcal{S}^j_0$ to $\hat{u}$ evaluated at $t = 1$.

The following theorem gives a necessary and sufficient condition for the existence of a decoupling family. This result is equivalent (for the case $D = 0$) to the corresponding result of Morse and Wonham [18, Theorem 6.2]. The proof is believed to be simpler and more constructive.

THEOREM 9.1. *There exists a family of input sets* $\{\mathcal{U}_i\}_{i=1}^p$ *which output controllably decouples* $\mathcal{S}$ *if and only if*

$$\text{rank } \bar{P}_i = m_i, \qquad i = 1, \cdots, p. \tag{9.1}$$

Proof. Necessity. Suppose $\{\mathcal{U}_i\}_{i=1}^p$ output controllably decouples $\mathcal{S}$. By Corollary 5.2 the set $\bar{\mathcal{U}}_i$ of all inputs u for which $\mathcal{S}^j_0[u] = 0$ for $j \neq i$ can be generated by the feedback law

$$u = -\bar{\Delta}^{\dagger}_{\alpha_i}\bar{\Gamma}_{\alpha_i}x + K_i v. \tag{9.2}$$

Moreover, $\bar{P}_i$ is the output controllability matrix of $\mathcal{S}^i$ subject to feedback of the form (9.2). Hence if condition (ii) of Definition 9.1 holds, $\bar{P}_i$ must have rank m_i since by condition (i), $\mathcal{U}_i \subset \bar{\mathcal{U}}_i$.

Sufficiency. If rank $\bar{P}_i = m_i$, then by definition, the set of input functions $\bar{\mathcal{U}}_i$ generated by the feedback law (9.2) with $x_0 = 0$ satisfies conditions (i) and (ii) of Definition 9.1.

Remark. As seen the the previous section, it is not generally true that all the $\bar{\mathcal{U}}_i$ can be generated simultaneously with a single feedback law. However, as was shown in Theorem 5.1, the set $\mathcal{U}_i$ can be generated in an open loop manner by the system representation $\mathcal{Z}^i$,

$$\dot{z}^i = (A - B\bar{\Delta}^{\dagger}_{\alpha_i}\bar{\Gamma}_{\alpha_i})z^i + BK_i v_i, \tag{9.3}$$

$$u^i = -\bar{\Delta}^{\dagger}_{\alpha_i}\bar{\Gamma}_{\alpha_i}z^i + K_i v_i \tag{9.4}$$

with $z^i(0) = 0$. It is clear, therefore, if we set

$$u = \sum_{i=1}^{p} u^i,$$

where u^i is the output of (9.3)–(9.4), decoupling is achieved with the dynamic control law defined by the direct sum of the system representations $\mathcal{Z}^i$.

More generally one can define a general dynamic control law by a pair of functions characterizing a dynamical system $\mathcal{T}$ coupled from $\mathcal{S}$:

$$z(t) = \mathcal{O}(t, t_0, z(t_0), x_{[t_0,t]}, v_{[t_0,t]}), \tag{9.5}$$

$$u(t) = \mathcal{B}(t, z(t), x(t), v(t)), \tag{9.6}$$

where $\mathcal{O}$ is the state transition function [22] of $\mathcal{T}$, $z(t) \in \Sigma'$, the state space of $\mathcal{T}$, $v \in \mathcal{V}$, the input space of the closed loop system and $\mathcal{B}$ is the readout function [22] of $\mathcal{T}$. The overall response of $\mathcal{T}$ can be represented by a single operator

$$u(t) = \mathcal{T}_{x_0,z_0}[t, v], \tag{9.7}$$

where it is understood that $t_0 = 0$, $x(0) = x_0$, $z(0) = z_0$ and $v = v_{[0,t]}$. Let v be partitioned as in (6.3) and let $\mathcal{T}^i_{x_0,z_0}[t, v^i] = \mathcal{T}_{x_0,z_0}[t, v]$ when $v^j = 0, j \neq i$. The closed loop system will be denoted as $\mathcal{S}[\mathcal{T}]$.

A general definition of dynamic decoupling can now be given.

DEFINITION 9.2. The control law (9.7) *output controllably decouples* $\mathcal{S}$ if and only if

(i) $\mathcal{S}^j_0[\mathcal{T}^i_{0,0}[t, v^i]] = 0$ for $j \neq i$,

and

(ii) the closed loop system $\mathcal{S}[\mathcal{T}]$ is output controllable.

The next theorem follows immediately from Theorem 5.1 and the subsequent remarks.

THEOREM 9.2. *There exists a control law of the type* (9.7) *which output controllably decouples* $\mathcal{S}$ *if and only if*

$$\operatorname{rank} \bar{P}_i = m_i, \qquad i = 1, \cdots, p. \tag{9.8}$$

Dynamic decoupling can always be achieved with a strictly stable closed loop system if (9.8) is satisfied and $\mathcal{S}$ is (state) controllable. To see this, first note that there is no loss of generality in assuming that A is strictly stable since if $\mathcal{S}$ is controllable a preliminary feedback law of the type $u = Fx + w$ can be used to give $A + BF$ any desired set of eigenvalues [19]. Also, it follows from Corollary 5.1 that the set of inputs $\bar{\mathcal{U}}_i$ can be generated by a system

$$\mathcal{Z}^i = (\hat{A}^i, \hat{B}^i, \hat{C}^i, \hat{D}^i)$$

with $\hat{A}^i$ having any desired set of eigenvalues. Consequently, the combined system $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ can have any desired set of eigenvalues since

$$\hat{A} = \begin{bmatrix} A & B\hat{C}^1 & B\hat{C}^2 & \cdots & B\hat{C}^p \\ 0 & \hat{A}^1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & \hat{A}^p \end{bmatrix}.$$

For completeness note that

$$\hat{B} = \begin{bmatrix} B\hat{D}^1 & B\hat{D}^2 & \cdots & B\hat{D}^p \\ \hat{B}^1 & 0 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \hat{B}^p \end{bmatrix},$$

$$\hat{C} = [C \vdots D\hat{C}^1 \vdots \cdots \vdots D\hat{C}^p], \qquad \hat{D} = [D\hat{D}^1 \vdots D\hat{D}^2 \vdots \cdots \vdots D\hat{D}^p].$$

It is of interest to observe that if the original A matrix is stable, no state feedback is required to achieve stable decoupling when condition (9.8) holds. In

many applications the state of $\mathscr{S}$ cannot be measured directly so that this type of decoupling law is quite important. These ideas can be extended by incorporating an observer [23], [24] into the dynamic decoupling law. It is well known that if $\mathscr{S}$ is controllable and observable, an observer of the type

$$\dot{z} = \check{A}z + \check{B}y \tag{9.9}$$

with

$$u = \check{F}z + w \tag{9.10}$$

can be found so that the closed loop system has any desired set of eigenvalues. It is clear, therefore, that if we can show that rank $\bar{P}_i$ is invariant under this type of feedback a necessary and sufficient condition for stable dynamic decoupling which utilizes only output measurements will result. To show that this is indeed the case, a generalization of Lemma 3.1 will be given.

Observe first that under feedback of the type (9.9)–(9.10), the closed loop system $\check{\mathscr{S}}$ is represented by the equations

$$\begin{bmatrix} \dot{x} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} A & B\check{F} \\ \check{B}C & (\check{A} + \check{B}D\check{F}) \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} + \begin{bmatrix} B \\ \check{B}D \end{bmatrix} w, \tag{9.11}$$

$$y = [C \vdots D\check{F}] \begin{bmatrix} x \\ z \end{bmatrix} + Dw. \tag{9.12}$$

If P is any operator associated with $\mathscr{S}$ and Q is the corresponding operator of $\check{\mathscr{S}}$, we shall say that

$$P \xrightarrow{(\check{A}, \check{B}, \check{F})} Q.$$

LEMMA 9.1. *The operators N_i and the matrices D_i of the structure algorithm are invariant under dynamic output feedback of the type* (9.9)–(9.10) *and*

$$\bar{C}_i \xrightarrow{(\check{A}, \check{B}, \check{F})} [\bar{C}_i \vdots \bar{D}_i\check{F}], \tag{9.13}$$

$$\tilde{C}_i \xrightarrow{(\check{A}, \check{B}, \check{F})} [\tilde{C}_i \vdots 0]. \tag{9.14}$$

The proof of this result follows in a manner essentially the same as that used to establish Lemma 3.1 and is therefore omitted.

Remark. Lemma 9.1 establishes the invariants of the structure algorithm applied to the whole output of $\mathscr{S}$. It is clear that similar results hold for the algorithm applied to a subset of the output. In particular, $\bar{\Delta}_{\alpha_i}$ is invariant so that K_i is also invariant and

$$\bar{\Gamma}_{\alpha_i} \xrightarrow{(\check{A}, \check{B}, \check{F})} [\bar{\Gamma}_{\alpha_i} \vdots \bar{\Delta}_{\alpha_i}\check{F}] \tag{9.15}$$

and

$$\tilde{\Gamma}_{\alpha_i} \xrightarrow{(\check{A}, \check{B}, \check{F})} [\tilde{\Gamma}_{\alpha_i} \vdots 0]. \tag{9.16}$$

LEMMA 9.2. rank $\bar{P}_i$ *is invariant under dynamic output feedback of the type* (9.9)–(9.10).

Proof. Using (9.11), (9.12) and (9.15), (9.16) we can see that

$$A - B\bar{\Delta}_{\alpha_i}\bar{\Gamma}_{\alpha_i} \xrightarrow{(\check{A},\check{B},\check{F})} \left[\begin{array}{c|c} A - B\bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Gamma}_{\alpha_i} & B(I - \bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Delta}_{\alpha_i})\check{F} \\ \hline \check{B}(C - D\bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Gamma}_{\alpha_i}) & A + \check{B}D(I - \bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Delta}_{\alpha_i})\check{F} \end{array}\right] \tag{9.17}$$

and

$$C^i - D^i\bar{\Delta}_{\alpha_i}\bar{\Gamma}_{\alpha_i} \xrightarrow{(\check{A},\check{B},\check{F})} [C^i - D^i\bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Gamma}_{\alpha_i} : D^i(I - \bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Delta}_{\alpha_i})\check{F}]. \tag{9.18}$$

Since the columns of $I - \bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Delta}_{\alpha_i}$ span the null space of $\bar{\Delta}_{\alpha_i}$ there exists a matrix M_i such that

$$I - \bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Delta}_{\alpha_i} = K_iM_i. \tag{9.19}$$

Let $\check{P}_i(0, K_i)$ be defined by

$$P_i(0, K_i) \xrightarrow{(\check{A},\check{B},\check{F})} \check{P}_i(0, K_i), \tag{9.20}$$

and for convenience reorder the columns of $P_i(0, K_i)$ so that

$$P_i(0, K_i) = [D^iK_i : (C^i - D^i\bar{\Delta}_{\alpha_i}^{\dagger}\bar{\Gamma}_{\alpha_i})Q_i(0, K_i)] \tag{9.21}$$

with $Q_i(0, K_i)$ given by (7.3). Using (9.17), (9.18) and (9.19) one can then see that $\check{P}_i(0, K_i)$ has the form

$$\check{P}_i(0, K_i) = P_i(0, K_i)\begin{bmatrix} I_{r_1} & J_{11} & \cdots & J_{1n} \\ 0 & I_{r_2} & \cdots & J_{2,n-1} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & I_{r_p} \end{bmatrix},$$

where $r_i = \operatorname{rank} K_i$ ($J_{11} = M_{11}\check{F}\check{B}D$, etc.). Since the matrix postmultiplying $P_i(0, K_i)$ is nonsingular, the desired result follows.

It is now clear by Lemma 9.2 and Theorem 9.2 that if rank $\bar{P}_i = m_i$ for $i = 1, \cdots, p$, and $\mathscr{S}$ is controllable and observable, then $\mathscr{S}$ can be output controllably decoupled by dynamic compensation and output feedback alone with the resulting closed loop system having any desired eigenvalues. The procedure to accomplish this is first to construct an observer of the type (9.9)–(9.10) to shift the eigenvalues of A (the eigenvalues of the observer can be placed arbitrarily). The closed loop system $\check{\mathscr{S}}$ resulting is then decoupled by the procedure given in the remarks following Theorem 9.1. The combined compensation for decoupling can be illustrated in block diagonal form as shown in Fig. 1, where $\mathscr{O}$ represents the observer and $\mathscr{P}$ the precompensator for decoupling.

The above results can be summarized by the following theorem.

THEOREM 9.3. *Let $\mathscr{S}$ be controllable and observable. Then there exists a dynamic control law of the type*

$$u = \hat{C}_1z_1 + \hat{C}_2z_2 + \hat{D}v,$$

where

$$\dot{z}_1 = \hat{A}_1z_1 + \hat{B}_1y \quad (\textit{observer})$$

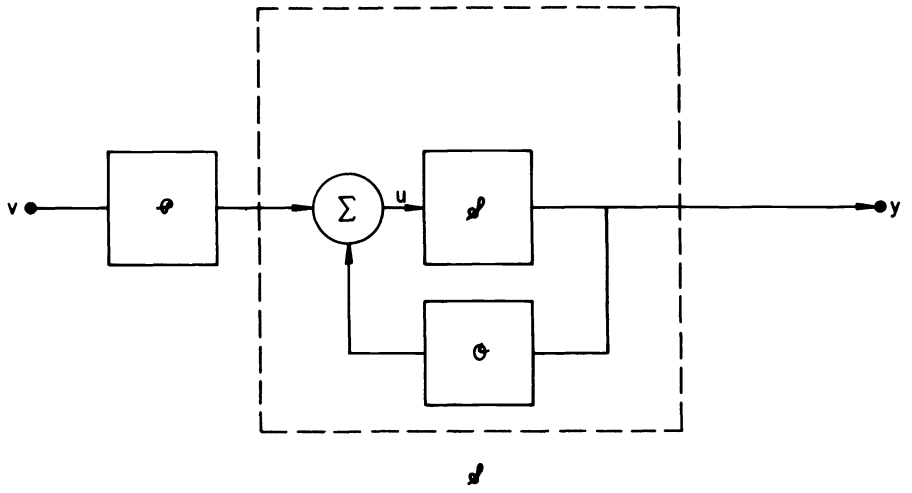


FIG. 1

and

$$\dot{z}_2 = \hat{A}_2 z_2 + \hat{B}_2 v \quad \textit{(decoupler)}$$

such that the closed loop system

$$\begin{bmatrix} \dot{x} \\ \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} A & B\hat{C}_1 & B\hat{C}_2 \\ \hat{B}_1 C & (\hat{A}_1 + \hat{B}_1 D\hat{C}_1) & \hat{B}_1 D\hat{C}_2 \\ 0 & 0 & \hat{A}_2 \end{bmatrix} \begin{bmatrix} x \\ z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} B\hat{D} \\ \hat{B}_1 D\hat{D} \\ B_2 \end{bmatrix} v,$$

$$y = [C \vdots D\hat{C}_1 \vdots D\hat{C}_2] \begin{bmatrix} x \\ z_1 \\ z_2 \end{bmatrix} + D\hat{D}v$$

is output controllably decoupled and strictly stable if and only if rank $\bar{P}_i = m_i$ *for* $i = 1, \cdots, p$.

Theorem 9.3 completely solves the "classical" decoupling problem of decoupling a system which is specified by its transfer function matrix (see Morgan [11] for references to the early literature) since a transfer function matrix is only a faithful representation of its minimal (controllable and observable) realizations. Note that for the case $r = m = p$ the condition on rank $\bar{P}_i$ is equivalent to invertibility-nonsingularity of the transfer function matrix.

The question of minimizing the compensating dynamical system for decoupling is still open. For the case where state feedback is allowed, some progress has been made by Morse and Wonham [18].

10. Concluding remarks. Several problems of interest related to input–output structure and decoupling are still outstanding. The most obvious is that of generalizing the results given here to the time-variable case. An indication of how such generalizations can be obtained is given in [8]. The problem of determining, by a

finite algorithm, whether or not a given system can be decoupled is still not resolved in general. The characterization theorems of § 7 should be useful in this respect. The related problem of decoupling with a minimal order compensator and output feedback is also open.

Acknowledgment. The authors would like to thank E. G. Gilbert, A. S. Morse and W. M. Wonham for several stimulating discussions on the decoupling problem.

REFERENCES

[1] L. Weiss, *On a question related to the control of linear systems*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 176–177.

[2] R. W. Brockett, *Poles, zeros and feedback: State space interpretation*, Ibid., AC-10 (1965), pp. 129–135.

[3] R. Sivan, *On zeroing the output and maintaining it zero*, Ibid., AC-10 (1965), pp. 193–194.

[4] L. M. Silverman, *Properties and application of inverse systems*, Ibid., AC-13 (1968), pp. 436–437.

[5] R. W. Brockett and M. D. Mesarovic, *The reproducibility of multivariable control systems*, J. Math. Anal. Appl., 11 (1965), pp. 548–563.

[6] D. C. Youla and P. Dorato, *On the inverse of linear dynamical systems*, Electrophysics Memo. PIBMR I-1319-66, Polytechnic Institute of Brooklyn, New York, 1966.

[7] M. K. Sain and J. L. Massey, *Invertibility of linear time-invariant systems*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 141–149.

[8] L. M. Silverman, *Inversion of multivariable linear systems*, Ibid., AC-14 (1969), pp. 270–276.

[9] W. M. Wonham and A. S. Morse, *Decoupling and pole assignment in linear multivariable systems: A geometric approach*, this Journal, 8 (1970), pp. 1–18.

[10] B. S. Morgan, Jr., *The synthesis of linear multivariable systems by state variable feedback*, Proc. 1964 JACC, Stanford, Calif., pp. 468–472.

[11] ———, *Multivariable systems*, 1965 IEEE International Convention Record, Part 6, pp. 87–95.

[12] P. L. Falb and W. A. Wolovich, *On the decoupling of multivariable systems*, Proc. 1967 JACC, Philadelphia, pp. 791–796.

[13] ———, *Decoupling in the design of multivariable control systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 651–659.

[14] W. A. Wolovich and P. L. Falb, *On the structure of multivariable systems*, this Journal, 7 (1969), pp. 437–451.

[15] E. G. Gilbert, *The decoupling of multivariable systems by state feedback*, this Journal, 7 (1969), pp. 50–63.

[16] Z. V. Rekasius, *Decoupling of multivariable systems by means of state variable feedback*, Proc. Third Allerton Conference on Circuit and System Theory, Monticello, Ill., 1965, pp. 439–448.

[17] A. S. Morse, *Output controllability and system synthesis*, this Journal, 9 (1971), pp. 143–148.

[18] A. S. Morse and W. M. Wonham, *Decoupling and pole assignment by dynamic compensation*, this Journal, 8 (1970), pp. 317–337.

[19] W. M. Wonham, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 660–665.

[20] E. Kreindler and P. E. Sarachik, *On the concepts of controllability and observability of linear systems*, Ibid., AC-9 (1964), pp. 129–136.

[21] L. M. Silverman, *Decoupling with state feedback and precompensation*, Ibid., to appear.

[22] R. E. Kalman, P. L. Falb and M. Arbib, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[23] D. G. Luenberger, *Observers for multivariable systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 190–197.

[24] J. J. Bongiorno, Jr. and D. C. Youla, *On observers in multivariable control systems*, Internat. J. Control, 8 (1968), pp. 221–243.

SIAM J. CONTROL
Vol. 9, No. 2, May 1971

EQUILIBRIUM FEEDBACK CONTROL IN LINEAR GAMES WITH QUADRATIC COSTS*

D. L. LUKES†

Abstract. Problems concerning equilibrium strategies for linear games are treated in a Hilbert space setting. First a nonlinear equation is derived for the equilibrium solutions in the player's allowed operator feedback spaces. The results obtained from the study of this equation are then used to compute the solution to a class of differential games and to establish the uniqueness of the solution.

1. An n-player game in Hilbert space.

1.1. Introduction. This paper studies the class of games modeled by the linear state equation

$$x = x_0 + \sum_1^n \mathscr{B}_j u_j \tag{1.1}$$

with *state variables* x, x_0 in a Hilbert space H; *control variables* u_j in a Hilbert space H_j, $j = 1, 2, \cdots, n$, and $\mathscr{B}_j : H_j \to H$ specified bounded linear transformations. The n players compute nonzero sum *costs* by means of (1.1) and prescribed quadratic functionals (1.2′).

The existence, uniqueness and stability of equilibrium solutions in $\sum_1^n \oplus H_j$ were investigated by Russell and Lukes [1] for games defined in terms of (1.1), (1.2′). That treatment includes the games in which (1.1) is an ordinary differential equation (a differential game). Varaiya [2] has developed an existence theory for the differential game with costs satisfying certain convexity conditions and with constraints upon the u_i.

The feedback synthesis of the equilibrium solution obtained for the differential game in [1] and further studied in [3] motivated the author's search for solutions to the game (1.1), (1.2′) in subspaces of $[H, \sum_1^n \oplus H_j]$, the Banach space of all bounded linear transformations from H into $\sum_1^n \oplus H_j$, rather than in $\sum_1^n \oplus H_j$ itself. The subsequent solutions presented below are probably more realistic in terms of their game-theoretic interpretation.

A nonlinear operator equation whose solutions are the equilibriums is derived in Theorem 1.1. Sufficient conditions for this equation to have a solution are then presented by Theorems 1.3, 1.5. To demonstrate the applicability of the results, the differential game is treated as an example, with operator subspaces taken to be matrix operators. In the proof of the local playability of the resulting game, Theorem 2.3, a method emerges for estimating the time interval over which the game has a solution. With the local playability established, an equilibrium matrix solution is thereby obtained, computed in terms of the solution to a system of quadratic matrix differential equations. This solution agrees with the results derived by Case [4] and others who have studied differential games

* Received by the editors July 23, 1970, and in revised form October 5, 1970.

† Department of Applied Mathematics and Computer Science, University of Virginia, Charlottesville, Virginia 22901. This research was supported by the National Aeronautics and Space Administration under Grant NGR 47-005-029.

by means of a Hamilton–Jacobi approach. The theory of the present article moreover establishes the uniqueness of those solutions.

The following abstract approach is intended to illuminate the underlying structure common to a general class of games including dynamic games of the above type.

1.2. The cost functionals. Player P_i, who commands control variable u_i

functional

$$\mathscr{C}_i' = |x - \bar{x}_i|^2_{\mathscr{W}_i} + |u_i|^2_{\mathscr{U}_i}, \qquad i = 1, 2, \cdots, n, \tag{1.2'}$$

in which target states $\bar{x}_i \in H$ and bounded self-adjoint linear operators $\mathscr{W}_i$, $\mathscr{U}_i$ are given. Of the associated quadratic forms we require $|x|^2_{\mathscr{W}_i} = (x, \mathscr{W}_i x) \geqq 0$ for all $x \in H$ and $|u_i|^2_{\mathscr{U}_i} = (u_i, \mathscr{U}_i u_i) > 0$ for all nonzero $u_i \in H_i$. By making the obvious preliminary change of control variables we can take the $\mathscr{U}_i$ to be identity operators and hereafter taking $\bar{x}_i = 0$ deal with the less cumbersome forms

$$\mathscr{C}_i = |x|^2_{\mathscr{W}_i} + |u_i|^2, \qquad i = 1, 2, \cdots, n. \tag{1.2}$$

1.3. Equilibrium feedback control. *Feedback controls* are defined to be bounded linear transformations $\mathscr{L}_i : H \to H_i$, $i = 1, 2, \cdots, n$, which give rise to the *abstract variation of parameters formulas*

$$x_i = \left(\mathscr{I} - \sum_{j \neq i}^{n} \mathscr{B}_j \mathscr{L}_j\right)^{-1} (x_0 + \mathscr{B}_i u_i) \tag{1.3}$$

with $x_0 \in H$, $u_i \in H_i$, $i = 1, 2, \cdots, n$, and

$$x = \left(\mathscr{I} - \sum_{1}^{n} \mathscr{B}_j \mathscr{L}_j\right)^{-1} x_0, \qquad x_0 \in H. \tag{1.4}$$

Hence if player P_i plays $u_i \in H_i$ in (1.1) while the remaining players play feedback controls $u_j = \mathscr{L}_j x$, then P_i incurs the cost computed from (1.2),

$$\begin{aligned} \mathscr{C}_i(\mathscr{L}_1, \mathscr{L}_2, \cdots, u_i, \cdots, \mathscr{L}_n) &= |x_i|^2_{\mathscr{W}_i} + |u_i|^2 \\ &= |x_0 + \mathscr{B}_i u_i|^2_{\Omega_i} + |u_i|^2, \end{aligned} \tag{1.5}$$

where

$$\Omega_i = \left(\mathscr{I} - \sum_{j \neq i}^{n} \mathscr{B}_j \mathscr{L}_j\right)^{-1*} \mathscr{W}_i \left(\mathscr{I} - \sum_{j \neq i}^{n} \mathscr{B}_j \mathscr{L}_j\right)^{-1}. \tag{1.6}$$

We shall restrict player P_i's feedback controls $\mathscr{L}_i \in L_i = L_i[H, H_i]$ where L_i is a closed subspace of $[H, H_i]$, the Banach space of all bounded linear transformations from H into H_i, $i = 1, 2, \cdots, n$. A *feedback system* $\mathscr{L} = (\mathscr{L}_1, \mathscr{L}_2, \cdots, \mathscr{L}_n)$ will refer to an element, $\mathscr{L} = \sum_1^n \oplus \mathscr{L}_i$ with $\mathscr{L}_i \in L_i$, in the space of transformations $\sum_1^n \oplus L_i$ from $\sum_1^n \oplus H$ into $\sum_1^n \oplus H_i$. We call a feedback system $(\hat{\mathscr{L}}_1, \hat{\mathscr{L}}_2, \cdots, \hat{\mathscr{L}}_n)$ an *equilibrium feedback system* (or simply an *equilibrium*) if it produces state and control responses defined by the equations

$$\hat{x} = x_0 + \sum_{1}^{n} \mathscr{B}_j(\hat{\mathscr{L}}_j \hat{x}), \tag{1.7}$$

$$\hat{u}_i = \hat{\mathscr{L}}_i \hat{x}, \qquad i = 1, 2, \cdots, n, \tag{1.8}$$

such that

$$\mathscr{C}_i(\hat{\mathscr{L}}_1, \hat{\mathscr{L}}_2, \cdots, \hat{u}_i, \cdots, \hat{\mathscr{L}}_n) \leqq \mathscr{C}_i(\hat{\mathscr{L}}_1, \hat{\mathscr{L}}_2, \cdots, u_i, \cdots, \hat{\mathscr{L}}_n) \tag{1.9}$$

for all $u_i \in H_i$, $i = 1, 2, \cdots, n$.

THEOREM 1.1. *A system of feedback controls $(\mathscr{L}_1, \mathscr{L}_2, \cdots, \mathscr{L}_n)$ is an equilibrium system if and only if it satisfies the equations*

$$\left[\mathscr{L}_i + \mathscr{B}_i^*\left(\mathscr{I} - \sum_{j \neq i}^{n} \mathscr{B}_j \mathscr{L}_j\right)^{-1*} \mathscr{W}_i\right]\left(\mathscr{I} - \sum_{1}^{n} \mathscr{B}_j \mathscr{L}_j\right)^{-1} x_0 = 0, \tag{1.10}$$

$i = 1, 2, \cdots, n$. The corresponding control response is given by the formula

$$u_i = -(\mathscr{I} + \mathscr{B}_i^* \Omega_i \mathscr{B}_i)^{-1} \mathscr{B}_i^* \Omega_i x_0, \tag{1.11}$$

where Ω_i is computed from the equilibrium system according to (1.6).

Proof. Assume $(\mathscr{L}_1, \mathscr{L}_2, \cdots, \mathscr{L}_n)$ is a feedback system satisfying (1.10). Define Ω_i by (1.6) and consider

$$\hat{u}_i = -(\mathscr{I} + \mathscr{B}_i^* \Omega_i \mathscr{B}_i)^{-1} \mathscr{B}_i^* \Omega_i x_0, \tag{1.12}$$

$i = 1, 2, \cdots, n$. Note that the inverse operator in (1.12) exists. Now let $u_i \in H_i$, $i = 1, 2, \cdots, n$, be arbitrary and compute

$$\begin{aligned} &\mathscr{C}_i(\mathscr{L}_1, \mathscr{L}_2, \cdots, u_i, \cdots, \mathscr{L}_n) - \mathscr{C}_i(\mathscr{L}_1, \mathscr{L}_2, \cdots, \hat{u}_i, \cdots, \mathscr{L}_n) - |u_i - \hat{u}_i|^2_{\mathscr{I} + \mathscr{B}_i^* \Omega_i \mathscr{B}_i} \\ &\quad = |x_0 + \mathscr{B}_i u_i|^2_{\Omega_i} + |u_i|^2 - |x_0 + \mathscr{B}_i \hat{u}_i|^2_{\Omega_i} - |\hat{u}_i|^2 \\ &\qquad - |u_i - \hat{u}_i|^2_{\mathscr{I} + \mathscr{B}_i^* \Omega_i \mathscr{B}_i} \\ &\quad = 2([\mathscr{B}_i^* \Omega_i x_0 + (\mathscr{I} + \mathscr{B}_i^* \Omega_i \mathscr{B}_i)\hat{u}_i], u_i - \hat{u}_i) = 0, \end{aligned} \tag{1.13}$$

the last equality following from (1.12). Hence with (1.13) holding for all $u_i \in H_i$, $i = 1, 2, \cdots, n$, the inequalities (1.9) are satisfied. The completion of the argument involves showing $\hat{u}_i$, $i = 1, 2, \cdots, n$, is the feedback response to $(\mathscr{L}_1, \mathscr{L}_2, \cdots, \mathscr{L}_n)$. We compute

$$\begin{aligned} &(\mathscr{I} + \mathscr{B}_i^* \Omega_i \mathscr{B}_i)\left[\hat{u}_i - \mathscr{L}_i\left(\mathscr{I} - \sum_{1}^{n} \mathscr{B}_j \mathscr{L}_j\right)^{-1} x_0\right] \\ &\quad = -\left[\mathscr{B}_i^* \Omega_i + (\mathscr{I} + \mathscr{B}_i^* \Omega_i \mathscr{B}_i)\mathscr{L}_i\left(\mathscr{I} - \sum_{1}^{n} \mathscr{B}_j \mathscr{L}_j\right)^{-1}\right] x_0 \\ &\quad = -\left[\mathscr{B}_i^* \Omega_i\left(\mathscr{I} - \sum_{1}^{n} \mathscr{B}_j \mathscr{L}_j\right) + (\mathscr{I} + \mathscr{B}_i^* \Omega_i \mathscr{B}_i)\mathscr{L}_i\right] \\ &\qquad \cdot \left(\mathscr{I} - \sum_{1}^{n} \mathscr{B}_j \mathscr{L}_j\right)^{-1} x_0 \\ &\quad = -\left[\mathscr{L}_i + \mathscr{B}_i^* \Omega_i\left(\mathscr{I} - \sum_{j \neq i}^{n} \mathscr{B}_j \mathscr{L}_j\right)\right]\left(\mathscr{I} - \sum_{1}^{n} \mathscr{B}_j \mathscr{L}_j\right)^{-1} x_0. \end{aligned} \tag{1.14}$$

Substitution of (1.6) into (1.14) and application of (1.10) gives

$$(\mathscr{I} + \mathscr{B}_i^*\Omega_i\mathscr{B}_i)\left[\hat{u}_i - \mathscr{L}_i\left(\mathscr{I} - \sum_1^n \mathscr{B}_j\mathscr{L}_j\right)^{-1} x_0\right] = 0. \tag{1.15}$$

The equation obtained upon cancellation of the nonsingular operator factor in (1.15) shows that $(\mathscr{L}_1, \mathscr{L}_2, \cdots, \mathscr{L}_n)$ meets the definition of an equilibrium given by (1.7)–(1.9).

To prove the converse let $\mathscr{L} = (\mathscr{L}_1, \mathscr{L}_2, \cdots, \mathscr{L}_n)$ satisfy (1.7)–(1.9). Since (1.9) holds for all $u_i \in H_i$ it holds in particular for the unique u_i which minimizes the right-hand cost term of (1.9). Using formula (1.5) we can easily compute the minimizing control to be the control given by formula (1.11). With (1.5) we may rewrite (1.9) as

$$|x_0 + \mathscr{B}_i\hat{u}_i|^2_{\Omega_i} - |x_0 + \mathscr{B}_i u_i|^2_{\Omega_i} + |\hat{u}_i|^2 - |u_i|^2 \leqq 0.$$

Then by substituting out u_i with (1.11) and $\hat{u}_i = \mathscr{L}_i(\mathscr{I} - \sum_1^n \mathscr{B}_j\mathscr{L}_j)^{-1}x_0$ which is the solution of (1.7)–(1.8), a rather long calculation based upon the expansion of the two differences of squares transforms the previous inequality into

$$\left|\left[\mathscr{L}_i\left(\mathscr{I} - \sum_1^n \mathscr{B}_j\mathscr{L}_j\right)^{-1} + (\mathscr{I} + \mathscr{B}_i^*\Omega_i\mathscr{B}_i)^{-1}\mathscr{B}_i^*\Omega_i\right]x_0\right|^2_{(\mathscr{I} + \mathscr{B}_i^\dagger\Omega_i\mathscr{B}_i)} \leqq 0.$$

Of course this implies that

$$\left[\mathscr{L}_i\left(\mathscr{I} - \sum_1^n \mathscr{B}_j\mathscr{L}_j\right)^{-1} + (\mathscr{I} + \mathscr{B}_i^*\Omega_i\mathscr{B}_i)^{-1}\mathscr{B}_i^*\Omega_i\right]x_0 = 0.$$

Elimination of Ω_i by (1.6) from this equation leads directly to (1.10). But $\mathscr{L}$ was an arbitrary equilibrium. This completes the proof that the equilibrium points of the game coincide with the solutions of (1.10).

Remark. We note (1.10) says that in order for (1.1) to be in an equilibrium feedback configuration each $\mathscr{L}_i$ must act upon the closed loop output to x_0 exactly as the adjoint of P_i's transfer function composed with $-\mathscr{W}_i$.

Thus the quest for an equilibrium leads to the problem of solving (1.10) for $\mathscr{L} = (\mathscr{L}_1, \mathscr{L}_2, \cdots, \mathscr{L}_n)$. The following fixed-point theorem [5, p. 216] will be useful in deriving sufficient conditions for the existence of a solution.

THEOREM 1.2. *Let F be a Banach space, V an open ball in F of center y_0 and radius γ. Let v be a mapping of V into F such that $|v(y_1) - v(y_2)| \leqq k|y_1 - y_2|$ for any pair of points y_1, y_2 of V, where k is a constant such that $0 \leqq k < 1$. Then, if $|v(y_0) - y_0| < \gamma(1 - k)$, there is one and only one point $z \in V$ such that $z = v(z)$.*

Notation. To prevent any confusion arising from the variety of norms to be dealt with, a word about our notation seems in order. Norms of vectors induced by inner products on R^m, H and H_i will be denoted as $|y|$, $|x|$ and $|u_i|$, etc. The norms of matrices and in general operators between Hilbert spaces will be denoted by double bars; e.g. $\|\mathscr{B}_i\| = \sup_{|u_i|=1} |\mathscr{B}_i u_i|$. However, for sums of spaces and operators we use norms such as $|z| = \sum_1^m |z_k|$ and $\|\mathscr{B}\| = \sum_1^n \|\mathscr{B}_i\|$ for $z = \sum_1^m \oplus z_k \in \sum_1^m \oplus H$ or $\mathscr{B} = \sum_1^n \oplus \mathscr{B}_i$, an operator from $\sum_1^n \oplus H$ into $\sum_1^n \oplus H_i$.

In some applications, (1.10) can be solved by setting the first operator factor equal to zero and solving

$$\mathscr{L}_i + \mathscr{B}_i^*\left(\mathscr{I} - \sum_{j\neq i}^{n} \mathscr{B}_j\mathscr{L}_j\right)^{-1*} \mathscr{W}_i = 0, \tag{1.16}$$

$i = 1, 2, \cdots, n$. Static games in which all operators are matrices often fall into this category. Other applications, e.g., differential games, will require that all factors in (1.10) be retained.

THEOREM 1.3. *A ball* $\|\mathscr{L}\| < \lambda$ *in* $\sum_1^n \oplus L_i$ *for which* $\mathscr{B}_i^*(\mathscr{I} - \sum_{j\neq i}^n \mathscr{B}_j\mathscr{L}_j)^{-1*} \cdot \mathscr{W}_i \in L_i$, $i = 1, 2, \cdots, n$, *contains a solution to* (1.16) *and hence to* (1.10) *if*

$$\|\mathscr{B}\|\,\|\mathscr{W}\|^{1/2} < (1 - \|\mathscr{B}\|\lambda)\left(1 - \frac{\|\mathscr{B}\|\,\|\mathscr{W}\|}{\lambda}\right)^{1/2}. \tag{1.17}$$

Proof. Consider $v(\mathscr{L}) = (v_1, v_2, \cdots, v_n)(\mathscr{L})$, where

$$v_i(\mathscr{L}) = -\mathscr{B}_i^*\left(\mathscr{I} - \sum_{j\neq i}^{n} \mathscr{B}_j\mathscr{L}_j\right)^{-1*} \mathscr{W}_i. \tag{1.18}$$

For $\mathscr{L}^1, \mathscr{L}^2$ in the ball $\|\mathscr{L}\| < \lambda$,

$$\begin{aligned} &v_i(\mathscr{L}^1) - v_i(\mathscr{L}^2) \\ &\quad = -\mathscr{B}_i^*\left(\mathscr{I} - \sum_{j\neq i}^{n} \mathscr{B}_j\mathscr{L}_j^1\right)^{-1*} \sum_{j\neq i}^{n} (\mathscr{L}_j^1 - \mathscr{L}_j^2)\mathscr{B}_j^*\left(\mathscr{I} - \sum_{j\neq i}^{n} \mathscr{B}_j\mathscr{L}_j^2\right)^{-1*} \mathscr{W}_i. \end{aligned} \tag{1.19}$$

Inequality (1.17) implies that $\|\mathscr{B}\|\lambda < 1$ which in turn we use to estimate (1.19) as

$$\begin{aligned} \|v_i(\mathscr{L}^1) - v_i(\mathscr{L}^2)\| &\leqq \frac{\|\mathscr{B}_i\|\,\|\mathscr{W}_i\| \sum_1^n \|\mathscr{L}_j^1 - \mathscr{L}_j^2\|\,\|\mathscr{B}_j\|}{(1 - \sum_1^n \|\mathscr{B}_j\|\,\|\mathscr{L}_j^1\|)(1 - \sum_1^n \|\mathscr{B}_j\|\,\|\mathscr{L}_j^2\|)} \\ &\leqq \frac{\|\mathscr{B}_i\|\,\|\mathscr{W}_i\|\,\|\mathscr{B}\|\,\|\mathscr{L}^1 - \mathscr{L}^2\|}{(1 - \|\mathscr{B}\|\lambda)^2}. \end{aligned} \tag{1.20}$$

Summing (1.20) we have

$$\begin{aligned} \|v(\mathscr{L}^1) - v(\mathscr{L}^2)\| &\leqq \textstyle\sum_1^n \|\mathscr{B}_i\|\,\|\mathscr{W}_i\|\,\|\mathscr{B}\|\,\|\mathscr{L}^1 - \mathscr{L}^2\|/(1 - \|\mathscr{B}\|\lambda)^2 \\ &\leqq \frac{\|\mathscr{B}\|^2\|\mathscr{W}\|}{(1 - \|\mathscr{B}\|\lambda)^2}\|\mathscr{L}^1 - \mathscr{L}^2\| = k\|\mathscr{L}^1 - \mathscr{L}^2\|, \end{aligned} \tag{1.21}$$

where we note from (1.17) that

$$k = \frac{\|\mathscr{B}\|^2\|\mathscr{W}\|}{(1 - \|\mathscr{B}\|\lambda)^2} < 1. \tag{1.22}$$

Thus we have shown

$$\|v(\mathscr{L}^1) - v(\mathscr{L}^2)\| \leqq k\|\mathscr{L}^1 - \mathscr{L}^2\| \tag{1.23}$$

for $\|\mathscr{L}^k\| < \lambda$, $k = 1, 2$. Computing $v_i(0) = -\mathscr{B}_i^*\mathscr{W}_i$ we find $\|v_i(0)\| \leqq \|\mathscr{B}_i\|\,\|\mathscr{W}_i\|$, and hence

$$\|v(0)\| \leqq \|\mathscr{B}\|\,\|\mathscr{W}\|. \tag{1.24}$$

Observe that (1.17) can be rewritten as

$$\|\mathscr{B}\|\,\|\mathscr{W}\| < \lambda\left[1 - \frac{\|\mathscr{B}\|^2\|\mathscr{W}\|}{(1 - \|\mathscr{B}\|\lambda)^2}\right]. \tag{1.25}$$

Combining (1.22), (1.24) and (1.25) shows

$$\|v(0) - 0\| \leqq \lambda(1 - k). \tag{1.26}$$

With (1.23) and (1.26) established, the proof can now be completed by direct application of Theorem 1.2.

Remark. We add that Theorem 1.2 further implies that the solution satisfying the conditions of Theorem 1.3 is unique. However, in general, the solutions to (1.10) and (1.16) are not unique. For example, the scalar system $\mathscr{B}_1 = \mathscr{W}_1 = 1$, $\mathscr{B}_2 = \mathscr{W}_2 = 2$ gives rise to equations (1.16),

$$l_1 + \frac{1}{1 - 2l_2} = 0, \qquad l_2 + \frac{4}{1 - l_1} = 0$$

having two solutions

$$l_1 = 4 \pm \sqrt{17}, \qquad l_2 = -3/2 \pm \sqrt{17}/2.$$

Looking at (1.16) we would expect to find a solution near $\mathscr{L} = 0$ if $\|\mathscr{B}\|$ or $\|\mathscr{W}\|$ is small. This is indeed verified by Theorem 1.3 since (1.17) will then have a solution $\lambda > 0$. Theorem 1.5 likewise represents an application of Theorem 1.2 with $y_0 = 0$. If we have prior knowledge about the location of a solution near a point other than the origin, then more appropriate results would be obtained using $y_0 \neq 0$.

LEMMA 1.4. *The operator function* $v(\mathscr{B}, \mathscr{L}) = (v_1, v_2, \cdots, v_n)(\mathscr{B}, \mathscr{L})$ *defined on the region* $\|\mathscr{B}\|\,\|\mathscr{L}\| < 1$ *by*

$$v_i(\mathscr{B}, \mathscr{L}) = -\left[\mathscr{L}_i \sum_1^n \mathscr{B}_j\mathscr{L}_j + \mathscr{B}_i^*\left(\mathscr{I} - \sum_{j\neq i}^n \mathscr{B}_j\mathscr{L}_j\right)^{-1*}\mathscr{W}_i\right]\left(\mathscr{I} - \sum_1^n \mathscr{B}_j\mathscr{L}_j\right)^{-1}$$

satisfies the estimate

$$\|v(\mathscr{B}, \mathscr{L}^1) - v(\mathscr{B}, \mathscr{L}^2)\| < \alpha\|\mathscr{B}\|\,\|\mathscr{L}^1 - \mathscr{L}^2\| \tag{1.27}$$

for $\|\mathscr{B}\| \leqq \beta$, $\|\mathscr{L}^1\| \leqq \lambda$, $\|\mathscr{L}^2\| < \lambda$, *where* β, λ *are any positive numbers such that* $\beta\lambda < 1$ *and* α *is the constant*

$$\alpha = \frac{3\lambda + 2\|\mathscr{W}\|\beta}{(1 - \beta\lambda)^3}.$$

Proof. As a notational convenience let $\mathscr{R}_k = (\mathscr{I} - \sum_1^n \mathscr{B}_j\mathscr{L}_j^k)^{-1}$ and $\mathscr{R}_{i,k} = (\mathscr{I} - \sum_{j\neq i}^n \mathscr{B}_j\mathscr{L}_j^k)^{-1}$ for an operator $\mathscr{L}^k \in L$, $i = 1, 2, \cdots, n$. The estimate

$$\left\|\left(\mathscr{I} - \sum_{j\neq i}^n \mathscr{B}_j\mathscr{L}_j\right)^{-1}\right\| \leqq \left(1 - \sum_{j\neq i}^n \|\mathscr{B}_j\|\,\|\mathscr{L}_j\|\right)^{-1}$$

$$\leqq \left(1 - \sum_1^n \|\mathscr{B}_j\|\,\|\mathscr{L}_j\|\right)^{-1} \leqq (1 - \|\mathscr{B}\|\,\|\mathscr{L}\|)^{-1}$$

implies that v is defined as an operator function on the region $\|\mathscr{B}\|\,\|\mathscr{L}\| < 1$, and in view of our assumptions that $\|\mathscr{B}\| \leqq \beta$, $\|\mathscr{L}^1\| \leqq \lambda$, $\|\mathscr{L}^2\| \leqq \lambda$, $\beta\lambda < 1$ leads to the bounds

$$\|\mathscr{R}_{i,k}\| \leqq (1 - \beta\lambda)^{-1}, \qquad \|\mathscr{R}_k\| \leqq (1 - \beta\lambda)^{-1} \tag{1.28}$$

for $k = 1, 2$; $i = 1, 2, \cdots, n$.

Recalling the definitions of v_i, $\mathscr{R}_{i,k}$, $\mathscr{R}_k$ we add and subtract to get

$$\begin{aligned} v_i(\mathscr{B}, \mathscr{L}^1) - v_i(\mathscr{B}, \mathscr{L}^2) = &\left[\mathscr{L}_i^2 \sum_1^n \mathscr{B}_j \mathscr{L}_j^2 + \mathscr{B}_i^* \mathscr{R}_{i,2}^* \mathscr{W}_i\right]\mathscr{R}_2 \\ &- \left[\mathscr{L}_i^2 \sum_1^n \mathscr{B}_j \mathscr{L}_j^2 + \mathscr{B}_i^* \mathscr{R}_{i,2}^* \mathscr{W}_i\right]\mathscr{R}_1 \\ &+ \left[\mathscr{L}_i^2 \sum_1^n \mathscr{B}_j \mathscr{L}_j^2 + \mathscr{B}_i^* \mathscr{R}_{i,2}^* \mathscr{W}_i\right]\mathscr{R}_1 \\ &- \left[\mathscr{L}_i^1 \sum_1^n \mathscr{B}_j \mathscr{L}_j^1 + \mathscr{B}_i^* \mathscr{R}_{i,1}^* \mathscr{W}_i\right]\mathscr{R}_1 \\ = &\left[\mathscr{L}_i^2 \sum_1^n \mathscr{B}_j \mathscr{L}_j^2 + \mathscr{B}_i^* \mathscr{R}_{i,2}^* \mathscr{W}_i\right][\mathscr{R}_2 - \mathscr{R}_1] \\ &+ \left[(\mathscr{L}_i^2 - \mathscr{L}_i^1) \sum_1^n \mathscr{B}_j \mathscr{L}_j^2 + \mathscr{L}_i^1 \sum_1^n \mathscr{B}_j (\mathscr{L}_j^2 - \mathscr{L}_j^1)\right]\mathscr{R}_1 \\ &+ \mathscr{B}_i^* [\mathscr{R}_{i,2}^* - \mathscr{R}_{i,1}^*] \mathscr{W}_i \mathscr{R}_1 . \end{aligned} \tag{1.29}$$

Rewriting the differences, we have

$$\mathscr{R}_2 - \mathscr{R}_1 = \mathscr{R}_2 \sum_1^n \mathscr{B}_j (\mathscr{L}_j^2 - \mathscr{L}_j^1) \mathscr{R}_1 , \tag{1.30}$$

$$\mathscr{R}_{i,2}^* - \mathscr{R}_{i,1}^* = \mathscr{R}_{i,2}^* \sum_{j \neq i}^n (\mathscr{L}_j^2 - \mathscr{L}_j^1) \mathscr{B}_j^* \mathscr{R}_{i,1}^* \tag{1.31}$$

in (1.29), and using (1.28) gives the upper estimate of $\|v_i(\mathscr{B}, \mathscr{L}^1) - v_i(\mathscr{B}, \mathscr{L}^2)\|$,

$$\begin{aligned} &\left[\|\mathscr{L}_i^2\| \beta\lambda + \|\mathscr{B}_i\| \left(\frac{1}{1 - \beta\lambda}\right) \|\mathscr{W}\|\right] \frac{\|\mathscr{B}\|\,\|\mathscr{L}^1 - \mathscr{L}^2\|}{(1 - \beta\lambda)^3} \\ &+ [\|\mathscr{L}_i^2 - \mathscr{L}_i^1\|\,\|\mathscr{B}\| \lambda + \|\mathscr{L}_i^1\|\,\|\mathscr{B}\|\,\|\mathscr{L}^2 - \mathscr{L}^1\|] \left(\frac{1}{1 - \beta\lambda}\right) \\ &+ \|\mathscr{B}_i\| \left(\frac{1}{1 - \beta\lambda}\right)^3 \|\mathscr{B}\|\,\|\mathscr{L}^2 - \mathscr{L}^1\|\,\|\mathscr{W}_i\| . \end{aligned} \tag{1.32}$$

But $\beta\lambda < 1$ and hence $1/(1 - \beta\lambda) > 1$. Consequently (1.32) produces the summed

upper estimate of $\|v(\mathscr{B}, \mathscr{L}^1) - v(\mathscr{B}, \mathscr{L}^2)\|$,

$$\left[\frac{\lambda(1-\beta\lambda)+\beta\|\mathscr{W}\|+2\lambda+\beta\|\mathscr{W}\|}{(1-\beta\lambda)^3}\right]\|\mathscr{B}\|\,\|\mathscr{L}^1-\mathscr{L}^2\| \leqq \left[\frac{3\lambda+2\|\mathscr{W}\|\beta}{(1-\beta\lambda)^3}\right]\|\mathscr{B}\|\,\|\mathscr{L}^1-\mathscr{L}^2\| \tag{1.33}$$

which completes the proof.

With each vector $Z=\sum_1^m \oplus z_k$ in $\sum_1^m \oplus H$ (m any positive integer) we associate a linear transformation $\mathscr{T}_Z(\cdot): L \to \sum_{i=1}^n \oplus \sum_{k=1}^m \oplus H_i$ given by

$$\mathscr{T}_Z(\mathscr{L}) = \sum_1^n \oplus \mathscr{T}_Z^i(\mathscr{L}) \quad \text{with } \mathscr{T}_Z^i(\mathscr{L}) = \sum_{k=1}^m \oplus \mathscr{L}_i z_k. \tag{1.34}$$

In the next theorem we again consider the operator function

$$v(\mathscr{B}, \mathscr{L}) = \sum_1^n \oplus v_i(\mathscr{B}, \mathscr{L}), \tag{1.35}$$

where

$$v_i(\mathscr{B}, \mathscr{L}) = -\left[\mathscr{L}_i \sum_1^n \mathscr{B}_j\mathscr{L}_j + \mathscr{B}_i^*\left(\mathscr{I} - \sum_{j\neq i}^n \mathscr{B}_j\mathscr{L}_j\right)^{-1*}\mathscr{W}_i\right]\left(\mathscr{I} - \sum_1^n \mathscr{B}_j\mathscr{L}_j\right)^{-1}.$$

THEOREM 1.5. *Suppose $\mathscr{T}_Z(\cdot)$ is a homeomorphism whose range $\mathscr{T}_Z(L)$ contains the vectors $v(\mathscr{B}, \mathscr{L})Z$ for all $\mathscr{L}$ in the ball $\|\mathscr{L}\| < \lambda$. If $\|\mathscr{B}\| < \beta$ where both*

$$\beta\lambda < 1 \tag{1.36}$$

and

$$\beta|Z|\,\|\mathscr{T}_Z^{-1}\|\{(1-\beta\lambda)^3\|\mathscr{W}\| + \lambda[3\lambda + 2\|\mathscr{W}\|\beta]\} < \lambda(1-\beta\lambda)^3, \tag{1.37}$$

then there exists a solution $\mathscr{L}$ in L to (1.10) *with $\|\mathscr{L}\| < \lambda$, holding for all $x_0 \in$ span $\{z_1, z_2, \cdots, z_m\}$.*

Proof. By adding and subtracting the term $\mathscr{L}_i$ and simplifying, we can rewrite (1.10) in the equivalent form

$$\mathscr{L}_i x_0 = v_i(\mathscr{B}, \mathscr{L})x_0. \tag{1.38}$$

Since we are assuming that $x_0 \in$ span $\{z_1, z_2, \cdots, z_m\}$, $x_0 = \sum_1^m c_k z_k$ for scalars c_k. Therefore it certainly is sufficient to solve

$$\sum_{k=1}^m \oplus \mathscr{L}_i z_k = \sum_{k=1}^m \oplus v_i(\mathscr{B}, \mathscr{L})z_k, \qquad i = 1, 2, \cdots, n. \tag{1.39}$$

Written in terms of $\mathscr{T}_Z(\cdot)$ defined by (1.34) this system is the same as the equation

$$\mathscr{T}_Z(\mathscr{L}) = \sum_{i=1}^n \oplus \sum_{k=1}^m \oplus v_i(\mathscr{B}, \mathscr{L})z_k = \sum_{i=1}^n \oplus v_i(\mathscr{B}, \mathscr{L}) \sum_{k=1}^m \oplus z_k = v(\mathscr{B}, \mathscr{L})Z. \tag{1.40}$$

From (1.34) we observe that the linear transformation $\mathscr{T}_Z(\cdot)$ is automatically

bounded with $\|\mathscr{T}_Z\| \leqq |Z|$. Since $\mathscr{T}_Z(\cdot)$ is assumed to be a homeomorphism, $\mathscr{T}_Z^{-1}$ is likewise a bounded linear transformation on its domain $\mathscr{T}_Z(L)$. Hence to prove the theorem we may deal with the equation

$$\mathscr{L} = \mathscr{T}_Z^{-1} v(\mathscr{B}, \mathscr{L}) Z \stackrel{\text{def}}{=} \tilde{v}(\mathscr{B}, \mathscr{L}). \tag{1.41}$$

With Lemma 1.4 we estimate

$$\begin{aligned} \|\tilde{v}(\mathscr{B}, \mathscr{L}^1) - \tilde{v}(\mathscr{B}, \mathscr{L}^2)\| &\leqq \|\mathscr{T}_Z^{-1}\| \, |Z| \, \|v(\mathscr{B}, \mathscr{L}^1) - v(\mathscr{B}, \mathscr{L}^2)\| \\ &\leqq \|\mathscr{T}_Z^{-1}\| \, |Z| \beta \left[\frac{3\lambda + 2\|\mathscr{W}\|\beta}{(1 - \beta\lambda)^3}\right] \|\mathscr{L}^1 - \mathscr{L}^2\| \\ &= k\|\mathscr{L}^1 - \mathscr{L}^2\|, \end{aligned} \tag{1.42}$$

with $k = \|\mathscr{T}_Z^{-1}\| \, |Z| \beta[(3\lambda + 2\|\mathscr{W}\|\beta)/(1 - \beta\lambda)^3]$ holding for $\|\mathscr{B}\| < \beta$, $\|\mathscr{L}^1\| < \lambda$, $\|\mathscr{L}^2\| < \lambda$ and $\beta\lambda < 1$. Inequalities (1.36)–(1.37) imply that $k < 1$. Using (1.41) and (1.35) we have

$$\begin{aligned} \|\tilde{v}(\mathscr{B}, 0) - 0\| &= \|\mathscr{T}_Z^{-1} v(\mathscr{B}, 0) Z\| \\ &\leqq \|\mathscr{T}_Z^{-1}\| \, |Z| \sum_1^n \|v_i(\mathscr{B}, 0)\| \leqq \|\mathscr{T}_Z^{-1}\| \, |Z| \sum_1^n \|\mathscr{B}_i^* \mathscr{W}_i\| \\ &\leqq \|\mathscr{T}_Z^{-1}\| \, |Z| \, \|\mathscr{W}\| \beta \leqq \lambda(1 - k), \end{aligned} \tag{1.43}$$

where the last inequality follows from (1.36) and division of (1.37) by $(1 - \beta\lambda)^3$. The remainder of the proof is a consequence of Theorem 1.2, which, in view of (1.42)–(1.43), can be applied to $\tilde{v}$.

Remark. Note that both (1.36) and (1.37) could be satisfied if β could independently be made small. This fact will be exploited in establishing the "local playability" of dynamic games.

2. Application to dynamic games. We are now interested in applying the results of the previous section to more concrete problems. As an example we choose a particular differential game. In the one player case it reduces to the optimal regulator problem which has been studied extensively in control theory [6], [7], [8] and has had useful applications in control engineering.

2.1. A linear-quadratic differential game. The data of the problem is given in the form of real matrix-valued functions $A(t)$, $B_i(t)$, $W_i(t) \geqq 0$, $\hat{W}_i \geqq 0$, $U_i(t) > 0$, and real, finite-dimensional, vector-valued functions $\bar{y}_i(t)$, $\hat{y}_i$, $i = 1, 2, \cdots, n$, all defined and continuous on a finite interval $t_0 \leqq t \leqq t_1$.

The dynamics of the game are given by the linear differential equation in $R^m \times [t_0, t_1]$,

$$\frac{dy}{dt} = A(t)y + \sum_1^n B_j(t)u_j, \tag{2.1}$$

where u_i is the m_i-dimensional control variable under the command of the ith player P_i. P_i is allowed to play any real, Borel measurable control command $u_i = u_i(t)$ satisfying

$$\int_{t_0}^{t_1} |u_i(t)|^2 \, dt < \infty. \tag{2.2}$$

With the initial state of the game $y(t_0) = y_0$ fixed, any such n-tuple of control commands $u(t) = (u_1, u_2, \cdots, u_n)(t)$ is called an *open loop strategy*. An open loop strategy together with the unique absolutely continuous solution of (2.1) that it generates will be called a *play* of the game.

Player P_i is also allowed to generate control commands by means of *feedback controls*. In the particular game that we are considering we allow $u_i = \mathscr{L}_i(t)y$, where $\mathscr{L}_i(t)$ is any real, continuous $m_i \times m$ matrix function. Hence P_i automatically generates control commands $u_i(t) = \mathscr{L}_i(t)y(t)$ in response to the action of the other players via $y(t)$, the solution to

$$\frac{dy}{dt} = [A(t) + B_i(t)\mathscr{L}_i(t)]y + \sum_{j \neq i}^{n} B_j(t)u_j(t). \tag{2.3}$$

In a similar manner any number of players can choose to play feedback control simultaneously.

P_i determines his cost incurred in a play of the game by a functional of the form

$$\mathscr{C}_i = \int_{t_0}^{t_1} [|y(t) - \bar{y}_i(t)|^2_{W_i(t)} + |u_i(t)|^2_{U_i(t)}]\, dt + |y(t_1) - \hat{y}_i|^2_{\hat{W}_i} \tag{2.4}$$

which he would like to minimize.

2.2. The differential game as a game in Hilbert space. In order to cast this game into the framework of § 1 we consider the Hilbert spaces

$$L_j^2 = \left\{ f(\cdot):[t_0, t_1] \to R^j, \int_{t_0}^{t_1} |f(t)|^2\, dt < \infty \right\} \tag{2.5}$$

and rewrite (2.1) using the variations of parameters formula

$$y(t) = \phi_0(t, t_0)y_0 + \sum_{1}^{n} \int_{t_0}^{t} \phi_0(t, \tau)B_j(\tau)u_j(\tau)\, d\tau, \tag{2.6}$$

where the fundamental matrix $\phi_0(t, \tau)$ is defined as the solution to

$$\frac{\partial \phi_0(t, \tau)}{\partial t} = A(t)\phi_0(t, \tau), \qquad \phi_0(\tau, \tau) = I_m. \tag{2.7}$$

We select

$$H = L_m^2 \oplus R^m, \qquad H_i = L^2_{m_i}, \qquad i = 1, 2, \cdots, n, \tag{2.8}$$

and define vectors x, x_0, $\bar{x}_i$, u_i by

$$\begin{aligned} x &= \begin{pmatrix} y(t) \\ y(t_1) \end{pmatrix} \in H, \qquad x_0 = \begin{pmatrix} \phi_0(t, t_0)y_0 \\ \phi_0(t_1, t_0)y_0 \end{pmatrix} \in H, \\ \bar{x}_i &= \begin{pmatrix} \bar{y}_i(t) \\ \hat{y}_i \end{pmatrix} \in H, \qquad u_i = u_i(t) \in H_i, \end{aligned} \tag{2.9}$$

and operators $\mathscr{B}_i : H_i \to H$ by

$$(\mathscr{B}_i u_i)(t) = \begin{pmatrix} \int_{t_0}^{t} \phi_0(t,\tau) B_i(\tau) u_i(\tau)\, d\tau \\ \int_{t_0}^{t_1} \phi_0(t_1,\tau) B_i(\tau) u_i(\tau)\, d\tau \end{pmatrix}, \tag{2.10}$$

$\mathscr{W}_i : H \to H$ by

$$(\mathscr{W}_i x)(t) = \begin{pmatrix} W_i(t) y(t) \\ \hat{W}_i y(t_1) \end{pmatrix}, \tag{2.11}$$

and finally $\mathscr{U}_i : H_i \to H_i$ by

$$(\mathscr{U}_i u_i)(t) = U_i(t) u_i(t), \qquad i = 1, 2, \cdots, n. \tag{2.12}$$

To complete the identification of the differential game with the abstract model of § 1 we must single out the L_i, the Banach spaces of operator feedbacks. Since the game allows feedback controls of the type $u_i = \mathscr{L}_i(t)y$ it is clear we must select the linear space of matrix operators corresponding to multiplication by $m_i \times 2m$ matrices of the type $[\mathscr{L}_i(t) \vdots \mathbf{0}]$ with $\mathscr{L}_i(t)$ of size $m_i \times m$ and continuous and $\mathbf{0}$ a zero matrix. For efficiency in notation we simply write $\mathscr{L}_i(t)$ and for later reference

$$L_i = \{\mathscr{L}_i(t) | m_i \times m, \text{continuous}\}, \tag{2.13}$$

$i = 1, 2, \cdots, n$. For the purpose of turning L_i into a complete normed linear space we consider the norms

$$\|\mathscr{L}_i\| = \sup_{|x|=1} |\mathscr{L}_i x| = \sup_{|x|=1} \left(\int_{t_0}^{t_1} |\mathscr{L}_i(t)x(t)|^2\, dt \right)^{1/2}$$

called the *uniform operator norm* and

$$\|\mathscr{L}_i\|' = \sup_{[t_0,t_1]} \|\mathscr{L}_i^*(t)\mathscr{L}_i(t)\|^{1/2}$$

called the *uniform norm.*

LEMMA 2.1. *The uniform norm and uniform operator norm agree on the space of matrix operators* (2.13) *and make L_i into a Banach space.*

Proof. It will be sufficient to compute the norms of the operators in L_i. Let $0 \neq \mathscr{L}_i \in L_i$, $x \in L_m^2$, and being careful not to confuse the various vector and matrix norms involved we compute

$$\begin{aligned} |\mathscr{L}_i x|^2 &= \int_{t_0}^{t_1} |\mathscr{L}_i(t)x(t)|^2\, dt \leqq \int_{t_0}^{t_1} \|\mathscr{L}_i^*(t)\mathscr{L}_i(t)\|\, |x(t)|^2\, dt \\ &\leqq \sup_{[t_0,t_1]} \|\mathscr{L}_i^*(t)\mathscr{L}_i(t)\| \int_{t_0}^{t_1} |x(t)|^2\, dt \\ &= \sup_{[t_0,t_1]} \|\mathscr{L}_i^*(t)\mathscr{L}_i(t)\|\, |x|^2 \end{aligned} \tag{2.14}$$

which shows that $\|\mathscr{L}_i\| \leqq \sup_{[t_0,t_1]} \|\mathscr{L}_i^*(t)\mathscr{L}_i(t)\|^{1/2}$.

Since $\|\mathscr{L}_i^*(t)\mathscr{L}_i(t)\|$ is a continuous function of t on $[t_0, t_1]$ there exists a $t_* \in [t_0, t_1]$ at which it takes on its maximum. Since $\mathscr{L}_i^*(t_*)\mathscr{L}_i(t_*)$ is a positive semidefinite symmetric matrix, there exists an eigenvector $x_* \in R^m$ with $|x_*| = 1$ such that

$$\|\mathscr{L}_i^*(t_*)\mathscr{L}_i(t_*)\| = x_* \cdot \mathscr{L}_i^*(t_*)\mathscr{L}_i(t_*)x_*. \tag{2.15}$$

We first examine the case where $t_* \in (t_0, t_1)$. Consider the sequence in L_m^2,

$$x_n(t) = \sqrt{n}\chi_n(t)x_*, \qquad n = 1, 2, \cdots, \tag{2.16}$$

in which $\chi_n(t)$ is the characteristic function of the interval $[t_*, t_* + 1/n] \cap [t_0, t_1]$. For this sequence we compute

$$|x_n|^2 = \int_{t_0}^{t_1} |x_n(t)|^2 \, dt = n \int_{t_0}^{t_1} \chi_n(t) \, dt \leqq 1 \tag{2.17}$$

for $n = 1, 2, \cdots$ (with equality holding for all large n) and

$$\begin{aligned} |\mathscr{L}_i x_n|^2 &= \int_{t_0}^{t_1} |\mathscr{L}_i(t)x_n(t)|^2 \, dt = n \int_{t_0}^{t_1} \chi_n(t) |\mathscr{L}_i(t)x_*|^2 \, dt \\ &= \int_{t^*}^{t^* + 1/n} |\mathscr{L}_i(t)x_*|^2 \, dt/(1/n) \end{aligned} \tag{2.18}$$

for all large n. Using (2.18) and the continuity of $|\mathscr{L}_i(t)x_*|$ we conclude

$$\lim_{n \to \infty} |\mathscr{L}_i x_n|^2 = |\mathscr{L}_i(t_*)x_*|^2. \tag{2.19}$$

From (2.17), (2.19) and (2.15) we get

$$\|\mathscr{L}_i\|^2 = \sup_{|x| \leqq 1} |\mathscr{L}_i x|^2 \geqq \lim_{n \to \infty} |\mathscr{L}_i x_n|^2 = \|\mathscr{L}_i^*(t_*)\mathscr{L}_i(t_*)\| \tag{2.20}$$

which shows

$$\|\mathscr{L}_i\| \geqq \sup_{[t_0, t_1]} \|\mathscr{L}_i^*(t)\mathscr{L}_i(t)\|^{1/2} \tag{2.21}$$

and hence that

$$\|\mathscr{L}_i\| = \sup_{[t_0, t_1]} \|\mathscr{L}_i^*(t)\mathscr{L}_i(t)\|^{1/2}. \tag{2.22}$$

A similar argument leads to (2.22) in the cases $t_* = t_0$ and $t_* = t_1$. But (2.22) shows that the norm generating the uniform operator topology on L_i agrees with the norm generating the uniform metric topology on L_i. The completeness of the latter space is well known. This concludes the proof.

With definitions (2.5)–(2.13) we may write (2.4) as

$$\mathscr{C}_i = |x - \bar{x}_i|_{\mathscr{W}_i}^2 + |u_i|_{\mathscr{U}_i}^2 \tag{2.23}$$

and have shown that the differential game given by (2.1), (2.4) and (2.13) is an example of the abstract model (1.1), (1.2′) presented in § 1. Therefore all the results developed there apply in particular to the differential game. Our attention is now focused upon the application of those results.

2.3. Local playability of the game. Let $[t_0, t_1]$ denote any compact subinterval of some underlying interval $[t_\alpha, t_\beta]$ on which $A(t)$, $B_i(t)$, $W_i(t) \geqq 0$, $U_i(t) > 0$ and $\bar{y}_i(t)$ are defined and continuous. The differential game is called *playable* on $[t_0, t_1]$ if there exists a system of equilibrium feedback matrices in the corresponding $L = \sum_1^n \oplus L_i$ for that interval. (Recall definitions (1.9) and (2.13).) If there exists a number $\Delta > 0$ such that the game is playable on every subinterval $[t_0, t_1]$ of length $t_1 - t_0 \leqq \Delta$, then we say the game is *locally playable* on $[t_\alpha, t_\beta]$. The next lemma will play an important role in proving the differential game is always locally playable.

LEMMA 2.2. *For $\mathscr{B}_i$ as defined by* (2.10),

$$\|\mathscr{B}\| \leqq \gamma(t_1 - t_0)^{1/2}, \tag{2.24}$$

where γ is a number giving the bound

$$\|\phi_0(t,\tau)B_i(\tau)\| \leqq \frac{\gamma}{(t_\beta - t_\alpha)^{1/2} + 1} \tag{2.25}$$

for $t_\alpha \leqq t, \tau \leqq t_\beta$, $i = 1, 2, \cdots, n$.

Proof. Being careful not to confuse the norms involved we estimate

$$\begin{aligned}
|\mathscr{B}_i u_i| &= \left[\int_{t_0}^{t_1} \left| \int_{t_0}^{t} \phi_0(t,\tau)B_i(\tau)u_i(\tau)\,d\tau \right|^2 dt\right]^{1/2} \\
&\quad + \left| \int_{t_0}^{t_1} \phi_0(t_1,\tau)B_i(\tau)u_i(\tau)\,d\tau \right| \\
&\leqq \left[\int_{t_0}^{t_1} \left(\int_{t_0}^{t} \|\phi_0(t,\tau)B_i(\tau)\|\,|u_i(\tau)|\,d\tau \right)^2 dt\right]^{1/2} \\
&\quad + \int_{t_0}^{t_1} \|\phi_0(t_1,\tau)B_i(\tau)\|\,|u_i(\tau)|\,d\tau \\
&\leqq \frac{\gamma}{(t_\beta - t_\alpha)^{1/2} + 1}[(t_1 - t_0)^{1/2} + 1] \int_{t_0}^{t_1} |u_i(\tau)|\,d\tau \\
&\leqq \gamma \int_{t_0}^{t_1} |u_i(\tau)|\,d\tau \leqq \gamma(t_1 - t_0)^{1/2}|u_i|.
\end{aligned} \tag{2.26}$$

Summing we get $|\mathscr{B}u| \leqq \gamma(t_1 - t_0)^{1/2}|u|$ and hence

$$\|\mathscr{B}\| = \sup_{|u|=1} |\mathscr{B}u| \leqq \gamma(t_1 - t_0)^{1/2} \tag{2.27}$$

which completes the proof.

It will be useful now to compute some of the operator functions which arose in the general theory of § 1. To accomplish this end let $(\mathscr{L}_1, \mathscr{L}_2, \cdots, \mathscr{L}_n)(t)$ be a system of feedback matrices from $L = \sum_1^n \oplus L_i$. Substituting $u_j = \mathscr{L}_j(t)y$ $(j \neq i)$ into (2.1) produces the equation

$$\dot{y}_i = \left[A(t) + \sum_{j \neq i}^{n} B_j(t)\mathscr{L}_j(t)\right] y_i + B_i(t)u_i \tag{2.28}$$

whose solution can be written as

$$y_i(t) = \phi_i(t, t_0)y_0 + \int_{t_0}^{t} \phi_i(t, \tau)B_i(\tau)u_i(\tau)\, d\tau, \tag{2.29}$$

where the fundamental matrix $\phi_i(t, \tau)$ is defined as the solution to

$$\frac{\partial \phi_i(t, \tau)}{\partial t} = \left[A(t) + \sum_{j \neq i}^{n} B_j(t)\mathscr{L}_j(t)\right]\phi_i(t, \tau) \tag{2.30}$$

with $\phi_i(t, t) = I_m$ for all $t, \tau \in [t_\alpha, t_\beta]$, $i = 1, 2, \cdots, n$. Equation (2.29) shows that

$$\left[(\mathscr{I} - \sum_{j \neq i}^{n} \mathscr{B}_j\mathscr{L}_j)^{-1}B_iu_i\right](t) = \begin{pmatrix} \int_{t_0}^{t} \phi_i(t, \tau)B_i(\tau)u_i(\tau)\, d\tau \\ \\ \int_{t_0}^{t_1} \phi_i(t_1, \tau)B_i(\tau)u_i(\tau)\, d\tau \end{pmatrix}, \tag{2.31}$$

$i = 1, 2, \cdots, n$, and consequently

$$\left[\mathscr{B}_i^*\left(\mathscr{I} - \sum_{j \neq i}^{n} \mathscr{B}_j\mathscr{L}_j\right)^{-1*} x\right](t) = \int_{t}^{t_1} B_i^*(t)\phi_i^*(\tau, t)y(\tau)\, d\tau + B_i^*(t)\phi_i^*(t, t_1)y(t_1) \tag{2.32}$$

for all

$$x(t) = \begin{pmatrix} y(t) \\ y(t_1) \end{pmatrix} \in H = L_m^2(t_0, t_1) \oplus R^m.$$

Recall that this operator function, as well as the following one, plays an important role in (1.10).

When all players play feedback, $u_j = \mathscr{L}_j(t)y$, in the integral form of

$$\dot{y} = A(t)y + \sum_{1}^{n} B_j(t)u_j, \tag{2.33}$$

$$y(t) = \phi_0(t, t_0)y_0 + \sum_{1}^{n} \int_{t_0}^{t} \phi_0(t, \tau)B_j(\tau)u_j(\tau)\, d\tau, \tag{2.34}$$

we have the solution $y(t) = \phi(t, t_0)y_0$ in which $\phi(t, \tau)$ is the fundamental matrix solution to

$$\frac{\partial \phi(t, \tau)}{\partial t} = \left[A(t) + \sum_{1}^{n} B_j(t)\mathscr{L}_j(t)\right]\phi(t, \tau) \tag{2.35}$$

with $\phi(t, t) = I_m$ for all $t, \tau \in [t_\alpha, t_\beta]$. This shows that

$$\left(\mathscr{I} - \sum_{1}^{n} \mathscr{B}_j\mathscr{L}_j\right)^{-1} x_0 = \begin{pmatrix} \phi(t, t_0)y_0 \\ \phi(t_1, t_0)y_0 \end{pmatrix}, \tag{2.36}$$

where

$$x_0 = \begin{pmatrix} \phi_0(t, t_0)y_0 \\ \phi_0(t_1, t_0)y_0 \end{pmatrix} \in H.$$

Recall that we made the identification $\mathscr{L}_i \leftrightarrow$ multiplication by $[\mathscr{L}_i(t) \,\vdots\, \mathbf{0}]$. In terms of the concrete operator functions given by (2.32) and (2.36), equation (1.10) for the system of equilibrium feedback matrices is

$$\mathscr{L}_i(t)\phi(t,t_0)y_0 + \int_t^{t_1} B_i^*(t)\phi_i^*(\tau,t)W_i(\tau)\phi(\tau,t_0)y_0\,d\tau + B_i^*(t)\phi_i^*(t,t_1)\,\hat{W}_i\phi(t_1,t_0)y_0 = 0, \tag{2.37}$$

$i = 1, 2, \cdots, n$. Note that (2.37) is a complicated functional equation in $\mathscr{L}$ since ϕ_i and ϕ depend upon $\mathscr{L}$ according to their defining equation (2.30) and (2.35). Nevertheless, as stated in the next theorem, a solution does exist—at least on short time intervals.

THEOREM 2.3. *The differential game with dynamics* (2.1), *costs* (2.4) *and feedback spaces* (2.13) *is locally playable on* $[t_\alpha, t_\beta]$ *with equilibrium feedback matrices which are independent of the initial state of the game. Moreover, such state-independent matrices are unique.*

Proof. We assume that the appropriate preliminary change of control variables has been made so that we can take $U_i(t) = I$, $i = 1, 2, \cdots, n$, and for simplicity treat the case in which $\bar{y}_i(t) = \hat{y}_i = 0$. The proof will follow from Theorem 1.5 and Lemma 2.2.

Recall that we are dealing with the Hilbert spaces

$$H = L_m^2(t_0,t_1) \oplus R^m, \qquad H_i = L_{m_i}^2(t_0,t_1)$$

and feedback operators identified as matrices $[\mathscr{L}_i(t) \,\vdots\, \mathbf{0}] \in L_i$. The "abstract initial state" is given as

$$x_0 = \begin{pmatrix} \phi_0(t,t_0)y_0 \\ \phi_0(t_1,t_0)y_0 \end{pmatrix} = \left(\begin{array}{c|c} \phi_0(t,t_0) & \mathbf{0} \\ \hline \mathbf{0} & \phi_0(t_1,t_0) \end{array}\right)\begin{pmatrix} y_0 \\ y_0 \end{pmatrix}. \tag{2.38}$$

In anticipation of applying Theorem 1.5 we consider $Z = \sum_1^{2m} \oplus z_k$, where z_k is the kth column of the $2m \times 2m$ matrix in (2.38). Suppose that we identify Z with that matrix. Then define $\mathscr{T}_Z$ in (1.32) according to the right matrix multiplication

$$\mathscr{T}_Z^i(\mathscr{L})(t) = [\mathscr{L}_i(t) \,\vdots\, \mathbf{0}]Z(t), \qquad i = 1, 2, \cdots, n. \tag{2.39}$$

Since $Z(t)$ is continuous and nonsingular it follows that $\mathscr{T}_Z(\cdot)$ is a homeomorphism and $\mathscr{T}_Z(L) = L$. In fact, $\mathscr{T}_Z^{-1} = \mathscr{T}_{Z^{-1}}$. Furthermore, from (2.32) and (2.36) we see that

$$v(\mathscr{B}, L)Z \subset L = \mathscr{T}_Z(L). \tag{2.40}$$

Remembering that the subinterval $[t_0, t_1]$ is arbitrary in $[t_\alpha, t_\beta]$ we conclude from continuity that $|Z|$ and $\|\mathscr{T}_Z^{-1}\|$ are bounded when regarded as functions of $t_0, t_1 \in [t_\alpha, t_\beta]$. Since it can easily be shown that

$$\|\mathscr{W}_i\| \leqq \sup_{[t_0,t_1]} \|W_i(t)\| + \|\hat{W}_i\|, \tag{2.41}$$

it follows that $\|\mathscr{W}\|$ is also bounded as a function of $t_0, t_1 \in [t_\alpha, t_\beta]$. Hence for any $\lambda > 0$ the hypothesis of Theorem 1.5 can be satisfied by first selecting $\beta > 0$

sufficiently small so that (1.36)–(1.37) hold for all $t_0, t_1 \in [t_\alpha, t_\beta]$. In view of Lemma 2.2 we can then maintain $\|\mathscr{B}\| < \beta$ by restricting $\Delta = t_1 - t_0 < \beta^2/\gamma^2$. The state-independence is clear since $x_0 \in \text{span}\{z_1, z_2, \cdots, z_{2m}\}$ for all $y_0 \in R^m$. The uniqueness argument will rely partly upon the calculations of the next section and is treated in the concluding comments.

Remark. By carrying out the last steps in the proof of Theorem 2.3 numerically one can compute the length of intervals $\Delta = t_1 - t_0$ over which the game with constraint $\|\mathscr{L}\| < \lambda$ is playable. In doing this the inequalities corresponding to (1.36)–(1.37) which must be solved can be somewhat simplified by, for example, replacing the first occurrence of the term $(1 - \beta\lambda)^3$ by 1, etc.

2.4. Calculation of the equilibrium feedback. The results up to this point show (2.37), the concrete form of (1.10), has a solution in L which moreover is a solution to

$$\mathscr{L}_i(t)\phi(t, t_0) + \int_t^{t_1} B_i^*(t)\phi_i^*(\tau, t)W_i(\tau)\phi(\tau, t_0)\, d\tau + B_i^*(t)\phi_i^*(t_1, t)\,\hat{W}_i\phi(t_1, t_0) = 0, \tag{2.42}$$

$i = 1, 2, \cdots, n$, for $\Delta = t_1 - t_0$ sufficiently small. We rewrite this equation as

$$\mathscr{L}_i(t) = -B_i^*(t)Q_i(t), \tag{2.43}$$

where

$$Q_i(t) = \int_t^{t_1} \phi_i^*(\tau, t)W_i(\tau)\phi(\tau, t_0)\phi^{-1}(t, t_0)\, d\tau + \phi_i^*(t_1, t)\,\hat{W}_i\phi(t_1, t_0)\phi^{-1}(t, t_0),$$

$i = 1, 2, \cdots, n$. Using elementary properties of fundamental matrices one can easily show that the matrix product $\phi(\tau, t_0)\phi^{-1}(t, t_0)$ is independent of t_0, and hence, as already anticipated in our notation, upon setting $t_0 = t$ we have

$$Q_i(t) = \int_t^{t_1} \phi_i^*(\tau, t)W_i(\tau)\phi(\tau, t)\, d\tau + \phi_i^*(t_1, t)\,\hat{W}_i\phi(t_1, t), \tag{2.44}$$

$i = 1, 2, \cdots, n$. Since ϕ_i and ϕ satisfy (2.30) and (2.35), respectively, from elementary properties of fundamental matrices it follows that

$$\frac{\partial\phi_i(\tau, t)}{\partial t} = -\phi_i(\tau, t)\left[A(t) + \sum_{j \neq i}^{n} B_j(t)\mathscr{L}_j(t)\right], \tag{2.45}$$

$$\frac{\partial\phi(\tau, t)}{\partial t} = -\phi(\tau, t)\left[A(t) + \sum_{1}^{n} B_j(t)\mathscr{L}_j(t)\right], \tag{2.46}$$

$i = 1, 2, \cdots, n$. By differentiating (2.44) with the aid of (2.45)–(2.46) and using (2.43) to substitute out the $\mathscr{L}_i(t)$, a long but otherwise simple calculation shows

that the $Q_i(t)$ satisfy the system of ordinary matrix differential equations

$$
\begin{aligned}
-\dot{Q}_i &= W_i(t) + A^*(t)Q_i + Q_iA(t) - Q_iB_i(t)B_i^*(t)Q_i \\
&\quad -Q_i\left[\sum_{j\neq i}^{n} B_j(t)B_j^*(t)Q_j\right] - \left[\sum_{j\neq i}^{n} Q_j^*B_j(t)B_j^*(t)\right]Q_i, \\
Q_i(t_1) &= \hat{W}_i, \qquad i = 1, 2, \cdots, n.
\end{aligned} \tag{2.47}
$$

The following uniqueness argument further shows $Q_i(t) = Q_i^*(t)$. The differential equation obtained from (2.47) by replacing the term Q_j^* by Q_j has a unique and hence symmetric solution. But that solution also satisfies (2.47), and hence again by uniqueness the solutions are the same.

Now we show that the $Q_i(t)$ are positive definite for $t < t_1$. Let $\mathscr{L}(t)$ be the equilibrium solution to (2.42). Then (2.43) and (2.47) hold for all $t < t_1$ and near t_1. Letting $y(t)$ be the equilibrium response to $\mathscr{L}(t)$; i.e., the solution to

$$
\dot{y} = \left[A(t) + \sum_{1}^{n} B_j(t)\mathscr{L}_j(t)\right]y \tag{2.48}
$$

with $y(t_0) = y_0$, using (2.43) and (2.47) one may easily verify

$$
\begin{aligned}
\frac{d}{dt}[y(t)\cdot Q_i(t)y(t)] &= -[|y(t)|^2_{W_i(t)} + |B_i^*Q_iy(t)|^2] \\
&= -[|y(t)|^2_{W_i(t)} + |\mathscr{L}_i(t)y(t)|^2].
\end{aligned} \tag{2.49}
$$

Integration of (2.49) gives

$$
|y_0|^2_{Q_i(t_0)} = \int_{t_0}^{t_1} [|y(t)|^2_{W_i(t)} + |\mathscr{L}_i(t)y(t)|^2]\,dt + |y(t_1)|^2_{\hat{W}_i} \tag{2.50}
$$

which shows that the quadratic form in y_0 associated with $Q_i(t)$ gives P_i's cost in the game with all players playing equilibrium feedback. Equation (2.50) also shows that each $Q_i(t)$ is positive definite for $t < t_1$.

2.5. Concluding comments. We have shown that the equilibrium feedback matrices whose existence was established by a fixed-point argument can be computed by integrating a system of differential equations, (2.47), for the cost matrices and then doing the matrix multiplication (2.43). One can easily show that the matrix feedbacks computed in this manner provide an equilibrium solution on the entire left-maximal interval of existence for (2.47).

The uniqueness of these initial-state-independent equilibrium matrices can now be argued. Of course equilibria in general are not unique. For example, if we choose $x_0 = 0$ in (1.10), then any system of operators for which the equation is defined constitute equilibrium solutions. Moreover, the example following Theorem 1.3 shows that even initial-state-independent equilibria in general are not unique. However Theorem 1.1 guarantees that any equilibrium must be a solution to (1.10) and in particular any initial-state-independent equilibrium solution to the differential game must be a solution to (2.42). But the technique used in computing the solution to that equation shows that it can have but one solution. That is, the uniqueness of the solution of (2.42) is a consequence of the

uniqueness of solutions to the differential equations of the type (2.47). These observations complete the proof of uniqueness stated in Theorem 2.3. We remark that when one considers the complexity of the operator calculations (that one in principle could do for the concrete operators arising in the differential game) which were necessary in proving the only if part of Theorem 1.1, it is understandable that previous investigators in differential games were not successful in treating the uniqueness question. Clearly the abstract approach of this paper makes a definite contribution here.

The existence of the solution to the differential game could have been established directly by construction of the solution from the solution of (2.47). However the fixed-point approach provides a means for estimating the interval of playability. From (2.42) we see that $\mathscr{L}_i(t_1) = -B_i^*(t_1)\hat{W}_i$. Hence if $\hat{W}_i = 0$, $i = 1, 2, \cdots, n$, then we can obtain an estimate using (1.37) as described in the remark following Theorem 2.3. In general one would expect to get improved estimates by utilizing this prior information as pointed out in the remark following Theorem 1.3. That is, the fixed-point theory should be applied about the point $-B_i^*(t_1)\hat{W}_i$.

The equilibrium feedbacks computed above agree with the solution derived in [4] using a Hamilton–Jacobi approach. In the one player game; i.e., $n = 1$, equation (2.47) reduces to the Kalman–Riccati equation

$$-\dot{Q} = W(t) + A^*(t)Q + QA(t) - QB(t)B^*(t)Q, \tag{2.51}$$

$Q(t_1) = \hat{W}$, studied in [6] in connection with the optimal regulator problem. This equation has a solution on every interval $[t_0, t_1] \subseteq [t_\alpha, t_\beta]$.

In [1] an equilibrium was found in $\sum_1^n \oplus H_i$, the space of open loop controls, and then synthesized by a matrix feedback system in L generated in the same way as the equilibrium in L but with an equation lacking the last term in (2.47). In general the two feedback systems are not the same and produce different control responses.

When the differential game is playable on every subinterval of $[t_\alpha, t_\beta]$ we call the game *globally playable.* A sufficient condition for global playability is

$$W_i(t) = \lambda_i W_0(t), \qquad \hat{W}_i = \lambda_i \hat{W}_0 \tag{2.52}$$

and

$$B_i(t)B_i^*(t) = \frac{1}{\lambda_i} S_0(t), \tag{2.53}$$

$i = 1, 2, \cdots, n$, for some choice of scalars $\lambda_i > 0$ and matrices $W_0(t)$, $\hat{W}_0$, $S_0(t)$. This result can be checked by using (2.52)–(2.53) to show $Q_i(t) = \lambda_i Q(t)$ is the required solution of (2.47) if we take $Q(t)$ to be the solution of (2.51) corresponding to the selection $W(t) = W_0(t)$, $\hat{W} = \hat{W}_0$ and $B(t)$ any symmetric square root of $(2n - 1)S_0(t)$.

As a final remark we point out that the results of this paper can be applied to the study of equilibrium solutions in other feedback spaces and bear upon games with dynamics described by other types of equations; e.g., integral equations.

REFERENCES

[1] D. L. LUKES AND D. L. RUSSELL, *A global theory for linear quadratic differential games*, J. Math. Anal. Appl., 33 (1971).

[2] P. P. VARAIYA, *N-person non-zero sum differential games with linear dynamics*, this Journal, 8 (1970), pp. 441–449.

[3] D. L. LUKES AND A. STRAUSS, *Two countable systems of differential inequalities with applications to the stability of linear quadratic differential games*, Tech. Summary Rep. 1003, Mathematics Research Center, Univ. of Wisconsin, Madison, 1969.

[4] J. H. CASE, *Equilibrium points of N-person differential games*, Tech. Rep. 1967-1 (Thesis), Dept of Industrial Engineering, Univ. of Michigan, Ann Arbor, 1967.

[5] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[6] D. L. LUKES, *Stabilizability and optimal control*, Funkcial. Ekvac., 11 (1968), pp. 39–50.

[7] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[8] D. L. LUKES, *Optimal regulation of nonlinear dynamical systems*, this Journal, 7 (1969), pp. 75–100.

SIAM J. CONTROL
Vol. 9, No. 2, May 1971

CONTROLLABILITY OF LINEAR OSCILLATORY SYSTEMS USING POSITIVE CONTROLS*

STEPHEN H. SAPERSTONE† AND JAMES A. YORKE‡

Abstract. A linear autonomous control process is considered where the null control is an extreme point of the restraint set Ω. In the event that $\Omega = [0, 1]$ (hence, scalar control) necessary and sufficient conditions are given so that the reachable set from the origin (in phase space) contains the origin as an interior point. For vector-valued controls with each component in $[0, 1]$, sufficient conditions are given so that the reachable set from the origin of a nonlinear autonomous control process contains the origin as an interior point.

1. Introduction and example. The results in this paper are best motivated by the following example of controllability. Can the motion of a simple pendulum be brought to rest in a finite time by the application of a unit force acting only in one direction? In terms of a differential equation, the problem can be stated as follows: Let $u:[0, \infty) \to \{0, 1\}$ be measurable. The linearized equation of motion of the pendulum is $\ddot{\theta} + k\theta = u$, where k is a positive constant of the motion and θ is the angular displacement of the pendulum from the vertical. Is there an open neighborhood V of the origin (in θ, $\dot{\theta}$ space) such that if $(\theta(0), \dot{\theta}(0)) \in V$, then there exist a controller $u(\cdot)$ and some $T > 0$ such that $(\theta(T), \dot{\theta}(T)) = (0, 0)$? The answer is yes. In fact all values of $\theta(0)$ and $\dot{\theta}(0)$ can be steered to the origin in finite time.

Similar questions of controllability may be asked of more complicated oscillatory (i.e., without real eigenvalues) systems having many degrees of freedom. In § 7 we shall analyze a double pendulum. Related questions for nonlinear systems are discussed in Corollary 6.3. A more general approach will be given in [5].

2. Definitions and statement of main result. Consider the system of linear differential equations in real Euclidean d-space R^d,

$$\dot{x} = Ax + bu, \tag{2.1}$$

where A is a real constant $d \times d$ matrix and x, b are real column d-vectors. Let R^+ denote the nonnegative real numbers, and Ω, the restraint set, be a nonempty interval in R. Let U_Ω be the set of all bounded measurable functions $u(\cdot)$, where $u:R^+ \to \Omega$. For each $u \in U_\Omega$ let $x(t) = x(t; u(\cdot))$ be the unique absolutely continuous function satisfying (2.1) such that $x(0; u(\cdot)) = 0$. Define the *reachable set* (*attainable set*) *at time* $t \geqq 0$ by

$$K_\Omega^+(t) = \{x(t; u(\cdot)) : u \in U_\Omega\}$$

and the *reachable set* by

$$K_\Omega^+ = \bigcup_{t \geqq 0} K_\Omega^+(t).$$

* Received by the editors March 23, 1970, and in revised form August 13, 1970.

† Center for Naval Analyses of the University of Rochester, Arlington, Virginia and University of Maryland, College Park, Maryland. Now at Department of Mathematics, Howard University, Washington, D.C. 20001. The work of this author was supported by a CNA Fellowship.

‡ Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, College Park, Maryland 20742. The work of this author was supported in part by the National Science Foundation under Grant NSF-GP-9347.

$K_\Omega^+(t)$ is convex and connected, and compact provided Ω is compact (see [4, p. 69]). We note that $x(t; u(\cdot))$ is given by

$$x(t; u(\cdot)) = \int_0^t e^{A(t-s)} bu(s)\, ds.$$

Let x' denote the transpose of the column vector $x \in R^d$. The *controllability matrix* of (2.1) is the $d \times d$ matrix whose columns are $b, Ab, A^2b, \cdots, A^{d-1}b$. We say that the origin can be *steered* (or *controlled*) to the point $x_0 \in R^d$ provided $x_0 \in K_\Omega^+(t_0)$ for some $t_0 < \infty$. The system (2.1) is called *locally controllable at the origin* if there exists $t_1 < \infty$ such that $K_\Omega^+(t_1)$ contains a neighborhood of the origin.

Let $\Omega = [0, 1]$. The main result we will prove is the following theorem.

THEOREM 2.1. *The system* (2.1) *with controls* $u(\cdot)$ *belonging to* $U_{[0,1]}$ *is locally controllable at the origin if and only if*

(i) *all eigenvalues of* A *have nonzero imaginary parts, and*

(ii) *the controllability matrix for* (2.1) *has rank* d.

Remark 2.2. Observe in Theorem 2.1 that if d is odd, then the system (2.1) is not controllable since a matrix A of odd order must have at least one real eigenvalue. Likewise A must be nonsingular. Thus we are led to consider only systems of an "oscillatory" nature.

Remark 2.3. We note that (2.1) is locally controllable at the origin for the class U_Ω if and only if there exists a neighborhood V of the origin such that every point of V can be steered to $x = 0$ in finite time. In fact, observe that λ is an eigenvalue of A if and only if $-\lambda$ is an eigenvalue of $-A$ and the vectors $\{b, Ab, A^2b, \cdots, A^{d-1}b\}$ span the same space as $\{-b, Ab, -A^2b, \cdots, (-1)^d A^{d-1}b\}$. It follows that (2.1) is locally controllable at the origin if and only if

$$\dot{x} = -Ax - bu \tag{2.2}$$

is. But $u(t)$ steers the origin to x_0 along a solution of (2.1) on $[0, t_0]$ if and only if $u(t_0 - t)$ steers x_0 to the origin along a solution of (2.2) on $[0, t_0]$. Thus every point of $K_{[0,1]}^+(t)$ for any $t \geqq 0$ may be steered to the origin in time t if and only if conditions (i) and (ii) of Theorem 2.1 hold.

Theorem 2.1 is related to the following well-known result in controllability. We give a new proof of this result and will make use of parts of the proof in establishing Theorem 2.1. Let $\Omega = [-\varepsilon, \varepsilon]$.

THEOREM 2.4. *For the system* (2.1) *with controls* $u(\cdot)$ *belonging to* $U_{[-\varepsilon,\varepsilon]}$, $x = 0$ *is interior to* $K_{[-\varepsilon,\varepsilon]}^+$ *if and only if the controllability matrix for* (2.1) *has rank* d.

Proof. Let $L^r = \text{span}\,\{b, Ab, \cdots, A^{r-1}b\}$ for $r = 1, 2, \cdots$. Since dimension $L^r \leqq d$, we may choose α (where $0 \leqq \alpha \leqq d$) to be the smallest integer such that $A^\alpha b \in L^\alpha$. For any $y \in L^\alpha$ there exist $a_0, a_1, \cdots, a_{\alpha-1}$ such that $y = a_0 b + a_1 Ab + \cdots + a_{\alpha-1}A^{\alpha-1}b$. It follows that $Ay = a_0 Ab + a_1 A^2 b + \cdots + a_{\alpha-1}A^\alpha b$ is in L^α. Since $Ay \in L^\alpha$, we have $A^2 y \in L^\alpha$, etc. In particular, $A^r b \in L^\alpha$ for all r and so $L^\alpha = L^s$ if $s \geqq \alpha$. Also, for any scalar function $u(t)$,

$$Ay + bu(t) \in L^\alpha \quad \text{if } y \in L^\alpha. \tag{2.3}$$

From (2.3) we have that for any function $u(t)$, equation (2.1) (which is a differential equation on R^d) may also be considered as a differential equation on the subspace

L^α in that $A: L^\alpha \to L^\alpha$ and $bu \in L^\alpha$. Since (2.1) is a differential equation on L^α, $x(t, u(\cdot)) \in L^\alpha$ for all t. When (2.1) is considered as a differential equation on R^d, we must still have $x(t, u(\cdot)) \in L^\alpha$. Hence $K_\Omega^+(t) \subseteq L^\alpha$ for all $t \geqq 0$.

Denote by $x(t)$ the solution of (2.1) with controller $u(t) \equiv \varepsilon$ and satisfying $x(0) = 0$. Choose $T > 0$. Let L be the linear subspace of R^d generated by $x(t)$ for all $t \in [0, T]$; that is,

$$L = \text{span}\,\{x(t) : t \in [0, T]\}.$$

Then for t and $t + \tau$ in $[0, T]$, we have

$$\tau^{-1}[x(t + \tau) - x(t)] \in L.$$

It follows that $\dot{x}(t) \in L$ for all $t \in [0, T]$. Similarly $\tau^{-1}[\dot{x}(t + \tau) - \dot{x}(t)] \in L$. Arguing inductively we see that all derivatives of $x(\cdot)$ are in L. Setting $t = 0$, (2.1) implies

$$\dot{x}(0) = A0 + b\varepsilon = b\varepsilon.$$

Taking higher derivatives of both sides of (2.1) and evaluating at $t = 0$ we get, by induction,

$$x^{(2)}(0) = A\dot{x}(0) = Ab\varepsilon,$$

$$\cdots$$

$$x^{(r)}(0) = Ax^{(r-1)}(0) = A^{r-1}b\varepsilon.$$

Therefore $b, Ab, \cdots, A^r b \in L$ for all $r > 0$, so $L^\alpha \subseteq L$.

For $\tau \in [0, T]$, let $u_\tau(t) = \varepsilon$ for $t \in [T - \tau, T]$ and $u_\tau(t) = 0$ elsewhere. Then $x(T, u_\tau(\cdot)) = x(\tau)$; hence $x(t) \in K_\Omega^+(T)$ for $0 \leqq t \leqq T$. We have thus far shown that $\{x(t) : 0 \leqq t \leqq T\} \subseteq K_\Omega^+(T) \subseteq L^\alpha \subseteq L$. Taking spans we get $L \subseteq L^\alpha \subseteq L$, and hence, $L = L^\alpha$.

Choose $0 < t_1 < t_2 < \cdots < t_\alpha < T$ such that $B \overset{\text{def}}{=} \{x(t_i)\}_{i=1}^\alpha$ is a linearly independent set. Such a set exists because $L^\alpha = L$ has dimension α and by definition of L, $\{x(t) : 0 \leqq t \leqq T\}$ spans L. Since B is a basis for L^α, there exists for each $y \in L^\alpha$ a unique set $\{c_1, c_2, \cdots, c_\alpha\}$ such that $y = c_1 x(t_1) + \cdots + c_\alpha x(t_\alpha)$. Now define $\|y\|_\alpha = |c_1| + \cdots + |c_\alpha|$. ($\|y\|_\alpha$ is a norm on L^α because it is the sum of the absolute values of the coordinates c_i using the basis B.) Since $\Omega = [-\varepsilon, \varepsilon]$ is convex, it is easy to verify that $K_\Omega^+(T)$ is convex. Furthermore if $y \in K_\Omega^+(T)$, then $-y \in K_\Omega^+(T)$ since $x(T, u(\cdot)) = -x(T, -u(\cdot))$. Now $x(t_i) = x(T, u_{t_i}(\cdot)) \in K_\Omega^+(T)$. Therefore $\pm x(t_i) \in K_\Omega^+(T)$. So, $\{y : \|y\|_\alpha \leqq 1\} \subset K_\Omega^+(T)$. Hence $K_\Omega^+(T)$ contains a neighborhood of $x = 0$ in the topology of L^α. Therefore $K_\Omega^+(T) \subset L^\alpha$ contains a neighborhood of $x = 0$ in the topology of R^d if and only if the rank of $\{b, Ab, \cdots, A^{d-1}b\}$ is d.

Remark 2.5. Since $x(T, u(\cdot))$ is linear in u, the point

$$y = \sum_{i=1}^{\alpha} c_i x(t_i) = \sum_{i=1}^{\alpha} c_i x(T, u_{t_i}(\cdot)) = x\left(T, \sum_{i=1}^{\alpha} c_i u_{t_i}(\cdot)\right)$$

(hence any point in L^α) can be reached by a controller $u(\cdot) = \sum_{i=1}^\alpha c_i u_{t_i}(\cdot)$ which is piecewise constant and has only α switching times $t_1, t_2, \cdots, t_\alpha$. (Of course, $|u| \leqq \varepsilon$ does not necessarily hold.) These switching times do not depend upon the

point y. If the controllability matrix has rank d, then $L^{\alpha} = R^d$ and the number of switches α is d.

Though both Theorems 2.1 and 2.4 characterize controllability in terms of the range of the control functions allowed, the theorems differ significantly with respect to the time required to reach any desired point in K_{Ω}^{+}. In particular for any arbitrary small time $t_1 > 0$, $K_{[-\varepsilon,\varepsilon]}^{+}(t_1)$ contains a neighborhood of $x = 0$ (see [4, p. 83]). Such a claim is false for the case in Theorem 2.1 when $\Omega = [0, \delta]$ for any $\delta > 0$. We can see from the simple pendulum in § 1 that $0 \in \operatorname{int} K_{\{0,1\}}^{+}(t_1)$ only if $t_1 > \pi k^{-1/2}$. Physical considerations require that the time to bring the pendulum to rest from each initial position in a neighborhood of the origin be as much as half a period. This can be easily verified by using the Pontryagin maximum principle. That this is also true for $\Omega = [0, \delta], \delta > 0$, follows from the "bang-bang" principle [3] which we now state. Let $E\Omega$ be the set of extreme points of Ω.

THEOREM 2.6. *Suppose the range Ω of the control function is compact. Then for any $t \geqq 0$, $K_{\Omega}^{+}(t) = K_{E\Omega}^{+}(t)$ for the system* (2.1). *Furthermore, if t_0 is the minimal time required to steer the origin to x_0 for $u \in U_{\Omega}$, then t_0 is also minimal for $u \in U_{E\Omega}$.*

Thus the reachable set at any time t for the motion of the simple pendulum is unaffected if $\Omega = [0, 1]$ is replaced by $\Omega = \{0, 1\}$. By virtue of Theorem 2.6 the oscillatory systems we describe can now be controlled by "off–on" controls.

In the example of the simple pendulum, the eigenvalues are $\pm ik^{1/2}$. Furthermore we can show that $K_{\{0,1\}}^{+} = R^2$. That the controllability matrix had rank 2 just depended on the fact that $b \neq 0$. The crucial aspect of generalizing the example to Theorem 2.1 lies in the oscillatory nature of the flow. But we even have controllability when the free system (with $u \equiv 0$) is completely unstable. That is, the eigenvalues of A can have positive real parts. The following result is an immediate corollary.

COROLLARY 2.7. *Consider* (2.1) *for a 2-dimensional system with $\Omega = \{0, 1\}$. Suppose $b \neq 0$. Then the system is locally controllable at the origin if and only if no eigenvalue of A is real.*

3. Preliminaries and lemmas.

LEMMA 3.1. *If the matrix A of the system* (2.1) *has a real eigenvalue λ (possibly zero), then $x = 0$ belongs to $\partial K_{[0,1]}^{+}$.*

Proof. Let λ be a real eigenvalue of A. Then λ is also an eigenvalue of A' (transpose of A), and there is some $v \in R^d$ ($v \neq 0$) such that $A'v = \lambda v$. We may assume $v \cdot b \geqq 0$ (replacing v by $-v$ if necessary). Let $x(t)$ be a solution of (2.1) for $t \geqq 0$, with $x(0) = 0$ for some $u \in U_{[0,1]}$. Write $\rho(t) = v \cdot x(t)$. Then

$$\frac{d}{dt}\rho(t) = v \cdot \dot{x}(t) = v \cdot Ax(t) + v \cdot bu(t)$$

$$= \lambda\rho(t) + (v \cdot b)u(t).$$

Since $u(t) \geqq 0$ for all t, and $\rho(0) = 0$, it follows that $\rho(t) \geqq 0$. Furthermore, if $v \cdot b = 0$, then $\rho(t) \equiv 0$. That is, for all $y \in K_{[0,1]}^{+}$, y is in the half-space $\{z \in R^d : v \cdot z \geqq 0\}$ when $v \cdot b > 0$ (and y is in the hyperplane $v \cdot z = 0$ if $v \cdot b = 0$). Considering the control function $u(t) \equiv 0$, we see that $0 \in K_{[0,1]}^{+}$. Hence $0 \in \partial K_{[0,1]}^{+}$.

From Lemma 3.1 for local controllability with $\Omega = [0, 1]$, we see that A must be nonsingular and of even dimension since matrices of odd dimension have

at least one real eigenvalue. Thus we are led to consider systems of even order $d = 2n$, such that the eigenvalues of A have nonzero imaginary parts. In particular let the distinct eigenvalues of A be

$$\lambda_k = \alpha_k + i\beta_k, \quad \lambda_k^* = \alpha_k - i\beta_k, \quad \alpha_k, \beta_k \in R^1, \quad \beta_k \neq 0 \quad \text{for } k = 1, 2, \cdots, s,$$

with multiplicities $m_1, m_2, \cdots, m_s$. (* denotes complex conjugate.) Let η be any nonzero row vector in R^{2n}. We find a convenient representation and important property for the product $\eta e^{A\tau}b$, $\tau \in R^+$. (We assume the controllability matrix has rank $2n$.)

LEMMA 3.2. *$\eta e^{A\tau}b$ is not identically zero on R^+ and can be written*

$$\eta e^{A\tau}b = \tau^m e^{\alpha\tau}\left\{\mu(\tau) + \sum_{k=1}^{s} h_k \sin(\beta_k\tau + \delta_k)\right\} \tag{3.1}$$

for constants m, α, δ_k, h_k (where not all h_k are zero) and where $\mu(\tau) \to 0$ as $\tau \to \infty$, and the summation is not identically zero.

Proof. On the contrary, assume $\eta e^{A\tau}b$ is identically zero. Then $e^{A\tau}b$ lies in some proper subspace V of R^{2n} for all $\tau \geqq 0$. Now $e^{A\tau}b$ is the unique solution to the differential equation $\dot{x} = Ax$, $x(0) = b$. We saw in the proof of Theorem 2.4 that $\dot{x}(0), x^{(2)}(0), \cdots, x^{(2n-1)}(0)$ lie in V; hence $b, Ab, \cdots, A^{2n-1}b$ belong to V. Since the controllability matrix has rank $2n$, it follows that $V = R^{2n}$ and therefore $\eta e^{A\tau}b \not\equiv 0$.

There exists an invertible matrix P such that $J = PAP^{-1}$ has Jordan normal form (cf. [2, p. 76]). Then $\eta e^{A\tau}b$ can be written

$$\eta P^{-1} e^{J\tau} P b.$$

Consider the complex vector $e^{J\tau}Pb$. It is the solution to the complex system $\dot{y} = Jy$, $y(0) = Pb$ and is of the form

$$\sum_{\lambda} e^{\lambda\tau} \sum_{j=1}^{m_\lambda} P_{\lambda,j}(\tau) v_{\lambda,j},$$

where the first sum is taken over all eigenvalues λ of A, $P_{\lambda,j}(\tau)$ is a polynomial in τ which depends on λ and has order $m_\lambda - 1$ (where m_λ is the multiplicity of the eigenvalue λ), and $v_{\lambda,j}$ is a (complex) vector depending on λ. Replacing $e^{\lambda\tau}$ by $e^{\alpha\tau}(\cos\beta\tau + i\sin\beta\tau)$, we observe that since $\eta e^{A\tau}b$ is real, it has the form

$$\sum_{k=1}^{s} e^{\alpha_k\tau}[R_k(\tau)\cos\beta_k\tau + S_k(\tau)\sin\beta_k\tau], \tag{3.2}$$

where $R_k(\tau), S_k(\tau)$ for $k = 1, 2, \cdots, s$ are real polynomials in τ and are not all zero. If $\alpha = \max\{\alpha_1, \alpha_2, \cdots, \alpha_s\}$, we can factor $e^{\alpha\tau}$ out of (3.2) along with an appropriate power of τ, say τ^m, and obtain

$$\eta e^{A\tau}b = \tau^m e^{\alpha\tau}\left\{\mu(\tau) + \sum_{k=1}^{s} h_k \sin(\beta_k\tau + \delta_k)\right\},$$

where not all the $\{h_k\}_{k=1}^s$ are zero and $\mu(\tau) \to 0$ as $\tau \to \infty$.

4. Almost periodic functions. Let $\mathbb{C}$ denote the field of complex numbers. Recall a continuous function $f: R \to \mathbb{C}$ is *almost periodic* (a.p.) if for any $\varepsilon > 0$ there exists a number $L(\varepsilon) > 0$ with the property that any interval of length $L(\varepsilon)$ of R

contains at least one point ξ such that $|f(t+\xi) - f(t)| < \varepsilon$ for all $t \in R$. The number ξ is called an *ε-translation number of f*. We note the following properties about a.p. functions [1]. If f and g are a.p. functions, then $a_1 f + a_2 g$ $(a_1, a_2 \in \mathbb{C})$ and f^* (the complex conjugate of f) are a.p. Furthermore f is bounded and uniformly continuous on R. Also, the integral

$$M\{f\} \overset{\text{def}}{=} \lim_{T\to\infty} \frac{1}{T} \int_0^T f(t)\,dt$$

exists and is called the *mean value of f* (see [1, p. 12]).

LEMMA 4.1. *Suppose f is an a.p. function such that $M\{f\} = 0$ and*

$$f(t) \leqq |\mu(t)| \quad \textit{for all } t \in R^+, \tag{4.1}$$

where $|\mu(t)| \to 0$ as $t \to \infty$. Then $f \equiv 0$.

Proof. First we show that $f(t) \leqq 0$ for $t \in R$. Suppose the contrary, that there exists t_0 such that $f(t_0) = m > 0$. Then there exists $\delta > 0$ such that $f(t) > 2m/3$ for all $t \in (t_0 - \delta, t_0 + \delta)$. There exists a number $L(m/3)$ such that any interval of length $L(m/3)$ contains at least one number of the form $t_0 + \xi$, where $|f(t+\xi) - f(t)| < m/3$. Choose $L = L(m/3) > 2\delta$. Any interval of length L contains at least one of the intervals $(t_0 + \xi - \delta, t_0 + \xi)$ or $(t_0 + \xi, t_0 + \xi + \delta)$. In each of these intervals, which are the intervals $(t_0 - \delta, t_0)$, $(t_0, t_0 + \delta)$ translated by ξ, we have $f(t) > m/3$. But this contradicts the assumption (4.1). Hence we must have $f(t) \leqq 0$ for all $t \in R^+$, which reduces our lemma to a result in [1, p. 36]. This completes the proof.

Functions of the form $f(t) = \sum_{k=1}^N a_k e^{iw_k t}$, where $a_k \in \mathbb{C}$, $w_k \in R$, $k = 1, 2, \cdots, N$, are a.p. In particular Im (f) is a.p.

5. Proof of Theorem 2.1. We have already proved that a necessary condition for $x = 0$ to be interior to $K^+_{[0,1]}$ is that no eigenvalue of A is real (Lemma 3.1). That the controllability matrix have rank $d = 2n$ is also necessary for $x = 0$ to be interior to $K^+_{[0,1]}$ follows directly from Theorem 2.4. It only remains to prove the sufficiency of the conditions (i) and (ii).

So assume the controllability matrix has rank $2n$ and no eigenvalue of A is real. We claim that for large enough t, $x = 0$ is interior to $K^+_{[0,1]}(t)$. Suppose not; i.e., for all $t \geqq 0$, $x = 0$ belongs to $\partial K_{[0,1]}(t)$. Since $K^+_{[0,1]}(t)$ is convex for any $t \geqq 0$, we can find supporting hyperplanes to $K^+_{[0,1]}(t)$ at $x = 0$. Let $N(t)$ denote the set of all unit outer normals of such supporting hyperplanes to $K^+_{[0,1]}(t)$ at $x = 0$. Then for each $t \geqq 0$, $N(t) \neq \varnothing$, since we have assumed $0 \in K^+_{[0,1]}(t)$. $N(t)$ clearly is bounded. Moreover, it is closed. In fact, let $\{\eta_i\}$ be a sequence in $N(t)$ converging to $\eta \in R^{2n}$. Then for any $x(t) \in K^+_{[0,1]}(t)$, $\eta_i \cdot x(t) \leqq 0$. Passing to the limit we have $\eta \cdot x(t) \leqq 0$ and $\|\eta\| = 1$. Thus $\eta \in N(t)$, and $N(t)$ is a nonempty compact subset of R^{2n}. Furthermore, if $t_1 < t_2$, then $N(t_2) \subset N(t_1)$ since $K^+_{[0,1]}(t)$ increases with increasing t. Consequently, for any finite set of times $0 \leqq t_1 < t_2 < \cdots < t_k < \infty$, $\bigcap_{i=1}^k N(t_i) = N(t_k) \neq \varnothing$. Hence $\bigcap_{t \geqq 0} N(t) \neq \varnothing$ from the finite intersection property of compact sets. Then there exists a unit outer normal vector η_0, and hence a supporting hyperplane to $K^+_{[0,1]}(t)$ good for all $t \geqq 0$. In particular, for any

controller $u(\cdot) \in U_{[0,1]}$ and corresponding solution $x(t; u(\cdot))$, we have

$$\eta_0 \cdot x(t; u(\cdot)) = \int_0^t \eta_0 e^{A(t-\tau)} b u(\tau)\, d\tau \leqq 0$$

for all $t \geqq 0$. It follows by continuity and special choice of $u(\cdot)$ that for all $\tau \geqq 0$,

$$\varphi(\tau) \overset{\text{def}}{=} \eta_0 e^{A\tau} b \leqq 0.$$

According to Lemma 3.2, $\varphi(\tau)$ is not identically zero and

$$\varphi(\tau) = \tau^m e^{\alpha\tau} \left\{ \mu(\tau) + \sum_{k=1}^{s} h_k \sin(\beta_k \tau + \delta_k) \right\},$$

where $|\mu(\tau)| \to 0$ as $\tau \to \infty$, and the summation is not identically zero. Denote this sum by $v(\tau)$. Observe that $v(\tau)$ is an a.p. function with mean $M\{v\} = 0$. We now claim that there exists a $\tau^* \in R^+$ such that $v(\tau^*) > |\mu(\tau^*)|$. If, on the contrary, no such τ^* exists, $v(t) \leqq |\mu(\tau)|$ for all $t \in R^+$. From Lemma 4.1 we see that $v(\tau) \equiv 0$. But we know this to be false. Therefore we conclude there exists $\tau^* \in R^+$ such that $v(\tau^*) > |\mu(\tau^*)|$. Thus $\varphi(\tau^*) > 0$, a contradiction. So we conclude that $x = 0$ is interior to $K^+_{[0,1]}(t)$ for some $t < \infty$.

6. Extensions and generalizations. Theorem 2.1 has a generalization in the event that the vector b is replaced by a matrix B of order $d \times q$. Let Ω_0 denote the unit cube in R^q, $\Omega_0 = \{(u_1, u_2, \cdots, u_q) \in R^q : 0 \leqq u_i \leqq 1,\ i = 1, 2, \cdots, q\}$ and define

$$U_{\Omega_0} = \{u : R^+ \to \Omega_0, u \text{ measurable}\}.$$

Consider

$$\dot{x} = Ax + Bu, \tag{6.1}$$

$u \in U_{\Omega_0}$. Then $K^+_{\Omega_0}(t)$ is defined accordingly. Denote by C the controllability matrix of (6.1), where C is now the matrix of order $d \times dq$ whose columns are the columns of $B, AB, A^2B, \cdots, A^{d-1}B$. We note that for any nonsingular matrix T of order $d \times d$, $\operatorname{rank}(TC) = \operatorname{rank}(C)$. We can now give sufficient conditions for controllability of (6.1) at the origin.

THEOREM 6.1. *Suppose all the eigenvalues of A have nonzero imaginary parts and that the controllability matrix for* (6.1) *has rank d. Then $x = 0$ is interior to $K^+_{\Omega_0}(t_1)$ for some $t_1 < \infty$.*

Proof. We proceed as in the proof of Theorem 2.1 to establish a unit normal vector, η_0. Denote $\eta_0 e^{A\tau} B$ by $\varphi(\tau)$. If the columns of B are denoted by $b_1, b_2, \cdots, b_q$, then the matrix φ satisfies

$$\varphi(\tau) = [\eta_0 e^{A\tau} b_1, \eta_0 e^{A\tau} b_2, \cdots, \eta_0 e^{A\tau} b_q].$$

We claim that not all of the terms $\eta_0 e^{A\tau} b_j$ for $j = 1, 2, \cdots, q$ are identically zero. In fact, suppose they were all zero. Then for all $\tau \geqq 0$ and every $j = 1, 2, \cdots, q$, $e^{A\tau} b_j$ lies in some proper subspace of R^d. Denote this subspace by V. Now each $e^{A\tau} b_j$ is the unique solution to the differential equation

$$\dot{x} = Ax, \qquad x(0) = b_j.$$

As in the proof of Theorem 2.1 we show $b_j, Ab_j, \cdots, A^{d-1}b_j$ lie in V for each $j = 1, 2, \cdots, q$. It follows that $\eta_0 B = \eta_0 AB = \cdots = \eta_0 A^{d-1}B = 0$, thus contradicting the fact that the controllability matrix C has rank d. As in the case of Theorem 2.1 we can show that if $x = 0$ is not interior to $K^+_{\Omega_0}(t)$ for any $t \geqq 0$, then each nonzero component $\varphi_j(\tau) \overset{\text{def}}{=} \eta_0 e^{A\tau} b_j$ of $\varphi(\tau)$ is never positive on R^+. But each $\varphi_j(\tau)$ has the form (3.1), and eventually is positive. From this contradiction we conclude that $x = 0$ is interior to $K^+_{\Omega_0}(t)$ for some $t_1 < \infty$, thus completing the proof.

It is clear from the above proof that the conditions for local controllability at the origin with positive controls are too strong. In fact, the following simple example illustrates how we can have $x = 0$ interior to $K^+_{\Omega_0}$ in the event (6.1) is a scalar equation.

Example 6.2. Consider the scalar equation

$$\dot{x} = \lambda x + b_1 u_1 + b_2 u_2$$

with $0 \leqq u_1, u_2 \leqq 1$. It may be seen that $0 \in \operatorname{int} K^+_{\Omega_0}$ if and only if $b_1 b_2 < 0$.

Theorem 6.1 can be extended to nonlinear systems which admit a linearization of the form (6.1).

COROLLARY 6.3. *Consider the following system:*

$$\dot{x} = f(x, u), \tag{6.2}$$

where $f : R^d \times R^q \to R^d$, f *is* C' *in* $R^d \times R^q$ *and* $u \in U_{\Omega_0}$. *Define*

$$A = \left.\frac{\partial f}{\partial x}\right|_{(0,0)}, \qquad B = \left.\frac{\partial f}{\partial u}\right|_{(0,0)}.$$

Suppose:

(i) $f(0, 0) = 0$,

(ii) *all eigenvalues of* A *have nonzero imaginary parts,*

(iii) $\operatorname{rank}[B \vdots AB \vdots A^2B \vdots \cdots \vdots A^{d-1}B] = d$.

Then $x = 0$ *is interior to* $K^+_{\Omega_0}$.

Proof. The proof is exactly the same as used by Markus and Lee in [4, p. 366]. In fact, nowhere in their proof do they require controllers which assume both positive and negative values. They only require that the solution to the linearized form of (6.2) be controllable to any point in some neighborhood of $x = 0$ (starting from $x = 0$). But conditions (ii) and (iii) guarantee this.

7. The double pendulum. As an example of controllability with positive controls, we consider the linearized equations of motion of a double pendulum (Fig. 7.1). Each mass m_k is fixed at the end of a rigid weightless rod of length L_k, $k = 1, 2$. Both masses are nonzero. The system swings without friction about the pivot points P_0 and P_1, where P_0 is fixed to a rigid structure. A time-varying force, $u(t)$, $0 \leqq u(t) \leqq 1$, $t \in R^+$, is applied simultaneously at masses m_1 and m_2 as indicated in Fig. 7.1.

The linearized equations of motion of the system are given by (with gravitational constant $g = 1$)

$$m_1 L_1 \ddot{\theta}_1 + (m_1 + m_2)\theta_1 - m_2\theta_2 = u,$$

$$m_2 L_2 \ddot{\theta}_2 + (m_1 + m_2)\theta_2 - (m_1 + m_2)\theta_1 = (m_1/m_2 - 1)u.$$

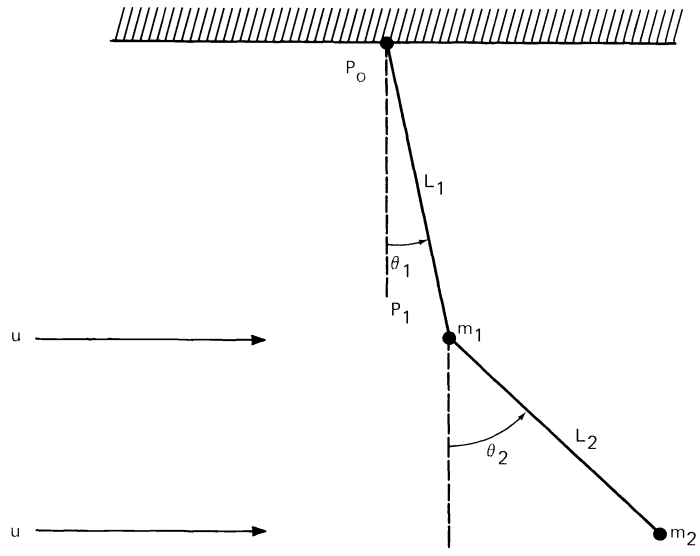


FIG. 7.1

Set $x_1 = \theta_1$, $x_2 = \dot{\theta}_1$, $x_3 = \theta_2$, $x_4 = \dot{\theta}_2$. In vector form we get $\dot{x} = Ax + bu$, where

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -k/L_1 & 0 & (k-1)/L_1 & 0 \\ 0 & 0 & 0 & 1 \\ k/L_2 & 0 & -k/L_2 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ (m_1 L_1)^{-1} \\ 0 \\ L_2^{-1}(m_2^{-1} - m_1^{-1}) \end{bmatrix}$$

and $k = 1 + m_2/m_1 > 1$.

The eigenvalues of A are pure imaginary. Denote them by $\pm i\omega_1, \pm i\omega_2$. We have (letting $\pm$ be $+$ for ω_1^2 and $-$ for ω_2^2)

$$\omega_1^2, \omega_2^2 = (2L_1L_2)^{-1}\{(L_1 + L_2)k \pm [(L_1 + L_2)^2k^2 - 4L_1L_2k]^{1/2}\}.$$

We take $\omega_1, \omega_2 > 0$. It is easy to show that $\omega_1^2 \neq \omega_2^2$. Define the matrix Q, whose jth row is given by

$$[(-1)^{j+1}\omega, 1, (-1)^{j+1}(1 - L_1\omega^2)/(L_1\omega), (1 - L_1\omega^2)/(L_1\omega^2)],$$

where $\omega = \omega_1$ for $j = 1, 2$ and $\omega = \omega_2$ for $j = 3, 4$. Under the coordinate transformation defined by Q, A becomes

$$\begin{bmatrix} 0 & \omega_1 & 0 & 0 \\ -\omega_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \omega_2 \\ 0 & 0 & -\omega_2 & 0 \end{bmatrix},$$

and b becomes $[b_1, b_1, b_2, b_2]$, where

$$b_j = (m_1 L_1)^{-1} + (1 - L_1\omega_j^2)(m_2^{-1} - m_1^{-1})/(L_1 L_2 \omega_j^2), \qquad j = 1, 2.$$

An easy computation shows that the determinant Δ of the controllability matrix is $4b_1^2 b_2^2 \omega_1 \omega_2 (\omega_1^2 - \omega_2^2)^2$. In general, $\Delta \neq 0$ except for certain critical values of the parameters L_1, L_2 and k. Under the assumption that $\Delta \neq 0$ we conclude that the reachable set for the double pendulum contains a neighborhood of the origin. To put it another way, for sufficiently small initial values of $(\theta_1(0), \dot{\theta}_1(0), \theta_2(0), \dot{\theta}_2(0))$, there exists an appropriate controller and a finite time $T \geqq 0$ such that $(\theta_1(T), \dot{\theta}_1(T), \theta_2(T), \dot{\theta}_2(T)) = (0, 0, 0, 0)$. Furthermore, one can show that if $T > \pi/\min(\omega_1, \omega_2)$, then $K^+_{[0,1]}(T)$ contains the origin as an interior point.

Acknowledgment. We would like to thank the referee who made many useful comments throughout the paper. In particular, his suggestions allowed us to simplify and clarify § 5 which was at best confusing.

REFERENCES

[1] A. S. BESICOVITCH, *Almost Periodic Functions*, Dover, New York, 1954.
[2] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
[3] R. CONTI, *Contribution to linear control theory*, J. Differential Equations, 1 (1965), pp. 427–445.
[4] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[5] J. A. YORKE, *The maximum principle and controllability of nonlinear equations*, to appear.

SIAM J. CONTROL
Vol. 9, No. 2, May 1971

ON THE CONVERGENCE OF APPROXIMATING SOLUTIONS FOR LINEAR DISTRIBUTED PARAMETER OPTIMAL CONTROL PROBLEMS*

HITOSHI SASAI† AND ETSUJRO SHIMEMURA‡

Abstract. Optimal control problems for distributed parameter systems, particularly systems described by partial differential equations, are often treated using mathematical function space techniques. As a result, the equations which define the optimal control are frequently obtained in abstract terms. Numerical solutions are obtained by approximating the abstract operations in a computationally feasible manner. In obtaining approximating solutions, the finite difference method is widely used.

After an approximate optimal control has been found, the question arises whether a sequence of these approximating optimal controls converges to an optimal control of the original system.

A condition of convergence of a sequence of approximating solutions of initial value problems by the finite difference method was given by H. F. Trotter. This theorem is concerned with a homogeneous system and does not give the condition of convergence of approximating optimal controls for a distributed parameter system.

In this paper the condition of convergence of a sequence of approximating solutions is given for the time-optimal control problem and the final value control problem for a class of linear systems with distributed parameters.

1. Introduction. Optimal control problems for distributed parameter systems, particularly systems described by partial differential equations, are often treated using mathematical function space techniques. As a result, the equations which define the optimal control are frequently obtained in abstract terms. Numerical solutions are obtained by approximating the abstract operations in a computationally feasible manner. In obtaining approximating solutions, the method of expanding in eigenfunctions [1] and the finite difference method are widely used.

After having found an approximate optimal control, the question arises whether a sequence of these approximating optimal controls converges to an optimal control of the original system. A condition of convergence of a sequence of approximating solutions of initial value problems by the finite difference method in which a differential operator is replaced by a difference operator was given by H. F. Trotter [2], [3]. This theorem is concerned with a homogeneous system and does not give the condition of convergence of approximating optimal controls for a distributed parameter system, but its result becomes very useful for our discussion as shown below.

In this paper the condition of convergence of a sequence of approximating solutions is given for the time-optimal control problem and the final value control problem for a class of linear systems with distributed parameters.

To discuss the convergence of approximating solutions, it is important to define a concept of metric. In the following, the space is considered to be a Hilbert space. Practically it would be convenient to consider the space to be $L^2(S)$, S being a region of Euclidean space with Lebesgue measure.

* Received by the editors December 16, 1969, and in final revised form November 13, 1970.

† Department of Aeronautical Engineering, Nagoya University, Nagoya, Japan.

‡ Department of Electrical Engineering, Waseda University, Tokyo, Japan.

2. Definitions and basic theorem. For convenience, the following notations, definitions and basic theorem are introduced. Let X be a Hilbert space and let X_n be a subspace of X. We consider a linear mapping $P_n: X \to X_n$ which satisfies the following conditions:

$$P_n^2 = P_n, \qquad \|P_n\| = 1,$$
$$\lim \|P_n f - f\| = 0 \quad \text{for every } f \in X.$$

DEFINITION. The *limit of a sequence of operators* $\{A_n\}$, where A_n is an operator on X_n, is the operator on X whose domain consists of $f \in X$, for which $\{A_n P_n f\}$ converges, and whose value is $\lim A_n P_n f$.

Let $\{S_n(t)\}$ be a sequence of strongly continuous semigroups of operators on X_n into X_n, and $\{A_n\}$ be the sequence of associated infinitesimal generators. Then we have the following theorem, which is due to H. F. Trotter.

THEOREM (Trotter [2]). *If the range of* $\lambda I - A$, *denoted by* $R(\lambda I - A)$, *is dense in* X *for some* $\lambda > K$ *and the following conditions are satisfied, where* M *and* K *are positive constants*:

(C) $\qquad A = \lim A_n \quad \textit{and} \quad D(A) \textit{ is dense in } X,$

(S) $\qquad \|S_n(t)\| \leqq Me^{Kt},$

then a closed extension of A *is the infinitesimal generator of* $S(t)$, *where* $S(t) = \lim S_n(t)$ *is a strongly continuous semigroup of operators on* X.

In this theorem it is demonstrated that if conditions (C) and (S) are satisfied, approximating solutions of approximating equations starting from given approximating initial states converge in the sense of norm to a solution of the original homogeneous partial differential equation, which satisfies a given initial condition.

3. Statement of the problem. Let us consider the following distributed parameter system:

$$\partial x(t, s)/\partial t = Ax(t, s) + u(t, s), \tag{1}$$

where for each t, $0 \leqq t < \infty$, $x(t, s)$ is an element of X, which is an arbitrary Hilbert space $X(\Omega)$, $s \in \Omega$, consisting of functions on a bounded domain Ω; A is an unbounded linear operator from X to X, for instance, A being a partial differential operator on Ω; and $u(t, s)$ is an element of X for each t and is bounded measurable in t. We denote $x(t, s)$ and $u(t, s)$ as $x(t)$ and $u(t)$, respectively. Here A is assumed to be an infinitesimal generator of the strongly continuous semigroup $S(t)$. Then (1) becomes

$$\dot{x}(t) = Ax(t) + u(t). \tag{1'}$$

Along with the system (1'), the following equation is considered:

$$\dot{x}_n(t) = A_n x_n(t) + u_n(t), \tag{2}$$

where $x_n(t)$ is an element of X_n for each t, $0 \leqq t < \infty$, $u_n(t)$ is an element of X_n for each t and is bounded measurable in t, and A_n is the operator from X_n into X_n which is an infinitesimal generator of a strongly continuous semigroup $S_n(t)$.

In the case of A_n being the difference operator on X_n into X_n, (1') is the original system which we consider and (2) is the approximating system given by the finite difference method. In what follows A_n is assumed to be a bounded linear operator.

As an example of the mapping P_n consider the following:

$$X = L_2(0,1), \quad x(s) \in X, \quad s \in [0,1],$$

$$P_n x(s) = \sum_0^{n-1} \alpha_k c_k(s),$$

where

$$\alpha_k = n \int_{k/n}^{k/n+1/n} x(s)\, ds,$$

$$C_k(s) = \begin{cases} 1, & k/n \leqq S \leqq k/n + 1/n, \\ 0, & k/n \geqq S \quad \text{or} \quad S \geqq k/n + 1/n. \end{cases}$$

In this case x_n is a space of finitely-valued functions.

The solution of (1′) is formally written as

$$x(t) = S(t)x(0) + \int_0^t S(t-\sigma)u(\sigma)\, d\sigma. \tag{3}$$

Since $u(t)$ is bounded and measurable, the integral on the right-hand side of (3) exists in the sense of Bochner. Sufficient conditions under which (3) represents the strong solution to (1′) (differentiable in the strong sense) have been given by Balakrishnan [4]. But since the right-hand side of (3) exists in the sense of Bochner, we define (3) as the solution of (1′) for any bounded measurable function $u(t)$ and a given initial condition $x(0)$; i.e., we deal with the mild solution [5] in this paper.

For simplicity's sake, one can assume without loss of generality that $x(0) = 0$ in the sequel. Throughout the following discussion, we assume that the conditions (C) and (S) are satisfied.

Now the mappings $L(T)$ and $L_n(T)$ are defined as follows:

$$L(T)u = \int_0^T S(T-\sigma)u(\sigma)\, d\sigma, \qquad u(t) \in X \quad \text{for each } t, \tag{4}$$

and

$$L_n(T)u_n = \int_0^T S_n(T-\sigma)u_n(\sigma)\, d\sigma, \qquad u_n(t) \in X_n \quad \text{for each } t. \tag{5}$$

Under the hypothesis of Trotter's theorem, it is shown that $\lim \|S(t)x - S_n(t)P_n x\| = 0$ for any $x \in X$, and again by the condition (S), the following equation is obtained for any $u \in X$:

$$\lim \|L(T)u - L_n(T)P_n u\| = 0.$$

The constraints C, C_n and the sets U, U_n of admissible controls are defined respectively as follows:

$$C = \{x \in X : \|x\| \leqq 1\},$$

$$C_n = \{x_n \in X_n : \|x_n\| \leqq 1\},$$

$$U = \{\text{bounded measurable function } u(t) : u(t) \in C \text{ for almost all } t\},$$

$$U_n = \{\text{bounded measurable function } u_n(t) : u_n(t) \in C_n \text{ for almost all } t\}.$$

From the property of P_n it can be shown that $P_nC = C_n$.

In this article we discuss the convergence of approximating solutions of the following Problems A and B, which are commonly seen in optimal control problems.

PROBLEM A. *Final value problem.*

(A1) The original final value problem in system (1′) is to find the admissible control minimizing $\|x(T) - y\| = \|L(T)u + S(T)x(0) - y\|$ for given $y \in X$ at a fixed time T, starting from a given initial state $x(0)$.

(A2) The approximating final value problem in system (2) is to find the admissible control minimizing $\|x_n(T) - P_ny\| = \|L_n(T)u_n + S_n(T)x_n(0) - P_ny\|$ for given $P_ny \in X_n$ at a fixed time T, starting from a given approximating initial state $P_nx(0)$, with $n \to \infty$.

Then our problem is reduced to, "Does a sequence of approximating optimal controls of the problem (A2) converge to an optimal control of the problem (A1)?"

PROBLEM B. *Time-optimal problem.*

(B1) The original time-optimal problem in system (1′) is to find the admissible control which transfers a system from a given initial state $x(0)$ to a given target state $x \in X$ in minimum time.

(B2) The approximating time-optimal problem is to find the admissible control which transfers a system from a given initial state $P_nx(0)$ to a given approximating target state $P_nx \in X_n$ in minimum time.

Then our problem is, "Does a sequence of these approximating time-optimal controls of the problem (B2) converge to a time-optimal control of the problem (B1)?"

4. Condition of convergence.

PROBLEM A. *Final value problem.* Let $u^0(t)$ be the optimal control of the original final value problem and $u_n^0(t)$ be the optimal control of the approximating final value problem at the nth degree of approximation.

Now we make the assumption that the approximating optimal solutions $u_n^0(t)$ exist and the conditions (C), (S) and

$$\text{(C*)} \qquad \lim \|S_n^*(t)P_nf - S^*(t)f\| = 0, \qquad f \in X,$$

are satisfied, where S^* and S_n^* are the adjoint operators of S and S_n considered on the space X and X_n respectively.

Moreover, let $B_2[X, T]$ be the space of strongly measurable functions $x(t)$ with range in X such that $\int_0^T \|x(t)\|^2\, dt < \infty$. $B_2[C, T]$ is the set of all $u(t)$ such that ess sup $\|u(t)\| \leqq 1$ and $B_2[C_n, T]$ is the set of all $u_n(t)$ such that ess sup $\|u_n(t)\| \leqq 1$. $B_2[C, T]$ is the admissible control set for the exact problem, and $B_2[C_n, T]$ is the admissible control set for the approximate problem.

LEMMA A1. *Let the conditions* (C), (S) *and* (C*) *be satisfied. Then there exists a weakly convergent subsequence* $\{u_{nj}^0(t)\}$ *of* $\{u_n^0(t)\}$ *in* $B_2[X, T]$, *and its weak limit* $u(t)$ *belongs to* $B_2[C, T]$. *Moreover* $u(t)$ *is an optimal control of the original final value problem.*

Proof. The first part of the lemma is easily shown from the fact that $u_n^0(t) \in B_2[C_n, T] \subseteq B_2[C, T]$, $B_2[X, T]$ is a reflexive Banach space, and the set $B_2[C, T]$ is bounded, closed and convex in $B_2[X, T]$.

Next we shall show that $L_n(T)u_n^0$ converges weakly to $L(T)u$. In the following we write $u_n^0(t)$ for $u_{nj}^0(t)$.

Since $\{X - X_n\} \perp X_n$ and P_n is a projection operator, for $f \in X$, we have

$$[L_n(T)u_n^0, f]_X = [L_n(T)u_n^0, P_n f]_X = [L_n(T)u_n^0, P_n f]_{X_n},$$

$$[u_n^0, S_n^*(\sigma)P_n f]_{X_n} = [u_n^0, S_n^*(\sigma)P_n f]_X,$$

where $[\cdot, \cdot]_X$ and $[\cdot, \cdot]_{X_n}$ denote inner product on X and X_n. Hence, we have

$$\begin{aligned}
[L_n(T)u_n^0 - L(T)u, f]_X &= [L_n(T)u_n^0, f]_X - [L(T)u_n^0, f]_X + [L(T)u_n^0 - L(T)u, f]_X \\
&= [L_n(t)u_n^0, P_n f]_{X_n} - [L(T)u_n^0, f]_X + [L(T)u_n^0 - L(T)u, f]_X \\
&= \int_0^T [S_n(T-\sigma)u_n^0(\sigma), P_n f]_{X_n}\, d\sigma \\
&\quad - \int_0^T [S(T-\sigma)u_n^0(\sigma), f]_X\, d\sigma \\
&\quad + \int_0^T [S(T-\sigma)(u_n^0(\sigma) - u(\sigma)), f]_X\, d\sigma \\
&= \int_0^T [u_n^0(\sigma), S_n^*(T-\sigma)P_n f]_{X_n}\, d\sigma \\
&\quad - \int_0^T [u_n^0(\sigma), S^*(T-\sigma)f]_X\, d\sigma \\
&\quad + \int_0^T [u_n^0(\sigma) - u(\sigma), S^*(T-\sigma)f]_X\, d\sigma \\
&= \int_0^T [u_n^0(\sigma), S_n^*(T-\sigma)P_n f - S^*(T-\sigma)f]_X\, d\sigma \\
&\quad + \int_0^T [u_n^0(\sigma) - u(\sigma), S^*(T-\sigma)f]_X\, d\sigma.
\end{aligned}$$

The first term of the right-hand side goes to zero as $n \to \infty$ by conditions (C*), (S), and the second term goes to zero as $n \to \infty$ since the integrand is bounded and converges to zero for almost all (a.a.) σ.

Next we shall show that $u(t)$ is optimal. It is clear that the following inequality is satisfied for any $v(t) \in U$:

$$\|L_n(T)u_n^0 - P_n y\| \leqq \|L_n(T)P_n v - P_n y\|.$$

By the resonance theorem [6] we have

$$\begin{aligned}\|L(T)u - y\| &\leqq \liminf \|L_n(T)u_n^0 - P_n y\| \\ &\leqq \liminf \|L_n(T)P_n v - P_n y\| \\ &= \|L(T)v - y\| .\end{aligned}$$

Hence $u(t)$ is optimal.

THEOREM A1. *Let the conditions* (C), (S) *and* (C*) *be satisfied. Suppose that* $P_n y \notin L_n(T)u_n$, $y \notin L(T)u$, *where* $u_n \in U_n$, $u \in U$ *and* $u^0(t)$ *is the unique optimal control of the original final value problem. Uniqueness has been assumed in this theorem. Then,* $u_n^0(t)$ *converges to* $u^0(t)$ *strongly for a.a. t.*

Proof. By Lemma A1, there exists a subsequence $\{u_{nj}^0(t)\}$ of $\{u_n^0(t)\}$ which converges weakly to an optimal control $u(t)$ in $B_2[X, T]$. $u^0(t)$ being unique, it follows that $u(t) = u^0(t)$ for a.a. t. Since $P_n y \notin L_n(T)u_n$, $y \notin L(T)u$ and $u^0(t)$, $u_n^0(t)$ are optimal, we see that [5]

$$\|u^0(t)\| = 1 \quad \text{for a.a. } t, \qquad \|u_n^0(t)\| = 1 \quad \text{for a.a. } t.$$

If we define the norm $|||x|||$ of the space $B_2[X, T]$ by $|||x|||^2 = \int_0^T \|x(t)\|^2 \, dt$, it follows that $|||u_{nj}^0||| = |||u^0|||$.

Therefore it can be shown that $\{u_{nj}^0(t)\}$ converges strongly to $u^0(t)$ in $B_2[X, T]$, because $B_2[X, T]$ is a Hilbert space, $\lim_{nj\to\infty} |||u_{nj}^0||| = |||u^0||| = T$ and $\{u_{nj}(t)\}$ converges weakly to $u^0(t)$. Any subsequence of $\{u_n^0(t)\}$ has its subsequence converging weakly to $u^0(t)$, and therefore it has the subsequence converging strongly to $u^0(t)$. Consequently $\{u_n^0(t)\}$ converges strongly to $u^0(t)$ in $B_2[X, T]$. By the definition of the norm of $B_2[X, T]$, it can be concluded that $u_n^0(t)$ converges to $u^0(t)$ for a.a. t.

PROBLEM B. *Time-optimal problem.*

LEMMA B1. *Let the conditions* (C), (S) *and* (C*) *be satisfied. Further, if the following conditions* (L1) *and* (L2) *are satisfied*:

(L1) *there exist some finite time* T_n *and a control* $u_n \in U_n$ *such that* $L_n(T_n)u_n = P_n x$ *for each n,*

(L2) T_n *converges to* T;

then there exists $u(t) \in U$ *such that* $L(T)u = x$.

Proof. We consider the space $B_2[X, T_0]$ and the set $B_2[C, T_0]$, where T_0 is a sufficiently large value. For $t \geqq T_n$, we put $u_n(t) = 0$. Clearly $u_n(t) \in B_2[C_n, T_0] \subseteq B_2[C, T_0]$. Since $B_2[X, T_0]$ is a reflexive space and $B_2[C, T_0]$ is closed, convex and bounded in $B_2[X, T_0]$, we can find a subsequence $\{u_{nj}(t)\}$ which converges weakly to $u(t) \in B_2[C, T_0]$ (i.e., $u(t) \in U$).

Next it will be shown that $L(T)u = x$. We prove that $[P_{nj}x, f]_X$ converges to $[x, f]_X$ and $[L_{nj}(T_{nj})u_{nj}, f]_X$ converges to $[L(T)u, f]_X$, where $f \in X$:

$$\begin{aligned}&[L_{nj}(T_{nj})u_{nj} - L(T)u, f]_X \\ &\quad = [L_{nj}(T_{nj})u_{nj}, P_{nj}f]_{X_{nj}} - [L(T)u_{nj}, f]_X + [L(T)u_{nj} - L(T)u, f]_X \\ &\quad = \int_T^{T_{nj}} [S_{nj}(T_{nj} - \sigma)u_{nj}(\sigma), P_{nj}f]_{X_{nj}} \, d\sigma + \int_0^T [S_{nj}(T_{nj} - \sigma)u_{nj}(\sigma), P_{nj}f]_{X_{nj}} \, d\sigma \\ &\qquad - \int_0^T [S(T - \sigma)u_{nj}(\sigma), f]_X \, d\sigma + \int_0^T [S(T - \sigma)(u_{nj}(\sigma) - u(\sigma)), f]_X \, d\sigma\end{aligned}$$

$$= \int_T^{T_{nj}} [S_{nj}(T_{nj} - \sigma)u_{nj}(\sigma), P_{nj}f]_{X_{nj}}\, d\sigma + \int_0^T [u_{nj}(\sigma), S_{nj}^*(T_{nj} - \sigma)P_{nj}f]_{X_{nj}}\, d\sigma$$

$$- \int_0^T [u_{nj}(\sigma), S^*(T - \sigma)f]_X\, d\sigma + \int_0^T [u_{nj}(\sigma) - u(\sigma), S^*(T - \sigma)f]_X\, d\sigma$$

$$= \int_T^{T_{nj}} [S_{nj}(T_{nj} - \sigma)u_{nj}(\sigma), f]_X\, d\sigma$$

$$+ \int_0^T [u_{nj}(\sigma), S_{nj}^*(T_{nj} - \sigma)P_{nj}f - S^*(T - \sigma)f]_X\, d\sigma$$

$$+ \int_0^T [u_{nj}(\sigma) - u(\sigma), S^*(T - \sigma)f]_X\, d\sigma.$$

The first term from the last equality sign goes to zero as $n \to \infty$ by condition (S) and the third term goes to zero since the integrand is bounded and converges to zero for a.a. σ. The second term goes to zero by condition (S), condition (C*) and the continuity of $S^*(t)$ in t. Hence $L(T)u = x$.

Remark. In order to obtain the condition (C*), we again make use of Trotter's theorem. Since condition (S) is always satisfied for $S_n^*(t)$, it is sufficient to check the validity of condition (C) for A_n^*.

LEMMA B2. *Let $S(t)$ be a strongly continuous group of operators. If x is attainable at time T with $u(t) \in U$, then for sufficiently large n, there exist $u_n(t) \in U_n$, such that $L_n(T + \varepsilon(n))u_n = P_n x$, where $\varepsilon(n) \to 0$ as $n \to \infty$.*

Proof. By $L(T)u = x$, the following inequality is obtained:

$$\|L_n(T)P_n u - P_n x\| \leqq \|L_n(T)P_n u - P_n L(T)u\| + \|P_n L(T)u - P_n x\| \leqq \varepsilon^2(n).$$

Thus we can set

$$P_n x = L_n(T)P_n u + \varepsilon^2(n)Z_n,$$

where $\|Z_n\| = 1$.

Since $S(t)$ is a group,

$$P_n x = \int_0^T S_n(T - \sigma)P_n u(\sigma)\, d\sigma + \varepsilon(n) \int_0^{\varepsilon(n)} S_n(T + \varepsilon(n) - \sigma)S_n(\sigma - \varepsilon(n) - T)Z_n\, d\sigma$$

$$= \int_{\varepsilon(n)}^{T+\varepsilon(n)} S_n(T + \varepsilon(n) - \sigma)P_n u(\sigma - \varepsilon(n))\, d\sigma + \int_0^{\varepsilon(n)} S_n(T + \varepsilon(n) - \sigma)w_n(\sigma)\, d\sigma,$$

where

$$w_n(\sigma) = \varepsilon(n)S_n(\sigma - \varepsilon(n) - T)Z_n.$$

By condition (S), we have $w_n(\sigma) \in U_n$ for sufficiently large n.

Finally we may only define a new admissible control $u_n(t)$ as follows:

$$u_n(t)\begin{cases} w_n(t), & 0 \leqq t \leqq \varepsilon(n), \\ P_n u(t - \varepsilon(n)), & \varepsilon(n) \leqq t \leqq T + \varepsilon(n). \end{cases}$$

THEOREM B1. *Let $S(t)$ be a strongly continuous group of operators, and let the conditions* (C), (S), (C*) *and the following two conditions be satisfied*:

(T1) *T_n^0 is an optimal time and u_n^0 is an associated optimal control such that $L_n(T_n^0)u_n^0 = P_n x$.*

(T2) *There exists $w(t) \in U$ such that $L(T)w = x$ at some finite time T.*

Then T_n^0 converges to an optimal time T^0 as $n \to \infty$.

Proof. By condition (T2) and Lemma B2, there exists a subsequence $\{T_{nj}^0\}$ which converges to T'. Hence we see that an optimal time exists, by condition (C*) and Lemma B1; i.e., there exists an admissible control $u \in U$ such that $L(T')u = x$.

If we denote the optimal time by T^0, the following inequality can be obtained by Lemma B2 and the optimality of T_{nj}^0:

$$T_{nj}^0 \leqq T^0 + \varepsilon(n).$$

Since T_{nj} converges to T', we have $T' \leqq T^0$ and hence $T' = T^0$ by the optimality of T^0.

The above proof is true for any convergent subsequence and hence T_n^0 converges to T^0.

LEMMA B3. *Let $S(t)$ be a holomorphic semigroup of operators, and let conditions* (C) *and* (S) *be satisfied. Suppose that $u(t)$ is continuously differentiable in $[0, T]$ and $\|A_n P_n x\| < 1$. If x is attainable at time T with $u(t) \in U$, where $x \in D(A)$, then for sufficiently large n, there exists $u_n(t) \in U_n$ such that $L_n(T + \gamma(n))u_n = P_n x$, where $\gamma(n) \to 0$ as $n \to \infty$.*

Proof. Let us define $y_n = L_n(T)P_n u$.

Since the function

$$x_n(t) = \frac{K - t}{K} y_n + \frac{t}{K} P_n x, \qquad 0 \leqq t \leqq K,$$

satisfies (2) in which we put

$$v_n(t) = \frac{1}{K}(P_n x - y_n) - \frac{K - t}{K}(A_n y_n - A_n P_n x) - A_n P_n x,$$

the control $v_n(t)$ transfers y_n to $P_n x$ in time K. By the assumption $x = L(T)u$, we have

$$\|P_n x - y_n\| = \|P_n L(T)u - L_n(T)P_n u\| = \varepsilon^2(n),$$

where $\varepsilon(n) \to 0$ as $n \to \infty$.

Next it will be shown that

$$\|A_n y_n - A_n P_n x\| = \eta^2(n),$$

where $\eta(n) \to 0$ as $n \to \infty$.

Since A_n is a bounded operator and $u(t)$ is continuously differentiable in $[0, T]$, we have $A_n S_n(T - \tau)P_n u(\tau) \in L_1(0, T)$ and $S_n(T - \tau)P_n \partial u(\tau)/\partial\tau \in L_1(0, T)$.

From the above facts and the boundedness of A_n it follows that $S_n(T - \tau)P_n u(\tau)$ is strongly differentiable and $\partial(S_n(T - \tau)P_n u(\tau))/\partial\tau \in L_1(0, T)$. Hence by the

continuity of $\partial(S_n(T-\tau)P_nu(\tau))/\partial\tau$ and again by the boundedness of A_n, we have

$$\begin{aligned} A_ny_n &= A_n\int_0^T S_n(T-\tau)P_nu(\tau)\,d\tau = \int_0^T A_nS_n(T-\tau)P_nu(\tau)\,d\tau \\ &= -\int_0^T \frac{\partial}{\partial\tau}(S_n(T-\tau)P_nu(\tau))\,d\tau + \int_0^T S_n(T-\tau)P_n\frac{\partial}{\partial\tau}u(\tau)\,d\tau \\ &= -P_nu(T) + S_n(T)P_nu(0) + \int_0^T S_n(T-\tau)P_n\frac{\partial}{\partial\tau}u(\tau)\,d\tau. \end{aligned}$$

Now we define Z_δ as follows:

$$Z_\delta = S(\delta)x = \int_0^T S(T+\delta-\tau)u(\tau)\,d\tau.$$

Since $S(t)$ is a holomorphic semigroup, i.e., $\|AS(t)\| < C/t$, where C is some constant, we see that $AS(T+\delta-\tau)u(\tau) \in L_1(0, T)$ and $S(T+\delta-\tau)\partial u(\tau)/\partial\tau \in L_1(0, T)$. From these and the holomorphic property of $S(t)$, it can be easily shown that $S(T+\delta-\tau)u(\tau)$ is strongly differentiable and $\partial(S(T+\delta-\tau)u(\tau))/\partial\tau \in L_1(0, T)$. Hence the next equality can be obtained by noticing that A is a closed operator:

$$\begin{aligned} AZ_\delta &= A\int_0^T S(T+\delta-\tau)u(\tau)\,d\tau \\ &= \int_0^A AS(T+\delta-\tau)u(\tau)\,d\tau \\ &= -S(\delta)u(T) + S(T+\delta)u(0) + \int_0^T S(T+\delta-\tau)\frac{\partial}{\partial\tau}u(\tau)\,d\tau. \end{aligned}$$

It is clear that the following inequality is satisfied:

$$\lim_{n\to\infty}\|A_ny_n - Ax\| \leqq \lim_{\delta\to 0}\lim_n \|A_ny_n - Az_\delta\| + \lim_{\delta\to 0}\|AZ_\delta - Ax\|.$$

Now by conditions (C) and (S), we have

$$\begin{aligned} \lim_{\delta\to 0}\lim_n \|A_ny_n - AZ_\delta\| &= \lim_{\delta\to 0}\lim_n \| -P_nu(T) + S_n(T)P_nu(0) \\ &\quad + \int_0^T S_n(T-\tau)P_n\frac{\partial}{\partial\tau}u(\tau)\,d\tau + S(\delta)u(T) - S(T+\delta)u(0) \\ &\quad - \int_0^T S(T+\delta-\tau)\frac{\partial}{\partial\tau}u(\tau)\,d\tau\| \\ &= \lim_{\delta\to 0}\| -u(T) + S(T)u(0) + \int_0^T S(T-\tau)\frac{\partial}{\partial\tau}u(\tau)\,d\tau \\ &\quad + S(\delta)u(T) - S(T+\delta)u(0) - \int_0^T S(T+\delta-\tau)\frac{\partial}{\partial\tau}u(\tau)\,d\tau\| = 0. \end{aligned}$$

On the other hand, by the fact that $x \in D(A)$ and strong continuity of $S(t)$, we

have that

$$\lim_{\delta \to 0} \|AZ_\delta - AX\| = \lim_{\delta \to 0} \|S(\delta)AX - AX\| = 0.$$

Consequently we have $\lim \|A_n y_n - A_n P_n x\| = \eta^2(n)$, since $\lim \|A_n y_n - Ax\| = 0$, and by the triangular inequality,

$$\|A_n y_n - A_n P_n x\| \leqq \|A_n y_n - Ax\| + \|Ax - A_n P_n x\|.$$

Since $\|A_n P_n x\| < 1$, if we choose $K = \gamma(n) = \max(\varepsilon(n), \eta(n))$, we have $v_n(t) \in U_n$ for sufficiently large n.

Finally we define a new admissible control $u_n(t)$ as follows:

$$u_n(t) = \begin{cases} P_n u(t), & 0 \leqq t \leqq T, \\ v_n(t - T), & T \leqq t \leqq T + \gamma(n). \end{cases}$$

Then we have the conclusion of Lemma B3, i.e., $P_n x = L_n(T + \gamma(n))u_n$.

Now we can easily get the following theorem from Lemma B3.

THEOREM B2. *Let $S(t)$ be a holomorphic semigroup of operators, and let the conditions* (C), (S), (C*) *and the following conditions be satisfied*:

(T1) *T_n^0 is an optimal time control and u_n^0 is an associated optimal control such that $L_n(T_n^0)u_n^0 = P_n x$.*
(T2) *There exists $u(t) \in U$ such that $L(T)u = x$ at some time T, where $x \in D(A)$.*
(T3) *$u(t)$ is continuously differentiable in $[0, T]$.*
(T4) $\|A_n P_n x\| < 1$.

Then T_n^0 converges to an optimal time T^0 as $n \to \infty$.

Remark. One should note that in the problem of getting to the origin, $\|A_n P_n x\| < 1$ is automatically satisfied, since $x = 0 \in D(A)$.

Remark. In the case of the space being a Hilbert space, it is clear that an optimal control sequence $\{u_n^0(t)\}$ converges to an optimal control for a.a. t, since the optimal control associated with the target x is unique [4], [5].

5. Conclusion. In this paper we have treated the convergence of approximating solutions for both the final value and time optimal control problems for linear distributed parameter systems. For the final value problem we have shown, under stated conditions, that where the sequence of approximating solutions exists, a subsequence converges weakly to the true optimal control (Lemma A1). In addition, if the desired system state is not in the range of certain system operations, and if the true optimal control is unique, then the preceding sequence converges strongly to the optimal control (Theorem A1).

For the time optimal control problem we have shown, under stated conditions, that the sequence of optimal times for the approximate problems exists and converges to the optimal time for the exact problem (Theorems B1, B2).

Theorem A1 is true even if $S(t)$ is not a group but a strongly continuous semigroup of operators. But, Lemma B2 and hence Theorem B1, Lemma B3 and hence Theorem B2 are valid only if $S(t)$ is a group of operators and a holomorphic semigroup of operators respectively.

In case $S(t)$ is a semigroup of operators, our theorems are also likely to be true, if we can choose more refined projection operators P_n and difference operators A_n.

Acknowledgment. The authors wish to thank Associate Professor H. Ishigaki, the School of Education, Waseda University, for many helpful suggestions and discussions.

REFERENCES

[1] E. I. AXELBAND, *Function space methods for the optimal control of a class of distributed parameter control systems*, Proc. Joint Automatic Control Conf., 1965, pp. 374–380.

[2] H. F. TROTTER, *Approximation of semi-groups of operators*, Pacific J. Math., 8 (1958), pp. 887–919.

[3] P. D. LAX, *Survey of the stability of linear finite difference equations*, Comm. Pure Appl. Math., 9 (1956), pp. 267–293.

[4] A. V. BALAKRISHNAN, *Optimal control problems in Banach spaces*, this Journal, 3 (1968), pp. 109–127.

[5] A. FRIEDMAN, *Optimal control in Banach spaces*, J. Math. Anal. Appl., 19 (1968), pp. 35–55.

[6] K. YOSHIDA, *Functional Analysis*, Springer-Verlag, Berlin, 1965.

SIAM J. CONTROL
Vol. 9, No. 2, May 1971

SOME EFFECTS OF MEASUREMENT UNCERTAINTY IN LINEAR MULTISTAGE GAMES*

W. W. WILLMAN†

Abstract. A class of linear multistage games is examined in which the players have noisy measurements of the state. If a linear solution exists for such a game, it is shown to be related to one of the solutions of the corresponding deterministic game by the certainty-equivalence principle. This solution is generally not the one ordinarily associated with this deterministic game. Results are obtained by constructing state variables from the players' measurement sequences and applying known control theoretic results to pairs of optimal control problems associated with linear game solutions.

1. Introduction. A certainty-equivalence principle holds for a class of multistage games in which the players have noisy measurements of the state. As in the control theory context, the games in this class have linear dynamics, quadratic payoffs, and measurements perturbed by sequences of independent additive Gaussian random variables. The measurement noise sequences affecting the players are statistically independent. Sufficient conditions for this certainty-equivalence result, and the corresponding control laws, can be expressed in terms of a set of implicit difference equations.

Unlike its counterpart in optimal control theory, however, this result does not apply to the state feedback control laws usually computed as a solution to the corresponding deterministic game, but to a more complicated pair of optimal control laws. One implication of this certainty-equivalence result is that the mean sample path generated by the optimal control laws in the stochastic game coincides with the optimal trajectory in the deterministic game.

The type of game examined here is an extension of a deterministic game solved earlier by Ho, Bryson and Baron [4]. Several other stochastic extensions of this game have already been investigated: by Behn and Ho [1] for the case in which one of the players has perfect measurements, by Rhodes and Luenberger [9] for the case in which one player has no measurements, and by Rhodes and Luenberger [10] for a case in which both players have noisy measurements and constrained state estimators. The mean optimal sample paths of these other extensions were shown by their investigators to possess the property described above. The solution to the game considered here, however, is not obtained as explicitly as in these other cases. The formal extension of the following results to the continuous time case is discussed in Willman [12].

2. Notation. All matrices and vectors in the following have real components. Capital letters are used to denote matrices, lower-case letters to denote vectors. It is also convenient to introduce the notation

$$\langle A \rangle_{i,j}$$

* Received by the editors April 16, 1970, and in revised form August 6, 1970.

† Operations Research Group, Naval Research Laboratory, Washington, D.C. 20390. This paper represents part of a dissertation submitted to Harvard University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The work was supported in part by the National Aeronautics and Space Administration under Contract NGR-22-007-068, by the Office of Naval Research under Contract N00014-67-A-0298-0006, and by the Division of Engineering and Applied Physics, Harvard University.

for the (i,j)th partition of a partitioned matrix A. Only one subscript is used for a partitioned vector, or a vertically or horizontally partitioned matrix. A^T is used to denote the transpose of a matrix A. I_r is used to denote the r-dimensional identity matrix when the dimension is not clear from the context.

3. A particular class of games. For greater clarity, the main results are derived for a restricted class of zero-sum games. In accordance with the original motivation, the two players are called the pursuer and evader. The extent to which the results generalize will be discussed later.

The generic game of interest here has dynamics defined by the following linear vector difference equation, where i is an integer index (to be called time) varying from 0 to $N-1$ for some fixed integer N:

$$x(i+1) = x(i) + G_p(i)u(i) - G_e(i)v(i) + n(i), \qquad i = 0, 1, \cdots, N-1, \tag{1}$$

where

x: state of dynamic system,
u: pursuer's control,
v: evader's control,
n: random variable to model uncertainty in the dynamics (process noise),
$x(0)$: Gaussian $(\bar{x}, P_0)$ random variable.

The players receive the following noisy measurements of the state:

$$z_p(i) = H_p(i)x(i) + w_p(i) \quad \text{(pursuer)}, \tag{2}$$

$$z_e(i) = H_e(i)x(i) + w_e(i) \quad \text{(evader)}. \tag{3}$$

The vectors of the sequence

$$\left\{ \begin{bmatrix} n(i) \\ \hline w_p(i) \\ \hline w_e(i) \end{bmatrix} \right\}$$

are independent zero-mean Gaussian random variables, statistically independent of $x(0)$, and have covariance matrices

$$\left[\begin{array}{c|c|c} Q(i) & 0 & 0 \\ \hline 0 & R_p(i) & 0 \\ \hline 0 & 0 & R_e(i) \end{array}\right].$$

The criterion associated with this game, which the pursuer should be regarded as trying to minimize and the evader as trying to maximize, is the quadratic form

$$J \triangleq \mathscr{E}\left\{x^T(N)S_Nx(N) + \sum_{i=0}^{N-1} [u^T(i)B(i)u(i) - v^T(i)C(i)v(i)]\right\}, \tag{4}$$

where $\mathscr{E}$ denotes prior expected value. The matrices $B(i)$, $R_p(i)$, $R_e(i)$ and $C(i)$ are symmetric and positive definite; S_N is symmetric and positive semidefinite.

The control values in equations (1) and (4) should be regarded as being generated from the available measurements by a pair of pursuit and evasion *strategies.*

DEFINITION. In the context of this game, a *pursuit strategy* is a function assigning a control sequence $\{u(i)\}$ to each measurement sequence $\{z_p(i)\}$. *Evasion strategies* are similarly defined.

DEFINITION. A pursuit strategy is *admissible* if and only if for $i = 0, \cdots, N-1$:

(i) the value of $u(i)$ is determined by the measurement subsequence $\{z_p(0), \cdots, z_p(i)\}$, a "nonanticipatory" requirement; and

(ii) $u(i)$ is Borel measureable as a function of the measurements, a technical requirement for the expectation operator in (4) to be meaningful.

This holds similarly for evasion strategies.

Strategies will be denoted by capital letters to distinguish them from the control values they determine. "$\mathscr{U}$" and "$\mathscr{V}$" will be used to denote the sets of admissible pursuit and evasion strategies, respectively.

The value of the criterion J depends only on the pair of strategies employed by the two players. As usual, a solution to this game is defined to be a particular pair of admissible pursuit and evasion strategies (U^*, V^*) such that

$$J(U^*, V) \leqq J(U^*, V^*) \leqq J(U, V^*) \tag{5}$$

for all admissible U and V.

Strategy pairs which satisfy this saddle point condition are called *minimax*, and have the property that each strategy of the pair optimizes against the other. Minimax strategy pairs are generally agreed to be the only reasonable kind of "optimum" solution for zero-sum games, although other distinct solution concepts become important in nonzero-sum games, as discussed in Starr and Ho [11]. A strategy is called minimax if it is a component of a minimax strategy pair.

Following Luce and Raiffa [8], it can be shown by two applications of the saddle-point condition that, if strategy pairs (U^1, V^1) and (U^2, V^2) are both minimax, then

$$J(U^1, V^2) \leqq J(U^1, V^1) \leqq J(U^2, V^1) \leqq J(U^2, V^2) \leqq J(U^1, V^2).$$

Therefore, all four of these strategy pairs are minimax, and each pair gives the same value of J.

4. The corresponding deterministic game. The main objective of this paper is to establish relationships between solutions to the above stochastic game and those of the corresponding deterministic game, in which no randomness is present. More precisely, the deterministic game corresponding to the preceding stochastic game is defined by the equations

$$\begin{aligned} &x(i+1) = x(i) + G_p(i)u(i) - G_e(i)v(i), \qquad i = 0, \cdots, N-1, \\ &x(0) = \bar{x} \quad \text{(dynamics)}, \end{aligned} \tag{6}$$

and

$$J_d \triangleq x^T(N)S_N x(N) + \sum_{i=0}^{N-1} [u^T(i)B(i)u(i) - v^T(i)C(i)v(i)] \quad \text{(criterion)}, \tag{7}$$

where the players have perfect measurements of the state. It is a simple extension of the results of Ho, Bryson and Baron [4] that the states and controls generated by any solution of this game are related by the equations

$$u(i) = -F_p(i)x(i), \qquad i = 0, \cdots, N-1, \tag{8}$$

and

$$v(i) = -F_e(i)x(i), \qquad i = 0, \cdots, N-1, \tag{9}$$

where the matrix sequences $\{F_p(i)\}$ and $\{F_e(i)\}$ are determined explicitly from the parameters defining the game by a matrix difference equation which is too intricate to be profitably displayed here.

Since (6), (8) and (9) determine a unique trajectory (the state and control variable sequences) for this game, there exists at most one minimax trajectory associated with such a game.

If this game has a solution, (8) and (9) can be used to define a pair of linear feedback strategies by interpreting x as a state measurement instead of a time function. This particular pair of strategies has two desirable properties. Each component strategy satisfies the principle of optimality in the sense that its restriction to any terminal segment $(k, \cdots, N-1)$ of the time sequence is also minimax in the subgame that "starts" at time k with *any* value of the state $x(k)$ which can be generated by an opponent's (possibly nonoptimal) strategy. Also, it is an easy modification of a result due to Berkovitz [2] to show that if any solution of this deterministic game exists, this strategy pair is also a solution. Because of these properties, this particular solution, referred to here as the *feedback solution*, is often the only one considered for this game.

However, the feedback solution is usually not the only one for the corresponding deterministic game, if it exists. For example, it is often true under these circumstances that mixtures of the feedback strategies and the corresponding "open loop" strategies, defined by (6), (8) and (9), are also minimax. In the stochastic game this indeterminacy of the solution often disappears in an interesting way.

It is appropriate to comment here that the above deterministic game is often called a *closed loop* game, since the players can base their control decisions on the current value of the state x. This is in distinction to the corresponding *open loop* game in which the players know the initial state exactly, and also the parameters defining the game, but have available no information on the current value of the state as the game progresses. Thus their control decisions must be a function of time only. The stochastic game under investigation here can be regarded as one in which the quality of information available to the players is intermediate between the open and closed loop extremes, although it will be seen later that it is perhaps more closely related to the closed loop type of game.

5. A useful construction. The basic approach used here is to obtain results about solutions to the stochastic game by examining pairs of associated stochastic optimal control problems. As it happens, it is useful to express these control problems in terms of new state variable sequences, each term of which is a vector with N partitions of equal dimension. These "enlarged state variables" are constructed, following Wonham [14], as

$$\langle \zeta_p(i) \rangle_j \triangleq \begin{cases} z_p(j), & j < i, \\ 0, & j \geqq i, \end{cases} \tag{10}$$

$$\langle \zeta_e(i) \rangle_j \triangleq \begin{cases} z_e(j), & j < i, \\ 0, & j \geqq i, \end{cases} \tag{11}$$

for $i = 0, \cdots, N$ and $j = 0, \cdots, N-1$. In terms of the information they contain, these enlarged state variables are essentially the sequences of initial segments of the players' measurement sequences.

6. Some preliminary results. In veiw of the particular nature of the stochastic game under consideration, it is not surprising that it is worthwhile to confine the search for minimax strategies to those which specify the players' controls as linear functions of their available measurements. It is convenient to begin this search by establishing some preliminary results concerning such linear strategies in this game.

LEMMA 1. *If the controls of the evader are determined by the strategy*

$$V' : v(i) = A_e(i)\bar{x} + \sum_{j=0}^{i} L_e(i,j) z_e(j), \qquad i = 0, \cdots, N-1, \tag{12}$$

where $\{A_e(i)\}$ *and* $\{L_e(i,j)\}$ *are arbitrary sequences of real matrices, then any admissible pursuit strategy* U^* *which minimizes* J *against* V' *(over* $\mathcal{U}$*) satisfies the conditions*

$$u(i) = K_p(i)\bar{x} + M_p(i)\hat{\sigma}_p(i), \qquad i = 0, \cdots, N-1, \tag{13}$$

where $\hat{\sigma}_p(i)$ *is generated from the pursuer's measurements by*

$$\hat{\sigma}_p(i) = N_p(i)\bar{x} + T_p(i)\hat{\sigma}_p(i-1) + E_p(i) z_p(i), \qquad \hat{\sigma}_p(-1) = \begin{bmatrix} \bar{x} \\ \hline 0 \end{bmatrix}, \tag{14}$$

for $i = 0, \cdots, N-1$ *and almost all measurement sequences* $\{z_p(i)\}$*, where the matrices* $\{K_p(i)\}$*,* $\{M_p(i)\}$*,* $\{N_p(i)\}$*,* $\{E_p(i)\}$ *and* $\{T_p(i)\}$ *are determined explictly from* $\{A_e(i)\}$ *and* $\{L_e(i,j)\}$ *by an effective procedure* (to be described in more detail in the proof).

Proof sketch. Define for $i, j = 0, \cdots, N-1$ the sequence of horizontally partitioned matrices $\{\Lambda_e(i)\}$ such that

$$\langle \Lambda_e(i) \rangle_j = \begin{cases} L_e(i,j), & j < i, \\ 0, & j \geqq i, \end{cases} \tag{15}$$

and the sequence of vertically partitioned matrices $\{\Theta_e(i)\}$ such that

$$\langle \Theta_e(i) \rangle_j = \begin{cases} I_r, & j = i, \\ 0, & j \neq i, \end{cases} \tag{16}$$

where r is the dimension of z_e.

It is tedious but straightforward to verify by substitution that, if the evader uses the strategy V', the dynamics of (1) are equivalent to

$$\begin{aligned} \begin{bmatrix} x(i+1) \\ \hline \zeta_e(i+1) \end{bmatrix} &= \left[\begin{array}{c|c} I - G_e(i)L_e(i,i)H_e(i) & -G_e(i)\Lambda_e(i) \\ \hline \Theta_e(i)H_e(i) & I \end{array}\right] \begin{bmatrix} x(i) \\ \hline \zeta_e(i) \end{bmatrix} + \begin{bmatrix} G_p(i) \\ \hline 0 \end{bmatrix} u(i) \\ &\quad - \begin{bmatrix} G_e(i)A_e(i) \\ \hline 0 \end{bmatrix} \bar{x} + \left[\begin{array}{c|c} I & -G_e(i)L_e(i,i) \\ \hline 0 & \Theta_e(i) \end{array}\right] \begin{bmatrix} n(i) \\ \hline w_e(i) \end{bmatrix}, \end{aligned} \tag{17}$$

where $\begin{bmatrix} x(0) \\ \hline \zeta_e(0) \end{bmatrix}$ is

$$\text{Normal}\left(\begin{bmatrix} \bar{x} \\ \hline 0 \end{bmatrix}, \begin{bmatrix} P_0 & 0 \\ \hline 0 & 0 \end{bmatrix} \right)$$

a priori, the criterion can be expressed as

$$\begin{aligned} J = \mathscr{E}\Bigg\{ & \left[\begin{array}{c|c} x^T(N) & \zeta_e^T(N) \end{array}\right] \left[\begin{array}{c|c} S_N & 0 \\ \hline 0 & 0 \end{array}\right] \begin{bmatrix} x(N) \\ \hline \zeta_e(N) \end{bmatrix} \\ & + \sum_{i=0}^{N-1} \Bigg(u^T(i)B(i)u(i) - \Bigg(\bar{x}^T A_e^T(i) + [x^T(i) \mid \zeta_e^T(i)] \begin{bmatrix} H_e^T(i)L_e^T(i,i) \\ \hline \Lambda_e^T(i) \end{bmatrix} \Bigg) \\ & \cdot C(i) \Bigg(\left[\begin{array}{c|c} L_e(i,i)H_e(i) & \Lambda_e(i) \end{array}\right] \begin{bmatrix} x(i) \\ \hline \zeta_e(i) \end{bmatrix} + A_e(i)\bar{x} \Bigg) \Bigg) \Bigg\} \\ & - \mathscr{E}\Bigg\{ \sum_{i=0}^{N-1} w_e^T(i)L_e^T(i,i)C(i) \Bigg(L_e(i,i)w_e(i) \\ & \qquad + 2\Bigg(A_e(i)\bar{x} + [L_e(i,i)H_e(i) \mid \Lambda_e(i)] \begin{bmatrix} x(i) \\ \hline \zeta_e(i) \end{bmatrix} \Bigg) \Bigg) \Bigg\}, \end{aligned} \tag{18}$$

and the pursuer's measurements are

$$z_p(i) = [H_p(i) \mid 0] \begin{bmatrix} x(i) \\ \hline \zeta_e(i) \end{bmatrix} + w_p(i). \tag{19}$$

Since $w_e(i)$ is statistically independent of $x(i)$, $\zeta_e(i)$, and $\bar{x}$ for any admissible strategy, the second term in (18) is independent of the control law. Therefore, the problem of minimizing $J(U, V')$ over U in $\mathscr{U}$ is equivalent to the stochastic control problem defined by (17), (19) and the first term of (18), where J is to be minimized.

As reformulated in terms of the enlarged state variables, this minimization problem has been converted into a stochastic optimal control problem with linear dynamics, quadratic cost, and additive independent Gaussian process and measurement noise sequences. This means that the well-known necessary conditions developed for this type of control problem developed by Joseph and Tou [5]

and Gunckel and Franklin [3] apply here. As applied to the particular example at hand, these necessary conditions are precisely (13) and (14), where the unspecified matrix sequences are determined from the matrices $\{A_e(i)\}$ and $\{L_e(i,j)\}$ and the parameters defining the game by a certain set of matrix difference equations with specified boundary conditions. Since the solution to this type of optimal control problem is well known, the details of these equations are not presented here. This completes the proof.

The random sequence $\{\hat{\sigma}_p(i)\}$ in equation system (14) is to be interpreted as the sequence of estimates of the composite vector $\sigma_p(i) \triangleq \begin{bmatrix} x(i) \\ \hline \zeta_e(i) \end{bmatrix}$ produced by the appropriate Kalman-Bucy filter for the dynamics and measurements of (17) and (19). This means, from well-known results due to Kalman [6], that $\hat{\sigma}_p(i)$ is the conditional expected value of $\sigma_p(i)$, given the sequence of measurements $\{z_p(0), \cdots, z_p(i)\}$ for $i = 0, \cdots, N-1$. For this reason, (13) and (14) will be referred to as the *certainty-equivalent* form of U^*.

COROLLARY. Since the preceding necessary conditions are satisfied by only one strategy, *there is at most one admissible pursuit strategy (up to a set of measurement sequences of measure zero) which minimizes J against V′ over* $\mathcal{V}$.

LEMMA 2. *In the context of Lemma* 1, *the strategy* U^* *can be expressed as*

$$U^*: u(i) = A_p(i)\bar{x} + \sum_{j=0}^{i} L_p(i,j)z_p(j), \qquad i = 0, \cdots, N-1, \tag{20}$$

where the matrices $\{A_p(i)\}$ *and* $\{L_p(i,j)\}$ *are explicitly determined from the matrices* $\{A_e(i)\}$ *and* $\{L_e(i,j)\}$ *by an effective procedure.*

Proof sketch. This proof merely consists of showing by induction that (20) is equivalent to (13) and (14) (i.e., gives the same control sequence for every measurement sequence) when $\{A_p(i)\}$ and $\{L_p(i,j)\}$ are computed from the matrices $\{K_p(i)\}$, $\{M_p(i)\}$, $\{N_p(i)\}$, $\{T_p(i)\}$ and $\{E_p(i)\}$ according to a certain set of matrix difference equations developed according to Kalman and Bucy [7]. Again, the details are not given here.

For future reference, the functional dependence of the matrices $\{A_p(i)\}$ and $\{L_p(i,j)\}$ in (20) on $\{A_e(i)\}$ and $\{L_e(i,j)\}$ will be expressed as

$$\begin{aligned} A_p(i) &= f_{p,i}(\{A_e(i)\}, \{L_e(i,j)\}), \\ L_p(i,j) &= g_{p,i,j}(\{A_e(i)\}, \{L_e(i,j)\}). \end{aligned} \tag{21}$$

Equation (20) will be called the *canonical form* of the pursuit strategy U^*.

Finally, it is useful to establish sufficient conditions for the pursuit strategy U^* to minimize the criterion against the evasion strategy V'. This is done by using the well-known sufficient conditions for "linear-quadratic-Gaussian" optimal control problems. As applied to the control problem specified by (17), (19), and the first term of (18), these sufficient conditions lead to the following result.

LEMMA 3. *In the context of the preceding lemmas, if the pursuit strategy* U^* *is determined from the evasion strategy* V' *according to* (20) *and* (21), *and if the matrices*

$$B(i) + G_p^T(i)S_p(i+1)G_p(i)$$

are positive definite for $i = 0, \cdots, N-1$, *then*

$$J(U^*, V') \leqq J(U, V') \quad \textit{for all } U \in \mathcal{U}.$$

The matrix $S_p(i+1)$ is the partition of the coefficient matrix for the quadratic form in x in the cost function for this optimal control problem at time i. This sequence of matrices is determined by the usual Riccati-like difference equation associated with this control problem. The details of this equation are straightforward and are not presented here.

It is possible to prove similarly a set of lemmas analogous to Lemmas 1, 2 and 3 with the roles of the pursuer and evader interchanged. These lemmas will henceforth be designated respectively as Lemmas 4, 5 and 6. In particular, let

$$U' : u(i) = A_p(i)\bar{x} + \sum_{j=0}^{i} L_p(i,j) z_p(j), \qquad i = 0, \cdots, N-1, \tag{22}$$

$$v(i) = K_e(i)\bar{x} + M_e(i)\hat{\sigma}_p(i), \tag{23}$$

$$\begin{gathered} \hat{\sigma}_e(i) = N_e(i)\bar{x} + T_e(i)\hat{\sigma}_e(i-1) + E_e(i) z_e(i), \\ \langle \hat{\sigma}_e(-1) \rangle_j \triangleq \begin{bmatrix} \bar{x} \\ \hline 0 \end{bmatrix}, \end{gathered} \tag{24}$$

$$V^* : v(i) = A_e(i)\bar{x} + \sum_{j=0}^{i} L_e(i,j) z_e(j), \qquad i = 0, \cdots, N-1, \tag{25}$$

and

$$\begin{aligned} A_e(i) &= f_{e,i}(\{A_p(i)\}, \{L_p(i,j)\}), \\ L_e(i,j) &= g_{e,i,j}(\{A_p(i)\}, \{L_p(i,j)\}) \end{aligned} \tag{26}$$

be the respective analogues of (12), (13), (14), (20) and (21).

7. Stochastic game solution. By combining the results of Lemmas 2 and 5, it follows that any solution to this stochastic game of the form

$$\begin{aligned} U^* : u(i) &= A_p(i)\bar{x} + \sum_{j=0}^{i} L_p(i,j) z_p(j) \quad \text{(pursuer)}, \\ V^* : v(i) &= A_e(i)\bar{x} + \sum_{j=0}^{i} L_e(i,j) z_e(j) \quad \text{(evader)}, \end{aligned} \tag{27}$$

$i = 0, \cdots, N-1$, must have coefficients $\{A_p(i)\}$, $\{A_e(i)\}$, $\{L_p(i,j)\}$ and $\{L_e(i,j)\}$ which satisfy both (21) and (26). Furthermore, by Lemmas 3 and 6, any strategy pair (U^*, V^*) of the form of (27) which satisfies these equations and also has the property that

$$B(i) + G_p^T(i) S_p(i+1) G_p(i) \tag{28}$$

and

$$C(i) + G_e^T(i) S_e(i+1) G_e(i) \tag{29}$$

are positive definite for $i = 0, \cdots, N-1$, where the matrices $\{S_p(i)\}$ and $\{S_e(i)\}$ are determined by the equations described earlier, is a game solution.

Equations (21) and (26) together can be regarded as a set of implicit difference equations to be solved for a set of matrix parameters defining a strategy pair (U^*, V^*) via (27). Not much is known at present about the existence or computation of solutions to this set of implicit equations. Numerical evidence indicates that linear minimax strategies of the form considered here do indeed exist for a variety of two-stage scalar games of this type, and that the implicit equations can be solved by iterative substitution in these cases. Also, it is shown in the next section that no such linear minimax strategy exists if there is no solution to the corresponding deterministic game.

The foregoing approach to finding minimax strategies, of course, is limited in that it eliminates from consideration all possibilities except strategy pairs of the form of (27). However, if a minimax strategy pair (U^*, V^*) of this form does exist, then it follows from the remark at the end of § 3 that any other minimax strategy pair results in the same value of the criterion J.

In this particular class of games, moreover, it is also true that if such a linear solution (U^*, V^*) exists, then any solution (U', V') is in fact the same strategy pair (except perhaps for a set of measurements of measure zero). This follows from the fact that (U', V^*) and (U^*, V') must also be minimax, which in turn implies that U' and V' satisfy the necessary conditions of Lemmas 1 and 4, respectively. Since U^* and V^* also satisfy these conditions, and since these conditions are satisfied by only one pair of strategies, $U' = U^*$ and $V' = V^*$.

This uniqueness argument, unfortunately, depends on the equivalence and interchangeability of all solutions, and hence does not generalize to the case of nonzero-sum games. This point will be discussed further in § 9.

8. Relation to the corresponding deterministic game. Although there is only one minimax trajectory in the corresponding deterministic game, assuming that solutions exist, there is usually a continuum of minimax strategies for each player. Two such minimax strategies always generate the minimax trajectory when the opponent uses a minimax strategy, but do not in general generate the same trajectory otherwise. It is shown in this section that if a linear solution exists for the stochastic game, minimax strategy pairs also exist for the corresponding deterministic game, *one* of which is related to the stochastic game solution by the certainty-equivalence principle.

It is assumed throughout the rest of this section that a stochastic game solution of the form of (27) has been found, and hence also a solution to (21) and (26).

DEFINITION. The *noiseless sample path* of the stochastic game for strategy pair (U, V) is defined as the trajectory generated by this strategy pair according to (1)–(3) when all the quantities defined as random variables assume their mean values; that is, when $x(0) = \bar{x}$, $n(i) \equiv 0$, $w_p(i) \equiv 0$, and $w_e(i) \equiv 0$.

Since all of the above equations and (27) are linear in these random variables, the noiseless sample path for the strategy pair (U^*, V^*) is the mean trajectory generated by this strategy pair, and hence describes the "average behavior" of the optimally controlled stochastic game. This particular sample trajectory

is useful in relating the solutions of the stochastic and corresponding deterministic game.

THEOREM. *If a linear solution* (U^*, V^*) *of the form of* (27) *exists for the stochastic game, then*:

(i) *A minimax strategy pair exists for the corresponding deterministic game which is related to* (U^*, V^*) *by the certainty-equivalence principle.*

(ii) *The noiseless sample path of the stochastic game for* (U^*, V^*) *coincides with the minimax trajectory of the corresponding deterministic game.*

Proof sketch. Construct the following two pursuit strategies for the corresponding deterministic game:

U_c^*—in which the pursuer's controls are generated by the canonical form of U^* (equation (20)) with the substitution of $H_p(i)x(i)$ for $z_p(i)$.

U_{ce}^*—in which the controls are generated by the certainty-equivalent form of U^* with the substitution of $\sigma_p(i)$ for $\hat{\sigma}_p(i)$ in (13). Equation (14) is superfluous in this strategy. $\sigma_p(i)$ is defined as $\begin{bmatrix} x(i) \\ \hline \zeta_e(i) \end{bmatrix}$ in this deterministic context, where $\zeta_e(i)$ is defined by (11) and (3) with $w_e(i) \equiv 0$. Note that $\sigma_p(i)$ is directly observable to the pursuer in this context.

By Lemma 2, the strategy U_c^* can also be defined by (13) and (14) with the substitution of $H_p(i)x(i)$ for $z_p(i)$ in (14).

Similarly define evasion strategies V_c^* and V_{ce}^* for the corresponding deterministic game.

The argument is now arranged into four intermediate results.

Result 1. The trajectory generated by (U_c^*, V_c^*) in the corresponding deterministic game is the same as the noiseless sample path of the optimally controlled stochastic game.

This result is verified by showing that (6) and (U_c^*, V_c^*) define the same trajectory as do (1), (2), (3), (20) and (25) when $x(0) = \bar{x}$, $n \equiv 0$, $w_p \equiv 0$ and $w_e \equiv 0$.

Result 2. In the optimally controlled stochastic game, $\hat{\sigma}_p \equiv \sigma_p$ and $\hat{\sigma}_e \equiv \sigma_e$ on the noiseless sample path.

If $e_p(i)$ is defined as $\sigma_p(i) - \hat{\sigma}_p(i)$, it obeys a linear difference equation whose inhomogeneous term is zero if $n \equiv 0$, $w_p \equiv 0$ and $w_e \equiv 0$. The initial value of e_p is proportional to $x(0) - \bar{x}$. Similarly this holds for the evader. The result follows by a simple induction argument on i.

Result 3. In the corresponding deterministic game, U_{ce}^* minimizes the criterion against V_c^*, and V_{ce}^* maximizes it against U_c^*.

This result is established by showing that the minimization of $J_d(U, V_c^*)$ over U is equivalent to solving the optimal control problem of Lemma 1 with all random variables replaced by their mean values. Since this control problem is of the classical "linear-quadratic-Gaussian" type, the certainty-equivalence principle applies to it. Therefore, by the definition of U_{ce}^*,

$$J_d(U_{ce}^*, V_c^*) \leqq J_d(U, V_c^*)$$

for all admissible U. An analogous argument shows that V_{ce}^* maximizes against U_c^*.

Result 4. The strategy pair (U_c^*, V_c^*) is minimax in the corresponding deterministic game.

By Results 1 and 2, the substitution of $H_p(i)x(i)$ for $z_p(i)$ in (14) generates $\hat{\sigma}_p \equiv \sigma_p$ in the corresponding deterministic game when the strategy pair (U_c^*, V_c^*) is played. It follows from the second definition of U_c^* that the same trajectory is produced in this game by using the (observable) value of σ_p in (13) instead of generating $\hat{\sigma}_p$ by this modification of (14). But this is equivalent to replacing the strategy U_c^* by U_{ce}^*. Therefore,

$$J_d(U_c^*, V_c^*) = J_d(U_{ce}^*, V_c^*)$$

since the value of J_d is determined by the trajectory. Similarly,

$$J_d(U_c^*, V_c^*) = J_d(U_c^*, V_{ce}^*).$$

These equalities and Result 3 imply that (U_c^*, V_c^*) satisfies the saddle-point condition.

Since the strategy pair (U_c^*, V_c^*) is minimax in the corresponding deterministic game, the trajectory it generates must be the (unique) minimax trajectory defined by (6), (8) and (9). Since (U_c^*, V_c^*) also generates the noiseless sample path, these two trajectories are identical.

By the argument for Result 4, both U_c^* and V_c^* can be replaced by U_{ce}^* and V_{ce}^* without changing the trajectory produced in the corresponding deterministic game. Therefore the strategy pair (U_{ce}^*, V_{ce}^*) is also minimax there. The strategies of this pair are related to U^* and V^* by the certainty-equivalence principle in the sense that one can be obtained from the other by replacing σ_p (or σ_e) by the estimate $\hat{\sigma}_p$ (or $\hat{\sigma}_e$). This completes the proof.

Although a certainty-equivalence principle holds for this class of games, it does not have the same significance as the one for optimal control problems. For one thing, it does not relate the stochastic game solution to the feedback solution of the corresponding deterministic game, but to a much more complicated solution. It has been verified by numerical counterexample [13] that this other solution is not the same as the feedback solution in general. A knowledge of the feedback solution for the corresponding deterministic game is therefore of little use in computing a solution of the stochastic game.

This certainty-equivalence result is also predicated on the existence of a linear solution to the stochastic game. It can be inferred from the results established here that such a solution cannot exist unless the feedback solution of the corresponding deterministic game exists, which is easy to determine. It is not known, however, whether or not the converse is true. It is also not known whether or not a solution of some nonlinear form ever exists for the stochastic game.

The fact that the noiseless sample path of a linear stochastic game solution coincides with the minimax trajectory of the corresponding deterministic game, referred to here as the *certainty-coincidence property*, is of computational significance. It implies that if the stochastic game has such a solution, its average minimax behavior can be determined by finding the minimax trajectory of the corresponding deterministic game, which is relatively easy. If the noise magnitudes are small, moreover, its minimax behavior can be expected to approximate closely this average trajectory most of the time. Thus, this property indicates that there is no discontinuity of behavior as the noise magnitudes go from small values to zero, although conditions for which this is true have not been established.

It should be emphasized that the certainty-coincidence property is not a statement about the behavior of the optimally controlled stochastic game as the noise variances approach zero, but rather about its behavior when the noise values are zero and the variances are not zero. Intuitively, it states that the presence of the noise does not make an optimal player more or less cautious on the average, provided that a linear solution exists in its presence.

9. Extensions. A careful examination of the arguments in the preceding sections reveals that most of the results obtained there extend to the many-player nonzero-sum case as well. Specifically, all of these results except those depending on the equivalence of different solutions hold for the general many-person nonzero-sum game with linear dynamics, quadratic criteria (in general not related) for the players to minimize, linear measurements for each player, and additive independent Gaussian perturbations in the dynamics and measurements. The appropriate solution concept here is the *Nash equilibrium solution*, which is an M-tuple of admissible strategies $(U_1^*, \cdots, U_M^*)$ for the M players such that the inequalities

$$J_k(U_1^*, \cdots, U_k^*, \cdots, U_M^*) \leqq J_k(U_1^*, \cdots, U_k, \cdots, U_M^*), \qquad k = 1, \cdots, M,$$

hold for all admissible strategy M-tuples $(U_1, \cdots, U_M)$, where J_k is the criterion of the kth player.

If they exist, Nash solutions linear in the measurements can be found in this more general context by solving a system of M implicit equation schemes, and verifying the positive definiteness of some matrices. Furthermore, the noiseless sample path generated by any such solution coincides with the unique Nash trajectory in the corresponding closed loop deterministic game (which is guaranteed to have a Nash solution if such a solution exists for the stochastic game). This last point is interesting because Starr and Ho [11] have shown that the Nash trajectories are in general distinct for the various combinations of open and closed loop assumptions about the several players in nonzero-sum games of this sort. The fact that the certainty-coincidence property takes the form that it does in this nonzero-sum context indicates that noisy measurements, no matter how poor, are more closely related to perfect measurements than to no measurements, at least as far as the Nash behavior of such games is concerned. If this phenomenon seems counterintuitive, it should be recalled that, in the general nonzero-sum game, Nash solutions do not have all the properties that one would intuitively expect of a completely satisfactory solution. Nash solutions are not in general minimax or noninferior, as they are in zero-sum games (see Starr and Ho [11]).

As mentioned earlier, the general nonequivalence and noninterchangeability of Nash solutions prevents the proof of uniqueness at the end of § 7 from being extended to nonzero-sum games. The lack of such a result is a more serious shortcoming here, however, precisely because of this nonequivalence. Nash strategies lose much of their significance in nonzero-sum games if they are not unique, because the rationale for playing any particular one of them becomes ambiguous.

Acknowledgment. The author is indebted to Professor Y. C. Ho of Harvard University for suggesting this area of research and for his help and encouragement during its development.

REFERENCES

[1] R. D. BEHN AND Y. C. HO, *On a class of linear stochastic differential games*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 227–239.

[2] L. D. BERKOVITZ, *A variational approach to differential games*, Advances in Game Theory, Annals of Math. Studies, No. 52, Princeton University Press, Princeton, 1964, pp. 127–174.

[3] T. F. GUNCKEL AND G. F. FRANKLIN, *A general solution for linear sampled-data control systems*, Trans. ASME Ser. D. J. Basic Engrg., 85 (1963), pp. 197–203.

[4] Y. C. HO, A. E. BRYSON AND S. BARON, *Differential games and optimal pursuit-evasion strategy*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 385–389.

[5] P. JOSEPH AND J. TOU, *On linear control theory*, AIEE Trans., Part II, Sept. 1961.

[6] R. E. KALMAN, *A new approach to linear filtering and prediction problems*, Trans. ASME Ser. D. J. Basic Engrg., 82 (1960), pp. 35–45.

[7] R. E. KALMAN AND R. BUCY, *New results in linear filtering and prediction*, Ibid., 83 (1961), pp. 95–108.

[8] R. D. LUCE AND H. RAIFFA, *Games and Decisions*, John Wiley, New York, 1957.

[9] I. B. RHODES AND D. G. LUENBERGER, *Differential games with imperfect state information*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 29–39.

[10] ———, *Nondeterministic differential games with constrained state estimators*, Ibid., AC-14 (1969), pp. 476–481.

[11] A. W. STARR AND Y. C. HO, *Nonzero-sum differential games*, Tech. Rep. 564, Division Engrg. Appl. Phys., Harvard Univ., Cambridge, Mass., 1968.

[12] W. W. WILLMAN, *Formal solutions for a class of stochastic pursuit-evasion games*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 504–509.

[13] ———, *On a class of stochastic pursuit-evasion games*, Tech. Rep. 585, Division Engrg. Appl. Phys., Harvard Univ., Cambridge, Mass., 1969.

[14] W. M. WONHAM, *On the separation theorem of stochastic control*, this Journal, 6 (1968), pp. 312–326.

SIAM J. CONTROL
Vol. 9, No. 2, May 1971

CLOSURE, LOWER CLOSURE, AND SEMICONTINUITY THEOREMS IN OPTIMAL CONTROL*

LAMBERTO CESARI†

Abstract. First we discuss in detail the concept of lower closure for Lagrange problems, and we extend in various ways previous closure and lower closure theorems. In particular, we show that the present lower closure theorems and related concepts are extensions to Lagrange problems of optimal control of well-known semicontinuity theorems for free problems of the calculus of variations and the related concept of seminormality. Finally, we prove by a new approach that the convexity condition usually requested in lower closure theorems is, in a suitable sense, both a necessary and sufficient condition for lower closure.

In previous papers [1, abcdefgh] we have given existence theorems for one- and multidimensional Lagrange and Mayer problems of optimal control based on closure and lower closure properties of the functional under consideration. In the present paper we discuss in detail some of the relevant concepts, in particular the concept of lower closure for Lagrange problems. Also, we extend in various ways previous closure and lower closure theorems. In particular, we show that the present lower closure theorems and related concepts are extensions to Lagrange problems of optimal control of well-known lower semicontinuity theorems for free problems and the concept of seminormality of Tonelli [7] and McShane [3]. Finally, we prove by a new approach that the convexity condition usually requested in lower closure theorems is—in a suitable sense—both a necessary and sufficient condition for lower closure.

We limit ourselves in this paper to one-dimensional problems. We deal, therefore, with the (Lagrange) problem of the minimum of an integral

$$I[x, u] = \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt,$$

in classes of pairs of functions $x(t) = (x^1, \cdots, x^n)$, $u(t) = (u^1, \cdots, u^m)$, $t_1 \leqq t \leqq t_2$, satisfying a system of ordinary differential equations of the form

$$dx/dt = f(t, x(t), u(t)) \quad \text{a.e. in } [t_1, t_2],$$

and constraints of the form

$$(t, x(t)) \in A, \quad u(t) \in U(t, x(t)) \quad \text{a.e. in } [t_1, t_2].$$

1. A closure theorem. We shall denote by t a scalar variable (time), and by $x = (x^1, \cdots, x^n)$ the state variables. Let A be a given subset of the tx-space E_{n+1}, and for any $(t, x) \in A$ let $Q(t, x)$ be a given subset of the z-space E_n, $z = (z^1, \cdots, z^n)$. A vector function $\chi(t) = (x^1, \cdots, x^n)$, $t_1 \leqq t \leqq t_2$, is said to be a solution of the orientor field

$$dx/dt \in Q(t, x), \tag{1.1}$$

* Received by the editors February 13, 1970, and in revised form August 10, 1970.

† Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48104. This research was supported in part by the United States Air Force Office of Scientific Research under Grant 69-1662.

provided:

(a) x is absolutely continuous in $[t_1, t_2]$;

(b) $(t, x(t)) \in A$ for all $t \in [t_1, t_2]$;

(c) $dx/dt \in Q(t, x(t))$ a.e. in $[t_1, t_2]$.

For any $(\bar{t}, \bar{x}) \in A$ and $\delta > 0$ we shall denote by $N_\delta(\bar{t}, \bar{x})$, or δ-neighborhood of $(\bar{t}, \bar{x})$ in A, the set of all $(t, x) \in A$ at a distance $\leqq \delta$ from $(\bar{t}, \bar{x})$. We say that the sets $Q(t, x)$ satisfy property (Q) at the point $(\bar{t}, \bar{x}) \in A$ provided

$$Q(\bar{t}, \bar{x}) = \cap_\delta \operatorname{cl} \operatorname{co} \cup Q(t, x),$$

where $\cup$ is taken for all $(t, x) \in N_\delta(\bar{t}, \bar{x})$. We say that the sets $Q(t, x)$ satisfy property (Q) in A if they satisfy this property at every $(\bar{t}, \bar{x}) \in A$. The weaker property (U) can be defined analogously by replacing the operation *cl co* above by the operation *cl*.

If $x(t)$, $a \leqq t \leqq b$, $y(t)$, $c \leqq t \leqq d$, are any two continuous vector functions with values in E_n, we denote as their ρ-distance the number $\rho(x, y) = |a - c| + |b - d| + \max |x(t) - y(t)|$, where the maximum is taken in $(-\infty, +\infty)$, and $x(t)$, $y(t)$ are thought of as extended in $(-\infty, +\infty)$ by continuity and constancy outside their intervals of definition.

Note that if $x(t)$, $t_1 \leqq t \leqq t_2$, is any continuous vector function, and $x_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, is any sequence of continuous vector functions with $\rho(x_k, x) \to 0$ as $k \to \infty$, then $(t_{1k}, x_k(t_{1k})) \to (t_1, x(t_1))$ and $(t_{2k}, x_k(t_{2k})) \to (t_2, x(t_2))$ as $k \to \infty$. Also, if A is a closed subset of the tx-space E_{n+1}, and $(t, x_k(t)) \in A$ for all $t \in [t_{1k}, t_{2k}]$ and all k, then we also have $(t, x(t)) \in A$ for all $t \in [t_1, t_2]$. Finally, if B is a closed subset of E_{2n+2} and $(t_{1k}, x_k(t_{1k}), t_{2k}, x_k(t_{2k})) \in B$ for all k, then $(t_1, x(t_1), t_2, x(t_2)) \in B$.

In [1a] we proved the following closure theorem.

THEOREM 1.1. *If A is closed, if the sets $Q(t, x)$ are closed, convex, and satisfy property* (Q) *at every point of A (with exception perhaps of a set of points whose t-coordinate lies in a set of measure zero on the t-axis), if $x_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, is a sequence of solutions of the orientor field* (1.1), *if $x(t)$, $t_1 \leqq t \leqq t_2$, is an absolutely continuous vector function, and $\rho(x_k, x) \to 0$ as $k \to \infty$, then $x(t)$, $t_1 \leqq t \leqq t_2$, is a solution of the orientor field* (1.1).

2. A closure theorem involving singular components. Let α be any integer, $0 \leqq \alpha \leqq n$, and if $x = (x^1, \cdots, x^n)$, let y denote the α-vector $y = (x^1, \cdots, x^\alpha)$ and z the $(n - \alpha)$-vector $z = (x^{\alpha+1}, \cdots, x^n)$, so that we can write $x = (y, z)$. Let A_0 be a subset of the ty-space $E_{\alpha+1}$, let $A = A_0 \times E_{n-\alpha}$, and for any $(t, y) \in A_0$ let $Q(t, y)$ be a given subset of the x-space E_n. We shall consider the orientor field

$$dx/dt \in Q(t, y). \tag{2.1}$$

We proved in [1a] the following closure theorem.

THEOREM 2.1. *If A_0 is closed, if the sets $Q(t, y)$ are closed, convex, and satisfy property* (Q) *at every point of A_0 (with exception perhaps of a set of points whose t-coordinate lies in a set of measure zero on the t-axis), if $x_k(t) = (y_k, z_k)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, is a sequence of solutions of the orientor field* (2.1) *and $x(t) = (y, z)$,*

$t_1 \leqq t \leqq t_2$, *is any vector function with* $y(t)$ *absolutely continuous in* $[t_1, t_2]$, $z(t) = Z(t) + S(t)$, $t_1 \leqq t \leqq t_2$, $Z(t)$ *absolutely continuous in* $[t_1, t_2]$, *and* $S(t)$ *singular, if* $\rho(y_k, y) \to 0$ *as* $k \to \infty$, *and* $z_k(t) \to z(t)$ *as* $k \to \infty$ *pointwise in* (t_1, t_2); *then* $X(t) = (y(t), Z(t))$, $t_1 \leqq t \leqq t_2$, *is a solution of the orientor field* (2.1).

3. Lower closure of functionals in integral form. As usual we denote by t the independent variable, by $x = (x^1, \cdots, x^n)$ the state variables, and by $u = (u^1, \cdots, u^m)$ the control variables.

As usual, let A be a closed subset of the tx-space $E_1 \times E_n$; for every $(t, x) \in A$ let $U(t, x)$ be a given subset of the u-space E_m; let M be the set of all (t, x, u) with $(t, x) \in A$, $u \in U(t, x)$; and let $\tilde{f}(t, x, u) = (f_0, f_1, \cdots, f_n) = (f_0, f)$ be a given continuous vector function on M. Let B be a closed subset of the $t_1 x_1 t_2 x_2$-space E_{2n+2}. We consider now the functional

$$I[x, u] = \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt. \tag{3.1}$$

We shall say that a pair $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, is *admissible* provided $x(t)$ is absolutely continuous in $[t_1, t_2]$, $u(t)$ is measurable in $[t_1, t_2]$, $(t, x(t)) \in A$ for $t \in [t_1, t_2]$, $u(t) \in U(t, x(t))$ a.e. in $[t_1, t_2]$, $dx/dt = f(t, x(t), u(t))$ a.e. in $[t_1, t_2]$, $f_0(t, x(t), u(t))$ is L-integrable in $[t_1, t_2]$, and $(t_1, x(t_1), t_2, x(t_2)) \in B$. Whenever we wish to disregard boundary conditions, we have only to take $B = E_{2n+2}$. We shall say that a vector function $x(t)$, $t_1 \leqq t \leqq t_2$, is an *admissible trajectory* if there is at least one vector function $u(t)$, $t_1 \leqq t \leqq t_2$, such that the pair x, u is admissible. Analogously, we shall say that a vector function $u(t)$, $t_1 \leqq t \leqq t_2$, is an *admissible strategy* if there is at least one vector function $x(t)$, $t_1 \leqq t \leqq t_2$, such that the pair x, u is admissible.

Let $x(t)$, $t_1 \leqq t \leqq t_2$, be any absolutely continuous vector function (which is the limit in the metric ρ of admissible trajectories). If, for any sequence $x_k(t)$, $u_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, of admissible pairs with $\rho(x_k, x) \to 0$, $\liminf I[x_k, u_k] < +\infty$ as $k \to \infty$, there is some measurable function $u(t)$, $t_1 \leqq t \leqq t_2$, such that $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, is admissible, and

$$I[x, u] \leqq \liminf_{k \to \infty} I[x_k, u], \tag{3.2}$$

then we say that $I[x, u]$ has the property of *lower closure* at the trajectory $x(t)$, $t_1 \leqq t \leqq t_2$.

Before we prove a sufficient condition for lower closure, the following remarks are needed. First, if x is the limit in the ρ-metric of admissible trajectories as assumed, then by the remark in § 1 we know that $(t, x(t)) \in A$ for all $t \in [t_1, t_2]$, and $(t_1, x(t_1), t_2, x(t_2)) \in B$.

Furthermore, if we know that the set M is closed, and that for every $(t, x) \in A$ the sets $Q(t, x) = f(t, x, U(t, x))$ are closed convex subsets of E_n satisfying property (Q) in A, then certainly $x'(t) \in Q(t, x(t))$ a.e. in $[t_1, t_2]$ by force of the closure Theorem 1.1; and then there is some measurable $u(t)$, $t_1 \leqq t \leqq t_2$, such that

$$u(t) \in U(t, x(t)), \qquad x'(t) = f(t, x(t), u(t)) \quad \text{a.e. in } [t_1, t_2], \tag{3.3}$$

by force of the implicit function theorem (see, e.g., [4]). As usual, we say that any such strategy $u(t)$ generates $x(t)$, $t_1 \leqq t \leqq t_2$. Obviously, in the concept of lower semicontinuity we require more; namely, we need a strategy u generating x for which (3.2) holds.

It may well occur that x is generated by some strategy $\bar{u}$ for which (3.2) does not hold. The following example displays two strategies u and $\bar{u}$, both generating the same trajectory x, such that (3.2) holds for u but not for $\bar{u}$.

Indeed, take $m = n = 1$, $t_1 = 0$, $t_2 = 1$, $f_0 = 1 + \cos \pi u$, $f = f_1 = \sin \pi u$, $u \in U = [-1 \leqq u \leqq 1]$, $x(t) = 0$, $0 \leqq t \leqq 1$, $A = E_2$. Now take $u_k(t) = \pm 2^{-1}$ according as $k^{-1}i \leqq t < k^{-1}i + (2k)^{-1}$ or $k^{-1} + (2k)^{-1} \leqq t < (i+1)k^{-1}$, $i = 0, 1, \cdots, k-1, k = 1, 2, \cdots$; and take $x_k(t) = t - k^{-1}i$ or $x_k(t) = k^{-1}(i+1) - t$, according as t is in one or the other set of intervals above. Then $x_k, u_k, k = 1, 2, \cdots$, is a sequence of admissible pairs, $0 \leqq x_k(t) \leqq (2k)^{-1}$, and $x_k \to x$ as $k \to \infty$ uniformly in $[0, 1]$. The trajectory $x(t) = 0$, $0 \leqq t \leqq 1$, is now generated by both $u(t) = 1$, $0 \leqq t \leqq 1$, and by $\bar{u}(t) = 0$, $0 \leqq t \leqq 1$. On the other hand,

$$I[x, u] = 0, \quad I[x, \bar{u}] = 2, \quad I[x_k, u_k] = 1, \qquad k = 1, 2, \cdots,$$

and thus relation (3.2) holds for u but not for $\bar{u}$.

As we shall see in § 7, the concept of lower closure introduced above contains as a particular case the usual concept of lower semicontinuity, in particular, the concept of lower semicontinuity for free problems.

4. Sufficient conditions for lower closure. Let A, $U(t, x)$, M, B, $f_0(t, x, u)$ and $f(t, x, u) = (f_1, \cdots, f_n)$ be defined as in § 3. For any $(t, x) \in A$ let $\tilde{Q}(t, x)$ be the set of all $\tilde{z} = (z^0, z^1, \cdots, z^n) = (z^0, z)$ with $z^0 \geqq f_0(t, x, u)$, $z = f(t, x, u)$ for some $u \in U(t, x)$.

THEOREM 4.1. *If the sets A, M, B are closed, and $f_0(t, x, u)$, $f(t, x, u) = (f_1, \cdots, f_n)$ are continuous on M, let us assume that the sets $\tilde{Q}(t, x)$ are closed, convex, and satisfy property* (Q) *at every point $(t, x) \in A$ with the exception perhaps of a set of points whose t-coordinate lies on a set of measure zero on the t-axis. Let us assume that, for some locally L-integrable scalar function $\psi(t) \geqq 0$ we have:*

$$(\psi) \qquad f_0(t, x, u) \geqq -\psi(t) \quad \textit{for all } (t, x, u) \in M,$$

with the exception perhaps of another set of points whose t-coordinate lies on a set of measure zero on the t-axis. Then the integral (3.1) *has the property of lower closure at every absolutely continuous vector function $x(t) = (x^1, \cdots, x^n)$, $t_1 \leqq t \leqq t_2$, which is the limit in the ρ-metric of admissible trajectories. In other words, for every absolutely continuous vector $x(t) = (x^1, \cdots, x^n)$, $t_1 \leqq t \leqq t_2$, and sequence $x_k(t)$, $u_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, of admissible pairs with $\rho(x_k, x) \to 0$, $\liminf I[x_k, u_k] < +\infty$ as $k \to \infty$, there is a measurable function $u(t)$, $t_1 \leqq t \leqq t_2$, such that $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, is admissible and $I[x, u] \leqq \liminf I[x_k, u_k]$.*

We proved Theorem 4.1 in [1a]. If i denotes the number $i = \liminf I[x_k, u_k]$ as $k \to \infty$, then it is not restrictive in this theorem to consider only a subsequence, say still $[k]$, with $I[x_k, u_k] \to i$ as $k \to \infty$. In the proof in [1a] of the lower closure Theorem 4.1 the functions

$$x_k^0(t) = \int_{t_{1k}}^{t} f_0(\tau, x_k(\tau), u_k(\tau))\, d\tau, \qquad t_{1k} \leqq t \leqq t_{2k}, \quad k = 1, 2, \cdots,$$

are taken into consideration, Helly's selection theorem is applied, and the closure Theorem 2.1 is used. This process to prove lower closure (or lower semicontinuity) (Cesari [1a]) was later used by E. J. McShane [3f], T. Nishiura [5], C. Olech [6].

By a well-known remark by C. S. Goodman, the hypothesis in Theorem 4.1 that the functions f_0, f are continuous on M can be replaced by the weaker assumption that f_0, f are continuous in x, u for every t and are measurable in t for every x, u. The proofs are essentially the same. Also, the following further remark concerning Theorem 4.1 may be of interest. To formulate it we shall denote by $[\bar{t}, T]$ a fixed interval containing all intervals $[t_{1k}, t_{2k}]$, $[t_1, t_2]$, and we shall extend $x_k(t)$, $x(t)$ on the whole interval $[\bar{t}, T]$ by continuity and constancy in each interval $[\bar{t}, t_{1k}]$, $[t_{2k}, T]$, $[\bar{t}, t_1]$, $[t_2, T]$. If in Theorem 4.1 we make the further assumption that the derivatives $x_k'(t)$ converge weakly in $L_1[\bar{t}, T]$ toward $x'(t)$ as $k \to \infty$, then the assumption concerning the sets $\tilde{\tilde{Q}}(t, x)$ satisfying property (Q) in Theorem 4.1 can be replaced by the following weaker assumption: There is a countable decomposition of $[\bar{t}, T]$ into disjoint measurable sets H_λ, $\lambda = 1, 2, \cdots$, such that, if A_λ denotes the set $A_\lambda = [(t, x) | (t, x) \in A, t \in H_\lambda]$, then the sets $\tilde{\tilde{Q}}(t, x)$ satisfy property (Q) in A_λ for almost every t, $\lambda = 1, 2, \cdots$. The assumption concerning the weak convergence of x_k' to x' in L_1 is usually satisfied in applications, but it is not requested in Theorem 4.1. A proof of the modification just mentioned is essentially contained in Cesari [1e, pp. 94–101, particularly p. 98]. The assumption of weak convergence of x_k' to x' in L_1 is essential. A counterexample, essentially due to A. Lasota and C. Olech [2], is reported in the Appendix at the end of this paper. On the other hand, under the hypothesis of weak convergence of the derivatives, the contention in Theorem 4.1 can be proved also under a different set of hypotheses (see C. Olech [6b]).

As we shall see in detail in § 7, Theorem 4.1 contains a number of lower semicontinuity theorems as corollaries.

Let us note here that whenever $f = u$, $n = m$ (e.g., for free problems); hence $dx/dt = u$, the functional can be written simply as $I[x]$, and the lower closure theorem reduces to a lower semicontinuity theorem.

COROLLARY 1. *Under the hypotheses of Theorem* 4.1 *with* $n = m$, $f = u$; *hence* $dx/dt = u$ *and*

$$\tilde{\tilde{Q}}(t, x) = \{(z^0, u) | z^0 \geqq f_0(t, x, u), u \in U(t, x)\} \subset E_{n+1};$$

if $x_k \to x$ *in the* ρ*-metric, and* $\liminf I[x_k] < \infty$, *then* $I[x] \leqq \liminf I[x_k]$.

Note that in Theorem 4.1 (as well as in the closure theorems, Theorems 1.1 and 2.1) no topology has been chosen for the measurable functions $u(t)$, $t_1 \leqq t \leqq t_2$, under consideration, and hence the question of what happens if the functions u_k converge toward some function u_0 does not arise. Here we may well assume that all intervals $[t_{1k}, t_{2k}]$, $[t_1, t_2]$ are all contained in a fixed interval $[\bar{t}, T]$, $\bar{t}$, T finite, and we may extend all functions u_k to the fixed interval $[\bar{t}, T]$ by defining them to be constant and equal to a fixed $\bar{u}$ (say, $\bar{u} = 0$) outside $[t_{1k}, t_{2k}]$. If we assume that all functions u_k are in L_p (that is, each component u_k^i is in $L_p[\bar{t}, T]$) and that the functions u_k converge weakly in L_p toward some element u_0 of L_p, then we may ask whether the pair x, u_0 is admissible, and whether the relation $I[x, u_0]$

$\leqq \liminf I[x_k, u_k]$ holds. The following corollaries of Theorem 4.1 give affirmative answers to the question being investigated under suitable assumptions.

To state Corollary 2 we shall need the sets

$$\tilde{\tilde{Q}}'(t, x) = \{(z^0, z, \zeta) | z^0 \geqq f_0(t, x, u), z = f(t, x, u), \zeta = u, u \in U(t, x)\}$$
$$= \{(z^0, z, u) | z^0 \geqq f_0(t, x, u), z = f(t, x, u), u \in U(t, x)\} \subset E_{n+m+1}.$$

COROLLARY 2. *Under the conditions of Theorem* 4.1 *with the sets* $\tilde{\tilde{Q}}$ *replaced by the sets* $\tilde{\tilde{Q}}'(t, x)$, *if the functions* u_k *are in* $(L_p)^m$, $p > 1$, *and* $u_k \to u_0$ *weakly in* $(L_p)^m$ *as* $k \to \infty$, *then the pair* x, u_0 *is admissible, and* $I[x, u_0] \leqq \liminf I[x_k, u_k]$. *The same statement holds for* $p = 1$ *provided we know in addition that the functions* u_k *are equiabsolutely integrable.*

Proof of Corollary 2. Let us consider the auxiliary control problem with n and m replaced by $n + m$ and m, the same control variable $u = (u^1, \cdots, u^m)$, the state variable $x = (x^1, \cdots, x^n)$, $y = (y^1, \cdots, y^m)$, same functional and constraints, and the $n + m$ differential equations

$$dx/dt = f(t, x, u), \qquad dy/dt = u.$$

Here A is replaced by the closed set $A' = A \times E_m$, and the sets $\tilde{\tilde{Q}}(t, x)$ by the sets $\tilde{\tilde{Q}}'(t, x)$ which are assumed to be closed, convex, and to satisfy property (Q). Now the functions

$$y_k(t) = \int_{t_{1k}}^{t} u_k(\tau)\, d\tau, \qquad t_{1k} \leqq t \leqq t_{2k}, \quad k = 1, 2, \cdots,$$

are equiabsolutely continuous and equibounded. We first extract a subsequence $[k_s]$ such that $I[x_{k_s}, u_{k_s}] \to i = \liminf I[x_k, u_k]$, and we further refine it in such a way that also $y_{k_s} \to y$ in the ρ-metric, where y is now absolutely continuous in $[t_1, t_2]$. By Theorem 4.1, there is a measurable $u(t)$, $t_1 \leqq t \leqq t_2$, such that x, u is admissible for the new problem and $I[x, u] \leqq i$; besides, $dx/dt = f(t, x(t), u(t))$, $dy/dt = u(t)$ a.e. in $[t_1, t_2]$. On the other hand, $u_{k_s} \to u_0$ weakly in L_p,

$$\int_{\bar{t}}^{t} u_{k_s}(\tau)\, d\tau \to \int_{\bar{t}}^{t} u_0(\tau)\, d\tau \quad \text{as } s \to \infty$$

for every t; hence, $y_{k_s} \to y(t)$ for every $t \in (t_1, t_2)$, or $y(t) = \int_{\bar{t}}^{t} u_0(\tau)\, d\tau$, and $u(t) = dy/dt = u_0(t)$ a.e. in $[t_1, t_2]$. This proves the corollary.

Note that in Corollary 2 the sequence $\|u_k\|_p$ is certainly bounded ($p \geqq 1$), and for $p > 1$, the functions u_k are certainly equiabsolutely integrable. This is the case even for $p = 1$ under suitable "growth" conditions. A suitable growth condition is the following one: (ε_0) For each $\varepsilon > 0$ there is some integrable function $\psi_\varepsilon(t) \geqq 0$, $\bar{t} \leqq t \leqq T$, such that

$$|u| \leqq \psi_\varepsilon(t) + \varepsilon f_0(t, x, u)$$

for all $(t, x, u) \in M$. An analogous growth condition has been considered in [1b] for a different purpose.

It is easy to see that in Corollary 2 the weak convergence of u_k in L_p for $p \geqq 1$ and the additional assumption that the u_k are equiabsolutely integrable if $p = 1$

can all be replaced by the single simple assumption: $y_k \to y$ in the ρ-metric, where

$$y_k(t) = \int_{t_{1k}}^{t} u_0(\tau)\, d\tau, \qquad t_{1k} \leqq t \leqq t_{2k}, \quad k = 1, 2, \cdots,$$

and

$$y(t) = \int_{t_1}^{t} u_0(\tau)\, d\tau, \qquad t_1 \leqq t \leqq t_2.$$

The proof is the same with obvious simplifications, and is still based on Theorem 4.1. For instance, for $m = 1$, the sequence $u_k(t) = k^{1/2} \sin kt$, $0 \leqq t \leqq 2\pi$, $k = 1, 2, \cdots$, converges in this sense to $u_0 = 0$, but the sequence u_k is not weakly convergent, say in L_1, since $\|u_k\|_1 = 4k^{1/2}$ is not bounded.

In Corollary 2 the hypothesis that the sets $\tilde{\tilde{Q}}'(t, x)$ be convex cannot be disregarded, as the following example shows. Take $n = 1$, $m = 1$, $f_0 = u$, $f = u^2$, $U = \{\pm 1\}$; hence $\tilde{\tilde{Q}}'$ is the fixed closed nonconvex set made up of the two half-straight-lines $(z = u = 1, z^0 \geqq 1)$ and $(z = 1, u = -1, z^0 \geqq -1)$. Also, take $t_{1k} = t_1 = 0$, $t_{2k} = t_2 = 1$, $u_k(t) = 1$ for $i/k \leqq t \leqq i/k + 1/2k$, $i = 0, 1, \cdots, k - 1$, $u_k(t) = -1$ in the complementary invervals, $x_k(t) = x(t) = t$, $0 \leqq t \leqq 1$, $k = 1, 2, \cdots$. Then $u_k \to u_0$ weakly in any L_p $(p \geqq 1)$ with $u_0(t) = 0$, $0 \leqq t \leqq 1$, and now the pair x, u_0 is not admissible since $0 \notin U$. On the other hand,

$$\tilde{\tilde{Q}} = \{(z^0, z) | z^0 \geqq -1, z = 1\}$$

is convex and all conditions of Theorem 4.1 are satisfied. If we take $u(t) = -1$, $0 \leqq t \leqq 1$, then the pair x, u is admissible with $I[x, u] = -1$, while $I[x_k, u_k] = 0$, $k = 1, 2, \cdots$.

An analogous example where both $\tilde{\tilde{Q}}$ and U are convex, where x, u_0 is admissible, and yet $I[x, u_0] > \liminf I[x_k, u_k]$, is as follows. Take $m = n = 1$, $U = E_1$, $f_0(-u) = f_0(u)$, $f(-u) = -f(u)$, $f_0 = 2 - u$ for $0 \leqq u \leqq 2$, $f_0 = 0$ for $u \geqq 2$, $f(u) = u$ for $0 \leqq u \leqq 1$, $f(u) = 2 - u$ for $1 \leqq u \leqq 2$, $f(u) = 0$ for $u \geqq 2$. Then $\tilde{\tilde{Q}}$ is the fixed closed convex set

$$\{(z^0, z) | z^0 \geqq |z| \text{ for } -1 \leqq z \leqq 1\} \subset E_2,$$

while $\tilde{\tilde{Q}}'$ is a fixed closed nonconvex set in E_3. If we take $t_{1k} = t_1 = 0, t_{2k} = t_2 = 1$, $k = 1, 2, \cdots$, and $u_k(t) = 1$ for $i/k \leqq t \leqq i/k + 1/2k$, $i = 0, 1, \cdots, k - 1$, $u_k(t) = -1$ in the complementary intervals, then $u_k \to u_0$ weakly in any L_p, $p \geqq 1$, with $u_0(t) = 0$, $0 \leqq t \leqq 1$. If we take $x_k(t) = t - i/k$, or $x_k(t) = (i + 1)/k - t$ according as t belongs to one or the other set of intervals above, then $x_k' = \bar{u}_k$, $x_k \to x$ uniformly and $x_k' \to x'$ weakly in any L_p as $k \to \infty$ with $x(t) = 0, 0 \leqq t \leqq 1$. The pair x, u_0 is admissible with $I[x, u_0] = 2$, while $I[x_k, u_k] = 1$ for all k. On the other hand, the conditions of Theorem 4.1 are all satisfied, and there is therefore some measurable u such that x, u is admissible and $I[x, u] \leqq \liminf I[x_k, u_k]$. One such u is $u(t) = 2$, $0 \leqq t \leqq 1$. Another example of the same type, where $\tilde{\tilde{Q}}$ is a fixed closed convex set, f_0 is convex in u, U is convex, but $\tilde{\tilde{Q}}'$ is fixed and closed but not convex, is as follows. Take $m = n = 1, f_0 = 0, f = u^2, U = [-1 \leqq u \leqq 1]$, so that $\tilde{\tilde{Q}} = \{(z^0, z) | z^0 \geqq 0, 0 \leqq z \leqq 1\}$. Take $u_k(t) = \pm 1$ as in the previous example, so that $u_k \to u_0$ weakly in any L_p with $u_0(t) = 0$, $0 \leqq t \leqq 1$. On the

other hand, $x_k(t) = x(t) = t$, $0 \leqq t \leqq 1$, $I[x_k, u_k] = 0$, $k = 1, 2, \cdots$, while x, u_0 is not admissible. For $u(t) = 1$, $0 \leqq t \leqq 1$, x, u is admissible and $I[x, u] = 0$.

Other examples of the same kind can be found in [1i, particularly p. 13] and [1j, particularly p. 53].

Note that the requirement in Corollary 2 that the sets $\tilde{\tilde{Q}}(t, x)$ be convex implies that, for every $(t, x) \in A$, the sets $U(t, x)$ are convex, $f(t, x, u)$ is linear in u (hence of the form $f = B(t, x)u + C(t, x)$, B an $n \times m$ matrix, C an $n \times 1$ matrix), and $f_0(t, x, u)$ is convex in u.

The following Corollary 3 for general Lagrange problems (with f linear in u) extends both Corollaries 1 and 2. To state Corollary 3 we shall need the sets

$$\tilde{\tilde{Q}}^*(t, x) = \{(z^0, \zeta) | z^0 \geqq f_0(t, x, u), \zeta = u, u \in U(t, x)\}$$
$$= \{(z^0, u) | z^0 \geqq f_0(t, x, u), u \in U(t, x)\} \subset E_{m+1}.$$

COROLLARY 3. *Under the conditions of Theorem* 4.1 *with the sets* $\tilde{\tilde{Q}}^*(t, x) \subset E_{m+1}$ *replacing the sets* $\tilde{\tilde{Q}}$, *if* f *is linear in* u (*or* $f = B(t, x)u + C(t, x)$, $(t, x) \in A$, $u \in U(t, x)$, *the matrices* B *and* C *with entries continuous in* A), *if the functions* u_k *are in* $(L_p)^m$, $p > 1$, *and* $u_k \to u_0$ *weakly in* $(L_p)^m$ *as* $k \to \infty$, *then the pair* x, u_0 *is admissible, and* $I[x, u_0] \leqq \liminf I[x_k, u_k]$. *The same statement holds for* $p = 1$ *provided we know in addition that the functions* u_k *are equiabsolutely integrable.*

Proof of Corollary 3. Let us consider the auxiliary control problem with $n + m$ control variables $u = (u^1, \cdots, u^m)$, $v = (v^1, \cdots, v^n)$, with $n + m$ state variables $x = (x^1, \cdots, x^n)$, $y = (y^1, \cdots, y^m)$, the same functional $I = \int_{t_1}^{t_2} f_0(t, x, u)\, dt$, with constraints $(t, x) \in A$, $u \in U(t, x)$, $v \in V = E_n$, and $n + m$ differential equations

$$dx/dt = v, \qquad dy/dt = u.$$

Here A is replaced by the closed set $A' = A \times E_m$, and the sets $Q(t, x)$ of Theorem 4.1 are replaced by the sets $Q^*(t, x) \times E_n \subset E_{n+m+1}$ which certainly are closed convex and satisfy property (Q). We take

$$v_k(t) = x_k'(t),$$

$$y_k(t) = \int_{t_{1k}}^{t} u_k(\tau)\, d\tau, \qquad t_{1k} \leqq t \leqq t_{2k}, \quad k = 1, 2, \cdots,$$

and then the functions $y_k(t)$ are equiabsolutely continuous and equibounded. We first extract a subsequence $[k_s]$ such that $I_{k_s} \to i = \liminf I_k$, and we further refine it in such a way that also $y_{k_s} \to y$ in the ρ-metric, where y is now absolutely continuous in $[t_1, t_2]$. By Theorem 4.1 there are measurable $u(t)$, $v(t)$, $t_1 \leqq t \leqq t_2$, such that x, y, u, v, is admissible for the new problem, and $I \leqq i$, besides

$$x'(t) = v(t), \qquad y'(t) = u(t) \in U(t, x(t)) \text{ a.e.},$$

$$I = \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt.$$

On the other hand, for every k, we have $(t, x_k(t)) \in A$, and

$$x_k'(t) = v_k(t), \quad x_k'(t) = B(t, x_k(t))u_k(t) + C(t, x_k(t)), \quad y_k'(t) = u_k(t)$$

a.e. in $[t_{1k}, t_{2k}]$; and hence in a fixed interval $[\bar{t}, T]$ after suitable extensions,

$[t_{1k}, t_{2k}] \subset [\bar{t}, T]$, $k = 1, 2, \cdots$. Then $B(t, x_k(t)) \to B(t, x(t))$, $C(t, x_k(t)) \to C(t, x(t))$ as $k \to \infty$ in the ρ-metric, as well as uniformly in $[\bar{t}, T]$, while $u_k \to u_0$ weakly in L_p. Then $x_k' \to x'$ weakly in L_p and hence

$$x'(t) = B(t, x(t))u_0(t) + C(t, x(t)) \quad \text{a.e. in } [t_1, t_2].$$

If we denote by W-lim the usual weak limit in L_p, we have also

$$u(t) = y'(t) = \underset{s\to\infty}{W\text{-lim}}\, y_{k_s}'(t) = \underset{k\to\infty}{W\text{-lim}}\, u_k(t) = u_0(t)$$

a.e. in $[t_1, t_2]$, or $u = u_0$ a.e., and

$$x'(t) = f(t, x(t), u_0(t)), \quad (t, x(t)) \in A, \quad u_0(t) \in U(t, x(t))$$

a.e. in $[t_1, t_2]$. Then

$$I[x, u_0] = I[x, u] \leqq i = \liminf_{k\to\infty} I[x_k, u_k].$$

This proves Corollary 3. Note that the hypothesis of convexity of $Q^*(t, x)$ certainly implies the convexity of $U(t, x)$.

In Corollary 3 for $p = 1$ the functions u_k are certainly equiabsolutely integrable under the growth condition (ε_0) mentioned below Corollary 2.

Note that if both f and f_0 are linear in u, or

$$f = B(t, x)u + C(t, x), \qquad f_0 = b(t, x)u + c(t, x),$$

where B, C, b, c are $n \times m$, $1 \times n$, $1 \times m$, 1×1 matrices with entries continuous in A, then in Corollary 3 the conditions on the sets Q^* can be dropped. In other words we have: If $x_k(t), u_k(t), t_{1k} \leqq t \leqq t_{2k}, k = 1, 2, \cdots$, is a sequence of admissible pairs with $x_k \to x$ in the ρ-metric and $u_k \to u_0$ weakly in L_p, $p > 1$, then x, u_0 is admissible and $I[x, u_0] = \lim I[x_k, u_k]$. The same holds for $p = 1$ if the functions u_k are known to be equiabsolutely integrable. The proof is the same as above with obvious simplifications.

Remark 1. Condition (ψ) in Theorem 4.1 and in the corollaries will be drastically reduced in Theorem 4.2 below. Simple conditions under which the sets $\tilde{\tilde{Q}}$ above, if convex, are closed and satisfy property (Q) as requested will be given in § 8.

Remark 2. In the sufficient condition for lower closure, Theorem 4.1, and in the corollaries, it is enough to request that the sets $\tilde{\tilde{Q}}(t, x)$ be closed, convex, and have property (Q) at the points $(t, x(t)) \in A$ for almost all $t \in [t_1, t_2]$. In this form, and under suitable regularity hypotheses, the convexity assumption of the sufficient condition, Theorem 4.1, for lower closure will be shown to be necessary also.

Remark 3. Condition (ψ) in Theorem 4.1 and in the corollaries is satisfied if, for instance, $f_0(t, x, u) \geqq 0$ for all $(t, x, u) \in M$, or $f_0(t, x, u) \geqq v$ for all $(t, x, u) \in M$, where v is some real constant. Nevertheless, condition (ψ) in Theorem 4.1 can be reduced. For instance, we may replace it by the following weaker assumption:

(ψ') For every compact subset A_0 of A there is a locally integrable function $\psi_0(t)$ (which may depend on A_0) such that $f_0(t, x, u) \geqq \psi_0(t)$ for all $(t, x) \in A_0$, $u \in U(t, x)$.

The proof is the same since we can include all trajectories x and x_k in a unique compact subset A_0 of A.

A more drastic generalization of Theorem 4.1 will be given below (Theorem 4.2), where we shall use the following much weaker form of the condition.

(ψ^*) For every $(\bar t, \bar x) \in A$ there are a neighborhood $N(\bar t, \bar x)$ of $(\bar t, \bar x)$ in A, a locally integrable function $\psi(t)$, and real numbers $b_1, \cdots, b_n$ (all $b_1, \cdots, b_n$ and ψ may depend on $\bar t$, $\bar x$, N) such that

$$\bar f(t, x, u) = f_0(t, x, u) - \sum_{j=1}^{n} b_j f_j(t, x, u) \geqq \psi(t) \tag{4.1}$$

for all $(t, x) \in N(\bar t, \bar x)$, $u \in U(t, x)$, with the exception perhaps of a set of points (t, x) whose t-coordinate lies on a set of measure zero on the t-axis.

Remark 4. We shall note here that, under condition (ψ^*), it is natural to consider the sets

$$\tilde{\tilde Q}(t, x) = \{(z^0, z) | z^0 \geqq f_0(t, x, u),\ z = f(t, x, u),\ u \in U(t, x)\},$$

or the analogous sets

$$\tilde{\tilde Q}^*(t, x) = \{(Z^0, Z) | Z^0 \geqq \bar f_0(t, x, u),\ Z = f(t, x, u),\ u \in U(t, x)\}.$$

It is easy to see that the sets $\tilde{\tilde Q}$ are closed, or convex, or satisfy property (Q) if and only if the same occurs for the sets $\tilde{\tilde Q}^*$. Indeed, the sets above are transformed into one another by the fixed affine transformation $Z^0 = z^0 - bz$, $Z = z$.

THEOREM 4.2. *Let A, B, $U(t, x)$, M, $f(t, x, u)$, $f_0(t, x, u)$ be as in Theorem* 4.1, *and let us assume that condition (ψ^*) holds. With $N(\bar t, \bar x)$ as in condition (ψ^*), and for every $(t, x) \in N(\bar t, x)$, let $\tilde{\tilde Q}(t, x)$ denote the set of all $\tilde z = (z^0, z^1, \cdots, z^n) = (z^0, z)$ with $z^0 \geqq \bar f_0(t, x, u)$, $z = f(t, x, u)$ for $u \in U(t, x)$, and assume that the sets $\tilde{\tilde Q}(t, x)$ are closed, convex, and satisfy property* (Q) *at all points $(t, x) \in N(\bar t, \bar x)$, with the exception perhaps of a set of points whose t-coordinate lies on a set of measure zero on the t-axis. Then the integral* (3.1) *has the property of lower closure at every absolutely continuous vector function $x(t)$, $t_1 \leqq t \leqq t_2$, which is the limit in the ρ-metric of admissible trajectories.*

Theorem 4.2 holds even if the continuity of f_0, f is replaced by the weaker assumption that f_0, f are continuous in x, u for every t, and measurable in t for every x, u. The same argument mentioned under Theorem 4.1 applies.

Proof. Let $x(t)$, $t_1 \leqq t \leqq t_2$, be any absolutely continuous function as in the text, and $x_k(t)$, $u_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, be a sequence of admissible pairs with $\rho(x_k, x) \to 0$, $\liminf I[x_k, u_k] < +\infty$. Let A_0 be a compact neighborhood (containing the graph of x and all x_k). By hypothesis, for every $(\bar t, \bar x) \in A_0$ there are numbers $\delta > 0$, $b_1, \cdots, b_n$ real, and a locally integrable function $\psi(t)$, $-\infty < t < +\infty$, such that $\bar f_0(t, x, u) \geqq \psi(t)$ for all $(t, x) \in N_{2\delta}(\bar t, \bar x)$ and $u \in U(t, x)$. We consider the smaller neighborhoods $N_\delta(\bar t, \bar x)$ which we think of as open (in A). These too form a cover of the compact set A_0. Thus, finitely many of these N_δ cover A_0, say $N_{\delta_\gamma}(t_\gamma, x_\gamma)$, $\gamma = 1, \cdots, s$. Let $\delta_\gamma > 0$, $b_{\gamma 1}, \cdots, b_{\gamma n}$, ψ_γ be the corresponding elements, so that

$$\bar f_{0\gamma}(t, x, u) = f_0(t, x, u) - \sum_{j=1}^{n} b_{\gamma j} f_j(t, x, u) \geqq \psi_\gamma(t)$$

for all $(t, x) \in N_{2\delta_\gamma}(t_\gamma, x_\gamma)$, $u \in U(t, x)$, $\gamma = 1, \cdots, s$, and $\bigcup_{\gamma=1}^{s} N_{\delta_\gamma}(t_\gamma, x_\gamma) \supset A_0$. Let $b = \max [|b_{\gamma j}|,\ j = 1, \cdots, n,\ \gamma = 1, \cdots, s]$, $\delta_0 = \min [\delta_\gamma,\ \gamma = 1, \cdots, s]$.

Since $(t, x(t)) \in A_0$ for all $t_1 \leqq t \leqq t_2$, we can divide the arc C_0: $x = x(t)$, $t_1 \leqq t \leqq t_2$, into finitely many subarcs, say C_σ, $\sigma = 1, \cdots, N$, each C_σ completely contained in some neighborhood $N_{\delta_\gamma}(t_\gamma, x_\gamma)$. Thus, we have for the arcs C_σ the representations C_σ: $x = x(t)$, $\tau_{\sigma-1} \leqq t \leqq \tau_\sigma$, with $t_1 = \tau_0 < \tau_1 < \cdots < \tau_N = t_2$, and each C_σ lies in a certain $N_{b_\gamma}(t_\gamma, x_\gamma)$ which now remains associated with C_σ. Since $\rho(x_k, x) \to 0$ as $k \to \infty$, and hence $t_{1k} \to t_1$, $t_{2k} \to t_2$, we see that for all k sufficiently large we have $t_{1k} < \tau_1 < \cdots < \tau_{N-1} < t_{2k}$. Thus, for all k sufficiently large, the arc C_k: $x = x_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, is divided into the same number N of subarcs, say C_{k0}: $x = x_k(t)$, $\tau_{\sigma-1} \leqq t \leqq \tau_\sigma$, $\sigma = 1, \cdots, N$, where now $\tau_0 = t_1$ must be replaced by t_{1k} and $\tau_N = t_2$ must be replaced by t_{2k}. Also, for all k sufficiently large, say for $k \geqq k_0$, the arc $C_{k\sigma}$ is completely contained in $N_{2\delta_\gamma}(t_\gamma, x_\gamma)$ for the same γ we have already associated with C_σ. Thus, for $k \geqq k_0$, C_σ lies in some $N_{\delta_\gamma}(t_\gamma, x_\gamma)$ and $C_{k\sigma}$ in $N_{2\delta_\gamma}(t_\gamma, x_\gamma)$. Also, $C_{k\sigma} \to C_\sigma$ as $k \to \infty$ in the sense that the ρ-distance approaches zero as $k \to \infty$. We shall now consider for each $\sigma = 1, \cdots, N$, the auxiliary functional

$$J = \int_{t'}^{t''} \bar{f}_0(t, x(t), u(t))\, dt$$

for all admissible pairs x, u with the graph of x lying in $N_{2\delta_\gamma}(t_\gamma, x_\gamma)$. Here by admissible we mean that the conditions listed in § 3 are satisfied with A replaced by $N_{2\delta_\gamma}(t_\gamma, x_\gamma)$, and of course $\bar{f}_0(t, x(t), u(t))$ L-integrable as usual.

For each σ we may now apply Theorem 4.1 to the arc C_σ, the sequence $C_{k\sigma}$, $k = 1, 2, \cdots$, and the functional J. We conclude that each C_σ is admissible and that

$$J[C_\sigma] \leqq \liminf_{k \to \infty} J[C_{\sigma k}], \qquad \sigma = 1, \cdots, N. \tag{4.2}$$

More precisely, for each σ, there is a measurable $u(t)$, $\tau_{\sigma-1} \leqq t \leqq \tau_\sigma$, such that the pair $x(t)$, $u(t)$, $\tau_{\sigma-1} \leqq t \leqq \tau_\sigma$, is admissible for the functional J; in particular, $u(t) \in U(t, x(t))$, $dx/dt = f(t, x(t), u(t))$, $\tau_{\sigma-1} \leqq t \leqq \tau_\sigma$ (a.e.), $\sigma = 1, \cdots, N$, and the expression

$$\bar{f}_0(t, x(t), u(t)) = f_0(t, x(t), u(t)) - \sum_{j=1}^{n} b_{\gamma j} f_j(t, x(t), u(t))$$

is L-integrable in $[\tau_{\sigma-1}, \tau_\sigma]$. Since the functions f_j here are certainly L-integrable in the same interval (as derivatives of the absolutely continuous functions $x^i(t)$ in $[\tau_{\sigma-1}, \tau_\sigma]$), we conclude that $f_0(t, x(t), u(t))$ itself is L-integrable in each $[\tau_{\sigma-1}, \tau_\sigma]$ and hence in the whole of $[t_1, t_2]$. We have proved that the pair $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, is admissible for the original integral I.

Now, given $\varepsilon > 0$, we deduce from (4.2) that there is some $\bar{k} \geqq k_0$ such that, for $k \geqq \bar{k}$, we have

$$J[C_{k\sigma}] - J[C_\sigma] > -\varepsilon/N, \quad \rho(C_{k\sigma}, C_\sigma) < \varepsilon/(Nnb), \qquad \sigma = 1, \cdots, N. \tag{4.3}$$

Now we have

$$\begin{aligned} I[x, u] &= \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt = \sum_{\sigma=1}^{N} \int_{\tau_{\sigma-1}}^{\tau_\sigma} f_0(t, x(t), u(t))\, dt \\ &= \sum_{\sigma=1}^{N} \left\{ \int_{\tau_{\sigma-1}}^{\tau_\sigma} \bar{f}_0(t, x(t), u(t))\, dt + \sum_{j=1}^{n} b_{\sigma j}[x^j(\tau_\sigma) - x^j(\tau_{\sigma-1})] \right\}, \end{aligned}$$

$$I[x_k, u_k] = \int_{t_{1k}}^{t_{2k}} f_0(t, x_k(t), u_k(t))\, dt = \sum_{\sigma=1}^{N} \int_{\tau_{\sigma-1}}^{\tau_\sigma} f_0(t, x_k(t), u_k(t))\, dt$$

$$= \sum_{\sigma=1}^{N} \left\{ \int_{\tau_{\sigma-1}}^{\tau_\sigma} \bar{f}_0(t, x_k(t), u_k(t))\, dt + \sum_{j=1}^{n} b_{\sigma j}[x_k^j(\tau_\sigma) - x^j(\tau_{\sigma-1})] \right\},$$

where we have written $b_{\sigma j}$ instead of $b_{\gamma j}$ where γ is the index we have associated to σ, and where we have written $\tau_1 = t_1$ and $\tau_N = t_2$ instead of t_{1k} and t_{2k} in the expression for $I[x_k, u_k]$. By taking the difference we now have, remembering relations (4.3),

$$I[x_k, u_k] - I[x, u] > -N(\varepsilon/N)$$

$$+ \sum_{\sigma=1}^{N} \sum_{j=1}^{n} b_{\sigma j}\{[x_k^j(\tau_\sigma) - x^j(\tau_\sigma)] + [x_k^j(\tau_{\sigma-1}) - x^j(\tau_{\sigma-1})]\},$$

where each bracket is now in absolute value less than or equal to $(Nnb)^{-1}\varepsilon$. We conclude that for all $k \geqq \bar{k}$ we have

$$I[x_k, u_k] - I[x, u] > -\varepsilon - Nnb[2(Nnb)^{-1}\varepsilon] = -3\varepsilon.$$

Because ε is arbitrary, we have proved the lower closure of I at x.

Remark 5. As shown in [1ab], lower closure theorems easily yield existence theorems for the minimum. Indeed, for A compact, we have only to consider any closed nonempty class Ω of admissible pairs x, u with $I[x, u] \leqq M$, and guarantee that the corresponding trajectories x are equiabsolutely continuous. To this purpose, growth conditions on f and f_0 such as (γ) or (ε) below suffice, or alternatively, a condition such as (μ) below (see [1ab] for details, for the case of A not compact, and for the simplifications occurring when the control space U is bounded).

Remark 6. We report below for the sake of comparison a lower closure and existence theorem proved by C. Olech [6a]. Let A, $U(t, x)$, M, B, $f_0(t, x, u) = (f_1, \cdots, f_n)$, $Q(t, x) \subset E_{n+1}$, be defined as in § 3. Let $\tilde{z} = (z^0, z) = (z^0, z^1, \cdots, z^n)$ be the usual variable in E_{n+1}. We shall denote by $\tilde{c} = (c^0, c) = (c^0, c^1, \cdots, c^n)$, $\tilde{d} = (d^0, d) = (d^0, d^1, \cdots, d^n)$ points in E_{n+1}, and by $\tilde{c} \cdot \tilde{z} = c^0 z^0 + c^1 z^1 + \cdots + c^n z^n$ the usual inner product in E_{n+1}. For any point $(t, x) \in A$ let $C(t, x)$ denote the set

$$C(t, x) = \{\tilde{c} \in E_{n+1} | \tilde{z} + \lambda \tilde{c} \in \tilde{\tilde{Q}}(t, x) \text{ for all } \lambda \geqq 0 \text{ and all } \tilde{z} \in \tilde{\tilde{Q}}(t, x)\}.$$

Obviously, $C(t, x)$ is a gone with vertex the origin in E_{n+1}. The set $C(t, x)$ is said to be the *asymptotic cone* of the set $\tilde{\tilde{Q}}(t, x)$. It is easy to see that if $\tilde{\tilde{Q}}(t, x)$ is convex, so is $C(t, x)$; if $\tilde{\tilde{Q}}(t, x)$ is closed, so is $C(t, x)$. Also, we shall denote by Γ the cone in E_{n+1} made up of only the positive half-straight-line $\Gamma = \{(c^0, c) | c^2 \geqq 0, c = 0\}$. Obviously, $\Gamma \subset \tilde{\tilde{Q}}(t, x)$ for all $(t, x) \in A$.

By the polar cone $C^0(t, x)$ of $C(t, x)$ is meant as usual the set

$$C^0(t, x) = \{\tilde{d} = (d^0, d) \in E_{n+1} | \tilde{d} \cdot \tilde{c} \leqq 0 \text{ for all } \tilde{c} \in C(t, x)\}.$$

Then, for $C(t, x) = \Gamma$, we have

$$C^0(t, x) = \Gamma^0 = \{\tilde{d} = (d^0, d) | d^0 \leqq 0, d \in E_n\},$$

and hence int Γ^0 is the set of all $\tilde{d} = (d^0, d)$ with $d^0 < 0$, $d \in E_n$.

THEOREM 4.3 (A lower closure theorem). *Let the sets M, B be closed and A compact, let $f_0(t, x, u)$, $f(t, x, u) = (f_1, \cdots, f_n)$ be continuous on M, and let us assume that the sets $\tilde{\tilde{Q}}(t, x)$ have the following properties*:

(a) *the sets $\tilde{\tilde{Q}}(t, x)$ are closed and convex*;

(b) *the sets $\tilde{\tilde{Q}}(t, x)$ satisfy property* (U) *in x for every t*;

(c) $C(t, x) = \Gamma$ *for all* $(t, x) \in A$;

(d) *for every $r > 0$ and $\tilde{c} = (c^0, c)$ with $c^0 < 0$ there is a locally integrable function $\psi(t; r, c)$, $t \in E_1$, which may depend on r and c, such that $\tilde{c} \cdot \tilde{z} \leqq \psi(t; r, c)$ for all $(t, x, u) \in M$ with $|x| \leqq r$ and all $\tilde{z} = (z^0, z) \in \tilde{\tilde{Q}}(t, x)$.*

Then the integral (3.1) *has the property of lower closure at every absolutely continuous vector function $x(t) = (x^1, \cdots, x^n)$, $t_1 \leqq t \leqq t_2$, which is the limit in the ρ-metric of admissible trajectories, and the minimum of the functional exists in every closed class Ω of admissible trajectories.*

As for Theorem 4.1 and Theorem 4.2 the hypothesis of continuity of f_0 and f can be replaced by the weaker assumption that f_0 and f are continuous in x, u for every t, and measurable in t for every x, u [see [6a]). Theorem 4.3 is proved by C. Olech in [6a] (reported here in the setting of pp. 175–176, in our notations, and for the problem in which we are interested). Condition (c) of Theorem 4.3 is not required in our Theorems 4.1 and 4.2, and condition (d) in Theorem 4.3 is more demanding than condition (ψ^*) required in Theorem 4.2. Indeed, if we take

$$b = (b_1, \cdots, b_n) = (-c^0)^{-1}c, \qquad \bar{\psi}(t) = (-c^0)^{-1}\psi,$$

then (d) can be written in the form $f_0 + b \cdot f \geqq \bar{\psi}$. In other words, in Theorem 4.3 it is required that for *every* $b \in E_n$ and $r > 0$ there is some $\bar{\psi}$ satisfying $f_0 + b \cdot f \geqq \bar{\psi}$. Condition ($\psi^*$) of Theorem 4.2 only requires that for every (t, x) there is *some* $b \in E_n$ such that $f_0 + b \cdot f \geqq \psi$ holds in a neighborhood of $(\bar{t}, \bar{x})$. The following examples illustrate these points. For instance, for $f_0 = (1 + u^2)^{1/2}$, $f = u$, $n = 1$, condition (c) is not satisfied; the same occurs for $f_0 = (1 + x^2u^2)^{1/2}$, $f = u$, $n = 1$. For $f_0 = 1 + |u|$, $f = |u|$, $n = 1$, condition (d) is not satisfied, while (ψ^*) is certainly satisfied.

It may be pointed out that condition (c) actually means that $f(t, x, u)$ is "of slower growth" than $f_0(t, x, u)$ in the following sense.

(γ) Given $\varepsilon > 0$ there is an $N \geqq 0$ such that $|f| \leqq \varepsilon f_0$ whenever $|f_0| \geqq N$.

A condition of this kind has been used in existence theorems, but lower closure Theorems 4.1 and 4.2 do not need a similar condition.

Another condition which has been used in existence theorems is the following one.

(ε) Given $\varepsilon > 0$ there is some function $\psi_\varepsilon(t) \geqq 0$ which is locally integrable and such that

$$|f(t, x, u)| \leqq \psi_\varepsilon(t) + \varepsilon f_0(t, x, u).$$

Under this condition then (d) holds. Indeed, given $b \in E_n$ take $\varepsilon > 0$ such that $\varepsilon|b| < 1$ and note that

$$f_0 + b \cdot f \geqq f_0 - |b||f| \geqq \varepsilon^{-1}(\varepsilon f_0 - |f|) \geqq \varepsilon^{-1}\psi(t).$$

Again, lower closure Theorems 4.1 and 4.2 require much less than (d) or (ε).

An example offered by C. Olech [5, p. 179] concerns the existence of the minimum and is similar to the one considered in [9, particularly § 2 and the first part of Example 4.1, p. 300], namely $\int_0^1 t^\alpha x'^2\,dt$, $0 < \alpha < 1$, where our existence theorems apply and use is made of condition (ε). Nevertheless, growth conditions (γ), or (ε), are not needed for existence in classes Ω satisfying constraints of the following type:

$$(\mu) \qquad \int_{t_1}^{t_2} |x'|^p\,dt \leqq M \quad \text{for some constants } p > 1,\ M > 0.$$

Condition (μ) and the less demanding conditions of a lower closure theorem such as Theorem 4.1 or 4.2 suffice.

Note that all conditions (γ), (ε), or (μ), and similar ones, only guarantee the equiabsolute integrability of the derivatives x_k' of the element x_k of a minimizing sequence $[x_k]$, so that we can extract a subsequence x_{k_s} whose elements x_{k_s} converge uniformly (since A is compact), and we could even request that the derivatives x_{k_s}' converge weakly in L_1 or L_p. In this sense we could state in Theorem 4.1 that it is enough in applications to know that the sets $\tilde{\tilde{Q}}(t, x)$ satisfy property (Q) in the subsets A_λ of a partition $[A_\lambda]$ of A into measurable sets as mentioned above. In Theorem 4.3 the combined hypotheses (a) to (d) guarantee that the sets $\tilde{\tilde{Q}}(t, x)$ satisfy property (Q) in x for almost every t, as proved by C. Olech [6a, pp. 168–169].

5. A variant of the lower closure property. Theorem 4.1 holds in a slightly stronger form. To formulate it we need, besides the sets $\tilde{\tilde{Q}}(t, x) \subset E_{n+1}$ of § 4, also the sets $Q(t, x) = f(t, x, U(t, x)) \subset E_n$. These sets $Q(t, x)$ are the projections on the z-space E_n of the sets $\tilde{\tilde{Q}}(t, x)$ of the z^0z-space E_{n+1}. Thus, if the sets $\tilde{\tilde{Q}}(t, x)$ are convex, so are the sets $Q(t, x)$. On the other hand, the sets $\tilde{\tilde{Q}}(t, x)$ may be closed, without the sets $Q(t, x)$ being so. This is shown by the example $n = 1$, $m = 1$, $U = \{u | -\infty < u < +\infty\}$, $f_0 = (1 + u^2)^{1/2}$, $f = \tan^{-1} u$, $-\pi/2 < f < \pi/2$. Then, Q and $\tilde{\tilde{Q}}$ are the fixed sets

$$Q = \{z | -\pi/2 < z < \pi/2\} \subset E_1,$$

$$\tilde{\tilde{Q}} = \{(z^0, z) | z^0 \geqq \sec z,\ -\pi/2 < z < \pi/2\} \subset E_2,$$

and $\tilde{\tilde{Q}}$ is closed, but Q is not. This example shows also that property (Q) for the sets $\tilde{\tilde{Q}}$ does not imply the same property for the sets Q. In the statement below we shall require that both the sets $\tilde{\tilde{Q}}(t, x)$ and the sets $Q(t, x)$ have property (Q).

THEOREM 5.1. *If we assume, in addition to the hypotheses of Theorem* 4.1, *that both the sets* $\tilde{\tilde{Q}}(t, x) \subset E_{n+1}$ *and the sets* $Q(t, x) \subset E_n$ *are closed, convex, and satisfy property* (Q) *at all points of* A *with the exception perhaps of a set of points whose* t*-coordinate lies on a set of measure zero on the* t*-axis, then for every sequence* $x_k(t)$, $u_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, *of admissible pairs, and any absolutely continuous vector function* $x(t) = (x^1, \cdots, x^n)$, $t_1 \leqq t \leqq t_2$, *with* $\rho(x_k, x) \to 0$ *as* $k \to \infty$, *there is a measurable function* $u(t)$, $t_1 \leqq t \leqq t_2$, *such that* $u(t) \in U(t, x(t))$, $x'(t) = f(t, x(t), u(t))$ *a.e. in* $[t_1, t_2]$, *and*

$$(5.1) \qquad I[x, u] \leqq \liminf_{k \to \infty} I[x_k, u_k].$$

If $\liminf I[x_k, u_k] < +\infty$, then certainly the pair x, u is admissible, and (5.1) holds.

An analogous variant of Theorem 4.2 also holds.

Proof of Theorem 5.1. First note that $f_0(t, x(t), u(t))$ is measurable in $[t_1, t_2]$ and is $\geqq -\psi(t)$; hence, $I[x, u]$ is finite or $+\infty$. If the second member of (5.1) is finite, then $I[x, u]$ must be finite, hence $f_0(t, x(t), u(t))$ must be L-integrable, and the conclusion of Theorem 5.1 reduces to the conclusion of Theorem 4.1 in the case under consideration. If the second member of (5.1) is $+\infty$, then (5.1) in itself is trivial, but we still have to prove that a measurable $u(t)$, $t_1 \leqq t \leqq t_2$, exists with $u(t) \in U(t, x(t))$, $x'(t) = f(t, x(t), u(t))$ a.e. in $[t_1, t_2]$. This, however, is a consequence of closure Theorem 1.1 applied to the absolutely continuous n-vector function x, the n-vector function f, and the sets $Q(t, x) \subset E_n$. Theorem 5.1 is thereby proved.

Finally, let us show by an example that an integral $I[x, u]$ may possess the properties of Theorem 4.1, and thus the property of lower closure as defined at the beginning of § 3, and yet not possess the stronger property of the present § 5.

Indeed, take $m = n = 1$, $U = \{u | -\infty < u < +\infty\}$, $f = \exp(u)$, $f_0 = \exp(u^2)$, $A = E_2$, and take $x(t) = 0$, $0 \leqq t \leqq 1$, $x_k(t) = k^{-1}t$, $0 \leqq t \leqq 1$, $k = 1, 2, \cdots$. Here $x_k \to x$ uniformly in $[0,1]$ as $k \to \infty$ and $I[x_k, u_k] = \exp(\log k)^2 \to +\infty$ as $k \to +\infty$. Obviously, there is no measurable $u(t)$, $0 \leqq t \leqq 1$, with $-\infty < u(t) < +\infty$, such that $0 = x'(t) = \exp(u(t))$ a.e. in $[0, 1]$. The integral I does not have the strong property represented by the conclusion of Theorem 5.1. Yet the integral I has the property of lower closure as defined in § 3 as a consequence of Theorem 4.1. Indeed, here

$$\tilde{\tilde{Q}} = \{(z^0, z) | z^0 \geqq \exp(u^2), z = \exp(u), u \in E_1\},$$

or

$$\tilde{\tilde{Q}} = \{(z^0, z) | z^0 \geqq \exp(\log z)^2, 0 < z < +\infty\},$$

is a fixed closed convex subset of E_2, and all conditions of Theorem 4.1 are satisfied. Instead, $Q = \{z | z = \exp(u), u \in E_1\}$ is the set $Q = \{z | 0 < z < +\infty\}$, a fixed convex set, and Q is not closed.

6. Criteria for property (Q) of the sets $\tilde{\tilde{Q}}(t, x)$. We assume here that the sets A, $U(t, x)$, M, $Q(t, x)$, $\tilde{\tilde{Q}}(t, x)$ are defined as usual, that the sets A and M are closed, and that the functions $f_0(t, x, u)$, $f(t, x, u) = (f_1, \cdots, f_n)$ are continuous on M.

(a) We say that a function $g(t, x, u)$ on M is of *slower growth* than $f_0(t, x, u)$ as $|u| \to \infty$ in a subset A_0 of A if, for every $\varepsilon > 0$, there is some number H, which may depend on ε, f_0 and A_0, such that $(t, x) \in A_0$, $|u| \geqq H$, $u \in U(t, x)$ implies $|g| \leqq \varepsilon f_0$.

THEOREM 6.1. *If* 1 *and* f *are of slower growth than* f_0 *as* $|u| \to \infty$ *in a neighborhood* $N_\delta(\bar{t}, \bar{x})$ *of* $(\bar{t}, \bar{x})$ *in* A, *and* $\tilde{\tilde{Q}}(\bar{t}, \bar{x})$ *is convex, then the sets* $\tilde{\tilde{Q}}(t, x)$ *satisfy property* (Q) *at* $(\bar{t}, \bar{x})$ (*in particular,* $\tilde{\tilde{Q}}(\bar{t}, \bar{x})$ *is closed*).

This statement is proved in Cesari [1b, (2.2.ii)]. Note that if 1 and f are of slower growth than f_0 as $|u| \to \infty$ in A, then not only the sets $\tilde{\tilde{Q}}(t, x)$ of Theorem 4.1 satisfy property (Q) in A, but also condition (ψ) of Theorem 4.1 is trivially satisfied with $\psi = \text{const}$.

(b) In Theorem 6.2 below we shall use a different set of hypotheses. At the beginning of § 5 we noticed that the sets $Q(t, x) \subset E_n$ are the projections of the sets $\tilde{\tilde{Q}}(t, x) \subset E_{n+1}$ on the z-space E_n; hence, the convexity of any set $\tilde{\tilde{Q}}(t, x)$ in E_{n+1} implies the convexity of the corresponding set $Q(t, x)$ in E_n. Nevertheless,

as we proved by an example at the beginning of § 5 the sets $\tilde{\tilde{Q}}(t, x)$ may be closed and even satisfy property (Q) at any given point $(\bar{t}, \bar{x})$ without this being the case for the sets $Q(t, x)$.

However, the following holds: If the sets $\tilde{\tilde{Q}}(t, x)$ satisfy property (Q) at $(\bar{t}, \bar{x})$, then

$$(*) \qquad (z^0, z) \in \bigcap_\delta \operatorname{cl} \operatorname{co} \bigcup \tilde{\tilde{Q}}(t, x)$$

implies $z \in Q(\bar{t}, \bar{x})$. Indeed, $(*)$ yields $(z^0, z) \in \tilde{\tilde{Q}}(\bar{t}, \bar{x})$ by property (Q) at $(\bar{t}, \bar{x})$, and then $z \in Q(\bar{t}, \bar{x})$.

We shall say that condition (α) holds at the point $(\bar{t}, \bar{x}) \in A$ provided:

$$(\alpha) \qquad (z^0, z) \in \bigcap_\delta \operatorname{cl} \operatorname{co} \bigcup \tilde{\tilde{Q}}(t, x) \quad \text{implies } z \in Q(\bar{t}, \bar{x}).$$

As mentioned, this condition is necessary for property (Q) of the sets $\tilde{\tilde{Q}}(t, x)$ at $(\bar{t}, \bar{x})$. This same condition (α) alone is not sufficient for property (Q) as the following example shows: Take $m = n = 1$, $U = E_1$, $f_0 = \exp(tu)$, $f = 0$, $0 \leqq t \leqq 1$. Then

$$\tilde{\tilde{Q}}(0) = \{(z^0, z) | z^0 \geqq 1, z = 0\}, \qquad \tilde{\tilde{Q}}(t) = \{(z^0, z) | z^0 > 0, z = 0\}$$

if $t > 0$, the sets $\tilde{\tilde{Q}}$ do not satisfy condition (Q) at $t = 0$, but condition (α) certainly holds at the same point. Note that condition (α) is trivially satisfied for free problems ($m = n$, $f = u$, $U = E_n$) since $Q = U = E_n$, and all points $z \in E_n$ are in Q.

Now we shall say that condition (X) holds at the point $(\bar{t}, \bar{x}) \in A$ provided the following holds.

(X) For every $\bar{z} \in Q(\bar{t}, \bar{x})$ there is at least one point $\bar{u} \in U(\bar{t}, \bar{x})$ with $\bar{z} = f(\bar{t}, \bar{x}, \bar{u})$ and the following property: Given $\varepsilon > 0$ there are numbers $\delta > 0$, and $r, b = (b_1, \cdots, b_n)$ real such that

$$(X') \qquad f_0(t, x, u) \geqq r + \sum_j b_j f_j(t, x, u)$$

for all $(t, x) \in N_\delta(\bar{t}, \bar{x})$ and $u \in U(t, x)$,

$$(X'') \qquad f_0(\bar{t}, \bar{x}, \bar{u}) \leqq r + \sum_j b_j f_j(\bar{t}, \bar{x}, \bar{u}) + \varepsilon.$$

As we have shown in [1c] this is a very weak requirement. For free problems, for instance, this condition reduces to a weak form of the well-known "seminormality convexity condition" (see below).

THEOREM 6.2. *If conditions* (α) *and* (X) *hold at a point* $(\bar{t}, \bar{x}) \in A$, *then the sets* $\tilde{\tilde{Q}}(t, x)$ *are convex, closed, and satisfy property* (Q) *at the point* $(\bar{t}, \bar{x})$.

This statement was proved in [1c].

(c) *A partial converse of Theorem* 6.2. We have already seen that the convexity of $\tilde{\tilde{Q}}(t, x)$ in E_{n+1} implies the convexity of $Q(t, x)$ in E_n. We shall denote by R the linear manifold in E_n of minimum dimension r containing $Q(t, x)$. Thus $Q(t, x) \subset R \subset E_n$, $0 \leqq r \leqq n$. As usual we denote by $\operatorname{int} Q(t, x)$ the set of all $z \in Q(t, x)$ which are in the interior of the convex set $Q(t, x)$ with respect to E_n. Analogously, we denote by $R\text{-int } Q(t, x)$ the set of all $z \in Q(t, x)$ which are in the interior of $Q(t, x)$ with respect to R. Then

$$\operatorname{int} Q(t, x) \subset R\text{-int } Q(t, x) \subset Q(t, x) \subset R \subset E_n.$$

If $Q(t, x)$ is reduced to a single point, then $\operatorname{int} Q(t, x) = \varnothing$, $R\text{-int } Q(t, x) = \varnothing$,

$r = 0$, $R = Q(t, x)$. If $Q(t, x)$ contains at least two points, then $1 \leqq r \leqq n$, and R-int $Q(t, x) \neq \varnothing$.

For any point $(\bar{t}, \bar{x}) \in A$ and $z \in Q(\bar{t}, \bar{x}) \subset E_n$ let us denote by $T(z; \bar{t}, \bar{x})$ the number, $-\infty \leqq T < +\infty$, defined by

$$\begin{aligned} T(z; \bar{t}, \bar{x}) &= \inf \{z^0 | (z^0, z) \in \tilde{\tilde{Q}}(\bar{t}, \bar{x}) \text{ for some } z^0 \in E_1\} \\ &= \inf \{z^0 | z^0 = f_0(\bar{t}, \bar{x}, u) \text{ for all } u \in U(\bar{t}, \bar{x}) \text{ with } f(\bar{t}, \bar{x}, u) = z\}. \end{aligned} \tag{6.1}$$

We shall consider $T(z; \bar{t}, \bar{x})$ as a function of z in $Q(t, x) \subset E_n$. We proved in [1c] the following properties of the function $T(z; \bar{t}, \bar{x})$.

(π_1) If $\tilde{\tilde{Q}}(t, x)$ is convex, then either $T(z; \bar{t}, \bar{x}) = -\infty$ for all $z \in R$-int $Q(\bar{t}, \bar{x})$, or $T(z; \bar{t}, \bar{x})$ is a convex real-valued function on the convex set $Q(\bar{t}, \bar{x}) \subset E_n$, is finite everywhere in $Q(\bar{t}, \bar{x})$, and continuous at every point $z \in R$-int $Q(t, x)$.

(π_2) If $\tilde{\tilde{Q}}(\bar{t}, \bar{x})$ is convex and closed, and $T(z; \bar{t}, \bar{x})$ is finite on $Q(\bar{t}, \bar{x})$, then min holds instead of inf in both relations (6.1), and $T(z; \bar{t}, \bar{x})$ is (continuous at each $z \in R$-int $Q(\bar{t}, \bar{x})$ and) lower semicontinuous at each $z \in Q(\bar{t}, \bar{x}) - R$-int $Q(\bar{t}, \bar{x})$.

In [1c] we proved by examples that the first alternative in (π_1) may well occur, and that, under the conditions of (π_2), $T(z: \bar{t}, \bar{x})$ may well be discontinuous at the points $z \in Q(\bar{t}, \bar{x}) - R$-int $Q(\bar{t}, \bar{x})$.

(π_3) Under the conditions of (π_2), $T(z; \bar{t}, \bar{x})$ has a supporting plane at every point $\bar{z} \in R$-int $Q(\bar{t}, \bar{x})$; that is, there are real numbers r, $b = (b_1, \cdots, b_n)$ such that $T(\bar{z}; \bar{t}, \bar{x}) \geqq r + b \cdot z$ for all $z \in Q(\bar{t}, \bar{x})$ and $T(z; \bar{t}, \bar{x}) = r + b \cdot \bar{z}$.

This last property can be reinterpreted by saying that for every $\bar{z} \in R$-int $Q(\bar{t}, \bar{x})$ there is at least one point $\bar{u} \in U(\bar{t}, \bar{x})$ and real numbers $r, b = (b_1, \cdots, b_n)$ such that

$$\begin{aligned} &T(z; \bar{t}, \bar{x}) \geqq r + b \cdot z \quad \text{with } z = f(\bar{t}, \bar{x}, u) \quad \text{for all } u \in U(\bar{t}, \bar{x}), \\ &T(\bar{z}; \bar{t}, \bar{x}) = r + b \cdot \bar{z} \quad \text{with } \bar{z} = f(\bar{t}, \bar{x}, \bar{u}). \end{aligned} \tag{6.2}$$

Finally, the same property (π_3) can be reinterpreted in turn by saying that for every $\bar{z} \in R$-int $Q(\bar{t}, \bar{x})$ there is at least one point $\bar{u} \in U(\bar{t}, \bar{x})$ and real numbers r, $b = (b_1, \cdots, b_n)$ such that

$$\begin{aligned} &f_0(\bar{t}, \bar{x}, u) \geqq r + b \cdot f(\bar{t}, \bar{x}, u) \quad \text{for all } u \in U(\bar{t}, \bar{x}), \\ &f_0(\bar{t}, \bar{x}, \bar{u}) = r + b \cdot f(\bar{t}, \bar{x}, \bar{u}). \end{aligned} \tag{6.3}$$

In [1c] we proved by examples that under the conditions of (π_2) the supporting plane for the convex set $\tilde{\tilde{Q}}(t, x)$ at points $(z^0, \bar{z})$ with $\bar{z} \in Q(\bar{t}, \bar{x}) - R$-int $Q(\bar{t}, \bar{x})$, $z^0 = T(z; \bar{t}, \bar{x})$, may well be vertical. Thus there may be no supporting plane (of the form $z^0 = r + b \cdot z$) for $T(z; \bar{t}, \bar{x})$ at the points $\bar{z} \in Q(\bar{t}, \bar{x}) - R$-int $Q(\bar{t}, \bar{x})$. At each of these points $\bar{z}$ we can only say that for every $\varepsilon > 0$ there are real numbers r, $b = (b_1, \cdots, b_n)$ depending on ε, $\bar{t}$, $\bar{x}$, $\bar{z}$ such that $T(z; \bar{t}, \bar{x}) \geqq r + b \cdot z$ for all $z \in Q(\bar{t}, \bar{x})$, and $T(\bar{z}; \bar{t}, \bar{x}) < r + b \cdot \bar{z} + \varepsilon$. Properties analogous to (6.2) and (6.3) then also hold.

Finally, in [1c] we proved the following partial converse of Theorem 6.2.

THEOREM 6.3. *If $\tilde{\tilde{Q}}(t, x)$ is closed and convex for all $(t, x) \in A$, if $\tilde{\tilde{Q}}(t, x)$ satisfies property* (Q) *at $(\bar{t}, \bar{x}) \in A$, and $T(z; \bar{t}, \bar{x})$ is finite on $Q(\bar{t}, \bar{x})$, then for every $\bar{z} \in Q(\bar{t}, \bar{x})$ and $\varepsilon > 0$ there are real numbers r, $b = (b_1, \cdots, b_n)$ and $\delta > 0$ such that*

$$T(z; t, x) \geqq r + b \cdot z - \varepsilon \quad \textit{for all} \quad z \in Q(t, x) \quad \textit{and} \quad (t, x) \in N_\delta(\bar{t}, \bar{x}),$$

$$T(\bar{z}; \bar{t}, \bar{x}) = r + b \cdot \bar{z}.$$

Hence, there is at least one point $\bar{u} \in U(\bar{t}, \bar{x})$ such that

$$T(z; t, x) \geqq r + b \cdot z \quad \text{for all } (t, x) \in N_\delta(\bar{t}, \bar{x}) \quad \text{and} \quad u \in U(t, x) \text{ such that } z = f(t, x, u),$$

$$T(\bar{z}; \bar{t}, \bar{x}) \leqq r + b \cdot \bar{z} + \varepsilon.$$

Also, we have

$$f_0(t, x, u) \geqq r + b \cdot f(t, x, u) \quad \text{for all } u \in U(t, x) \quad \text{and} \quad (t, x) \in N_\delta(\bar{t}, \bar{x}),$$

$$f_0(\bar{t}, \bar{x}, \bar{u}) \leqq r + b \cdot f(\bar{t}, \bar{x}, \bar{u}) + \varepsilon.$$

Thus, under the mild restriction of assuming $T(z; t, x)$ finite, we see that condition (Q) for the sets $\tilde{\tilde{Q}}(t, x)$ and the set of conditions (α) and (X) are equivalent. Under the mild restriction above, we have obtained a characterization of property (Q) for the sets $\tilde{\tilde{Q}}(t, x)$.

(d) *The case of f linear in u.* We shall assume here that A is a given closed subset of the tx-space E_{1+n}, that $U = E_m$, that $f_0(t, x, u)$ and $f(t, x, u) = (f_1, \cdots, f_n)$ are continuous on $M = A \times E_m$, and that f is linear in u; that is,

$$f_i(t, x, u) = \sum_{j=1}^{m} b_{ij}(t, x)u^j + c_i(t, x), \qquad i = 1, \cdots, n,$$

or

$$f(t, x, u) = B(t, x)u + C(t, x),$$

where B, C are $n \times m$ and $n \times 1$ matrices with entries continuous in A. For every compact subset A_0 of A, the functions b_{ij}, c_i are continuous and bounded on A_0; hence, there are constants G_0, F_0 such that $|f(t, x, u)| \leqq G_0|u| + F_0$ for all $(t, x) \in A_0$ and $u \in E_n$.

THEOREM 6.4. *If $f_0(t, x, u)$ is convex in u, and f is linear in u with $U = E_m$, then the sets $\tilde{\tilde{Q}}(t, x)$ are convex.*

THEOREM 6.5. *If A is closed, $U = E_m$, $M = A \times E_m$, if $f_0(t, x, u)$ is continuous on M, convex in u, and "seminormal" in u at a point $\bar{x} \in A$ (see definition* (SN) *in* (e) *below), if $f(t, x, u) = B(t, x)u + C(t, x)$, where the matrices B, C have entries continuous in A, then the sets*

$$\tilde{\tilde{Q}}(t, x) = \{(z^0, z) | z^0 \geqq f_0(t, x, u), z = f(t, x, u), u \in E_m\}$$

satisfy property (Q) *at* $(\bar{t}, \bar{x})$.

A proof was given in [1c]. This statement for f linear in u, or $f = B(t, x)u + C(t, x)$, is much stronger than the analogous statement, Theorem 6.1. Indeed, we would deduce from Theorem 6.1 an analogous statement as Theorem 6.5 under a growth condition $f_0(t, x, u) \geqq \Phi(|u|)$ with $\Phi(\zeta)/\zeta \to +\infty$ as $\zeta \to +\infty$.

(e) *The free problem $m = n, f = u, U = E_n$.* Here the sets Q reduce to the fixed, closed, and convex set $Q = U = E_n$. The sets $\tilde{\tilde{Q}}(t, x)$ reduce here to

$$\tilde{\tilde{Q}}(t, x) = \{(z^0, u) | z^0 \geqq f_0(t, x, u), u \in E_n\}.$$

These sets are closed whenever f_0 is continuous, and convex whenever $f_0(t, x, u)$ is convex in u. As mentioned, condition (α) is trivially satisfied. Condition (X) at a

point $(\bar{t}, \bar{x}) \in A$ reduces to the following simple (and well-known) requirement:

(X_f) (*weak seminormality condition*). We say that the real-valued function $f_0(t, x, u), (t, x) \in A, u \in E_n$, is weakly seminormal at the point $(\bar{t}, \bar{x}) \in A$ provided, for every $\bar{u} \in E_n$ and $\varepsilon > 0$, there are numbers $\delta > 0$, and $r, b = (b_1, \cdots, b_n)$ real such that

(X'_f) $\quad f_0(t, x, u) \geqq r + b \cdot u$ for all $(t, x) \in N_\delta(\bar{t}, \bar{x})$ and $u \in E_n$,

(X''_f) $\quad f_0(\bar{t}, \bar{x}, \bar{u}) \leqq r + b \cdot \bar{u} + \varepsilon$.

For this concept and a number of variants, see L. Tonelli [7a, b] and E. J. McShane [3a, b]. Theorem 6.2 now yields the following theorem.

THEOREM 6.6. *For free problems* $(m = n, f = u, U = E_n)$, *if* A *is closed, if* $f_0(t, x, u)$ *is continuous on* $M = A \times E_n$ *and convex in* u, *and if* f_0 *is weakly seminormal at a point* $(\bar{t}, \bar{x}) \in A$, *then the sets* $\tilde{\tilde{Q}}(t, x)$ *satisfy property* (Q) *at* $(\bar{t}, \bar{x})$.

Convexity of f_0 alone does not imply that the sets $\tilde{\tilde{Q}}(t, x)$ have property (Q) in A. This is shown by the following simple example. Take $n = 1$, $f_0(t, u) = tu$, $0 \leqq t \leqq 1, u \in U = E_1$. Then f_0 is continuous and convex in u for every t, but at every $t, 0 \leqq t \leqq 1$, we have

$$\tilde{\tilde{Q}}(t) = \{(z^0, u) | z^0 \geqq tu, u \in E_1\},$$

a half-plane in E_2, while $\bigcap_\delta \operatorname{cl} \operatorname{co} \tilde{\tilde{Q}}(t; \delta)$ is the entire plane E_2. Thus the sets $\tilde{\tilde{Q}}$ do not satisfy property (Q) at any $t, 0 \leqq t \leqq 1$. Clearly, the function f_0 is not weakly seminormal.

Let us consider now the seminormality condition, which is a somewhat stronger requirement than the weak seminormality condition. The seminormality condition too was used by Tonelli [7] and McShane [3].

(SN) (*Seminormality condition*). We say that the real-valued function $f_0(t, x, u)$, $(t, x) \leqq A, u \in E_n$, is seminormal at the point $(\bar{t}, \bar{x}) \in A$ provided, for every $\bar{u} \in E_n$, there are numbers $\delta > 0, \nu > 0$, and $r, b = (b_1, \cdots, b_n)$ real such that

(SN') $\quad f_0(t, x, u) \geqq r + b \cdot u + \nu|u - \bar{u}|$ for all $(t, x) \in N_\delta(\bar{t}, \bar{x})$ and $u \in E_n$,

(SN'') $\quad f_0(\bar{t}, \bar{x}, \bar{u}) \leqq r + b \cdot \bar{u} + \varepsilon$.

The seminormality condition has a very simple elegant characterization.

THEOREM 6.7. *For free problems* $(m = n, f = u, U = E_n)$, *if the real-valued function* $f_0(t, x, u)$, $(t, x) \in A, u \in E_n$, *is continuous on* $A \subset E_n$, *and convex in* u *at some* $(t, x) \in A$, *then* f_0 *is seminormal at* (t, x) *if and only if for no* $u, u_1 \in E_n, u_1 \neq 0$, *it occurs that*

$$f_0(t, x, u) = 2^{-1}[f_0(t, x, u + \lambda u_1) + f_0(t, x, u - \lambda u_1)] \quad \textit{for all} \quad \lambda \geqq 0.$$

A proof of this statement under smoothness hypotheses was given by L. Tonelli [7a, b]. A proof under the sole hypotheses of continuity and convexity stated in Theorem 6.7 can be found in L. Turner [8] and is reported in [1c]. Note that, if we denote by $\tilde{Q}(t, x)$ the set $\{(z^0, u) | z^0 = f_0(t, x, u), u \in E_n\}$, then $\tilde{Q}(t, x)$ is often denoted as the "figurative" of f_0 (at the point $(t, x) \in A$). Theorem 6.7 then states that f_0 is seminormal at $(\bar{t}, \bar{x})$ if and only if the figurative contains no straight line. In particular, if say $f_0(\bar{t}, \bar{x}, u) \to +\infty$ as $|u| \to +\infty$, and $f_0(\bar{t}, \bar{x}, u)$ is convex in u, then the figurative $\tilde{Q}(\bar{t}, \bar{x})$ cannot contain any straight line, f_0 is seminormal at $(\bar{t}, \bar{x})$, f_0 is weakly seminormal, and certainly the sets $\tilde{\tilde{Q}}(t, x)$ satisfy property (Q) at $(\bar{t}, \bar{x})$.

7. Lower semicontinuity. The rather general concept of lower closure defined in § 3 is a natural extension of the usual concept of lower semicontinuity. Indeed, the definition of lower closure in § 3 reduces to the usual concept of lower semicontinuity whenever the strategy u is "determined" by the (admissible) trajectory x, and then the functional (3.1) can be thought of as depending on the (admissible) trajectory x only:

$$I[x] = \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt. \tag{7.1}$$

This occurs, for instance, for free problems where $u(t) = x'(t)$ (a.e.). The purpose of the present § 7 and the next one, § 8, is to clarify the concepts, to deduce theorems of lower semicontinuity from our previous theorems of lower closure in § 4 and § 5, and to show that the usual theorems of lower semicontinuity for free problems are corollaries of our theorems of lower closure. We already presented a statement of this sort in Corollary 1 of § 4.

(a) It may happen that the data A, $U(t, x)$, B, $f_0(t, x, u)$, $f(t, x, u) = (f_1, \cdots, f_n)$ are so arranged that, for any admissible pair $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, the trajectory x uniquely determines the strategy u (a.e. in $[t_1, t_2]$). Then the functional (7.1) can be thought of as being defined for every admissible trajectory x, and we may denote it as $I[x]$. In this section we shall refer to these systems, for the sake of brevity, as TDS systems, or systems in which any admissible trajectory uniquely determines the corresponding strategy.

For all these systems the concept of lower closure (§ 3) reduces to the one of lower semicontinuity. Let $x(t)$, $t_1 \leqq t \leqq t_2$, be any absolutely continuous vector function which is the limit in the ρ-metric of a sequence of admissible trajectories $x_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, with $\rho(x_k, x) \to 0$ and $\liminf I[x_k] < +\infty$ as $k \to \infty$ (thus, of course $t_{1k} \to t_1$, $t_{2k} \to t_2$). The functional (7.1) is said to be *lower* semicontinuous at x provided, from any such sequence we can conclude that x is admissible, and that $I[x] \leqq \liminf I[x_k]$ as $k \to \infty$.

(b) For general TDS systems Theorems 4.1 and 5.1 reduce to the following ones.

THEOREM 7.1. *For TDS systems, and under the same conditions of Theorem* 4.1, *let* $x(t)$, $t_1 \leqq t \leqq t_2$, *be any absolutely continuous function with* $(t, x(t)) \in A$ *for all* $t \in [t_1, t_2]$, *and let* x *be the uniform limit of admissible trajectories* $x_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, *with* $\rho(x_k, x) \to 0$, $\liminf I[x_k] < +\infty$ *as* $k \to \infty$. *Then,* x *is admissible and* $I[x] \leqq \liminf I[x_k]$.

THEOREM 7.2. *For TDS systems, and under the conditions of Theorem* 5.1, *let* $x(t)$, $t_1 \leqq t \leqq t_2$, *be any absolutely continuous function with* $(t, x(t)) \in A$ *for all* $t \in [t_1, t_2]$ *and let* x *be the uniform limit of admissible trajectories* $x_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, *that is,* $\rho(x_k, x) \to 0$ *as* $k \to \infty$. *Then,* x *is admissible and* $I[x] \leqq \liminf I[x_k]$.

In the last statement we understand that there is a measurable function $u(t)$, $t_1 \leqq t \leqq t_2$, such that $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, satisfy all conditions for admissibility but perhaps the L-integrability of $f_0(t, x(t), u(t))$ in $[t_1, t_2]$, and that this condition also is satisfied whenever $\liminf I[x_k] \leqq +\infty$.

Theorem 4.2 also has its counterpart here, but we leave its formulation to the reader.

8. Theorems of lower semicontinuity for free problems. Let us consider here free problems, that is, systems with $m = n$, $f = u$, $U = E_n$; hence, the strategy $u(t) = x'(t)$ (a.e.) is determined by the trajectory (a.e.). If A and B are closed sets as usual, then $M = A \times E_n$ is also closed, and $f_0(t, x, u)$ is a given continuous scalar function on M. Here a function $x(t) = (x^1, \cdots, x^n)$, $t_1 \leqq t \leqq t_2$, is an *admissible trajectory* provided x is absolutely continuous in $[t_1, t_2]$, $(t, x(t)) \in A$ for all $t \in [t_1, t_2]$, $(t_1, x(t_1), t_2, x(t_2)) \in B$, and $f_0(t, x(t), x'(t))$ is L-integrable in $[t_1, t_2]$. Then, the cost functional is

$$I[x] = \int_{t_1}^{t_2} f_0(t, x(t), x'(t))\, dt. \tag{8.1}$$

The corresponding sets $Q(t, x)$ and $\tilde{\tilde{Q}}(t, x)$ have already been discussed in § 6, part (d), and the concept of weak seminormality has been introduced there.

Our general statement, Theorem 4.2, in conjunction with Theorem 6.5 yields the following theorem.

THEOREM 8.1 (A theorem of lower semicontinuity for free problems). *For free problems* ($m = n$, $f = u$, $U = E_n$), *if* A *is closed, if* $f_0(t, x, u)$ *is continuous on* $M = A \times E_n$, *convex in* u, *and weakly seminormal with respect to* u *in* A, *then the functional* (8.1) *has the property of lower semicontinuity; that is, if* $x(t) = (x^1, \cdots, x^n)$, $t_1 \leqq t \leqq t_2$, *is an absolutely continuous function which is the limit in the* ρ*-metric of admissible trajectories* $x_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, *with* $\rho(x_k, x) \to 0$ *and* $\liminf I[x_k] < +\infty$ *as* $k \to +\infty$, *then* x *is admissible and* $I[x] \leqq \liminf I[x_k]$.

The condition of weak seminormality is certainly satisfied if $f_0(t, x, u)$ is continuous in (t, x, u), convex in u for every (t, x), and $f_0(t, x, u) \to +\infty$ as $|u| \to +\infty$ for every $(t, x) \in A$.

Theorem 8.1 is due to L. Tonelli [7a] who proved it for f_0 of class C' in u. A proof under the present sole continuity hypotheses was given by L. Turner [8]. The lower semicontinuity Theorem 8.1 is here a corollary of Theorem 4.2 for lower closure of general Lagrange problems.

Theorem 8.1 without the hypothesis of weak seminormality is not true, as the following simple example shows. Take $n = 2$, $A = E_3$, $f_0 = yx' - xy'$, x, y state variables. Then f_0 is certainly convex in (x', y'), namely linear. Nevertheless, $I = \int_{t_1}^{t_2} (yx' - xy')\, dt$ is not lower semicontinuous. Indeed, if we take C: $x = 0$, $y = 0$, $0 \leqq t \leqq 2\pi$, and C_k: $x = k^{-1} \cos k^2 t$, $y = k^{-1} \sin k^2 t$, $0 \leqq t \leqq 2\pi$, $k = 1, 2, \cdots$, then $C_k \to C$, $I[C_k] = -2\pi$, $k = 1, 2, \cdots$, and $I[C] = 0$. An analogous example for $n = 1$ has been given by Tonelli [7b, vol. 2, pp. 390–392]. Nevertheless, Tonelli proved that, for $n = 1$ and f_0 continuous in (t, x, x') with continuous first order partial derivatives $f_{0x'}$ and $f_{0x'x'}$, Theorem 8.1 holds without the weak seminormality requirement [7a, pp. 205–206]. Again, the example above shows that this is not the case for $n \geqq 2$. (See, for analogous examples, McShane [3b].)

9. Convexity as a necessary and sufficient condition for lower closure. We are now in a position to prove the statement we mentioned in § 4, Remark 2, that the convexity of the sets $\tilde{\tilde{Q}}$, that is, the convexity part of property (Q), is essentially a necessary and sufficient condition for lower closure.

The sufficiency part is covered by Theorems 4.1, 4.2 and 5.1. For the necessity part we shall prove that the convexity of the sets $\tilde{\tilde{Q}}$ is essentially necessary for lower closure.

We shall need a few more definitions. Here again, as in §§ 1–7, A is a closed subset of the tx-space E_{1+n}, $U(t,x) \subset E_m$, M defined as usual is a closed set of the txu-space E_{1+n+m}, $f_0(t,x,u)$, $f(t,x,u) = (f_1, \cdots, f_n)$ are continuous on M, and I denotes the functional for Lagrange problems

$$I[x,u] = \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt. \tag{9.1}$$

In this section we shall denote as admissible any pair $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, satisfying all requirements listed in § 3 (disregarding boundary conditions, or equivalently, taking $B = E_{2n+2}$). The functional $I[x,u]$ is defined for every admissible pair x, u.

Also, we shall assume below that $f(t,x,u)$ is locally Lipschitzian with respect to x in M.

Given a point $(\bar{t}, \bar{x}) \in A$, we shall say that a convex combination of points in E_n,

$$\begin{aligned} &\bar{z} = \sum_{j=1}^{n+1} \lambda_j z_j, \qquad \sum_{j=1}^{n+1} \lambda_j = 1, \\ &\bar{z} = f(\bar{t}, \bar{x}, \bar{u}), \qquad z_j = f(\bar{t}, \bar{x}, u_j), \qquad j = 1, \cdots, n+1, \end{aligned} \tag{9.2}$$

is *generic* at $(\bar{t}, \bar{x})$ provided:

(a) there is some $\delta > 0$ such that $\bar{u}$, $u_j \in U(t,x)$, $j = 1, \cdots, n+1$, for all $(t,x) \in N_\delta(\bar{t}, \bar{x})$;

(b) $\Delta = \det(F_{ij}, i,j = 1, \cdots, n+1) \neq 0$, where $F_{ij} = f_i(\bar{t}, \bar{x}, u_j)$ for $i = 1, \cdots, n$, $j = 1, \cdots, n+1$; $F_{ij} = 1$ for $i = n+1, j = 1, \cdots, n+1$;

(c) $0 < \lambda_j < 1, j = 1, \cdots, n+1$.

For instance, for $m = n = 1$, $f = u$, $U = E_1$, $\lambda_1 = \lambda_2 = 1/2$, $\bar{u} = 0$, $u_1 = 1$, $u_2 = -1$, we have $\Delta = \det(1, -1; 1, 1) = 2 \neq 0$, and (a), (b), (c) are satisfied. For free problems ($m = n$, $f_i = u^i$, $i = 1, \cdots, n$, $U = E_n$, $z = u$), (a) is always trivially satisfied, and near any convex combination there are as many generic convex combinations as we want.

The statement below, Theorem 9.1, depicts a situation where there is no "lower semicontinuity" at suitable trajectories issued from a point $(\bar{t}, \bar{x}) \in A$ when the convexity requirement is not satisfied.

THEOREM 9.1. *Let A, $U(t,x)$, M be defined as usual and closed, and let $f_0(t,x,u)$, $f(t,x,u) = (f_1, \cdots, f_n)$ be continuous on M and locally Lipschitzian with respect to x. If $(\bar{t}, \bar{x})$ is any point interior to A, and if $\bar{z} = \sum_{j=1}^{n+1} \lambda_j z_j$ is a convex combination of points of E_n, with $\bar{z} = f(\bar{t}, \bar{x}, \bar{u})$, $z_j = f(\bar{t}, \bar{x}, u_j)$, $j = 1, \cdots, n+1$, which is generic at $(\bar{t}, \bar{x})$, and*

$$f_0(\bar{t}, \bar{x}, \bar{u}) > \sum_{j=1}^{n+1} \lambda_j f_0(\bar{t}, \bar{x}, u_j), \tag{9.3}$$

then there are admissible pairs $x(t), u(t)$, $\bar{t} \leqq t \leqq \bar{t} + \delta$, and $x_k(t), u_k(t)$, $\bar{t} \leqq t \leqq \bar{t} + \delta$,

$k = 1, 2, \cdots$, *such that* $x_k \to x$ *as* $k \to \infty$ *uniformly in* $[\bar{t}, \bar{t} + \delta]$, *and* $I[x, u] > \liminf I[x_k, u_k]$.

Proof. By force of (a) there is a $\delta_0 > 0$ such that $\bar{u}, u_j \in U(t, x)$ for all $(t, x) \in N_\delta(\bar{t}, \bar{x})$, $j = 1, \cdots, \nu$. Take any number δ' with $0 < \delta' < \delta_0(n^2 + 1)^{-1/2}$, and note that all points (t, x) with $\bar{t} \leqq t \leqq \bar{t} + \delta'$, $|x - \bar{x}| \leqq \delta'$ are at a distance from $(\bar{t}, \bar{x})$ which is $\leqq (\delta'^2 + n^2\delta'^2)^{1/2} = \delta'(n^2 + 1)^{1/2} < \delta_0$. We can take $\delta' > 0$ sufficiently small so that all mentioned points are in A.

Let us consider the differential equation

$$dx/dt = f(t, x, \bar{u}), \tag{9.4}$$

with initial condition $x(\bar{t}) = \bar{x}$. Having assumed f continuous on M and locally Lipschitzian with respect to x in M, we know that $f(t, x, \bar{u})$ is continuous and uniformly Lipschitzian with respect to x in the set $\{(t, x) | \bar{t} \leqq t \leqq t + \delta', |x - \bar{x}| \leqq \delta\}$. By differential equation theory we know that for some δ sufficiently small, $0 < \delta \leqq \delta'$, there is one and only one absolutely continuous solution $x(t)$, $\bar{t} \leqq t \leqq \bar{t} + \delta$, to (9.4) with $x(\bar{t}) = \bar{x}$ and $|x(t) - \bar{x}| \leqq \delta'$ for all $\bar{t} \leqq t \leqq \bar{t} + \delta$.

Now let us consider the system of $n + 1$ linear equations in the $n + 1$ unknowns $p_1(t), \cdots, p_{n+1}(t)$:

$$\begin{aligned} &\sum_{j=1}^{n+1} p_j(t) f(t, x(t), u_j) - f(t, x(t), \bar{u}) = 0, \\ &\sum_{j=1}^{n+1} p_j(t) - 1 = 0, \qquad \bar{t} \leqq t \leqq \bar{t} + \delta. \end{aligned} \tag{9.5}$$

Note that for $t = \bar{t}$ these equations have the trivial solution $p_j(\bar{t}) = \lambda_j$, $j = 1, \cdots, n + 1$. The functional determinant of these equations is

$$\Delta(t) = \det (F_{ij}(t), i, j = 1, \cdots, n + 1),$$

with $F_{ij}(t) = f_i(t, x(t), u_j)$, $i = 1, \cdots, n, j = 1, \cdots, n + 1$, and $F_{ij} = 1$, $i = n + 1$, $j = 1, \cdots, n + 1$. Hence, $\Delta(\bar{t}) \neq 0$ by force of (b).

Thus, we can take $\delta > 0$ sufficiently small so that equations (9.5) have a continuous solution $p(t) = (p_1, \cdots, p_{n+1})$ in $[\bar{t}, \bar{t} + \delta]$. Since $p_j(\bar{t}) = \lambda_j$, and $0 < \lambda_j < 1$, we can take $\delta > 0$ so small that we have also $0 \leqq p_j(t) \leqq 1, \bar{t} \leqq t \leqq \bar{t} + \delta$, $j = 1, \cdots, n + 1$.

Note that $x(t)$, $\bar{t} \leqq t \leqq \bar{t} + \delta$, is an absolutely continuous solution of the differential system

$$\frac{dx}{dt} = \sum_{j=1}^{n} p_j(t) f(t, x(t), u_j), \qquad \bar{t} \leqq t \leqq \bar{t} + \delta, \tag{9.6}$$

with initial value $x(\bar{t}) = \bar{x}$, since the second member coincides with $f(t, x(t), \bar{u})$ in $[\bar{t}, \bar{t} + \delta]$.

Note that, by hypothesis,

$$f_0(\bar{t}, \bar{x}, \bar{u}) > \sum_{j=1}^{n+1} \lambda_j f_0(\bar{t}, \bar{x}, u_j). \tag{9.7}$$

If we denote by $2\sigma > 0$ the difference between the first and second member in (9.7), we see that by simple continuity argument, we can take $\delta > 0$ sufficiently

small so that

$$f_0(t, x(t), \bar{u}) > \sum_{j=1}^{n} p_j(t) f_0(t, x(t), u_j) + \sigma$$

for all $\bar{t} \leqq t \leqq \bar{t} + \delta$.

Note that we have here a generalized system $x(t)$, $p(t)$, $v(t)$, $\bar{t} \leqq t \leqq \bar{t} + \delta$, with $v(t) = \{u^{(j)}(t) = u_j, j = 1, \cdots, n+1\}$, $\bar{t} \leqq t \leqq \bar{t} + \delta$, and

$$I[x, p, v] = \int_{\bar{t}}^{\bar{t}+\delta} \sum_{j=1}^{n+1} p_j(t) f_0(t, x(t), u_j)\, dt$$

$$< \int_{\bar{t}}^{\bar{t}+\delta} f_0(t, x(t), \bar{u})\, dt - \sigma\delta = I[x, u] - \sigma\delta,$$

where u denotes here the constant strategy $u(t) = \bar{u}$, $\bar{t} \leqq t \leqq \bar{t} + \delta$.

Now we shall apply the general theorem of approximation of generalized solutions by means of usual solutions proved in [1a, § 14, (i), p. 416]. To do this we need to reduce drastically the control space since in that theorem it is assumed that the control space has to depend on t only. In the present situation, however, we can take as auxiliary control space U^* the fixed space made up of the $n + 2$ points $\bar{u}$, $u_j, j = 1, \cdots, n+1$, of E_m; these are the only points we need in the control space. Then, by force of [1a, § 14, (i)] we know that there exists a sequence of admissible pairs $x_k(t), u_k(t)$, $\bar{t} \leqq t \leqq \bar{t} + \delta$, $k = 1, 2, \cdots$, with $x_k \to x$ uniformly in $[\bar{t}, \bar{t} + \delta]$ as $k \to \infty$, and $I[x_k, u_k] \to I[x, p, v]$. Thus, for k sufficiently large we certainly have $|x_k(t) - x(t)| < \delta'$ and thus $(t, x_k(t)) \in A$ for all $\bar{t} \leqq t \leqq \bar{t} + \delta$, and $I[x_k, u_k] < I[x, p, v] + \sigma\delta/2$; hence,

$$I[x_k, u_k] < I[x, u] - \sigma\delta/2$$

for all k sufficiently large. Theorem 9.1 is thereby proved.

The statement below, Theorem 9.2, depicts a situation where there is no "lower semicontinuity" at a given trajectory x in A, that is, at a given admissible pair $x(t), u(t)$, $t_1 \leqq t \leqq t_2$, where $u(t)$ may be bounded or unbounded. The statement is similar to Theorem 9.1. Since we shall use the same theorem [1a, § 14, (i)] concerning the approximation of generalized solutions by means of usual solutions, we have to make sure that the conditions of that theorem are satisfied. For the case in which the strategy $u(t)$ is bounded, the local Lipschitz condition suffices; for the case in which $u(t)$ is unbounded we shall need the following assumption (S) which is of the type considered by McShane in [3f] for the same purpose of extending to unbounded strategies statements proved for bounded strategies.

(S) There is a number $\delta > 0$ and an L-integrable function $S(t)$, $t_1 \leqq t \leqq t_2$, such that $t \in [t_1, t_2]$, $|x' - x(t)| \leqq \delta$, $|x'' - x(t)| \leqq \delta$ implies

$$|f(t, x', u(t)) - f(t, x'', u(t))|, |f_0(t, x', u(t)) - f_0(t, x'', u(t))| \leqq |x' - x''| S(t).$$

We are now in a position to state and prove the following theorem.

THEOREM 9.2. *Let A, $U(t, x)$, M be defined as usual and closed, and let $f_0(t, x, u)$, $f(t, x, u) = (f_1, \cdots, f_n)$ be continuous on M and locally Lipschitzian with respect to x. Let $x(t), u(t), t_1 \leqq t \leqq t_2$, be any admissible pair, whose trajectory x is interior to A. Let us assume that there is a subset E of positive measure in $[t_1, t_2]$ such that for*

every $\bar{t} \in E$ there is a convex combination of points of E_n:

$$\begin{aligned} &\bar{z} = \sum_{j=1}^{n+1} \lambda_j z_j, \qquad \sum_{j=1}^{n+1} \lambda_j = 1, \\ &\bar{z} = f(\bar{t}, \bar{x}, \bar{u}), \quad z_j = f(\bar{t}, \bar{x}, u_j), \quad \bar{x} = x(\bar{t}), \quad \bar{u} = u(\bar{t}), \end{aligned} \tag{9.8}$$

which is generic at $(\bar{t}, \bar{x})$, and such that

$$f_0(\bar{t}, \bar{x}, \bar{u}) > \sum_{j=1}^{n+1} \lambda_j f_0(\bar{t}, \bar{x}, u_j). \tag{9.9}$$

Finally, if $u(t)$, $t_1 \leqq t \leqq t_2$, is unbounded, we assume that condition (S) *holds. Then, there is a sequence of admissible pairs $x_k(t)$, $u_k(t)$, $t_1 \leqq t \leqq t_2$, $k = 1, 2, \cdots$, such that $x_k \to x$ as $k \to \infty$ uniformly in $[t_1, t_2]$ and $I[x, u] > \liminf I[x_k, u_k]$.*

Proof. By a suitable reduction we may well assume E to be a closed, hence compact, subset of $[t_1, t_2]$. Also, since $u(t)$ is measurable in $[t_1, t_2]$, and hence continuous on compact subsets of measure as close to $t_2 - t_1$ as we want, we may well assume that $u(t)$ is continuous on E. If a point $\bar{t} \in E$ and hence relations (9.8), (9.9) hold for certain points u_j and numbers λ_j, $j = 1, \cdots, n+1$, let us prove that there is some closed interval $I = [\bar{t} - \delta, \bar{t} + \delta]$ and functions $p_j(t)$, $t \in E \cap I$, such that $p_j(\bar{t}) = \lambda_j$, $j = 1, \cdots, n+1$, and

$$\begin{aligned} &f(t, x(t), u(t)) = \sum_{j=1}^{n+1} p_j(t) f(t, x(t), u_j), && t \in E \cap I, \\ &1 = \sum_{j=1}^{n+1} p_j(t), && t \in E \cap I, \\ &f_0(t, x(t), u(t)) > \sum_{j=1}^{n+1} p_j(t) f_0(t, x(t), u_j), && t \in E \cap I, \\ &0 < p_j(t) < 1, \quad j = 1, \cdots, n+1, && t \in E \cap I, \\ &\bar{u} \in U(t, x(t)), \quad u_j \in U(t, x(t)), \quad j = 1, \cdots, n+1, && t \in I. \end{aligned} \tag{9.10}$$

Indeed, the first two equations of (9.10), thought of as equations in $p_1, \cdots, p_{n+1}$, are satisfied at $t = \bar{t}$ with $p_j(\bar{t}) = \lambda_j$, $j = 1, \cdots, n+1$, and the functional determinant $\Delta(t)$ is $\neq 0$ at $t = \bar{t}$. To see that we are applying the usual implicit function theorem properly, we should first choose a fixed interval I_0 as I above, then modify $u(t)$ on $I_0 - E$ so as to make u continuous in the whole of I_0, and then determine the interval $I \subset I_0$ by the usual implicit function theorem of calculus on the system so modified. We then disregard the values chosen for $u(t)$ in $I_0 - E$. Finally, we can further reduce I if necessary so as to satisfy also the third, fourth, and fifth of relations (9.10). Note that $\bar{u} = \bar{u}(\bar{t}) \in U(t, x(t))$ together with $u(t) \in U(t, x(t))$ hold for all $t \in I$.

By the process indicated we associate with each point $\bar{t} \in E$ a closed interval $I = [\bar{t} - \delta, \bar{t} + \delta]$ with the properties above. Hence, finitely many of these intervals cover E, say $I_1, \cdots, I_{N_0}$. The endpoints of these intervals can be used to define a subdivision of $[t, t_2]$ into finitely many parts $J_1, \cdots, J_N$. Let $\{J\}'$ be the collection of those J_s which are parts of at least one interval I above, and then let us make

a choice of the corresponding points u_j and multipliers $p_j(t), t \in J_s, j = 1, \cdots, n+1$, for which all relations (9.10) hold. Let $\{J\}''$ be the collection of the remaining intervals J_s.

Note that now relations (9.10) hold in each closed interval $J_s \in \{J\}'$ and for the points u_j and multipliers $p_j(t)$ chosen for the closed interval J_s. By a continuity argument and the closedness of the intervals J_s we conclude that there are some constants $\sigma > 0, \sigma_0 > 0$ such that

$$f(t, x(t), u(t)) = \sum_{j=1}^{n+1} p_j(t) f(t, x(t), u_j), \qquad t \in E \cap J_s,$$

$$1 = \sum_{j=1}^{n+1} p_j(t), \qquad t \in E \cap J_s,$$

$$f_0(t, x(t), u(t)) \geqq \sum_{j=1}^{n+1} p_j(t) f_0(t, x(t), u_j) + \sigma, \qquad t \in E \cap J_s,$$

$$0 < \sigma_0 \leqq p_j(t) \leqq 1 - \sigma_0 < 1, \qquad t \in E \cap J_s,$$

$$\bar{u} \in U(t, x(t)), \qquad u_j \in U(t, x(t)), \qquad t \in J_s,$$

for every $J_s \in \{J\}'$.

Let $\mu > 0$ denote the measure of E.

Now let us define $p_j(t), u_j(t)$ on the whole of $[t_1, t_2]$. For $t \in E$, hence $t \in E \cap J_s$ for some s, $J_s \in \{J\}'$, let us take for $p_j(t), u_j(t) = u_j$, the values already chosen in J_s; for $t \in H = [t_1, t_2] - E$ let us take $p_j(t) = (n+1)^{-1}, u_j(t) = u(t), j = 1, \cdots, n+1$. It is immaterial what choice we make at the endpoints of the intervals J_s. If we consider the differential system

$$\frac{dy}{dt} = \sum_{j=1}^{n+1} p_j(t) f(t, y, u_j(t)), \qquad t_1 \leqq t \leqq t_2,$$

we see that $x(t), t_1 \leqq t \leqq t_2$, is a solution since the second member coincides with $f(t, x(t), u(t))$ for all $t \in [t_1, t_2]$. In other words, $x(t)$, $p(t) = (p_1(t), \cdots, p_{n+1}(t))$, $u(t) = (u_1(t), \cdots, u_{n+1}(t))$ is a generalized system.

Now we shall apply the general theorem of approximation of generalized solutions by means of usual solutions proved in [1a, § 14, (i)]. We need to reduce drastically the control space since in [1a, § 14, (i)] it is assumed that such a control space has to depend on t only. For every $\bar{t} \in E$, hence $\bar{t} \in E \cap J_s, J_s \in \{J\}'$, we take as auxiliary control space $U^*(\bar{t})$ the space made up of the $n+1$ points $u_j, j = 1, \cdots, n+1$. These are the only points we need, since $\bar{z} = f(\bar{t}, \bar{x}, \bar{u})$, $\bar{x} = x(\bar{t})$, $\bar{u} = u(\bar{t})$ is used here only as the convex combination $\sum_{j=1}^{n+1} p_j(\bar{t}) f(\bar{t}, \bar{x}, u_j)$. For every $\bar{t} \in [t_1, t_2] - E$, we take $U^*(\bar{t}) = \{u(\bar{t})\}$, made up, that is, of the sole point $u(\bar{t})$. No other points u are needed.

Let A_0 be a compact neighborhood of the graph G of the trajectory x, so that $G \subset \operatorname{int} A_0 \subset A_0 \subset A$. Let M^* be the set of all (t, x, u) with $(t, x) \in A_0, u \in U^*(t)$. If the original strategy $u(t)$ was bounded, then M^* is a closed bounded set, and f_0 and f are Lipschitzian on M^*. If the original strategy $u(t)$ was unbounded, then M^* is only a closed set, and f_0 and f are locally Lipschitzian on M^*, and property (S) is satisfied. In either case the theorem in [1a, § 14, (i)] holds, and hence there is a

sequence

$$x_k(t) = (x_k^1, \cdots, x_n^k), \qquad u_k(t) = (u_k^1, \cdots, u_k^n), \qquad t_1 \leqq t \leqq t_2, \quad k = 1, 2, \cdots,$$

of (usual) admissible pairs with $x_k \to x$ as $k \to \infty$ uniformly in $[t_1, t_2]$ and such that $\lim I[x_k, u_k] = I[x, p, v]$, or

$$\lim \int_{t_1}^{t_2} f_0(t, x_k(t), u_k(t))\, dt = \int_{t_1}^{t_2} \sum_{j=1}^{n+1} p_j(t) f_0(t, x(t), u_j(t))\, dt.$$

On the other hand,

$$\begin{aligned} I[x, u] &= \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt = \left(\int_H + \int_E \right) f_0\, dt \\ &\geqq \int_H f_0(t, x(t), u(t))\, dt + \int_E \left[\sum_{j=1}^{n+1} p_j(t) f_0(t, x(t), u_j) + \sigma \right] dt \\ &= I[x, p, v] = \sigma\mu = \lim I[x_k, u_k] + \sigma\mu. \end{aligned}$$

This proves Theorem 9.2.

Remark. Theorems 9.1 and 9.2 correspond to analogous statements proved by Tonelli concerning the necessity of the convexity hypothesis for lower semicontinuity in free problems (Tonelli [7]) (see also L. Turner [8]). The present treatment of the question of convexity as a necessary condition for lower closure differs from these essentially because of the use of the theorem of approximation of generalized solutions by means of usual solutions which simplifies an otherwise rather difficult argument in the Lagrange problems under consideration. The present treatment differs also from the recent one by P. Brunovsky [10] because of the use of the same theorem and because of a completely different set of underlying hypotheses.

Appendix. We give the counterexample mentioned in § 4, which is essentially due to A. Lasota and C. Olech [2]. Let $n = 1$ and $A = [0, 1] \times E_1$, let C be a closed Cantor subset of $[0, 1]$ whose measure $m(C)$ is positive, and let $C' = [0, 1] - C$. Then C' is the countable union of disjoint subintervals of $[0, 1]$. Let $s(t)$ be a continuous function on C', which is positive, integrable on C', and which tends to $+\infty$ whenever t tends to an end of any interval component of C'. Let $m = 1$ and define

$$U(t, x) = U(t) = \begin{cases} \{-1\} & \text{if } t \in C, \\ \{u | u \geqq s(t)\} & \text{if } t \in C'. \end{cases}$$

Let $f(t, x, u) = u$. Then $Q(t, x) = Q(t) = U(t)$ and we note that the sets $Q(t)$ so defined satisfy property (Q) in either $A_1 = C \times E_1$ or $A_2 = C' \times E_1$, a decomposition of A into disjoint measurable subsets as described in § 4. Let us extend the function s by taking $s(t) = 0$ when $t \in C$, and consider the decomposition of $[0, 1]$ into k intervals of equal lengths:

$$J_{ks} = [t_{k,s-1}, t_{ks}], \qquad s = 1, \cdots, k.$$

Then $t_{ks} = s/k$. Define u_k by taking

$$u_k(t) = s(t) + v_k(t),$$

where $v_k(t) = -1$ if $t \in C$ and

$$v_k(t) = m(C \cap J_{ks})/m(C' \cap J_{ks})$$

if $t \in C' \cap J_{ks}$. Then u_k is integrable in $[0, 1]$ and $u_k(t) \in U(t)$ for every $t \in [0, 1]$ and k. Let x_k denote the trajectory corresponding to u_k with $x_k(0) = 0$. Thus

$$x_k(t) = \int_0^t v_k(\tau)\, d\tau + \int_0^t s(t)\, d\tau = x(t) + y_k(t).$$

It is easy to see that $y_k(t_{ks}) = 0$ and that $|y_k(t)| \leqq 1/k$. Hence $x_k \to x$ uniformly in $[0, 1]$, where $x'(t) = s(t)$ and $x(0) = 0$. Now x is not an admissible solution of the orientor field under consideration. If we take $f_0 = 0$, we obtain a situation where the closure property of Theorem 4.1 is not true, with the property (Q) satisfied only in A_1 and A_2 separately. As mentioned in § 4, whenever $x_k' \to x'$ weakly, then it is enough to know that property (Q) is satisfied at each set A_k of a decomposition $[A_1, A_2, \cdots]$ of A into countably many disjoint measurable subsets as described in § 4. The counterexample above was suggested by the referee.

REFERENCES

[1a] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints. I and II*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412, 413–429.

[1b] ———, *Existence theorems for optimal controls of the Mayer type*, this Journal, 6 (1968), pp. 517–552.

[1c] ———, *Seminormality and upper semicontinuity in optimal control*, J. Optimization Theory and Applications, 6 (1970), pp. 114–137.

[1d] ———, *Existence theorems for multidimensional problems of optimal control*, Differential Equations and Dynamic Systems, Academic Press, New York, 1967, pp. 115–132.

[1e] ———, *Existence theorems for multidimensional Lagrange problems*, J. Optimization Theory and Applications, 1 (1967), pp. 87–112.

[1f] ———, *Sobolev spaces and multidimensional Lagrange problems of optimization*, Ann. Scuola Norm. Sup. Pisa, 22 (1960), pp. 193–227.

[1g] ———, *Multidimensional Lagrange problems of optimization in a fixed domain and an application to a problem of magnetohydrodynamics*, Arch. Rational Mech. Anal., 29 (1968), pp. 81–104.

[1h] ———, *Existence theorems for abstract multidimensional control problems*, J. Optimization Theory and Applications, 6 (1970), pp. 210–236.

[1i] ———, *An existence theorem in problems of optimal control*, this Journal, 3 (1965), pp. 7–22.

[1j] ———, *Un teorema di esistenza in problemi di controlli ottimi*, Ann. Scuola Norm. Sup. Pisa (3), 19 (1965), pp. 35–78.

[2] A. LASOTA AND C. OLECH, *On Cesari's semicontinuity condition for set valued mappings*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 16 (1968), pp. 711–716.

[3a] E. J. MCSHANE, *On the semicontinuity of integrals in the calculus of variations*, Ann. of Math., 33 (1932), pp. 460–484.

[3b] ———, *Existence theorems for ordinary problems of the calculus of variations*, Ann. Scuola Norm. Sup. Pisa (2), 3 (1934), pp. 181–211, pp. 287–315.

[3c] ———, *Semicontinuity of integrals in the calculus of variations*, Duke Math. J., 2 (1936), pp. 597–616.

[3d] ———, *Some existence theorems for problems in the calculus of variations*, Ibid., 4 (1938), pp. 132–156.

[3e] ———, *A navigation problem in the calculus of variations*, Amer. Math. J., 59 (1937), pp. 327–334.

[3f] ———, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.

[4] E. J. MCSHANE AND R. B. WARFIELD, *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.

[5] T. NISHIURA, *On an existence theorem for optimal control*, this Journal, 5 (1967), pp. 532–544.

[6a] C. OLECH, *Existence theorems for optimal problems with vector valued cost functions*, Trans. Amer. Math. Soc., 136 (1969), pp. 157–180.

[6b] ———, *Existence theorems for optimal control problems involving multiple integrals*, J. Differential Equations, 6 (1969), pp. 512–526.

[7a] L. TONELLI, *Sugli integrali del calcolo delle variazioni in forma ordinaria*, Ann. Scuola Norm. Sup. Pisa (2), 3 (1934), pp. 401–450 = Opere Scelte, vol. 3, Edizioni Cremonese, Roma, 1962, pp. 192–254.

[7b] ———, *Fondamenti di Calcolo delle Variazioni*, 2 vols., Zanichelli, Bologna, 1921–23.

[8] L. TURNER, *The direct method in calculus of variations*. Doctoral thesis, Purdue University, Lafayette, Ind., 1957.

[9] L. CESARI, J. R. LA PALM AND T. NISHIURA, *Remarks on some existence theorems for optimal control*, J. Optimization Theory Appl., 3 (1969), pp. 296–305.

[10] P. BRUNOVSKY, *On the necessity of a certain convexity condition for lower closure of control problems*, this Journal, 6 (1968), pp. 174–185.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

TRACKING VIA FEEDBACK FOR SYSTEMS WITH IRRATIONAL TRANSFER FUNCTION*

M. I. FREEDMAN† AND R. GLASSEY‡

Abstract. Given stable convolution filters G and H, possibly with irrational transfer functions, find a feedback filter C such that for all inputs (discrete stationary processes $(x_n)_0^\infty$ in an ensemble with fixed covariance $(\Gamma_k)_{-\infty}^{\infty}$) the closed-loop system $G(I + CG)^{-1}$ will eventually track the open system H in optimal fashion. The filter C may be stable or unstable but must satisfy suitable realizability conditions. Focus will be on the case where G is not of minimum phase. In this case the system must satisfy an additional restriction, called M-stability, which will exclude the occurrence of "pole-zero" cancellation in the product CG.

1. Introduction. After the fundamental work of Wiener [1], Wiener and Masani [2] and Kolmogorov [3] on the frequency domain approach to prediction and filtering of random processes the viewpoint of many control oriented researchers, following the lead of Kalman–Bucy [4], turned to the time domain. This viewpoint has led to great success in dealing with models described by a system consisting of a finite number of ordinary differential equations influenced by white noise. Application of the Wiener–Hopf technique does not require this "finiteness" condition on the model description, i.e., the transfer function of the operators involved may be irrational.

From one point of view Kalman and Bucy's notion of state estimator, when coupled with the optimal state variable feedback technique (see [5]), amounts to the use of (possibly-unstable) feedback about an open-loop operator in order to cause the overall system to behave in accordance with an a priori prescribed plan.

The applications the authors of this present paper are mainly interested in arise from situations where partial differential equations are involved. As such the transfer functions involved are generally not rational. Thus, a frequency domain approach seems appropriate.

The main problem considered in this paper is heuristically as follows (see Fig. 1):

Given stable convolution filters G and H, possibly with irrational transfer functions, find a feedback filter C such that for all inputs (discrete stationary processes $(x_n)_0^\infty$ in an ensemble with fixed covariance $(\Gamma_k)_{-\infty}^{\infty}$) the closed-loop system $G(I + CG)^{-1}$ will eventually track the open system H in optimal fashion. The filter C may be stable or unstable but must satisfy suitable realizability conditions. Focus will be on the case where G is not of minimum phase. In this case the system must satisfy an additional restriction, called M-stability, which will exclude the occurrence of "pole-zero" cancellation in the product CG. See § 2 for details.

* Received by the editors May 11, 1970, and in revised form October 28, 1970.

† NASA Electronics Research Center, Cambridge, Massachusetts, and Department of Mathematics, Boston University, Boston, Massachusetts 02215.

‡ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

The results found are explicit and computational. They involve the spectral density function $\Gamma(\theta)$ for the processes $(x_n)_0^\infty$ and a Hermite interpolation procedure. Conditions are derived for the existence of an optimal C and a formula for this C is given. Finally an expression for the "tracking error" is determined.

As a final remark we mention that discrete rather than continuous time systems are considered. This allows the convenience of dealing with functions analytic on the interior of the unit disc rather than in a half-plane. As such, certain technical and notational points are simplified, particularly in § 6 and § 7. The translation of the results to the continuous time situation offers no real obstacle.

2. General preliminaries. This short preliminary section begins by briefly reviewing facts about the l_p-spaces and related power series. A few definitions specific to this paper, such as the notion of M-stability, are then presented. The section ends with some very basic remarks about discrete time series.

2.1. Remarks about l_p-spaces and related series. For $1 < p < \infty$, l_p will denote the space of infinite sequences of complex numbers $\{a_k\}_{k=-\infty}^{\infty}$ such that $\sum_{k=-\infty}^{\infty} |a_k|^p < \infty$ and l_p^+ will denote the subspace of l_p consisting of those sequences $\{a_k\}_{k=-\infty}^{\infty}$ with $a_k = 0$ for $k < 0$.

l_p and l_p^+ are Banach spaces (with the norm $\|\{a_k\}\|_p = (\sum_{k=-\infty}^{\infty} |a_k|^p)^{1/p}$) and their properties are both standard and well known. In this paper our dealings will be with the cases $p = 1$ and $p = 2$. l_1 and l_1^+ are Banach algebras under convolution "$*$", where $\{a_k\}_{k=-\infty}^{\infty} * \{b_k\}_{k=-\infty}^{\infty} = \{c_k\}_{k=-\infty}^{\infty}$ with $c_k = \sum_{m=-\infty}^{\infty} a_{k-m} b_m$ and with the identity $I = \{a_k\}_{k=-\infty}^{\infty}$, where $a_0 = 1$, $a_k = 0$ for $k \neq 0$. Also l_2 and l_2^+ are Hilbert spaces with the inner product $\langle \{a_k\}_{k=-\infty}^{\infty}, \{b_k\}_{k=-\infty}^{\infty} \rangle = \sum_{k=-\infty}^{\infty} a_k \bar{b}_k$. We remark that $l_1 \subset l_2$ and $l_1 \subset l_2^+$.

As a shorthand we shall alternatively denote a sequence $\{a_k\}_{k=-\infty}^{\infty}$ by the capital letter A, $\{b_k\}_{k=-\infty}^{\infty}$ by B, etc. AB will refer to $\{a_k\}_{k=-\infty}^{\infty} * \{b_k\}_{k=-\infty}^{\infty}$ where defined. We note that any element $A \in l_1$ induces a bounded linear map $\mathbf{A}: l_p \to l_p$ (for any p, $1 \leqq p$) given by $\mathbf{A}(B) = AB$ for $B \in l_p$ and it is by virtue of this association that we often call elements of l_1, operators.

DEFINITION 2.2. A sequence $A = \{a_k\}_{k=-\infty}^{\infty}$ will be called *causal* if $a_k = 0$ for $k < 0$ and *stable* (alternatively *bounded*) if $A \in l_1$.

For $1 \leqq p < \infty$, $L_p(d\theta)$ will denote the Banach space of complex-valued measurable functions $f(e^{i\theta})$ with $|f|^p$ summable on the unit circle, while for $p = \infty$, $L_\infty(d\theta)$ will denote the space of the essentially bounded complex-valued unit circle. Recall that since the unit circle represents a finite measure space, $L_\infty(d\theta) \subset L_2(d\theta) \subset L_1(d\theta)$.

Given $A = \{a_k\}_{k=-\infty}^{\infty}$ in l_1 we shall denote by $A(e^{i\theta})$ the continuous function on the unit circle with kth Fourier coefficient equal to a_k, i.e., $A(e^{i\theta}) = \sum_{k=-\infty}^{\infty} a_k e^{ik\theta}$, where this series converges absolutely. Further if $A \in l_1^+$ we define $A(z) = \sum_{k=0}^{\infty} a_k z_k$ for z complex, $|z| < 1$. This series converges uniformly absolutely on $|z| < 1$.

If $A = \{a_k\}_{k=-\infty}^{\infty}$ is in l_2, the series $A(e^{i\theta}) = \sum_{k=-\infty}^{\infty} a_k e^{ik\theta}$ converges in $L_2(d\theta)$-norms and so represents a square integrable function on the unit circle. In this case, if $A \in l_2^+$, then $A(z) = \sum_{k=0}^{\infty} a_k z^k$ converges uniformly absolutely on any compact subset of $|z| < 1$ and represents an analytic function on $|z| < 1$. The

Hilbert subspace of $L_2(d\theta)$ corresponding to A in l_2^+ is often denoted by H_2 (Hardy space).

DEFINITION 2.3. For A an element of either l_1^+ or l_2^+ we shall call $A(z)$ the $1/z$-*transform* of A.[1]

For this paper we require a few more facts about the l_p-spaces ($p = 1, 2$) and their transforms:

(a) $(AB)(e^{i\theta}) = A(e^{i\theta})B(e^{i\theta})$ a.e. for A in l_1 and for B in either l_1 or l_2.

(b) If A, B are in l_2, then $\langle A, B\rangle = (1/(2\pi)) \int_0^{2\pi} A(e^{i\theta})\overline{B(e^{i\theta})}\, d\theta$ (the Plancherel theorem).

(c) Given $A \in l_1$, then a necessary and sufficient condition for A to have an inverse in l_1^+ is that $A(z) \neq 0$ on $|z| \leqq 1$, and if this is so, $A^{-1}(z) = 1/A(z)$ (Wiener's theorem, see [6, pp. 72]).

This completes the short summary of standard facts about l_p, l_p^+, $p = 1, 2$, which will be needed.

2.4. The notion of model-stability.

DEFINITION 2.5. Given $G \in l_1^+$, $C \in l_1$ and $Y \in l_1^+$ the 3-tuple $\{G, C; Y\}$ will be termed a *feedback triplet* if $I + CG$ is invertible in l_1 with $Y = (I + CG)^{-1}$.

DEFINITION 2.6. A *feedback triplet* $\{G, C; Y\}$ will be called *model-stable* (or simply M-stable) if given any $G_1 \in l_1^+$ then for all real ε sufficiently small in magnitude, $Y_\varepsilon \overset{\text{def}}{=} (I + C(G + \varepsilon G_1))^{-1}$ remains in l_1^+.

This definition is of importance in the sequel. In § 8, given a fixed $G \in l_1^+$, we desire to minimize a certain integro-quadratic form over an appropriate class of feedback operators $\mathscr{C}$. This class $\mathscr{C}$ must be such as to assure that the "closed-loop system" (i.e., the system defined via feedback, with impulse response $G(I + CG)^{-1}$) corresponds to a legitimate "physical situation." What we require is that not only must $G(I + CG)^{-1}$ be causal and stable (lie in l_1^+) but that under sufficiently small perturbations in G the new closed-loop system must remain causal and stable. The exact definition of the class $\mathscr{C}$ must wait until § 4. For now we conclude with the following easily checked remark which motivated Definition 2.6.

Remark 2.7. In the situation $G \in l_1^+$ and $C \in l_1$ with $G(e^{i\theta})$ and $C(e^{i\theta})$ both representable as rational functions (of the variable $e^{i\theta}$), then M-stability is equivalent to the assumption that no "pole-zero" cancellation occurs (in the interior of the unit disc) in the definition of Y as $(I + CG)^{-1}$.

2.8. Remarks on discrete time series. Let Ω be a space having a Borel field of subsets over which a probability measure P is defined. Let $L_2(\Omega)$ be the set of complex-valued P-measurable functions f on Ω for which $\int_\Omega |f(w)|^2\, dP(\omega) < \infty$; then $L_2(\Omega)$ is a Hilbert space with inner product given by

$$E(f, g) = \int_\Omega f(\omega)\overline{g(\omega)}\, dP(\omega).$$

[1] In engineering literature the z-transform of a causal sequence $\{a_k\}_{k=-\infty}^{\infty} = A$ is usually defined to be $\sum_{k=0}^{\infty} a_k z^{-k}$ making $A(z)$ analytic in the exterior of the unit disc. For our analysis it was decided that the convenience of dealing with functions analytic on the interior of the disc was great enough to justify the definition of $A(z)$ as given below. To emphasize the relationship between $A(z)$ here and the usual z-transform we call our transform the "$1/z$-transform."

DEFINITION 2.9. (a) An element f of $L_2(\Omega)$ is called a *random variable.*

(b) A sequence $(x_n)_0^\infty$ with x_n a random variable for each n is called a *stochastic process* (S.P.).

(c) A stochastic process $(x_n)_0^\infty$ is called *stationary* (wide-sense) if for each integer k with $k + n \geqq 0$, $E(x_{n+k}, x_n) = \Gamma_k$, where Γ_k is a complex number independent of n.

(d) A *spectral density function* $\Gamma(\theta)$ for a stationary S.P. is an a.e. nonnegative function $\Gamma(\theta)$ defined and integrable on the unit circle which has Γ_k as its kth Fourier coefficient., i.e.,

$$\Gamma_k = \frac{1}{2\pi}\int_0^{2\pi} e^{-ik\theta}\Gamma(\theta)\,d\theta \qquad \text{for } -\infty < k < \infty.$$

Not every stationary S.P. has a spectral density function, but throughout this paper we hypothesize their existence for the stationary S.P.'s we consider. (In general, $\Gamma_k = (1/(2\pi))\int_0^{2\pi} e^{-ik\theta}\,dF(\theta)$, $-\infty < k < \infty$, where $F(\theta)$ is a bounded nondecreasing right-continuous function on $[0, 2\pi]$ with $F(0) = 0$. This $F(\theta)$ is called the spectral distribution function of the stationary stochastic process and its existence follows from Bochner's theorem (see [6, p. 42]).)

3. *M*-stability. In this section an alternative characterization of M-stability for a feedback triplet $\{G, C; Y\}$ is presented (assuming $G^{-1} \in l_1$ exists). This new description involves explicitly only the causal operators G and Y and proves useful in later sections.

A lemma based on the variational nature of the M-stability definition is considered first.

LEMMA 3.1. *The feedback triplet $\{G, C; Y\}$ is M-stable if and only if $Y^{n+1}C^n \in l_1^+$ for all integers $n \geqq 1$ (true for $n = 0$ by definition of feedback triplet).*

Proof. Let G_1 be in l_1^+ and for ε a sufficiently small real number let $Y_\varepsilon \triangleq (I + C(G + \varepsilon G_1))^{-1}$ be an element of l_1. For ε sufficiently small, Y_ε has an l_1-convergent Taylor expansion, namely:

$$\begin{aligned} Y_\varepsilon &= (I + CG)^{-1} - \varepsilon CG_1(I + CG)^{-2} + \varepsilon^2 C^2 G_1^2 (I + CG)^{-3} - \cdots \\ &= Y - \varepsilon CY^2 G_1 + \varepsilon^2 C^2 G_1^2 Y^3 - \cdots. \end{aligned} \tag{3.2}$$

Now $\{G, C; Y\}$ will be M-stable provided that for all sufficiently small ε, Y_ε lies in l_1^+. Clearly this will be the case if $Y^{n+1}C^n$ is in l_1^+ for $n \geqq 1$. Conversely let G_1 be the identity I and assume that for all ε sufficiently small the corresponding Y_ε lies in l_1^+. An easy argument shows that $Y^{n+1}C^n$ lies in l_1^+ for $n \geqq 1$, as follows:

Consider $(1/\varepsilon)(Y_\varepsilon - Y)$ which is for each sufficiently small nonzero ε an element of l_1^+. From (3.2) it is clear that as $\varepsilon \to 0$ this expression has an l_1-norm limit equal to Y^2C (as G_1 is I here) and so Y^2C must therefore lie in the subspace l_1^+. The induction scheme showing $Y^{n+1}C^n \in l_1^+$ for general $n \geqq 1$ follows these same simple lines.

THEOREM 3.3. *Let G^{-1} exist in l_1. Then a necessary and sufficient condition for the feedback triplet $\{G, C; Y\}$ to be M-stable is that $G^{-1}(Y - I)$ lie in l_1^+.*

The proof to be presented is of a computational nature and depends to an extent on the following.

Remark 3.4. Given A, B in l_1^+ with $A^{-1} \in l_1$, then a necessary and sufficient condition for $A^{-1}B$ to lie in l_1^+ is that $B(z)/A(z)$ be analytic in $|z| < 1$; i.e., any zero of $A(z)$ in $|z| < 1$ of multiplicity $k \geqq 1$ is likewise a zero of $B(z)$ of at least multiplicity k.

Proof of Remark 3.4. For any integer k,

$$\int_0^{2\pi} e^{ik\theta} \frac{B(e^{i\theta})}{A(e^{i\theta})} d\theta = \frac{1}{2\pi i} \oint_{|z|=1} z^{k-1} \frac{B(z)}{A(z)} dz. \tag{3.5}$$

Now $A^{-1}B \in l_1^+$ if and only if the integral on the left in (3.5) vanishes for each positive k. But the right-hand integral of (3.5) shows that this is exactly the condition that $B(z)/A(z)$ be analytic in $|z| < 1$.

Proof of Theorem 3.3. *Sufficiency.* Suppose the feedback triplet $\{G, C; Y\}$ is M-stable. By Remark 3.4 we must show for each root a of $G(z)$, $|a| < 1$, of multiplicity $k \geqq 1$ that $(Y - 1)^{(j)}(a) = 0$, $j = 0, \cdots, k - 1$. Assume at first that $G(a) = 0$ for a complex, $|a| < 1$. We shall show that $Y(a) = 1$.

To begin consider the expression $Y(I + CG) = I$. It follows that $Y^2 + Y^2CG = Y$ and hence for $|z| < 1$,

$$Y^2(z) + (Y^2CG)(z) = Y(z)$$

or

$$(Y(z))^2 + (Y^2C)(z)G(z) = Y(z) \tag{3.6}$$

since $Y^2C \in l_1^+$ and as such has a "$1/z$-transform."

Since we are assuming $G(a) = 0$ it follows that $(Y(a))^2 = Y(a)$ and therefore $Y(a) = 0$ or $Y(a) = 1$. We proceed to show that the possibility $Y(a) = 0$ is unstable.

Assume therefore that $Y(a) = 0$. For $k \geqq 1$,

$$\begin{aligned} Y^{k+2}C^{k+1}G &= Y^{k+2}C^k(I + CG) - Y^{k+2}C^k \\ &= Y^{k+1}C^k - Y^{k+2}C^k. \end{aligned}$$

So

$$Y^{k+2}C^{k+1}G = (Y^{k+1}C^k)(I - Y). \tag{3.7}$$

Since $Y^{k+2}C^{k+1}$ and $Y^{k+1}C^k$ are both elements of l_1^+ it follows on taking transforms that

$$(Y^{k+2}C^{k+1})(z)G(z) = (Y^{k+1}C^k)(z)(1 - Y(z)). \tag{3.8}$$

Therefore since $G(a) = 0$ and we are assuming $Y(a) = 0$, (3.8) yields $(Y^{k+1}C^k)(a) = 0$ for $k \geqq 1$.

Still under the assumption $Y(a) = 0$ and proceeding by induction we assume

$$Y(a) = Y'(a) = \cdots = Y^{(n-1)}(a) = 0$$

and

$$(Y^{k+1}C^k)(a) = (Y^{k+1}C^k)'(a) = \cdots = (Y^{k+1}C^k)^{(n-1)}(a) = 0$$

for some $n \geqq 0$ and all $k \geqq 0$.

Differentiating (3.8) n-times with respect to z and using the above induction assumptions yields $(Y^{k+1}C^k)^{(n)}(a) = 0$ for all $k \geqq 0$.

Returning to (3.6) we likewise differentiate that expression n times and evaluate at $z = a$. Using the induction assumptions once again yields $Y^{(n)}(a) = 0$ and completes the induction procedure.

Therefore $Y^{(n)}(a) = 0$ for $n = 0, 1, 2, \cdots$. Since $Y(z)$ is analytic in the unit disc $|z| < 1$, it follows that $Y(z) = 0$ in a neighborhood of the origin and so by continuation in the unit disc. This, of course, contradicts the fact that Y is the l_1-inverse of $I + CG$.

It follows, therefore, that $Y(a) = 1$ is the correct conclusion to draw from (3.6).

Assume next that G has a zero at $z = a$ of order $k \geqq 2$, i.e., $G(a) = G'(a) = \cdots = G^{(k-1)}(a) = 0$. We know that $Y(a) = 1$. We additionally now show that $Y'(a) = Y''(a) = \cdots = Y^{(k-1)}(a) = 0$.

To proceed, for each l, $0 < l \leqq k - 1$, differentiate (3.6) l times with respect to z and evaluate at $z = a$. Then

$$\frac{d^l}{dz^l}(Y^2(z))\bigg|_{z=a} = Y^{(l)}(a) \tag{3.9}$$

since G has a kth order zero at $z = a$. Now $2Y(a)Y'(a) = Y'(a)$ shows that $Y'(a) = 0$ since $Y(a) = 1$; proceeding inductively one concludes easily that $Y''(a) = \cdots = Y^{(k-1)}(a) = 0$ completing this half of the proof.

Necessity. Assuming that $G^{-1}(Y - I) \in l_1^+$ we must show that the feedback triplet $\{G, C; Y\}$ is M-stable. The characterization of M-stability given in Lemma 3.1 proves useful here, namely: $\{G, C; Y\}$ will be M-stable if and only if $Y^{n+1}C^n \in l_1^+$ for all integers $n \geqq 1$. Now, by definition, $Y = (I + CG)^{-1}$ so $YCG = I - Y$ and $G^{-1} = -YC$. Therefore $YC \in l_1^+$ and in general for any $n \geqq 1$, $Y(YC)^n = Y^{n+1}C^n \in l_1^+$, completing the proof.

Remark 3.10. In engineering literature the term "minimum phase" is often applied to the situation where $G(z) \neq 0$ in $|z| \leqq 1$. It follows from Theorem 3.3 that given $G \in l_1^+$ of minimum phase, then for any $C \in l_1$ and $Y \in l_1^+$ with $Y = (I + CG)^{-1}$ the feedback triplet $\{G, C; Y\}$ is automatically M-stable, as $(Y(z) - 1)/G(z)$ is analytic on $|z| < 1$ in this case. Our interest in this paper will focus on $G \in l_1^+$ which is *not* of minimum phase.

DEFINITION 3.11. Let $C \in l_1^+$ with $G^{-1} \in l_1$ and define

$$\mathscr{W} = \left\{ C \in l_1 \;\middle|\; \begin{array}{l} Y = (I + CG)^{-1} \text{ exists in } l_1^+ \text{ and the} \\ \text{feedback triplet } \{G, C; Y\} \text{ is } M\text{-stable} \end{array} \right\}.$$

The subset $\mathscr{W}$ will find use in § 8 on minimization.

4. Quasi-extendability. In certain situations we shall wish to associate causal but possibly unstable sequences with stable but possibly noncausal ones (see Definition 2.2). The notion of quasi-extendability is pertinent in this regard.

DEFINITION 4.1. A causal sequence $C = \{c_k\}_{k=-\infty}^{\infty}$ (with $c_k = 0$, $k < 0$) is called *quasi-extendable* if the power series $C(z) = \sum_{k=0}^{\infty} c_k z^k$ is analytic on $|z| < \varepsilon$, ε sufficiently small, and $C(z)$ has a meromorphic extension denoted $C^E(z)$ to $|z| \leqq 1$ such that $C^E(e^{i\theta})$ is continuous for $\theta \in [0, 2\pi]$ and is the Fourier series of some C^E in l_1. C^E will be called the *quasi-extension* of C.

Example 4.2. Let c satisfy $|c| < 1$. Define $B = \{b_k\}_{k=-\infty}^{\infty}$ by

$$b_k = \begin{cases} -c^{-(k+1)}, & k \geqq 0, \\ 0, & k \leqq -1. \end{cases}$$

Then

$$B(z) = \sum_{k=0}^{\infty} b_k z^k = -\frac{1}{c} - \frac{1}{c^2 z} - \frac{1}{c^2 z^2} - \cdots,$$

so $B(z) = (z - c)^{-1}$ for $|z| < |c|$. Let $A = \{a_k\}_{k=-\infty}^{\infty}$ in l_1 be defined by

$$a_k = \begin{cases} 0, & k \geqq 0, \\ c^{-(k+1)}, & k \leqq -1. \end{cases}$$

The $(z - c)^{-1}$ has a meromorphic extension to $|z| \leqq 1$ and it is clear that $B^E = A$.

In § 8 we shall consider a minimization problem taken over a class of operators occurring in feedback. These operators will necessarily be causal due to physical constraints but need not necessarily be stable (bounded). More specifically, we shall have to deal with operators that are quasi-extendable. To help explain the sort of feedback equations involved we proceed with the following definition and theorem.

DEFINITION 4.3. Let $G \in l_1^+$ with $G(e^{i\theta})$ nonzero for $\theta \in [0, 2\pi]$. Define

$$\mathscr{C} = \{C | C \text{ is causal and quasi-extendable to } C^E \text{ and } C^E \in \mathscr{W}\}.$$

Now let $\{x_n\}_0^{\infty}$ lie in l_2^+. Consider the pair of *feedback equations* for $\{e_n\}_0^{\infty}$, $\{f_n\}_0^{\infty}$ in l_2^+ given, for $n \geqq 0$, by

$$\begin{aligned} e_n &= x_n - \sum_{j=0}^{n} c_{n-j} f_j, \\ f_n &= \sum_{l=0}^{n} g_{n-l} e_l. \end{aligned} \tag{4.4}$$

Symbolically we write (see Fig. 1)

$$\mathbf{e} = \mathbf{x} - C\mathbf{f}, \qquad \mathbf{f} = G\mathbf{e},$$

where $\mathbf{e} = \{e_n\}_{n=0}^{\infty}$, $\mathbf{f} = \{f_n\}_{n=0}^{\infty}$ and $\mathbf{x} = \{x_n\}_{n=0}^{\infty}$.

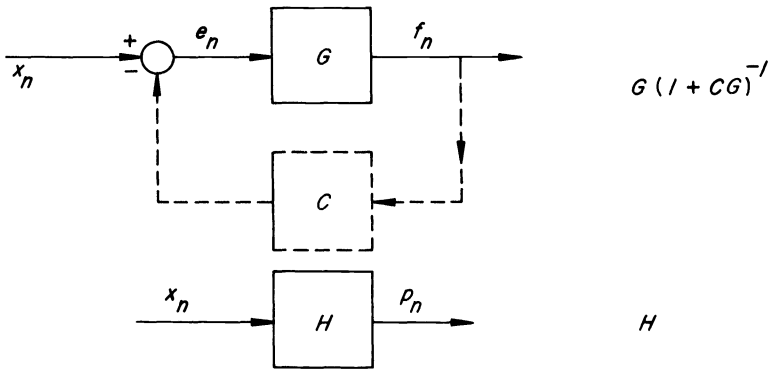


FIG. 1

THEOREM 4.5. *Assume feedback equations* (4.6) *relating* $\mathbf{e}$, $\mathbf{f}$, $\mathbf{x} \in l_2^+$. *Then if* $C = \{c_n\}_{n=-\infty}^{\infty}$ *lies in* $\mathscr{C}$ *it follows that* $\mathbf{e}$ *and* $\mathbf{f}$ *are uniquely specified by* $\mathbf{x}$ *and, in fact, there exists* $D = \{d_k\}_{k=-\infty}^{\infty} \in l_1^+$ *such that* $\mathbf{e} = D\mathbf{x}$ *or* $e_n = \sum_{k=0}^{n} d_{n-k}x_k$ *(i.e., the closed-loop system* $(I + CG)^{-1}$ *is causal and stable). Further if* $G \in l_1^+$ *is replaced by* $G^* \in l_1^+$ *sufficiently close to* G *in* l_1*-norm, then there will likewise exist* $D^* \in l_1$ *with* $\mathbf{e}^* = D^*\mathbf{x}$, *where* $\mathbf{e}^*$ *satisfies the modified* (4.4) ($\mathbf{x}$ *fixed and* G *replaced by* G^*).

Proof. Let $D = \{d_k\}_{k=-\infty}^{\infty} = (I + C^E G)^{-1}$. The rest of the proof follows from the definition of $\mathscr{C}$.

Remark 4.6. Theorem 4.5 suggests that $\mathscr{C}$ is an appropriate class of causal but possibly unstable operators for use in feedback loops. The operators in $\mathscr{C}$ produce closed-loop systems and do so in a physically acceptable way, in that the overall system remains insensitive to slight inaccuracies in the open-loop model.

In the remainder of this section we prove a result concerning quasi-realizability.

THEOREM 4.7. *Let* $\{G, C; Y\}$ *be an M-stable feedback triplet. Then if* $Y(0) \neq 0$, C *is the quasi-extension of some causal sequence* $A = \{b_k\}_{k=-\infty}^{\infty}$, *i.e.,* $C = A^E$.

Proof. $Y(z)$ being analytic on $|z| < 1$, continuous and nonzero on $|z| = 1$ can have only a finite number of zeros within the unit disc. As such, $Y(z)$ may be factored into a product $B(z)H(z)$ valid for $|z| \leqq 1$. Here $H(z)$ is a nonzero analytic function on the disc, arising from an element $H \in l_1^+$, while $B(z)$ is given by the finite product $B(z) = z^p \prod_{i=1}^{n} (z - \alpha_1)^{k_i}$ corresponding to the roots of $Y(z)$ in $|z| < 1$ with appropriate multiplicity. However in this case $p = 0$, as $Y(0) \neq 0$, i.e., $B(z) = \prod_{i=1}^{n} (z - \alpha_i)^{k_i}$. Now let B be that element of l_1^+ corresponding to $B(e^{i\theta})$. It is clear that B^{-1} exists in l_1 and from Example 4.2 one sees that B^{-1} is the quasi-extension of a causal operator (being the product of causal operators). But H is invertible in l_1^+ since $H(z) \neq 0$ for $|z| \leqq 1$. So $H^{-1} \in l_1^+$ is its own quasi-extension. Similarly, $G^{-1}(I - Y) \in l_1^+$ is its own quasi-extension. The result now follows as $C = G^{-1}(I - Y)B^{-1}H^{-1}$.

In keeping with the above we make the following definition.

DEFINITION 4.8. Assume $G \in l_1^+$ with $G^{-1} \in l_1$. Then let $\mathscr{W}$ be as in § 3 and define $\mathscr{W}_0 = \{C \in \mathscr{W} \mid Y(0) \neq 0\}$. (Note that if $G(0) = 0$, then $Y(0) = 1$ and so $\mathscr{W}_0 = \mathscr{W}$.)

5. A density theorem. $\mathscr{W}$ and $\mathscr{W}_0$ have previously been defined. The following additional definitions are also needed in this section.

DEFINITION 5.1. Assume $G \in l_1^+$ with $G^{-1} \in l_1$ and define $T = \{Y \in l_2^+ \mid G^{-1}(Y - I) \in l_2^+\}$ and $T_0 = \{Y \in T \mid Y(0) \neq 0\}$.

Note that $T_0 = T$ if $G(0) = 0$.

ASSUMPTION 5.2. From this point on throughout this paper we shall assume, without always explicitly stating so, that the symbol G represents an operator in l_1^+ with $G(z)$ nonzero on $|z| = 1$, i.e., with G^{-1} lying in l_1.

Remark 5.3. T_0 *is dense in* T (*in the* l_2*-induced topology*).

Remark and Notation 5.4. For $A = \{a_k\}_{k=-\infty}^{\infty}$, $A \in l_p^+$, $1 \leqq p \leqq \infty$, and for any r, $0 < r < 1$, define A_r in $l_1^+ \subset l_2^+$ by

$$A_r = \begin{cases} a_k r^k, & k \geqq 0, \\ 0, & k < 0. \end{cases}$$

Then the Fourier series for A_r, $A_r(e^{i\theta})$ equals $A(re^{i\theta})$, where $A(z)$, $|z| \leqq 1$, represents

as usual the $1/z$-transform of A. Now if $p = 2$, then as $r \to 1$, $A_r \to A$ in l_2-norm (as follows from Parseval's theorem, since $A_r(e^{i\theta}) \to A(e^{i\theta})$ in $L_2(d\theta)$-norm; see [7, p. 32]). Note also that if $A \in l_1^+$, then as $r \to 1$, $A_r \to A$ in l_1-norm (by the Lebesgue bounded convergence theorem).

THEOREM 5.5. *$(I + \mathscr{W}G)^{-1}$ is dense in T in the l_2-induced topology.*

Proof. Let $Y \in T$. Then there exists a sequence $\{r_k\}$ of numbers, $0 < r_k < 1$, with $r_k \to 1$ such that Y_{r_k} has an l_1-inverse for each r_k. This is so because the condition under which a given Y_r will fail to have such an inverse is that $Y(z)$ has a zero on $|z| = r$. $Y(z)$, being analytic, may only have a zero on countably many such circles.

Fixing on this sequence $\{r_k\}$, define $C_{r_k} = G_{r_k}^{-1}(I - Y_{r_k})Y_{r_k}^{-1}$. The C_{r_k} all lie in l_1. The feedback triplets $\{G_{r_k}, C_{r_k}; Y_{r_k}\}$ are all M-stable since $G_{r_k}^{-1}(I - Y_{r_k}) = (G^{-1}(I - Y))_{r_k} \in l_1^+$ as seen from the assumption $Y \in T$ and Remark 5.4.

To complete the proof one must check that the C_{r_k} lie in $\mathscr{W}$ for r_k sufficiently close to 1 and that as $r_k \to 1$, $(I + C_{r_k}G)^{-1}$ tends to Y with respect to the l_2-norm. We omit the lengthy but straightforward computations.

COROLLARY 5.6. (a) *$(I + \mathscr{W}_0G)^{-1}$ is dense in T_0 with respect to the l_2-induced topology.*

(b) *$(I + \mathscr{W}G)^{-1}$ is dense in $T \cap l_1^+$ with respect to the l_1-induced topology.*

(c) *$(I + \mathscr{W}_0G)^{-1}$ is dense in $T_0 \cap l_1^+$ (which is in turn dense in $T \cap l_1^+$) with respect to the l_1-induced topology.*

Essentially the same computations as in Theorem 5.5 are needed.

6. Factorization of the spectral density function θ. Let $\Gamma(\theta)$ be any nonnegative function Lebesgue integrable on the unit circle. In § 8 such a $\Gamma(\theta)$ will represent the spectral density function of a stationary (invariate) stochastic process. We define

$$L_2(\Gamma) = \left\{ Y(e^{i\theta}) \,\middle|\, \begin{array}{l} Y(e^{i\theta}) \text{ is measurable on the unit circle} \\ \text{and is square integrable with respect} \\ \text{to the measure } \Gamma(\theta)\, d\theta \end{array} \right\}$$

and also $L_2^+(\Gamma)$ = the closure in $L_2(\Gamma)$ of trigonometric polynomials of the form $P(e^{i\theta}) = \sum_{k=0}^{n} a_k e^{ik\theta}$.

A classical theorem of prediction theory (attributable to Szegö, Kolmogorov and Krein in various forms; see [7]) states that for $\Gamma(\theta)$ as given,

$$\exp\left\{\frac{1}{2\pi}\int_0^{2\pi} \log \Gamma(\theta)\, d\theta\right\} = \min_{Y(e^{i\theta}) \in L_2^+(\Gamma)} \left\{\frac{1}{2\pi}\int_0^{2\pi} |e^{i\theta}Y(e^{i\theta}) - 1|^2 \Gamma(\theta)\, d\theta\right\}. \tag{6.1}$$

Furthermore, the theorem also states that if the right-hand side of (6.1) is zero (perfect mean square prediction for lag 1), then $(1/(2\pi))\int_0^{2\pi} \log \Gamma(\theta)\, d\theta = -\infty$ and conversely, i.e., the equality (6.1) remains true in this case.

It is also known (see Doob [8, p. 577]) that $(1/(2\pi))\int_0^{2\pi} \log \Gamma(\theta)\, d\theta > -\infty$ is a necessary and sufficient condition for $\Gamma(\theta)$ to be expressible as the absolute value square of an H^2-function, nonzero for $|z| < 1$, i.e., as the product of an H_2-function and its conjugate.

Our concern will be with the somewhat unorthodox minimization problem described below.

First let $T' = \{Y(e^{i\theta}) \in L_2^+(\Gamma) | [G(e^{i\theta})]^{-1}[Y(e^{i\theta}) - 1] \in L_2^+(\Gamma)\}$. Here $G(e^{i\theta})$ is as in Assumption 5.2. We shall consider the problem

$$\min\left\{\frac{1}{2\pi}\int_0^{2\pi} |G(e^{i\theta})Y(e^{i\theta}) - V(e^{i\theta})|^2\Gamma(\theta)\,d\theta \,|\, Y(e^{i\theta}) \in T'\right\},$$

where $V(e^{i\theta})$ lies in $L_2^+(\Gamma)$.

This minimization problem is somewhat different from the one of item (6.1) and arises from an attempt to utilize feedback in accomplishing a certain other minimization procedure described in § 8. The restriction $Y \in T'$ will amount to assuring (see Theorem 3.3) that minimization takes place only over M-stable feedback triplets.

In this present section we restrict ourselves to proving the following analogue of the Szegö–Kolmogorov–Krein theorem quoted above.

THEOREM 6.2. *Let $G \in l_1^+$ with $G(z) \neq 0$ on $|z| = 1$ and let $V(e^{i\theta}) \in L_2^+(\Gamma)$. If the problem*

$$\min\left\{\frac{1}{2\pi}\int_0^{2\pi} |G(e^{i\theta})Y(e^{i\theta}) - V(e^{i\theta})|^2\Gamma(\theta)\,d\theta \,|\, Y(e^{i\theta}) \in T'\right\} \tag{6.3}$$

has a nonzero minimum, then $(1/(2\pi))\int_0^{2\pi} \log \Gamma(\theta)\,d\theta > -\infty$, i.e., $\Gamma(\theta)$ factors into the product of an H^2-function, nonzero for $|z| < 1$, and its conjugate.

Remark 6.4. Note that the integrability of $\log \Gamma(\theta)$ does not in our case imply that (6.3) is nonzero. In fact, the latter will certainly be zero whenever $G(e^{i\theta})Y(e^{i\theta}) = V(e^{i\theta})$ has a solution $Y(e^{i\theta}) \in T'$ and this depends in no essential way on $\log \Gamma(\theta)$. Note also that (6.3) actually does have a minimum and not simply an infimum as $G^{-1} \in l_1$.

The proof of Theorem 6.2 will be preceded by the following.

LEMMA 6.5. *Let $\{K_j\}_{j=0}^m$ and $\{L_j\}_{j=1}^m$ be complex numbers and assume that*

$$K(\theta) = \sum_{j=0}^{m} K_j \cos j\theta + \sum_{j=1}^{m} L_j \sin j\theta \geqq 0 \qquad \text{for all } \theta \in [0, 2\pi].$$

Then

$$\int_0^{2\pi} \log K(\theta)\,d\theta > -\infty. \tag{6.6}$$

Proof. The equation $K(\theta) = 0$ has only a finite number of isolated roots in $[0, 2\pi]$. Clearly if this number is zero, then (6.6) holds. Otherwise the only way that (6.6) might possibly fail to hold true would be that for at least one of these roots θ_0 and any arbitrary small neighborhood N of θ_0, $\int_N |\log K(\theta)|\,d\theta = \infty$. That this possibility actually fails to occur can be seen as follows:

Given θ_0, a root of $K(\theta) = 0$, and any sufficiently small neighborhood N of θ_0 there exists a positive integer n and a function $Q(\theta)$ continuous on $[0, 2\pi]$ such that $K(\theta) = |\theta - \theta_0|^n |Q(\theta)|$ and $Q(\theta) \neq 0$ for $\theta \in N$. Let δ be the diameter of N and let M be an upper bound on $|\log Q(\theta)|$ for $\theta \in N$. Then

$$\begin{aligned}\int_N |\log K(\theta)|\,d\theta &\leqq n\int_{\theta_0-\delta}^{\theta_0+\delta} \log|\theta - \theta_0|\,d\theta + 2M\delta \\ &\leqq 2n\int_0^{\delta} \log \xi\,d\xi + 2M\delta < \infty.\end{aligned}$$

Proof of Theorem 6.2. To begin we note that the map $f: T' \to L_2^+(\Gamma)$ given by $f(Y(e^{i\theta})) = [G(e^{i\theta})]^{-1}(Y(e^{i\theta}) - 1)$ is one-to-one and onto and so the minimum of (6.3) must be equal to

$$\text{(6.7)} \quad \min\left\{\frac{1}{2\pi}\int_0^{2\pi} |G(e^{i\theta})(1 + G(e^{i\theta})u(e^{i\theta}) - V(e^{i\theta}))|^2\Gamma(\theta)\,d\theta|\ u(e^{i\theta}) \in L_2^+(\Gamma)\right\}.$$

Now (6.7) corresponds to the problem of finding the unique element $w^*(e^{i\theta})$ in the closed submanifold $S = \{w(e^{i\theta}) = G^2(e^{i\theta})u(e^{i\theta}) | u(e^{i\theta}) \in L_2^+(\Gamma)\}$ of $L_2^+(\Gamma)$ which is closest (in $L_2(\Gamma)$-norm) to the element $v(e^{i\theta}) - G(e^{i\theta})$ of $L_2^+(\Gamma)$. Clearly this element $w^*(e^{i\theta})$ determines uniquely $u^*(e^{i\theta})$ in $L_2^+(\Gamma)$ with $G^2(e^{i\theta})u^*(e^{i\theta}) = w^*(e^{i\theta})$.

From the projection theorem for Hilbert space it follows that a necessary and sufficient condition for $u^*(e^{i\theta})$ to be the element of $L_2^+(\Gamma)$ minimizing (5.7) is that

$$\text{(6.8)} \quad \int_0^{2\pi} [G^2(e^{i\theta})u^*(e^{i\theta}) - (v(e^{i\theta}) - G(e^{i\theta}))]\overline{G^2(e^{i\theta})\eta(e^{i\theta})}\Gamma(\theta)\,d\theta = 0$$

for all $\eta(e^{i\theta}) \in L_2^+(\Gamma)$, i.e., that $G^2(e^{i\theta})u^*(e^{i\theta}) - (v(e^{i\theta}) - G(e^{i\theta}))$ is orthogonal to S.

Translating back to the minimization problem (6.3) over T' we see that (6.3) must have a unique minimum for some $Y(e^{i\theta}) \in T'$ and that a necessary and sufficient condition for $Y^*(e^{i\theta})$ to be this minimal element is that

$$\text{(6.9)} \quad \int_0^{2\pi} [G(e^{i\theta})Y^*(e^{i\theta}) - V(e^{i\theta})]\overline{G^2(e^{i\theta})\eta(e^{i\theta})}\Gamma(\theta)\,d\theta = 0$$

for all $\eta(e^{i\theta}) \in L_2^+(\Gamma)$.

To proceed with the proof we note that since $G \in l_1^+$ with $G(z) \neq 0$ on $|z| = 1$ it follows that $G(z)$ has only a finite number p (counting multiplicity) of roots in $|z| < 1$. We may factor $G(z)$ into a product $B(z)H(z)$ (as was done to $Y(z)$ in the proof of Theorem 4.7), where $B(z)$ is a polynomial of degree p with roots and multiplicities corresponding to those of $G(z)$, while $H(z)$ is continuous and nonzero on the closed disc and analytic on the open disc. Let

$$\eta(e^{i\theta}) = [G(e^{i\theta})Y^*(e^{i\theta}) - V(e^{i\theta})][H(e^{i\theta})]^{-2}B^2(e^{i\theta})e^{2pi\theta}r(e^{i\theta}),$$

where $r(e^{i\theta})$ is an arbitrary element of $L_2^+(\Gamma)$. Then $\eta(e^{i\theta})$ is in $L_2^+(\Gamma)$ since the negative Fourier coefficients of $[H(e^{i\theta})]^{-2}$ vanish as well as the negative Fourier coefficients of $B^2(e^{i\theta})e^{2pi\theta}$.

Using this $\eta(e^{i\theta})$ in (6.9) gives

$$\text{(6.10)} \quad \int_0^{2\pi} |G(e^{i\theta})Y^*(e^{i\theta}) - V(e^{i\theta})|^2|B(e^{i\theta})|^4\Gamma(\theta)e^{2pi\theta}r(e^{i\theta})\,d\theta = 0$$

for $r(e^{i\theta}) \in L_2^+(\Gamma)$. Making the choices $e^{ik\theta}$, $k \geqq 0$, for $r(e^{i\theta})$ in (6.10) yields

$$\int_0^{2\pi} |G(e^{i\theta})Y^*(e^{i\theta}) - V(e^{i\theta})|^2|B(e^{i\theta})|^4\Gamma(\theta)e^{-il\theta}\,d\theta = 0$$

for $l \geqq 2p$ and by conjugation the same is true for $l \leqq -2p$. From these orthogonality conditions it follows that

$$\text{(6.11)} \quad |G(e^{i\theta})Y^*(e^{i\theta}) - V(e^{i\theta})|^2|B(e^{i\theta})|^4\Gamma(\theta) = K(\theta)$$

for some $K(\theta) = \sum_{j=0}^{p} k_j \cos j\theta + \sum_{j=1}^{p} L_j \sin j\theta$ with $K(\theta)$ nonnegative (the K_j and L_j are real). Define

$$S(e^{i\theta}) = \frac{(G(e^{i\theta})Y^*(e^{i\theta}) - V(e^{i\theta}))^{-1}K(\theta)e^{-i\theta}}{\overline{B^2(e^{i\theta})}}.$$

We claim that $S(e^{i\theta})$ lies in H_2. First $|S(e^{i\theta})|^2 = K(\theta)\Gamma(\theta)$ as follows from (6.11) and therefore $S(e^{i\theta}) \in L_2(d\theta)$ since $\Gamma(\theta) \in L_1(d\theta)$ (while of course $K(\theta) \in L_\infty(d\theta)$). Now $S(e^{i\theta})$ will lie in H_2 if

$$\int_0^{2\pi} S(e^{i\theta})e^{ik\theta}\,d\theta = 0 \qquad \text{for } k = 1, 2, \cdots.$$

To see this we conjugate (6.9) to find that for any $\eta(e^{i\theta})$ in $L_2^+(\Gamma)$,

$$\int_0^{2\pi} \overline{[G(e^{i\theta})Y^*(e^{i\theta}) - V(e^{i\theta})]}G^2(e^{i\theta})\Gamma(\theta)\eta(e^{i\theta})\,d\theta = 0. \tag{6.12}$$

Now for $k \geqq 1$,

$$\int_0^{2\pi} S(e^{i\theta})e^{ik\theta}\,d\theta = \int_0^{2\pi} \frac{[G(e^{i\theta})Y^*(e^{i\theta}) - V(e^{i\theta})]K(\theta)}{\overline{B^2(e^{i\theta})}} e^{i(k-1)\theta}\,d\theta \tag{6.13}$$

as seen from using (6.11). But (6.13) is zero for $k \geqq 1$ as it corresponds to (6.12) with $\eta(e^{i\theta}) = [H(e^{i\theta})]^{-2}e^{i(k-1)\theta}$ which is in $L_2^+(\Gamma)$ for each $k \geqq 1$. Thus $S(e^{i\theta})$ is a nonzero function in H_2. Now the well-known Jensen inequality of classical complex variable theory (see [7, p. 68]) assures us that

$$\int_0^{2\pi} \log|S(e^{i\theta})|\,d\theta > -\infty.$$

But since $|S(e^{i\theta})|^2 = K(\theta)\Gamma(\theta)$ it now follows that $\int_0^{2\pi} \log K(\theta)\,d\theta + \int_0^{2\pi} \log \Gamma(\theta)\,d\theta > -\infty$. By Lemma 6.5, $\log K(\theta)$ is integrable and so $\int_0^{2\pi} \log \Gamma(\theta)\,d\theta > -\infty$. Hence also $\Gamma(\theta)$ factors.

7. A frequency domain minimization problem. On the basis of the results in § 6 we shall assume in this section that the nonnegative function $\Gamma(\theta)$, Lebesgue integrable on the unit circle, also satisfies the condition $\int_0^{2\pi} \log \Gamma(\theta)\,d\theta > -\infty$. Let $R(e^{i\theta})$ denote the H_2-function, with $R(z)$ nonzero for $|z| < 1$ such that $\Gamma(\theta) = |R(e^{i\theta})|^2$ a.e.

We note that the integrability of $\log \Gamma(\theta)$ assures that $R(e^{i\theta}) \neq 0$ a.e. in $[0, 2\pi]$. Also the Hilbert spaces H_2 and $L_2^+(\Gamma)$ are isometric under the linear map $Y(e^{i\theta}) \to Y(e^{i\theta})[R(e^{i\theta})]^{-1}$.

In this section we shall investigate properties of the minimization problem

$$\min\left\{\frac{1}{2\pi}\int_0^{2\pi} |G(e^{i\theta})Y(e^{i\theta}) - V(e^{i\theta})|^2\Gamma(\theta)\,d\theta \;\middle|\; Y(e^{i\theta}) \in T'\right\}, \tag{7.1}$$

where $V(e^{i\theta})$ in $L_2^+(\Gamma)$ is fixed. (That this problem actually has a minimum was seen in § 6.)

Here G once again represents an operator in l_1^+ with $G(z)$ nonzero on $|z| = 1$. As such, $G(z)$ can have only a finite number P of roots in $|z| < 1$. As in § 6 we may

write $G(z) = B(z)H(z)$, where $H(z)$ is a continuous nonvanishing function on the closed unit disc analytic on the open disc (which is the $1/z$-transform of an element H in l_1^+) while $B(z) = \prod_{j=1}^{n} (z - a_j)^{k_j}$, where $\{a_j\}_{j=1}^{n}$, a_j complex, and $\{k_j\}_{j=1}^{n}$, k_j nonnegative integers, represent the roots of $G(z)$ in $|z| < 1$ and their multiplicities respectively (including possible roots at $z = 0$).

Letting $k(e^{i\theta}) = V(e^{i\theta})R(e^{i\theta})$ we have $k(e^{i\theta}) \in H_2$. It is convenient to consider the minimization problem (7.1) in the alternate form

$$\min\left\{\frac{1}{2\pi}\int_0^{2\pi} |G(e^{i\theta})s(e^{i\theta}) - k(e^{i\theta})|^2\, d\theta \,\middle|\, s(e^{i\theta}) \in H_2 \text{ with } s(e^{i\theta})[R(e^{i\theta})]^{-1} \in T'\right\}. \tag{7.2}$$

The advantage is that here we deal with underlying Hilbert spaces $L_2(d\theta)$ and H_2. The minimum value attained by (7.2) is clearly the same as the minimum of (7.1), and if $s^*(e^{i\theta})$ is the point where (7.2) attains its minimum then $Y^*(e^{i\theta}) = s^*(e^{i\theta})[R(e^{i\theta})]^{-1}$ is the point in T' where (7.1) does so.

A necessary and sufficient condition for an element $s(e^{i\theta})$ of H_2 to satisfy $s(e^{i\theta})[R(e^{i\theta})]^{-1} \in T'$ is that $s(a_j) = R(a_j)$, $s^1(a_j) = R^1(a_j)$, $\cdots$, $s^{(k_j-1)}(a_j) = R^{(k_j-1)}(a_j)$ for $j = 1, \cdots, n$, i.e., that $f(z) = (s(z) - R(z))/G(z)$ is analytic in $|z| < 1$.

A procedure based on Hermite interpolation will be exhibited for finding the optimal $s^*(e^{i\theta})$. In this direction we first prove the following.

THEOREM 7.3. *A necessary and sufficient condition for the minimization problem* (7.2) *to attain its minimum at a function* $s^*(e^{i\theta})$ *in* H_2 *with* $s^*(e^{i\theta})[R(e^{i\theta})]^{-1}$ *in* T' *is that there exist a trigonometric polynomial of order* $2P - 1$ *or less,* $M(e^{i\theta}) = \sum_{k=0}^{2P-1} m_k e^{ik\theta}$, *such that*

$$[G(e^{i\theta})s^*(e^{i\theta}) - k(e^{i\theta})]\overline{B^2(e^{i\theta})}e^{2Pi\theta} = M(e^{i\theta}) \quad a.e. \tag{7.4}$$

Proof. Sufficiency. In § 6, (6.9) is a necessary and sufficient condition for (7.1) to have its minimum at $Y^*(e^{i\theta})$ in T'. Translating (6.9) to the corresponding condition for (7.2) yields that a necessary and sufficient condition for (7.2) to have its minimum at $s^*(e^{i\theta})$ in H_2 with $s^*(e^{i\theta})[R(e^{i\theta})]^{-1}$ in T' is that

$$\int_0^{2\pi} [G(e^{i\theta})s^*(e^{i\theta}) - k(e^{i\theta})]\overline{G^2(e^{i\theta})\eta(e^{i\theta})}\, d\theta = 0 \tag{7.5}$$

for all $\eta(e^{i\theta})$ in H_2.

Letting $\eta(e^{i\theta}) = [H(e^{i\theta})]^{-2}r(e^{i\theta})$ for $r(e^{i\theta})$ an arbitrary element of H_2, we may rewrite (7.5) as

$$\int_0^{2\pi} [G(e^{i\theta})s^*(e^{i\theta}) - k(e^{i\theta})]\overline{B^2(e^{i\theta})}e^{2Pi\theta}\overline{e^{2Pi\theta}r(e^{i\theta})}\, d\theta = 0 \tag{7.6}$$

for all $r(e^{i\theta})$ in H_2. Now the term $M(e^{i\theta}) = [G(e^{i\theta})s^*(e^{i\theta}) - k(e^{i\theta})]\overline{B^2(e^{i\theta})}e^{2Pi\theta}$ must lie in H_2, as $B^2(e^{i\theta})e^{2Pi\theta}$ has all negative Fourier coefficients zero. From (7.6) it follows that

$$\int_0^{2\pi} M(e^{i\theta})e^{-ik\theta}\, d\theta = 0 \qquad \text{for all } k \geqq 2P \tag{7.7}$$

as seen by making the appropriate choices for $r(e^{i\theta})$. But (7.7) together with the fact that $M(e^{i\theta})$ is in H_2 implies that $M(e^{i\theta})$ is a trigonometric polynomial of degree $\leqq 2P-1$, as said in the statement of the theorem.

Conversely, if $[G(e^{i\theta})s^*(e^{i\theta}) - k(e^{i\theta})]\overline{B^2(e^{i\theta})}e^{2Pi\theta} = M(e^{i\theta})$ for some trigonometric polynomial $M(e^{i\theta})$ of degree $\leqq 2P-1$, then (7.7) must hold, and since the $\{e^{ik\theta}\}_{k=0}^{\infty}$ span H_2, (7.6) also must hold. The choice $r(e^{i\theta}) = [H(e^{i\theta})]^2\eta(e^{i\theta})$ then yields (7.5), the previously established necessary and sufficient condition for $s^*(e^{i\theta})$ to be the minimizing function for (7.2).

Remark 7.8. (a) For a given $s^*(e^{i\theta})$ in H_2 the validity of (7.4) for some $M(e^{i\theta})$ a trigonometric polynomial of degree $\leqq 2P-1$ combined with the conditions $s^*(a_j) = R(a_j), \cdots, s^{*(k_j-1)}(a_j) = R^{(k_j-1)}(a_j)$ for $j = 1, \cdots, n$ suffices to determine $s^*(e^{i\theta})$ uniquely. This is true because (7.2) has a unique minimum. The procedure by which these conditions determine $s^*(e^{i\theta})$ will be spelled out in Theorem 7.11 below.

(b) For notational convenience let $B_1(e^{i\theta}) = \overline{B(e^{i\theta})}e^{Pi\theta} = \prod_{i=1}^{n}(1-\bar{a}_ie^{i\theta})^{k_i}$ for $\theta \in [0, 2\pi]$ and similarly $B_1(z) = \prod_{i=1}^{n}(1-\bar{a}_iz)^{k_i}$ for $|z| \leqq 1$. Now $B_1(z)$ so given and likewise $[B_1(z)]^{-1}$ are continuous nonvanishing functions on $|z| \leqq 1$ which are analytic on $|z| < 1$. The condition $[G(e^{i\theta})s^*(e^{i\theta}) - k(e^{i\theta})]B_1^2(e^{i\theta}) = M(e^{i\theta})$ a.e. implies that for $|z| < 1$,

$$[G(z)s^*(z) - k(z)]B_1^2(z) = M(z). \tag{7.9}$$

Definition and Remark 7.10. Let $F \in l_1^+$ with $F(e^{i\theta}) \neq 0$, $\theta \in [0, 2\pi]$. Then $F(z)$ has a finite number q (counting multiplicity) of roots in $|z| < 1$. Let $\{b_j\}_{j=1}^{n}$, $\{l_j\}_{j=1}^{n}$ be the roots and their multiplicities respectively. For any function $W(e^{i\theta})$ in H_2, $(W(z))_F$ (read $W(z)$ mod F) will denote that unique polynomial in z of order $\leqq q-1$ such that $(W(z) - (W(z))_F)/F(z)$ is analytic in $|z| < 1$.

Clearly $(W(z))_F \equiv 0$ if $F(z)$ has no zeros in $|z| < 1$. In general, $(W(z))_F$ represents a *Hermite interpolation* with respect to conditions specified by the roots of $F(z)$ and as such $(W(z))_F$ depends only on the values $W(b_j), W^1(b_j), \cdots, W^{(l_j-1)}(b_j)$, $j = 1, \cdots, n$. An explicit construction of $(W(z))_F$ is as follows:

Let $v(z) = \prod_{i=1}^{n}(z-b_i)^{l_i}$ and define

$$r_{ik}(z) = v(z)\frac{(z-z_i)^{k-l_i}}{k!}\frac{d^{(l_i-1-k)}}{dz^{(l_i-1-k)}}\left[\frac{(z-z_i)^{l_i}}{v(z)}\right]_{z=z_i}$$

for $1 \leqq i \leqq n$, $0 \leqq k \leqq l_i - 1$. Then

$$(W(z))_F = \sum_{i=1}^{n}[W(b_i)r_{i0}(z) + W^1(b_i)r_{i1}(z) + \cdots + W^{(l_i-1)}(b_i)r_{i,l_i-1}(z)].$$

(This formula appears in [9].)

THEOREM 7.11. *The function $s^*(e^{i\theta})$ for which the variational problem* (7.2) *attains its minimum is given by*

$$s^*(e^{i\theta}) = \frac{L(e^{i\theta}) + B_1^2(e^{i\theta})k(e^{i\theta})}{G(e^{i\theta})B_1^2(e^{i\theta})} \qquad \text{for } \theta \in [0, 2\pi],$$

where $L(z)$ is that polynomial of degree $\leqq 2P-1$ given by $L(z) = ([G(z)R(z) - k(z)]B_1^2(z))_{G^2}$.

Proof. $s^*(e^{i\theta})$ as defined clearly lies in $L_2(d\theta)$. Actually $s^*(e^{i\theta})$ also lies in H_2. This can be checked by showing that $[L(z) + B_1^2(z)k(z)]/G(z)B_1^2(z)$ is analytic in $|z| < 1$. But by the definition of $L(z)$,

$$\frac{[G(z)R(z) - k(z)]B_1^2(z) - L(z)}{G^2(z)} \tag{7.12}$$

is analytic in $|z| < 1$.

Multiplying (7.12) by the analytic function $G(z)[B_1(z)]^{-2}$ shows that

$$R(z) - \frac{L(z) + B_1^2(z)k(z)}{G(z)B_1^2(z)}$$

is analytic in $|z| < 1$, and since $R(z)$ is analytic in $|z| < 1$ it follows that $(L(z) + B_1^2(z)k(z))/G(z)B_1^2(z)$ is, i.e., $s^*(e^{i\theta})$ is in H_2.

Clearly then $[G(z)s^*(z) - k(z)]B_1^2(z) = L(z)$ for $|z| \leqq 1$ and substituting for $L(z)$ in (7.12) yields that $[R(z) - s^*(z)]B_1^2(z)/G(z)$ is analytic in $|z| < 1$. Multiplying this last expression by the analytic function $[B_1(z)]^{-2}$ shows that $(R(z) - s^*(z))/G(z)$ is analytic in $|z| < 1$. Hence $s^*(e^{i\theta})[R(e^{i\theta})]^{-1}$ is in T'. The hypothesis (7.4) now holds with $M(e^{i\theta}) = L(e^{i\theta})$ and so Theorem 7.3 assures that $s^*(e^{i\theta})$ so defined is the optimal for variational problem (7.2).

THEOREM 7.13. *The minimizing element* $Y^*(e^{i\theta})$ *in* T' *for variational problem* (7.1) *is given by*

$$Y^*(e^{i\theta}) = \frac{L(e^{i\theta}) + B_1^2(e^{i\theta})V(e^{i\theta})R(e^{i\theta})}{R(e^{i\theta})G(e^{i\theta})B_1^2(e^{i\theta})}, \qquad \theta \in [0, 2\pi],$$

where $L(z)$ *is that polynomial of degree* $\leqq 2P - 1$ *given by* $L(z) = ((G(z) - V(z))R(z)B_1^2(z))_{G^2}$.

Proof. $R(z)Y^*(z) = s^*(z)$ and $R(z)V(z) = k(z)$.

Remark 7.14. (a) Suppose $R(e^{i\theta})$ derives from $R \in l_1^+$ with R^{-1} also in l_1^+ (i.e., with $R(z) \neq 0$ on $|z| \leqq 1$). Then if $V(e^{i\theta})$ likewise comes from $V \in l_1^+$ it follows that the minimizing function $Y^*(e^{i\theta})$ for problem (7.1) must also arise from some $Y^* \in l_1^+$.

(b) Alternatively, suppose there exist positive constants K_1 and K_2 with $0 < K_1 \leqq \Gamma(\theta) \leqq K_2$ a.e. Then $R(e^{i\theta})$ and $[R(e^{i\theta})]^{-1}$ both lie in $L_\infty(d\theta) \cap H_2$ and if $V(e^{i\theta}) \in L_\infty(d\theta) \cap H_2$ it follows that $Y^*(e^{i\theta}) \in L_\infty(d\theta) \cap H_2$.

(c) Alternatively, if $R(z)$ is rational in z and $V(z)$ is rational in z, then $Y^*(z)$ will be a rational function in z.

The minimum value of the variational problems (7.1) or (7.2) which we shall henceforth denote by $J_{\min}$ can now be computed from $Y^*(e^{i\theta})$ or $s^*(e^{i\theta})$, for

$$J_{\min} = \frac{1}{2\pi}\int_0^{2\pi} |G(e^{i\theta})Y^*(e^{i\theta}) - V(e^{i\theta})|^2\Gamma(\theta)\,d\theta$$

$$= \frac{1}{2\pi}\int_0^{2\pi} |G(e^{i\theta})s^*(e^{i\theta}) - k(e^{i\theta})|^2\,d\theta.$$

There is an alternative often more attractive computational method based on contour integration and residue theory. This method does not involve preliminary

computation of either $Y^*(e^{i\theta})$ or $s^*(e^{i\theta})$ (or for that matter computation of $L(z)$).

To begin note that from (7.4) one has

$$J_{\min} = \frac{1}{2\pi}\int_0^{2\pi}\left|\frac{L(e^{i\theta})}{B_1^2(e^{i\theta})}\right|^2 d\theta,$$

where $L(e^{i\theta})$ is the trigonometric polynomial of degree $\leqq 2P-1$ defined in Theorem 7.11. Now $B_1(e^{i\theta}) = e^{Pi\theta}\overline{B(e^{i\theta})}$. So $|B_1(e^{i\theta})| = |B(e^{i\theta})|$ for $\theta \in [0, 2\pi]$. It follows that

$$J_{\min} = \frac{1}{2\pi}\int_0^{2\pi}\left|\frac{L(e^{i\theta})}{B^2(e^{i\theta})}\right|^2 d\theta.$$

The function $L(e^{i\theta})[B(e^{i\theta})]^2$ is a rational (trigonometric) function which has a Laurent expansion in $e^{i\theta}$ with only negative powers, i.e., all nonnegative Fourier coefficients vanish.

Letting $L(e^{i\theta})[B(e^{i\theta})]^{-2} = \sum_{l=1}^{\infty} s_l e^{-il\theta}$ we have

$$s_k = \frac{1}{2\pi i}\oint_{|z|=1}\frac{L(z)}{B^2(z)}dz \qquad \text{for } k = 1, 2, \cdots,$$

and so by Parseval's theorem, $J_{\min} = \sum_{k=1}^{\infty}|s_k|^2$.

The calculation of the s_k as indicated by the above contour integral seems to involve prior computation of the polynomial $L(z)$. It is this computation we shall be able to avoid. In this direction consider the equation $[G(z)s^*(z) - k(z)]B_1^2(z) = L(z)$ valid for $|z| \leqq 1$. Multiplying through by $z^{k-1}[B(z)]^{-2}$ for any $k \geqq 1$ and integrating around the unit circle gives

$$2\pi i s_k = \oint_{|z|=1} z^{k-1}\frac{s^*(z)G(z)B_1^2(z)}{B^2(z)}dz - \oint_{|z|=1} z^{k-1}\frac{k(z)B_1^2(z)}{B^2(z)}dz \tag{7.15}$$

for $k = 1, 2, \cdots$. Recalling that $(s^*(z) - R(z))/G(z) = A(z)$ for some analytic function $A(z)$ on $|z| < 1$ (actually $A(e^{i\theta}) \in H_2$) we have that

$$\begin{aligned}\oint_{|z|=1} z^{k-1}\frac{s^*(z)G(z)B_1^2(z)}{B^2(z)}dz &= \oint_{|z|=1} z^{k-1}\frac{R(z)G(z)B_1^2(z)}{B(z)}dz \\ &\quad + \oint_{|z|=1} z^{k-1}\frac{A(z)G^2(z)B_1^2(z)}{B^2(z)}dz\end{aligned} \tag{7.16}$$

for $k = 1, 2, \cdots$. Now the second integral on the right in (7.16) vanishes, as the function $A(z)G^2(z)B_1^2(z)/B(z)$ is analytic in $|z| < 1$.

Therefore (7.15) may be rewritten as

$$2\pi i s_k = \oint_{|z|=1} z^{k-1}\frac{R(z)G(z)B_1^2(z)}{B^2(z)}dz - \oint_{|z|=1} z^{k-1}\frac{k(z)B_1^2(z)}{B^2(z)}dz. \tag{7.17}$$

For convenience we now let $T(z)$ be the rational function given by

$$T(z) = \frac{B(z)}{B_1(z)} = \prod_{i=1}^{n}\left(\frac{z-a_i}{1-\bar{a}_i z}\right)^{k_i}.$$

Then making a substitution in (7.17) we have proved the following theorem.

THEOREM 7.18. *The minimum value taken by the variational problems* (7.1) *or* (7.2) *is given by*

$$J_{\min} = \sum_{k=1}^{\infty} |s_k|^2,$$

where

$$s_k = \frac{1}{2\pi i}\oint_{|z|=1} z^{k-1}\frac{[R(z)G(z) - k(z)]}{[T(z)]^2}dz$$

$$= \frac{1}{2\pi i}\oint_{|z|=1} z^{k-1}\frac{[R(z)(G(z) - V(z))]}{[T(z)]}dz.$$

Here

$$T(z) = \prod_{i=1}^{n}\left(\frac{z - a_i}{1 - \bar{a}_i z}\right)^{k_i},$$

the $\{a_i\}_{i=1}^n$, $\{k_i\}_{i=1}^n$ *being the roots of* $G(z)$ *in* $|z| < 1$ *and their multiplicities respectively.*

Remark 7.19. The rational analytic function $T(z)$ (on $|z| < 1$) is a (finite) Blaschke product for $G(z)$ in the sense of [7, p. 66]; i.e., $|T(e^{i\theta})| = 1$ for $\theta \in [0, 2\pi]$, $G(z)/T(z)$ is analytic in $|z| < 1$ and $T(z)$ is a finite product of bilinear functions in z.

8. Minimization via feedback. Let $(x_n)_0^\infty$ be a stationary (wide sense) stochastic process with a given known covariance sequence $(\Gamma_n)_{-\infty}^{\infty}$. Let $G = \{g_n\}_{-\infty}^{\infty}$ be in l_1^+. Assume $G(e^{i\theta}) = \sum_{n=0}^{\infty} g_n e^{in\theta}$ is nonzero for $\theta \in [0, 2\pi]$. As in § 4 define

$$\mathscr{C} = \left\{ C = \{c_k\}_{k=-\infty}^{\infty} \;\middle|\; \begin{array}{l} \text{(i) } C \text{ is causal and quasi-extendable,} \\ \text{(ii) } Y^E = (1 + C^E G)^{-1} \text{ exists in } l_1^+, \\ \text{(iii) } \{G, C; Y\} \text{ is an } M\text{-stable feedback triplet} \end{array} \right\}.$$

Consider the feedback system defining (nonstationary) stochastic processes $(e_n)_0^\infty$ and $(f_n)_0^\infty$ given by the equations

$$e_n = x_n - \sum_{j=0}^{n} c_{n-j} f_j \quad \text{and} \quad f_n = \sum_{l=0}^{n} g_{n-l} e_l, \tag{8.1}$$

where $C = \{c_k\}_{k=-\infty}^{\infty}$ is an element of $\mathscr{C}$. Here we regard (8.1) as a *closed-loop* input-output system with $(x_n)_0^\infty$ to be viewed as the *input* and $(f_n)_0^\infty$ as the *output*. (See Fig. 1.)

Also let $H = \{h_k\}_{k=-\infty}^{\infty}$ with h_k real for each h be a fixed element of l_1^+. We define the (nonstationary) stochastic process $(p_n)_0^\infty$ by the equation

$$p_n = \sum_{j=0}^{n} h_{n-j} x_j. \tag{8.2}$$

Here $(x_n)_0^\infty$ is viewed as the *input* and $(p_n)_0^\infty$ as the output of the system (*open loop*) described by (8.2) (see again Fig. 1).

The problem treated in this section is that of selecting a *feedback operator* C in $\mathscr{C}$ which most nearly (in a statistical sense) causes the system of (8.1) to *track* the system of (8.2), i.e., to find that C in $\mathscr{C}$, when it exists, which minimizes $\lim_{n\to\infty} \mathscr{E}([f_n - p_n]^2)$. This limit will be shown to exist.

Let us in what follows emphasize the dependence of the f_n on C by writing $f_n(C)$.

The problem just briefly described may be rephrased in a frequency domain setting. We denote by $\Gamma(\theta)$ the spectral density function (which we assume to exist) for the S.P. $(x_n)_0^\infty$. As in § 2, $\Gamma(\theta)$ is an a.e. defined nonnegative function integrable on the unit circle.

LEMMA 8.3. *Let $C \in \mathscr{C}$ be given. Assume the S.P.'s $(C_n)_0^\infty$ and $(f_n(C))_0^\infty$ satisfy* (8.1) *while the S.P. $(p_n)_0^\infty$ satisfies* (8.2). *Then $\lim_{n\to\infty} \mathscr{E}([f_n(C) - p_n]^2)$ exists, and denoting this limit by $\mathscr{E}^*(C)$, we have*

$$\mathscr{E}^*(C) = \frac{1}{2\pi}\int_0^{2\pi} \left| \frac{G(e^{i\theta})}{1 + C^E(e^{i\theta})G(e^{i\theta})} - H(e^{i\theta}) \right|^2 \Gamma(\theta)\, d\theta,$$

where $C^E \in l_1$ is the quasi-extension of C. (*See Definition* 4.1.)

Proof. We begin by showing that if $\{k_n\}_{n=-\infty}^{\infty}$ is any element of l_1^+ and $(m_n)_0^\infty$, the S.P., is given by $m_n = \sum_{j=0}^n k_{n-j}x_j$, then

$$\lim_{n\to\infty} \mathscr{E}([m_n]^2) = \frac{1}{2\pi}\int_0^{2\pi} |K(e^{i\theta})|^2 \Gamma(\theta)\, d\theta, \tag{8.4}$$

where $K(e^{i\theta}) = \sum_{n=0}^\infty k_n e^{in\theta}$.

To see this note that

$$\mathscr{E}([m_n]^2) = \sum_{j=0}^n \sum_{l=0}^n k_{n-j}k_{n-l}E(x_j, x_l).$$

But

$$\mathscr{E}(x_j, x_l) = \Gamma_{j-l} = \frac{1}{2\pi}\int_0^{2\pi} e^{-i(j-l)\theta}\Gamma(\theta)\, d\theta \qquad \text{for any } j, l \geqq 0.$$

Therefore,

$$\begin{aligned}
\mathscr{E}([m_n]^2) &= \sum_{j=0}^n \sum_{l=0}^n \left\{ \frac{1}{2\pi}\int_0^{2\pi} k_{n-j}e^{-ij\theta}k_{n-l}e^{il\theta}\Gamma(\theta)\, d\theta \right\} \\
&= \sum_{j=0}^n \sum_{l=0}^n \left\{ \frac{1}{2\pi}\int_0^{2\pi} k_{n-j}e^{i(n-j)\theta}k_{n-l}e^{-i(n-l)\theta}\Gamma(\theta)\, d\theta \right\} \\
&= \frac{1}{2\pi}\int_0^{2\pi} |K_n(e^{i\theta})|^2\Gamma(\theta)\, d\theta,
\end{aligned}$$

where $k_n(e^{i\theta}) = \sum_{j=0}^n k_j e^{ij\theta}$. Therefore letting $n \to \infty$ gives the result.

Now by a simple modification of Theorem 4.5 to handle S.P.'s as inputs and outputs it follows that $e_n = \sum_{j=0}^n d_{n-j}x_j$, where $\{d_k\}_{k=-\infty}^{\infty} \in l_1^+$ with $\{d_k\}_{k=-\infty}^{\infty} = Y^E = (I + C^E G)^{-1}$. Therefore since $f_n = \sum_{l=0}^n g_{n-l}e_l$ and $p_n = \sum_{l=0}^n k_{n-l}x_l$ we have that

$$f_n - p_n = \sum_{j=0}^n \left(\sum_{l=0}^n g_{n-l}d_{l-j} - k_{n-j} \right) x_j.$$

But clearly $G(I + C^E G)^{-1} - H$ is that element of l_1^+ given by $\{s_n\}_{n=0}^{\infty}$ with $s_n = \sum_{l=0}^{n} g_{n-l} d_l - k_n$ for each $n \geqq 0$. The conclusion of the lemma therefore follows by employing (8.4) with $(m_n)_0^\infty = (f_n - p_n)_0^\infty$.

Remark 8.5. For any integer $n \geqq 0$,

$$\inf_{C \in \mathscr{C}} \mathscr{E}^*(C) = \inf_{\substack{C(e^{i\theta}), \\ C \in \mathscr{W}_0}} \left\{ \frac{1}{2\pi} \int_0^{2\pi} \left| \frac{G(e^{i\theta})}{1 + C(e^{i\theta})G(e^{i\theta})} - H(e^{i\theta}) \right|^2 \Gamma(\theta)\, d\theta \right\}. \tag{8.6}$$

Proof. The proof is immediate.

Remark 8.7. In the case where an actual minimum is attained by the left-hand side of (8.6) at some $A \in \mathscr{C}$, then the right-hand side of (8.6) will also have a minimum and will realize it at the quasi-extension A^E of A, A^E lying in $\mathscr{W}_0$. Conversely, if a minimum of the right-hand side of (8.6) is attained at some $C \in \mathscr{W}_0$, then the left-hand side of (8.6) will attain a minimum and it will be at that unique C_* in $\mathscr{C}$ with $C_*^E = C$.

In fact, $C_* = \{c_k^*\}_{k=0}^{\infty}$, where

$$c_k^* = \frac{1}{2\pi i} \oint_{|z|=\varepsilon} z^{-(k+1)} C(z)\, dz$$

provided $\varepsilon > 0$ is sufficiently small (so that $|z| \leqq \varepsilon$ lies within the domain of analyticity of $C(z)$).

Utilizing the density theorem of § 4 we may now state the following theorem.

THEOREM 8.8. *For any $C \in \mathscr{C}$ assume that S.P.'s $(e_n)_0^\infty$ and $(f_n(C))_0^\infty$ satisfy* (8.1) *while S.P. $(p_n)_0^\infty$ satisfies* (8.2). *Then the following quantities are equal*:

(i) $\inf_{C \in \mathscr{C}} \mathscr{E}^*(C)$,

(ii) $\inf_{\substack{C(e^{i\theta}), \\ C \in \mathscr{W}_0}} \frac{1}{2\pi} \int_0^{2\pi} \left| \frac{G(e^{i\theta})}{1 + C(e^{i\theta})G(e^{i\theta})} - H(e^{i\theta}) \right|^2 \Gamma(\theta)\, d\theta$,

(iii) $\inf_{\substack{C(e^{i\theta}), \\ C \in \mathscr{W}}} \frac{1}{2\pi} \int_0^{2\pi} \left| \frac{G(e^{i\theta})}{1 + C(e^{i\theta})G(e^{i\theta})} - H(e^{i\theta}) \right|^2 \Gamma(\theta)\, d\theta$,

(iv) $\inf_{\substack{Y(e^{i\theta}), \\ Y \in T \cap l_1^+}} \frac{1}{2\pi} \int_0^{2\pi} |G(e^{i\theta})Y(e^{i\theta}) - H(e^{i\theta})|^2 \Gamma(\theta)\, d\theta$,

(v) $\min_{Y(e^{i\theta}) \in T'} \frac{1}{2\pi} \int_0^{2\pi} |G(e^{i\theta})Y(e^{i\theta}) - H(e^{i\theta})|^2 \Gamma(\theta)\, d\theta$,

where, as in § 6,

$$T' = \{ Y(e^{i\theta}) \in L_2^+(\Gamma) | [G(e^{i\theta})]^{-1} (Y(e^{i\theta}) - 1) \in L_2^+(\Gamma) \}.$$

Proof. By Remark 8.5, (i) = (ii). Corollary 5.8 states that $(I + \mathscr{W}_0 G)^{-1}$ and $(I + \mathscr{W} G)^{-1}$ are both dense in $T \cap l_1^+$ with respect to the l_1-induced topology. This is enough to ensure that (ii) = (iii) = (iv). Finally $\{ Y(e^{i\theta}) | Y \in T \cap l_1^+ \}$ is dense in T' with respect to the $L_2(\Gamma)$-induced topology as seen from the definition of $L_2(\Gamma)$. This shows that (iv) = (v), completing the proof.

We now easily show that if $\inf_{C\in\mathscr{C}} \mathscr{E}^*(C)$ is strictly positive (*imperfect tracking*), the function by $\Gamma(\theta)$ must necessarily be integrable on the unit circle.

THEOREM 8.9. *For any $C\in\mathscr{C}$ assume that S.P.'s $(e_n)_0^\infty$ and $(f_n(C))_0^\infty$ satisfy* (8.1) *while S.P. $(p_n)_0^\infty$ satisfies* (8.2). *Then if* $\inf_{C\in\mathscr{C}} \mathscr{E}^*(C) > 0$ (this expression is independent of n, $n \geqq 0$) *it follows that* $\int_0^{2\pi} \log \Gamma(\theta)\, d\theta > -\infty$.

Proof. By using Theorem 8.8 it follows under the hypothesis of this theorem that

$$\min_{Y(e^{i\theta})\in T'} \left\{ \frac{1}{2\pi} \int_0^{2\pi} |G(e^{i\theta})Y(e^{i\theta}) - H(e^{i\theta})|^2 \Gamma(\theta)\, d\theta \right\} > 0.$$

But this is exactly the condition needed in order to call Theorem 6.2 into play. By that theorem, $\int_0^{2\pi} \log \Gamma(\theta)\, d\theta > -\infty$.

We are mainly interested in the case $\inf \mathscr{E}^*(C) > 0$ (*imperfect tracking*). Without actually making this assumption we shall still assume integrability of $\log \Gamma(\theta)$.

Assumption and Remark 8.10. Following the indication of Theorem 8.9 we assume from this point on that $\int_0^{2\pi} \log \Gamma(\theta)\, d\theta > -\infty$. Equivalently, $\Gamma(\theta) = |R(e^{i\theta})|^2$ a.e., where $R(e^{i\theta})$ is an H_2-function with $R(z)$ nonzero on $|z| < 1$. We shall however additionally assume in the remainder of this section that $R(e^{i\theta})$ is the Fourier series for some $R \in l_1^+$ with $R(e^{i\theta}) \neq 0$ for $\theta \in [0, 2\pi]$. This assumption yields the following immediate consequences:

(i) $L_2(\Gamma) = L_2(d\theta)$, $L_2^+(\Gamma) = H_2$ and $T' = \{Y(e^{i\theta}) | Y \in T\}$.

(ii) $L_2(\Gamma)$ is isometric with $L_2(d\theta)$ and $L_2^+(\Gamma)$ is isometric with H_2 under the map $Y(e^{i\theta}) \to Y(e^{i\theta})R(e^{i\theta})$.

(iii) The set

$$\left\{ \frac{1}{2\pi} \int_0^{2\pi} |G(e^{i\theta})Y(e^{i\theta}) - H(e^{i\theta})|^2 \Gamma(\theta)\, d\theta \;\middle|\; Y(e^{i\theta}) \in T' \right\} \tag{8.11}$$

attains a minimum at some $Y^*(e^{i\theta})$ corresponding to $Y \in l_1^+$, i.e., the inf in (iv) of Theorem 8.8 may be changed to a min.

Note that (iii) follows directly from Remark 7.14(a) since we are assuming that $H(e^{i\theta})$ is the Fourier series for $H \in l_1^+$.

Remark 8.12. The assumptions in (8.10) are made in order to simplify certain details in assembling our results. The type of problem we are dealing with has one inherent untidiness as follows:

Minimization problems involving integrated quadratic forms are almost always taken over spaces of square integrable functions (Hilbert spaces). However, feedback relations involving convolution operators are most readily expressible for "l_1-operator algebras" (Banach algebras). More specifically, without the assumptions on $R(e^{i\theta})$ of (8.10) the element $Y_*(e^{i\theta})$ in T' minimizing (8.11) need not be the Fourier series of any Y_* in l_1^+. However, if this $Y_*(e^{i\theta})$ is to arise as $(1 + C(e^{i\theta})G(e^{i\theta}))^{-1}$ for some $C(e^{i\theta})$ related to $C \in \mathscr{W}$, then $Y_*(e^{i\theta})$ would have to lie in l_1^+.

To conclude this remark we point out that another way around this difficulty exists based on part (b) of Remark 7.14. To take this route would have involved

modification in our definition of M-stability, quasi-realizability, etc., so as to be able to handle L_2-filters.

We are now ready to state our main theorem.

THEOREM 8.13. *Let $(x_n)_0^\infty$ be a stationary stochastic process with a given known covariance sequence $(\Gamma_n)_{-\infty}^{\infty}$ and such that the corresponding spectral density function $\Gamma(\theta)$ satisfies the conditions of Assumption* 8.10. *For any $C \in \mathscr{C}$ let $(e_n(C))_0^\infty$ and $(f_n(C))_0^\infty$ denote S.P.'s satisfying feedback equations* (8.1) *and let $(p_n)_0^\infty$ be the S.P. satisfying* (8.2). *Let the polynomial $L(z)$ be given by the interpolation procedure*:

$$L(z) = ((G(z) - H(z))R(z)B_1^2(z))_{G^2},$$

where $B_1(z)$ and $(\cdot)_{G^2}$ are as defined in § 7. *Denoting $L(z) + R(z)H(z)B_1^2(z)$ by $D(z)$ for $|z| \leqq 1$ we find the following statements are true*:

(a) *If $D(e^{i\theta}) \neq 0$ for any $\theta \in [0, 2\pi]$ and if $D(0) \neq 0$, define*

$$C_{\#}(e^{i\theta}) = \frac{R(e^{i\theta})G(e^{i\theta})B_1^2(e^{i\theta}) - L(e^{i\theta}) - R(e^{i\theta})H(e^{i\theta})F_1^2(e^{i\theta})}{G(e^{i\theta})(L(e^{i\theta}) + R(e^{i\theta})H(e^{i\theta})B_1^2(e^{i\theta}))}. \tag{8.14}$$

Then $C_{\#}(e^{i\theta})$ so defined is the Fourier series of $C_{\#}$ in $\hat{\mathscr{W}}_0$ and

$$\frac{1}{2\pi}\int_0^{2\pi} |G(e^{i\theta})[1 + C_{\#}(e^{i\theta})G(e^{i\theta})]^{-1} - H(e^{i\theta})|^2\Gamma(\theta)\,d\theta$$

$$= \min_{\substack{C(e^{i\theta}),\\ C\in\hat{\mathscr{W}}_0}} \left\{\frac{1}{2\pi}\int_0^{2\pi} |G(e^{i\theta})[1 + C(e^{i\theta})G(e^{i\theta})]^{-1} - H(e^{i\theta})|^2\Gamma(\theta)\,d\theta\right\},$$

i.e., (ii) (*and hence* (iii)) *of Theroem* 8.8 *attains an actual minimum at this $C_{\#}$.*

Further let $C_* \in \mathscr{C}$ be given by $C_* = \{c_n^*\}_{n=0}^{\infty}$, where

$$c_n^* = \frac{1}{2\pi i}\int_{|z|=\varepsilon} z^{-(n+1)}C(z)\,dz$$

provided $\varepsilon > 0$ is sufficiently small (*see Remark* 8.7). *Then $\mathscr{E}^*(C_*) = \min_{C\in\mathscr{C}} E^*(C)$, i.e.*, (i) *of Theorem* 8.8 *attains an actual minimum at $C_* \in \mathscr{C}$.*

(b) *Under the hypotheses of the theorem it follows that*

$$\inf_{C\in\mathscr{C}} \mathscr{E}^*(C) = \sum_{k=1}^{\infty} |s_k|^2,$$

where

$$s_k = \frac{1}{2\pi i}\oint_{|z|=1} z^{k-1}\frac{[R(z)(G(z) - H(z))]}{[T(z)]^2}\,dz.$$

Here as in Theorem 7.18,

$$T(z) = \prod_{i=1}^{n}\left(\frac{z - a_i}{1 - \bar{a}_i z}\right)^{k_i},$$

the $\{a_i\}_{i=1}^n$, $\{k_i\}_{i=1}^n$ being the roots of $G(z)$ in $|z| < 1$ and their multiplicities, respectively. ($T(z)$ is a finite Blaschke product for $G(z)$; see [7. p. 66].)

Proof. (a) Let

$$Y_{\#}(e^{i\theta}) = \frac{L(e^{i\theta}) + B_1^2(e^{i\theta})H(e^{i\theta})R(e^{i\theta})}{R(e^{i\theta})G(e^{i\theta})B_1^2(e^{i\theta})} \qquad \text{for } \theta \in [0, 2\pi].$$

Then from Theorem 7.13, $Y_{\#}(e^{i\theta})$ lies in T' and $Y_{\#}(e^{i\theta})$ minimizes (v) of Theorem 8.8. Also by Remark 7.14(a), $Y_{\#}(e^{i\theta})$ is the Fourier series for some $Y_{\#} \in l_1^+$ and so $Y_{\#} \in T \wedge l_1^+$ and minimizes (iv) of Theorem 8.8.

Now assuming that $D(e^{i\theta}) \neq 0$ for $\theta \in [0, 2\pi]$ is equivalent to stating that $Y_{\#}(e^{i\theta}) \neq 0$ for $\theta \in [0, 2\pi]$ so that $Y_{\#}^{-1}$ exists in l_1^+. Then $C_{\#}(e^{i\theta})$ as given in (8.14) equals $[G(e^{i\theta})]^{-1}[1 - Y_{\#}(e^{i\theta})][Y_{\#}(e^{i\theta})]^{-1}$ so that $C_{\#} \in l_1$. Now $Y_{\#} = (I + C_{\#}G)^{-1}$ and the feedback triplet $\{G, C_{\#}; Y_{\#}\}$ is M-stable as $Y_{\#} \in T$. Therefore $C_{\#} \in \mathscr{W}$ and $C_{\#}$ clearly minimizes (iii) of Theorem 8.8.

If, additionally, $D(0) \neq 0$ or equivalently $Y_{\#}(0) \neq 0$, then Theorem 4.7 assures that $C_{\#} \in \mathscr{W}_0$ and part (a) of this theorem follows directly.

(b) The proof of (b) is immediate from Theorem 7.18 and the fact that (i) and (v) of Theorem 8.8 are equal.

Remark 8.15. In the cases where $D(e^{i\theta}) = 0$, some $\theta \in [0, 2\pi]$ or $D(0) = 0$, there is no $C \in \mathscr{C}$ minimizing $\mathscr{E}^*(C)$ but minimizing sequences can be constructed. These are mainly of theoretical interest and we omit the details.

Remark 8.16. Part (a) of Theorem 8.13 represents an explicit formula for the solution of our minimization via feedback problem (under the conditions $D(e^{i\theta}) \neq 0$ for $\theta \in [0, 2\pi]$ and $D(0) \neq 0$) and this together with the minimum mean square error formula (see Theorem 8.13 (b)) constitute our main results.

REFERENCES

[1] N. WIENER, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*, MIT Press, Cambridge, Mass., 1950.

[2] N. WIENER AND P. MASANI, *The prediction theory of multivariate stochastic processes, Parts I, II*, Acta Math., 98 (1957), pp. 111–150; 99 (1958), pp. 93–137.

[3] A. KOLMOGOROV, *Stationary sequences in Hilbert space*, Bull. Math. Univ. Moscow, 2 (1941), no. 6, 44 pp.

[4] R. E. KALMAN AND R. S. BUCY, *New results in linear filtering and prediction theory*, Trans. ASME Ser. D J. Basic Engrg., 83 (1961), pp. 95–108.

[5] W. M. WONHAM, *Random differential equations in control theory*, Probabilistic Methods in Applied Math, vol. II, A. T. Bharucha-Reid, ed., Academic Press, New York, 1969.

[6] L. H. LOOMIS, *An Introduction to Abstract Harmonic Analysis*, Van Nostrand, New York, 1953.

[7] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, N.J., 1962.

[8] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[9] P. J. DAVIS, *Interpolation and Extrapolation*, Blaisdell, New York, 1963.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

GEOMETRIC THEORY OF TIME-OPTIMAL CONTROL*

OTOMAR HÁJEK†

Abstract. Properties of the minimal-time function are considered for the problem of reaching the origin within (linear autonomous finite-dimensional) control systems. The function's continuity is first established, leading to the construction of optimal feedback controls for normal systems. For systems where the control matrix has maximal rank (= dimension of state space), the minimal-time function is locally Lipschitzian, and hence differentiable almost everywhere.

1. Introduction. In this paper we will consider the control system

$$\dot{x} = Ax + Bu, \tag{S}$$

where A, B are constant $n \times n$ and $n \times m$ matrices respectively, and the controls u are restricted by $|u_i| \leqq 1$ (where $u' = (u_1, \cdots, u_m)$; primes denote transposition). We will be concerned with the problem of reaching 0 from $x \in R^n$ in minimal time, i.e., of finding (time-)optimal solutions.

This section merely reviews terminology and notation. In § 2 we prove a fundamental theorem on continuity of the minimal-time function. Most of our results are more or less distant outgrowths of this; among the latter belongs a new characterization of normality, Corollary 7. In § 3 the principal result is Theorem 10, existence of optimal feedback controls for normal systems. A number of intriguing questions connected with this result (see the remarks following Theorem 10) are still unresolved.

The last two sections are really appendices. Section 4 emphasizes the connection between control theory and dynamical system theory. In § 5 we treat a rather special class of control systems, for which quite direct methods yield much information concerning the minimal-time function.

Geometric aspects of such problems, for the far more difficult nonlinear case, were treated in [5]; among other things, there is an apparent relationship between our Proposition 14 and Theorem 6.3 in [5] (or rather its proof: note $\varepsilon < \eta$, p. 321).

We will use the terminology and notation of [2] where available, and of [3] where not. In particular, the reachable set at time $t \geqq 0$ is

$$\mathscr{R}(t) = \left\{ \int_0^t e^{-As} Bu(s)\, ds : |u_i(s)| \leqq 1, u \in \mathscr{L}_1[0, t] \right\}$$

(a not particularly happy choice of term: the set of points reached from the origin in time t is actually $e^{At}\mathscr{R}(t)$); and the reachable set is

$$\mathscr{R} = \bigcup \{ \mathscr{R}(t) : t \geqq 0 \}.$$

It is known that $\mathscr{R}(t)$ is compact, convex, symmetric about 0 and has $\mathscr{R}(s) \subset \mathscr{R}(t)$ whenever $0 \leqq s \leqq t$ [2, p. 46]. It is easily shown that $\mathscr{R}(t)$ satisfies the

* Received by the editors August 4, 1970, and in final revised form, December 15, 1970.

† Department of Mathematics and Statistics, Case Western Reserve University, Cleveland, Ohio 44106. This work was supported in part by the National Science Foundation under Grant GP-22689.

stricter monotonicity condition

$$\mathscr{R}(s) \subset \operatorname{Int} \mathscr{R}(t) \quad \text{if } 0 \leqq s < t \tag{1}$$

if and only if $\mathscr{R}$ is open, and also if and only if (S) is controllable [2, p. 72 ff.].

Without the restriction $|u_i| \leqq 1$ we obtain the controllability space

$$\mathscr{C}(t) = \left\{ \int_0^t e^{-As} B u(s)\, ds : u \in \mathscr{L}[0, t] \right\},$$

$$\mathscr{C} = \bigcup \{ \mathscr{C}(t) : t \geqq 0 \};$$

$\mathscr{C}(t)$ and $\mathscr{C}$ are linear subspaces of R^n, whereupon, easily, $\mathscr{C}(t) = \mathscr{C}$ for $t > 0$ [3, p. 97ff.].

One (possibly) new concept will be useful, that of the $(n + 1)$-dimensional reachability set

$$\mathscr{R}^* = \{(x, t) \in R^{n+1} : x \in \mathscr{R}(t)\}.$$

Its connection with the sets $\mathscr{R}(t)$ is obvious from

$$\mathscr{R}^* \cap (R^n \times \{t\}) = \mathscr{R}(t) \times \{t\};$$

in particular, $\mathscr{R}^*$ is closed (with compact "slabs"), symmetric about the t-axis and contains it; usually $\mathscr{R}^*$ is not convex.

It is immediate that $x \in \mathscr{R}(t)$ (equivalently, $-x \in \mathscr{R}(t)$) if and only if x can be steered to 0 in time t by some admissible controls. The minimal-time function T is defined [3, p. 145] by

$$T(x) = \inf \{ t \geqq 0 : x \in \mathscr{R}(t) \}; \tag{2}$$

thus $0 \leqq T \leqq +\infty$ with $T(x) < +\infty$ if and only if $x \in \mathscr{R}$.

If grad $T(x)$ exists, then

$$\max \{(-\operatorname{grad} T(x), Ax + Bu) : u \in R^m, |u_i| \leqq 1\} = 1,$$

and, if T has a total differential in an open set, then there

$$f(x) = -\operatorname{sgn}(B' \operatorname{grad} T(x))$$

is an optimal feedback control (see [3, p. 146], modulo notational changes). It is precisely this unwarranted ex post assumption of differentiability that we wish to avoid in the present paper.

2. The minimal-time function.

THEOREM 1. *The minimal-time function $T : \mathscr{R} \to R^1$ is continuous, with $\mathscr{R}$ open in the linear space $\mathscr{C} \subset R^n$.*

Proof. Directly from (2) and the closedness of the sets $\mathscr{R}(t)$, T is lower semicontinuous. For controllable systems, upper semicontinuity follows from (1). For the remaining case, use Kálman's decomposition [3, p. 99]: up to a linear equivalence, (S) may be written in the form

$$\dot{x}_1 = A_{11} x_1 + A_{12} x_2 + B_1 u,$$

$$\dot{x}_2 = \qquad\quad A_{22} x_2 \qquad\quad ;$$

furthermore, the system

$$(S_1) \qquad \dot{y} = A_{11}y + B_1u$$

(corresponding to $x_2 \equiv 0$) is controllable. Obviously $\mathscr{R}(t) = \mathscr{R}_1(t) \times \{0\}$, so that

$$T\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{cases} T_1(x_1) & \text{if } x_2 = 0, \\ +\infty & \text{if } x_2 \neq 0. \end{cases}$$

Since (S_1) is controllable, T_1 is continuous, and hence so is T; T_1 is finite precisely on $\mathscr{R}_1$, an open subset of $\mathscr{C}_1 = \mathscr{C}$ [2, p. 79].

COROLLARY 2. *If $x \in \partial\mathscr{R}$, the boundary relative to $\mathscr{C}$, then*

$$\lim_{y \to x} T(y) = +\infty.$$

Thus $T : \mathscr{C} \to R^1 \cup \{+\infty\}$ is also continuous.

Proof. Assume $x_p \to x$, $T(x_p) \leqq t$. Then $x_p \in \mathscr{R}(t)$, and hence $x \in \mathscr{R}(t) \subset \mathscr{R}$ from closedness.

COROLLARY 3. *For all $t \geqq 0$ we have*

$$\{x : T(x) \leqq t\} = \mathscr{R}(t), \qquad \{x : T(x) = t\} = \partial\mathscr{R}(t)$$

(boundary relative to $\mathscr{C}$).

Proof. As concerns the first formula, obviously

$$\mathscr{R}(t) \subset \{x : T(x) \leqq t\} \subset \bigcap_{p=1}^{\infty} \mathscr{R}\left(t + \frac{1}{p}\right);$$

thus it is sufficient to show that the last set is contained in the first. Let x be in the intersection; thus there are admissible controls $u_p : [0, t + 1/p] \to R^m$ such that

$$x = \int_0^{t+1/p} e^{-As} B u_p(s)\, ds = \int_0^t + \int_t^{t+1/p} = x_p + y_p.$$

Here $x_p \in \mathscr{R}(t)$, and the remainder terms $y_p \to 0$ since all coordinates of the u_p's are bounded by 1. Thus $x_p \to x$, so that x is the closed set $\mathscr{R}(t)$.

For the second formula note first that

$$\{x : T(x) < t\} \subset \operatorname{Int} \mathscr{R}(t)$$

from the continuity of T; and we need only verify the converse inclusion for $t > 0$. If $x \in \operatorname{Int} \mathscr{R}(t)$, then $x \in \operatorname{Int} \mathscr{R}(t - \varepsilon)$ for small $\varepsilon > 0$ [2, p. 47], and hence $T(x) \leqq t - \varepsilon < t$.

Remark 1. Consider the boundary $\partial\mathscr{R}^*$ of the $(n + 1)$-dimensional reachable set, relative to $\mathscr{C} \times R^1$. Then Theorem 1 shows that any line in R^{n+1} parallel to the t-axis intersects $\partial\mathscr{R}^*$ at most once. Thus the set $\mathscr{R}^*$, though often not convex, nevertheless exhibits a form of convexity in the t-direction.

Remark 2. The assertion of Theorem 1 may be invalid for time-dependent systems, or for target sets other than the origin.

Remark 3. The assertion of Corollary 2 is false if $\mathscr{C}$ is replaced by R^n. In point of fact, $T : R^n \to R^1 \cup \{+\infty\}$ is continuous if and only if (S) is controllable (i.e.,

$\mathscr{C} = R^n$). For if (S) is not controllable, one may take $x_p \to x$ with $x \in \mathscr{R}$, $x_p \notin \mathscr{C} \supset \mathscr{R}$, whereupon $T(x_p) = +\infty > T(x)$.

Remark 4. From Corollary 3 it follows that the minimal-time function completely determines the performance of system (S) (interpreting "performance" as the correspondence to $t \mapsto \mathscr{R}(t)$). In particular, the sets $\{x : T(x) \leqq t\}$ are strictly convex if and only if (S) is normal [2, p. 65].

In the next theorem we will need the following approximation result (this is probably well-known, easily proved by standard methods, and amenable to generalization).

LEMMA 4. *For every* $\varepsilon > 0$ *and* $t \geqq 0$ *there exists* $\delta > 0$ *such that*

$$\operatorname{dist}(\mathscr{R}_1(t), \mathscr{R}(t)) < \varepsilon,$$

where $\mathscr{R}_1(t)$ *is the reachability set of any system*

$$\dot{x} = A_1 x + B_1 u \tag{S_1}$$

with $|A - A_1| < \delta, |B - B_1| < \delta$.

THEOREM 5. *Given a sequence of systems*

$$\dot{x} = A_p x + B_p u \tag{S_p}$$

with $A_p \to A$, $B_p \to B$, *then their minimum-time functions* T_p *converge to* T *in the following sense*:

$$T_p(x_p) \to T(x) \quad \text{whenever } x_p \to x, \quad x_p \in \mathscr{R} \cup \mathscr{R}_p, \quad x \in \mathscr{R}.$$

In particular, $T_p \to T$ *uniformly on compact subsets of* $\mathscr{R}$.

Proof. Assume that, on the contrary, $T_p(x_p)$ is bounded away from $T(x_0)$ for some subsequence of the p's. Thus there are two cases: either

$$T_p(x_p) < t_0 < T(x_0) \tag{3}$$

for some t_0 and subsequence, or similarly with opposite inequalities. In the first case we have $x_p \in \mathscr{R}_p(t_0)$; from Lemma 4, $y_p \in \mathscr{R}(t_0)$ for some y_p with $y_p - x_p \to 0$. But then $T(y_p) \leqq t_0$ and $y_p \to x_0$; since $t_0 < T(x_0)$, this contradicts Theorem 1.

In the second case we have

$$T(x_0) < t_0 < T_p(x_p). \tag{4}$$

Then $x_0 \in \operatorname{Int} \mathscr{R}(t_0)$, the interior relative to $\mathscr{C}$. From Lemma 4, there exist $y_p \in \mathscr{R}_p(t_0)$ with $y_p \to x_0$. Since $x_p \notin \mathscr{R}(t_0)$ by assumption (4), on the segment $x_p y_p$ (or $x_0 y_p$, if $y_p \notin \mathscr{R}(t_0)$) there is a point $z_p \in \partial\mathscr{R}(t_0)$; necessarily, $z_p \to x_0$. Let c_p be a unit exterior normal to a supporting hyperplane at z_p to $\mathscr{R}_p(t_0)$. Then

$$(c_p, x - z_p) \leqq 0 \qquad \text{for all } x \in \mathscr{R}_p(t_0). \tag{5}$$

Choose a convergent subsequence $c_p \to c \neq 0$. For any $w \in \mathscr{R}(t_0)$ there exist $w_p \in \mathscr{R}_p(t_0)$ with $w_p \to w$; taking limits in (5) with $x = w_p$, we conclude that

$$(c, w - x_0) \leqq 0 \qquad \text{for all } w \in \mathscr{R}(t_0).$$

However, this contradicts $x_0 \in \operatorname{Int} \mathscr{R}(t_0)$, and completes the proof.

Remark 5. In the situation of this theorem we cannot assert that

$$T_p(x_p) \to T(x) \quad \text{whenever } x_p \to x.$$

Actually this holds if and only if (S) is controllable. Indeed, if (S) is not controllable, take $(S_p) \equiv (S)$, and proceed as in Remark 2 above. On the other hand, if (S) is controllable, then $x_p \to x \in \mathscr{R}$ implies $x_p \in \mathscr{R}$ ultimately.

Remark 6. Theorem 1 is, of course, a special case of Theorem 5; however, the former was needed in the proof of the latter.

Remark 7. Theorem 5 admits this pessimistic interpretation: by changing the system slightly one cannot improve performance too much.

PROPOSITION 6. *In the situation of Theorem* 5, *let the* y_p *be optimal solutions of* (S_p), *and let* $y_p(0) \to x_0 \in \mathscr{R}$. *Then some subsequence converges to an optimal solution* y *of* (S), *uniformly on compact subsets of* $[0, T(x_0))$.

Proof. We have

$$y_p(0) \to x_0, \quad y_p(\theta_p) = 0 \qquad \text{for } \theta_p = T_p(y_p(0)).$$

From Theorem 5, $\theta_p \to \theta = T(x_0)$. Using weak sequential compactness of the space of admissible controls, we conclude that $\{y_p\}$ has a subsequence $\{y_q\}$ converging, uniformly on compact subsets of $[0, \theta)$, to a solution y of (S) through x_0. Since the y_q were optimal,

$$T_q(y_q(t)) = T_q(y_q(0)) - t \quad \text{for } 0 \leqq t \leqq \theta_q;$$

thus, from Theorem 5 again,

$$T(y(t)) = T(x_0) - t \quad \text{for } 0 \leqq t < \theta. \tag{6}$$

In particular, y is optimal for (S), at least on $[0, \theta)$. Finally, $\lim_{t \to \theta -} y(t) = 0$ follows from (6): $y(t) \in \mathscr{R}(\theta - t)$.

COROLLARY 7. (S) *is normal if and only if it satisfies the following condition. Whenever the* (S_p) *are as in Theorem* 5, *and the* y_p *are optimal solutions of* (S_p) *through* $y_p(0) \to x_0 \in \mathscr{R}$, *necessarily the sequence* $\{y_p\}$ *converges (whereupon the limit is an optimal solution of* (S) *and convergence is uniform on compact subsets of* $[0, T(x_0))$).

3. Optimal feedback controls. We will study an auxiliary concept associated with a given system (S). For every $x \in \mathscr{R}$ let $E(x)$ denote the set of all unit exterior normals of supporting hyperplanes at x to $\mathscr{R}(\theta)$, where $\theta = T(x)$ (so that $x \in \partial\mathscr{R}(\theta)$, see Corollary 3).

LEMMA 8. $E(x)$ *is a nonvoid compact subset of the* $(n-1)$*-sphere* S^{n-1}. *Furthermore,*

$$\begin{gathered} E(-x) = -E(x), \\ \limsup E(y) \subset E(x) \quad \text{as } y \to x \text{ in } \mathscr{R}. \end{gathered} \tag{7}$$

If (S) *is controllable, then* $E(x)$ *contains antipodal vectors only for* $x = 0$.

Proof. The first assertion actually concerns existence of supporting hyperplanes. For the second use symmetry of $\mathscr{R}(\theta)$ about 0. The third follows easily from the continuity of T and a compactness argument. For the last observe that (S) is proper, so that $\operatorname{Int} \mathscr{R}(\theta) \neq \varnothing$ if $\theta > 0$, i.e., if $0 \neq x \in \mathscr{R}$.

It follows that the set of all the exterior normals,

$$\{tc : t > 0, c \in E(x)\},$$

is a wedge in R^n, and actually a proper cone if (S) is controllable and $x \neq 0$.

Since $E(x)$ is nonvoid, for each x one can choose $c(x) \in E(x)$. Obviously a continuous selection need not be possible (unless, e.g., the $\mathscr{R}(t)$ have no corners). However, one can at least make a measurable selection.

LEMMA 9. *There exists a measurable function* $c : \mathscr{R} \to R^n$ *such that*

$$c(x) \in E(x) \quad \textit{for all } x \in \mathscr{R}.$$

Proof. This paraphrases part of the proof of Filippov's theorem [2, p. 31]. For each $x \in \mathscr{R}$ choose $c(x) \in E(x)$ with minimal first coordinate $c_1(x)$ (apply compactness of $E(x)$, from Lemma 8); if there is more than one such element in $E(x)$, choose that whose second coordinate $c_2(x)$ is also minimal; and so on. Directly from (7), $c_1(\cdot)$ is upper semicontinuous on $\mathscr{R}$. By Lusin's theorem, for every $\varepsilon > 0$ there exists a subset $M_1 \subset \mathscr{R}$ on which $c_1(\cdot)$ is continuous, and has meas $\mathscr{R} - M_1 < \varepsilon/n$. Again from (7), $c_2(\cdot)$ is upper semicontinuous on M_1, so there is a subset $M_2 \subset M_1$ on which both $c_1(\cdot)$, $c_2(\cdot)$ are continuous, and meas $M_1 - M_2 \subset \varepsilon/n$. Continuing in this fashion we find a set $M \subset \mathscr{R}$ on which all coordinates of $c_k(\cdot)$ are continuous, and meas $\mathscr{R} - M < \varepsilon$. Since $\varepsilon > 0$ was arbitrary, c is measurable (this is the trivial converse to Lusin's theorem).

THEOREM 10. *If* (S) *is normal, there exists a measurable function* $f : \mathscr{R} \to R^m$ *which is an optimal feedback control for* (S) *in the following sense: In addition to* (S), *consider the autonomous differential equation*

$$\dot{y} = Ay + Bf(y). \tag{F}$$

Then each optimal solution of (S) *is a solution of* (F) *in* $\mathscr{R}$; *as a partial converse, each solution of* (F) *is a solution (possibly not optimal) of* (S).

Proof. We will first construct f and then verify the assertion on the solutions.

Let $c : \mathscr{R} \to R^n$ be as described in Lemma 9. For each $x \in \mathscr{R}$ set

$$f(x) = \lim_{s \to 0+} -\operatorname{sgn}(B'e^{-A's}c(x)). \tag{8}$$

Since each coordinate of $B'e^{-A's}c(x)$ is analytic in s (for fixed x), $\operatorname{sgn}(B'e^{-A's}c(x))$ is piecewise constant in s, so that the limit exists. Finally, f is measurable since c is such. (Remark: the "obvious" choice would be $f(x) = -\operatorname{sgn}(B'c(x))$; the reason for the extra factor and limit is technical, and will become apparent in the course of the proof.)

Now let $x : [0, \theta] \to R^n$ be an optimal solution of (S) through $x(0) = x_0 \in \mathscr{R}$, $\theta = T(x_0)$; we wish to show that simultaneously it satisfies (F). Take arbitrarily s, $0 \leqq s < \theta$; there is an optimal control through $x(s) \in \mathscr{R}$ at time s, which may be taken in the form

$$u_s(t) = -\operatorname{sgn}(B'e^{-A'(t-s)}c), \qquad s \leqq t \leqq \theta,$$

for any choice of c in $E(x(s))$ (i.e., $-c \in E(-x(s))$, see [2, p. 51]); we choose $c = c(x(s))$. Since (S) is normal, the response to u_0 (u_s for $s = 0$) is precisely x; furthermore, on $[s, \theta]$, u_0 is again an optimal control for $x(s)$ at initial time s (see (6)). Again from normality, optimal controls are essentially unique; thus, for each individual s and almost all $t \geqq s$ we have

$$u_0(t) = u_s(t) = -\operatorname{sgn}(B'e^{-A'(t-s)}c(x(s))).$$

Now take limits $t \to s+$ over these t's and obtain (see (8))

$$u_0(s+) = f(x(s)) \quad \text{for } s \in [0, \theta).$$

Since u_0 is piecewise constant, we have $u_0(s) = f(x(s))$ for almost all s; therefore the response x to u_0 satisfies

$$\dot{x}(s) = Ax(s) + Bu(s) = Ax(s) + Bf(x(s))$$

almost everywhere. As x is absolutely continuous, it is a solution of (F).

Finally, that solutions of (F) solve (S) is almost immediate. Indeed, any solution y of (F) is the response of (S) to $v(t) = f(y(t))$, and obviously this is an admissible control (see (8) for the bound, and also Lemma 9 for measurability). This completes the proof of Theorem 10.

Remark 8. This raises further questions.

It would be desirable to know whether there is a complete converse in Theorem 10, more precisely, whether all solutions of (F) are optimal solutions of (S), or equivalently, whether (F) has uniqueness of solutions to the initial value problem into positive time.

In the positive case the converse assertion may even hold without normality; a proof might proceed via Proposition 6 and an approximation of (S) by normal systems.

4. Semidynamical systems. Assume that (S) is normal. Then with every point $x \in \mathscr{R}$ we may associate an optimal solution $y:[0, +\infty) \to R^n$ of (S) through $y(0) = x$, whereupon y is determined uniquely (let $y(t) = 0$ identically for $t \geqq T(x)$). Thus there is a well-defined mapping $\pi:\mathscr{R} \times [0, +\infty) \to \mathscr{R}$,

$$\pi(x, t) = y(t) \quad \text{for } t \geqq 0.$$

It is obvious that

$$\pi(x, 0) = x, \qquad \pi(\pi(x, t), s) = \pi(x, t + s);$$

and it is easily shown that π is continuous. Therefore [1, p. 12] we have the following theorem.

THEOREM 11. *If* (S) *is normal, then* π *is a global semidynamical system on* $\mathscr{R}$.

Note that even in the most reasonable cases, π need not have negative uniqueness (i.e., solutions are not determined uniquely by initial data in the negative time direction); see, e.g., the situation indicated in [3, p. 10].

An examination of the relation between (S) and π is intriguing. Thus, 0 is the only critical point of π, and it is asymptotically stable, with $\mathscr{R}$ as its region of attraction. Furthermore, the minimum-time function of (S) (or rather $-T$) is a Lyapunov function for π, strictly decreasing along trajectories outside 0, and with $T(x) = 0$ if and only if $x = 0$, $T(x) \to +\infty$ as x approaches the boundary of $\mathscr{R}$ (Corollary 2). Finally, Theorem 10 can be reformulated in dynamical terms thus: the semidynamical system π admits a differential representation.

We conclude this section with an application of dynamical systems to control theory. The question treated may be indicated as follows: If x is any point on $\partial\mathscr{R}(t)$ and $t \geqq s \geqq 0$, then obviously an optimal solution through x can be followed down till it meets $\mathscr{R}(s)$. The converse assertion, concerning backing out of $\mathscr{R}(s)$ to reach $\partial\mathscr{R}(t)$, though plausible, does not have an obvious proof.

PROPOSITION 12. *Every optimal solution on* $[0, +\infty)$ *can be extended to an optimal solution defined on* R^1.

Proof. First consider the special case that (S) is normal, so that semidynamical system theory may be applied. Consider any optimal solution $x: I \to R^n$ which is inextensible. According to [1, p. 32], there are three possibilities. Either $I = R^1$, and there is nothing to prove; or

$$I = [\alpha, +\infty), \quad \alpha > -\infty, \quad x(\alpha) \text{ is a start point}; \tag{9}$$

or finally,

$$I = (\alpha, +\infty), \alpha > -\infty, \quad \text{and} \quad x(t) \text{ has no cluster points in } \mathscr{R} \text{ as } t \to \alpha+. \tag{10}$$

We proceed to eliminate the latter two alternatives. Since (S) is controllable, $\mathscr{R}$ is open in R^n, and hence π is a semidynamical system on an n-manifold. According to [1, p. 124], π has no start points; thus (9) is impossible.

Finally, consider case (10). Since the right-hand side of (S) admits a linear estimate $|A|\,|x| + |B|$, we cannot have $x(t) \to \infty$ as $t \to \alpha+$ (this is a minor modification of a standard result, see, e.g., [4, Theorem 3, Chap. 1]). Thus there is $x(t_p) \to y$ for some $t_p \to \alpha+$, $y \in \bar{\mathscr{R}} - \mathscr{R}$. But then, from (6),

$$T(x(t_p)) = T(x(0)) - t_p \leqq T(x_0) - \alpha < +\infty,$$

contradicting Corollary 2. This completes the proof in the case of normal (S).

Next, consider the general case, and let $x: [0, +\infty) \to R^n$ be an optimal solution through $x(0) = y_0 \in \mathscr{R}$. Since normal systems are generic (see [3, p. 100], and recall that finite intersections of open dense sets are again such), there exist normal systems

$$\dot{x} = A_p x + B_p u, \qquad A_p \to A, \quad B_p \to B. \tag{S_p}$$

From Lemma 4, there exist $y_p \to y_0$ with $y_p \in \mathscr{R}_p$. We have just shown that (S_p) has an optimal solution $x_p: R^1 \to R^n$ through y_p. Now, some subsequence of the x_p's converges uniformly on compact subsets of $(-\infty, 0]$. Indeed, if the u_p are corresponding controls, then a subsequence converges weakly on compact subsets of $(-\infty, 0]$; the limit is also an admissible control, and the $B_p u_p$ converge weakly to Bu. Denote the limit by x again. Since x_p is optimal,

$$T_p(x_p(t)) = T_p(y_p) - t \qquad \text{for } t \leqq 0.$$

Take limits, and apply Theorem 5:

$$T(x(t)) = T(y_0) - t \qquad \text{for } t \leqq 0;$$

thus x is indeed an optimal solution of (S). This concludes the proof.

5. Systems with uncoupled controls. In this section only we make the following overall convention.

CONVENTION 13. *The control matrix* B *has* rank $B = n$.

A system is such if and only if it is linearly equivalent to a system with control matrix I, after possible omission of ineffective or duplicated controls; i.e., linearly

equivalent to a system with uncoupled controls

$$\dot{x}_i = \sum_j a_{ij}x_j + u_i, \qquad 1 \leqq i \leqq n.$$

Note that automatically such systems are controllable, without any assumption on A (but not necessarily normal, e.g., if the minimal polynomial of A has degree $< n$).

PROPOSITION 14. *The $(n+1)$-dimensional reachable set $\mathscr{R}^*$ contains a small cone with vertex at 0. More precisely, there exist $\lambda > 0$, $\varepsilon > 0$, such that*

$$x \in \mathscr{R}(\lambda|x|) \quad \textit{whenever } |x| \leqq \varepsilon.$$

Furthermore, λ may be taken arbitrarily close to

$$1/\min\left\{\sum_i |b_i'c| : c \in R^n, |c| = 1\right\},$$

where $B = (b_1, \cdots, b_m)$.

Proof. First fix any $\lambda > 0$, and assume an $x \in R^n$ with $x \notin \mathscr{R}(\lambda|x|)$. Since each $\mathscr{R}(t)$ contains 0, some point y on the segment $0x$ is on the boundary of $\mathscr{R}(\lambda|x|)$; and, of course, $|y| \leqq |x| > 0$.

Corresponding to y, there is a control, to reach y in time $\lambda|x|$, of a special type [2, p. 51]: for some $c \in R^n$ with $|c| = 1$, if $u(t) = \operatorname{sgn}(B'e^{-A't}c)$,

$$y = \int_0^{\lambda|x|} e^{-As}Bu(s)\,ds.$$

It follows that

$$(c, y) = \int_0^{\lambda|x|} \sum_{i=1}^m |(B'e^{-A's}c)_i|\,ds.$$

The assumptions on magnitudes yield $(c, y) \leqq |x|$; thus

$$|x| \geqq \lambda|x| \cdot \mu(\lambda|x|), \qquad x \neq 0, \tag{11}$$

where $\mu(\cdot)$ is defined by

$$\mu(t) = \inf\left\{\sum_{i=1}^m |(B'e^{-A's}c)_i| : |c| = 1, 0 \leqq s \leqq t\right\}.$$

Evidently μ is nonincreasing. To show that it is strictly positive, assume that

$$\sum_{i=1}^m |(B'e^{-A's_p}c_p)_i| \to 0 \quad \text{as } p \to \infty;$$

we may take $s_p \to s$, $c_p \to c$, $|c| = 1$. The limit relation shows that $c \in R^n$ is perpendicular to all columns of $e^{-As}B$, contradicting

$$\operatorname{rank} e^{-As}B = \operatorname{rank} B = n, \quad c \neq 0.$$

Now choose any $\varepsilon = \delta/\lambda$, $\delta > 0$. Then, for all x with $|x| \leqq \varepsilon$, we have $\mu(\lambda|x|) \geqq \mu(\delta) > 0$, and (11) yields $\lambda \leqq 1/\mu(\delta)$. In other words, for $\lambda > 1/\mu(\delta)$, necessarily no $x \in R^n$ with $|x| \leqq \varepsilon$ can have $x \notin \mathscr{R}(\lambda|x|)$.

Evidently, we may take λ arbitrarily close to $1/\mu(0)$, at the cost of decreasing ε.

Remark 9. For Proposition 14 above, constancy of A, B may be weakened to: A is summable, B is continuous at 0, rank $B(0) = n$. However, subsequent results will require autonomy.

Remark 10. Obvious estimates yield that $\mathscr{R}^*$ is also contained within a cone with vertex at 0 (locally at 0). This suggests that the numerical value of the actual opening, of $\mathscr{R}^*$ at 0, might be a significant qualitative measure of performance of the system.

Remark 11. Consideration of the case $A = 0$ suggests that our rank condition is also necessary; but this will not be needed here. The condition cannot be replaced by normality alone. Indeed, for the two-dimensional system

$$\dot{x} = y, \quad \dot{y} = u,$$

the three-dimensional reachable set intersects the (x, t)-plane in $\{(x, t) : |x| \leqq \frac{1}{4}t^2\}$; thus $\mathscr{R}$ contains no small cone with vertex at 0, and the minimal-time function is not locally Lipschitzian at 0 (cf. Theorem 17). (Actually, $T(x, y) = y + \sqrt{2y^2 + 4x}$.)

Remark 12. It will be seen from subsequent results that good upper estimates of λ are of considerable interest (while ε is unimportant). For the one-dimensional system

$$\dot{x} = x + u, \tag{12}$$

$\mathscr{R}(t)$ is the segment $[-(1 - e^{-t}), (1 - e^{-t})]$; thus the largest opening is 1. In this case also $\min\{\cdots\} = 1$, so that our estimate of λ is exact. Somewhat more generally, it can be shown that, if B is square nonsingular, then λ may be taken arbitrarily close to $|B^{-1}|$.

Throughout this section, λ and ε will be taken as described in Proposition 14.

COROLLARY 15. *For any $x \in \mathscr{R}(t)$ and y with $|y - x| \leqq \varepsilon \exp - |A|t$ we have*

$$y \in \mathscr{R}(t + |y - x|\lambda \exp |A|t).$$

Thus $\mathscr{R}^*$ contains, with every x, a small cone with vertex at x and axis parallel to the t-axis.

Proof. This follows from Proposition 14 and the following elementary observation: if $x \in \mathscr{R}(t)$, $z \in \mathscr{R}(s)$, then

$$x + e^{-At}z \in \mathscr{R}(t + s).$$

Next we turn to the study of the minimal-time function T.

THEOREM 16. *For all $x, y \in \mathscr{R}$,*

$$|T(x) - T(y)| \leqq |x - y|\lambda \exp(|A| \max(T(x), T(y))). \tag{13}$$

Consequently $T : \mathscr{R} \to R^1$ is locally Lipschitzian, and the Lipschitz constant near x may be taken arbitrarily close to $\lambda \exp |A|T(x)$. Finally, grad $T(x)$ *exists almost everywhere in $\mathscr{R}$.*

Proof. We first show that, for each $x \in \mathscr{R}$,

$$T(y) - T(x) \leqq |y - x|\lambda \exp |A|T(x) \quad \text{if } |y - x| < \varepsilon \exp(-|A|T(x)).$$

Indeed, choose any y as indicated, and then $t > T(x)$ with $|y - x| < \varepsilon \cdot \exp(-|A|t)$. Then $x \in \mathscr{R}(t)$, and, from Corollary 15,

$$T(y) \leqq t + |y - x| \lambda \exp |A| t;$$

finally, let $t \to T(x)$.

In proving (13) we may assume that $T(y) < T(x)$. Since $\{x : T(x) \leqq t\}$ is convex, the entire segment P from y to x has

$$T(z) \leqq T(x) \qquad \text{for } z \in P.$$

Now apply the previous estimate to a finite decomposition of P into segments of length $< \varepsilon \exp(-|A|T(x))$, noting that the Lipschitz constants are all $\leqq \lambda \cdot \exp |A| T(x)$.

Finally, according to [6, p. 311, Theorem (14.2) (ii)] a Lipschitzian function has total differential almost everywhere.

PROPOSITION 17. *$\mathscr{R}$ contains the ball*

$$U = \{x \in R^n : |x| < 1/(\lambda |A|)\}$$

(with $U = R^n$ if $A = 0$); in U,

$$T(x) \leqq \frac{1}{|A|} \log \frac{1}{1 - \lambda |A| \, |x|}, \tag{14}$$

so that, within any ball of radius ξ in U,

$$|T(x) - T(y)| \leqq |x - y| \frac{\lambda}{1 - \lambda |A| \xi}.$$

Proof. We obtain estimate (14) first. From Theorem 16, the function S,

$$S(x) = \exp(-|A| T(x)) \quad \text{for } x \in \mathscr{R},$$

is Lipschitzian, with Lipschitz constant $\lambda |A|$ (first locally only). Thus

$$|S(x) - S(0)| \leqq \lambda |A| \, |x|,$$

and the estimate follows.

To show that T is defined throughout U it is sufficient (since U is connected and $\mathscr{R}$ open) to prove that $\mathscr{R} \cap U$ is closed in U. Thus, let $x_p \to x$, $x_p \in \mathscr{R}$, $x \in U$. From the estimate, the $T(x_p)$ are bounded, say by t. Thus $x_p \in \mathscr{R}(t)$ and hence $x \in \mathscr{R}(t) \subset \mathscr{R}$.

Remark 13. For the example $\dot{x} = x + u$ treated earlier, the estimates of both $\mathscr{R}$ and T are sharp.

REFERENCES

[1] N. P. BHATIA AND O. HÁJEK, *Local Semi-Dynamical Systems*, Lecture Notes in Mathematics 90, Springer, Berlin, 1969.

[2] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[3] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[4] V. V. NIEMYCKIĬ AND V. V. STEPANOV, *Qualitative Theory of Differential Equations*, 2nd ed., Gostechizdat, Moscow, 1947.

[5] E. ROXIN, *A geometric interpretation of Pontryagin's maximum principle*, Nonlinear Differential Equations and Nonlinear Mechanics, J. P. LaSalle and S. Lefschetz, eds., Academic Press, New York, 1963, pp. 303–324.

[6] S. SAKS, *Theory of the Integral*, Monografie Matematyczne VII, Warszawa, 1937. (English transl.)

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

ANOTHER PROOF OF THE LIAPUNOV CONVEXITY THEOREM*

JAMES A. YORKE†

Abstract. A new proof of the Liapunov convexity theorem is presented.

1. Introduction. Let x be in R^n and let $A(t)$ be a continuous $n \times n$ matrix. Let $b:[0, T] \to R^n$ be continuous. LaSalle's well-known bang-bang control theorem [5] for linear systems states that if $X(t)$ is the fundamental matrix (with $X(0) = I$, the identity matrix) for $x'(t) = A(t)x$, then the "reachable set at time T" for

$$y'(t) = A(t)y(t) + b(t)u(t), \qquad y(0) = 0 \tag{1}$$

is compact and convex and the reachable set does not change if we restrict ourselves to bang-bang controls. That is:

Write $E = [0, T]$. Let $\mathcal{F}$ be the set of integrable functions $u:E \to [-1, +1]$, and let $\mathcal{F}^*$ be $\{u \in \mathcal{F} : |u(t)| = 1 \text{ for almost all } t\}$. If we let $f(t) = X(T)X^{-1}(t)b(t)$, then for $u \in \mathcal{F}$, the solution $y(t)$ at time T satisfies $y(T) = \int_E f(t)u(t)\,dt$. Define

$$\mathcal{I}u = \int_E f(t)u(t)\,dt, \tag{2}$$

and for any set $\mathcal{S}$ of functions $u(\cdot)$ write $\mathcal{I}\mathcal{S} = \{\mathcal{I}u : u \in \mathcal{S}\}$. Then $\mathcal{I}\mathcal{F}^*$ is the set of reachable points using bang-bang controls.

THEOREM. *$\mathcal{I}\mathcal{F}^*$ is compact and convex and equals $\mathcal{I}\mathcal{F}$.*

Many proofs and extensions of this result have been given. The result (actually a minor variation) was first proved by A. M. Liapunov [3] in 1940. The shortest proof was given by Lindenstrauss [1] who used the Krein–Milman theorem. See Remark 2. Other proofs appear in [2], [4], [6]. The referee has pointed out that Olech used an induction argument in a way similar to that used here in Case 2. See [9, p. 89] and [10, p. 42].

2. A proof of the theorem. For a convex set $U \subset R^n$, the *dimension d of U* (dim U) may be defined to be the dimension of the smallest hyperplane H that contains U. For $U \subset H \subset R^n$, define $\partial_H U = \overline{H - U} \cap \overline{U}$.

LEMMA. *Let $U \subset H$ be convex and let H be a hyperplane with $p \in \partial_H U$. Assume the dimension of U equals the dimension of H. Then there is a linear functional $h:R^n \to R$ which is not constant on H and satisfies* $\sup_{x \in U} h(x) = h(p)$.

See, for example, [7], for the result that such an $h:H \to R$ exists. The domain of h can then easily be extended to R^n.

* Received by the editors September 8, 1970, and in revised form December 4, 1970.

† Institute for Fluid Dynamics and Applied Mathematics, University of Maryland, College Park, Maryland 20740. This research was supported in part by a Columbia University undergraduate scholarship, by the National Science Foundation with a Graduate Fellowship and under Grants GP-9347 and GP-17601.

Proof of Theorem. Let J be a measurable subset of E and let $\phi : J \to \{-1, +1\}$ be measurable. We will say then that (ϕ, J) is a *restriction.* Let

$$\mathscr{F}(\phi, J) = \{u \in \mathscr{F} : u(t) = \phi(t) \text{ for almost all } t \in J\},$$

$$\mathscr{F}^*(\phi, J) = \{u \in \mathscr{F}^* : u(t) = \phi(t) \text{ for almost all } t \in J\}.$$

We actually will prove the following stronger result.

PROPOSITION. *Let (ϕ, J) be a restriction. Then $\mathscr{I}\mathscr{F}(\phi, J) = \mathscr{I}\mathscr{F}^*(\phi, J)$.*

The proposition reduces to the theorem when J is allowed to be trivial (i.e., have measure 0). Note that $\mathscr{F}(\phi, J)$ is a convex closed bounded subset of $L^2(E; R^n)$ so $\mathscr{F}(\phi, J)$ is weakly compact and its continuous image $\mathscr{I}\mathscr{F}(\phi, J)$ is therefore convex and compact. ($\mathscr{I}$ is continuous and linear so it is weakly continuous.)

Choose some restriction (ϕ, J) and write $\mathscr{F}(\phi, J) = \mathscr{F}_0$ and $\mathscr{F}^*(\phi, J) = \mathscr{F}_0^*$. Since $\mathscr{F}_0^* \subset \mathscr{F}_0$, it suffices to prove that $\mathscr{I}\mathscr{F}_0^* \supset \mathscr{I}\mathscr{F}_0$. Choose $p \in \mathscr{I}\mathscr{F}_0$. We now prove $p \in \mathscr{I}\mathscr{F}_0^*$. Let d be the dimension of $\mathscr{I}\mathscr{F}_0$. Then $\mathscr{I}\mathscr{F}_0$ lies in a d-dimensional hyperplane H. If $d = 0$, $\mathscr{I}\mathscr{F}_0$ contains only one point, so $\{p\} = \mathscr{I}\mathscr{F}_0^*$ and the result is true. We now prove the proposition by induction on d.

Assumption. The proposition is true if $\dim \mathscr{I}\mathscr{F}(\phi, J) < d$.

Case 1. Suppose $p \in \partial_H \mathscr{I}\mathscr{F}_0$. Letting $U = \mathscr{I}\mathscr{F}_0$, choose h to be the linear function in the lemma. Let $J_0 = \{t \in E - J : h(f(t)) \neq 0\}$ and $J_1 = J_0 \cup J$. Let

$$\phi_1(t) = \begin{cases} \phi(t) & \text{for } t \in J, \\ \operatorname{sgn} h(f(t)) & \text{for } t \in J_0. \end{cases} \tag{3}$$

Write $\mathscr{F}_1 = \mathscr{F}(\phi_1, J_1)$ and $\mathscr{F}_1^* = \mathscr{F}^*(\phi_1, J_1)$. For $u \in \mathscr{F}_0$,

$$\begin{aligned} h(\mathscr{I}u) &= \int_E u(t)h(f(t)) = \int_{J_1} u(t)h(f(t)) \\ &= \int_J \phi(t)h(f(t)) + \int_{J_0} u(t)h(f(t)) \\ &\leqq \int_J \phi(t)h(f(t)) + \int_{J_0} |h(f(t))| = \int_E \phi_1(t)h(f(t)). \end{aligned} \tag{4}$$

It may be seen that we have equality in (4) for $u \in \mathscr{F}_0$ if and only if $u(t) = \operatorname{sgn} h(f(t))$ for almost all $t \in J_0$. Since $h(p) = \sup_{u \in \mathscr{F}_0} h(\mathscr{I}u)$, we have $h(p) \geqq h(\mathscr{I}u)$ for all $u \in \mathscr{F}_0$. For $u \in \mathscr{F}_0$, $h(p) = h(\mathscr{I}u)$ if and only if $u \in \mathscr{F}_1$. Since h is nonconstant on H, the set $H_p = \{x \in H : h(x) = h(p)\}$ has dimension $d - 1$. Since $\mathscr{I}\mathscr{F}_1 \subset H_p$, $\dim \mathscr{I}\mathscr{F}_1 < d$. Since $p \in \mathscr{I}\mathscr{F}_1$ we may use the assumption and conclude $p \in \mathscr{I}\mathscr{F}_1^*$, which proves the result because $\mathscr{I}\mathscr{F}_1^* \subset \mathscr{I}\mathscr{F}_0^*$.

Case 2. Suppose $p \in \operatorname{int}_H \mathscr{I}\mathscr{F}_0$, where H is the smallest hyperplane containing $\mathscr{I}\mathscr{F}_0$. We now show that by extending the definition of ϕ to ϕ_τ which is defined on the larger set $J \cup [0, \tau]$, the set $\mathscr{I}\mathscr{F}(\phi_\tau, J^\tau)$ decreases continuously as τ is increased until for some σ, $p \in \partial_H \mathscr{I}\mathscr{F}(\phi_\sigma, J^\sigma)$, reducing the problem to Case 1. For $\tau \in E$ let J^τ denote $J \cup [0, \tau]$ and define ϕ_τ on J^τ by $\phi_\tau(t) = \phi(t)$ for $t \in J$ and $\phi_\tau(t) = +1$ for $t \in [0, \tau] - J$. Consider the restriction (ϕ_τ, J^τ) and write $\mathscr{F}_\tau$ for $\mathscr{F}(\phi_\tau, J^\tau)$. If $\tau_1 > \tau_2$, then $\mathscr{F}_{\tau_1} \subset \mathscr{F}_{\tau_2} \subset \mathscr{F}_0$. $\mathscr{I}\mathscr{F}_\tau$ is of course a convex set and we may let σ ($\leqq T$) be the supremum of $\tau \in E$ for which $p \in \mathscr{I}\mathscr{F}_\tau$. We claim $p \in \partial_H \mathscr{I}\mathscr{F}_\sigma$.

If $\tau_1 > \tau_2$ and $p \in \mathscr{I}\mathscr{F}_{\tau_2}$ and, say $p = \mathscr{I}u_p$ for some $u_p \in \mathscr{F}_{\tau_2}$, then there is $u_1 \in \mathscr{F}_{\tau_1}$ such that $u_1 = u_p$ except perhaps on $[\tau_2, \tau_1]$; hence $\mathscr{I}u_1 \in \mathscr{I}\mathscr{F}_{\tau_1}$ and if δ is the distance of p from $\mathscr{I}\mathscr{F}_{\tau_1}$,

$$\delta \leqq |\mathscr{I}u_1 - p| \leqq \int_{\tau_2}^{\tau_1} |f(s)|\,|u_1(s) - u_p(s)|\,ds \leqq 2|\tau_1 - \tau_2|M,$$

where $M = \sup_E |f|$. From this "continuity" of $\mathscr{I}\mathscr{F}_\tau$, $p \in \partial_H \mathscr{I}\mathscr{F}_\sigma$. By considering the restriction (ϕ_σ, J^σ) instead of (ϕ, J), we have reduced the situation to Case 1. The proof of the theorem is complete since the induction argument for dimension d is complete.

Remark 1. It is known that the theorem is true if $f(s)\,ds$ is replaced by a nonatomic vector measure but is false for more general measures. This proof uses the fact that $f(s)\,ds$ is a nonatomic measure in Case 2 since otherwise $\mathscr{I}\mathscr{F}(\phi_\tau, J^\tau)$ does not have to be a continuous set function of τ (in the Hausdorff metric).

Remark 2. This proof is similar to Lindenstrauss' proof in certain ways. It differs primarily in Case 2 where we have avoided the Krein–Milman theorem.

Acknowledgment. The author would like to thank H. Hermes and L. Markus for their helpful comments.

REFERENCES

[1] JORAM LINDENSTRAUSS, *A short proof of Liapunoff's convexity theorem*, J. Math. Mech., 15 (1966), pp. 971–972.

[2] H. HALKIN, *Some further generalizations of a theorem of Lyapunov*, Arch. Rational Mech. Anal., 17 (1964), pp. 272–277.

[3] A. LIAPOUNOFF, *Sur les fonctions-vecteurs complètement additives*, Bull. Acad. Sci. URSS Sér. Math., 4 (1940), pp. 465–478.

[4] P. R. HALMOS, *The range of a vector measure*, Bull. Amer. Math. Soc., 54 (1948), pp. 416–421.

[5] J. P. LASALLE, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. 5, Princeton Univ. Press, Princeton, N.J., 1959, pp. 1–24.

[6] C. OLECH, *On the range of an unbounded vector-valued measure*, Math. Systems Theory, 2 (1968), pp. 251–256.

[7] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control*, John Wiley, New York, 1967.

[8] J. ÉTIENNE, *Sur une démonstration du bang-bang principle*, Bull. Soc. Roy. Sci. Liège, 37 (1968), no. 11–12, pp. 551–556.

[9] C. OLECH, *Extremal solutions of a control system*, J. Differential Equations, 2 (1966), pp. 74–101.

[10] A. BALAKRISHNAN AND LUCIEN W. NEUSTADT, eds., *Mathematical Theory of Control*, Academic Press, New York, 1967.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

ON THE SOLUTIONS OF A STOCHASTIC CONTROL SYSTEM*

TYRONE DUNCAN† AND PRAVIN VARAIYA‡

Abstract. The control system considered in this paper is modeled by the stochastic differential equation

$$dx(t, \omega) = f(t, x(\cdot, \omega), u(t, \omega))\, dt + dB(t, \omega),$$

where B is n-dimensional Brownian motion, and the control u is a nonanticipative functional of $x(\cdot, \omega)$ taking its values in a fixed set U. Under various conditions on f it is shown that for every admissible control a solution is defined whose law is absolutely continuous with respect to the Wiener measure μ, and the corresponding set of densities on the space C forms a strongly closed, convex subset of $L^1(C, \mu)$. Applications of this result to optimal control and two-person, zero-sum differential games are noted. Finally, an example is given which shows that in the case where only some of the components of x are observed, the set of attainable densities is not weakly closed in $L^1(C, \mu)$.

1. Introduction and contents. A stochastic control problem is defined by the specification of the stochastic differential equation which models the system dynamics, the information available to the controller and the corresponding set of admissible control laws, and the cost incurred by each control law. Of theoretical interest is the "existence" problem, which means determining in terms of the above three defining characteristics a class of control problems for which there exist control laws achieving minimum cost. Published results ([1], [2], [3], see especially the excellent survey article [4] of Fleming) differ from one another and are not usually comparable because either the models are different or the set of admissible control laws is different.

There are two basic steps involved in obtaining an existence result. The first step involves determining conditions which guarantee that a solution of the stochastic differential equation is defined for every admissible control law. The next step involves the search for a topology under which the set of solutions (or an equally good substitute) is compact, and the cost function is lower semicontinuous. Thus, for instance, Fleming and Nisio [1] consider stochastic differential equations of the form

$$dx(t) = f(t, x(\cdot))\, u(t)\, dt + \sigma(t, x(\cdot), B(\cdot))\, dB(t), \qquad 0 \leqq t < \infty,$$

where $u(t)$ is any process taking values in the unit cube, and independent of future increments $B(t_2) - B(t_1)$, $t \leqq t_1 \leqq t_2$, of the Brownian motion B. Various conditions on f, σ are imposed to guarantee a solution for every admissible control. It is then shown that the set of laws of all the solutions of the differential equation

* Received by the editors August 25, 1970, and in revised form January 11, 1971.

† Computer Information and Control Engineering Department, University of Michigan, Ann Arbor, Michigan 48104. The work of this author was done while he was visiting the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley, and was supported by the Office of Naval Research under Contract N0014-67-A-081-0021, NR041-376.

‡ Department of Electrical Engineering and Computer Sciences, Electronics Research Laboratory, University of California, Berkeley, California 94720. The work of this author was supported by the U.S. Army Research Office—Durham under Contract DAHC04-67-0046.

corresponding to the different control laws is compact in the Prohorov metric. Beneš [3] considers stochastic differential equations of the form

$$dx(t,\omega) = f(t, x(\cdot,\omega), u(t,\omega))\,dt + dB(t,\omega), \qquad 0 \leqq t \leqq 1, \tag{1}$$

where f is measurable with respect to its arguments and continuous in u. The control law is any nonanticipative, measurable functional $u(t,\omega) = \psi(t, x(\cdot,\omega))$ which takes values in a compact set U. He assumes that f satisfies a linear growth condition

$$|f(t, x(\cdot,\omega), u)|^2 \leqq K(1 + \|x(\cdot,\omega)\|^2),$$

where $\|x(\cdot,\omega)\| = \sup\{|x(t,\omega)| \quad |0 \leqq t \leqq 1\}$.

The existence of a solution to (1) for every control law is guaranteed by a result of Girsanov [5] (see Corollary 3 below). The resulting law is absolutely continuous with respect to the Wiener measure μ on the space C of all continuous functions from $[0, 1]$ into R^n. Beneš shows that if $f(t, x(\cdot,\omega), U)$ is convex for every $t \in [0, 1]$ and $x(\cdot,\omega) \in C$, then the set of densities corresponding to all the admissible control laws is a convex and strongly closed (hence weakly compact) subset of $L^1(C, \mu)$.

In this paper, we show that the above result holds if the linear growth condition is replaced by the growth condition

$$|f(t, x(\cdot,\omega), u)| \leqq f_0(\|x(\cdot,\omega)\|), \tag{2}$$

where $f_0 : R \to R$ is increasing, and the condition

$$\int_C \exp\left[\int_0^1 \langle f(t, B, u), dB(t)\rangle - \frac{1}{2}\int_0^1 |f(t, B, u)|^2\,dt\right]\mu(dB) = 1 \tag{3}$$

for every admissible control law. An example is given to show that (2) does not imply (3). The linear growth condition implies (3) (see Corollary 3). Condition (3) also follows from (2), if the drift term f in (1) has a delay (see Corollary 4). Finally we show that in the important case where the control is allowed to depend only on some components of the state x, the set of densities is not always weakly closed in $L^1(C, \mu)$.

In § 2 we give some preliminary results and definitions, and in § 3 we present the main result on weak compactness of the attainable densities. In § 4 we give conditions which guarantee (3), in § 5 we present applications to optimal control and stochastic differential games, and in the final section we present the negative example for the problem with partial observations.

2. Preliminaries. In the main, we adopt the notations and definitions of Beneš [3]. Consider the stochastic differential equation

$$\begin{aligned} dx(t) &= f(t, x, u(t, x))\,dt + dB(t), \qquad 0 \leqq t \leqq 1, \\ x(0) &= 0, \end{aligned} \tag{1'}$$

where $B(t)$ is a standard n-dimensional Brownian motion process with continuous sample paths, $x(t)$ is the state of the system and $u(t, x)$ is the control law which takes values in a compact subset U of R^m. To state the precise conditions which f, u must satisfy we need the following definition.

DEFINITION 1. (a) Let C be the Banach space of all continuous functions $z:[0, 1] \to R^n$ with norm $\|z\| = \max\{|z(t)| \mid 0 \leqq t \leqq 1\}$, where $|y|$ is the Euclidean norm of $y \in R^n$.

(b) For each $t \in [0, 1]$ let $\mathscr{S}_t$ be the smallest σ-field of subsets of C which contains all sets of the form $\{z \mid z(\tau) \in A\}$, where $\tau \in [0, t]$ and A is a Borel subset of R^n.

(c) Let $\mathscr{S} = \mathscr{S}_1$.

We shall define the solution of (1) in such a way that the sample paths of x are continuous (and have no explosions), so that f is a map from $[0, 1] \times C \times U \to R^n$. We impose throughout the following conditions on f.

C1. f is measurable with respect to the product σ-algebra $\mathscr{B} \otimes \mathscr{S} \otimes \mathscr{B}_U$, where $\mathscr{B}$ ($\mathscr{B}_U$) is the set of Borel measurable subsets of $[0, 1]$ (U).

C2. For fixed $t \in [0, 1)$, $f(t, \cdot, \cdot)$ is measurable with respect to the product σ-algebra $\mathscr{S}_t \otimes \mathscr{B}_U$.

C3. For fixed $(t, z) \in [0, 1] \times C$, $f(t, z, \cdot)$ is continuous on U.

C4. There exists an increasing function $f_0 : R \to R$ such that $|f(t, z, u)| \leqq f_0(\|z\|)$ for all (t, z, u).

C5. $f(t, z, U)$ is closed and convex for every (t, z).

DEFINITION 2. (a) An *admissible control* (*law*) is any map $u:[0, 1] \times C \to U$ which is measurable with respect to $\mathscr{B} \otimes \mathscr{S}$, and for each fixed $t \in [0, 1]$, $u(t, \cdot)$ is measurable with respect to $\mathscr{S}_t$. Let $\mathscr{U}$ be the set of all admissible control laws.

(b) For each $u \in \mathscr{U}$, the *drift corresponding to* u is the function $g = g_u:[0, 1] \times C \to R^n$ defined by

$$g(t, z) = f(t, z, u(t, z)).$$

Let $\mathscr{G} = \{g_u \mid u \in \mathscr{U}\}$.

(c) For $g \in \mathscr{G}$ and $N \geqq 0$, let $g^N:[0, 1] \times C \to R^n$ be defined by

$$g^N(t, z) = \begin{cases} g(t, z) & \text{if } |z(\tau)| \leqq N \quad \text{for } \tau \leqq t, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathscr{G}^N = \{g^N \mid g \in \mathscr{G}\}$.

DEFINITION 3. A function $\psi:[0, 1] \times C \to R^n$ will be said to be *causal* if it is $\mathscr{B} \otimes \mathscr{S}$ measurable, and if for each fixed $t \in [0, 1]$, $\psi(t, \cdot)$ is measurable with respect to $\mathscr{S}_t$.

From [6, Lemmas 1, 2] we can obtain the following useful characterization of $\mathscr{G}$. Condition C3 is needed only for Lemma 1. The reader should be warned that the proof of the "only if" part of Lemma 1 involves a nontrivial synthesis problem (lemma of Fillipov).

LEMMA 1. *A causal function* $g:[0, 1] \times C \to R^n$ *belongs to* $\mathscr{G}$ *if and only if* $g(t, z) \in f(t, z, U)$ *for all* (t, z).

It will prove convenient to work with sets larger than $\mathscr{G}$, $\mathscr{G}^N$.

DEFINITION 4. Let Φ be the set of all causal maps $\phi:[0, 1] \times C \to R^n$ such that $|\phi(t, z)| \leqq f_0(\|z\|)$ for all (t, z). Let $\Phi^N = \{\phi \mid \phi \in \Phi, |\phi(t, z)| \leqq N \text{ for all } (t, z)\}$.

Throughout the rest of this paper let Ω be a fixed space and let $\mathscr{A}_t$, $0 \leqq t \leqq 1$, be a fixed, increasing family of σ-fields of subsets of Ω. Let $\mathscr{A} = \mathscr{A}_1$. We say that $z(t)$ or $z(t, \omega)$, $0 \leqq t \leqq 1$, is a family of n-dimensional random variables on $(\Omega, \mathscr{A}_t)$ if for each t, $z(t, \cdot)$ is a map from Ω into R^n which is measurable with respect to $\mathscr{A}_t$.

We shall need to consider various probability measures on $\mathscr{A}$. If $z(t)$, $0 \leqq t \leqq 1$, is a family of n-dimensional random variables on $(\Omega, \mathscr{A}_t)$ and we wish to consider the stochastic process generated by $z(t)$ corresponding to a particular probability measure P on $\mathscr{A}$, we will say that $z(t)$, $0 \leqq t \leqq 1$, is an n-dimensional stochastic process on $(\Omega, \mathscr{A}_t, P)$. Finally let P_0 be a distinguished probability measure on $\mathscr{A}$, and let $x(t, \omega)$, $0 \leqq t \leqq 1$, be a fixed n-dimensional, Brownian motion process on $(\Omega, \mathscr{A}_t, P_0)$ with almost all sample paths $x(\cdot, \omega) \in C$. We assume that the σ-fields $\mathscr{A}_t$ are complete with respect to P_0.

DEFINITION 5. Let $\psi : [0, 1] \times C \to R^n$ be a causal function such that

$$\int_0^1 |\psi(t, z)|^2 \, dt < \infty \quad \text{for all } z \in C. \tag{4}$$

Then $\zeta^t(\psi)$, $0 \leqq t \leqq 1$, is the stochastic process on $(\Omega, \mathscr{A}_t, P_0)$ with continuous sample paths, defined by

$$\zeta^t(\psi) = \int_0^t \langle \psi(\tau, x), dx(\tau) \rangle - \frac{1}{2} \int_0^t |\psi(\tau, x)|^2 \, d\tau. \tag{5}$$

For convenience, let $\zeta(\psi) = \zeta^1(\psi)$. (In (5), the first integral is to be interpreted as an Ito stochastic integral.)

The results of this section are immediate consequences of the work of Girsanov [5].

THEOREM 1 (Existence). *Let $\psi : [0, 1] \times C \to R^n$ be a causal function such that* (4) *holds.*

(i) *Then,*

$$\int_\Omega \exp [\zeta(\psi)] P_0(d\omega) \leqq 1.$$

(ii) *Suppose*

$$\int_\Omega \exp [\zeta(\psi)] P_0(d\omega) = 1, \tag{6}$$

and define the probability measure P_ψ on $\mathscr{A}$ by

$$P_\psi(A) = \int_A \exp [\zeta(\psi)] P_0(d\omega), \qquad A \in \mathscr{A}.$$

Then the stochastic process $B(t)$ defined on $(\Omega, \mathscr{A}_t, P_\psi)$ by

$$B(t, \omega) = x(t, \omega) - \int_0^t \psi(\tau, x(\cdot, \omega)) \, d\tau, \qquad 0 \leqq t \leqq 1,$$

is a Brownian motion.

(iii) *If ψ is bounded, then* (6) *holds.*

Proof. Parts (i), (ii) and (iii) are immediate consequences of Lemma 2, Theorem 1, and Lemma 1, respectively, of [5].

Theorem 1 immediately gives us a sufficient condition for the existence of a solution to (1). For, let $u \in \mathcal{U}$, and let g be the drift corresponding to u. If

$$\int_{\Omega} \exp[\zeta(g)]P_0(d\omega) = 1,$$

then the stochastic process $x(t)$ on $(\Omega, \mathcal{A}_t, P_g)$ satisfies the equation

$$x(t) = \int_0^t g(\tau, x)\,d\tau + \text{Brownian motion}.$$

LEMMA 2. *Let $\psi \in \Phi$. Let $y(t)$, $0 \leqq t \leqq 1$, be a stochastic process on $(\Omega, \mathcal{A}_t, P)$ with continuous sample paths, such that the stochastic process $B(t)$ on $(\Omega, \mathcal{A}_t, P)$ defined by*

$$B(t) = y(t) - \int_0^t \psi(\tau, y)\,d\tau, \qquad 0 \leqq t \leqq 1, \tag{7}$$

is a Brownian motion. Then, the measure ν induced by y on $(C, \mathcal{S})$ is mutually absolutely continuous with respect to the Wiener measure μ, and

$$\frac{d\mu}{d\nu}(y) = \exp\left[-\int_0^1 \langle \psi(t, y), dB(t)\rangle - \frac{1}{2}\int_0^1 |\psi(t, y)|^2\,dt\right]. \tag{8}$$

Proof. Since $|\psi(\cdot, z)| \leqq f_0(\|z\|)$, it follows from Lemma 7 of [5], that the measure μ on $(C, \mathcal{S})$ defined by

$$\mu(S) = \int_S \exp\left[-\int_0^1 \langle \psi, dB\rangle - \frac{1}{2}\int_0^1 |\psi|^2\,dt\right] d\nu$$

coincides with the Wiener measure. It is easy to see that

$$\exp\left[-\int_0^1 \langle \psi, dB\rangle - \frac{1}{2}\int_0^1 |\psi|^2\,dt\right] > 0,$$

ν-almost everywhere. The result follows.

COROLLARY 1. *Let $\psi \in \Phi$, and let $y(t)$, $0 \leqq t \leqq 1$, satisfy the hypothesis of Lemma 2. Then*

$$\int_C \exp\left[\int_0^1 \langle \psi(t, z), dz(t)\rangle - \frac{1}{2}\int_0^1 |\psi(t, z)|^2\,dt\right]\mu(dz) = 1. \tag{9}$$

COROLLARY 2. *Let $\psi \in \Phi$, and let $y(t)$, $0 \leqq t \leqq 1$, satisfy the hypothesis of Lemma 2. Then the measure ν on $(C, \mathcal{S})$ induced by y is uniquely specified by ψ and is given by*

$$\nu(S) = \int_S \exp\left[\int_0^1 \langle \psi(t, z), dz(t)\rangle - \frac{1}{2}\int_0^1 |\psi(t, z)|^2\,dt\right]\mu(dz). \tag{10}$$

Proof. The corollaries follow from (8) and the identity $dB = dy - \psi(t, y)\,dt$.

3. Main results.

DEFINITION 6. For any subset $\Sigma \subset \Phi$, let $\mathcal{D}(\Sigma)$ be the subset of $L^1(\Omega, \mathcal{A}, P_0)$ defined by

$$\mathcal{D}(\Sigma) = \{\exp \zeta(\phi) | \phi \in \Sigma\}.$$

PROPOSITION 1. *$\mathscr{D}(\Phi^N)$ is a bounded subset of $L^2(\Omega, \mathscr{A}, P_0)$.*

Proof. If $\phi \in \Phi^N$, then by definition, $|\phi| \leqq N$. By Lemma 1 of [5], it follows that

$$\int_\Omega \exp 2\zeta^t(\phi) P_0(d\omega) \leqq \exp tN^2.$$

For the rest of this paper let E_0 denote expectation with respect to the probability measure P_0. Also if $\gamma \in L^1(\Omega, \mathscr{A}, P_0)$, then $E_0(\gamma|\mathscr{A}_t)$ denotes the conditional expectation of γ with respect to $\mathscr{A}_t$.

The proofs of Lemmas 3 and 4 are simple modifications of the proofs of Theorems 4 and 3, respectively, of [3]. They are presented here for completeness and because we shall need to refer to parts of the proofs later.

LEMMA 3. *$\mathscr{D}(\Phi^N)$ is a closed subset of $L^2(\Omega, \mathscr{A}, P_0)$.*

Proof. Let ϕ_n, $n = 1, 2, 3, \cdots$, be a sequence from Φ^N and let ρ be such that

$$\lim_{n\to\infty} E_0|\rho - \exp \zeta(\phi_n)|^2 = 0 \tag{11}$$

and

$$\lim_{n\to\infty} \exp \zeta(\phi_n) = \rho \quad \text{a.s. } P_0. \tag{12}$$

First of all $\rho > 0$ a.s. P_0. Because, let $A = \{\omega|\rho(\omega) = 0\}$. Then from (12),

$$\lim_{n\to\infty} \zeta(\phi_n)(\omega) = -\infty \quad \text{for } \omega \in A. \tag{13}$$

Also,

$$\zeta(\phi_n) = \int_0^1 \langle \phi_n(s, x), dx(s)\rangle - \frac{1}{2}\int_0^1 |\phi_n(s, x)|^2\, ds$$

and $|\phi_n| \leqq N$, so that from (13),

$$\lim_{n\to\infty} \int_0^1 \langle \phi_n(s, x), dx(s)\rangle = -\infty \quad \text{on } A.$$

But

$$E_0\left[\int_0^1 \langle \phi_n, dx(s)\rangle\right]^2 = E_0 \int_0^1 |\phi_n(s)|^2\, ds \leqq N^2,$$

so that $P_0(A) = 0$. By Ito's representation [7], there is a causal map

$$\psi : [0, 1] \times C \to R^n$$

with

$$\int_0^1 |\psi(t, z)|^2\, dt < \infty$$

for z in C, such that

$$\rho = 1 + \int_0^1 \langle \psi(t, x), dx(t)\rangle \quad \text{a.s. } P_0.$$

Let

$$\rho(t) = E_0(\rho|\mathscr{A}_t) = 1 + \int_0^t \langle \psi(s, x), dx(s)\rangle.$$

Then, by Jensen's inequality,

$$\int_0^1 E_0|\rho(t) - \exp \zeta^t(\phi_n)|^2\, dt \leqq \int_0^1 E_0|\rho - \exp \zeta(\phi_n)|^2\, dt,$$

which converges to zero, so that taking subsequences if necessary we can assume that

$$\rho(t) = \lim_{n\to\infty} \exp \zeta^t(\phi_n) \quad \text{a.s. } l \otimes P_0, \tag{14}$$

where l denotes Lebesgue measure on $[0, 1]$. Next, by Ito's differentiation rule,

$$\exp \zeta(\phi_n) = 1 + \int_0^1 \exp \zeta^t(\phi_n)\langle \phi_n(t), dx(t)\rangle \quad \text{a.s. } P_0$$

so that

$$E_0 \int_0^1 |\exp \zeta^t(\phi_n)\phi_n(t) - \psi(t)|^2\, dt = E_0|\exp \zeta(\phi_n) - \rho|^2$$

converges to zero, and therefore, taking subsequences if necessary, we can assume that

$$\psi(t) = \lim_{n\to\infty} \exp \zeta^t(\phi_n)\phi_n(t) \quad \text{a.s. } l \otimes P_0.$$

Since $\rho(t) > 0$ a.s. P_0, we see using (14) that

$$\psi(t)/\rho(t) = \lim_{n\to\infty} \phi_n(t) \quad \text{a.s. } l \otimes P_0. \tag{15}$$

It follows that there is a causal map $\phi : [0, 1] \times C \to R^n$,

$$\phi(t, x(\cdot, \omega)) = \lim_{n\to\infty} \phi_n(t, x(\cdot, \omega)) \quad \text{a.s. } l \times P_0$$

and

$$\rho(t) = 1 + \int_0^t \rho(s)\langle \phi(s, x), dx(s)\rangle.$$

From Ito's differential rule we see that

$$d(\log \rho(t)) = \langle \phi(t), dx(t)\rangle - \tfrac{1}{2}|\phi(t)|^2\, dt,$$

and hence,

$$\rho = \exp \zeta(\phi).$$

Because of (15) we can assume that $|\phi| \leqq N$, so that the lemma is proved.

LEMMA 4. *$\mathscr{D}(\Phi^N)$ is a convex set.*

Proof. Let $\phi_i \in \Phi^N$, $\lambda_i \geqq 0$, $i = 1, 2$, with $\lambda_1 + \lambda_2 = 1$. By Ito's differentiation rule,

$$d[\exp \zeta^t(\phi_i)] = \exp \zeta^t(\phi_i)\langle \phi_i(t), dx(t)\rangle.$$

Define

$$\rho(t) = \lambda_1 \exp \zeta^t(\phi_1) + \lambda_2 \exp \zeta^t(\phi_2).$$

Then

$$d\rho(t) = \lambda_1 \exp \zeta^t(\phi_1)\langle \phi_1(t), dx(t)\rangle + \lambda_2 \exp \zeta^t(\phi_2)\langle \phi_2(t), dx(t)\rangle$$

which we can rewrite as

$$d\rho(t) = \rho(t)\langle \phi(t), dx(t)\rangle, \tag{16}$$

where

$$\phi(t) = \frac{\lambda_1 \exp \zeta^t(\phi_1)}{\sum_{i=1}^2 \lambda_i \exp \zeta^t(\phi_i)}\phi_1(t) + \frac{\lambda_2 \exp \zeta^t(\phi_2)}{\sum_{i=1}^2 \lambda_i \exp \zeta^t(\phi_i)}\phi_2(t).$$

Evidently $\phi \in \Phi^N$ since $\phi(t, z)$ is a convex combination of $\phi_1(t, z)$ and $\phi_2(t, z)$. By Ito's differentiation rule from (16) we obtain

$$d(\log \rho(t)) = \langle \phi(t), dx(t)\rangle - \tfrac{1}{2}|\phi(t)|^2\, dt \tag{17}$$

and $\rho(0) = 1$ from (14) so that integrating (17) yields

$$\log \rho(t) = \int_0^t \langle \phi(s), dx(s)\rangle - \frac{1}{2}\int_0^t |\phi(s)|^2\, ds.$$

Hence $\rho(1) = \exp \zeta(\phi)$ and the lemma is proved.

We now state our main result and develop the proof through a sequence of lemmas.

THEOREM 2. (i) *$\mathscr{D}(\mathscr{G})$ is a convex set.*

(ii) *Let*

$$\mathscr{G}^0 = \{g | g \in \mathscr{G}, E_0(\exp \zeta(g)) = 1\}.$$

Then, $\mathscr{D}(\mathscr{G}^0)$ is a closed and convex subset of $L^1(\Omega, \mathscr{A}, P_0)$.

We shall develop the proof through a sequence of lemmas.

LEMMA 5. *$\mathscr{D}(\mathscr{G})$ is convex.*

Proof. Let $g_i(t, z) = f(t, z, u_i(t, z))$ with $u_i \in \mathscr{U}_i$, $i = 1, 2$, and let $\lambda_i \geqq 0$, with $\lambda_1 + \lambda_2 = 1$. By Ito's differentiation rule,

$$d(\exp \zeta^t(g_i)) = \exp \zeta^t(g_i)\langle g_i(t), dx(t)\rangle, \qquad i = 1, 2.$$

Define

$$\rho(t) = \lambda_1 \exp \zeta^t(g_1) + \lambda_2 \exp \zeta^t(g_2).$$

Then if we repeat the proof of Lemma 4 we can conclude that (noting $\rho(0) = 1$)

$$\rho(1) = \lambda_1 \exp \zeta(g_1) + \lambda_2 \exp \zeta(g_2) = \exp \zeta(\phi),$$

where $\phi(t, z)$ is a convex combination of $g_1(t, z)$ and $g_2(t, z)$. Since $g_i(t, z) \in f(t, z, U)$, and since this set is convex by condition C5, we see that

$$\phi(t, z) \in f(t, z, U)$$

and hence $\phi \in \mathscr{G}$ by Lemma 1. The lemma is proved.

LEMMA 6. *$\mathscr{D}(\mathscr{G}^0)$ is convex.*

Proof. The set

$$\mathscr{R} = \{\rho | \rho \in L^1(\Omega, \mathscr{A}, P_0), \rho \geqq 0, E_0\rho = 1\}$$

is convex, and

$$\mathscr{D}(\mathscr{G}^0) = \mathscr{D}(\mathscr{G}) \cap \mathscr{R},$$

so that the result follows from Lemma 5.

Next let g_n, $n = 1, 2, \cdots$, be a sequence from $\mathscr{G}^0$, and let ρ be in $L^1(\Omega, \mathscr{A}, P_0)$ such that

$$\lim_{n\to\infty} \exp \zeta(g_n) = \rho \quad \text{a.s. } P_0 \quad \text{and in } L^1(\Omega, \mathscr{A}, P_0). \tag{18}$$

For each positive integer N, let

$$g_n^N(t, z) = \begin{cases} g_n(t, z) & \text{if } |z(\tau)| \leqq N \quad \text{for } \tau \leqq t, \\ 0 & \text{otherwise,} \end{cases}$$

and for $N = 1, 2, 3, \cdots$, inductively select subsequences g_k^N, $k \in K_N$, and $\phi^N \in \Phi^N$ as follows:

For $N = 1$, let g_k^1, $k \in K_1$, be a subsequence of g_n^1, $n = 1, 2, 3, \cdots$, and let $\phi^1 \in \Phi^1$ be such that

$$\exp \zeta(\phi^1) = \underset{k\in K_1}{\text{w. lim}} \exp \zeta(g_k^1).$$

(Here and in the remainder w. lim means the weak limit in $L^1(\Omega, \mathscr{A}, P_0)$.) From Lemmas 4 and 5, $\mathscr{D}(\Phi^N)$ is a weakly, sequentially compact subset of $L^2(\Omega, \mathscr{A}, P)$ and $g_n^N \in \Phi^N$ so that the above selection makes sense.

Suppose g_k^N, $k \in K_N$, and $\phi^N \in \Phi^N$ are defined. Then let g_k^{N+1}, $k \in K_{N+1}$, be a sequence of g_k^N, $k \in K_N$, and let $\phi^{N+1} \in \Phi^{N+1}$ be such that

$$\exp \zeta(\phi^{N+1}) = \underset{k\in K_{N+1}}{\text{w. lim}} \exp \zeta(g_k^{N+1}).$$

LEMMA 7. *Let $C_N^t = \{z | z \in C, |z(\tau)| \leqq N \text{ for } \tau \leqq t\}$. Then for $i \geqq 0$,*

$$\phi^{N+1}(t, z) = \phi^N(t, z) \quad \text{for } 0 \leqq t \leqq 1, \quad z \in C_N^1.$$

Proof. First of all from

$$\exp \zeta(\phi^N) = \underset{k\in K_N}{\text{w. lim}} \exp \zeta(g_k^N),$$

it is immediate that

$$E_0(\exp \zeta(\phi^N) | \mathscr{A}_t) = \underset{k\in K_N}{\text{w. lim}} E_0(\exp \zeta(g_k^N) | \mathscr{A}_t).$$

Secondly since

$$E_0(\zeta(\phi^N)) = 1, \qquad E_0(\zeta(g_k^N)) = 1,$$

it follows that a.s. P_0,

$$\exp \zeta^t(\phi^N) = E_0(\exp \zeta(\phi^N) | \mathscr{A}_t), \qquad \exp \zeta^t(g_k^N) = E_0(\exp \zeta(g_k^N) | \mathscr{A}_t),$$

and hence,

$$\exp \zeta^t(\phi^N) = \text{w.}\lim_{k\in K_N} \exp \zeta^t(g_k^N). \tag{19}$$

Next let

$$\Omega_N^t = \{\omega | \omega \in \Omega, x(\cdot, \omega) \in C_N^t\}.$$

By definition, for $i \geqq 0$,

$$g_k^N(\tau, x(\cdot, \omega)) = g_k^{N+i}(\tau, x(\cdot, \omega)) \quad \text{for } \tau \leqq t, \quad \omega \in \Omega_N^t,$$

so that from (19) for $i \geqq 0$,

$$\exp \zeta^t(\phi^{N+i})(\omega) = \exp \zeta^t(\phi^N)(\omega), \qquad \tau \leqq t, \quad \omega \in \Omega_N^t.$$

The result now follows if we note that

$$0 = \int_{\Omega_N^1} |\exp \zeta(\phi^{N+i}) - \exp \zeta(\phi^N)|^2 P_0(d\omega) = \int_{\Omega_N^1} \left[\int_0^1 \exp \zeta^t(\phi^{N+i}) \langle \phi^{N+i}(t), dx(t) \rangle \right.$$

$$\left. - \int_0^1 \exp \zeta^t(\phi^N) \langle \phi^N(t), dx(t) \rangle \right]^2 P_0(d\omega)$$

$$= \int_{\Omega_N^1} \left[\int_0^1 \exp 2\zeta^t(\phi^N) |\phi^{N+i}(t, x(\cdot, \omega)) - \phi^N(t, x(\cdot, \omega))|^2 \, dt \right] P_0(d\omega)$$

so that since $\exp \zeta^t(\phi^N) > 0$ a.s. P_0, we must have

$$\int_{\Omega_N^1} \left[\int_0^1 |\phi^{N+i}(t, x(\cdot, \omega)) - \phi^N(t, x(\cdot, \omega))|^2 \, dt \right] P_0(d\omega) = 0,$$

and the lemma is proved.

Because of Lemma 7 we can define a causal function $\phi : [0, 1] \times C \to R^n$ such that

$$\phi(t, z) = \phi^{N+i}(t, z) \quad \text{for } 0 \leqq t \leqq 1, \quad \|z\| \leqq N, \quad i \geqq 0. \tag{20}$$

From the proof of Lemma 8, and from (18) it follows that

$$\rho = \exp \zeta(\phi) \quad \text{a.s. } P_0.$$

Lemma 8 completes the proof of Theorem 2.

LEMMA 8.

$$\phi \in \mathcal{G}.$$

Proof. Because of (20) and Lemma 1 it is enough to show that

$$\phi^N(t, z) \in f(t, z, U) \quad \text{for } 0 \leqq t \leqq 1, \quad \|z\| \leqq N. \tag{21}$$

Recall that

$$\exp \zeta(\phi^N) = \text{w.}\lim_{k\in K_1} \exp \zeta(g_k^N).$$

From the properties of weak L^2-convergence it is known that there is a convex combination of the $\exp \zeta(g_k^N)$ which converges to $\exp \zeta(\phi^N)$ in the L^2-norm topology.

More precisely, for each n, there are nonnegative numbers $\lambda_1^n, \cdots, \lambda_n^n$ with $\lambda_1^n + \cdots + \lambda_n^n = 1$ such that

$$\lim_{n\to\infty} E_0 \left| \exp \zeta(\phi^N) - \sum_{i=1}^{n} \lambda_i^n \exp \zeta(g_i^N) \right|^2 = 0. \tag{22}$$

Let

$$h_n(t) = \sum_{i=1}^{n} \lambda_i^n \exp \zeta(g_i^N).$$

Repeating the proof of Lemma 4, we can conclude that

$$h_n(t) = \exp \zeta^t(\eta_n) \quad \text{a.s. } P_0,$$

where $\eta_n(t, z)$ is a convex combination of $g_1^N(t, z), \cdots, g_n^N(t, z)$. In particular, from the convexity of $f(t, z, U)$ and the fact that $g_i^N(t, z) = g_i(t, z) \in f(t, z, U)$ for $\|z\| \leqq N$, it follows that for $\|z\| \leqq N$,

$$\eta_n(t, z) \in f(t, z, U). \tag{23}$$

Next

$$\begin{aligned} E_0|\exp \zeta(\phi^N) - \exp \zeta(\eta_n)|^2 &= E_0 \left| \int_0^1 \exp \zeta^t(\phi^N) \langle \phi^N(t), dx(t) \rangle \right. \\ &\quad \left. - \int_0^1 \exp \zeta^t(\eta_n) \langle \eta_n(t), dx(t) \rangle \right|^2 \\ &= E_0 \int_0^1 |\exp \zeta^t(\phi^N)\phi^N(t) - \exp \zeta^t(\eta_n)\eta_n(t)|^2 \, dt \end{aligned}$$

converges to zero by (22). Taking subsequences if necessary we see that

$$\exp \zeta^t(\phi^N)(\omega)\phi^N(t, x(\cdot, \omega)) = \lim_{n\to\infty} \exp \zeta^t(\eta_n)(\omega)\eta_n(t, x(\cdot, \omega)) \quad \text{a.s. } l \otimes P_0, \tag{24}$$

where l denotes Lebesgue measure on $[0, 1]$.

Also a.s. P_0,

$$\exp \zeta^t(\phi^N) = E_0(\zeta(\phi^N)|\mathscr{A}_t), \qquad \exp \zeta^t(\eta_n) = E_0(\zeta(\eta_n)|\mathscr{A}_t),$$

so that from (22),

$$\lim_{n\to\infty} \int_0^1 E_0|\exp \zeta^t(\phi^N) - \exp \zeta^t(\eta_n)|^2 \, dt = 0,$$

and hence taking subsequences if necessary, we have

$$\exp \zeta^t(\phi^N)(\omega) = \lim_{n\to\infty} \exp \zeta^t(\eta_n)(\omega) \quad \text{a.s. } l \otimes P_0.$$

Since $\exp \zeta^t(\phi^N) > 0$ a.s. P_0, we conclude from (24) that

$$\phi^N(t, x(\cdot, \omega)) = \lim_{n\to\infty} \eta_n(t, x(\cdot, \omega)) \quad \text{a.s. } l \otimes P_0,$$

and hence from (23), and the fact that $f(t, z, U)$ is closed, we see that

$$\phi^N(t, x(\cdot, \omega)) \in f(t, x(\cdot, \omega), U) \quad \text{for } \|x(\cdot, \omega)\| \leqq N, \quad \text{a.s. } l \otimes P_0.$$

From Corollary 1 of Lemma 2 and from Theorem 2 we obtain Theorem 3.

THEOREM 3. *Suppose f satisfies* C1–C5 *of* § 2.

(i) *For an admissible control $u \in \mathcal{U}$, there exists a solution to* (1) *with continuous sample paths (without explosions) if and only if*

$$E_0 \exp \zeta(g_u) = 1.$$

(ii) *The set of densities* $\{\exp \zeta(g_u) | E_0 \exp \zeta(g_u) = 1\}$ *is a convex set, which is closed in the norm topology of* $L^1(\Omega, \mathcal{A}, P_0)$.

4. Sufficient conditions for $E_0 \exp \zeta(\phi) = 1$.

LEMMA 9. *Let $\phi : [0, 1] \times C \to R^n$ be a causal map such that $\int_0^1 |\phi(t, z)|^2 \, dt < \infty$ for all z in C. Define $T_\phi : C \to C$ by*

$$T_\phi(z)(t) = z(t) - \int_0^t \phi(\tau, z) \, d\tau. \tag{25}$$

Suppose that for each $N > 0$ there is $M > 0$ such that $\| T_\phi(z) \| \leqq N$ implies $\| z \| \leqq M$. Then,

$$E_0 \exp \zeta(\phi) = 1.$$

Proof. The proof is immediate from Lemma 7 of [5].

As a consequence of Lemma 9, we can obtain the following sufficient conditions. The first result is due to Beneš [3].

COROLLARY 3. *Let $\phi : [0, 1] \times C \to R^n$ be a causal map and suppose there is a constant K such that*

$$|\phi(t, z)| \leqq K(1 + \max_{1 \leqq \tau \leqq t} |z(\tau)|).$$

Then,

$$E_0 \exp \zeta(\phi) = 1.$$

Proof. Let $T_\phi(z)(t) = y(t)$, and let $\gamma(t) = \max_{0 \leqq \tau \leqq t} |z(\tau)|$. Then, from (25),

$$\gamma(t) \leqq |y(t)| + \int_0^t K(1 + \gamma(\tau)) \, d\tau$$

$$\leqq (\| y \| + K) + \int_0^t K\gamma(\tau) \, d\tau.$$

By the Bellman–Gronwall inequality,

$$\| z \| = \gamma(1) \leqq (\exp K)\gamma(0) + (\exp K)(\| y \| + K)$$

$$\leqq (\exp K)(2\| y \| + K),$$

and the result follows from Lemma 9.

The next result is useful if we have a control system with delay.

COROLLARY 4. *Let $\phi : [0, 1] \times C \to R^n$ be a causal map such that for some $\delta > 0$,*

$$|\phi(t, z)| \leqq f_0(\max_{0 \leqq \tau \leqq t - \delta} |z(\tau)|),$$

where $f_0 : R \to R$ is increasing. Then,

$$E_0 \exp \zeta(\phi) = 1.$$

Proof. Let y, γ be defined as in the previous proof. Then,

$$\begin{aligned}\gamma(\delta) &\leqq \|y\| + f_0(\gamma(0)),\\ \gamma(2\delta) &\leqq \|y\| + f_0(\gamma(\delta)) \leqq \|y\| + f_0(\|y\| + f_0(\gamma(0)))\\ &= f_1(\|y\|, \gamma(0)) \quad \text{say}.\end{aligned}$$

By induction,

$$\gamma(i\delta) \leqq f_i(\|y\|, \gamma(0)),$$

where f_i is increasing in each argument. Evidently if $(m-1)\delta < 1 \leqq m\delta$, we see that

$$\gamma(1) = \|z\| \leqq f_m(\|y\|, |z(0)|),$$

and the result follows from Lemma 9.

Remark. McKean [8, p. 66] has shown that if $\delta > 0$, then all solutions of the one-dimensional diffusion equation

$$dx(t) = |x|^{1+\delta}\, dt + dB(t), \qquad 0 \leqq t < \infty,$$

explode with probability 1. It follows that condition (6) is a nontrivial restriction.

5. Applications. Consider a control system

$$dx(t, \omega) = f(t, x(\cdot, \omega), u(t, x(\cdot, \omega)))\, dt + dB(t, \omega),$$

where the control u takes values in a set U and f obeys the conditions C1–C5 of § 2. Let us impose an additional restriction.

C6. For every admissible $u \in \mathcal{U}$,

$$E_0 \exp \zeta(g_u) = 1,$$

or equivalently (and directly in terms of u) for

$$\rho_u(z) = \exp\left[\int_0^1 \langle f(t, z, u(t, z)), dz(t)\rangle - \frac{1}{2}\int_0^1 |f(t, z, u(t, z))|^2\, dt\right]: \tag{26}$$

C6′.

$$\int_C \rho_u(z)\mu(dz) = 1.$$

Instead we can limit ourself to the subset $\mathcal{U}^0$ consisting of those u in $\mathcal{U}$ which satisfy C6′.

Next let $L : C \to R$ be a bounded function, measurable with respect to $\mathcal{S}$. L is the *cost function* and assigns to every $u \in \mathcal{U}^0$ the cost

$$J(u) = \int_C L(z)\rho_u(z)\mu(dz). \tag{27}$$

THEOREM 4. *Suppose $\mathcal{U}^0$ is nonempty. Then, there exists $u^* \in \mathcal{U}^0$ such that*

$$J(u^*) \leqq J(u) \quad \textit{for all } u \in \mathcal{U}^0.$$

Proof. By Theorem 3, the set $\{\rho_u | u \in \mathcal{U}^0\}$ is a strongly closed, convex subset of $L^1(C, \mathcal{S}, \mu)$. Hence it is weakly compact. Since $\int_C L(z)\rho_u(z)\mu(dz)$ is linear and continuous in ρ_u, the result follows.

Let us note that a cost functional of the type (27) allows for a variable endpoint problem as follows. Let $\mathcal{T}$ be a closed subset of $[0, 1] \times R^n$ which includes the set $\{1\} \times R^n$. Let $\lambda : [0, 1] \times C \to R^n$ be a bounded, causal function, and to each $u \in \mathcal{U}^0$ assign the cost

$$J(u) = \int_C \left[\int_0^{t(z)} \lambda(t, z)\, dt \right] \rho_u(z)\mu(dz),$$

where $t(z) = \min\{\tau | z(\tau) \in \mathcal{T}\}$. The term in brackets is clearly of the form $L(z)$ in (27).

If the cost also depends on the control u, then sometimes we can add an extra coordinate to the state vector and get an equivalent cost depending only on the state. See [3] for details.

As a second application consider a zero-sum stochastic differential game, with two players I and II, with controls $u_1(t) \in U_1$ and $u_2(t) \in U_2$ respectively, and dynamics given by

$$dx(t) = f(t, x, u_1(t), u_2(t))\, dt + dB(t).$$

Suppose that f splits as

$$f(t, x, u_1, u_2) = \begin{pmatrix} f_1(t, x, u_1) \\ f_2(t, x, u_2) \end{pmatrix}.$$

Assume that f satisfies C1–C5 with C5 now restated as: $f_1(t, z, U_1)$ and $f_2(t, z, U_2)$ are closed and convex for each (t, z). As before, we define the admissible controls for player i, as all causal maps $u_i : [0, 1] \times C \to U_i$, $i = 1, 2$. Let $\mathcal{U}_i^0$ consist of those admissible controls u_i which satisfy

$$\int_C \rho_{u_i}^i(z)\mu(dz) = 1,$$

where

$$\rho_{u_i}^i = \exp\left[\int_0^1 \langle f_i(t, z, u_i(t, z)), dz_i(t) \rangle - \frac{1}{2} \int_0^1 |f_i(t, z, u_i(t, z))|^2\, dt \right]. \tag{28}$$

Here we have split $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$ to be compatible with $f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$. Let $L : C \to R$ be a bounded function, measurable with respect to $\mathcal{S}$, and to each pair $(u_1, u_2) \in \mathcal{U}_1^0 \times \mathcal{U}_2^0$ assign to player I the payoff

$$J(u_1, u_2) = \int_C L(z)\rho_{(u_1, u_2)}(z)\mu(dz). \tag{29}$$

THEOREM 5. *Suppose* $\mathcal{U}_i^0$ *is nonempty for* $i = 1, 2$. *Then, there exist* $u_i^* \in \mathcal{U}_i^0$, $i = 1, 2$, *such that*

$$J(u_1, u_2^*) \leqq J(u_1^*, u_2^*) \leqq J(u_1^*, u_2) \quad \text{for all } u_i \in \mathcal{U}_i^0, \quad i = 1, 2.$$

Proof. From the definition (26) of $\rho_{(u_1,u_2)}$ and the definition (28) we see that

$$J(u_1, u_2) = \int_C L(z)\rho_{u_1}^1(z)\rho_{u_2}^2(z)\mu(dz). \tag{30}$$

Next, from Theorem 3, the sets $\{\rho_{u_i}^i | u_i \in \mathcal{U}_i^0\}$ are convex, closed subsets of $L^1(C, \mathcal{S}, \mu)$, hence weakly compact. Finally the integral in (30) is concave (in fact linear) and continuous in $\rho_{u_1}^1$ for fixed $\rho_{u_2}^2$, and convex (in fact linear) and continuous in $\rho_{u_2}^2$ for fixed $\rho_{u_1}^1$. Hence from the well-known results on two-person zero-sum games the existence of a saddle point (u_1^*, u_2^*) follows.

6. Partial observations: A negative example. Again consider the stochastic differential equation

$$dx(t) = f(t, x, u)\,dt + dB(t). \tag{1}$$

The conditions on f are as before, but now suppose that we consider the important case where the control u can only depend upon the past history of the last m $(m < n)$ components of x. More precisely, let Q_t be the sub-σ-algebra of $\mathcal{S}_t$ generated by all sets of the form

$$\{z | z \in C, z_i(\tau) \in A\},$$

where $\tau \leqq t$, A is a Borel subset of R and $n - m + 1 \leqq i \leqq n$. Let $\mathcal{U}_m$ be the set of all causal maps $u : [0, 1] \times C \to U$ such that $u(t, \cdot)$ is measurable with respect to Q_t. Evidently,

$$\mathcal{U}_m \subset \mathcal{U} = \mathcal{U}_n.$$

First of all it should be clear from the proof of Lemma 4 that the set $\{\exp \zeta(g_u) | u \in \mathcal{U}_m\}$ may fail to be convex. For, consider the two-dimensional system

$$dx_1 = u\,dt + dB_1, \qquad dx_2 = dB_2,$$

where the control u is allowed to depend only on x_2 and must take values in the set $U = [-1, 1]$. Now let u_1 and u_2 be control laws defined as follows:

$$u_1(t, x_2) \equiv 0,$$

$$u_2(t, x_2) = \begin{cases} 0, & 0 \leqq t < \frac{1}{2}, \\ \operatorname{sgn}(x_2(\frac{1}{2})), & \frac{1}{2} \leqq t \leqq 1. \end{cases}$$

It is easy to calculate that

$$\zeta^t(g_{u_1}) \equiv 0$$

so that $\exp \zeta^t(g_{u_1}) \equiv 1$ and

$$\zeta^t(g_{u_2}) = \begin{cases} 0, & 0 \leqq t < \frac{1}{2}, \\ \operatorname{sgn}(x_2(\frac{1}{2}))(x_1(t) - x_1(\frac{1}{2})) - \frac{1}{2}(t - \frac{1}{2}), & t \geqq \frac{1}{2}. \end{cases}$$

From the proof of Lemma 4 we know that the function u_λ, given by $u_\lambda(t) \equiv 0$, $0 \leqq t \leqq \frac{1}{2}$, and

$$u_\lambda(t) = \frac{\lambda \exp \zeta^t(g_{u_2})}{\lambda \exp \zeta^t(g_{u_2}) + (1 - \lambda)} u_2(t), \qquad \tfrac{1}{2} \leqq t \leqq 1,$$

satisfies

$$\exp \zeta^t(g_{u_\lambda}) \equiv (1 - \lambda) \exp \zeta^t(g_{u_1}) + \lambda \exp \zeta^t(g_{u_2}).$$

Since $d \exp \zeta^t(g_{u_\lambda}) = \exp \zeta^t(g_{u_\lambda}) \langle g_{u_\lambda}(t), dx(t) \rangle$, g_{u_λ} (and hence u_λ) is uniquely defined by the previous equation. Since for $0 < \lambda < 1$, u_λ depends on x_1, the assertion is proved.

Next, from Theorem 3 we see that the set $\{\exp \zeta(g_u) | u \in \mathcal{U}_m\}$ is weakly compact in $L^1(\Omega, \mathcal{A}, P_0)$ if and only if it is weakly closed. We give a simple example to show that in general we do not have weak closure.

Consider the two-dimensional system $x = (x_1, x_2)$, with $u \in R$ depending only on x_2,

$$dx_1(t) = f(t, x_1)u + dB_1(t), \qquad dx_2(t) = dB_2(t),$$

where

$$f(t, x_1) = \begin{cases} 0, & t \leqq \frac{1}{2}, \\ 2, & t > \frac{1}{2}, \quad x_1(\frac{1}{2}) > 0, \\ 1, & t > \frac{1}{2}, \quad x_1(\frac{1}{2}) \leqq 0. \end{cases}$$

The control set is $U = [-1, 1]$. We shall define a sequence of control laws $u_n(t, x_2)$ such that

$$u_n(t, x_2) = \begin{cases} 0, & t \leqq \frac{1}{2}, \\ \gamma_n(x_2(\frac{1}{2})), & t > \frac{1}{2}, \end{cases}$$

where the functions γ_n are defined later. It follows that

$$\begin{aligned} \zeta_n = \zeta(g_{u_n}) &= \int_0^1 f(t, x_1)u_n(t)\, dx_1(t) - \frac{1}{2}\int_0^1 f^2(t, x_1)u_n^2(t)\, dt \\ &= \alpha\beta\gamma_n - \tfrac{1}{4}\beta^2\gamma_n^2, \end{aligned}$$

where

$$\alpha = x_1(1) - x_1(\tfrac{1}{2}), \qquad \beta = \begin{cases} 2 & \text{if } x_1(\frac{1}{2}) > 0, \\ 1 & \text{if } x_1(\frac{1}{2}) \leqq 0. \end{cases}$$

Therefore,

$$\exp \zeta_n = \exp(\alpha\beta\gamma_n) \exp(-\tfrac{1}{4}\beta^2 \gamma_n^2). \tag{31}$$

We shall select γ_n such that $|\gamma_n| \equiv 1$, so that (31) simplifies to

$$\exp \zeta_n = \exp(\alpha\beta\, \gamma_n) \exp(-\tfrac{1}{4}\beta^2). \tag{32}$$

We define γ_n as follows:

Let $\xi : R \to R$ be a measurable function such that under P_0, $\xi(x_2(\frac{1}{2}))$ is uniformly distributed over $[0, 1]$. For each integer $n \geqq 0$, define $\eta_n : [0, 1] \to \{-1, 1\}$ by

$$\eta_n(\xi) = \begin{cases} 1 & \text{if } \dfrac{2m}{2n} \leqq \xi < \dfrac{2m+1}{2n}, \quad m = 0, 1, \cdots, n-1, \\ -1 & \text{if } \dfrac{2m+1}{2n} \leqq \xi < \dfrac{2m+2}{2n}, \quad m = 0, 1, \cdots, n-1. \end{cases}$$

Finally, let

$$\gamma_n(x_2(\tfrac{1}{2})) = \eta_n(\xi(x_2(\tfrac{1}{2}))).$$

LEMMA 10. *$\exp \zeta_n$ converges to $\frac{1}{2}[\exp(\alpha\beta) + \exp(-\alpha\beta)] \exp(-\frac{1}{4}\beta^2)$ in the weak topology of $L^1(\Omega, \mathscr{A}, P_0)$.*

Proof. Let A_α, A_β, A_ξ be Borel subsets of R and let I_A denote the indicator function of a set A. Let

$$\Pi_n = \int_\Omega I_{A_\alpha}(\alpha(\omega)) I_{A_\beta}(\beta(\omega)) I_{A_\xi}(\xi(\omega))(\exp \zeta_n)(\omega) P_0(d\omega).$$

Now under P_0 the random variables α, β, ξ are independent, so that

$$\Pi_n = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I_{A_\alpha}(\alpha) I_{A_\beta}(\beta) I_{A_\xi}(\xi) \exp(\alpha\beta\eta_n(\xi)) \exp\left(-\frac{1}{4}\beta^2\right) P_\xi(d\xi) P_\beta(d\beta) P_\alpha(d\alpha), \tag{33}$$

where P_α, P_β, P_ξ are the marginal distributions of α, β, ξ respectively. From the way η_n is defined and the fact that ξ is uniformly distributed on $[0, 1]$ it follows that for fixed α, β,

$$\lim_{n\to\infty} \int_{-\infty}^{t} \exp(\alpha\beta\eta_n(\xi)) \exp\left(-\frac{1}{4}\beta^2\right) P_\xi(d\xi)$$

$$= \int_{-\infty}^{t} \frac{1}{2}[\exp(\alpha\beta) + \exp(-\alpha\beta)] \exp\left(-\frac{1}{4}\beta^2\right) P_\xi(d\xi)$$

uniformly for $t \in (-\infty, \infty)$. It follows that $\exp(\alpha\beta\eta_n(\xi)) \exp(-\frac{1}{4}\beta^2)$ converges to $\frac{1}{2}[\exp(\alpha\beta) + \exp(-\alpha\beta)] \exp(-\frac{1}{4}\beta^2)$ weakly in $L^1(R, P_\xi)$. Since the integrands in (33) are uniformly integrable, it follows that

$$\lim_{n\to\infty} \Pi_n = \int_{R^3} I_{A_\alpha}(\alpha) I_{A_\beta}(\beta) I_{A_\xi}(\xi) \tfrac{1}{2}[\exp(\alpha\beta) + \exp(-\alpha\beta)] \cdot \exp(-\tfrac{1}{4}\beta^2) P_\xi(d\xi) P_\beta(d\beta) P_\alpha(d\alpha).$$

From this it follows easily that $\exp(\alpha\beta\eta_n(\xi)) \exp(-\frac{1}{4}\beta^2)$ converges to $\frac{1}{2}[\exp(\alpha\beta) + \exp(-\alpha\beta)] \exp(-\frac{1}{4}\beta^2)$ weakly in $L^1(R^3, P_\alpha \otimes P_\beta \otimes P_\xi)$ and the lemma is proved.

Next, by direct calculation we can show that the two-dimensional drift $\hat{g}$, where $\hat{g}(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $0 \leqq t \leqq \frac{1}{2}$, and

$$\hat{g}(t) = \begin{pmatrix} \beta \dfrac{\exp 2(x_1(t) - x_1(\frac{1}{2})) - 1}{\exp 2(x_1(t) - x_1(\frac{1}{2})) + 1} \\ 0 \end{pmatrix}, \qquad \frac{1}{2} \leqq t \leqq 1,$$

satisfies

$$\exp \zeta(\hat{g}) = \tfrac{1}{2}[\exp(\alpha\beta) + \exp(-\alpha\beta)] \exp(-\tfrac{1}{4}\beta^2). \tag{34}$$

If we define $\rho(t) = E_0(\exp \zeta(\hat{g}) | \mathscr{A}_t)$, then

$$d\rho(t) = \rho(t)\langle \hat{g}(t), dx(t) \rangle, \qquad 0 \leqq t \leqq 1,$$

so that (34) characterizes $\hat{g}$ uniquely. Hence any control law $\hat{u}$ such that $\exp \zeta(g_u) = \exp \zeta(\hat{g})$ must satisfy $g_u = \hat{g}$ so that $\hat{u}$ must depend on x_1. Therefore the set of densities $\exp \zeta(g_u)$ with u depending only on x_2 is not weakly closed in $L^1(\Omega, \mathscr{A}, P_0)$.

Incidentally this example also shows that to guarantee weak closure, the convexity condition C5 is necessary, for even though $u_n(t) \in \{-1, 1, 0\}$ for all t, it is not the case for $\hat{u}(t)$.

Acknowledgments. It should be obvious to anyone who has read [3] our great debt to V. E. Beneš. E. Wong clarified many subtle points about Ito's calculus and suggested the proofs in § 2. M. Davis caught many errors in earlier proofs. It is a pleasure to acknowledge our gratitude to them.

REFERENCES

[1] W. H. Fleming and M. Nisio, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777–794.

[2] H. J. Kushner, *On the existence of optimal stochastic controls*, this Journal, 3 (1965), pp. 463–474.

[3] V. E. Beneš, *Existence of optimal stochastic control laws*, to appear.

[4] W. H. Fleming, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–509.

[5] I. V. Girsanov, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theor. Probability Appl., 5 (1960), pp. 285–301.

[6] V. E. Beneš, *Existence of optimal strategies based on specified information, for a class of stochastic decision problems*, this Journal, 8 (1970), pp. 179–188.

[7] K. Ito, *Multiple Wiener integral*, J. Math. Soc. Japan, 3 (1951), pp. 158–161.

[8] H. P. McKean, Jr., *Stochastic Integrals*, Academic Press, New York and London, 1969.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

ALMOST SURE BOUNDEDNESS OF RANDOMLY SAMPLED SYSTEMS*

R. G. AGNIEL† AND E. I. JURY‡

Abstract. This paper discusses the almost sure boundedness of linear and nonlinear randomly sampled systems. It is shown that if an autonomous linear randomly sampled system exhibits almost sure asymptotic stability, then the system is almost surely bounded input–bounded output. Moreover, for a bounded input, the second moment of the output remains bounded and this bound is easily computable.

It is also found that linear or nonlinear systems which are almost surely asymptotically stable for a null input remain almost surely bounded when the input consists of an uncorrelated noise with finite variance.

Introduction. Recently, more and more attention has been devoted to sampled data systems with a randomly variable sampling period. The sampling intervals may be random for physical reasons, when they are contaminated by noise because of cheap equipment ("jitter") or when the sampling is governed by a random process such as a sonar or radar echo. In certain cases, a randomly variable period is advantageous for economical reasons, for instance, when a time-shared computer controls several plants.

Moreover, randomly sampled systems were considered by many authors such as Kalman [3], Bergen [4], Bharucha [5] and others. In [3], Kalman repeatedly applies the Schwarz and Minkowski inequalities to prove that for an autonomous system which shows stability of the second moment of the output, this second moment remains bounded when a bounded input is applied to the system.

Following Bucy's idea [6], [7] of using the submartingale convergence theorem, in this paper we will prove that not only the second moment remains bounded, but the output is almost surely bounded as well.

Moreover, it will be proved that a large class of linear or nonlinear randomly sampled systems, shown to be almost surely asymptotically stable with a null input, remain almost surely bounded when the input consists of an uncorrelated noise with finite variance.

1. Stochastic stability concepts [8]. Let $t_0 < t_1 < t_2 < t_3 < \cdots < t_n < t_{n+1} < \cdots$ be a discrete sequence of strictly ordered random times, and let $\mathbf{X}(n, \mathbf{X}_0, n_0)$ be a discrete sequence of random vectors at time t_n such that $\mathbf{X}(n_0, \mathbf{X}_0, n_0) = \mathbf{X}_0$ and satisfying the vector difference equation

$$\mathbf{X}(n+1, \mathbf{X}_0, n_0) = f[\mathbf{X}(n, \mathbf{X}_0, n_0), \mathbf{r}_n, \mathbf{u}_n],$$

where the r_n's are a sequence of random vectors and the u_n's a sequence of input vectors.

* Received by the editors December 12, 1969, and in final revised form December 3, 1970. This research was sponsored by the Air Force Office of Scientific Research, Office of Aerospace Research, under AFOSR Grant AF-AFOSR-68-1463.

† Direction Commerciale Groupe Tubes Electroniques, Thomson-CSF, Paris, France. Formerly with Electronics Research Laboratory, University of California, Berkeley, California.

‡ Department of Electrical Engineering and Computer Science, and Electronics Research Laboratory, University of California, Berkeley, California 94720.

Let $\mathbf{0}$ be an equilibrium solution, i.e.,

$$f[\mathbf{0}, \mathbf{r}_u, \mathbf{0}] = \mathbf{0} \quad \text{for all } n.$$

The equilibrium solution $\mathbf{X}_e \equiv \mathbf{0}$ is said:

(a) to be *almost surely stable* if given $\varepsilon, \varepsilon' > 0$, there exists $\delta(\varepsilon, \varepsilon', n_0)$ such that $\|\mathbf{X}_0\| < \delta$ implies

$$P[\sup_{n \geqq n_0} \|\mathbf{X}(n, \mathbf{X}_0, n_0)\| > \varepsilon'] < \varepsilon,$$

where $P[\cdot]$ denotes the probability and $\|\cdot\|$ is a vector norm;

(b) to be *almost surely asymptotically stable* if

(i) it is almost surely stable,

(ii) there exists $\delta' > 0$ such that $\|\mathbf{X}_0\| < \delta'$ implies

$$\|\mathbf{X}(n, \mathbf{X}_0, n_0)\| \xrightarrow[n]{\text{a.s.}} 0;$$

(c) to show *stability of the p-th moment* if given $\varepsilon > 0$, there exists $\delta(\varepsilon, n_0)$ such that $\|\mathbf{X}_0\| < \delta$ implies

$$\sup_{n \geqq n_0} E[\|\mathbf{X}(n, \mathbf{X}_0, n_0)\|^p] < \varepsilon;$$

(d) to show *asymptotic stability of the p-th moment* if

(i) it shows stability of the pth moment,

(ii) there exists $\varepsilon' > 0$ such that $\|\mathbf{X}_0\| < \delta'$ implies

$$\lim_{n \to \infty} E[\|\mathbf{X}(n, \mathbf{X}_0, n_0)\|^p] = 0;$$

(e) to be *almost surely bounded input–bounded output* if for any $\|\mathbf{X}_0\| < \infty$ and any almost surely bounded input,

$$\sup_n \|\mathbf{X}(n, \mathbf{X}_0, n_0)\| < \infty \quad \text{a.s.}$$

In what follows, any of these modes of stability will be said to hold absolutely if true for a null input.

2. Stability of linear randomly sampled systems.

2.1. Description of the system. Consider the closed loop sampled data system comprised of a random sampling and hold device followed by a time invariant plant and unity feedback (Fig. 1). This system has the following properties:

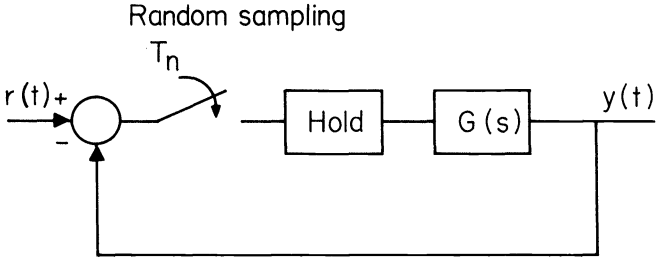


FIG. 1. *Linear randomly sampled system*

(a) *Sampling process*: The successive sampling intervals are stochastically independent, identically distributed, and are denoted by $T_0, T_1, T_2, \cdots, T_k, \cdots$. The sampling times are denoted by $t_0, t_1, t_2, \cdots, t_k, \cdots$ (Fig. 2). The different distributions and their practical motivation are described by Agniel [8]. Since the successive periods are identically distributed, the expectation of any function of T_k is independent of k.

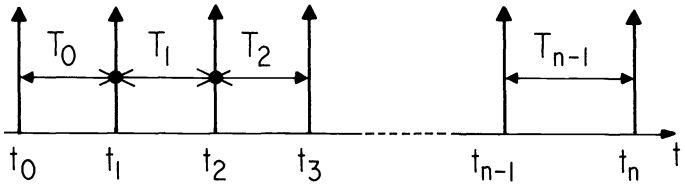


FIG. 2. *Sampling process*

(b) *Equation of the system*: Following Kalman and Bertram's approach [9], we have the system equations

$$\mathbf{X}(k+1) = [G(k) + \mathbf{h}(k)\mathbf{a}']\mathbf{X}(k) + \mathbf{h}(k)r(t_k),$$

$$y(k) = \mathbf{a}'\mathbf{X}(k), \qquad k = 0, 1, 2, 3, \cdots, \tag{1}$$

$$\mathbf{X}(0) = \mathbf{X}_0,$$

where $\mathbf{X}(k)$ is the state vector at time t_k,

$G(k) \triangleq G(T_k)$ is an $n \times n$ matrix continuous in T_k,

$\mathbf{h}(k) \triangleq \mathbf{h}(T_k)$ is an n-dimensional vector continuous in T_k,

$\mathbf{a}$ is an n-dimensional constant vector,

$r(t_k)$ is the input at time t_k.

2.2. Absolute stability of linear randomly sampled systems. Suppose the input $r(t) \equiv 0$ for all $t < 0$. Let

$$V_k = \mathbf{X}'(k)B\mathbf{X}(k),$$

where B is a symmetric positive definite matrix.

Taking V_k as a stochastic Lyapunov function and by a trivial application of Bucy's results [6], [7], the linear system described in § 2.1(b) is almost surely asymptotically stable and shows asymptotic stability of the second moment if there exist positive definite matrices B and Q such that

$$E[\{G(k) + \mathbf{h}(k)\mathbf{a}'\}'B\{G(k) + \mathbf{h}(k)\mathbf{a}'\}] - B = -Q. \tag{2}$$

It was shown by Agniel and Jury [10] (or our Appendix B) that this can be done provided the eigenvalues of

$$E[\{G(k) + \mathbf{h}(k)\mathbf{a}'\} \otimes \{G(k) + \mathbf{h}(k)\mathbf{a}'\}]$$

lie inside the unit circle, $\otimes$ denoting the first Krönecker product of matrices.

By using the same reasoning as in Kalman and Bertram [11], we see that condition (2) is equivalent to

$$E[\{G(k) + \mathbf{h}(k)\mathbf{a}'\}'B\{G(k) + \mathbf{h}(k)\mathbf{a}'\}] - \frac{1}{\eta}B \leqq 0, \tag{3}$$

where η is a scalar defined by

$$\frac{1}{\eta} = \lambda_{\min}[E\{[(G(k) + \mathbf{h}(k)\mathbf{a}')'B(G(k) + \mathbf{h}(k)\mathbf{a}')]B^{-1}\}], \tag{4}$$

where $\eta > 1$.

In conclusion, (2) implies for a null input that

$$E^{\mathscr{B}_k}V_{k+1} - V_k/\eta \leqq 0, \quad \eta > 1, \qquad k = 0, 1, 2, 3, \cdots, \tag{5}$$

where $E^{\mathscr{B}_k}(\cdot)$ denotes the conditional expectation given the σ-field $\mathscr{B}_k$,[1] which is the σ-field generated by $\mathbf{X}(1), \mathbf{X}(2), \cdots, \mathbf{X}(k)$. (See Bucy [6], [7] or Loève [12].)

2.3. Stability for a bounded input. Suppose the system is almost surely asymptotically stable and shows asymptotic stability of the second moment for a null input, i.e.,[2]

$$\sup_i |\lambda_i\{E[(G(k) + \mathbf{h}(k)a') \otimes (G(k) + \mathbf{h}(k)a')]\}| < 1. \tag{6}$$

Suppose the system is now driven by a bounded deterministic input.

Consider the same Lyapunov function $V_k = \mathbf{X}'(k)B\mathbf{X}(k)$. We have

$$E^{\mathscr{B}_k}V_{k+1} - \frac{1}{\eta}V_k = E^{\mathscr{B}_k}\mathbf{X}'(k+1)B\mathbf{X}(k+1) - \frac{1}{\eta}\mathbf{X}'(k)B\mathbf{X}(k). \tag{7}$$

Since the sampling intervals are independent, and by properties of conditional expectations (see Loève [12, p. 350]), we get (using (1))

$$\begin{aligned} E^{\mathscr{B}_k}V_{k+1} - \frac{1}{\eta}V_k &= \mathbf{X}'(k)E[(G(k) + \mathbf{h}(k)\mathbf{a}')'B(G(k) + \mathbf{h}(k)\mathbf{a}')]\mathbf{X}(k) \\ &\quad + 2E[\mathbf{h}'(k)B(G(k) + \mathbf{h}(k)\mathbf{a}')]\mathbf{X}(k)r(t_k) \\ &\quad + E[\mathbf{h}'(k)B\mathbf{h}(k)]r^2(t_k) - \frac{1}{\eta}\mathbf{X}'(k)B\mathbf{X}(k), \end{aligned} \tag{8}$$

which implies

$$E^{\mathscr{B}_k}V_{k+1} - \frac{1}{\eta}V_k - E[\mathbf{h}'(k)B\mathbf{h}(k)]r^2(t_k) = \mathbf{X}'(k)\theta\mathbf{X}(k) + 2\boldsymbol{\gamma}'\mathbf{X}(k)r(t_k) \tag{9}$$

with

$$\theta = E[(G(k) + \mathbf{h}(k)\mathbf{a}')'B(G(k) + \mathbf{h}(k)\mathbf{a}')] - \frac{1}{\eta}B, \tag{10}$$

$$\boldsymbol{\gamma}' = E[\mathbf{h}'(k)B(G(k) + \mathbf{h}(k)\mathbf{a}')]. \tag{11}$$

Equation (9) can be rewritten as

$$E^{\mathscr{B}_k}V_{k+1} - \frac{1}{\eta}V_k - [E\mathbf{h}'(l)B\mathbf{h}(k) + \rho]r^2(t_k) = [\mathbf{X}'(k), r(t_k)]\begin{bmatrix} \theta\boldsymbol{\gamma} \\ \boldsymbol{\gamma}' - \rho \end{bmatrix}\begin{bmatrix} \mathbf{X}(k) \\ r(t_k) \end{bmatrix}. \tag{12}$$

[1] In this case, $\mathscr{B}_k$ is the σ-field generated by $\mathbf{X}(k)$ only. The derivation remains valid for sampling periods which are not independent.

[2] Condition (6) is necessary and sufficient for absolute stability of the second moment [5]. However, it is only sufficient for a.s. absolute stability [5], [10].

Using the results of Gantmacher [13, pp. 46, 306], we see that the right-hand side of (12) is negative definite if

(i) θ is negative definite, which is always verified by our assumption (the system is absolutely asymptotically stable, i.e., inequality (3) is satisfied);

(ii) $\rho > -\gamma'\theta^{-1}\gamma > 0$.

Therefore,

$$E^{\mathscr{B}_k} V_{k+1} - \frac{1}{\eta} V_k - C_k < 0 \quad \text{a.s. for all } k \tag{13}$$

with $C_k = [E\{h'(k)B\mathbf{h}(k)\} + \rho]r^2(t_k) \geqq 0$, and since $|r(t_k)| \leqq M < \infty$, we have

$$E^{\mathscr{B}_k} V_{k+1} - \frac{1}{\eta} V_k - C < 0 \quad \text{a.s. for all } k \tag{14}$$

with $C = \sup_k C_k \leqq [E\{\mathbf{h}'(k)B\mathbf{h}(k)\} + \rho]M^2 < \infty$.

It is not necessary to compute θ^{-1} since a close bound for ρ can be found by writing

$$0 < \rho \leqq -[\lambda_{\min}(\theta)]^{-1}\gamma'\gamma.$$

The properties of systems satisfying inequality (14) will be studied in § 4 and the following theorem will be proved.

THEOREM. *For the system described in* § 2.1:

1. V_k *is almost surely bounded for all* k.
2. $E(V_k)$ *is bounded for all* k *and*

$$EV_k \leqq \sup\left[C\frac{\eta}{\eta - 1}, V_0\right].$$

3. $\|\mathbf{X}(k)\|$ *is almost surely bounded for all* k.
4. $E[\|X(k)\|^2]$ *is bounded for all* k *and*

$$E[\|\mathbf{X}(k)\|^2] \leqq [\min_i \lambda_i(B)]^{-1} \sup\left[C\frac{\eta}{\eta - 1}, \mathbf{X}_0'B\mathbf{X}_0\right].$$

It will now be shown that linear or nonlinear absolutely stable systems satisfy inequality (14) when the input is an uncorrelated noise with finite variance.

3. Randomly sampled systems with a noise as an input.

3.1. Description of the input. Randomly sampled systems where the input is uncorrelated noise with zero mean and finite variance will now be investigated. Such noise is different from white noise since the variance is finite, and then does not require infinite energy.

This input noise will be denoted by $r(t)$ and it satisfies the following conditions:

(i) $r(t)$ has zero mean, i.e., $E[r(t)] = 0$ for all t;

(ii) $r(t)$ is uncorrelated, i.e., $E[r(t)r(t + \tau)] = 0$ for all t and all $\tau > 0$;

(iii) $r(t)$ has finite variance, i.e.,

$$0 \leqq E[r^2(t)] \leqq C_1 \leqq \infty \quad \text{for all } t,$$

C_1 being a positive constant;

(iv) the input and the sampling process are independent.

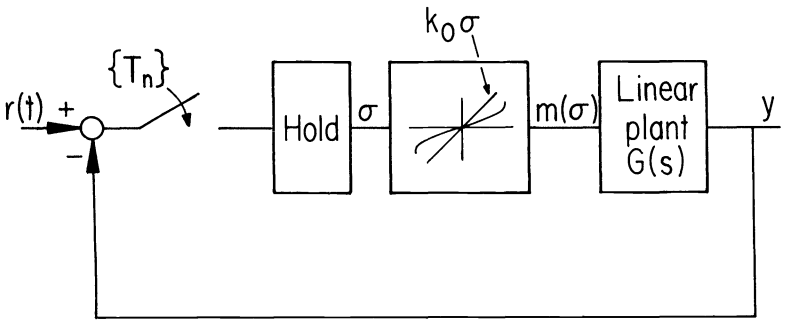


FIG. 3. *Nonlinear randomly sampled system*

CLAIM. *For linear or nonlinear randomly sampled systems shown to be almost surely asymptotically stable for a null input,* (14) *is verified when the input consists of the noise described in* § 3.1.

This claim will be proved for a particular configuration. Consider a closed loop sampled data system with a random sample and hold element followed by a time invariant nonlinearity and a plant with unity feedback (Fig. 3). The system satisfies the following assumptions:

(i) the sampling intervals are independent and identically distributed;

(ii) the nonlinearity is differentiable, of sector type and has bounded slope, i.e., if $m(\sigma)$ is the output of the nonlinearity for an input σ,

$$|dm/d\sigma| < K, \tag{15}$$

$$0 \leqq \sigma m(\sigma) \leqq k_0\sigma^2; \tag{16}$$

(iii) the input noise is described in § 3.1;

(iv) the linear plant is time invariant, i.e. (see for instance Kalman and Bertram [9]),

$$\mathbf{X}(k+1) = G(k)\mathbf{X}(k) + \mathbf{h}(k)m(\sigma_k) \tag{17}$$

with the same notation as in § 2.1, and

$$m(\sigma_k) = m[\mathbf{a}'\mathbf{X}(k) + r(t_k)]. \tag{18}$$

Consider the stochastic Lyapunov function $V_k = \mathbf{X}'(k)B\mathbf{X}(k)$, where B is a positive definite matrix.

If the system has been proven to be almost surely asymptotically stable for a null input, then there exists[3] a positive definite B such that (see Agniel and Jury [10])

$$E^{\mathscr{B}_k}V_{k+1} - \frac{1}{\eta}V_k \leqq 0 \quad \text{a.s. for all } k, \eta > 1, \tag{19}$$

[3] The necessity of the existence of a Lyapunov function can be proved by an analogy with the Massera theorem (see [11, p. 379]).

i.e.,

$$\mathbf{X}'(k)\left\{E[G'(k)BG(k)] - \frac{1}{\eta}B\right\}\mathbf{X}(k) + 2E[h'(k)BG(k)]\mathbf{X}(k)m(a'\mathbf{X}(k)) + E[\mathbf{h}'(k)B\mathbf{h}(k)]m^2(\mathbf{a}'\mathbf{X}(k)) \leqq 0 \quad \text{for all } k \quad \text{a.s.} \tag{20}$$

If the input $r(t)$ is as described in § 3.1, then

$$\begin{aligned} E^{\mathscr{B}_k}V_{k+1} = E^{\mathscr{B}_k}\{&[\mathbf{X}'(k)G'(k)BG(k)\mathbf{X}(k)] \\ &+ 2\mathbf{h}'(k)BG(k)\mathbf{X}(k)m[\mathbf{a}'\mathbf{X}(k) + r(t_k)] \\ &+ \mathbf{h}'(k)B\mathbf{h}(k)m^2[\mathbf{a}'\mathbf{X}(k) + r(t_k)]\}. \end{aligned} \tag{21}$$

The matrix $G'(k)BG(k)$, the vector $\mathbf{h}'(k)BG(k)$ and $\mathbf{h}'(k)B\mathbf{h}(k)$ are Borel functions of T_k; moreover, the σ-field induced by T_k and $\mathscr{B}_k$ are independent by assumption. By properties of conditional expectations (see Loève [12, pp. 349, 350]),

$$\begin{aligned} E^{\mathscr{B}_k}V_{k+1} = &\mathbf{X}'(k)E[G'(k)BG(k)]\mathbf{X}(k) \\ &+ 2E[\mathbf{h}'(k)BG(k)]\mathbf{X}(k)E^{\mathscr{B}_k}[m(\mathbf{a}'\mathbf{X}(k) + r(t_k))] \\ &+ E[\mathbf{h}'(k)B\mathbf{h}(k)]E^{\mathscr{B}_k}[m^2(\mathbf{a}'\mathbf{X}(k) + r(t_k))] \quad \text{a.s.} \end{aligned} \tag{22}$$

Since the nonlinearity is differentiable, we can apply MacLaurin's formula to get

$$m[\mathbf{a}'\mathbf{X}(k) + r(t_k)] = m[\mathbf{a}'\mathbf{X}(k)] + \dot{m}(\theta)r(t_k) \tag{23}$$

with $|\dot{m}| = |dm/d\sigma| \leqq K < \infty$ by assumption (from (15)) and $\theta \in [\mathbf{a}'\mathbf{X}(k), \mathbf{a}'\mathbf{X}(k) + r(t_k)]$.

Since the sequences $\{r(t_k)\}$ and $\{T_k\}$ are independent by assumption, and by using the properties of $r(t_k)$ and the fact that $m(\mathbf{a}'\mathbf{X}(k))$ is $\mathscr{B}_k$-measurable, we obtain (see Loève [12, pp. 348, 349, 350])

$$E^{\mathscr{B}_k}\{m^2[\mathbf{a}'\mathbf{X}(k) + r(t_k)]\} \leqq m^2[\mathbf{a}'\mathbf{X}(k)] + K^2C_1 \quad \text{a.s.}, \tag{24}$$

$$E^{\mathscr{B}_k}[m(\mathbf{a}'\mathbf{X}(k) + r(t_k))] = m(\mathbf{a}'\mathbf{X}(k)) \quad \text{a.s.}; \tag{25}$$

then

$$\begin{aligned} E^{\mathscr{B}_k}V_{k+1} - \frac{1}{\eta}V_k \leqq &\mathbf{X}'(k)\left\{E[G'(k)BG(k)] - \frac{1}{\eta}B\right\}\mathbf{X}(k) \\ &+ 2E[\mathbf{h}'(k)BG(k)]\mathbf{X}(k)m[\mathbf{a}'\mathbf{X}(k)] \\ &+ E[\mathbf{h}'(k)B\mathbf{h}(k)][m^2(\mathbf{a}'\mathbf{X}(k)) + K^2C_1] \quad \text{for all } k, \quad \text{a.s.}, \end{aligned} \tag{26}$$

and from (19),

$$E^{\mathscr{B}_k}V_{k+1} - \frac{1}{\eta}V_k - E[\mathbf{h}'(k)B\mathbf{h}(k)]K^2C_1 < 0 \quad \text{for all } k, \quad \text{a.s.} \tag{27}$$

We obtain an equation of type (14) with

$$C = E[\mathbf{h}'(k)B\mathbf{h}(k)]K^2C_1 > 0.$$

The same procedure can be applied successfully to a large class of system configurations and Lyapunov functions; in particular,

- linear randomly sampled systems,
- nonlinear randomly sampled systems with nonlinear element in the feedback loop,
- nonlinear randomly sampled systems with a stochastic Lyapunov function in the Luré form as considered by Agniel and Jury [10], Kushner and Tobias [15].

It can be shown that the constant C in inequality (14) is proportional to C_1 in most cases, and the physical interpretation of this remark will be illustrated in § 5.

For linear systems, this property remains true if the system has an output consisting of a bounded signal with a noise as described in § 3.1. Of course, this property is not true for nonlinear systems, but for some simple inputs, one can sometimes circumvent this difficulty by considering the deterministic signal as an initial condition.

It will be shown in the next section that for the class of systems considered which have been shown to satisfy equation (14):

1. The sequence V_k $(k = 0, 1, \cdots)$ is almost surely bounded;
2. EV_k is bounded for all k and $EV_k \leqq \sup [C(\eta/\eta - 1)), V_0]$;
3. $\|\mathbf{X}(k)\|$ remains almost surely bounded.

4. Properties of systems described by $E^{\mathscr{B}k}V_{k+1} - (1/\eta)V_k - C < 0$ a.s. $k = 0, 1, 2, \cdots, n$.

4.1. Properties of V_k.

THEOREM 1. *V_k is almost surely bounded.*

Proof. As it is, $\{V_k\}$ is not supermartingale. Consider $\{W_k\}$ defined by

$$W_k = V_k - C\frac{\eta}{\eta - 1}. \tag{28}$$

Obviously, W_k and V_k are measurable with respect to the same σ-field $\mathscr{B}_k$. Since conditional expectations are a.s. linear operators and since the conditional expectation of a constant is a.s. equal to this constant (see Loève [12, pp. 347, 348]), inequality (14) becomes

$$E^{\mathscr{B}k}W_{k+1} - \frac{1}{\eta}W_k < 0 \quad \text{a.s.}, \tag{29}$$

i.e., W_k is a supermartingale satisfying an equation similar to (5). It is deduced that

$$E(-W_0) \leqq E(-W_1) < \cdots \leqq E(-W_k) \leqq \cdots, \tag{30}$$

and recalling that V_k is positive for all k, then

$$E(-W_k) = E(W_k^- - W_k^+) = C\frac{\eta}{\eta - 1} - EV_k, \tag{31}$$

$$E(|W_k|) = E(W_k^+ + W_k^-) \leqq C\frac{\eta}{\eta - 1} + EV_k. \tag{32}$$

Summing up, one has

$$E(-W_k)^+ \leqq C\frac{\eta}{\eta - 1} < \infty \quad \text{since } \eta > 1. \tag{33}$$

The submartingale convergence theorem (see Appendix A) can now be applied to the submartingale $\{-W_k\}$ to obtain

$$-W_k = C\frac{\eta}{\eta - 1} - V_k \xrightarrow[k]{\text{a.s.}} W < \infty, \tag{34}$$

i.e.,

$$V_k \xrightarrow[k]{\text{a.s.}} V < \infty \quad \text{a.s.}, \tag{35}$$

and from (28) and (30), $V_k < \infty$ a.s. for all k.

THEOREM 2. *EV_k remains bounded for all k and*

$$EV_k \leqq \sup\left[C\frac{\eta}{\eta - 1}, V_0\right].$$

Proof. Inequality (14) implies (see Loève [12, p. 341])

$$\eta EV_k < EV_{k-1} + \eta C, \tag{36}$$

and by induction,

$$EV_k < \frac{V_0}{\eta^k} + C\frac{\eta}{\eta - 1}\left(1 - \frac{1}{\eta^{k+1}}\right). \tag{37}$$

By a straightforward optimization procedure, (37) implies

$$EV_k \leqq \sup\left[C\frac{\eta}{\eta - 1}, V_0\right] \quad \text{for all } k. \tag{38}$$

4.2. Properties of $\mathbf{X}(k)$.

THEOREM 3. *$\|\mathbf{X}(k)\|$ is almost surely bounded for all k.*

The proof is trivial since it is known that V_k is a stochastic Lyapunov function if there exist nondecreasing functions $\alpha(\cdot)$ and $\beta(\cdot)$ such that $\alpha(0) = \beta(0) = 0$,

$$\alpha(\|\mathbf{X}(k)\|) \leqq V_k \leqq \beta(\|\mathbf{X}(k)\|). \tag{39}$$

If $V_k = \mathbf{X}'(k)B\mathbf{X}(k)$, then $(\mathbf{X}'(k)B\mathbf{X}(k))^{1/2}$ is a norm for $\mathbf{X}(k)$, and then $V_k = \|\mathbf{X}(k)\|^2 < \infty$ for all k, a.s.

THEOREM 4. *$E\|\mathbf{X}(k)\|$ is bounded for all k.*

In the general case, since V_k is positive definite, an $\alpha(\cdot)$ which is convex can be found such that

$$\alpha(\|\mathbf{X}(k)\|) \leqq V_k \quad \text{for all } k. \tag{40}$$

By applying Jensen's inequality,

$$\alpha(E\|\mathbf{X}(k)\|) \leqq EV_k \leqq \sup\left[V_0, C\frac{\eta}{\eta - 1}\right] < \infty, \tag{41}$$

and the procedure is as in Theorem 3.

However, in most cases of interest, V_k is a quadratic form or a Luré form (see Agniel and Jury [10]); for a quadratic form, if $V = \mathbf{X}'B\mathbf{X}$,

$$E[\|\mathbf{X}(k)\|^2] \leqq [\min_i \lambda_i(B)]^{-1} \sup \left[C\frac{\eta}{\eta - 1}, \mathbf{X}_0'B\mathbf{X}_0 \right], \tag{42}$$

i.e., a bound has been found for the *second moment* of $\|\mathbf{X}(k)\|$. The same kind of result can be obtained for a Luré form.

5. Conclusions.

5.1. Linear randomly sampled systems. For periodically sampled systems, it is well known that if (see (1))

$$|\lambda_i\{G(T)\}| < 1, \tag{43}$$

then:

(i) the unforced system is asymptotically stable in the large,

(ii) there exists a unique solution to the matrix equation

$$G'(T)BG(T) - B = -Q, \tag{44}$$

i.e., given Q positive definite, there exists a unique positive definite B such that

(iii) the system is bounded input–bounded output stable.

For randomly sampled systems with independent identically distributed sampling intervals, it is shown that if

$$|\lambda_i\{E[G(k) \otimes G(k)]\}| < 1, \tag{45}$$

then:

(i) the unforced system is almost surely asymptotically stable and also shows stability of the second moment (see Kalman [3], Bharucha [5]);

(ii) the matrix equation

$$E[G'(k)BG(k)] - B = -Q \tag{46}$$

has a unique solution and Q positive definite implies B positive definite (see Agniel and Jury [10], or our Appendix B);

(iii) the system is almost surely bounded input–bounded output stable and also shows boundedness of the second moment for a bounded input (see § 2 and § 4).

In other words, with respect to almost sure stability and stability of the first and second moments, condition (45) is similar to condition (43) for periodically sampled systems.

It was shown by Agniel and Jury [10] that condition (45) is a necessary assumption to show the stability of nonlinear randomly sampled systems. It will be proved in a forthcoming paper that this condition is of interest to show almost sure stability by a Popov-type method.

5.2. Nonlinear randomly sampled systems. It was shown that provided the autonomous system was proven to be almost surely stable by a Lyapunov function approach, a noise with zero mean and finite energy introduced bounded disturbances.

The bound for EV_k is $\sup [C(\eta/(\eta - 1)), V_0]$. This introduces an interesting physical interpretation, i.e., the bound is given by the major cause of disturbances,

V_0 representing the initial conditions and $C(\eta/(\eta - 1))$ the size of the input. Moreover, it is not difficult to show (see Agniel [8]) that the bigger η is, the faster the system converges to its equilibrium for a null input and the smaller is the parameter domain of stability. Here, the bound $C(\eta/(\eta - 1))$ is continuously decreasing when we increase $\eta > 1$, i.e., when we make the system more and more stable.

5.3. Related work. In his book, Kushner [14] gives a comprehensive study of equations of type (14) for finite time stability and excursion times.

Appendix A.

SUBMARTINGALE CONVERGENCE THEOREM (see Loève [12, p. 393]). *Let the random variables X_n form a submartingale sequence.*

(i) *If* $\sup EX_n^+ < \infty$, *then* $X_n \xrightarrow{\text{a.s.}} X < \infty$ *with*

$$EX \leqq \sup EX_n^+,$$

$$E|X| \leqq \sup E|X_n|.$$

(ii) $X_n \xrightarrow{r} X$ *where* $r \geqq 1$ *if and only if the* $|X_n|^r$ *are uniformly integrable, and then* $X_n \xrightarrow{\text{a.s.}} X$.

Appendix B. We wish to solve

$$E[G'(k)BG(k)] - B = -Q. \tag{B.1}$$

Problem. Given a positive definite matrix Q, under what conditions can we find a positive definite matrix B satisfying (B.1)?

Existence and uniqueness. Although $G(k)$ is not necessarily symmetric, $G'(k)BG(k)$ is symmetric and it does not affect the generality of the proof to assume B and Q symmetric (see Gantmacher [13, p. 294]). If b_{ij} and q_{ij} are the general elements of B and Q respectively, (B.1) can be rewritten as

$$[(E[G(k) \otimes G(k)]) - I] \begin{bmatrix} b_{11} \\ b_{12} \\ \vdots \\ b_{1n} \\ b_{22} \\ b_{23} \\ \vdots \\ b_{2n} \\ b_{33} \\ \vdots \\ b_{3n} \\ \vdots \\ b_{nn} \end{bmatrix} = - \begin{bmatrix} q_{11} \\ q_{12} \\ \vdots \\ q_{1n} \\ q_{22} \\ q_{23} \\ \vdots \\ q_{2n} \\ q_{33} \\ \vdots \\ q_{3n} \\ \vdots \\ q_{nn} \end{bmatrix}, \tag{B.2}$$

where $G(k) \otimes G(k)$ is the first Krönecker product of $G(k)$ by itself. From (B.2) we see that given the q_{ij}'s we can determine the b_{ij}'s provided

$$\lambda_i[E\{G(k) \otimes G(k)\}] \neq 1. \tag{B.3}$$

Since B and Q are taken symmetric, we can determine B completely from Q.

Condition for a positive definite B. Consider the linear system described by

$$\begin{aligned} &\mathbf{X}(k+1) = G(k)\mathbf{X}(k), \\ &\mathbf{X}(0) = \mathbf{X}_0, \qquad\qquad k = 0, 1, 2, 3, \cdots, \end{aligned} \tag{B.4}$$

where $\mathbf{X}(k)$ is an n-dimensional vector with components $x_1(k), x_2(k), \cdots, x_n(k)$ and $G(k)$ is an $n \times n$ matrix continuous in T_k. If B is positive definite, condition (B.1) implies that the system (B.4) shows stability of the second moment, and it is necessary that (see Kalman [3], Bharucha [5])

$$\sup_i |\lambda_i\{E[G(k) \otimes G(k)]\}| < 1; \tag{B.5}$$

then,

$$E[x_i(k)x_j(k)] \to 0 \quad \text{as } k \to \infty \quad \text{for all } i, j = 1, 2, \cdots, n. \tag{B.6}$$

Conversely, assume that (B.5) is satisfied and B is not positive definite. Then there exists a state $\mathbf{X}_0$ such that $\mathbf{X}_0' B \mathbf{X}_0 < 0$. This implies that

$$E[\mathbf{X}'(1)B\mathbf{X}(1)] = \mathbf{X}_0' B \mathbf{X}_0 - \mathbf{X}_0' Q \mathbf{X}_0 < 0$$

since Q is positive definite.

By induction,

$$E[\mathbf{X}'(k)B\mathbf{X}(k)] < \mathbf{X}_0' B \mathbf{X}_0 < 0 \quad \text{for all } k = 0, 1, 2, \cdots,$$

which contradicts (B.6). Therefore B must be positive definite.

In conclusion, given a positive definite matrix Q, there exists a unique positive definite matrix B satisfying (B.1) provided (B.5) is satisfied.

REFERENCES

[1] G. A. Bekey, J. M. Biddle and A. J. Jacobson, *The effect of a random sampling interval on a sampled-data model of the human operator*, Proc. 3rd Conference on Manual Control, Rep. SP-144, NASA, 1967.

[2] J. D. Forester, *A stochastic revised sampled data model of the human operator*, S.M. Thesis Rep. MVT-68-2, Manned Vehicle Laboratory, Center for Space Research, MIT, Cambridge, Mass., 1968.

[3] R. E. Kalman, *Analysis and synthesis of linear dynamical systems operating on randomly sampled data*, Doctoral thesis, Columbia University, New York, 1957.

[4] A. R. Bergen, *Stability of systems with randomly time varying parameters*, IRE Trans. Automatic Control, AC-5 (1960), pp. 265–269.

[5] B. H. Bharucha, *On the stability of randomly varying systems*, Doctoral thesis, University of California, Berkeley, 1961.

[6] R. S. Bucy, *Stability and positive supermartingales*, J. Differential Equations, 1 (1965), pp. 151–155.

[7] R. S. Bucy and P. D. Joseph, *Filtering for Stochastic Processes with Application to Guidance*, Interscience, New York, 1968.

[8] R. G. Agniel, *Stability of randomly sampled systems*, Doctoral thesis, University of California, Berkeley, 1969.

[9] R. E. KALMAN AND J. E. BERTRAM, *A unified approach to the theory of sampling systems*, J. Franklin Institute, 267 (1959), pp. 405–436.
[10] R. G. AGNIEL AND E. I. JURY, *Stability of nonlinear randomly sampled systems*, Allerton Conference, Urbana, Illinois, 1969, pp. 710–719.
[11] R. E. KALMAN AND J. E. BERTRAM, *Control system analysis and design via the second method of Lyapunov. I: Continuous time systems; II: Discrete time systems*, Trans. ASME Ser. D. J. Basic Engrg., 82 (1960), pp. 371–400.
[12] M. LOÈVE, *Probability Theory*, 3rd ed., Van Nostrand, Princeton, N.J., 1963.
[13] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1959.
[14] H. J. KUSHNER, *Stochastic Stability and Control*, Academic Press, New York, 1967.
[15] H. J. KUSHNER AND L. TOBIAS, *On the stability of randomly sampled systems*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 219–324.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

THE SOLUTION OF A QUADRATIC PROGRAMMING PROBLEM USING SYSTEMATIC OVERRELAXATION*

COLIN W. CRYER†

Abstract. Let **A** be a real symmetric positive definite $n \times n$ matrix and **b** a real column n-vector. We consider the following problem: Find real column n-vectors **x** and **y** such that

$$\mathbf{Ax} = \mathbf{b} + \mathbf{y},$$

$$\mathbf{x}^T\mathbf{y} = 0, \quad \mathbf{x} \geqq 0, \quad \mathbf{y} \geqq 0.$$

Problems of this type occur when the method of Christopherson is used to solve free boundary problems for journal bearings. In such cases, **A** is a "finite difference" matrix.

We present a method for solving the above problem which is a modification of systematic overrelaxation. This method is particularly suitable when **A** is a finite difference matrix.

1. Introduction. Let $\mathbf{A} = (a_{ij})$ be a real symmetric positive definite $n \times n$ matrix and $\mathbf{b} = (b_i)$ a real column n-vector. We shall be concerned with the following problem.

Problem 1. Find real column n-vectors $\mathbf{x} = (x_i)$ and $\mathbf{y} = (y_i)$ such that

$$\mathbf{Ax} - \mathbf{y} = \mathbf{b}, \tag{1.1}$$

$$\mathbf{x}^T\mathbf{y} = 0, \tag{1.2}$$

$$\mathbf{x} \geqq 0, \qquad \mathbf{y} \geqq 0. \tag{1.3}$$

It is known that Problem 1 is equivalent to a quadratic programming problem, Problem 2 (see § 2). Both Problems 1 and 2 have been extensively studied from the viewpoint of linear and quadratic programming (Cottle and Dantzig [1], Hadley [5, p. 212], and Lemke [6]) and there are many methods available for solving these problems.

Our interest in Problem 1 arose because problems of this type occur when the method of Christopherson is used to solve free boundary problems for journal bearings (Cryer [3]). In such cases, Problem 1 has certain features which are unusual in nonlinear programming problems:

(i) **A** is a large matrix, perhaps a 10,000 × 10,000 matrix.

(ii) **A** is a "finite difference" matrix. Typically, each row of **A** will have no more than five nonzero elements. However, $\mathbf{A}^{-1}$ is a full matrix.

(iii) Because of the physical significance of the solution vector **x**, most of the components x_i may be expected to be positive.

When these features are present, the conventional methods for solving Problems 1 and 2 have substantial disadvantages.

In § 3, we introduce a method for solving Problem 1 which is particularly suitable when **A** is a "finite difference" matrix, since the method is a modified version of S.O.R. (systematic overrelaxation). In § 3, we prove that this method

* Received by the editors February 9, 1970, and in revised form November 30, 1970.

† Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706. This work was supported by the Office of Naval Research under Contract N00014-67-A-0004.

converges, and in § 4, we study how the rate of convergence depends upon the relaxation parameter.

2. Existence and uniqueness of solution. It is convenient to introduce the following quadratic programming problem.

Problem 2. Find a column n-vector $\mathbf{x}$ which maximizes

$$f(\mathbf{x}) = \mathbf{b}^T\mathbf{x} - (\mathbf{x}^T\mathbf{A}\mathbf{x})/2, \tag{2.1}$$

subject to the constraints

$$\mathbf{x} \geqq 0. \tag{2.2}$$

THEOREM 2.1. *Problems* 1 *and* 2 *are equivalent: if* $\{\mathbf{x}, \mathbf{y}\}$ *is a solution of Problem* 1, *then* $\mathbf{x}$ *is a solution of Problem* 2; *if* $\mathbf{x}$ *is a solution of Problem* 2, *then* $\{\mathbf{x}, \mathbf{A}\mathbf{x} - \mathbf{b}\}$ *is a solution of Problem* 1.

There exists a unique solution to Problem 2 (*and hence to Problem* 1).

Proof. Since $\mathbf{A}$ is positive definite, the equivalence of Problems 1 and 2 follows from the Kuhn–Tucker theory (Hadley [5, pp. 212–214]).

Since $\mathbf{A}$ is positive definite, $f(\mathbf{x})$ is strictly concave (Hadley [5, p. 213]). Hence, since $\mathbf{x} = 0$ is a "feasible" solution of Problem 2, there exists a unique solution to Problem 2. The proof of the theorem is therefore complete.

3. Application of S.O.R. We study the following algorithm for solving Problem 1.

ALGORITHM 1. Choose a column n-vector $\mathbf{x}^{(0)} = (x_i^{(0)})$, where $\mathbf{x}^{(0)} \geqq 0$. Choose a *relaxation parameter* ω, where $0 < \omega < 2$.

Generate a sequence of column n-vectors $\mathbf{x}^{(k)} = (x_i^{(k)})$, $\mathbf{r}^{(k)} = (r_i^{(k)})$, $\mathbf{y}^{(k)} = (y_i^{(k)})$, $k = 1, 2, \cdots$, using the equations,

$$r_i^{(k+1)} = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij}x_j^{(k)}, \tag{3.1}$$

$$x_i^{(k+1)} = \max\{0, x_i^{(k)} + \omega r_i^{(k+1)}/a_{ii}\}, \tag{3.2}$$

$$y_i^{(k+1)} = -r_i^{(k+1)} + a_{ii}(x_i^{(k+1)} - x_i^{(k)}). \tag{3.3}$$

(We remind the reader that we have assumed that $\mathbf{A}$ is positive definite so that $a_{ii} > 0$ for $1 \leqq i \leqq n$.)

Algorithm 1 is a generalization of methods used by Christopherson [2] and Gnanadoss and Osborne [4]; a brief account of the history of the algorithm is given in Cryer [3].

Algorithm 1 can be interpreted in two ways. On the one hand, Algorithm 1 consists of applying S.O.R. to the equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ with the proviso that the vectors $\mathbf{x}^{(k)}$ should be nonnegative. On the other hand, as will be seen in the proof of Theorem 3.1, $f(\mathbf{x}^{(k+1)}) \geqq f(\mathbf{x}^{(k)})$, so that Algorithm 1 is a method for maximizing $f(\mathbf{x})$ subject to the restraint that $\mathbf{x} \geqq 0$. Of course, it is not surprising that two interpretations of Algorithm 1 exist, since it has been known for a long time (Temple [10]) that there is a connection between relaxation methods and the minimization of quadratic forms.

THEOREM 3.1. *Let* $\mathbf{x}^{(k)}$ *and* $\mathbf{y}^{(k)}$ *be generated by Algorithm* 1. *Then* $\mathbf{x}^{(k)} \to \mathbf{x}$ *and* $\mathbf{y}^{(k)} \to \mathbf{y}$, *where* $\{\mathbf{x}, \mathbf{y}\}$ *is the solution of Problem* 1.

Proof. The method of proof is similar to that used by Schechter [8], [9].

For any column n-vector $\mathbf{u}$ let

$$G(\mathbf{u}) = -2f(\mathbf{u}) = \mathbf{u}^T\mathbf{A}\mathbf{u} - 2\mathbf{u}^T\mathbf{b}. \tag{3.4}$$

Then, direct computation shows that if $\mathbf{u}$ and $\mathbf{v}$ are column n-vectors,

$$G(\mathbf{u}) - G(\mathbf{v}) = (\mathbf{u} - \mathbf{v})^T\mathbf{A}(\mathbf{u} - \mathbf{v}) + 2(\mathbf{u} - \mathbf{v})^T(\mathbf{A}\mathbf{v} - \mathbf{b}). \tag{3.5}$$

It is convenient to introduce the vectors $\mathbf{x}^{(k,l)} = (x_i^{(k,l)})$, where

$$x_i^{(k+1,l)} = \begin{cases} x_i^{(k+l)} & \text{if } 1 \leqq i \leqq l, \\ x_i^{(k)} & \text{if } l < i \leqq n \end{cases} \tag{3.6}$$

for $k \geqq 0$ and $0 \leqq l \leqq n$. Then,

$$\mathbf{x}^{(k+1,0)} = \mathbf{x}^{(k)}, \qquad \mathbf{x}^{(k+1,n)} = \mathbf{x}^{(k+1)}, \tag{3.7}$$

and, from (3.1),

$$r_i^{(k+1)} = [\mathbf{b} - \mathbf{A}\mathbf{x}^{(k+1,i-1)}]_i. \tag{3.8}$$

Let

$$\omega_{k+1,i} = \begin{cases} [x_i^{(k+1)} - x_i^{(k)}]a_{ii}/r_i^{(k+1)} & \text{if } r_i^{(k+1)} \neq 0, \\ \omega & \text{if } r_i^{(k+1)} = 0. \end{cases} \tag{3.9}$$

Then, noting (3.2),

$$x_i^{(k+1)} = x_i^{(k)} + \omega_{k+1,i}r_i^{(k+1)}/a_{ii}, \tag{3.10}$$

and

$$0 \leqq \omega_{k+1,i} \leqq \omega. \tag{3.11}$$

Using (3.5), (3.6), (3.8) and (3.10), we find that

$$\begin{aligned} G(\mathbf{x}^{(k+1,i)}) - G(\mathbf{x}^{(k+1,i-1)}) &= [\mathbf{x}^{(k+1,i)} - \mathbf{x}^{(k+1,i-1)}]^T\mathbf{A}[\mathbf{x}^{(k+1,i)} - \mathbf{x}^{(k+1,i-1)}] \\ &\quad + 2[\mathbf{x}^{(k+1,i)} - \mathbf{x}^{(k+1,i-1)}]^T[\mathbf{A}\mathbf{x}^{(k+1,i)} - \mathbf{b}], \end{aligned}$$

and

$$\begin{aligned} G(\mathbf{x}^{(k+1,i)}) - G(\mathbf{x}^{(k+1,i-1)}) &= a_{ii}[x_i^{(k+1)} - x_i^{(k)}]^2 - 2[x_i^{(k+1)} - x_i^{(k)}]r_i^{(k+1)} \\ &= -\omega_{k+1,i}(2 - \omega_{k+1,i})[r_i^{(k+1)}]^2/a_{ii}. \end{aligned} \tag{3.12}$$

Remembering that $0 < \omega < 2$, it follows from (3.11) and (3.12) that $G(\mathbf{x}^{(k+1,i)}) \leqq G(\mathbf{x}^{(k+1,i-1)})$. Therefore, the sequence $\{G(\mathbf{x}^{(k,i)})\}$ is monotone decreasing. But $\mathbf{A}$ is positive definite so that $G(\mathbf{u})$ is strictly convex and hence bounded below.

Consequently, there is a constant, G_∞ say, such that

$$G(\mathbf{x}^{(k,l)}) \downarrow G_\infty. \tag{3.13}$$

Next, we prove that

$$|x_i^{(k+1)} - x_i^{(k)}| \leqq [a(-1 + 2/\omega)]^{1/2}[G(\mathbf{x}^{(k+1,i-1)}) - G(\mathbf{x}^{(k+1,i)})]^{1/2}, \tag{3.14}$$

where

$$a = \min_i a_{ii}. \tag{3.15}$$

If $x_i^{(k+1)} = x_i^{(k)}$, then (3.14) is trivially true, so that we need only consider the case when $x_i^{(k+1)} \neq x_i^{(k)}$. But then, from (3.10), $\omega_{k+1,i} \neq 0$, so that, from (3.10), (3.11), (3.12) and (3.15),

$$\begin{aligned} G(\mathbf{x}^{(k+1,i-1)}) - G(\mathbf{x}^{(k+1,i)}) &= (-1 + 2/\omega_{k+1,i})a_{ii}[x_i^{(k+1)} - x_i^{(k)}]^2 \\ &\geqq (-1 + 2/\omega)a[x_i^{(k+1)} - x_i^{(k)}]^2. \end{aligned} \tag{3.16}$$

Inequality (3.14) follows immediately from (3.16).

Noting (3.13), it follows from (3.14) that

$$x_i^{(k+1)} - x_i^{(k)} \to 0 \quad \text{as} \quad k \to \infty, \qquad 1 \leqq i \leqq n. \tag{3.17}$$

Now, let $\mathbf{x}$ be any limit point of the sequence $\{\mathbf{x}^{(k)}\}$. Then there is an increasing sequence of integers $\{k_p\}$, $p = 1, 2, \cdots$, such that

$$\mathbf{x}^{(k_p)} \to \mathbf{x} \quad \text{as} \quad p \to \infty. \tag{3.18}$$

From (3.1), (3.3) and (3.18), we have that, as $p \to \infty$,

$$\begin{aligned} \mathbf{r}^{(k_p)} &\to \mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}, \\ \mathbf{y}^{(k_p)} &\to \mathbf{y} = -\mathbf{r}. \end{aligned} \tag{3.19}$$

We assert that

$$\mathbf{x} \geqq 0, \qquad \mathbf{r} \leqq 0. \tag{3.20}$$

That $\mathbf{x} \geqq 0$ follows immediately from the fact that $\mathbf{x}^{(k)} \geqq 0$ for all k. To prove that $\mathbf{r} \leqq 0$, suppose that this is not the case. Then there is an $\varepsilon > 0$ and integers i_0 and k_0' such that $r_{i_0}^{(k_p)} \geqq \varepsilon$ for $k_p \geqq k_0'$. Hence, from (3.2),

$$x_{i_0}^{(k_p)} - x_{i_0}^{(k_p-1)} \geqq \varepsilon\omega/a_{i_0 i_0}$$

for $k_p \geqq k_0'$. But this contradicts (3.17).

Next, we show that

$$\mathbf{r}^T\mathbf{x} = 0. \tag{3.21}$$

Suppose that this is not the case. Then, noting (3.20), we see that there is an $\varepsilon > 0$ and integers i_0 and k' such that $r_{i_0}^{(k_p)} \leqq -\varepsilon$ and $x_{i_0}^{(k_p)} \geqq \varepsilon$ for $k_p \geqq k'$. It follows from (3.2) that if $k_p \geqq k'$, then $x_{i_0}^{(k_p-1)} \geqq x_{i_0}^{(k_p)}$ and

$$|x_{i_0}^{(k_p)} - x_{i_0}^{(k_p-1)}| \geqq \omega\varepsilon/a_{i_0 i_0}.$$

But this contradicts (3.17).

From (3.19), (3.20) and (3.21), it follows that $\{\mathbf{x}, \mathbf{y}\}$ satisfies (1.1) through (1.3) and is the (unique) solution of Problem 1.

To complete the proof of the theorem we must show that the sequence $\{\mathbf{x}^{(k)}\}$ has at least one limit point. But this is a consequence of the fact that (see (3.13)) $\mathbf{x}^{(k)} \in R$ for all k, where R is the compact set

$$R = \{\mathbf{x}; G(\mathbf{x}) \leqq G(\mathbf{x}^{(0)})\}.$$

4. Determination of the optimum relaxation parameter. It is natural to ask how the convergence of Algorithm 1 depends upon ω, and whether there is a value of ω for which the rate of convergence is maximized. In this section, we partially answer these questions.

Since we make use of the theory of S.O.R., we first summarize this theory.

Let

$$\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}, \tag{4.1}$$

and

$$\tilde{\mathbf{A}} = \tilde{\mathbf{D}} - \tilde{\mathbf{E}} - \tilde{\mathbf{F}}, \tag{4.2}$$

where $\tilde{\mathbf{x}}$ is a column m-vector, $\tilde{\mathbf{D}}$ is a diagonal $m \times m$ matrix, and $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{F}}$ are respectively strictly upper and lower triangular $m \times m$ matrices. Let

$$\mathfrak{L}_\omega(\tilde{\mathbf{A}}) = (\tilde{\mathbf{D}} - \omega\tilde{\mathbf{E}})^{-1}\{(1 - \omega)\tilde{\mathbf{D}} + \omega\tilde{\mathbf{F}}\}. \tag{4.3}$$

For a given relaxation parameter ω and initial guess $\tilde{\mathbf{x}}^{(0)}$, let $\tilde{\mathbf{x}}^{(k)}$, $k = 1, 2, \cdots$, denote the iterates generated by S.O.R. applied to (4.1). Let

$$\tilde{\mathbf{e}}^{(k)} = \tilde{\mathbf{x}}^{(k)} - \tilde{\mathbf{x}}. \tag{4.4}$$

Then (Varga [11, p. 59]),

$$\tilde{\mathbf{e}}^{(k+1)} = \mathfrak{L}_\omega(\tilde{\mathbf{A}})\tilde{\mathbf{e}}^{(k)}. \tag{4.5}$$

From (4.5) it can be seen that $\tilde{\mathbf{e}}^{(k)}$ depends upon $\tilde{\mathbf{A}}$, $\tilde{\mathbf{e}}^{(0)}$ and ω. The *asymptotic rate of convergence* corresponding to $\tilde{\mathbf{A}}$ and ω is (Varga [11, p. 67])

$$R_\infty[\mathfrak{L}_\omega(\tilde{\mathbf{A}})] = -\log[\rho[\mathfrak{L}_\omega(\tilde{\mathbf{A}})]], \tag{4.6}$$

where $\rho[\mathfrak{L}_\omega(\tilde{\mathbf{A}})]$ denotes the spectral radius of $\mathfrak{L}_\omega(\tilde{\mathbf{A}})$. Equivalently,

$$R_\infty[\mathfrak{L}_\omega(\tilde{\mathbf{A}})] = -\log[\sup_{\tilde{\mathbf{e}}^{(0)}} \limsup_{k\to\infty} \|\tilde{\mathbf{e}}^{(k)}\|^{1/k}]. \tag{4.7}$$

The optimum relaxation parameter, $\omega_b = \omega_b(\tilde{\mathbf{A}})$, is defined by means of the relation (Varga [11, p. 109])

$$R_\infty[\mathfrak{L}_{\omega_b}(\tilde{\mathbf{A}})] = \max_{0<\omega<2} R_\infty[\mathfrak{L}_\omega(\tilde{\mathbf{A}})]. \tag{4.8}$$

For certain classes of matrices $\tilde{\mathbf{A}}$, notably 2-cyclic consistently ordered matrices, $\omega_b(\tilde{\mathbf{A}})$ is known in terms of the eigenvalues of the Jacobi matrix corresponding to $\tilde{\mathbf{A}}$ (Varga [11, p. 110]).

Next, we introduce some notation. We set

$$Z = \{1, 2, \cdots, n\}. \tag{4.9}$$

Let $T \subset Z$, $\mathbf{B} = (B_{ij})$ be an $n \times n$ matrix and $\mathbf{z} = (z_i)$ be an n-vector. Then $|T|$ denotes the number of elements of T; $\mathbf{B}(T)$ is the $|T| \times |T|$ submatrix of $\mathbf{B}$ obtained

by deleting those elements B_{ij} for which $i \notin T$ or $j \notin T$; and $\mathbf{z}(T)$ is the $|T| \times 1$ subvector of $\mathbf{z}$ obtained by deleting those elements z_i for which $i \notin T$. Finally,

$$Z(\mathbf{z}) = \{i \in Z;\ z_i \neq 0\}. \tag{4.10}$$

We are now ready to consider Algorithm 1. Let $\{\mathbf{x}, \mathbf{y}\}$ be the solution of Problem 1, and let $\{\mathbf{x}^{(k)}, \mathbf{y}^{(k)}\}$ be generated using Algorithm 1. We set

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}. \tag{4.11}$$

From (3.1) through (3.3) and (4.11) we see that $\mathbf{e}^{(k)}$ depends upon $\mathbf{A}, \mathbf{b}, \mathbf{e}^{(0)}$, and ω. By Theorem 3.1, $\mathbf{e}^{(k)} \to 0$ as $k \to \infty$ for any $\mathbf{e}^{(0)}$.

Corresponding to (4.7) and following Ortega and Rockoff [7] we define the asymptotic rate of convergence corresponding to $\mathbf{A}$, $\mathbf{b}$, and ω, to be

$$R(\mathbf{A}, \mathbf{b}, \omega) = -\log \{\sup_{\mathbf{e}^{(0)}} \limsup_{k \to \infty} \|\mathbf{e}^{(k)}\|^{1/k}\}. \tag{4.12}$$

LEMMA 4.1. *Given $\mathbf{x}^{(0)}$, there is an integer k_0 such that for $k \geqq k_0$,*

$$\begin{aligned} &x_i^{(k)} > 0 && \text{if } i \in X, \\ &x_i^{(k)} = 0 \quad \text{and} \quad y_i^{(k)} > 0 && \text{if } i \in Y, \end{aligned} \tag{4.13}$$

where $X = Z(\mathbf{x})$ and $Y = Z(\mathbf{y})$.

Proof. Let $\mathbf{x}^{(0)}$ be given. Let $\mathbf{r}^{(k)}$ be as in Algorithm 1. Since $\mathbf{x}^{(k)} \to \mathbf{x}$ and $\mathbf{r}^{(k)} \to -\mathbf{y}$, it follows that there is an $\varepsilon > 0$ and an integer k_1 such that if $k \geqq k_1$, then

$$\begin{aligned} &x_i^{(k)} > 0 && \text{if } i \in X, \\ &y_i^{(k)} > 0 \quad \text{and} \quad r_i^{(k)} < -\varepsilon && \text{if } i \in Y. \end{aligned}$$

Noting (3.2), we see that (4.13) holds if $k_0 \geqq k_1 + 1 + [\max_i a_{ii} x_i^{(k_1)}]/(\varepsilon\omega)$.

THEOREM 4.2. *Let $\mathbf{A}$ and $\mathbf{b}$ be such that*

$$|x_i| + |y_i| > 0, \qquad 1 \leqq i \leqq n, \tag{4.14}$$

where $\{\mathbf{x}, \mathbf{y}\}$ is the solution of Problem 1. Let $X = Z(\mathbf{x})$.

Then,

$$R(\mathbf{A}, \mathbf{b}, \omega) = \begin{cases} \infty & \text{if } X \text{ is empty}, \\ R_\infty(\mathfrak{L}_\omega[\mathbf{A}(X)]) & \text{otherwise}. \end{cases} \tag{4.15}$$

Proof. Let $\mathbf{x}^{(0)}$ be given. Then it follows from Lemma 4.1, (3.1), (3.2) and (4.14), that, for $k \geqq k_0$,

$$\begin{aligned} \mathbf{e}^{(k+1)}(X) &= \mathfrak{L}_\omega(\mathbf{A}(X))\mathbf{e}^{(k)}(X), \\ \mathbf{e}^{(k+1)}(Y) &= 0. \end{aligned}$$

The theorem follows from (4.7) and (4.12).

Condition (4.14) is satisfied by "almost all" $\mathbf{A}$ and $\mathbf{b}$, and the following theorem covers an important subclass of the remaining problems.

THEOREM 4.3. *Let*

$$0 < \omega \leqq 1 \tag{4.16}$$

and

$$a_{ij} \leqq 0 \qquad \text{for } i \neq j. \tag{4.17}$$

Then, $R[\mathbf{A}, \mathbf{b}, \omega] \leqq R_\infty[\mathfrak{L}_\omega(\mathbf{A}[T])]$, *where* $T = Z - Z(\mathbf{y})$.

Proof. Let $\mathbf{x}^{(0)}$ be given. Using an idea due to Gnanadoss and Osborne [4], we see from Lemma 4.1, (3.1) and (3.2), that for $k \geqq k_0$,

$$\begin{aligned} \mathbf{e}^{(k+1)}(T) &= \mathbf{C}^{(k+1)}\mathbf{e}^{(k)}(T), \\ \mathbf{e}^{(k+1)}(Y) &= 0. \end{aligned} \tag{4.18}$$

Here, $\mathbf{C}^{(k+1)}$ is a $|T| \times |T|$ matrix such that

$$\mathbf{C}^{(k+1)} = \prod_{l=1}^{|T|} (\mathbf{H}^{(k+1,l)}\mathbf{L}^{(l)}); \tag{4.19}$$

$\mathbf{L}^{(l)}$ are $|T| \times |T|$ matrices such that

$$\mathfrak{L}_\omega[\mathbf{A}(T)] = \prod_{l=1}^{|T|} \mathbf{L}^{(l)}; \tag{4.20}$$

and $\mathbf{H}^{(k+1,l)}$ is a $|T| \times |T|$ diagonal matrix with diagonal elements equal to either 0 or 1. In particular, when $T = Z$, then $\mathbf{L}^{(l)} = (L_{ij}^{(l)})$ and $\mathbf{H}^{(k+1,l)} = \operatorname{diag}(H_{ii}^{(k+1,l)})$, where

$$L_{ij}^{(l)} = \begin{cases} 1 & \text{if } i = j \text{ and } i \neq l, \\ 1 - \omega & \text{if } i = j = l, \\ -\omega a_{lj}/a_{ll} & \text{if } i = l \text{ and } j \neq l, \\ 0 & \text{otherwise,} \end{cases}$$

$$H_{ii}^{(k+1,l)} = \begin{cases} 1 & \text{if } i \neq l, \\ 1 & \text{if } i = l \text{ and } x_l^{(k+1)} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

From (4.16) and (4.17), it follows that $\mathbf{L}^{(l)} \geqq 0$; that is, the elements of $\mathbf{L}^{(l)}$ are nonnegative. Hence, we see from (4.19) and (4.20) that $0 \leqq \mathbf{C}^{(k+1)} \leqq \mathfrak{L}_\omega[\mathbf{A}(T)]$. The theorem follows immediately.

On the basis of Theorems 4.2 and 4.3 we make the following conjecture.

CONJECTURE 4.4. $R[\mathbf{A}, \mathbf{b}, \omega] \leqq R_\infty[\mathfrak{L}_\omega(\mathbf{A}[T])]$, *where* $T = Z - Z(\mathbf{y})$, *and* $\{\mathbf{x}, \mathbf{y}\}$ *is the solution of Problem* 1.

Theorems 4.2 and 4.3 provide some help in choosing ω so as to maximize the rate of convergence of Algorithm 1.

If (4.14) holds, then we see from Theorem 4.2 that we should set $\omega = \omega_{\text{opt}}$, where

$$\omega_{\text{opt}} = \omega_b[\mathbf{A}(X)]. \tag{4.21}$$

Of course, (4.21) does not give ω_{opt} explicitly, since, in general, neither X nor $\omega_b[\mathbf{A}(X)]$ is known explicitly. However:

(i) There are several methods of estimating an optimum overrelaxation parameter $\omega_b(\tilde{\mathbf{A}})$. One approach is to obtain rigorous a priori bounds on the

spectral radius of the Jacobi matrix corresponding to $\tilde{\mathbf{A}}$. Another approach is to choose a value for the overrelaxation parameter ω, iterate a number of times, and use the information gained to estimate $\omega_b(\tilde{\mathbf{A}})$. Details and references are given by Varga [11, p. 283].

(ii) If $\mathbf{A}$ is 2-cyclic, consistently ordered, and satisfies (4.17), then it follows from the Perron–Frobenius theory for nonnegative matrices (Varga [11, p. 26]) that, for any $X \subset Z$, $\omega_b(\mathbf{A}[X]) \leqq \omega_b(\mathbf{A})$. Remembering that it is in general better to overestimate ω_{opt} rather than underestimate ω_{opt} (Varga [11, p. 114]), we can estimate ω_{opt} by $\omega_b(\mathbf{A})$.

(iii) After performing k iterations of Algorithm 1, the set X can be estimated, according to Lemma 4.1, by the set $Z(\mathbf{x}^{(k)})$.

In [3], some of these ideas were used to estimate ω_{opt} for Christopherson's problem for $n = 64$. The estimated value of ω_{opt} was $\omega^* = 1.906$, and with $\omega = \omega^*$ it was found that 146 iterations were needed to reduce the residuals $r_i^{(k+1)}$ (see (3.1)) to less than 10^{-7}. The iterations were also performed with $\omega = 1.0, 1.1, 1.2, \cdots, 1.9$; the maximum number of iterations needed was 811 (for $\omega = 1.0$), and the minimum number needed was 70 (for $\omega = 1.8$).

If (4.14) does not hold, then we can say much less about the choice of ω. However, if $\mathbf{A}$ is 2-cyclic and consistently ordered and (4.17) is satisfied, then, from the Perron–Frobenius theory for nonnegative matrices, it follows that for any $T \subset Z$, $R_\infty[\mathfrak{L}_\omega(\mathbf{A}(T))]$ is a monotone decreasing function of ω for $0 < \omega \leqq 1$. Hence, the results of Theorem 4.3 suggest that we should choose $\omega \geqq 1$.

Acknowledgment. It is a pleasure to thank Professor J. B. Rosen for his help.

REFERENCES

[1] R. W. Cottle and G. B. Dantzig, *Complementary pivot theory of mathematical programming*, Mathematics of the Decision Sciences, Part I, G. B. Dantzig and A. F. Veinott, Jr., eds., American Mathematical Society, Providence, R.I., 1968.

[2] D. G. Christopherson, *A new mathematical method for the solution of film lubrication problems*, Proc. Inst. Mech. Engrgs., 146 (1941), pp. 126–135.

[3] C. W. Cryer, *The method of Christopherson for solving free boundary problems for infinite journal bearings by means of finite differences*, Tech. Rep. 72, Computer Sciences Dept., University of Wisconsin, Madison, 1969.

[4] A. A. Gnanadoss and M. R. Osborne, *The numerical solution of Reynolds' equation for a journal bearing*, Quart. J. Mech. Appl. Math., 17 (1964), pp. 241–246.

[5] G. Hadley, *Nonlinear and Dynamic Programming*, Addison-Wesley, Reading, Mass., 1964.

[6] C. E. Lemke, *On complementary pivot theory*, Mathematics of the Decision Sciences, Part I, G. B. Dantzig and A. F. Veinott, Jr., eds., American Mathematical Society, Providence, R.I., 1968.

[7] J. M. Ortega and M. L. Rockoff, *Nonlinear difference equations and Gauss-Seidel type iterative methods*, SIAM J. Numer. Anal., 3 (1966), pp. 497–513.

[8] S. Schechter, *Iteration methods for nonlinear problems*, Trans. Amer. Math. Soc., 104 (1962), pp. 179–189.

[9] ———, *Relaxation methods for convex problems*, SIAM J. Numer. Anal., 5 (1968), pp. 601–612.

[10] G. Temple, *The general theory of relaxation methods applied to linear systems*, Proc. Roy. Soc. Ser. A., 169 (1939), pp. 476–500.

[11] R. S. Varga, *Matrix Iterative Analysis*. Prentice-Hall, Englewood Cliffs, N.J., 1962.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

PERTURBATIONS OF LINEAR CONTROL SYSTEMS*

JERALD P. DAUER†

Abstract. The purpose of this paper is to consider the controllability of linear control systems which are perturbations (with respect to L^p-norms) of controllable linear systems. We show that the set of all completely controllable linear systems is open and dense in the set of all linear control systems.

1. Introduction. The problem of controllability of a linear control system

$$\dot{x} = A(t)x + B(t)u \qquad (\dot{x} = dx/dt) \tag{1}$$

defined on a bounded interval has been studied by several authors [1]–[4]. Kalman, Ho and Narendra [1], Weiss [2], and Silverman and Meadows [3] gave three basic necessary and sufficient conditions when A and B are continuous. Conti [4] studied the problem for $A \in L^1$ and $B \in L^p$, $1 \leqq p < \infty$. Related results were obtained by Youla [5], who studied realizations of weighting patterns with A, $B \in L^2$ (see also [6], [7]). In addition, when A and B are autonomous, Lee and Markus [8, p. 100] showed that the set of completely controllable linear autonomous systems is open and dense in the set of all linear autonomous control systems.

The object of this paper is to generalize this result of Lee and Markus to non-autonomous linear systems. In § 3 we show openness with respect to L^p-norms for linear systems defined on bounded intervals. We give a density theorem in § 4 which also proves an assertion of Kalman, Ho and Narendra [1]. They commented that intuitively linear systems are completely controllable. In fact, as they point out, there are many linear systems that are not completely controllable, such as $\dot{x}(t) = 2$. However, if the system represents a physical process that involves approximated parameters, the result of § 4 shows that we can assume that the system is completely controllable.

2. Preliminaries. Suppose system (1) is defined on a bounded interval $I = [t_0, t_1]$. The state vector function x is n-dimensional, the control vector function u is m-dimensional, and A and B are $n \times n$ and $n \times m$ matrix functions which are (Lebesgue) integrable on I. Let X be the fundamental matrix solution of $z = A(t)z$ such that $X(t_0)$ is the identity matrix.

If f is a real-valued measurable function defined on I and r is such that $1 \leqq r < \infty$, then define

$$\|f\|_r = \left(\int_I |f(s)|^r \, ds \right)^{1/r}.$$

For $r = \infty$, define $\|f\|_r = \text{ess sup } \{|f(s)| : s \in I\}$. Let L^r be the set of all such functions f satisfying $\|f\|_r < \infty$.

System (1) is said to be *completely controllable in* L^r if for every $x_0, x_1 \in E^n$, Euclidean n-space, there exists a control function $u \in L^r$ such that the solution x of

$$\dot{x} = A(t)x + B(t)u(t),$$

$$x(t_0) = x_0$$

satisfies $x(t_1) = x_1$.

* Received by the editors June 9, 1970, and in revised form December 3, 1970.

† Department of Mathematics, University of Kansas, Lawrence, Kansas. Now at Department of Mathematics, University of Nebraska, Lincoln, Nebraska 68508.

Let p be given, $1 \leqq p \leqq \infty$, and let q satisfy $1/p + 1/q = 1$. For $y \in E^n$, let y^* denote the transpose of y.

The following result is due to Conti [4].

THEOREM 1. *Let $1 < q \leqq \infty$ and assume $B \in L^p$. System* (1) *is completely controllable in L^q if and only if*

$$\inf\left\{\left(\int_I |y^* X(t_1)X^{-1}(s)B(s)|^p \, ds\right)^{1/p} : |y^*| = 1\right\} > 0.$$

A set of vector functions $\{x_1, \cdots, x_k\}$ is said to be linearly independent over I if for every nonzero vector $(a_1, \cdots, a_k) \in E^k$ there exists a subset J of I with positive (Lebesgue) measure such that

$$\sum_{i=1}^{k} a_i x_i(t) \neq 0 \qquad \text{for all } t \in J.$$

Weiss [2] and Kalman, Ho and Narendra [1] proved results similar to the following two theorems. Their proofs can be extended to these results.

THEOREM 2. *Assume $B \in L^\infty$. System* (1) *is completely controllable in L^1 if and only if the rows of the matrix function $\{X^{-1}(t)B(t)\}$ are linearly independent over I.*

THEOREM 3. *Let $2 \leqq p \leqq \infty$ and assume $B \in L^p$. System* (1) *is completely controllable in L^p if and only if*

$$W = \int_I X^{-1}(s)B(s)B(s)^* X^{-1}(s)^* \, ds$$

is positive definite.

In the next section, we will use the following two lemmas in connection with the preceding two theorems.

LEMMA 1. *Let $\{x_1, \cdots, x_k\}$ be a set of measurable vector functions in E^n which is linearly independent over I. Then there exists $\varepsilon > 0$ such that any set of vector functions $\{y_1, \cdots, y_k\}$, satisfying $|y_i(t) - x_i(t)| < \varepsilon$ for $1 \leqq i \leqq k$ and almost every $t \in I$, is linearly independent over I.*

Proof. Let the function f be defined by

$$f(a, G(t)) = \sum_{i=1}^{k} a_i x_i(t),$$

where $a = (a_1, \cdots, a_k) \in E^k$ and $G(t) = (x_1(t), \cdots, x_k(t))$ is an $n \times k$ matrix function.

Fix an arbitrarily chosen $a \in \partial S_1(0)$, the boundary of the open ball with center 0 and radius 1. Then there exists a set $J_a \subseteq I$ with positive measure and a number $b_a > 0$ such that $|f(a, G(t))| \geqq b_a$ for all $t \in J_a$. Further, by Lusin's theorem, there exists a compact set $K_a \subseteq I$ such that G is continuous on K_a and such that the set $K_a \cap J_a$ has positive measure.

Since f is uniformly continuous on the compact set

$$\mathfrak{A} = \partial S_1(0) \times \{n \times k \text{ matrices } H : |H - G(t)| \leqq 1 \text{ for some } t \in K_a\},$$

there exists $\varepsilon_a > 0$ such that if $t \in K_a$, $(c, H_c) \in \mathfrak{A}$ and $|a - c| + |H_c - G(t)| < \varepsilon_a$,

then

$$|f(a, G(t)) - f(c, H_c)| < b_a/2.$$

Hence if $t \in K_a \cap J_a$, we have $|f(c, H_c)| \geqq b_a/2$.

Covering the compact set $\partial S_1(0)$ with the open balls $S_{\varepsilon_{a/2}}(a)$, for $a \in \partial S_1(0)$, gives us a finite subcover with centers $a_1, \cdots, a_r$. Let $b_{a_i} = b_i$, $\varepsilon_{a_i} = \varepsilon_i$ and $\varepsilon = (1/2k) \min \{\varepsilon_i : 1 \leqq i \leqq r\}$.

Without loss of generality, suppose $|y_i(t) - x_i(t)| < \varepsilon$ for all $t \in I$. Let $H(t) = (y_1(t), \cdots, y_k(t))$.

Then, given any $a \in \partial S_1(0)$, there exists $i \in \{1, \cdots, r\}$ such that $|a - a_i| < \varepsilon_i/2$. Hence, $|a - a_i| + |H(t) - G(t)| < \varepsilon_i$ for all $t \in K_{a_i} \cap J_{a_i}$. Therefore, $|f(a, H(t))| \geqq b_i/2 > 0$ for all $t \in K_{a_i} \cap J_{a_i}$.

Taking $a \in E^k$, $a \neq 0$, we have $a/|a| \in \partial S_1(0)$. Using the above argument, there exists $i \in \{1, \cdots, r\}$ such that

$$\frac{1}{|a|} \sum_{j=1}^{k} a_j y_j(t) \geqq \frac{b_i}{2} > 0 \qquad \text{for all} \quad t \in K_{a_i} \cap J_{a_i}.$$

Hence $\{y_1, \cdots, y_k\}$ is linearly independent.

The following lemma can be easily shown.

LEMMA 2. *Suppose W is a positive definite $n \times n$ matrix. There exists $\varepsilon > 0$ such that if the $n \times n$ matrix V satisfies $|W - V| < \varepsilon$, then V is positive definite.*

3. Openness. In this section, we will assume that system (1) is completely controllable in various L^r spaces. We then examine the controllability of linear systems which are perturbations of system (1). Throughout this section, C and D will denote measurable $n \times n$ and $n \times m$ matrix functions.

The first result considers the natural setting of controllability in L^q, where $B \in L^p$.

THEOREM 4. *Suppose $A \in L^1$ and $B \in L^p$, $1 \leqq p \leqq \infty$, and assume that system (1) is completely controllable in L^q. There exists $\varepsilon > 0$ such that if*

$$\|A - C\|_p + \|B - D\|_p < \varepsilon,$$

then the system

$$\dot{y} = C(t)y + D(t)u \tag{2}$$

is completely controllable in L^q.

Since the interval I is bounded, we have the following result.

COROLLARY. *Assume the conditions of Theorem 4. For every r, $p \leqq r \leqq \infty$, there exists $\varepsilon > 0$ such that if*

$$\|A - C\|_r + \|B - D\|_r < \varepsilon,$$

then system (2) is completely controllable in L^q.

Proof of Theorem 4. We first show that if $\|A - C\|_1$ is sufficiently small, then the system

$$\dot{y} = C(t)y + B(t)u \tag{3}$$

is completely controllable in L^q.

Let Y be the fundamental matrix solution of $\dot{z} = C(t)z$ such that $Y(t_0)$ is the identity matrix. Then

$$\dot{Y}(t) = A(t)Y(t) + [C(t) - A(t)]Y(t).$$

Applying the variation of parameters formula, we have

$$Y(t) = X(t)X^{-1}(t_0)Y(t_0) + X(t)\int_{t_0}^{t} X^{-1}(s)[C(s) - A(s)]Y(s)\,ds.$$

Letting $M = \max_{t\in I}|X(t)|$ and $N = \max_{t\in I}|X^{-1}(t)|$, we have

$$|Y(t) - X(t)| \leqq M^2N\|A - C\|_1 + MN\int_{t_0}^{t} |A(s) - C(s)|\,|Y(s) - X(s)|\,ds.$$

Thus, by Grönwall's inequality, we have

$$|Y(t) - X(t)| \leqq M^2N\|A - C\|_1\, e^{(MN\|A-C\|_1)}. \tag{4}$$

Similarly, using adjoint equations,

$$|Y^{-1}(t) - X^{-1}(t)| \leqq MN^2\|A - C\|_1\, e^{(MN\|A-C\|_1)}. \tag{5}$$

Assume $q > 1$, and let

$$\delta = \inf\left\{\left(\int_{t_0}^{t_1} |y^*X(t_1)X^{-1}(s)B(s)|^p\,ds\right)^{1/p} : |y^*| = 1\right\}.$$

Theorem 1 states that $\delta > 0$ and that if $|y^*| = 1$, we have

$$\|y^*X(t_1)X^{-1}B\|_p \geqq \delta > 0.$$

Hence $\|B\|_p \neq 0$. Therefore, using (4) and (5), we can choose $\varepsilon > 0$ such that if $\|A - C\|_1 < \varepsilon$, then

$$\frac{\delta}{2\|B\|_p} > M\|Y^{-1} - X^{-1}\|_\infty + \|Y - X\|_\infty N + \|X - Y\|_\infty\|Y^{-1} - X^{-1}\|_\infty.$$

Hence, for every y^* such that $|y^*| = 1$, we have

$$\|y^*Y(t_1)Y^{-1}B\|_p \geqq \|y^*X(t_1)X^{-1}B\|_p - \frac{\delta}{2} \geqq \frac{\delta}{2} > 0.$$

Theorem 1 shows that system (3) is completely controllable in L^q for $1 < q \leqq \infty$.

If $q = 1$, then $B \in L^\infty$. By Theorem 2, the rows of the matrix function $\{X^{-1}(t)B(t)\}$ are linearly independent over I. Let $\tilde{\varepsilon} > 0$ be given by Lemma 1 for this function. Then, for sufficiently small $\|A - C\|_1$, equation (5) gives

$$|Y^{-1}(t) - X^{-1}(t)| < \frac{\tilde{\varepsilon}}{\|B\|_\infty} \qquad \text{for all } t \in I.$$

Using Lemma 1 and Theorem 2, we have that system (3) is completely controllable in L^q for $q = 1$.

Next, we show that if $\|B - D\|_p$ is sufficiently small, then the system

$$\dot{y} = A(t)y + D(t)u \tag{6}$$

is completely controllable in L^q.

Assume that $q > 1$ and let M, N and δ be as above. Let $\varepsilon = \delta/(2MN)$. Then, for $\|B - D\|_p < \varepsilon$ and every y^* with $|y^*| = 1$, we have

$$\|y^*X(t_1)X^{-1}D\|_p \geqq \|y^*X(t_1)X^{-1}B\|_p - \frac{\delta}{2} \geqq \frac{\delta}{2} > 0.$$

Theorem 1 shows that system (6) is completely controllable in L^q for $1 < q \leqq \infty$.

The proof of the result for $q = 1$ is similar to that of the corresponding result in the first part of the proof. Here, we need

$$\|B - D\|_\infty < \frac{\tilde{\varepsilon}}{\|X^{-1}\|_\infty}.$$

Again, we use Lemma 1 and Theorem 2 to show system (6) completely controllable. Combining the results on systems (3) and (6) gives Theorem 4.

The following theorem considers controllability in a smaller set of controls than the preceding result.

THEOREM 5. *Suppose $A \in L^1$ and $B \in L^p$, $2 \leqq p \leqq \infty$, and assume that system (1) is completely controllable in L^p. There exists $\varepsilon > 0$ such that if*

$$\|A - C\|_q + \|B - D\|_q < \varepsilon,$$

then the system

$$\dot{y} = C(t)y + D(t)u \tag{2}$$

is completely controllable in L^p.

COROLLARY. *Assume the conditions of Theorem 5. For every r, $q \leqq r \leqq \infty$ (in particular for $r = p$), there exists $\varepsilon > 0$ such that if*

$$\|A - C\|_r + \|B - D\|_r < \varepsilon,$$

then system (2) is completely controllable in L^p.

Proof of Theorem 5. As in the proof of Theorem 4, we will first consider system (3).

Let Y, M and N be as in the proof of Theorem 4. Define V by

$$V = \int_{t_0}^{t_1} Y^{-1}(s)B(s)B(s)^*Y^{-1}(s)^* \, ds.$$

If W is defined as in Theorem 3, then we have

$$|W - V| \leqq (2N + \|Y^{-1} - X^{-1}\|_\infty)\|Y^{-1} - X^{-1}\|_\infty \int_{t_0}^{t_1} |B(s)|^2 \, ds.$$

Since I is bounded, we have $\|B\|_p < \infty$ implies $\|B\|_2 < \infty$. Thus, using (5), we can make $|W - V|$ as small as we wish by making $\|A - C\|_1$ sufficiently small. Since W is positive definite, Lemma 2 and Theorem 3 show that system (3) is completely controllable for $\|A - C\|_1$ sufficiently small.

Now, consider system (6). Define $P(t) = B(t) - D(t)$, let W be as in Theorem 3, and let

$$U = W - \int_{t_0}^{t_1} X^{-1}(s)B(s)P(s)^*X^{-1}(s)^* \, ds - \int_{t_0}^{t_1} X^{-1}(s)P(s)B(s)^*X^{-1}(s)^* \, ds.$$

Then

$$|W - U| = 2\|X^{-1}\|_\infty^2 \|B\|_p \|P\|_q.$$

Since W is positive definite, it follows from Lemma 2 that if $\|B - D\|_q = \|P\|_q$ is sufficiently small, then U is positive definite. Noting that

$$\int_{t_0}^{t_1} X^{-1}(s)P(s)P(s)^*X^{-1}(s)^* \, ds$$

is positive semidefinite, we have that if $\|B - D\|_q$ is sufficiently small, then

$$\int_{t_0}^{t_1} X^{-1}(s)D(s)D(s)^*X^{-1}(s)^* \, ds = U + \int_{t_0}^{t_1} X^{-1}(s)P(s)P(s)^*X^{-1}(s)^* \, ds$$

is positive definite. Theorem 3 shows that system (6) is completely controllable for $\|B - D\|_q$ sufficiently small. Combining these results gives us Theorem 5.

4. Density. The following is a result on the denseness of completely controllable systems. The generality of this result is exhibited in the corollary.

THEOREM 6. *Consider system* (1), *where* $A, B \in L^1$. *For each* $\varepsilon > 0$ *there exists a measurable* $n \times m$ *matrix function* D *such that*

$$|B(t) - D(t)| < \varepsilon \qquad \text{for all } t \in I$$

and such that the system

$$\dot{y} = A(t)y + D(t)u$$

is completely controllable in L^∞.

Further, given any $\delta > 0$, *we can choose* D *such that the* (*Lebesgue*) *measure of* $\{t \in I : D(t) \neq B(t)\}$ *is less than* δ. *Also, if* B *is* (*piecewise*) *continuous, we can choose* D *to be* (*piecewise*) *continuous. If* B *is constant we can choose* D *constant.*

Proof. Let $\varepsilon, \delta > 0$ be given. Since B is measurable, there exists a closed set $T_1 \subseteq I$ of positive measure such that B is continuous on T_1. Hence there exists $T \subseteq T_1$ and $\bar{t} \in T$ such that the measure of T is positive, but less than δ, and such that

$$|B(t) - B(\bar{t})| < \varepsilon/2 \qquad \text{for all } t \in T.$$

Let D_1 be a matrix whose entries are algebraically independent over the rational numbers and which satisfies

$$|B(\bar{t}) - D_1| < \varepsilon/2.$$

Thus (the determinant) $\det(D_1) \neq 0$ and

$$|B(t) - D_1| < \varepsilon \qquad \text{for all } t \in T.$$

Define the matrix function D by

$$D(t) = \begin{cases} B(t) & \text{for } t \in I \setminus T, \\ D_1 & \text{for } t \in T. \end{cases}$$

Then $\det(X(t_1)X^{-1}(s)D(s)) = \det(X(t_1)X^{-1}(s)) \det(D(s)) \neq 0$ for all $s \in T$. Therefore, for each $s \in T$ the rows of the matrix $\{X(t_1)X^{-1}(s)D(s)\}$ are linearly independent. Hence, for each y^*, $|y^*| = 1$ and each $s \in T$, we have

$$y^*X(t_1)X^{-1}(s)D(s) \neq 0.$$

Since the function $f(s, y^*) = y^*X(t_1)X^{-1}(s)D(s)$ is uniformly continuous on the compact set $\mathfrak{A} = T \times \{y^* : |y^*| = 1\}$, there exists $(\bar{s}, \bar{y}^*) \in \mathfrak{A}$ such that

$$0 < \alpha = |f(\bar{s}, \bar{y}^*)| \leqq |f(s, y^*)| \qquad \text{for all } (s, y^*) \in \mathfrak{A}.$$

Hence, for each y^*, $|y^*| = 1$, we have

$$\int_T |y^*X(t_1)X^{-1}(s)D(s)|\, ds \geqq \alpha \cdot (\text{measure of } T) > 0.$$

Since $D \in L^1$, Theorem 1 gives the desired result.

If B is (piecewise) continuous, we can assume that $T = [t_2, t_3]$ for some $t_2, t_3 \in I$. By defining D_1 on a subinterval of T we can easily make D a (piecewise) continuous function. If B is constant, define $D(t) = D_1$.

COROLLARY. *Consider system* (1) *with* $A \in L^1$ *and* $B \in L^p$, $1 \leqq p \leqq \infty$. *For each pair* r, v, *with* $1 \leqq r, v \leqq \infty$, *and each* $\varepsilon > 0$ *there exists a measurable matrix function* D *such that*

$$\|B - D\|_r < \varepsilon$$

and such that the system

$$\dot{y} = A(t)y + D(t)u$$

is completely controllable in L^v.

5. Remark. Suppose A, $B \in L^1$, and let completely controllable mean completely controllable in L^∞. Theorems 4 and 6 show that the set of completely controllable linear systems is open and dense in the set of all linear control systems with respect to the L^p-norm for $1 \leqq p \leqq \infty$. This statement is also true if we replace the words linear systems by continuous linear systems, piecewise continuous linear systems or constant linear systems.

Acknowledgment. This paper is a chapter of the author's doctoral thesis written under the guidance of Professor Fred S. Van Vleck at the University of Kansas.

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] L. WEISS, *The concepts of differential controllability and differential observability*, J. Math. Anal. Appl., 10 (1965), pp. 442–449.
[3] L. M. SILVERMAN AND A. E. MEADOWS, *Controllability and observability in time-variable linear systems*, this Journal, 5 (1967), pp. 64–73.
[4] R. CONTI, *Contributions to linear control theory*, J. Differential Equations, 1 (1965), pp. 427–445.
[5] D. C. YOULA, *The synthesis of linear dynamical systems from prescribed weighting patterns*, SIAM J. Appl. Math., 14 (1966), pp. 527–549.
[6] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.
[7] L. WEISS AND R. E. KALMAN, *Contributions to linear system theory*, Internat. J. Engrg. Sci., 3 (1965), pp. 141–171.
[8] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

BOUNDARY VALUE CONTROL THEORY OF THE HIGHER-DIMENSIONAL WAVE EQUATION, PART II*

DAVID L. RUSSELL†

Abstract. The present work extends controllability results obtained for the wave equation in two or more space variables in an earlier article [13]. In the earlier paper approximate controllability was established for time $T > 2T_0$, where T_0 is a constant related to the wave propagation speeds in the medium, but only for three or fewer space dimensions. In the present article we establish this result for an arbitrary space dimension. We also examine the controllability problem for $T = 2T_0$, the "critical time," and show that here controllability depends upon certain relationships between the coefficients of the partial differential equation and the shape of the spatial domain under consideration.

1. Introduction. This paper is a sequel to [13], where we began our study of the approximate controllability of the higher-dimensional wave equation with boundary value controls. There, and here, we let Ω be a bounded, open, connected domain in R^n whose boundary, Γ, is an analytic (or C^∞ and piecewise analytic) $(n-1)$-dimensional surface in R^n. We parametrize Γ with an $(n-1)$-dimensional vector variable s and indicate points on Γ by $x(s)$. Integrals over Ω are written as $\int_\Omega (\cdot)\, dx$, while integrals over Γ are written $\int_\Gamma (\cdot)\, ds$. Taking $\tilde{\Gamma}$ to be a relatively open subset of Γ and T a positive number, we define an admissible control to be a function $f: \Gamma \otimes [0, T] \to R^1$ such that $f \in C^\infty(\Gamma \otimes [0, T])$ and f vanishes identically outside a compact subset of $\tilde{\Gamma} \otimes (0, T)$.

For all such admissible controls f we let $w^f(x, t)$ solve the linear hyperbolic mixed initial boundary value problem

$$\rho(x) w^f_{tt} - \sum_{i,j=1}^{n} (\alpha_{ij}(x) w^f_i)_j = 0 \quad \text{in } \Omega \otimes [0, T], \tag{1.1}$$

$$w^f_x(x(s), t) A(x(s)) \eta(x(s)) = f(s, t) \quad \text{on } \Gamma \otimes [0, T], \tag{1.2}$$

$$w^f(x, 0) \equiv w^f_t(x, 0) \equiv 0, \quad x \in \Omega. \tag{1.3}$$

The subscripts t and i denote partial differentiation with respect to t and x^i (the ith component of $x \in R^n$) respectively. The subscript x indicates the gradient vector of the function to which it is applied. The vector $\eta(x(s))$ is the outward unit normal to Γ at $x(s) \in \Gamma$. The real analytic functions $\rho(x), \alpha_{ij}(x), i, j = 1, 2, \cdots, n$, are such that

$$\alpha_{ij}(x) = \alpha_{ji}(x),$$

$$\rho(x) \geqq \rho_0 > 0,$$

$$v' A(x) v \geqq \delta_0 \|v\|^2, \qquad \delta_0 > 0,$$

* Received by the editors September 16, 1970.

† Departments of Mathematics and Computer Sciences, University of Wisconsin, Madison, Wisconsin 53706 and Honeywell, Inc., St. Paul, Minnesota. This research was supported in part by the National Science Foundation under Grant GP-20858.

in some open set which includes $\Omega \cup \Gamma$. Here $A(x)$ is the $n \times n$ symmetric matrix whose entries are $\alpha_{ij}(x)$.

From [3] and [8] we learn that (1.1), (1.2), (1.3) has a unique C^∞-solution in $\Omega \otimes [0, T]$. Thus we may let R_T denote the set of all terminal states $(w^f(\cdot, T), w_t^f(\cdot, T))$. The set R_T is a subspace of the Hilbert space $H_E(\Omega)$ of finite energy states with inner product

$$\langle (u, u_t); (v, v_t) \rangle_E = \int_\Omega [\rho(x) u_t(x) v_t(x) + u_x(x) A(x) v_x(x)'] dx$$

(here $'$ denotes the transpose of a vector) and norm

$$\|(v, v_t)\|_E = (\langle (v, v_t); (v, v_t) \rangle_E)^{1/2}.$$

The gradients u_x, v_x are defined in the sense of the theory of distributions. To avoid an indefinite inner product, two states which differ by $(c, 0)$, where c is a constant function on Ω, are identified. However, we will continue to speak of elements of $H_E(\Omega)$ as "states" rather than as "equivalence classes of states."

The control system (1.1), (1.2) is said to be approximately controllable in time T if R_T is dense in $H_E(\Omega)$, i.e., if the validity of the equation

$$\langle (w^f(\cdot, T), w_t^f(\cdot, T)); (\hat{v}, \hat{v}_t) \rangle_E = 0$$

for all $f \in R_T$ implies that $(\hat{v}, \hat{v}_t) = (c, 0)$, a zero energy state in $H_E(\Omega)$.

In [13] we showed that Ω, Γ, ρ and A determine a positive number T_0 such that:

(i) if $T < 2T_0$, the system (1.1), (1.2), (1.3) is not approximately controllable in time T;

(ii) if $T > 2T_0$ and $n \leqq 3$, then the system is approximately controllable in time T.

We will refer to $2T_0$ as the *critical time*. When $n = 1$ it is known (see [5], [14], [15], e.g.) that approximate controllability continues to hold for $T = 2T_0$.

The purpose of the present paper is two-fold. First, we show in § 2 that if $T > 2T_0$, approximate controllability holds without any restriction on the dimension n. Second, we show in the remaining sections that if $n \geqq 2$, approximate controllability may or may not hold for $T = 2T_0$, the critical time, depending on certain relationships between Γ, ρ and A. Because the proofs for $T = 2T_0$ are very detailed, they are given only for special examples. In the concluding remarks we describe the form which a general theory of critical time approximate controllability would take.

2. A new proof of approximate controllability for $T > 2T_0$. The theorem which we will prove in this section replaces Theorem 4 in [13]. The new result has the advantage of being valid for all positive integers n. Many of the details of the proof are the same as in the earlier result. Therefore we will concentrate on the essential differences, referring the reader to [13] for complete treatment of parts common to both proofs.

Let $(\hat{v}, \hat{v}_t)$ be a finite energy state, i.e., $\|(\hat{v}, \hat{v}_t)\|_E < \infty$, and assume that $(\hat{v}, \hat{v}_t)$ is orthogonal to all states $(w^f(\cdot, T), w_t^f(\cdot, T))$ in R_T relative to the energy inner

product. Thus

$$\langle (w^f(\cdot, T), w_t^f(\cdot, T)); (\hat{v}, \hat{v}_t)\rangle_E = \int_\Omega [\rho(x) w_t^f(x, T)\hat{v}_t(x) + w_x^f(x, T)A(x)\hat{v}_x'(x)]\, dx = 0 \tag{2.1}$$

for all admissible controls f. We let $v(x, t)$ be the generalized solution of the mixed problem

$$\rho(x)v_{tt} - \sum_{i,j=1}^{n} (\alpha_{ij}(x)v_i)_j = 0 \quad \text{in } \Omega \otimes [0, T], \tag{2.2}$$

$$v_x(x(s), t)A(x(s))\eta(x(s)) = 0, \quad (x(s), t) \in \Gamma \otimes [0, T], \tag{2.3}$$

$$v(x, T) = \hat{v}(x), \qquad v_t(x, T) = \hat{v}_t(x). \tag{2.4}$$

The existence of such a solution is proved, e.g., in [8] and [10], where it is likewise shown that $v(\cdot, t)$ and $v_t(\cdot, t)$ define continuous functions from $[0, T]$ into $H^1(\Omega)$ and $H^0(\Omega) = L^2(\Omega)$, respectively. (Recall that if m is a nonnegative integer, then $H^m(\Omega)$ consists of real functions $u(x)$ whose derivatives of order $\leqq m$, taken in the sense of the theory of distributions, lie in $L^2(\Omega)$. $H^m(\Omega)$ is a Hilbert space with inner product

$$(u, \hat{u})_{H^m(\Omega)} = \sum_{0 \leqq \|p\| \leqq m} \int_\Omega [D^p u(x) D^p \hat{u}(x)]\, dx.$$

Here p is an n-vector with nonnegative integer components $p_1, p_2, \cdots, p_n$, $\|p\| = p_1 + p_2 + \cdots + p_n$, and D^p denotes $\partial^{\|p\|}/(\partial x^1)^{p_1}(\partial x^2)^{p_2} \cdots (\partial x^n)^{p_n}$).

As in [13] we smooth the solution $v(x, t)$ by a process of antidifferentiation and formation of finite differences. The innovation lies in the way in which the antiderivatives are defined. We consider the elliptic operator

$$Bu = \frac{1}{\rho(x)} \sum_{i,j=1}^{n} (\alpha_{ij}(x)u_i)_j$$

which is defined on functions $u \in C^2(\bar{\Omega})$ ($\bar{\Omega} = \Omega \cup \Gamma$) satisfying the boundary conditions

$$u(x(s))A(x(s))\eta(x(s)) = 0, \qquad x(s) \in \Gamma.$$

This unbounded operator is symmetric with respect to the inner product $(u, \hat{u})_\rho = \int_\Omega u(x)\hat{u}(x)\rho(x)\, dx$ and has an unbounded self-adjoint (with respect to that inner product) extension, which we shall still call B, defined on a domain D dense in $L^2(\Omega)$. (See, e.g., [4], [6].) Moreover, if $(u, 1)_\rho = 0$, then there is a positive number λ_0, the smallest eigenvalue of B except 0, such that

$$\|Bu\|_\rho \geqq \lambda_0 \|u\|_\rho.$$

From this it follows that if we let $\hat{B}$ denote the restriction of B to $D \cap \{u \in L^2(\Omega) | (u, 1)_\rho = 0\}$, then $\hat{B}^{-1}$ is defined, bounded and self-adjoint on $\{u \in L^2(\Omega) | (u, 1)_\rho = 0\}$, which we shall call $\hat{D}$.

From the work of Lions–Magenes [9, p. 165 ff.] it is known that if $g \in \hat{D} \cap H^m(\Omega)$, $m \geqq 0$, then $\hat{B}^{-1}g \in \hat{D} \cap H^{m+2}(\Omega)$ and the mapping $g \in \hat{D} \cap H^m(\Omega) \to \hat{B}^{-1}g \in \hat{D} \cap H^{m+2}(\Omega)$ is continuous with respect to the norms $\|\cdot\|_{H^m(\Omega)}$, $\|\cdot\|_{H^{m+2}(\Omega)}$.

We return to $(\hat{v}, \hat{v}_t)$ and let c_1 and c_2 be real constants such that

$$\int_\Omega (\hat{v}(x) - c_1)\rho(x)\,dx = \int_\Omega (\hat{v}_t(x) - c_2)\rho(x)\,dx = 0. \tag{2.5}$$

Then

$$\tilde{v}(x,t) = v(x,t) - c_1 - c_2(t - T)$$

satisfies (2.2) and (2.3) and $\tilde{v}(\cdot\,,t) \in \hat{D} \cap H^1(\Omega)$, $t \in [0, T]$. Likewise,

$$\tilde{v}_t(x,t) = v_t(x,t) - c_2$$

is such that $\tilde{v}_t(\cdot\,,t) \in \hat{D} \cap H^0(\Omega)$, $t \in [0, T]$. We define, for each nonnegative integer k,

$$D^{-2k}\tilde{v} = \hat{B}^{-k}\tilde{v}, \qquad D^{-2k+1}\tilde{v} = \hat{B}^{-k}\tilde{v}_t$$

and conclude from the above cited work in [9] that for a nonnegative integer m,

$$D^{-m}\tilde{v}(\cdot\,,t) \in \hat{D} \cap H^{m+1}(\Omega), \qquad t \in [0, T],$$

and that $D^{-m}\tilde{v}(\cdot\,,t)$ is a continuous function of t relative to the norm $\|\cdot\|_{H^{m+1}(\Omega)}$. Since $\tilde{v}(\cdot\,,t)$ is a generalized solution of $\tilde{v}_{tt} = B\tilde{v}$ (i.e., (2.2)) one can verify without difficulty that $D^{-m}\tilde{v}$ satisfies the same equation (in the strict sense if $m > 0$) and that

$$\frac{d^m}{dt^m}(D^{-m}\tilde{v}(\cdot\,,t)) = \tilde{v}(\cdot\,,t).$$

Next we define

$$D^{-m}v(\cdot\,,t) = D^{-m}\tilde{v}(\cdot\,,t) + c_1\frac{(t - T)^m}{m!} + c_2\frac{(t - T)^{m+1}}{(m + 1)!} \tag{2.6}$$

and verify that

$$\frac{d^m}{dt^m}(D^{-m}v(\cdot\,,t)) = v(\cdot\,,t).$$

It is not in general true that $D^{-m}v(\cdot\,,t)$ is a solution of $v_{tt} = Bv$. But since c_1 and c_2 are constants, it is clear that we still have

$$D^{-m}v(\cdot\,,t) \in H^{m+1}(\Omega), \qquad t \in [0, T], \quad m \geqq 0. \tag{2.7}$$

We now refer to the theorem of Sobolev (see, e.g., [1, p. 32]) which states that if $v \in H^m(\Omega)$ and if l is a positive integer strictly less than $m - n/2$, then $v \in C^l(\Omega)$. Moreover, there is a constant K, independent of v, such that

$$\|v\|_{C^l(\bar{\Omega})} \leqq K\|v\|_{H^m(\Omega)}. \tag{2.8}$$

We choose $m = 2k$ to be a positive integer such that $m - n/2 > 1$. Then from (2.7) and the Sobolev theorem we have

$$D^{-m}v(\cdot\,,t) \in C^2(\bar{\Omega}), \qquad D^{-m+1}v(\cdot\,,t) = D^{-m}v_t(\cdot\,,t) \in C^1(\bar{\Omega}).$$

The continuity of $D^{-m}v(\cdot\,,t)$, $D^{-m+1}v(\cdot\,,t)$, as functions of t, with respect to $\|\cdot\|_{H^{m+1}(\Omega)}$, $\|\cdot\|_{H^m(\Omega)}$, respectively, combined with (2.8), then shows that $D^{-m}v(\cdot\,,t) \in C^2(\bar{\Omega} \otimes [0, T])$.

Now, for $\delta > 0$, we define

$$\Delta(D^{-m}v(\cdot,t)) = D^{-m}v(\cdot,t+\delta) - D^{-m}v(\cdot,t), \qquad t \in [0, T-\delta],$$

$$\Delta^k(D^{-m}v(\cdot,t)) = \Delta(\Delta^{k-1}(D^{-m}v(\cdot,t))), \qquad t \in [0, T-k\delta].$$

Noting (2.6), the fact that $D^{-m}\tilde{v}$ solves $\tilde{v}_{tt} = B\tilde{v}$, and the fact that $v \in C^2(\bar{\Omega} \otimes [0, T])$, we see that $\Delta^m(D^{-m}v(\cdot,t))$, which we will call $v^\delta(\cdot,t)$, is such that $v^\delta(x,t)$ is a C^2-solution of (2.2), (2.3) in $\bar{\Omega} \otimes [0, T]$.

The rest of the proof proceeds much as in [13] and we will give an outline only. The interested reader should consult the earlier paper for details, noting that there $\tilde{\Gamma}$ of this paper was called Γ_1.

Using the divergence theorem one shows that (2.1) implies (with D denoting $\partial/\partial t$)

$$\int_{\tilde{\Gamma} \otimes [0,T]} [D^{-m+1}v(x(s),t)D^m f(s,t)]\, dx\, dt = 0 \tag{2.9}$$

for all admissible controls f. This implies that $D^{-m+1}v(x(s),t) = (D^{-m}v(x(s),t))_t$ is a polynomial in t of degree at most $m-1$ whose coefficients are C^1-functions of $x(s)$, for $(x(s),t) \in \tilde{\Gamma} \otimes [0, T]$. Then

$$(\Delta^m(D^{-m}v(x(s),t)))_t = \Delta^m((D^{-m}v(x(s),t))_t) \equiv 0, \qquad (x(s),t) \in \tilde{\Gamma} \otimes [0, T-m\delta].$$

This, combined with the fact that $\Delta^m(D^{-m}v)$ satisfies the boundary condition (2.3), enables us to use the Holmgren–Fritz John uniqueness theorem [7] to show that $(\Delta^m(D^{-m}v))_t$ must vanish identically for $(x,t) \in K(\tilde{\Gamma}, 0, T-m\delta)$, the intersection of the forward cone of influence of $\tilde{\Gamma}$ at time 0 with the backward cone of influence of $\tilde{\Gamma}$ at time $T-m\delta$. If $T > 2T_0$, the set $K(\tilde{\Gamma}, 0, T-m\delta)$ includes a set $\bar{\Omega} \otimes [(T/2)-\varepsilon, (T/2)+\varepsilon]$ for some $\varepsilon > 0$, provided $\delta > 0$ is sufficiently small. (See figures in [13].) Thus,

$$(\Delta^m(D^{-m}v(x,t)))_t \equiv 0, \qquad (x,t) \in \bar{\Omega} \otimes [(T/2)-\varepsilon, (T/2)+\varepsilon],$$

which clearly implies

$$(\Delta^m(D^{-m}v(x,t)))_{tt} \equiv 0, \qquad (x,t) \in \bar{\Omega} \otimes [(T/2)-\varepsilon, (T/2)+\varepsilon].$$

Since $\Delta^m(D^{-m}v)$ is a C^2-solution of (2.2), (2.3) we conclude that

$$v(x,t) \equiv v(x), \qquad (x,t) \in \bar{\Omega} \otimes [(T/2)-\varepsilon, (T/2)+\varepsilon],$$

where $v(x)$ is a C^2-solution of the elliptic boundary value problem

$$\sum_{i,j=1}^{n} (\alpha_{ij}(x)u_i)_j = 0, \qquad x \in \Omega, \tag{2.10}$$

$$u_x(x(s))A(x(s))\eta(x(s)) = 0, \qquad x(s) \in \Gamma. \tag{2.11}$$

But the only solutions of (2.10), (2.11) have the form

$$u(x) = c, \quad \text{a constant}, \quad x \in \Omega.$$

Thus

$$\Delta^m(D^{-m}v(x,t)) = c, \qquad (x,t) \in \Omega \otimes [(T/2)-\varepsilon, (T/2)+\varepsilon],$$

so that $D^{-m}v(x,t)$ is a polynomial in t of degree at most m whose coefficients are C^2-functions of x for $x \in \Omega$. Then $v(x,t) = D^m(D^{-m}v(x,t))$ is a constant in $\Omega \otimes [(T/2) - \varepsilon, (T/2) + \varepsilon]$. In particular,

$$(v(\cdot, T/2), v_t(\cdot, T/2)) = (c, 0),$$

a zero energy state. Applying the conservation of energy principle, which is valid for generalized solutions of (2.2), (2.3), we infer that

$$(v(\cdot, T), v_t(\cdot, T)) = (v, v_t) = (c, 0).$$

We see therefore that if (2.1) holds for all admissible controls f, so that (v, v_t) is orthogonal, relative to the energy inner product $\langle \cdot\,; \cdot \rangle_E$, to every state in R_T, then $\|(v, v_t)\|_E = 0$ and (v, v_t) is the null element in $H_E(\Omega)$. We have proved this without making any special assumptions on n, the dimension of the space in which Ω lies. Thus Theorem 4 of [13] can be replaced by the following stronger theorem.

THEOREM 4a. *The system* (1.1), (1.2) *is approximately controllable in time* T *if* $T > 2T_0$.

Combined with Theorem 2 of [13], which states that the system (1.1), (1.2) is not approximately controllable in time T if $T < 2T_0$, we see that we are justified in referring to $2T_0$ as the critical time. We will see in the sequel that, if $n \geqq 2$, critical time approximate controllability is a rather delicate question.

3. The critical time control problem. We are going to study the problem for a particular partial differential equation in certain special domains. In § 6 we will indicate a more general theory.

In R^n, $n \geqq 2$, we consider "rectangles" Σ_r, $r = 1, 2, \cdots, n$, of dimension $n - r$, defined by

$$\Sigma_r = \{x = (x^1, x^2, \cdots, x^n) \in R^n | x^i = 0,\ i = 1, \cdots, r,\ 0 \leqq x^j \leqq 1,\ j = r+1, \cdots, n\}.$$

Of course, Σ_n is just the origin in R^n. For all real ξ we define

$$\tilde{\rho}(\xi) = \exp(1 - 1/\xi^2) \quad (\equiv 0 \text{ if } \xi = 0),$$

and for all $x = (x^1, x^2, \cdots, x^n) \in R^n$ we put

$$\rho(x) = \tilde{\rho}(x^1) + \tilde{\rho}(x^2) + \cdots + \tilde{\rho}(x^n).$$

We define domains $\Omega_r \subseteq R^n$ as follows:

$$\Omega_r = \{x \in R^n | \inf_{y \in \Sigma_r} \rho(x - y) < 1\}.$$

Then Ω_r is an open, bounded, simply connected region in R^n whose boundary

$$\Gamma_r = \{x \in R^n | \inf_{y \in \Sigma_r} \rho(x - y) = 1\}$$

is an n-dimensional surface of class C^∞ which is piecewise analytic.

In Ω_r we consider a boundary value control problem for the ordinary wave equation:

$$w_{tt}^f - \sum_{i=1}^{n} w_{ii}^f = 0 \quad \text{in } \Omega_r \otimes [0, T], \tag{3.1}$$

$$w_x^f(x(s), t)\eta(x(s)) = f(s, t), \quad (x(s), t) \in \Gamma_r \otimes [0, T], \tag{3.2}$$

$$w^f(x, 0) \equiv w_t^f(x, 0) \equiv 0, \quad x \in \Omega_r. \tag{3.3}$$

We take $\tilde{\Gamma} = \Gamma_r$, i.e., admissible controls are C^∞-functions whose supports are compact subsets of the interior of $\Gamma_r \otimes [0, T]$. Thus control forces operate over the whole boundary of Ω_r.

For (3.1) there is a universal wave propagation speed, 1. Thus, given an instant t_0, the forward cone of influence of Γ_r at time t_0 is given by

$$K^+(\Gamma_r, t_0) = \{(x, t) \in \Omega_r \otimes [t_0, +\infty] \| \inf_{y \in \Gamma_r} \|x - y\| \leqq t - t_0\} \tag{3.4}$$

and the backward cone of influence of Γ_r at time t_0 is

$$K^-(\Gamma_r, t_0) = \{(x, t) \in \Omega_r \otimes (-\infty, t_0] | (x, 2t_0 - t) \in K^+(\Gamma_r, t_0)\}.$$

(In (3.4) $\| \cdot \|$ denotes the Euclidean norm in R^n.) We define, for $T > 0$,

$$K(\Gamma_r, 0, T) = K^+(\Gamma_r, 0) \cap K^-(\Gamma_r, T).$$

As shown in [13, § 3], the critical time T_0 has the property that

$$\Omega_r \otimes \{T_0\} \subseteq K(\Gamma_r, 0, 2T_0)$$

but $\Omega_r \otimes \{T/2\}$ is not a subset of $K(\Gamma_r, 0, T)$ if $T < 2T_0$. In the present case it follows that $T_0 = 1$, and hence the critical time is $T = 2$, because

$$\sup_{x \in \Omega_r} \{\inf_{y \in \Gamma_r} (\|x - y\|)\} = 1.$$

We shall prove two theorems regarding approximate controllability of (3.1), (3.2) in the critical time $T = 2$. We give these theorems the numbers 5 and 6 since they complement the four theorems proved in [13] and § 2 of the present paper.

THEOREM 5. *If $r = 1$, the system* (3.1), (3.2) *is not approximately controllable in the critical time $T = 2$.*

THEOREM 6. *If $2 \leqq r \leqq n$, the system* (3.1), (3.2) *is approximately controllable in the critical time $T = 2$.*

The reader should be aware that these theorems apply for $n \geqq 2$ only. When $n = 1$ the analogue of Theorem 5 is not true, for it has already been shown in [5], [14], [15] that we do have critical time approximate controllability in this case.

In order to prove Theorems 5 and 6 we need certain results from the theory of distributions.

4. Distributions in $H^{-1}(\Omega_r)$ with support in Σ_r. As in § 2, we denote by $H^1(\Omega_r)$ real-valued functions $v(x)$ defined on Ω_r which lie in $H^0(\Omega_r) = L^2(\Omega_r)$ and have first order partial derivatives, defined in the sense of the theory of

distributions, which also lie in $H^0(\Omega_r)$. With the inner product

$$(u, v)_{H^1(\Omega_r)} = \int_{\Omega_r} \left[u(x)v(x) + \sum_{i=1}^{n} u_i(x)v_i(x) \right] dx$$

(again the subscript i refers to partial differentiation with respect to x^i) $H^1(\Omega_r)$ is a Hilbert space. We have

$$H^1(\Omega_r) \subseteq H^0(\Omega_r)$$

and for each $v \in H^1(\Omega_r)$,

$$\|v\|_{H^1(\Omega_r)} \geqq \|v\|_{H^0(\Omega_r)},$$

which shows that the injection mapping of $H^1(\Omega_r)$ into $H^0(\Omega_r)$ is continuous.

We will now indicate the construction of a third Hilbert space $H^{-1}(\Omega_r)$ with

$$H^1(\Omega_r) \subseteq H^0(\Omega_r) \subseteq H^{-1}(\Omega_r),$$

and the injection of $H^0(\Omega_r)$ into $H^{-1}(\Omega_r)$ is likewise continuous. To begin, let $u \in H^0(\Omega_r)$. We define a continuous linear functional on $H^0(\Omega_r)$:

$$l_u(v) = (u, v)_{H^0(\Omega_r)}, \qquad v \in H^0(\Omega_r). \tag{4.1}$$

Now if $v \in H^1(\Omega_r)$,

$$|l_u(v)| \leqq \|u\|_{H^0(\Omega_r)}\|v\|_{H^0(\Omega_r)} \leqq \|u\|_{H^0(\Omega_r)}\|v\|_{H^1(\Omega_r)}$$

and we conclude that (4.1) also defines l_u as a continuous linear functional on $H^1(\Omega_r)$. It follows that there is a unique element $\hat{u} \in H^1(\Omega_r)$ such that

$$l_u(v) = (\hat{u}, v)_{H^1(\Omega_r)}. \tag{4.2}$$

We define

$$\|u\|_{H^{-1}(\Omega_r)} = \|\hat{u}\|_{H^1(\Omega_r)}. \tag{4.3}$$

Now for all $u \in H^0(\Omega_r)$,

$$\begin{aligned}
\|u\|_{H^{-1}(\Omega_r)} &= \sup_{\substack{v \in H^1(\Omega_r) \\ v \neq 0}} \frac{|(\hat{u}, v)_{H^1(\Omega_r)}|}{\|v\|_{H^1(\Omega_r)}} \\
&= \sup_{\substack{v \in H^1(\Omega_r) \\ v \neq 0}} \frac{|(u, v)_{H^0(\Omega_r)}|}{\|v\|_{H^1(\Omega_r)}} \leqq \sup_{\substack{v \in H^1(\Omega_r) \\ v \neq 0}} \frac{|(u, v)_{H^0(\Omega_r)}|}{\|v\|_{H^0(\Omega_r)}} \\
&= \sup_{\substack{v \in H^0(\Omega_r) \\ v \neq 0}} \frac{|(u, v)_{H^0(\Omega_r)}|}{\|v\|_{H^0(\Omega_r)}} = \|u\|_{H^0(\Omega_r)},
\end{aligned}$$

the second last equality being true because $H^1(\Omega_r)$ is dense in $H^0(\Omega_r)$ relative to the topology induced by the norm $\|\cdot\|_{H^0(\Omega_r)}$.

We define $H^{-1}(\Omega_r)$ to be the completion of $H^0(\Omega_r)$ relative to the norm $\|\cdot\|_{H^{-1}(\Omega_r)}$. Now $\|u\|_{H^{-1}(\Omega_r)} = \|\hat{u}\|_{H^1(\Omega_r)}$ holds for $u \in H^0(\Omega_r)$, which is clearly dense in $H^{-1}(\Omega_r)$, and this relationship extends (see [12]) to an isometry $u \leftrightarrow \hat{u}$

between $H^{-1}(\Omega_r)$ and $H^1(\Omega_r)$. The space $H^{-1}(\Omega_r)$ is a Hilbert space with

$$(u, v)_{H^{-1}(\Omega_r)} = (\hat{u}, \hat{v})_{H^1(\Omega_r)}. \tag{4.4}$$

The elements ϕ of $H^{-1}(\Omega_r)$ correspond to distributions l_ϕ of order at most 1 (see [16]) on Ω_r.

We are now ready to prove two lemmas which will be of great importance in the proofs of Theorems 5 and 6.

LEMMA 1. *If $n \geqq 2$, there exists a nontrivial element $\phi \in H^{-1}(\Omega_1)$ such that:*
(i) *the support of l_ϕ is a subset of Σ_1;*
(ii) *if c is a constant function on Ω_1, then $l_\phi(c) \equiv (\hat{\phi}, c)_{H^1(\Omega_1)} = 0$.*

LEMMA 2. *If $2 \leqq r \leqq n$, there is no nontrivial distribution in $H^{-1}(\Omega_r)$ with support in Σ_r.*

Proof of Lemma 1. Let ψ denote a real-valued function of $n-1$ variables $x^2, x^3, \cdots, x^n$ such that, with $\tilde{\Sigma}_1$ defined by

$$\tilde{\Sigma}_1 = \{\tilde{x} = (x^2, \cdots, x^n) \in R^{n-1} | (0, \tilde{x}) \in \Sigma_1\},$$

$\psi \in C^2(\tilde{\Sigma}_1)$ vanishes outside a compact subset of the interior of $\tilde{\Sigma}_1$, and

$$\int_{\tilde{\Sigma}_1} \psi(\tilde{x})\, d\tilde{x} = 0, \qquad \int_{\tilde{\Sigma}_1} (\psi(\tilde{x}))^2\, d\tilde{x} \neq 0. \tag{4.5}$$

For positive integers $k = 4, 5, 6, \cdots$ define

$$\tilde{\theta}_k(\xi) = \begin{cases} 0, & -1 \leqq \xi \leqq -3/4, \\ -(1/2)(\xi + 3/4)^2, & -3/4 \leqq \xi \leqq -1/4, \\ -(1/2)\xi - 1/4, & -1/4 \leqq \xi \leqq 1/k, \\ (k/4)\xi^2 + 1/(4k) - 1/4, & -1/k \leqq \xi \leqq 1/k, \\ \tilde{\theta}_k(-\xi), & 1/k \leqq \xi \leqq 1. \end{cases} \tag{4.6}$$

Then, for $x \in \Omega_1$, put

$$\theta_k(x) = \theta_k(x^1, \tilde{x}) = \begin{cases} 0 & \text{if } \tilde{x} \notin \tilde{\Sigma}_1, \\ \tilde{\theta}_k(x^1)\psi(\tilde{x}) & \text{if } x \in \tilde{\Sigma}_1. \end{cases} \tag{4.7}$$

Then θ_k is defined as a function of class C^2 in Ω_1 for $k = 4, 5, 6, \cdots$. Now compute, for any $v \in H^1(\Omega_1)$,

$$\begin{aligned} \int_{\Omega_1} \frac{\partial \theta_k(x)}{\partial x^1} \frac{\partial v(x)}{\partial x^1}\, dx &= -\int_{\Omega_1} \frac{\partial^2 \theta_k(x)}{(\partial x^1)^2} v(x)\, dx \\ &= -\left[\int_{[-3/4, -1/4] \otimes \tilde{\Sigma}} -\psi(\tilde{x}) v(x)\, dx + \int_{[-1/k, 1/k] \otimes \tilde{\Sigma}} \frac{k}{2} \psi(\tilde{x}) v(x)\, dx \right. \\ &\qquad \left. + \int_{[1/4, 3/4] \otimes \tilde{\Sigma}} -\psi(\tilde{x}) v(x)\, dx \right]. \end{aligned}$$

(In each case the set described beneath the integral is the domain of integration

for that integral.) Thus

$$\int_{[-1/k,1/k]\otimes\tilde{\Sigma}} \frac{k}{2}\psi(\tilde{x})v(x)\,dx = \int_{[-3/4,-1/4]\otimes\tilde{\Sigma}} \psi(\tilde{x})v(x)\,dx + \int_{[1/4,3/4]\otimes\tilde{\Sigma}} \psi(\tilde{x})v(x)\,dx - \int_{\Omega_1} \frac{\partial\theta_k(x)}{\partial x^1}\frac{\partial v(x)}{\partial x^1}\,dx$$

and, for $k = 4, 5, 6, \cdots$, $j = 4, 5, 6, \cdots$,

$$\int_{[-1/k,1/k]\otimes\tilde{\Sigma}} \frac{k}{2}\psi(\tilde{x})v(x)\,dx - \int_{[-1/j,1/j]\otimes\tilde{\Sigma}} \frac{j}{2}\psi(\tilde{x})v(x)\,dx = \int_{\Omega_1}\left(\frac{\partial\theta_j(x)}{\partial x^1} - \frac{\partial\theta_k(x)}{\partial x^1}\right)\frac{\partial v(x)}{\partial x^1}\,dx.$$

Applying the Schwarz inequality, we have

$$\begin{aligned}\left|\int_{[-1/k,1/k]\otimes\tilde{\Sigma}} \frac{k}{2}\psi(\tilde{x})v(x)\,dx - \int_{[-1/j,1/j]\otimes\tilde{\Sigma}} \frac{j}{2}\psi(\tilde{x})v(x)\,dx\right| &\leqq \left\|\frac{\partial\theta_j}{\partial x^1} - \frac{\partial\theta_k}{\partial x^1}\right\|_{H^0(\Omega_1)}\left\|\frac{\partial v}{\partial x^1}\right\|_{H^0(\Omega_1)} \\ &\leqq \|\theta_j - \theta_k\|_{H^1(\Omega_1)}\|v\|_{H^1(\Omega_1)}.\end{aligned} \tag{4.8}$$

An inspection of (4.6), (4.7) readily shows that

$$\|\theta_j - \theta_k\|_{H^1(\Omega)} = \varepsilon_{jk},$$

where

$$\lim_{\substack{j\to\infty \\ k\to\infty}} \varepsilon_{jk} = 0.$$

Let us put

$$\phi_k(x) = \phi_k(x^1, \tilde{x}) = \begin{cases} \dfrac{k}{2}\psi(\tilde{x}), & (x^1, \tilde{x}) \in \left[-\dfrac{1}{k}, \dfrac{1}{k}\right] \otimes \tilde{\Sigma}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.9}$$

Then (4.8) shows that the continuous linear functionals l_{ϕ_k}, defined on $H^1(\Omega_1)$ as in (4.1), (4.2), have the property that

$$|l_{\phi_k}(v) - l_{\phi_j}(v)| \leqq \varepsilon_{jk}\|v\|_{H^1(\Omega_1)}$$

which implies (cf. (4.2)) that

$$\|\hat{\phi}_k - \hat{\phi}_j\|_{H^1(\Omega_1)} \leqq \varepsilon_{jk},$$

and therefore, from (4.3),

$$\|\phi_k - \phi_j\|_{H^{-1}(\Omega_1)} \leqq \varepsilon_{jk}.$$

Thus, in $H^{-1}(\Omega_1)$, $\{\phi_k\}$ is a Cauchy sequence and has a limit $\phi \in H^{-1}(\Omega_1)$. It remains only to show that ϕ has properties (i) and (ii) stated in Lemma 1.

Let $v \in C^\infty(\Omega_1)$ have support K which is a compact subset of $\Omega_1 - \Sigma_1$. Then, for all sufficiently large k, $K \cap ([-1/k, 1/k] \otimes \tilde{\Sigma})$ is empty and

$$l_{\phi_k}(v) = (\phi_k, v)_{H^0(\Omega_1)} = (\hat{\phi}_k, v)_{H^1(\Omega_1)} = 0.$$

Since ϕ_k converges to ϕ in $H^{-1}(\Omega_1)$, $\hat{\phi}_k$ converges to $\hat{\phi}$ in $H^1(\Omega_1)$, by virtue of the isometry discussed just prior to (4.4). Therefore

$$l_\phi(v) = (\hat{\phi}, v)_{H^1(\Omega_1)} = \lim_{k \to \infty} (\hat{\phi}_k, v) = 0.$$

Thus l_ϕ vanishes when applied to $v \in C^\infty(\Omega_1)$ with support K not meeting Σ_1, and we have shown that the support of l_ϕ must be a subset of Σ_1.

Similarly, for $k = 4, 5, 6, \cdots, c$ constant,

$$l_{\phi_k}(c) = (\phi_k, c)_{H^0(\Omega_1)} = \int_{[-1/k, 1/k] \otimes \tilde{\Sigma}_1} \frac{k}{2} \psi(\tilde{x}) c \, dx^1 \, d\tilde{x}$$

$$= c \int_{\tilde{\Sigma}_1} \psi(\tilde{x}) \, d\tilde{x} = 0,$$

from (4.5). Thus part (ii) of Lemma 1 is proved. The second part of (4.5) readily shows that ϕ is nontrivial, and the proof of Lemma 1 is complete.

Proof of Lemma 2. For $p > 0$ we define

$$h_p(x^1, x^2, \cdots, x^r) = 1 - [(x^1)^2 + (x^2)^2 + \cdots + (x^r)^2]^{1/p}. \tag{4.10}$$

We compute

$$\sum_{i=1}^{r} \left(\frac{\partial h_p}{\partial x^i} \right)^2 = \frac{4}{p^2} [(x^1)^2 + (x^2)^2 + \cdots + (x^r)^2]^{2/p - 1}. \tag{4.11}$$

Integrating (4.11) over the unit ball in R^r we obtain the integral

$$4\omega_{r-1}/(4p + (r - 2)p^2),$$

where ω_{r-1} is the integral of 1 over the $(r - 1)$-dimensional sphere of radius 1. Thus we see that if B is any bounded open set in R^r, then $h_p \in H^1(B)$ for $p > 0$ and

$$\lim_{p \to +\infty} \|h_p\|_{H^1(B)} = 0. \tag{4.12}$$

(Note that $r \geqq 2$ is necessary for these conclusions.)

Given $x = (x^1, \cdots, x^r, x^{r+1}, \cdots, x^n) \in R^n$, let us set $y = (x^1, \cdots, x^r)$, $z = (x^{r+1}, \cdots, x^n)$. Each distribution l defined on R^{n-r} has a natural extension to a distribution $\hat{l}$ defined on R^n. If $\hat{v}(= \hat{v}(y, z)) \in C^\infty(R^n)$, we let $\hat{v}$ be defined on R^{n-r} by $v(z) = v(0, z)$. Then $\hat{l}(\hat{v}) = l(v)$ defines the extension $\hat{l}$ of l.

A result in [9, p. 78] shows that if $\phi \in H^{-1}(\Omega_r)$ then the distribution l_ϕ associated with ϕ can be expressed in the form

$$l_\phi(u) = (g_0, u)_{H^0(\Omega_r)} + \sum_{i=1}^{n} \left(g_i, \frac{\partial u}{\partial x^i} \right)_{H^0(\Omega_r)}, \qquad u \in C^\infty(\Omega_r),$$

where $g^i \in H^0(\Omega_r)$, $i = 0, 1, \cdots, n$. This shows that l_ϕ is a distribution or order at most 1 (i.e., if the v_k are C^∞-functions converging to zero in $C^1(\Omega_r)$ as $k \to \infty$, then $\lim_{k \to \infty} l_\phi(v_k) = 0$).

Thus if we take $\phi \in H^{-1}(\Omega)$, l_ϕ has order at most 1. A theorem in [16, p. 99 ff.] then shows that if the support of l_ϕ is a subset of Σ_r, there are distributions $l_0, l_1, l_2, \cdots, l_r$ defined on R^{n-r} with support in $\tilde{\Sigma}_r = \{z | (0, z) \in \Sigma_r\}$ such that

$$l_\phi = \hat{l}_0 + \sum_{i=1}^{r} \frac{\partial \hat{l}_i}{\partial x^i}.$$

Let $\psi(z) \in C^\infty(R^{n-r})$ have support K contained in some small neighborhood of $\tilde{\Sigma}_r$ in R^{n-r}. Define $v(= v(y, z))$ in Ω_r by

$$v(x) = v(y, z) = h_4(y)\psi(z), \tag{4.13}$$

where h_4 is given by (4.10). Then $v \in H^1(\Omega_r)$. Since $\phi \in H^{-1}(\Omega_r)$, the linear functional l_ϕ can be defined on all of $H^1(\Omega_r)$ and we have

$$\begin{aligned} l_\phi(v) &= \hat{l}_0(v) + \sum_{i=1}^{r} \left(\frac{\partial \hat{l}_i}{\partial x^i}\right)(v) \\ &= \hat{l}_0(v) - \sum_{i=1}^{r} \hat{l}_i\left(\frac{\partial v}{\partial x^i}\right). \end{aligned}$$

Let $e_j = (0, \cdots, 0, 1, 0, \cdots, 0)$ in R^r, the 1 being in the jth position. Define

$$v_\varepsilon(x) = v_\varepsilon(y, z) = v(y + \varepsilon e_j, z).$$

One verifies without difficulty that

$$\lim_{\varepsilon \to 0} \|v - v_\varepsilon\|_{H^1(\Omega_r)} = 0. \tag{4.14}$$

On the other hand,

$$\begin{aligned} l_\phi(v_\varepsilon) &= \hat{l}_0(v_\varepsilon) - \sum_{i=1}^{r} \hat{l}_i\left(\frac{\partial v_\varepsilon}{\partial x^i}\right) \\ &= h_4(0 + \varepsilon e_j) l_0(\psi) - \sum_{i=1}^{r} \frac{\partial h_4}{\partial x^i}(0 + \varepsilon e_j) l_i(\psi) \\ &= h_4(\varepsilon e_j) l_0(\psi) - \frac{\partial h_4}{\partial x^i}(\varepsilon e_j) l_j(\psi) \\ &= (1 - \varepsilon^{1/2}) l_0(\psi) + \tfrac{1}{2}\varepsilon^{-1/2} l_j(\psi). \end{aligned}$$

Thus if $l_j(\psi)$ is different from zero,

$$\lim_{\varepsilon \to 0} l_\phi(v_\varepsilon) = +\infty$$

and then (4.14) shows that l_ϕ cannot be a continuous linear functional on $H^1(\Omega_r)$, contrary to our assumption that $\phi \in H^{-1}(\Omega_r)$. We conclude that $l_j(\psi) = 0$. Since this is true for all such ψ and the support of l_j is a subset of $\tilde{\Sigma}_r$, we conclude that $l_j = 0$. We can do this for $j = 1, 2, \cdots, r$ and conclude that

$$l_\phi = \hat{l}_0.$$

Now, for $p > 0$, define $v_p(x)$ as in (4.13), replacing 4 by p. Compute

$$l_\phi(v_p) = h_p(0) l_0(\psi) = l_0(\psi)$$

for all $p > 0$, since $h_p(0) = 1$. But (4.12) is easily seen to imply that

$$\lim_{p \to \infty} \|v_p\|_{H^1(\Omega_r)} = 0$$

and thus l_ϕ cannot be a continuous linear functional on $H^1(\Omega_r)$ unless $l_0(\psi) = 0$. Therefore, since l_ϕ is a continuous linear functional on $H^1(\Omega_r)$, $l_0(\psi) = 0$ for all ψ of the form prescribed above. Then the fact that l_0 has support in $\tilde{\Sigma}_r$ shows that $l_0 = 0$. We have now shown that

$$l_\phi = 0,$$

and Lemma 2 has been proved.

Remarks. Some readers may find the dual role of l_ϕ, as a distribution of order $\leqq 1$ and as a linear functional on $H^1(\Omega_r)$, slightly confusing. Given $\phi \in H^{-1}(\Omega_r)$ there is associated with it a unique element $\hat{\phi} \in H^1(\Omega_r)$ and for all $v \in H^1(\Omega_r)$,

$$l_\phi(v) = (\hat{\phi}, v)_{H^1(\Omega_r)}.$$

This also defines l_ϕ as a continuous linear functional on $C^\infty(\Omega_r)$, since convergence in $C^\infty(\Omega_r)$ implies convergence in $H^1(\Omega_r)$. Thus l_ϕ is also a distribution in the sense of Schwartz [16].

One can easily see that Lemma 1, part (i) continues to hold for $n = 1$. (Just put $\phi = \delta$, the Dirac distribution.) But (ii) cannot hold for $n = 1$. The function ψ cannot be constructed as in (4.5). This explains why Theorem 5 is true for $n \geqq 2$ but not for $n = 1$.

5. Proof of Theorem 5. A result in Lions–Magenes [9, p. 202] states that if $\tilde{\phi} \in H^{-1}(\Omega)$ satisfies (ii) of Lemma 1, then there is a unique function $\tilde{v} \in H^1(\Omega)$ with $\int_{\Omega_1} \tilde{v}(x)\, dx = 0$ such that, in the sense of the theory of distributions,

$$\sum_{i=1}^{n} \tilde{v}_{ii} = \tilde{\phi} \quad \text{in } \Omega_1, \tag{5.1}$$

$$\tilde{v}_x(x(s))\eta(x(s)) = 0, \qquad x(s) \in \Gamma_1. \tag{5.2}$$

The sense in which (5.2) holds is also explained in [9]. In our applications $\tilde{v}$ is harmonic outside a compact subset of Ω_1 and (5.2) holds in the classical sense. Moreover, there is a constant $M > 0$ such that

$$\|\tilde{v}\|_{H^1(\Omega_1)} \leqq M \|\phi\|_{H^{-1}(\Omega_1)}. \tag{5.3}$$

Let the functions ϕ_k be defined on Ω_1 as in (4.9) and let $\tilde{v}^k$ be the corresponding solutions of (5.1), (5.2) with $\tilde{\phi}$ replaced by ϕ_k. Also, let $\tilde{v}$ satisfy (5.1), (5.2) with $\tilde{\phi}$ replaced by the element $\phi \in H^{-1}(\Omega_1)$ constructed in Lemma 1. Since $\lim_{k \to \infty} \|\phi - \phi_k\|_{H^{-1}(\Omega)} = 0$, (5.3) implies that

$$\lim_{k \to \infty} \|\tilde{v} - \tilde{v}^k\|_{H^1(\Omega_1)} = 0.$$

It is clear that $\tilde{v}$ cannot be a constant on Ω_1; therefore $(\tilde{v}, 0)$ is a nonzero energy state. We let $v(x, t)$, $v^k(x, t)$, $k = 4, 5, 6, \cdots$, be generalized solutions in $\Omega_1 \otimes [0, 2]$ of

$$v_{tt} - \sum_{i=1}^{n} v_{ii} = 0, \tag{5.4}$$

$$v_x(x(s), t)\eta(x(s)) = 0, \qquad (x(s), t) \in \Gamma_1 \otimes [0, 2], \tag{5.5}$$

satisfying

$$v(x, 1) \equiv \tilde{v}(x), \quad v_t(x, 1) \equiv 0, \quad v^k(x, 1) \equiv \tilde{v}^k(x), \quad v_t^k(x, 1) \equiv 0.$$

By the principle of conservation of energy, $(v(\cdot, 2), v_t(\cdot, 2))$ is also a nonzero energy state.

Let f be an admissible control. Then the support of f lies in a set $\Gamma_1 \otimes [\delta, 2 - \delta]$ for some $\delta > 0$. Since the support of ϕ_k is $[-1/k, 1/k] \otimes \tilde{\Sigma}_1$, $\tilde{v}^k$ is harmonic in $\Omega_1 - ([-1/k, 1/k] \otimes \tilde{\Sigma}_1)$. Then, by a familiar uniqueness result in the theory of hyperbolic partial differential equations (see, e.g., [2]),

$$v^k(x, t) \equiv \tilde{v}^k(x), \quad v_t^k(x, t) \equiv 0 \qquad \text{for } |t - 1| \leqq \inf_{y \in [-1/k, 1/k] \otimes \tilde{\Sigma}_1} \{\|x - y\|\}.$$

Thus, for sufficiently large k, $v_t^k(x(s), t) \equiv 0$, $(x(s), t) \in \Gamma_1 \otimes [\delta, 2 - \delta]$ and an application of the divergence theorem (cf. Theorem 1 in [13]) shows that

$$\begin{aligned} &\int_{\Omega_1} \left[w_t^f(x, 2)v_t^k(x, 2) + \sum_{i=1}^{n} w_i^f(x, 2)v_i^k(x, 2) \right] dx \\ &\qquad = \int_{\Gamma_1 \otimes [0,2]} v_t^k(x(s), t) f(s, t)\, ds = 0. \end{aligned} \tag{5.6}$$

(The solution $w^f \in C^\infty(\Omega_1 \otimes [0, 2])$ and it is proved in [10] that $v^k \in H^2(\Omega^1 \otimes [0, 2])$. This enables one to use the divergence theorem without difficulty.)

Noting that

$$\lim_{k \to \infty} \|v_t(\cdot, 2) - v_t^k(\cdot, 2)\|_{H^0(\Omega_1)} = 0,$$

$$\lim_{k \to \infty} \|v_i(\cdot, 2) - v_i^k(\cdot, 2)\|_{H^0(\Omega_1)} = 0, \qquad i = 1, 2, \cdots, n,$$

we conclude from (5.6) that

$$\int_{\Omega_1} \left[w_t^f(x, 2)v_t(x, 2) + \sum_{i=1}^{n} w_i^f(x, 2)v_i(x, 2) \right] dx = 0.$$

Since f is an arbitrary admissible control we have shown that the nonzero energy state $(v(\cdot, 2), v_t(\cdot, 2))$ lies in $R_2^\perp$ and thus that R_2 is not dense in $H_E(\Omega_1)$ relative to the norm $\|\cdot\|_E$. Thus Theorem 5 is proved.

6. Proof of Theorem 6. Much of the work necessary to prove Theorem 6 has already been done in § 2 in the proof of Theorem 4a. We again assume that (v, v_t) is a finite energy state which satisfies (2.1) (with $\rho \equiv 1$, $A \equiv I$ and $\Omega = \Omega_r$) and we let $v(x, t)$ be the generalized solution of (5.4), (5.5) satisfying the terminal conditions (2.4). The solution v is smoothed by the same process of forming antiderivatives and finite time differences as described in (2.5)–(2.8) ff. The divergence theorem can again be used to obtain (2.9) (with $\tilde{\Gamma}$ replaced by Γ_r), and thus, via the Holmgren–Fritz John uniqueness theorem [7] to prove that $(\Delta^m(D^{-m}v))_t$ must vanish identically for $(x, t) \in K(\Gamma_r, 0, T - m\delta)$, the intersection of the forward cone of influence of Γ_r at time 0 with the backward cone of influence of Γ_r at time $T - m\delta$.

The essential difference between the proof of Theorem 6 and that of Theorem 4a lies in the fact that when $T = 2$, the critical time, $K(\Gamma_r, 0, 2 - m\delta)$ does not include any set $\bar{\Omega}_r \otimes [1 - \varepsilon, 1 + \varepsilon]$ for any $\varepsilon > 0$, no matter how small we take $\delta > 0$ to be.

If $\delta > 0$ is small, the functions $\Delta^m(D^{-m}v(x, 1))$ are defined and twice continuously differentiable for $x \in \bar{\Omega}_r$. Now the operator Δ depends on δ, and we define

$$v^\delta(x) = \delta^{-m}\Delta^m(D^{-m}v(x, 1)), \qquad x \in \bar{\Omega}_r.$$

The continuity of $v(\cdot, t)$ as a mapping from R^1 into $H^1(\Omega_r)$ enables one to show by elementary means that

$$\lim_{\delta \to 0} \|v^\delta(x) - v(x, 1)\|_{H^1(\Omega_r)} = 0. \tag{6.1}$$

Now $\delta^{-m}\Delta^m(D^{-m}v(x, t)) \equiv v^\delta(x, t)$ is twice continuously differentiable in $\bar{\Omega}_r \otimes [0, 2 - m\delta]$ and there satisfies

$$\sum_{i=1}^{n} v^\delta_{ii} = v^\delta_{tt}$$

and boundary conditions of the form (5.5). Thus the functions

$$g^\delta(x) = v^\delta_{tt}(x, T_0)$$

are, for $\delta > 0$, continuous in $\bar{\Omega}_r$ and we have

$$\sum_{i=1}^{n} v^\delta_{ii}(x) \equiv g^\delta(x), \qquad x \in \bar{\Omega}_r.$$

Now $v^\delta_t(x, t)(= \delta^{-m}\Delta^m(D^{-m}v)_t(x, t))$ has been shown to vanish in $K(\Gamma_r, 0, 2 - \delta)$, which implies that $v^\delta_{tt}(x, t)$ vanishes there also. We conclude therefore that

$$g^\delta(x) \equiv v^\delta_{tt}(x, T_0) \equiv 0, \qquad x \in \Omega^\delta_r, \tag{6.2}$$

where

$$\Omega^\delta_r = \{x \in \Omega_r | (x, 1) \in K(\Gamma_r, 0, 2 - m\delta) \cap (\Omega_r \otimes \{1\})\}. \tag{6.3}$$

The sets Ω^δ_r are monotone increasing as $\delta \to 0$ with the property

$$\bigcap_{\delta > 0} (\Omega_r - \Omega^\delta_r) = \Sigma_r. \tag{6.4}$$

Let $u \in H^1(\Omega_r) \subseteq H^0(\Omega_r)$. Since $g^\delta \in C^0(\bar{\Omega}_r) \subseteq H^0(\Omega_r)$ we can form the inner product $(g^\delta, u)_{H^0(\Omega_r)}$. Integrating by parts we find that

$$\begin{aligned} |(g^\delta, u)_{H^0(\Omega_r)}| &\equiv \left| \left(\sum_{i=1}^{n} v^\delta_{ii}, u \right)_{H^0(\Omega_r)} \right| \\ &= \left| - \sum_{i=1}^{n} (v^\delta_i, u_i)_{H^0(\Omega_r)} \right| \leqq \|v^\delta\|_{H^1(\Omega_r)} \|u\|_{H^1(\Omega_r)}. \end{aligned}$$

Thus g^δ is an element of $H^0(\Omega_r)$ which defines, via $(g^\delta, u)_{H^0(\Omega_r)}$, a continuous linear functional l_{g^δ} on $H^1(\Omega_r)$ for which

$$\|l_{g^\delta}\| \leqq \|v^\delta\|_{H^1(\Omega_r)}.$$

There is an element $\hat{g}^\delta \in H^1(\Omega_r)$ such that

$$l_{g^\delta}(u) = (\hat{g}^\delta, u)_{H^1(\Omega_r)}, \qquad u \in H^1(\Omega_r).$$

Then, reasoning as in § 4, we have

$$\|g^\delta\|_{H^{-1}(\Omega_r)} = \|\hat{g}^\delta\|_{H^1(\Omega_r)} \leqq \|v^\delta\|_{H^1(\Omega_r)}.$$

Similarly for $\delta_1 > 0$, $\delta_2 > 0$,

$$\|g^{\delta_1} - g^{\delta_2}\|_{H^{-1}(\Omega_r)} \leqq \|v^{\delta_1} - v^{\delta_2}\|_{H^1(\Omega_r)}.$$

Now if we take a sequence $\{\delta_k\}$ of positive numbers with $\lim_{k\to\infty} \delta_k = 0$, we have

$$\lim_{k\to\infty} \|v^{\delta_k} - v(\cdot, 1)\|_{H^1(\Omega_r)} = 0$$

from (6.1). Thus

$$\lim_{\substack{k\to\infty \\ j\to\infty}} \|g^{\delta_k} - g^{\delta_j}\|_{H^{-1}(\Omega_r)} = \lim_{\substack{k\to\infty \\ j\to\infty}} \|v^{\delta_k} - v^{\delta_j}\|_{H^1(\Omega_r)} = 0$$

and we see that $\{g^{\delta_k}\}$ is Cauchy in $H^{-1}(\Omega_r)$, converging to an element $g \in H^{-1}(\Omega_r)$.

Let l_g be the distribution (also linear functional on $H^1(\Omega_r)$) associated with g. We claim that the support of g is contained in Σ_r. For, if $u \in C^\infty(\Omega_r)$ has support K which does not meet Σ_r, then (6.4) shows that K is a subset of Ω_r^δ for sufficiently small $\delta > 0$. Then

$$l_g(u) = \lim_{k\to\infty} l_{g^{\delta_k}}(u) = (g^{\delta_k}, u)_{H^0(\Omega_r)} = 0,$$

as we see from (6.2). Thus $l_g(u)$ vanishes whenever the support of $u \in C^\infty(\Omega_r)$ does not meet Σ_r and we conclude that the support of l_g lies in Σ_r.

In § 4 we showed that if $g \in H^{-1}(\Omega_r)$ and l_g has support in Σ_r, $2 \leqq r \leqq n$, then $g = 0$. Thus

$$0 = \|g\|_{H^{-1}(\Omega_r)} = \lim_{k\to\infty} \|g^{\delta_k}\|_{H^{-1}(\Omega_r)},$$

and for every $u \in H^1(\Omega_r)$,

$$\lim_{k\to\infty} l_{g^{\delta_k}}(u) = (g^{\delta_k}, u)_{H^0(\Omega_r)} = 0. \tag{6.5}$$

Setting $u = -v(\cdot, 1)$ in (6.5), we have

$$0 = \lim_{k\to\infty} (g^{\delta_k}, -v(\cdot, 1))_{H^0(\Omega_r)} = \lim_{k\to\infty} \left(\sum_{i=1}^n (v_i^{\delta_k}, v_i(\cdot, 1))_{H^0(\Omega_r)} \right).$$

Since v^{δ_k} converges to $v(\cdot, 1)$ in $H^1(\Omega_r)$ we have

$$0 = \lim_{k\to\infty} \left(\sum_{i=1}^n (v_i^{\delta_k}, v_i(\cdot, 1))_{H^0(\Omega_r)} \right) = \sum_{i=1}^n (v_i(\cdot, 1), v_i(\cdot, 1))_{H^0(\Omega_r)}$$

and we conclude that there is a constant c such that

$$v(x, T_0) \equiv c, \qquad x \in \Omega_r.$$

Since $v_t(\cdot, t)$ is a continuous mapping from $[0, 2]$ into $H^0(\Omega_r)$ one can show by elementary means that

$$\lim_{\delta \to 0} \|v_t^\delta(\cdot, 1) - v_t(\cdot, 1)\|_{H^0(\Omega_r)} = 0. \tag{6.6}$$

But $v_t^\delta(x, 1) \equiv 0$ for $x \in \Omega_r^\delta$ as we see from (6.3) and the fact that $v_t^\delta(x, t) \equiv 0$ in $K(\Gamma_r, 0, 2 - m\delta)$. Combined with (6.4) this shows that

$$\lim_{\delta \to 0} v_t^\delta(x, 1) = 0 \qquad \text{a.e. in } \Omega_r$$

and then (6.6) shows that

$$v_t(\cdot, T_0) = 0 \qquad \text{in } H_0(\Omega_r).$$

Thus $(v(\cdot, T_0), v_t(\cdot, T_0)) = (c, 0)$ is a zero energy state. The conservation of energy principle then shows that $(v(\cdot, 2), v_t(\cdot, 2))$ is likewise a zero energy state and, reasoning as in the proof of Theorem 4a, we see that (3.1), (3.2) is approximately controllable in time $T = 2$. Thus Theorem 6 is proved.

Remark. The Holmgren–Fritz John uniqueness theorem [7] cited here and in § 2 was originally proved under the assumption that the boundary Γ of Ω is analytic. The boundaries Γ_r of the sets Ω_r constructed in § 3 do not have this property—they are C^∞ and piecewise analytic. However, the results of [7] can be extended to such boundaries with very little difficulty. If $\Gamma_r = \Gamma_r^1 \cup \cdots \cup \Gamma_r^S$, where the Γ_r^k are relatively closed in Γ_r with disjoint relative interiors $\mathring{\Gamma}_r^k$, and if each $\mathring{\Gamma}_r^k$ is an analytic surface, then $(\Delta^m(D^{-m}v))_t \equiv 0$ on $\mathring{\Gamma}_r^k$ implies, via [7], that this identity continues to hold in $K(\mathring{\Gamma}_r^k, 0, 2 - m\delta)$. But the interior of $K(\mathring{\Gamma}_r^k, 0, 2 - m\delta)$ is included in the set $\cup_k K(\mathring{\Gamma}_r^k, 0, 2 - m\delta)$ and thus the continuous function $(\Delta^m(d^{-m}v))_t$ also vanishes in $K(\Gamma_r, 0, 2 - m\delta)$, as we need for our proof.

7. Concluding remarks. While Theorems 5 and 6 are stated for special domains Ω_r and a special hyperbolic partial differential equation, it is not difficult to extrapolate these results to systems of the form (1.1), (1.2) in more general domains Ω with boundary Γ which includes a relatively open subset $\tilde{\Gamma}$ whereon control is exercised.

Given the critical time $T = 2T_0$ one forms sets $K(\Gamma_1, 0, 2T_0 - m\delta)$ as in the proof of Theorem 6. (See [13] for complete description.) Then we form the sets

$$\Omega^\delta = \{x | (x, T_0) \in K(\Gamma_1, 0, 2T_0 - m\delta) \cap [\Omega \otimes \{T_0\}]\}.$$

As δ tends to zero the sets Ω^δ increase. The complementary sets $\Omega - \Omega^\delta$ decrease and we put

$$\Sigma = \bigcap_{\delta > 0} (\Omega - \Omega^\delta).$$

The dimension of Σ is what is critical. If Σ contains a smooth manifold of dimension $n - 1$, the system will not be controllable in time $T = 2T_0$, for one can construct a distribution $\phi \in H^{-1}(\Omega)$ with properties (i) and (ii) of Lemma 1, solve

$$\sum_{i,j=1}^{n} (\alpha_{ij}(x)\tilde{v}_i)_j = \phi,$$

set

$$v(x, T_0) \equiv \tilde{v}(x), \qquad v_t(x, T_0) \equiv 0 \tag{7.1}$$

and then let $v(x, t)$ be the generalized solution of (2.2), (2.3) satisfying (7.1). The state $(v(\cdot, 2T_0), v_t(\cdot, 2T_0))$ will then lie in $R^{\perp}_{2T_0}$ relative to the energy inner product in $H_E(\Omega)$. If Σ has dimension $n-2$ or less one can show, as in Lemma 2, that Σ cannot be the support of a nontrivial distribution in $H^{-1}(\Omega)$ and prove critical time controllability as in Theorem 6.

It is clear that in the "typical" case Σ will have dimension less than $n-1$. In fact, Σ will be a single point in many instances. It seems reasonable to conjecture that Σ cannot have dimension greater than $n-2$ if Γ is an analytic surface. Thus critical time approximate controllability is the rule, not the exception.

The results of [13] and the present paper leave the theory of *approximate* boundary value controllability of systems (1.1), (1.2) in a fairly satisfactory state. However, much remains to be done. Perhaps the most important task is that of characterizing all finite energy states which can be reached (from a zero initial state) in a time $T \geqq 2T_0$ using controls $f \in L^2(\Gamma_1 \otimes [0, T])$. A first step is to consider $f \in C^{\infty}(\Gamma_1 \otimes [0, T])$ as in the present paper and try to bound $\|f\|_{L^2(\Gamma_1 \otimes [0,T])}$ in terms of $w^f(\cdot, T)$ and its derivatives. Similar work has already been done in [5], [15] for the wave equation in one space dimension. Results in this direction would enable one to undertake a systematic study of the applicability of the quadratic criterion to hyperbolic boundary value control problems, as has been done, e.g., in [11] for the case of spatially distributed controls.

REFERENCES

[1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, Princeton, N.J., 1965.

[2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics, vol. II, Partial Differential Equations*, Interscience, New York, 1962.

[3] G. F. D. DUFF, *Mixed problems for hyperbolic equations of general order*, Canad. J. Math., 11 (1959), pp. 195–221.

[4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part II: Spectral Theory*, Interscience, New York, 1963.

[5] J. J. GRAINGER, *Boundary-value control of distributed systems characterized by hyperbolic differential equations*, Doctoral thesis, Dept. of Electrical Engineering, University of Wisconsin, Madison, 1967.

[6] G. HELLWIG, *Differential Operators of Mathematical Physics*, Addison-Wesley, Reading, Mass., 1964.

[7] FRITZ JOHN, *On linear partial differential equations with analytic coefficients—Unique continuation of data*, Comm. Pure Appl. Math., 2 (1949), pp. 209–253.

[8] P. D. LAX, *On Cauchy's problem for hyperbolic equations and the differentiability of solutions of elliptic equations*, Ibid., 8 (1955), pp. 615–633.

[9] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, vol. 1, Dunod, Paris, 1968.

[10] ———, *Problèmes aux limites non homogènes et applications*, vol. 2, Dunod, Paris, 1968.

[11] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

[12] P. H. RABINOWITZ, *Periodic solutions of nonlinear hyperbolic partial differential equations*, Comm. Pure Appl. Math., 20 (1967), pp. 145–204.

[13] D. L. RUSSELL, *Boundary value control of the higher dimensional wave equation*, this Journal, 9 (1971), pp. 29–42.

[14] ———, *On boundary- value controllability of linear symmetric hyperbolic systems*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 312–321.

[15] ———, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–560.

[16] L. SCHWARTZ, *Théorie des distributions*, Hermann, Paris, 1950.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

RECURSIVE IDENTIFICATION OF LINEAR SYSTEMS*

J. RISSANEN†

Abstract. Let the three matrices $\Sigma(N) = (G(N), F(N), H(N))$ define a linear constant system of least degree which realizes the set of numbers $f_1, \cdots, f_N$ regarded as a partial impulse response of a system. An algorithm has been developed for recursively calculating the minimal partial realizations for each $N = 1, 2, \cdots$ such that

$$\cdots \sum(N-k) \subseteq \sum(N) \subseteq \cdots .$$

This algorithm differs from the previous ones, such as that of B. L. Ho's, in that there is a recursion on N as well. Because of this, no a priori guess of the order of the system is required. Moreover, an addition of terms to the initial sequence causes the computation of only a few new elements. When combined with another algorithm for factoring covariance matrices the given algorithm permits a recursive identification of linear random systems. No earlier recursive identification methods seem to appear in the literature. Finally, a categorical description of the abstract realizations is given.

1. Introduction. Let $A_1, A_2, \cdots$ be the impulse response of a linear constant system of the type

$$\begin{aligned} x(k+1) &= Fx(k) + Gu(k), \\ y(k) &= Hx(k), \end{aligned} \tag{1.1}$$

where $x(k) \in X = R^n$, $u(k) \in U = R^p$ and $y(k) \in Y = R^m$ for some integers n, p, and m. In other words,

$$A_i = HF^{i-1}G, \qquad i = 1, 2, \cdots . \tag{1.2}$$

The so-called realization or identification problem is one of recovering the maps or matrices (G, F, H) from the sequence $A_1, A_2, \cdots$. The problem is classical, but the first good solution was not found until 1965 when B. L. Ho, in his doctoral dissertation, presented a neat algorithmic or nearly algorithmic solution to the problem (see [1]). At about the same time related algorithms were described also by L. Silverman and D. Youla [2], [10].

Subsequently, the problem was extended to the interesting case where only a finite sequence $A_1, \cdots, A_N$ was given. The same algorithm was shown to give a solution [3], [4]. In that form the problem was a generalization of the Padé approximation problem, and the algorithm provided a solution which in many respects is superior to the classical Cauchy–Jacoby formulas.

In this paper we consider the following more general variant of the same problem: Given the partial sequence $A_1, \cdots, A_N$, for each $N = 1, 2, 3, \cdots$ find a sequence of minimal partial realizations

$$\Sigma(N) = (G(N), F(N), H(N)) \tag{1.3}$$

such that

$$\cdots \Sigma(N') \subseteq \Sigma(N) \subseteq \cdots \quad \text{if } N' < N.$$

* Received by the editors November 3, 1970.
† IBM, Inc., San Jose, California 95114.

Here the inclusion symbol means that the matrices in $\Sigma(N')$ appear as submatrices of the corresponding ones in $\Sigma(N)$. Thus, not only do we look for an algorithmic solution to the partial realization problem for some N, but we want a recursion on N as well. Moreover, we look for a solution where each extension of the sequence of the A_i's can be met by calculating just a few new elements in the corresponding realization.

We point out that if Ho's algorithm is used in this case, then each partial realization must be calculated anew. Even when applied to a single partial sequence our algorithm turns out to be simpler than that of Ho's.[1]

In §§ 2 and 3 we give a self-contained abstract discussion of the realization problem. We give a categorical characterization of the fundamental notions of the minimal state space and the associated canonical factorization and realization.

In the final section we present an application where the problem naturally leads to finding the partial realizations for an indefinite number of values N.

The discussion throughout is restricted to the case where the input and the output spaces are one-dimensional. This is done to avoid introducing messy irrelevant notations and indices. The reader should have little trouble extending the results to the general case.

2. Linear input-output systems. We begin by giving a condensed but self-contained exposition of how linear systems are characterized by their input-output properties. Our approach differs from that of Kalman's [3] above all in that the important notions of canonical factorization and the associated minimal state space are defined in categorical terms. This better emphasizes the universality of these constructs; see also [6].

Let U and Y be one-dimensional vector spaces over the field R of the real numbers; i.e., both of them may be identified with R. Let T_- denote the set of nonpositive integers, and let $\Omega = \hom_c(T_-, U)$ denote the linear space of all sequences $\omega: T_- \to U$ of finite support; i.e., ω has only finitely many nonzero components. Finally, let $\Gamma = \hom(T_+, Y)$ denote the linear space of all functions or sequences $\gamma: T_+ \to Y$, where T_+ is the set of positive integers.

The linear space $L = \Omega \oplus \Gamma$ admits a ring structure with convolution as the product:

$$(a * b)_n = \sum_{i \in T} a_{n-i} b_i, \qquad n \in T = T_- \cup T_+, \quad a, b \in L. \tag{2.1}$$

This product is well-defined since the sum contains only finitely many nonzero terms for each n. Moreover, the product is commutative, and the ring has the element $(\cdots, 0, 1, 0, \cdots) = e$, 1 in the zeroth position, as the identity.

Any element f of Γ defines a linear mapping

$$f: \Omega \to \Gamma \tag{2.2}$$

[1] When this paper was written Professor Kalman called our attention to a recent paper by Zeiger, *Some computational aspects of Ho's algorithm* (we have not found it published), in which Ho's algorithm was improved by a special factorization of the so-called Hankel matrix. That factorization is related to the one discussed in § 4. However, we exploit the factorization in a different way and the result is an altogether new algorithm with features not obtained by that of Zeiger's which still basically remains similar to Ho's algorithm.

by the ring product

$$(f(\omega))_n = (f * \omega)_n, \qquad n \in T_+. \tag{2.3}$$

We take any such mapping f as an input-output description of a linear constant system. This is justified, for if $f_i = HF^{i-1}G$, then any system (1.1) defines an element of Γ, and as will be shown shortly, any f in Γ admits a representation (1.1).

The element $e_1 = (\cdots, 0, 1, 0)$ of Ω defines a left shift, say σ_Ω:

$$\sigma_\Omega : \Omega \to \Omega, \qquad \sigma_\Omega(\omega) = e_1 * \omega. \tag{2.4}$$

Similarly, it defines a left shift σ_Γ on Γ:

$$\sigma_\Gamma : \Gamma \to \Gamma, \quad (\sigma_\Gamma(\gamma))_n = (e_1 * \gamma)_n, \quad n \in T_+. \tag{2.5}$$

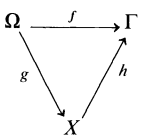
The commutativity of the ring implies that a linear mapping f of Γ satisfies

$$f\sigma_\Omega = \sigma_\Gamma f, \tag{2.6}$$

which in [3] was taken as the definition of linear constant systems.

We now turn to the question of how to associate a state space, above all a minimal one, to the mapping f. This question is central to the whole theory of systems described by their input-output properties, and we shall give two equivalent characterizations of the minimal state space and the associated canonical factorization.

Consider any factorization of f:

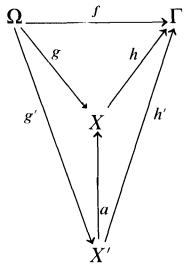


DIAG. 2.7

If g is surjective and h injective, the factorization is said to be *canonical* with a *minimal* state space X. Such factorizations obviously exist; $X = \Omega/\ker f$, with g the natural projection, gives one. Moreover, all canonical factorizations are equivalent in the sense that $hg = h'g'$ implies the existence of a unique isomorphism $q : X \to X'$ such that $g' = qg$ and $h = h'q$ (see [3], [5]).

These facts suggest an alternative definition of the canonical factorization and the associated state space which incorporates the universal properties of these constructs virtually without any further proving. We form the category $\mathscr{L}_s$ of linear systems as follows. The objects are Diag. 2.7 with g surjective; Ω, f, and Γ are the same in all the objects. An object may then be denoted by (g, X, h). As the set of morphisms, $\hom[(g', X', h'), (g, X, h)]$, take all linear maps such that Diag. 2.8 commutes. Compositions and the identities are the obvious ones, and the category axioms [9] for $\mathscr{L}_s$ are satisfied. Define a *terminal object* in this category to be a canonical factorization through the associated minimal state space.

We could give a standard construct as a limit for the terminal objects, which would prove their existence. But since we must anyway show that this definition



DIAG. 2.8

gives the canonical factorization in the same sense as the previous one, we avoid such constructs by proving the following theorem.

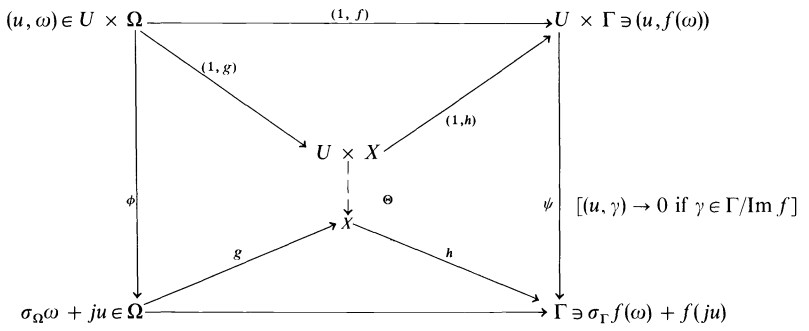
THEOREM 2.9. *The factorization* $f = hg$ *through* $X = \Omega/\ker f$ *gives a terminal object in* $\mathscr{L}_s$.

Proof. We must show that there exists a unique a in Diag. 2.8, where X now is taken as $\Omega/\ker f$ and g the associated natural projection. Let z belong to X'. Define $a: z \to g(\omega)$, where ω is any point in Ω such that $g'(\omega) = z$; one such exists since g' is surjective. This definition gives a function, since if also $g'(\omega') = z$, then $f(\omega) = h'g'(\omega) = h'g'(\omega') = f(\omega')$, and $\omega - \omega' \in \ker f$. Hence, $g(\omega) = g(\omega')$. That a is linear and satisfies $ag' = g$ is clear. a is the only such mapping since g' is surjective, which means that $g = ag' = a'g'$ implies $a = a'$.

The case where the dimension of the minimal state space is finite is of particular interest. This is the case which we shall be primarily concerned with.

3. Minimal realizations. By a minimal or canonical realization of the system described by an input-output mapping (2.2) is meant the intrinsic state description (1.1) of the same system characterized by the three mappings $\Sigma = (G, F, H)$ and the associated spaces, U, Y, X with X minimal. Before describing how such a realization is constructed recursively, we quickly define them in abstract terms.

The problem is to indicate in Diag. 2.7 how input sequences are successively built into longer ones and to show how this induces the state transitions. Let $j: U \to \Omega$ by $u \to (\cdots, 0, u)$ denote the natural injection. Consider Diag. 3.1, where X is minimal.



DIAG. 3.1

The mapping ϕ induces the mappings Θ and ψ, Θ unique, such that the diagram commutes. The latter is defined in the diagram, and the former results from the fact that h is injective:

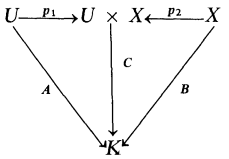
$$\Theta = k\psi(1, h), \tag{3.2}$$

where $k:\Gamma \to X$ satisfies $kh = 1_X$.

Θ can be written as

$$\Theta(u, x) = Fx + Gu. \tag{3.3}$$

To see this, recall first that by $U \times X$ is meant the direct sum $U \xrightarrow{p_1} U \times X \xleftarrow{p_2} X$, where p_1 and p_2 are the canonical injections. Hence, there is a bijection: $\hom(U \times X, K) \leftrightarrow \hom(U, K) \times \hom(X, K)$ in Diag. 3.4



DIAG. 3.4

for any linear space K. Taking $K = X$ and $C = \Theta$ gives (3.3).

More explicitly, the mapping G is given by Diag. 3.1 as

$$G(u) = \Theta(u, 0) = gj(u) \tag{3.5}$$

and F as

$$F(x) = \Theta(0, x) = k\sigma_\Gamma h(x). \tag{3.6}$$

If $P_1:\Gamma \to Y$, $P_1(y_1, y_2, \cdots) = y_1$, then the last mapping in (1.1) is given by

$$H = P_1 h:X \to Y. \tag{3.7}$$

By Theorem 2.9 any other canonical realization is isomorphic to the one given above. The preceding formulas form the basis for the algorithm to be described in the next section.

4. Realization algorithm. We need a few preliminary results. Let $b_1 = f(e) = (f_1, f_2, \cdots) \in \Gamma$ and $b_i = \sigma_\Gamma^{i-1} b_1$. We arrange the entries in the so-called Hankel matrix:

$$A = \begin{bmatrix} f_1 & f_2 & \cdots & f_n & \cdots \\ f_2 & f_3 & \cdots & f_{n+1} & \cdots \\ \text{—} & \text{—} & \text{—} & \text{—} & \\ f_m & f_{m+1} & \cdots & f_{m+n-1} & \cdots \\ \hline \end{bmatrix}. \tag{4.1}$$

Let $A(m, n)$ denote the dashed submatrix of A.

Observe first from Diag. 2.7 that since h is injective in a canonical factorization, the dimension of X, or the order of the system, is the same as the dimensionality of the image of f. Since the set $\{\sigma_\Omega^i e\}$ spans Ω, the image of f is spanned by the set $\{b_i\}$. The order of the system is then the cardinality of the linearly independent vectors in the set $\{b_i\}$. We need the following lemma.

LEMMA 4.2. *If* $\dim\{b_1, b_2, \cdots\} = n$, *then* $b_1, b_2, \cdots, b_n$ *are the linearly independent vectors in* $\{b_i\}$. *In addition,* $A(n, n)$ *is nonsingular and* $A(n, m)$, $m \geqq n$, *has rank* n.

Proof. Suppose that b_{k+1} is linearly dependent on $b_1, \cdots, b_k$; i.e., $(\sigma_\Gamma^k + \sum_1^k c_i \sigma_\Gamma^{i-1})b_1 = 0$. Then, σ_Γ of that same expression also vanishes, or $b_{k+2} + c_k b_{k+1} + \cdots + c_1 b_2 = 0$, and also b_{k+2} is linearly dependent on $b_1, \cdots, b_k$. Hence, the first n must be linearly independent.

Consider the last row in $A(n+1, n)$. Since b_{n+1} is linearly dependent on $b_1, \cdots, b_n$, the rank of $A(n+1, n)$ is still n. By symmetry the same is true about $A(n, n+1)$. Hence the ranks of $A(n, m)$ for $m \geqq n$ are no greater than n. But since they clearly can neither be less than n they all must be equal to n.

The algorithm is based on a factorization of $A(n, m)$ of the following type:

$$A(n, m) = P(n, n)Q(n, m), \qquad m \geqq n, \quad \text{rank } A(n, m) \geqq n - 1, \tag{4.3}$$

where $P(n, n)$ is lower triangular with 1's on the diagonal; i.e.,

$$\begin{bmatrix} f_1 \cdots f_i & f_{i+1} & \cdots f_m \\ f_2 \cdots f_{i+1} & f_{i+2} & \cdots \\ & & \\ & f_{i+n-1} & \cdots \\ f_n \cdots f_{i+n-1} & f_{i+n} & \cdots f_{n+m-1} \end{bmatrix} \tag{4.4}$$

$$= \begin{bmatrix} 1 & & & \\ p_{21} & 1 & & \\ & & & \\ & & & \\ p_{n1} \cdots p_{n,n-1} & & 1 & \end{bmatrix} \begin{bmatrix} q_{11} \cdots q_{1i} & q_{1,i+1} & \cdots q_{1m} \\ q_{21} \cdots q_{2i} & & \cdots \\ & & \\ & & \\ q_{n1} \cdots q_{ni} & q_{n,i+1} & \cdots q_{n,m} \end{bmatrix}.$$

The factors are not unique, a fact which we shall take advantage of to obtain factors with further desired features. By setting certain elements $q_{ij} = 0$ we shall be able to calculate the p_{ij}'s recursively one by one. Moreover, an addition of rows and columns to $A(n, m)$ will not change the numbers already calculated. A further property of $Q(n, m)$ is that the first $n-1$ rows are linearly independent and the last row is zero if the rank of $A(n, m) = n - 1$. With a minor modification the algorithm would work regardless of what the rank of $A(n, m)$ is. Since we took U and Y to be one-dimensional we shall have no need for that case, however.

The factorization algorithm runs as follows.

Step 1. Set $q_{1i} = f_i$ for all i. If $n = 1$ we are done: $P(1, 1) = (1)$.

Step 2. In the other event, proceeding recursively, we have at the ith step or row all the p_{jk}'s and q_{jk}'s, $j = 0, 1, \cdots, i-1$, determined. Let $s(j)$ be the least

integer such that $q_{j,s(j)} \neq 0$, $j < n$. $s(j)$ exists because of the rank condition on $A(n, m)$. Set $q_{k,s(j)} = 0$ for $k > j$. Equation (4.4), then, leads to a set of $i - 1$ equations, one for each column $s(j)$, $j = 1, \cdots, i - 1$. Because of the previous conditions, the unknowns $p_{i1}, \cdots, p_{i,i-1}$ can be solved recursively one by one from these equations. The submatrix $P(i, i)$ with (4.4) determine the remaining elements of the ith row of $Q(n, m)$, which completes the cycle.

As an example, consider

$$A(4,5) = \begin{bmatrix} 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 2 & 1 & 3 \\ 1 & 2 & 1 & 3 & 2 \\ 2 & 1 & 3 & 2 & 3 \end{bmatrix}. \tag{4.5}$$

We have $s(1) = 1$. Set $q_{i1} = 0, i > 1$. Then

$$p_{21} \cdot 1 + 1 \cdot q_{21} = 1, \qquad p_{21} = 1.$$

Further, $q_{22} = 0$, $q_{23} = 1$, $q_{24} = -1$, $q_{25} = 2$. Then, since $s(2) = 3$, $q_{i3} = 0$, $i > 2$, and the first and the third column give

$$p_{31}q_{11} = 1 \quad \text{or} \quad p_{31} = 1,$$
$$1 \cdot q_{13} + p_{32}q_{23} = f_5 = 1 \quad \text{or} \quad p_{32} = 0.$$

Continuing we get the result

$$A(4,5) = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ 1 & 0 & 1 & \\ 2 & 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 2 & 1 \\ 0 & 0 & 1 & -1 & 2 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{4.6}$$

Define

$$P_*(n-1, n-1) = \begin{bmatrix} p_{21} & 1 & & \\ p_{31} & p_{32} & 1 & \\ \vdots & & & \ddots \\ p_{n1} & \cdots & & p_{n,n-1} \end{bmatrix}, \qquad G(n-1) = \begin{bmatrix} q_{11} \\ q_{21} \\ \vdots \\ q_{n-1,1} \end{bmatrix}, \tag{4.7}$$

$H(n - 1) = (1, 0, \cdots, 0)$, $n - 1$ elements.

THEOREM 4.8. *Given $f_1, f_2, \cdots$, let $A(n - 1, n - 1)$ be nonsingular and m any integer such that $A(n, m)$ has rank $n - 1$. Then $\Sigma(n - 1) = (G(n - 1), F(n - 1), H(n - 1))$, $X(n - 1) = R^{n-1}$, where*

$$F(n-1) = P^{-1}(n-1, n-1)P_*(n-1, n-1) \tag{4.9}$$

is a minimal realization of the sequence $f_1, \cdots, f_{n+m-1}$.

Proof. By the factoring algorithm the last row of $Q(n, m)$ is zero. Hence by writing the equality between the dashed columns of $A(n, m)$ in (4.4) we get with (4.9)

$$F(n-1) \begin{bmatrix} q_{1i} \\ \vdots \\ q_{n-1,i} \end{bmatrix} = \begin{bmatrix} q_{1,i+1} \\ \vdots \\ q_{n-1,i+1} \end{bmatrix} \quad \text{for all } i. \tag{4.10}$$

Denote the elements of R^{n-1} by $x = \operatorname{col}(x_1, \cdots, x_{n-1})$ and consider the equations

$$\begin{aligned} x(i+1) &= F(n-1)x(i) + G(n-1)u(i), \\ y(i) &= H(n-1)x(i). \end{aligned} \tag{4.11}$$

With $x(0) = 0$ and $u(0) = 1, u(i) = 0, i > 0$, the first equation in (4.11) describes the consecutive states which are just the columns of $Q(n-1, m)$ in (4.4), or (4.10), and the second gives $y(i) = q_{1i} = f_i$ for $i = 1, \cdots, m$. We still have to show that (4.11) realizes the rest of the f_i's. This will follow from the special form of $F(n-1)$ which it inherits from $P_*(n-1)$, namely,

$$F(n-1) = \begin{bmatrix} f_{11} & 1 & 0 & \cdots & 0 \\ f_{21} & f_{22} & 1 & 0 \cdots & 0 \\ \vdots & & & & \vdots \\ & & & & 1 \\ f_{n-1,1} & \cdots & & & f_{n-1,n-1} \end{bmatrix}. \tag{4.12}$$

Indeed, applying formula (4.11) for $i = m+1, \cdots, m+n-1$ extends $Q(n, m)$ to $\bar{Q}(n, m+n-1)$ (the last row is extended as a zero row). Multiplying the result by $P(n, n)$ we get the corresponding extension of $A(n, m)$ to $\bar{A}(n, m+n-1)$, say. But due to the special shape of $F(n-1)$, $\bar{A}(n, m+n-1)$ will have the following elements:

$$\begin{bmatrix} f_1 & \cdots & f_m & \cdots & f_{m+n-1} \\ \vdots & & & \iddots & \\ f_n & \cdots & f_{m+n-1} & & \end{bmatrix}.$$

This means that $y(i) = q_{1i} = f_i$ for all $i \leqq m+n-1$.

Any realization of the sequence $f_1, \cdots, f_{m+n-1}$ extends it indefinitely. If such a realization has order $k < n-1$, then $A(n-1, n-1)$ has rank k by Lemma 4.2 which contradicts the assumptions. Hence, (4.11) is minimal.

We are now ready to describe the realization algorithm. It works under the hypothesis that from some source we can pick the numbers $f_1, \cdots, f_N$ for any $N = 1, 2, \cdots$.

Step 1. Let k be the least integer for which $f_k \neq 0$. Take $N = 2k+1$ and form $A(k+1, k+1)$. It has rank $\geqq k$.

Step 2. Apply the factoring algorithm and find $P(k+1, k+1)$ and $Q(k+1, k+1)$ (equation (4.4)). If the last row of Q is nonzero, the rank of $A(k+1, k+1)$ is $k+1$. Increase N by 2, form $A(k+2, k+2)$ and continue the factorization. If the last row of $Q(k+2, k+2)$ is still nonzero, increase N by 2 and repeat until, say, for $N = 2n-1$, the last row of $Q(n, n)$ is zero. Such an n exists by Lemma 4.2 if the numbers f_i admit a finite order realization.

Step 3. From formulas (4.7) and (4.9) calculate the partial realization $\Sigma(n-1)$. Observe that the inverse of $P(n-1, n-1)$ can also be calculated recursively since this matrix together with its inverse is lower triangular.

Step 4. Increase N by 1. Continue the factorization for $A(n, n+1)$. If the last row of $Q(n, n+1)$ remains zero, increase N by 1, and repeat. If the last row remains zero for all m, we have found the realization. (This cannot, of course, be decided, so that the algorithm would never stop. A stopping rule is introduced by setting an upper limit for m.)

Step 5. If for some $m(>n)$ the last element in the last row of $Q(n, m)$ becomes nonzero, the last picked element f_{m+n-1} is not realized by the partial realization $\Sigma(n-1)$. In this case, pick one new point, f_{m+n}, and form $A(n+1, m)$. Continue the factorization, pick one new point and repeat until either $Q(n', m)$ for some least $n' \leqq m$ has last row zero or $n' = m$ and the last row is nonzero. In the previous case go to Step 3. In the latter case go to Step 2.

Comments. Due to the special form of the matrices (4.7) and (4.12) this algorithm has the crucial "nesting" property (for k, see (4.13)):

$$\cdots \Sigma(n-k) \subseteq \Sigma(n) \subseteq \cdots, \quad \text{if } k < 1;$$

where the inclusion sign means that the matrices in $\Sigma(n-k)$ are submatrices of the corresponding ones in $\Sigma(n)$. Hence, at each step only a few new elements as a function of the old ones and the new picked elements f_i need be calculated. In fact, we have established a recursion of the type,

$$\begin{aligned} \Sigma(n+k) &= \Pi_1(n, \Sigma(n), f_{2n+1}, \cdots, f_m), \\ m &= \Pi_2(n, \Sigma(n), f_{2n+1}, \cdots, f_{m-1}), \\ k &= \Pi_3(n, \Sigma(n), f_{2n+1}, \cdots, f_m). \end{aligned} \tag{4.13}$$

Another advantage of this algorithm is that the formulas (4.7) and (4.9) are actually simpler than those in the Ho algorithm [3].

To illustrate the algorithm we give the sequence of partial realizations corresponding to the example in (4.5) and (4.6). In Step 2, $n = 2$, and in Step 3 we get $\Sigma(1) = ((1), (1), (1))$. In Step 4 an addition of $f_4 = 2$ makes $q_{23} = 1 \neq 0$. In Step 5 pick $f_5 = 1$, and form $A(3, 3)$. Returning to Step 2, we pick two new points, $f_6 = 3$ and $f_7 = 2$. This time the last row of $Q(4, 4)$ is zero, and we calculate $\Sigma(3)$ following Step 3:

$$G(3) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad F(3) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{bmatrix}, \qquad H(3) = (1, 0, 0),$$

which realizes all the numbers given in $A(4, 5)$.

5. Identification of random systems. In this section we briefly describe an application of the realization algorithm where its recursive nature proves to be particularly useful.

Consider a system whose output forms a stationary random process $\{y_t\}$, $t = 1, 2, \cdots$. Take $E(y_t) = 0$ for all t, and denote the covariances by $r_i = E(y_t y_{t-i})$. Form the infinite symmetric covariance matrix, $R = (r_{ij})$, $r_{ij} = r_{i-j}$.

Assuming the process to be of full rank, i.e., $R > 0$, we can factor R recursively [7], [8] into the product of a lower triangular matrix B and its transpose:

$$R = BB', \tag{5.1}$$

where

$$B = \begin{bmatrix} b_{00} & & & \\ b_{10} & b_{11} & & \\ & & & \\ b_{n0} & b_{n1} & \cdots & b_{nn} \\ & \cdots & \cdots & \end{bmatrix}. \tag{5.2}$$

The recursive equations for the b_{ij}'s are obtained by comparing elements in (5.1) starting on the top row and progressing down row by row. Observe that the result is just a special case of the factoring algorithm in § 4.

THEOREM 5.3. *If R satisfies Szego's criterion,*

$$\int_{|z|=1} \log\left(\sum_{-\infty}^{\infty} r_i z^i\right) dz > -\infty,$$

then $b_{ij} \to f_{i-j}$ as $i \to \infty$.

We omit the proof, which is a simple modification of the proof of the related theorem, Theorem 6 in [8].

Applying the realization algorithm to the sets of numbers $f_1, \cdots, f_N$. $N = 1, 2, \cdots$, we get a series of realizations. If the process is generated by a finite order system, this series leads to the minimal realization.

Since we actually must use the approximations b_{ij} for f_{i-j}, the realizations will be approximations, too. Because of this, the precise rank condition in the algorithm should be replaced by some rule such as: the last row of $Q(n, m)$ is considered as zero if $|q_{ni}|/\|q_i\| \leqq \varepsilon$, where q_i is the ith column of $Q(n-1, m)$. Since the last row represents the nth components of the state vectors, the error so introduced admits a clear interpretation.

The above described scheme seems to compare favorably with the commonly used procedures in that all of the computations progress in an orderly, recursive fashion without any trial and error involved.

The outlined procedure has the defect that the computation of (5.1) involves a growing amount of memory even when $R(z) = \sum_{-\infty}^{\infty} r_i z^i$ is a rational function. We sketch an approach which in some respect is an improved version of a related one by P. Faurre (see, e.g., [11]); see also [13].

One first applies the realization algorithm to the sequences of numbers $r_0, r_1, \cdots, r_N$. This gives the realization (K, F, H) of order n, say, and the rational function

$$R(z) = H(zI - F)^{-1}FK + K'F'(z^{-1}I - F')^{-1}H' + HK = \frac{P(z)}{Q(z)Q(z^{-1})}.$$

The polynomial $P(z) = \sum_{-m}^{m} p_i z^i$, $m \leqq n$, satisfies (5.3). Hence, the factorization obtained by replacing R in (5.1) by the Toeplitz matrix P associated with $P(z)$ which has only $2m + 1$ nonzero diagonals converges and gives in the limit

$$P(z) = P_1(z)P_1(z^{-1}).$$

The transfer function of the system is $P_1(z)/Q(z)$. We leave the details of this

procedure, above all the proof that the result is a "minimum phase" system, to another context. Such a result, in effect, was announced in [12], which reference was communicated to us by H. Aasnaes.

Acknowledgment. We are indebted to Professor R. Kalman, Dr. F. Palermo and Professor T. Kailath for stimulating discussions during the preparation of this paper.

REFERENCES

[1] B. L. Ho AND R. E. KALMAN, *Effective construction of linear state-variable models from input/output functions*, Proc. Third Allerton Conference, Urbana, Illinois, 1966, pp. 449–459.

[2] D. C. YOULA, *The synthesis of linear dynamic systems from prescribed weighting patterns*, SIAM J. Appl. Math., 14 (1966), pp. 527–549.

[3] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1968.

[4] ———, *Lectures on Controllability and Observability*, Centro Internazionale Matematico Estivo, Bologna, Italy, 1968.

[5] H. P. ZEIGER, *Ho's algorithm, commutative diagrams, and the uniqueness of minimal linear systems*, Information and Control, 11 (1967), pp. 71–79.

[6] M. A. ARBIB AND H. P. ZEIGER, *On the relevance of abstract algebra to control theory*, Automatika, 5 (1969), pp. 589–606.

[7] J. RISSANEN, *An algebraic approach to the problems of linear prediction and identification*, Res. Rep. RJ468, IBM, Yorktown Heights, N.Y., 1967.

[8] J. RISSANEN AND L. BARBOSA, *Properties of infinite covariance matrices and stability of optimum predictors*, Information Sci., 1 (1969), pp. 221–236.

[9] S. MACLANE AND G. BIRKHOFF, *Algebra*, Macmillan, New York, 1967.

[10] L. SILVERMAN, *Representation and realization of time-variable linear systems*, Tech. Rep. 94, Dept. of Electrical Engineering, Columbia Univ., New York, 1966.

[11] P. FAURRE AND J. P. MARMORAT, *Un algorithme de realization stochastique*, C. R. Acad. Sci. Paris, Sér. A, 268 (1969), pp. 978–981.

[12] F. BAUER, *Ein directes Iterationsverfahren zur Hurwitz-Zerlegung eines Polynoms*, Mitt. Math. Inst. Tech. Hochschule München, A.E.Ü., 9 (1955), pp. 285–290.

[13] T. KAILATH AND R. GEESEY, *An innovations approach to least squares estimation, Part IV: Recursive estimation given the covariance functions*, IEEE Trans. Automatic Control, to appear.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

GLOBAL VARIATION CRITERIA FOR STABILITY OF LINEAR TIME-VARYING SYSTEMS*

Y. V. VENKATESH†

Abstract. The system under study is a linear feedback system consisting of a time-invariant block G in the forward path and a time-varying gain $k(t)$ in the feedback path. Improved sufficient conditions for its asymptotic stability are derived by a combination of Brockett's factorization technique of generating Lyapunov functions and the Krasovskii–Corduneanu theorem. The resulting bound on $((dk/dt)/k)$ resembles but is distinct from the average criterion of Freedman and Zames [1].

1. Introduction. Consider the feedback system of Fig. 1 which is governed by the linear differential equation

$$p(D)y + k(t)q(D)y = 0 \quad \text{on the interval} \quad [t_0, \infty), \tag{1}$$

where

$$p(D) = D^n + p_{n-1}D^{n-1} + \cdots + p_0,$$
$$q(D) = q_mD^m + q_{m-1}D^{m-1} + \cdots + q_0$$

are constant coefficient differential operators with the order n of $p(D)$ at least one higher than the order m of $q(D)$.

Let $y = x_1$, $x_2 = dx_1/dt, \cdots, x_n = dx_{n-1}/dt$; and $\mathbf{x} = \text{col}\,[x_1, x_2, \cdots, x_n]$. Then (1) can be written as the vector differential equation

$$\frac{d\mathbf{x}}{dt} = A_0\mathbf{x} - k(t)\mathbf{bc}'\mathbf{x} \triangleq A(t)\mathbf{x}, \tag{2}$$

where A_0 is a stable matrix having the form:

$$A_0 = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -p_0 & -p_1 & -p_2 & -p_3 & \cdots & -p_{n-1} \end{bmatrix},$$

and $\mathbf{b}$, $\mathbf{c}$ are n-vectors given by

$$\mathbf{b} = \text{col}\,[0, 0, \cdots, 1],$$
$$\mathbf{c} = \text{col}\,[-q_0, -q_1, \cdots, -q_m, \cdots, 0].$$

The gain $k(t)$ is assumed to be absolutely continuous on the interval $[t_0, \infty)$. Let $G(s)$ be the transfer function of the forward block, i.e., $G(s) = q(s)/p(s)$.

* Received by the editors January 2, 1969, and in final revised form December 7, 1970.
† Department of Electrical Engineering, Indian Institute of Science, Bangalore 12, India.

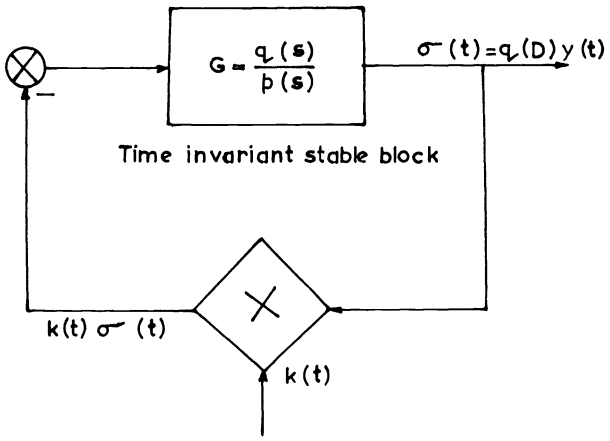


FIG. 1. *A linear time-varying feedback system*

Remark. The stability properties of the null solution (n.s.) of (1) are identical with the stability properties of the n.s. of (2) and vice versa.

ASSUMPTION. The n.s. of (1) is asymptotically[1] stable (a.s.) for every constant function $k(t) = K$ in $[0, \infty)$.

PROBLEM. Find conditions for the n.s. of (1) to be a.s. for every (absolutely) continuous function $k(t)$ with values in $[0, \infty)$.

Literature on the preceding problem is too vast to receive due acknowledgment here. For all unbounded dk/dt, the circle criteria of Zames [2], Narendra and Goldwyn [3] and Sandberg [4] are well known. The circle criterion is applicable even when the feedback is nonlinear and time-varying. Using a positive multiplier approach similar to that of Zames [2], Brockett and Forys [5] and others introduced a bound on the rate of variation $((dk/dt)/k)$ permitting the hypothesis on $G(s)$ to be weakened. Freedman and Zames [1] obtained an explicit stability condition free of multipliers in terms of a bound on an average of $((dk/dt)/k)$, namely,

$$\sup_{t \geqq 0} \frac{1}{T} \int_t^{t+T} \left| \frac{dk(\tau)}{d\tau} \Big/ k(\tau) \right| d\tau < 4\beta_{sh} \tag{3}$$

for some $T > 0$, where β_{sh} is a constant determined from the Nyquist diagram of $G(s)$. What is noteworthy here is that the use of an implicit multiplier has been avoided. However, condition (3) is not necessary and weaker conditions may be possible.

2. Solution of the main problem. The present objective is to explore a route towards obtaining sufficient conditions for the a.s. of (1) based on Brockett's technique [9] of generating a Lyapunov function and the Krasovskii–Corduneanu theorem [6, pp. 56–57], [7]. In fact, a new condition on $((dk/dt)/k)$ is derived below which is distinct from the modulus condition (3) of Freedman and Zames [1]. In order to state the main results of the paper, the following definitions and lemmas will be needed.

[1] See Definitions 3 and 4 below.

DEFINITION 1. A complex-valued function $Z(s)$ of a complex variable s is called a *positive real* (*strictly positive real*) *function of the argument* s, if $Z(s)$ is real for real values of s, and for $\operatorname{Re} s > 0$ (where Re denotes the real part) is analytic and satisfies the inequality

$$\operatorname{Re} Z(s) \geqq 0 \quad (>0).$$

The following notation will be used: $\|\mathbf{x}\|$ denotes the norm of $\mathbf{x}$, where $\|\mathbf{x}\|^2 = \mathbf{x}'\mathbf{x}$; $\mathbf{x}_0$ denotes $\mathbf{x}(t_0)$; $\mathbf{x}(t; t_0, \mathbf{x}_0)$ denotes the solution of (2) which takes the value $\mathbf{x}_0$ for $t = t_0$.

DEFINITION 2. The null solution of (2) is said to be *stable* if for every $\varepsilon > 0$ there exists $\delta(\varepsilon; t_0)$ such that if $\|\mathbf{x}_0\| < \delta(\varepsilon; t_0)$, then $\|\mathbf{x}(t; t_0, \mathbf{x}_0)\| < \varepsilon$ for $t \geqq t_0$.

DEFINITION 3. The null solution of (2) is said to be *asymptotically stable* if it is stable and, in addition, there exists a $\delta_0(t_0) > 0$ with the property that if $\|\mathbf{x}_0\| < \delta_0$, then $\lim_{t\to\infty} \mathbf{x}(t; t_0, \mathbf{x}_0) = 0$.

DEFINITION 4. The null solution of (2) is said to be *exponentially stable* if there exist positive constants $\varepsilon_1, \varepsilon_2$, such that, for $t \geqq t_0$,

$$\|\mathbf{x}(t; t_0, \mathbf{x}_0)\| < \varepsilon_2 \|\mathbf{x}_0\| \exp[-\varepsilon_1(t - t_0)].$$

LEMMA 1. *The asymptotic stability of* (1) *is equivalent to the asymptotic stability of*

$$p(D)n(D)y + k(t)q(D)n(D)y = 0, \tag{4}$$

provided $n(D)z = 0$ *represents an asymptotically stable system.*

Proof. Since the system described by $n(D)z = 0$ is asymptotically stable, $n(s) = 0$ has all its zeros in the half-plane $\operatorname{Re} s < 0$.

If $\phi(t)$ is a solution of (4), then $n(D)\phi(t)$ is a solution of (1). Further, if $n(D)\phi(t)$ tends to zero as $t \to \infty$, so does $\phi(t)$ showing that the stability of (1) implies the stability of (4).

To prove the converse, note that if $\psi(t)$ is a solution of (1), then there is a function $y(t)$, satisfying $n(D)y = \psi(t)$ and such that $y(t)$ is a solution of (4). Now, because of the a.s. of the system described by (4), $\psi(t)$ tends to zero as $t \to \infty$. But $n(s) = 0$ has all its zeros in the half-plane $\operatorname{Re} s < 0$. Consequently, $y(t)$, treated as the output of a system with the transfer function $1/n(s)$, also tends to zero as $t \to \infty$.

LEMMA 2. *A real function of a complex variable* $Z(s) = m(s)/n(s)$, *where* $m(s)$ *and* $n(s)$ *are finite polynomials in* s, *is positive real if and only if*:

(i) $n(s) + m(s)$ *has no zeros in the closed right half-plane* ($\operatorname{Re} s \geqq 0$), *and*

(ii) $\operatorname{Re} Z(j\omega) \geqq 0$ *for all real* ω.

Proof. See Weinberg and Slepian [8].

A. If $u(s)$ is a polynomial with real coefficients, let $\operatorname{Ev} u(s) = \{u(s) + u(-s)\}/2$ denote its even part. Note that for real ω, $\operatorname{Ev} u(j\omega) = \operatorname{Re} u(j\omega)$. If $u(s)$ is even and $\operatorname{Re} u(j\omega) \geqq 0$ for all real ω, then according to a theorem of Wiener, there exists a unique polynomial $v(s)$ with real positive coefficients such that $v(s)\, v(-s) = u(s)$, and $v(s)$ has zeros only in the closed left half-plane $\operatorname{Re} s \leqq 0$. Let $[u(s)]^{(+)}$ stand for $v(s)$ and $[u(s)]^{(-)}$ for $v(-s)$. The latter is also known as the right half-plane spectral factor, or, at times, as the negative spectral factor of the even polynomial $u(s)$. Observe that $[u(s)]^{(-)}$ has zeros only in the closed right half-plane $\operatorname{Re} s \geqq 0$.

B. If $u_1(s)$ and $u_2(s)$ are two polynomials with real coefficients and $\operatorname{Re}[u_1(j\omega)/u_2(j\omega)] \geqq 0$ for all real ω, then clearly $\operatorname{Re}[u_1(j\omega)u_2(-j\omega)] \geqq 0$ for all real ω, and there is a unique factor $[\operatorname{Ev} u_1(s)u_2(-s)]^{(-)}$.

Similarly, if β is a real constant and $\operatorname{Re} Z(j\omega - \beta)G(j\omega - \beta) \geqq 0$ for all real ω, then

$$r_1(s) = \{\operatorname{Ev} m(s-\beta)q(s-\beta)n(-s-\beta)p(-s-\beta)\}^{(-)} \tag{5}$$

is uniquely defined. Moreover, if α is a real constant and $\operatorname{Re} Z(j\omega - \alpha) \geqq 0$ for all real ω, then

$$r_2(s) = \{\operatorname{Ev} m(s-\alpha)n(-s-\alpha)\}^{(-)} \tag{6}$$

is also uniquely defined. The polynomials $r_1(s)$ and $r_2(s)$ have their zeros in the closed right half-plane only.

C. Let $f(\mathbf{x}(t))$ be a given (scalar) function of trajectories $\mathbf{x}(t)$ on $[0, \infty)$, for example, of solutions of (2). Suppose the integral

$$I = \int_{t_1}^{t_2} f(\mathbf{x}(\tau))\, d\tau \tag{7}$$

exists and is identical for all trajectories $\mathbf{x}(t)$ satisfying $\mathbf{x}(t_1) = \mathbf{x}_1$ and $\mathbf{x}(t_2) = \mathbf{x}_2$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are constant vectors. Then the integral (7) will be represented by the notation

$$I = \int_{t(\mathbf{x}_1)}^{t(\mathbf{x}_2)} f(\mathbf{x}(\tau))\, d\tau.$$

Now suppose that at time t, the state is $\mathbf{x}$ and at $(t + \Delta t)$, the state is $\mathbf{x}_0$. Then

$$I(\mathbf{x}) - I(\mathbf{x}_0) = \int_{t}^{t+\Delta t} f(\mathbf{x}(\tau))\, d\tau.$$

The time derivative of I along the solutions of (2) is obtained by allowing $\Delta t \to 0$ in $\{I(\mathbf{x}) - I(\mathbf{x}_0)\}/\Delta t$ and substituting for $\mathbf{x}(t)$ the solution of (2):

$$\left.\frac{dV}{dt}\right|_{(2)} = f(\mathbf{x}(t)),$$

where now $\mathbf{x}(t)$ is the solution of (2).

D. Let $Z(s - \beta)G(s - \beta)$ be assumed to be positive real for some $\beta \geqq 0$. Then by Lemma 2, the solution $z'(t)$ of the equation

$$m(D-\beta)q(D-\beta)z' + n(D-\beta)p(D-\beta)z' = 0 \tag{8}$$

is asymptotically stable. The order η of this equation is greater than or equal to n. Let $z'(t) = z(t)\exp(\beta t)$. Then $z(t)$ satisfies the ηth order equation

$$m(D)q(D)z + n(D)p(D)z = 0, \tag{9}$$

and is asymptotically stable. Further, let

$$z_1(t) = z(t), \quad z_2(t) = \frac{dz_1}{dt}, \quad z_3(t) = \frac{dz_2}{dt}, \cdots, z_\eta(t) = \frac{dz_{\eta-1}}{dt},$$

and

$$\mathbf{z}(t) = \operatorname{col}[z_1, \cdots, z_\eta].$$

Note that the order of (4) is also η. The state vector $\mathbf{x}(t)$ of (1) is a subspace of $\mathbf{x}^*(t)$, the state vector of (4).

LEMMA 3. *If* $Z(s-\beta)G(s-\beta)$ *is positive real for some* $\beta \geqq 0$, *and* $m(s-\beta)q(s-\beta)$ *and* $n(s-\beta)p(s-\beta)$ *have no zeros on the imaginary axis, then there exists a positive definite function* $V_1(\mathbf{z}, t)$ *quadratic in* $\mathbf{z}$ *defined by*

$$V_1(\mathbf{z}, t) = e^{-2\beta t}\int_{t(\mathbf{0})}^{t(\mathbf{z})} \{[m(D-\beta)q(D-\beta)(ze^{\beta t})][n(D-\beta)p(D-\beta)(ze^{\beta t})] - (r_1(D)ze^{\beta t})^2\}\, d\tau, \tag{10}$$

where $z(t)$ *is any solution of* (9) *on* $[0, \infty)$.

Proof. As shown in [9], the integral in (10) is path independent. It remains to show that $V_1(\mathbf{z}, t)$ is positive definite. To this end, rewrite (10) letting $z' \triangleq ze^{\beta t}$ to obtain

$$V_1(\mathbf{z}, t) = e^{-2\beta t}\int_{t(\mathbf{0})}^{t(\mathbf{z})} \{[m(D-\beta)q(D-\beta)z'][n(D-\beta)p(D-\beta)z'] - [r_1(D)z']^2\}\, d\tau. \tag{11}$$

But z' satisfies (8), from which (11) takes the form

$$V_1(\mathbf{z}, t) = -e^{-2\beta t}\int_{t(\mathbf{0})}^{t(\mathbf{z})} \{[m(D-\beta)q(D-\beta)z']^2 + [r_1(D)z']^2\}\, d\tau \tag{12}$$

or

$$V_1(\mathbf{z}, t) = -e^{-2\beta t}\int_{t(\mathbf{0})}^{t(\mathbf{z})} \{[n(D-\beta)p(D-\beta)z']^2 + [r_1(D)z']^2\}\, d\tau. \tag{13}$$

Further, by the asymptotic stability of (8), (12) and (13) are valid for $t(\mathbf{0}) = \infty$, $t(\mathbf{z}) = 0$. Hence,

$$V_1(\mathbf{z}, t) = e^{-2\beta t}\int_0^{\infty} \{[m(D-\beta)q(D-\beta)z']^2 + [r_1(D)z']^2\}\, d\tau$$

or

$$V_1(\mathbf{z}, t) = e^{-2\beta t}\int_0^{\infty} \{[n(D-\beta)p(D-\beta)z']^2 + [r_1(D)z']^2\}\, d\tau,$$

from which it is evident that V_1 is positive semidefinite. Suppose V_1 is not positive definite. Then

$$m(D-\beta)q(D-\beta)z' = -n(D-\beta)p(D-\beta)z' = r_1(D)z' = 0$$

for some nonzero z'. That is, $m(s-\beta)q(s-\beta)$, $n(s-\beta)p(s-\beta)$ and $r_1(s)$ have common factors. This can only happen if $m(\cdot)\, q(\cdot)$ and $n(\cdot)\, p(\cdot)$ have imaginary zeros, a situation ruled out by hypothesis. This proves the positive definiteness of $V_1(\mathbf{z}, t)$.

E. Observing that $m(D-\beta)q(D-\beta)(ze^{\beta t}) = e^{\beta t}m(D)q(D)z(t)$, we can write (10) as

$$V_1(\mathbf{z},t) = e^{-2\beta t}\int_{t(\mathbf{0})}^{t(\mathbf{z})} e^{2\beta\tau}\{[m(D)q(D)z][n(D)p(D)z] - [r_1(D+\beta)z]^2\}\,d\tau. \tag{14}$$

The time derivative of $V_1(\mathbf{z},t)$ is then given by

$$\frac{dV_1(\mathbf{z},t)}{dt} = -2\beta V_1(\mathbf{z},t) + [m(D)q(D)z][n(D)p(D)z] - [r_1(D+\beta)z]^2. \tag{15}$$

In order to find the value of $dV_1(\mathbf{z},t)/dt$ when $z(t)$ is the solution of (4), replace $\mathbf{z}(t)$ by $\mathbf{x}^*(t)$ and $z(t)$ by $y(t)$ in (15) and use (4) (after multiplication by $m(D)q(D)y$) to obtain

$$\left.\frac{dV_1(\mathbf{x}^*,t)}{dt}\right|_{(4)} = -2\beta V_1(\mathbf{x}^*,t) - k(t)[m(D)q(D)y][n(D)q(D)y] - [r_1(D+\beta)y]^2. \tag{16}$$

F. Let it be assumed that $Z(s-\alpha)$ is positive real (p.r.) for some $\alpha \geqq 0$. Then by Lemma 2, the solution $W(t)$ of the equation

$$m(D-\alpha)(e^{\alpha t}q(D)W) + n(D-\alpha)(e^{\alpha t}q(D)W) = 0 \tag{17}$$

is asymptotically stable. Let η_1 be the order of (17). It is easy to verify that $\eta_1 \geqq n$ and the order of (8) is equal to η_1 or, at the most, greater than η_1 by 1. Further, let $\mathbf{W}(t)$ be the state vector for (17).

LEMMA 4. *If $Z(s-\alpha)$ is p.r. for some $\alpha \geqq 0$, then there exists a positive semi-definite quadratic form in $\mathbf{W}$ defined by*

$$V_2(\mathbf{W},t) = e^{-2\alpha t}\int_{t(\mathbf{0})}^{t(\mathbf{W})} \{[m(D-\alpha)(e^{\alpha\tau}q(D)W)][n(D-\alpha)(e^{\alpha\tau}q(D)W)] - [r_2(D)(e^{\alpha\tau}q(D)W)]^2\}\,d\tau. \tag{18}$$

Proof. The proof is similar to the proof of Lemma 3.

G. The time derivative of (18) is given by

$$\frac{dV_2(\mathbf{W},t)}{dt} = -2\alpha V_2(\mathbf{W},t) + [m(D)q(D)W][n(D)q(D)W] - [r_2(D+\alpha)q(D)W]^2. \tag{19}$$

Its value along the trajectories of (4) is obtained by substituting $y(t)$, the solution of (4), for $W(t)$ in (19):

$$\left.\frac{dV_2(\mathbf{W},t)}{dt}\right|_{(4)} = -2\alpha V_2(\mathbf{x}^*,t) + [m(D)q(D)y][n(D)q(D)y] - [r_2(D+\alpha)q(D)y]^2. \tag{20}$$

It can be verified that

$$\frac{d}{dt}[k(t)V_2(\mathbf{W},t)]\Big|_{(4)} = \left\{\left(\frac{dk}{dt}\Big/k\right) - 2\alpha\right\}k(t)V_2(\mathbf{x}^*,t)$$

$$+k(t)[m(D)q(D)y][n(D)q(D)y] - k(t)[r_2(D+\alpha)q(D)y]^2. \tag{21}$$

The proof of the following lemma is obvious.

LEMMA 5. *Let $\delta_1(t)$, $\delta_2(t)$ be bounded real functions on $[0,\infty)$. If $V_1(\mathbf{x}^*,t) \geqq 0$, $V_2(\mathbf{x}^*,t) \geqq 0$ and $V_{12}(\mathbf{x}^*,t) = \delta_1(t)V_1(\mathbf{x}^*,t) + \delta_2(t)V_2(\mathbf{x}^*,t)$, then*

$$V_{12}(\mathbf{x}^*,t) \leqq \sup_{t\geqq 0}[\delta_1(t),\delta_2(t)](V_1+V_2)(\mathbf{x}^*,t).$$

H. Consider

$$[r_1(D+\beta)y]^2 = \{[\mathrm{Ev}\, m(D)q(D)n(-D)p(-D)]^{(-)}y\}^2,$$

which is nonnegative and quadratic in $\mathbf{x}^*$. From Lemma 3, $V_1(\mathbf{x},t)$ is positive definite and quadratic in $\mathbf{x}^*$. By a well-known property of quadratic forms, there exists a nonnegative constant γ_1 such that

$$[r_1(D+\alpha)y]^2 \geqq \gamma_1 V_1(\mathbf{x}^*,t).$$

Similarly, there exists a nonnegative constant γ_2 such that

$$[r_2(D+\alpha)q(D)y]^2 \geqq \gamma_2 V_2(\mathbf{x}^*,t).$$

Let γ be equal to the minimum of the two numbers γ_1,γ_2; and let $V(\mathbf{x}^*,t) = V_1(\mathbf{x}^*,t) + k(t)V_2(\mathbf{x}^*,t)$. Therefore, in view of the fact that $k(t)$ is nonnegative,

$$[r_1(D+\beta)y]^2 + k(t)[r_2(D+\alpha)q(D)y]^2 \geqq \gamma V(\mathbf{x}^*,t). \tag{22}$$

I. The time derivative of $V(\mathbf{x}^*,t)$ along the solution of (4) is obtained by adding (16) and (21) to give

$$\frac{dV(\mathbf{x}^*,t)}{dt}\Big|_{(4)} = -2\beta V_1(\mathbf{x}^*,t) + \left[\left(\frac{dk}{dt}\Big/k\right) - 2\alpha\right]k(t)V_2(\mathbf{x}^*,t)$$

$$-[r_1(D+\beta)y]^2 - k(t)[r_2(D+\alpha)q(D)y]^2. \tag{23}$$

Let $\theta = ((dk/dt)/k)$ and

$$\sup_{t\geqq 0}(-2\beta-\gamma,\theta(t)-2\alpha-\gamma) = -\xi(t).$$

Then, using Lemma 5, from (23) one obtains

$$\frac{dV(\mathbf{x}^*,t)}{dt}\Big|_{(4)} \leqq -\xi(t)V(\mathbf{x}^*,t). \tag{24}$$

J. DEFINITION. The solutions of (2) are said *to have Property* K–C if there exist a positive constant η_0 and a real function $\mu(t)$ on $[t_0,\infty)$ such that

$$\|\mathbf{x}(t)\| \leqq \|\mathbf{x}(t_0)\|\eta_0 e^{-[\mu(t)-\mu(t_0)]/2}, \qquad t \geqq t_0. \tag{25}$$

LEMMA 6 (The Krasovskii–Corduneanu theorem). *The solutions of* (2) *have Property* K–C *if there exist a positive definite and decrescent quadratic form*

$v(\mathbf{x}, t) = \mathbf{x}'P(t)\mathbf{x}$, and a real-valued function $\lambda(t)$ on $[t_0, \infty)$ such that the derivative of $v(\mathbf{x}, t)$ along the solutions of (2) satisfies the inequality

$$\left.\frac{dv}{dt}\right|_{(2)} \leqq -\lambda(t)v. \tag{26}$$

Proof. Because $v(\mathbf{x}, t)$ is positive definite and decrescent, there exist positive constants α_1 and α_2 such that

$$\alpha_1\mathbf{x}'\mathbf{x} \leqq \mathbf{x}'P(t)\mathbf{x} \leqq \alpha_2\mathbf{x}'\mathbf{x}.$$

Integration of (26) gives

$$v(\mathbf{x}, t) \leqq v(\mathbf{x}_0, t) \exp\left(-\int_{t_0}^{t} \lambda(\tau)\, d\tau\right).$$

Consequently,

$$\alpha_1\mathbf{x}'\mathbf{x} \leqq v(\mathbf{x}, t) \leqq v(\mathbf{x}_0, t) \exp\left(-\int_{t_0}^{t} \lambda(\tau)\, d\tau\right)$$

$$\leqq \alpha_2\|\mathbf{x}_0\|^2 \exp\left(-\int_{t_0}^{t} \lambda(\tau)\, d\tau\right),$$

from which

$$\|\mathbf{x}\|^2 \leqq \frac{\alpha_2}{\alpha_1}\|\mathbf{x}_0\|^2 \exp\left(-\int_{t_0}^{t} \lambda(\tau)\, d\tau\right).$$

Therefore, the solutions of (2) have Property K–C with $\eta_0 = \sqrt{\alpha_2/\alpha_1}$ and $\lambda(t)$ equal to the derivative of $\mu(t)$.

COROLLARY 1. *If $\int_{t_0}^{t} \lambda(\tau)\, d\tau$ increases without bound as $t \to \infty$, then the system (2) is asymptotically stable.*

COROLLARY 2. *If $(1/T)\int_{t_0}^{t_0+T} -\lambda(\tau)\, d\tau \leqq -\gamma$ for some constants $\gamma > 0$ and $T > 0$, then*

$$\|\mathbf{x}\| \leqq \|\mathbf{x}_0\|\eta_0 \exp[-\gamma(t - t_0)/2], \qquad t \geqq t_0,$$

and the system is exponentially stable.

K. Let $\zeta(t)$ be a nonnegative (integrable and bounded) function on $[t_0, \infty)$, and $h(t) = \exp -\int_{t_0}^{t} \zeta(\tau)\, d\tau$. Assume that the integral $\int_{t_0}^{t} \zeta(\tau)\, d\tau \leqq M < \infty$ for all t in $[t_0, \infty)$, and

$$0 < \varepsilon \leqq \lim_{t\to\infty} \int_{t_0}^{t} \zeta(\tau)\, d\tau \leqq M < \infty.$$

Then $h(t)$ is a bounded positive function. Note that

$$\left(\frac{dh(t)}{dt} \Big/ h(t)\right) = -\zeta(t),$$

which is nonpositive.

Let

$$V_0(\mathbf{x}^*, t) = h(t)\{V_1(\mathbf{x}^*, t) + k(t)V_2(\mathbf{x}^*, t)\}, \tag{27}$$

where $V_1(\mathbf{x}^*, t)$, $V_2(\mathbf{x}^*, t)$ are as defined in Lemmas 3 and 4, respectively. The time derivative of $V_0(\mathbf{x}^*, t)$ along the solutions of (4) is given by

$$\begin{aligned}\left.\frac{dV_0(\mathbf{x}^*, t)}{dt}\right|_{(4)} &= -\zeta(t)V_0(\mathbf{x}^*, t) + h(t)\{-2\beta V_1(\mathbf{x}^*, t) \\ &\quad + [\theta(t) - 2\alpha]k(t)V_2(\mathbf{x}^*, t) - [r_1(D + \beta)y]^2 \\ &\quad - k(t)[r_2(D + \alpha)q(D)y]^2\} \\ &\leqq \sup_{t \geqq 0}\{[-2\beta - \gamma, \theta(t) - 2 - \gamma] - \zeta(t)\}V_0(\mathbf{x}^*, t).\end{aligned} \tag{28}$$

3. Main results.

THEOREM 1. *If*

(a) $Z(s)$ (*the "multiplier"*) *is a function of the complex variable* s *such that* $Z(s - \alpha)$ *is positive real for some* $\alpha \geqq 0$;

(b) $Z(s - \beta)G(s - \beta)$ *is positive real for some* $\beta \geqq 0$, *and* $m(s - \beta)q(s - \beta)$, $n(s - \beta)p(s - \beta)$ *have no imaginary zeros*; *and*

(c) *for some positive constant* γ,

$$-\int_{t_0}^{t} \sup_{\tau \geqq 0}\left[-2\beta - \gamma, \left(\frac{dk}{d\tau}\Big/k\right) - 2\alpha - \gamma\right] d\tau \tag{29}$$

increases without bound as $t \to \infty$,
then the null solution of (1) *is asymptotically stable.*

THEOREM 2. *If the hypotheses* (a) *and* (b) *of Theorem* 1 *are valid,* M *is a positive constant,* $\theta(t)$ *and* $[\theta(t) + 2(\beta - \alpha)]^+$ *denote respectively* $((dk/dt)/k)$ *and the positive values of* $[\theta(t) + 2(\beta - \alpha)]$, *and*

(d)

$$\text{(i)} \qquad \frac{1}{T}\int_{t_0}^{t_0+T} [\theta(\tau) + 2(\beta - \alpha)]^+ d\tau \leqq M < \infty \tag{30}$$

for all finite $T > 0$,

$$\text{(ii)} \qquad \lim_{T\to\infty} \frac{1}{T}\int_{t_0}^{t_0+T} [\theta(\tau) + 2(\beta - \alpha)]^+ d\tau \leqq 2\beta + \gamma - \nu$$

for some $\nu > 0$,

then the null solution of (1) *is exponentially stable.*

Proof of Theorem 1. As a Lyapunov function candidate for (4) and hence for (1), choose

$$V(\mathbf{x}^*, t) = V_1(\mathbf{x}^*, t) + k(t)V_2(\mathbf{x}^*, t),$$

where $V_1(\mathbf{x}^*, t)$ and $V_2(\mathbf{x}^*, t)$ are defined by (10) and (18) respectively. Clearly, $V(\mathbf{x}^*, t)$ is positive definite, radially unbounded, has continuous first partial derivatives, and satisfies decrescent conditions by virtue of the boundedness of $k(t)$. Its time derivative along the solutions of (4) was shown to satisfy inequality (24). Invoking the Krasovskii–Corduneanu theorem (Corollary 1), we conclude that hypothesis (c) implies asymptotic stability.

Proof of Theorem 2. As a Lyapunov function candidate for (4), choose

$$V_0(\mathbf{x}^*, t) = h(t)\{V_1(\mathbf{x}^*, t) + k(t)V_2(\mathbf{x}^*, t)\},$$

where $V_1(\mathbf{x}^*, t)$, $V_2(\mathbf{x}^*, t)$ are defined by (10), (18) respectively, and $h(t)$ is defined in subsection K above. $V_0(\mathbf{x}^*, t)$ is positive definite, radially unbounded, has continuous partial derivatives, and satisfies decrescent conditions by virtue of the boundedness of $k(t)$ and $h(t)$. Its time derivative along the solutions of (4) was shown to satisfy inequality (28). Invoking Corollary 2 of the Krasovskii–Corduneanu theorem, we conclude that hypothesis (d) implies exponential stability.

Remarks. (a) The present global condition (29) for asymptotic stability of (1) differs from the Freedman and Zames average condition (3) in that the integrand of (29) may assume negative values, whereas the integrand of (3) may not.

(b) (Note that $\alpha \geqq \beta$.) The average condition (30) for exponential stability of (1) allows larger positive variations of $\theta(t)$ over a finite interval than (3) and the negative lobes of $\theta(t)$ do not enter into the integrand of (30).

(c) The condition on $G(s)$ in Theorems 1 and 2 can be replaced by an explicit geometric condition, free of multipliers (for details see [1] and [9]).

REFERENCES

[1] M. Freedman and G. Zames, *Logarithmic variation criteria for the stability of systems with time-varying gains*, this Journal, 6 (1968), pp. 487–507.

[2] G. Zames, *On the stability of nonlinear time-varying feedback systems*, Proc. NEC, 20 (1964), pp. 725–730.

[3] K. S. Narendra and R. M. Goldwyn, *A geometrical criterion for the stability of certain nonlinear nonautonomous systems*, IEEE Trans. Circuit Theory, CT–11 (1964), pp. 406–407.

[4] I. W. Sandberg, *A frequency domain condition for the stability of systems containing a single time varying nonlinear element*, Bell System Tech. J., 43 (1964), pp. 1601–1638.

[5] R. W. Brockett and L. J. Forys, *On the stability of systems containing a time-varying gain*, Proc. Second Allerton Conference, University of Illinois, Urbana, 1964, pp. 413–430.

[6] N. N. Krasovskii, *Stability of Motion*, Stanford University Press, Stanford, Calif., 1963.

[7] C. Corduneanu, *The application of differential inequalities to the theory of stability*, An. Sti. Univ. "Al. I. Cuza" Iasi Sect. I (N.S.), 6 (1960), pp. 47–58.

[8] L. Weinberg and P. Slepian, *Positive real matrices*, J. Math. Mech., 9 (1960), pp. 71–83.

[9] R. W. Brockett and J. L. Willems, *Frequency-domain stability criteria, I*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 255–261.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

A NOTE ON THE NECESSARY CONDITIONS FOR OPTIMAL STRATEGIES IN A CLASS OF NONCOOPERATIVE N-PERSON DIFFERENTIAL GAMES*

I. G. SARMA AND U. R. PRASAD†

Abstract. In this note, we show that the main results of our previous paper [1], the properties of the value function vector and the equilibrium point principle stated in Theorems 1 and 2 of [1], are extendable to a slightly larger but considerably more realistic class of games.

In general, an N-person game can have multiple equilibrium points. The equilibrium points of a two-person zero-sum game (called the saddle points), if nonunique, are automatically both interchangeable and equivalent. This is *not* the case for a general N-person game. However, if all the equilibrium points of an N-person game are interchangeable (without necessarily being equivalent), then all of them constitute the Nash noncooperative solution of the game. These notions are discussed by Nash [2] and Luce and Raiffa [3].

In our original paper [1], the solution of an N-person game is characterized by (2.4) and the resulting payoff function is assumed single-valued on $\mathscr{R}$. This is equivalent to assuming the twin conditions of interchangeability and equivalence which are valid in general *only* for two-person zero-sum games. To emphasize that two-person zero-sum ideas are not applicable when dealing with N-person games, we introduce the more realistic concept known as the Nash noncooperative solution for these games. This solution also satisfies (2.4) but the resulting payoff function can be discontinuous across certain well-defined manifolds in $\mathscr{R}$. The following simple example bears out these points.

Example. The state of the game satisfies the following differential equations:

$$\begin{aligned} \dot{x}_1 &= -x_2, \\ \dot{x}_2 &= u^1 + cu^2 \end{aligned} \tag{1}$$

with u^1 and u^2 constrained as

$$|u^1| \leqq 1; \qquad |u^2| \leqq 1. \tag{2}$$

The terminal surface $\mathscr{T}_1 \cup \mathscr{T}_2$ is given in terms of the parameter $\sigma = (x_{2_f}, t_f)$ as

$$\begin{aligned} \mathscr{T}_1 &= \left\{ (x_{1_f}, x_{2_f}) : x_{1_f} = \frac{x_{2_f}^2}{2(1+c)}; x_{2_f} \geqq 0 \right\}, \\ \mathscr{T}_2 &= \left\{ (x_{1_f}, x_{2_f}) : x_{1_f} = \frac{x_{2_f}^2}{2}; x_{2_f} \geqq 0 \right\}. \end{aligned} \tag{3}$$

The playing space is in between $\mathscr{T}_1$ and $\mathscr{T}_2$ as shown in Fig. 1. The payoff functionals

* Received by the editors April 7, 1970, and in revised form December 10, 1970.

† Department of Electrical Engineering, Indian Institute of Technology, Kanpur, India.

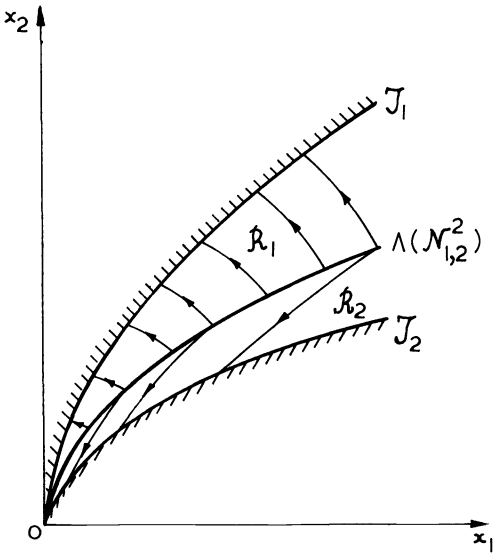


FIG. 1. *Nash noncooperative solution of example*

of the players are given by

$$J^1[x_0, \mathbf{u}] = \phi^1(\sigma) + \int_0^{t_f} dt,$$
$$J^2[x_0, \mathbf{u}] = \phi^2(\sigma) + \int_0^{t_f} \{|u^1| + b|u^2|\}\, dt, \tag{4}$$

where

$$\phi^1(\sigma) = \begin{cases} \dfrac{x_{2_f}}{(1+c)}, & x_f \in \mathcal{T}_1, \\ x_{2_f}, & x_f \in \mathcal{T}_2, \end{cases}$$
$$\phi^2(\sigma) = \begin{cases} \dfrac{x_{2_f}(1+b)}{(1+c)}, & x_f \in \mathcal{T}_1, \\ x_{2_f}, & x_f \in \mathcal{T}_2. \end{cases} \tag{5}$$

The constants b and c in (1) and (4) are related by

$$2 < c < b. \tag{6}$$

The optimal control actions should necessarily minimize the Hamiltonians for the players. Hence, we have

$$H^1 = 1 - \lambda_1^1 x_2 + \lambda_2^1(u^1 + cu^2),$$
$$H^2 = |u^1| + b|u^2| - \lambda_1^2 x_2 + \lambda_2^2(u^1 + cu^2), \tag{7}$$

and

$$u^{1*} = -\operatorname{sgn} \lambda_2^1,$$
$$u^{2*} = -\operatorname{dez}(\lambda_2^2 c/b), \tag{8}$$

where we define

$$\text{sgn } z = \begin{cases} +1, & z > 0, \\ -1, & z < 0, \end{cases}$$

$$\text{dez } z = \begin{cases} +1, & z > 1, \\ 0, & |z| < 1, \\ -1, & z < -1. \end{cases} \tag{9}$$

The adjoint equations are given for $l = 1, 2$ by

$$\begin{aligned} \dot{\lambda}_1^l &= 0, \\ \dot{\lambda}_2^l &= \lambda_1^l. \end{aligned} \tag{10}$$

As u^{1*} and u^{2*} assume the only values of ± 1 and 0, in view of (9), the term $H_{u^m}^l U_x^m$ corresponding to the other player's optimal strategy is absent in (10).

Applying the transversality condition (4.4) at $\mathscr{T}_1$, we have

$$\begin{aligned} \frac{1}{1+c} - \lambda_1^1\left(\frac{x_{2_f}}{1+c}\right) - \lambda_2^1 &= 0, \\ 1 - \lambda_1^1 x_{2_f} + \lambda_2^1(u^1 + cu^2) &= 0, \\ \frac{1+b}{1+c} - \lambda_1^2\left(\frac{x_{2_f}}{1+c}\right) - \lambda_2^2 &= 0, \\ |u^1| + b|u^2| - \lambda_1^2 x_{2_f} + \lambda_2^2(u^1 + cu^2) &= 0 \end{aligned} \tag{11}$$

with the quantities referred to time t_f. Solving (11) to be consistent with (6), (8) and (10) yields

$$\begin{aligned} \lambda_1^1(t_f) &= \frac{1}{x_{2_f}}; & \lambda_2^1(t_f) &= 0, \\ \lambda_1^2(t_f) &= \frac{2+c+b}{(2+c)x_{2_f}}; & \lambda_2^2(t_f) &= \frac{b}{2+c} \end{aligned} \tag{12}$$

and

$$u^1(t) = 1; \quad u^2(t) = 0 \quad \text{for } t < t_f. \tag{13}$$

By a similar application at $\mathscr{T}_2$, we have

$$u^1(t) = 1; \quad u^2(t) = -1 \quad \text{for } t < t_f. \tag{14}$$

The resulting paths are shown in Fig. 1. It is clear that the strategy of player 1 is continuous on $\mathscr{R}$. The second player's strategy is discontinuous giving rise to his dispersal surface Λ. This surface divides the playing space into two regions, $\mathscr{R}_1$ and $\mathscr{R}_2$. (Because of this, this switching surface could be denoted by $\mathscr{N}_{1,2}^2$.) For starting points on Λ, the value to the second player W^2 must be the same

whether the paths reach $\mathscr{T}_1$ or $\mathscr{T}_2$. By actually equating the values, we get

$$\Lambda = \left\{(x_1, x_2) : x_1 = \frac{\eta}{2} x_2^2, x_2 \geqq 0\right\}, \tag{15}$$

where η is given by

$$\frac{b+1}{c-1} - \left(\frac{2+b-c}{c-1}\right)\left(\frac{-1+(c-1)\eta}{c-2}\right)^{1/2} = -1 + (2+c+b)\left(\frac{1+\eta}{(1+c)(2+c)}\right)^{1/2}. \tag{16}$$

The optimality of this solution can be conclusively shown by verifying the Bellman equations for W^1 and W^2. It can be further verified easily that for starting points on Λ, the value function W^1 of player 1 is greater for the paths reaching $\mathscr{T}_2$ in comparison with the paths ending on $\mathscr{T}_1$. This result has no counterpart in two-person zero-sum differential games, for obvious reasons.

Now, we make the necessary changes in the results of our original paper to include these situations. Let $\mathbf{U}^*$ be a playable N-tuple which introduces a regular decomposition on $\mathscr{R}$. The hypersurface $\mathscr{N}^l_{i_1,i_2}$ which (if it exists) separates $\mathscr{R}_{i_1}$ and $\mathscr{R}_{i_2}$ and across which U^{l*} is discontinuous, is called the dispersal surface of player l (R. Isaacs, [4]). The payoff $\mathbf{P}(t, x, \mathbf{U}^*)$ is single-valued in $\mathscr{R}$ except on the dispersal surfaces. On the dispersal surface $\mathscr{N}^l_{i_1,i_2}$ of player l, the lth component of $\mathbf{P}(t, x, \mathbf{U}^*)$ is assumed to be independent of the two optimal paths branching into $\mathscr{R}_{i_1}$ and $\mathscr{R}_{i_2}$. Then $\mathbf{U}^*$ is said to be the Nash noncooperative solution relative to the normal form defined by the strategy sets $U^l_1, l = 1, 2, \cdots, N$, if the addition $\mathbf{P}(t, x, \mathbf{U}^*)$ satisfies (2.4) of [1] with the interpretation that suitable one-sided limits are taken whenever necessary. The rest of the assumptions on the optimal paths remain the same as in the paper. The solution $\mathbf{U}^*$ in this sense consists of interchangeable (but not necessarily equivalent) equilibrium points, whenever multiple paths arise.

For the enlarged class of games having the Nash noncooperative solution, Theorem 1 holds with the difference that the value function $\mathbf{W}$ need not be continuous across the N manifolds. As explained earlier, W^l will however be continuous across $\mathscr{N}^l_{i_1,i_2}$. Analogous to [5], it is straightforward to show that for any (t, x) on $\mathscr{N}^l_{i_1,i_2}$, the following condition holds:

$$H^l(t, x, (\mathbf{U}^*; U^l_{i_1}), \lambda^l_{i_1})\, dt - \lambda^l_{i_1}\, dx = H^l(t, x, (\mathbf{U}^*; U^l_{i_2}), \lambda^l_{i_2})\, dt - \lambda^l_{i_2}\, dx, \tag{17}$$

where dt and dx are any differentials on the manifold and the subscripts i_1, i_2 indicate the appropriate one-sided limits.

It may be noted that in (4.3), it is tacitly assumed that the terminal surface has the parametric representation

$$t = T_{ij_i}(\sigma); \qquad x = X_{ij_i}(\sigma). \tag{18}$$

Similarly, in (4.6), the following parametric representation is assumed for the

manifold $\mathscr{M}_{ik}$:

$$t = T_{ik}(\sigma); \qquad x = X_{ik}(\sigma). \tag{19}$$

These are clear from the context.

REFERENCES

[1] I. G. SARMA, R. K. RAGADE AND U. R. PRASAD, *Necessary conditions for optimal strategies in a class of noncooperative N-person differential games*, this Journal, 7 (1969), pp. 637–644.
[2] J. NASH, *Non-cooperative games*, Ann. of Math., 54 (1951), pp. 286–295.
[3] R. O. LUCE AND H. RAIFFA, *Games and Decisions*, John Wiley, New York, 1957.
[4] RUFUS ISAACS, *Differential Games*, John Wiley, New York, 1965.
[5] L. D. BERKOVITZ, *Necessary conditions for optimal strategies in a class of differential games and control problems*, this Journal, 5 (1967), pp. 1–24.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

EXISTENCE OF OPTIMAL STOCHASTIC CONTROL LAWS*

V. E. BENEŠ†

Abstract. We give an approach to optimal control of systems described by stochastic functional-differential equations of Itô's type: $dx(t) = f(t, x, u(t, x))\,dt + dw(t)), 0 \leqq t \leqq 1$, with a cost functional $k(u) = E\int_0^1 c(t, x, u(t, x))\,dt$ to be minimized. Here $w(\cdot)$ is Brownian motion, f and c are nonanticipative functionals describing system dynamics and cost rate respectively, and $u(t, \cdot)$ is a causal control law, to be chosen, taking values in a space Γ of control points, and depending at t at most on given information about the past $\{x(s), s \leqq t\}$.

A key role is played by the information available for control. This is described by giving, for each t, a sub-σ-algebra G_t of the σ-algebra S_t over the continuous functions $C = C[0, 1]$ generated by sets $\{y : y(s) \in A\}$ with $0 \leqq s \leqq t$ and A Borel in R^d. S_t is the σ-algebra corresponding to knowing the whole past of the trajectory prior to t. The set $\mathscr{U}$ of admissible control laws consists of functions $u : [0, 1] \times C \to \Gamma$ which are Lebesgue in t, and G_t-measurable in y for each t. There is a σ-algebra G over $[0, 1] \times C$ such that admissibility is equivalent to G-measurability.

Our formulation is based on a result of Girsanov (Teoriya Veroyatnostei, 5 (1960), p. 285): For φ a nonanticipative functional of Brownian motion w, the transformed measure $d\tilde{P} = \exp\zeta(\varphi)\,dP$ with

$$\zeta(\varphi) = \int_0^1 \varphi\,dw - \frac{1}{2}\int_0^1 |\varphi|^2\,dt$$

makes the functions $w(\cdot) - \int_0^{\cdot}\varphi\,dt$ a Wiener process, provided $E\exp\zeta(\varphi) = 1$. This result suggests and justifies taking, for the solution $x(\cdot)$ of the system equations, the process determined by Girsanov's device with $\varphi = f(t, w, u(t, w))$; in this case we say that u attains the density $\exp\zeta(\varphi)$.

The control problem is reformulated as a search for admissible u that achieve $\inf_{u\in\mathscr{U}} E\exp\zeta(\varphi)\int_0^1 c\,dt$. That is, with each $u \in \mathscr{U}$ we associate, as the solution of the system equations to be considered for the purpose of our minimization of $k(\cdot)$, the functions $w(\cdot)$ under the measure $\exp\zeta(\varphi)\,dP$, with the justification that under this measure

$$w(t) - \int_0^t f(s, w, u(s, w))\,ds$$

is a Wiener process.

Novelty of the approach lies in these features: (i) control is closed loop; (ii) admissible controls need not be smooth; (iii) the Radon–Nikodym derivative used by Girsanov directly gives a measure corresponding to a solution of the system equations.

As a principal result, we prove that if Γ is compact metric, if $f(t, y, u)$ grows at most linearly with $y(t)$, and if $G_t = S_t$ (i.e., if the whole past is available for control), then the set of densities $\{\exp\zeta(\varphi) : \varphi = f(t, w, u(t, w)), u \in \mathscr{U}\}$ attainable by the admissible control laws is convex, and there exists an optimal control law $u^* \in \mathscr{U}$ achieving $\inf_{u\in\mathscr{U}} E\exp\zeta(\varphi)\int_0^1 c\,dt$.

1. Introduction.

We give a new approach to two aspects of the optimal control of systems described by stochastic functional-differential equations. These aspects are the *formulation* of the control problem, and the *existence of* optimal control laws. Particular emphasis is placed on the role of the information about the past of the trajectory that is available to the controller as a basis for control decisions.

Novelty of the approach lies in four features: (i) A generalized notion of what is a solution of the stochastic system equations is used: a Radon–Nikodym

* Received by the editors May 19, 1970, and in revised form October 19, 1970.

† Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07974.

derivative functional used by Girsanov [1] directly gives a measure corresponding to a solution. (ii) Control is closed loop. (iii) No restriction of the admissible control laws to be (for example) Lipschitz is imposed; a control law is admissible if it is unambiguous, nonanticipative, suitably measurable, and restricted in its dependence to the available information; it does not have to be smooth. (iv) A general description of the information available to the controller is assumed; to our knowledge, only the case of complete information about the past carries with it sufficient convexity to yield closure and existence results by the usual methods.

The systems of principal interest here are those governed by a stochastic functional-differential equation

$$dx(t) = f(t, x, u(t, x))\, dt + dw(t), \qquad t \in [0, 1], \tag{1}$$

where $w(\cdot)$ is a d-dimensional Brownian motion, f is a d-vector-valued nonanticipative functional, and the "control" u is a function (with values in a prescribed space Γ) whose value at t may depend at most on specified information about the past of $x(\cdot)$ prior to t. The tasks to be tackled are first, (because u need not be smooth) to formulate a suitable meaning for the "existence" of solutions to (1), second, to pose a minimization problem for cost functionals

$$k(u) = E \int_0^1 c(t, x, u(t, x))\, dt, \qquad c \geqq 0,$$

under the constraint (1) and the information restriction on u, and third, to prove within this formulation that optimal admissible controls exist. Cost criteria involving final values of smooth functions and random stopping times can also be handled by our approach, as the reader can verify.

We first informally describe our approach to the existence problem by considering the system (1) with $u(\cdot, \cdot)$ chosen and fixed, and we set

$$h(t, y) = f(t, y, u(t, y))$$

for y in the space C of continuous R^d-valued functions. We also assume that a Brownian motion $w(\cdot)$ is given on a probability space of points ω. In many cases of interest, e.g., if $h(t, \cdot)$ is Lip in y uniformly in t, (1) has a unique solution [2], [3], a stochastic process $x(t, \omega)$ with continuous sample paths; the measure ν which $x(\cdot, \omega)$ induces on C is absolutely continuous with respect to Wiener measure μ, with Radon–Nikodym derivative [3]:

$$\frac{d\nu}{d\mu}(y) = \exp\left\{\int_0^1 h(t, y)\, dy(t) - \frac{1}{2}\int_0^1 |h(t, y)|^2\, dt\right\}, \qquad y \in C. \tag{2}$$

In such a case we do not need to solve (1) to find the cost of using control law $u(\cdot, \cdot)$; the cost is simply

$$k(u) = \int_C \int_0^1 c(t, y, u(t, y))\, dt \frac{d\nu}{d\mu}(y)\, d\mu(y). \tag{3}$$

These circumstances suggest directly using a Radon–Nikodym derivative functional like (2) to *define* a solution of (1), to single it out as the relevant one,

and thereby to give the definite form (3) to the cost of using the control $u(\cdot,\cdot)$. That this is possible was shown by the late I. V. Girsanov [1] in a 1960 paper that is only now attracting the attention it deserves.

Girsanov considered a nonanticipating Brownian functional $\varphi(t,\omega)$, square-integrable almost surely, and defined the functional

$$e^{\zeta_s^t(\varphi)} = \exp\left\{\int_s^t \varphi(t,\omega)\,dw(t) - \frac{1}{2}\int_s^t |\varphi(t,\omega)|^2\,dt\right\}. \tag{4}$$

He then showed that if $E\exp\zeta_s^t(\varphi) = 1$, then the transformed measure $d\tilde{P} = \exp\zeta_s^t(\varphi)\,dP$ makes the function

$$w(\cdot,\omega) - w(s,\omega) - \int_s^{\cdot} \varphi(u,\omega)\,du$$

a Wiener process on $[s,t]$, i.e., it makes $w(\cdot,\omega) - w(s,\omega)$ a solution of the equation

$$w(t) - w(s) - \int_s^t \varphi(s,\omega)\,ds = \text{Wiener process}.$$

In other words, the functional "solves" the stochastic differential equation $dx = \varphi(t,\omega)\,dt + dw$ by giving a transformation of the basic measure into one corresponding to a solution. Girsanov's result is a generalization of Cameron and Martin's translation theorem [4] in which φ was not random.

We propose to use Girsanov's theorem to accomplish our first task, the clarification of the existence of solutions, as follows: to obtain the stochastic process corresponding to use of $u(\cdot,\cdot)$ as control law, we put

$$\varphi(t,\omega) = f(t, w(\omega), u(t, w(\omega))), \qquad \zeta(\varphi) = \zeta_0^1(\varphi), \tag{5}$$

in Girsanov's functional (4), and we take the functions $w(\cdot,\omega)$ under the measure $\exp\zeta(\varphi)\,dP$ as the required process. This procedure gives the conveniently explicit form

$$E\,e^{\zeta(\varphi)}\int_0^1 c(t, w(\omega), u(t, w(\omega)))\,dt \tag{6}$$

for the cost of using $u(\cdot,\cdot)$, and obviates solving the system equation (1), salutary effects indeed; the cost can be computed by numerical integration over Wiener space!

Our second task, of posing a minimization problem, is now straightforward. We describe the information available for control in terms of a σ-algebra G on $[0,1]\times C$, and we can characterize the admissible controls as the G-measurable functions taking values in the control space Γ. The control problem becomes this: To minimize (6) over admissible controls $u(\cdot,\cdot)$, with the understanding (5).

With a definite minimization problem at hand, we can pass to the third task, proving that optimal admissible control laws exist. Our basic result is that if the entire past of the trajectory is available for control, then optimal admissible control laws exist, provided that $f(t,y,u)$ is continuous in the control variable u, that it does not grow faster than linearly in $y(t)$, that $f(t,y,\Gamma)$ is convex, and that Γ is compact metric. In this case it is possible to construct, out of an arbitrary

minimizing sequence of control laws, a new minimizing sequence for which the corresponding stochastic processes converge to a process obtained by using an optimal control law. Closure and existence theorems, analogous to those used in deterministic optimal control, turn out to be difficult to prove, harder than in the nonstochastic case. This additional difficulty is due in large part to the role of the information available to the controller. As in the nonstochastic problems, convexity and continuity are basic in existence proofs, as is a version of Filippov's lemma [5].

2. Historical remarks. The problem of the existence of optimal stochastic control laws for processes of diffusion type has been studied by Kushner [6] for Lipschitz closed loop controls and measurable open loop control entering linearly. Fleming and Nisio [2] considered open loop controls entering multiplicatively, and proved existence theorems using Prokhorov's topology. The case of incomplete information was broached by Fleming [7] in a paper on control of partially observable diffusions: only some of the state-vector components are observed.

The functionals e^{ζ} originated with the work [4] of Cameron and Martin on Wiener measure. After Itô extended stochastic integrals to random but non-anticipative integrands, they reappeared in the work of Ventcel [8] on additive functionals, in that of Skorokhod [3] on differentiability of measures corresponding to diffusion processes, and in Girsanov's work. They were first used in control theory by Mortensen [9]; Kailath [10], Duncan [11], and Kallianpur and Striebel [12] have noted their relevance to estimation and filtering. They are implicit in the work [13] of Stroock and Varadhan on diffusion processes with continuous coefficients, and explicit in McKean's exposition [14] of stochastic differentials and integrals.

3. Formulation. Let Γ be a compact metric space of control points, and let $C = C[0, 1]$ denote the space of continuous functions $y(\cdot)$ with $y:[0, 1] \to R^d$. The sets $\{y(\cdot) \in C : y(s) \in A\}$ for $0 \leqq s \leqq t \leqq 1$ and A Borel in R^d, generate a σ-algebra S_t of C-subsets. This is the σ-algebra representing knowledge of the past from 0 to t. We shall suppose that the system dynamics are given by a function $f:[0, 1] \times C \times \Gamma \to R^d$ with these properties:

(i) $f(t, y, \cdot)$ is continuous on Γ for each t, y;

(ii) f is nonanticipative in the strong sense that $f(t, \cdot, u)$ is measurable with respect to S_t for each $t, u \in [0, 1] \times \Gamma$;

(iii) $|f(t, y, u)|^2 \leqq \kappa(1 + |y(t)|^2)$, κ a constant, for every $t, y, u \in [0, 1] \times C \times \Gamma$,

(iv) $f(\cdot, y, u)$ is Lebesgue measurable for $y, u \in C \times \Gamma$.

We take the view that an admissible control law is a function $u:[0, 1] \times C \to \Gamma$ with the interpretation that $u(t, \cdot)$ indicates what point of the control space Γ is to be exercised as control at time t, and with the proviso that $u(t, \cdot)$ depend only on whatever information (about the past of the trajectory) the controller is allowed to know, remember, and use at time t. The restrictions on $u(\cdot, \cdot)$ representing the pattern of available information will be described mathematically by the concept of measurability with respect to a σ-algebra. Roughly speaking, if $\mathscr{A}_1$ and $\mathscr{A}_2$ are two σ-algebras over the same space, and if $\mathscr{A}_1 \subseteq \mathscr{A}_2$, then the functions measurable on $\mathscr{A}_1$ are less complicated, ("depend on less") than those which are measurable with respect to $\mathscr{A}_2$ but not $\mathscr{A}_1$.

The information pattern will be described by specifying, for each $t \in [0,1]$, a σ-algebra $G_t \subseteq S_t$, and by imposing the condition that for each t $u(t,\cdot)$ be G_t-measurable.

Thus the *admissible controls* will be all those functions $u:[0,1] \times C \to \Gamma$ such that

(i) $u(\cdot, y)$ is Lebesgue for $y \in C$,

(ii) $u(t,\cdot)$ is G_t-measurable for $t \in [0,1]$.

It is possible and convenient to express the property of being an admissible control in terms of a single σ-algebra. This is done as follows: consider the measurable subsets E of $[0,1] \times C$ such that

(i) every t-section of E is a G_t-set, for $t \in [0,1]$,

(ii) every y-section of E is a Lebesgue set, for $y \in C$.

It is easy to verify that G is an algebra; since G is closed under monotone limits, it is a σ-algebra; it can then be proved that a function h on $[0,1] \times C$ is G-measurable if $h(t,\cdot)$ is G_t-measurable for fixed $t \in [0,1]$ and $h(\cdot, y)$ is Lebesgue measurable for $y \in C$. Thus measurability with respect to G concisely expresses the requirement of admissibility.

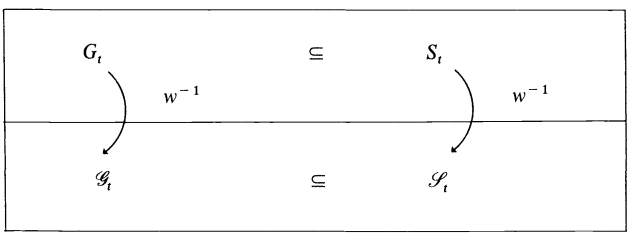
The preceding assumptions have concerned the information available for control and the system dynamics. The reader should note, though, that although σ-algebras over C were used, no probabilistic machinery has been introduced yet; this is now done.

We assume as given a probability space $(\Omega, P, \mathscr{B})$ of points $\omega \in \Omega$, with $P(A)$ the probability of a $\mathscr{B}$-set A. On this space is defined a measurable separable Brownian motion $\{w(t,\omega), 0 \leqq t \leqq 1, \omega \in \Omega\}$ taking values in R^d, with continuous sample paths.

There is a set $\Omega_0 \in \mathscr{B}$ of full measure such that $w(\cdot,\omega) \in C$ for $\omega \in \Omega_0$. The process induces a measurable map $w:\Omega_0 \to C$ according to the formula $w(\omega) = w(\cdot,\omega)$. This is seen as follows: every set $\{y \in C : y(t) \in A\}$ for A Borel is in S_1; every set $\{\omega : w(t,\omega) \in A\}$ for A Borel is in $\mathscr{B}$; but

$$\{\omega : w(t,\omega) \in A\} \cap \Omega_0 = w^{-1}\{y \in C : y(t) \in A\},$$

and so $w^{-1}S_1 \subseteq \mathscr{B}$. The classes $w^{-1}G_t$, $t \in [0,1]$, and $w^{-1}S_t$, $t \in [0,1]$, are all σ-algebras; they will provide us with a way of doing all our work in the probability space $(\Omega, P, \mathscr{B})$ and then returning, for our control laws, to the space C. Setting $\mathscr{G}_t = w^{-1}G_t$ and $\mathscr{S}_t = w^{-1}S_t$, we have the following diagram:



algebras over C

induced algebras over Ω_0

express what is available for control

express the total past

Let $\varphi:[0,1]\times\Omega_0\to[0,1]\times C$ according to the formula $\varphi(t,\omega)=t,w(\omega)$. Define $\mathscr{F}=\varphi^{-1}G$, where G is the σ-algebra expressing admissibility. $\mathscr{F}$ is a σ-algebra of $[0,1]\times\Omega_0$ sets, and will be used to express the requirement of admissibility in terms of functions of t,ω rather than t,y for $y\in C$. Indeed, we shall prove the existence of optimal control laws by first expressing them as $\mathscr{F}$-measurable t,ω functions, and then (properly, now) as G-measurable t,y functions, using the following result.

LEMMA. *If $h:[0,1]\times\Omega_0\to\Gamma$ compact metric is $\mathscr{F}$-measurable, then there exists a G-measurable $u:[0,1]\times C\to\Gamma$ such that*

$$u(t,w(\omega))=h(t,\omega),\qquad \omega\in\Omega_0.$$

This result is elementary (see [15, p. 185]).

The following convenient terminology is used:

The set $\mathscr{A}$ of *admissible drifts* consists of all functions $g:[0,1]\times\Omega\to R^d$ of the form

$$g(t,\omega)=f(t,w(\omega),u(t,w(\omega))),\qquad u\in\mathscr{U},$$

i.e., where $u:[0,1]\times C\to\Gamma$ is an admissible control law.

The set $\mathscr{A}$ of *admissible drifts* consists of all functions $g:[0,1]\times\Omega\to R^d$ form

$$\zeta(\omega)=e^{\zeta(g)},\qquad g\in\mathscr{A},$$

i.e., where g is an admissible drift.

It is to be noted that admissible drifts are random processes and attainable densities are random variables.

$(\Omega,\mathscr{B},P)$ is the probability space on which is defined a Wiener process (Brownian motion) $\{w(t,\omega),\omega\in\Omega\}$ in d-dimensions. We propose to cut through all the questions of existence and uniqueness of solutions to (1), and at the same time obtain and use a representation of the criterion value achieved by a given admissible control law u, by taking, as "the solution" of (1), the process obtained by taking the functions $w(\cdot,\omega)$ under the measure

$$d\tilde{P}=e^{\zeta(g)}\,dP,\qquad g=f(t,w,u(t,w)),$$

where it is assumed (and for suitable f, proved) that $\tilde{P}(\Omega)=1$, and where

$$\zeta(g)_\omega=\int_0^1 f(t,w(\omega),u(t,w(\omega)))\,dw(t)-\int_0^1|f(t,w(\omega),u(t,w(\omega)))|^2\,dt.$$

This procedure provides a solution of (1) in the sense that under the transformed measure $\tilde{P}$,

$$w(t,\omega)-\int_0^t f(s,w(\omega),u(s,w(\omega)))\,ds$$

is a Wiener process. If we define

$$W(t,\omega)=w(t,\omega)-\int_0^t f(s,w(\omega),u(s,w(\omega)))\,ds,$$

then $W(\cdot, \omega)$ is a Wiener process under $\tilde{P}$. If in this context we give the new name $x(t, \omega)$ to $w(t, \omega)$, then the definition of $W(\cdot, \omega)$ tells us that

$$x(t, \omega) = \int_0^t f(s, x(\omega), u(s, x(\omega)))\, ds + W(t, \omega).$$

With measure $\tilde{P}$ this asserts that x is the solution of an equation with drift $f(s, x, u(s, x))$ and Brownian noise W, which is what we wanted to find. The correspondence between Ω and the solution curves is the same as before, but now $\tilde{P}$ gives their distribution. This procedure does not assign, to each sample function $w(\cdot, \omega)$ of the original Wiener process, a "solution function" $x(\cdot, \omega)$ constructed out of $w(\omega)$.

It is possible and useful to eliminate the dependence of the criterion on the control law through the cost function $c(\cdot, \cdot, \cdot)$. In the nonstochastic case this is done by adding one more differential equation

$$dx_0(t) = c(t, x, u(t, x))\, dt$$

and minimizing $x_0(1)$. In our stochastic case we can make an analogous simplification, provided that c is nonanticipative, a reasonable condition. Heuristically: we replace the d-vector function f by the $(1 + d)$-vector function c, f; we add another dimension w_0 of Brownian motion, independent of w, to w to get a $(1 + d)$-dimensional motion

$$z = (w_0, w) = w_0, w_1, \cdots, w_d;$$

and we swap the integral

$$\int_0^1 c\, dt$$

for the end-value $w_0(1)$. If now $\xi = \xi_{c,f}$ is defined by

$$\xi = \int_0^1 h(t, z)\, dz(t) - \frac{1}{2}\int_0^1 |h(t, z)|^2\, dt$$

with $h(t, z) = c(t, w(\omega), u(t, w(\omega))), f(t, w(\omega), u(t, w(\omega)))$, then the functions $z(\cdot)$ under the measure $\exp \xi\, dP$ will, if $E\, e^{\xi} = 1$, have the property that

$$z(t) - \int_0^t h(s, z)\, ds$$

forms a Wiener process of $d + 1$ dimensions. Replacement of the time-integral by $w_0(1)$ is justified by the fact that

$$E \int_0^1 c(t, w(\omega), u(t, w(\omega)))\, dt\, e^{\xi(f)} = E w_0(1)\, e^{\xi}.$$

To see this, note that

$$\xi = \zeta(f) + \int_0^1 c(t, w, u(t, w))\, dw_0(t) - \frac{1}{2}\int_0^1 |c(t, w, u(t, w))|^2\, dt$$

and that under the measure $\exp \xi \, dP$ the function

$$w_0(t) - \int_0^t c(s, w, u(s, w)) \, ds$$

is a Wiener process in one dimension; hence,

$$0 = E\left(w_0(1) - \int_0^1 c(s, w, u(s, w)) \, ds \right) e^{\xi}$$

$$= E w_0(1) \, e^{\xi} - E \int_0^1 c \, ds \, e^{\zeta(f)} E\{ e^{\xi - \zeta(f)} | \mathscr{S}_1 \}.$$

If $c(\cdot, \cdot, \cdot)$ satisfies growth conditions similar to those on $f(\cdot, \cdot, \cdot)$, and $\mathscr{S}_1$ denotes the σ-algebra corresponding to the w-process over $[0, 1]$, the conditional expectation on the right will be 1.

Thus, we may and shall assume henceforth that the cost rate $c(t, y, u)$ is simply the zero-index component $f_0(t, y, u)$ of a new $(1 + d)$-dimensional "right-hand side" $f(t, y, u)$, with $f: [0, 1] \times C \times \Gamma \to R^{1+d}$.

Finally we discuss the question of a prescribed initial value $x(0) = a$ for our solution $x(\cdot)$ of

$$dx = f(t, x, u(t, x)) \, dt + dw.$$

As presently formulated this initial condition is $x(0) = 0$, since the Wiener process starts at 0, and we took for $x(\cdot)$ the functions $w(\cdot)$ under the measure $e^{\zeta(g)} \, dP$. The initial condition $x(0) = a \neq 0$ can be realized by taking instead the functions $w(\cdot) + a$ under the same measure; if desired, g can be redefined so as to be a functional of $w(\cdot) + a$ instead of $w(\cdot)$; this is a simple shift. As will be shown, the important thing is that $\exp\{(w(t) + a) \cdot \theta - \theta \cdot \int_0^t g \, ds\}$ be a martingale with respect to $e^{\zeta(g)} \, dP$ for all θ; since this process differs from the previous one in having a factor $\exp a \cdot \theta$, the result is clear. In view of these facts we shall assume $x(0) = 0$ henceforth.

We thus arrive at this formulation of an optimal control problem for stochastic functional-differential equations: To minimize the integral

$$\int w_0(1) \, e^{\zeta(g)} \, dP = E \, w_0(1) \, e^{\zeta(g)}$$

subject to the condition that g be an admissible drift, i.e., have the form

$$g(t, \omega) = f(t, w(\omega), u(t, w(\omega))),$$

where $w(\cdot)$ is a Wiener process, $u(\cdot, \cdot)$ is an admissible control, and

$$f_0(t, y, v) = c(t, y, v), \qquad t, y, v \in [0, 1] \times C \times \Gamma,$$

i.e., the zero component of f is the cost rate c. In this form of the problem we minimize the average of the value of $x_0(\cdot)$ at the endpoint 1; the functional $\exp \zeta(g)$ determines what this averaging is.

The functions $z(t) = E\{e^{\zeta(\varphi)} | \mathscr{S}_t\}$ have many interesting properties, described by McKean [14]; among them is the fact that they are the unique solution of the

stochastic differential equation

$$dz(t) = z(t)\varphi(t)\,dw(t), \qquad 0 \leqq t \leqq 1. \tag{7}$$

This allows us to rephrase the control problem as that of minimizing $E\,x_0(1)z(1)$ subject to (7) and the same conditions on φ as before.

4. Martingale proof of Girsanov's theorem. We have proposed basing an approach to optimal stochastic control on an absolute continuity theorem of Girsanov [1, Theorem 1]. In this section we prove this theorem by a direct martingale method, and note its applicability under our assumptions. Girsanov proved that if φ is a nonanticipative Brownian functional with $\int_0^1 |\varphi|^2\,dt < \infty$ a.s., and if $E \exp \zeta(\varphi) = 1$, then the process

$$w(t) - \int_0^t \varphi\,ds$$

is a Wiener process under the measure $\exp \zeta(\varphi)\,dP$. We shall put his result into a more symmetric form.

THEOREM 1. *Let φ be a nonanticipative Brownian functional with $\int_0^1 |\varphi|^2\,dt < \infty$ a.s. The following are equivalent:*

(i) *$w(t) - \int_0^t \varphi\,ds$ is a Wiener process under $e^{\zeta(\varphi)}\,dP$;*

(ii) *$E\,e^{\zeta(\varphi+\theta)} = 1$ for every constant $\theta \in R^d$;*

(iii) *$E\,e^{\zeta(\varphi)} = 1$.*

Proof of (i) implies (ii): $d\tilde{P} = e^{\zeta(\varphi)}\,dP$ makes $W(t) = w(t) - \int_0^t \varphi\,ds$ a Wiener process, so for $\theta \in R^d$,

$$\tilde{E} \exp\{\theta \cdot W(1) - \tfrac{1}{2}|\theta|^2\} = 1.$$

But

$$\begin{aligned}\zeta(\varphi + \theta) &= \zeta(\varphi) + \zeta(\theta) - \int_0^1 \varphi \cdot \theta\,ds \\ &= \zeta(\varphi) + \theta \cdot w(1) - \tfrac{1}{2}|\theta|^2 - \theta \cdot \int_0^1 \varphi\,ds \\ &= \zeta(\varphi) + \theta \cdot W(1) - \tfrac{1}{2}|\theta|^2.\end{aligned}$$

So

$$E \exp\{\zeta(\varphi + \theta)\} = \tilde{E}\{\theta \cdot W(1) - \tfrac{1}{2}|\theta|^2\} = 1.$$

Proof of (ii) implies (i): Let $E\,e^{\zeta(\varphi)} = 1$. Then since $e^{\zeta_0^t(\varphi)}$ is a supermartingale [14, p. 25] we have a.s., for $0 \leqq s \leqq t \leqq 1$,

$$E\{e^{\zeta_0^t(\varphi)} | \mathscr{S}_s\} \leqq e^{\zeta_0^s(\varphi)},$$

$$E\,e^{\zeta_0^t(\varphi)} = EE\{e^{\zeta_0^t(\varphi)} | \mathscr{S}_s\} \leqq \mathrm{E}\,e^{\zeta_0^s(\varphi)}.$$

Thus $E\,e^{\zeta_0^t(\varphi)} = 1$. Further for $s < t$,

$$0 = 1 - E\,e^{\zeta_0^t(\varphi)} = E\,e^{\zeta_0^s(\varphi)} E\{e_s^t(\varphi) - 1 | \mathscr{S}_s\}.$$

Since $e^{\zeta_0^s(\varphi)} > 0$ a.s. we have

$$E\{e_s^t(\varphi) | \mathscr{S}_s\} = 1 \quad \text{a.s.}$$

In particular (ii) implies that for any $\theta \in R^d$,

$$E\{e^{\zeta_s^t(\varphi+\theta)}|\mathscr{S}_s\} = 1 \quad \text{a.s.}$$

It follows from Doob [22, Theorem 11.9, p. 384] that if a continuous process $y(\cdot)$ has the property that

$$e^{\theta \cdot y(t) - |\theta|^2 t/2}$$

is a martingale for every $\theta \in R^d$, then $y(\cdot)$ is a Wiener process. Thus it is enough to show that

$$\exp\left\{\theta \cdot w(t) - \theta \cdot \int_0^t \varphi \, ds - \frac{1}{2}|\theta|^2 t\right\}$$

is a martingale with respect to $e^{\zeta(\varphi)}\, dP$, i.e., that for $0 \leqq s \leqq t \leqq 1$,

$$\tilde{E}\left\{\exp\left\{\theta \cdot [w(t) - w(s)] - \theta \cdot \int_s^t \varphi \, du - |\theta|^2(t-s)/2\right\}\middle|\mathscr{S}_s\right\} = 1 \quad \text{a.s.}$$

The conditional expectation above is an $\mathscr{S}_s$-measurable function $\psi(\omega)$, $\omega \in \Omega$, such that for $B \in \mathscr{S}_s$,

$$\begin{aligned}
\int_B \psi(\omega)\, e^{\zeta(\varphi)}\, dP &= \int_B \exp\left\{\theta \cdot [w(t) - w(s)] - \theta \cdot \int_s^t \varphi \, du - \tfrac{1}{2}|\theta|^2(t-s)\right\} e^{\zeta(\varphi)}\, dP \\
&= \int_B e^{\zeta_t^1(\varphi) + \zeta_s^t(\varphi+\theta) + \zeta_0^s(\varphi)}\, dP \\
&= \int_B E\{e^{\zeta_t^1(\varphi)}|\mathscr{S}_t\}\, e^{\zeta_s^t(\varphi+\theta) + \zeta_0^s(\varphi))}\, dP \\
&= \int_B E\{e^{\zeta_s^t(\varphi+\theta)}|\mathscr{S}_s\}\, e^{\zeta_0^s(\varphi)}\, dP \\
&= \int_B e^{\zeta_0^s(\varphi)}\, dP \\
&= \int_B e^{\zeta_0^1(\varphi)}\, dP = \tilde{P}(B),
\end{aligned}$$

because the conditional expectations above all equal 1 a.e. It follows that $\psi = 1$ a.e.

To encompass Girsanov's theorem and complete the proof of Theorem 1 we must show that (iii) implies (ii). Accordingly let $E \exp \zeta(\varphi) = 1$, and let φ_N bounded be such that

$$\int_0^1 |\varphi_N - \varphi|^2\, dt \to 0 \quad \text{a.e.,} \qquad N \to \infty.$$

Then also

$$\int_0^1 |\varphi_N + \theta - \varphi - \theta|^2\, dt \quad \text{a.e.,}$$

and thus $\zeta(\varphi_N + \theta) \to \zeta(\varphi + \theta)$ in probability. Together with the hypothesis $E \exp \zeta(\varphi) = 1$, this implies by Girsanov's elementary Lemma 5 [1] that

$$e^{\zeta(\varphi_N + \theta)} \to e^{\zeta(\varphi + \theta)} \quad \text{in } L_1.$$

Each $\varphi_N + \theta$ is bounded, so

$$1 = E\, e^{\zeta(\varphi_N + \theta)} \to E\, e^{\zeta(\varphi + \theta)}, \qquad N \to \infty.$$

This completes the proof of Theorem 1.

Girsanov's theorem can be viewed as stating that the Wiener measure μ on C is a fixed point of the composition of two transformations of measures, the first corresponding to weighting by $e^{\zeta(\varphi)}$, the second to translation by $-\int_0 \varphi\, ds$. Since $e^{\zeta(\varphi)}$ is $\mathscr{S}_1$-measurable and integrable, it corresponds to an S_1-measurable μ-integrable functional η on C according to the formula $\int_A \eta\, d\mu = \int_{w^{-1}(A)} e^{\zeta(\varphi)}\, dP$, $A \in S_1$. For $y \in C$ set $Ty(t) = y(t) - \int_0^t \varphi(s, y)\, ds$. For measures ν on C define $W\nu$ and νT^{-1} for $A \in S_1$ by

$$W\nu(A) = \int_A \eta\, d\nu,$$

$$\nu T^{-1}(A) = \nu(T^{-1}A).$$

Then Girsanov's theorem states that $\mu = (W\mu)T^{-1}$.

5. Some preliminary results. In view of the role of condition $E \exp \zeta(\varphi) = 1$ in application of Girsanov's theorem, it is particularly important to clarify the conditions under which it obtains for attainable densities $\exp \zeta(g)$, $g \in \mathscr{A}$. We prove the following lemma.

LEMMA 0. *$E \exp \zeta(\alpha g) = 1$ for all $g \in \mathscr{A}$ and $\alpha \geqq 0$.*

Proof. The proof to be given is the same for all $\alpha \geqq 0$, so only the case $\alpha = 1$ is described. Let $g \in \mathscr{A}$, and

$$y(t, \omega) = w(t, \omega) - \int_0^t g(s, \omega)\, ds$$

so that in Girsanov's terminology [1] $y(\cdot, \omega)$ is an Itô process with respect to $w(\cdot, \omega)$ corresponding to the matrix I and the vector $g(\cdot, \omega)$. The result will follow from Girsanov's Lemma 7 provided that we find for each $\varepsilon > 0$, an integer $N(\varepsilon)$ and a monotone system of open sets $C_N(t)$ with

$$C_N(t) \in S_t, \qquad C_N(s) \subseteq C_N(t) \quad \text{for} \quad t > s$$

and such that:

(a) $\{\omega : y(\cdot, \omega) \in C_N(t)\} \in \mathscr{S}_t$,

(b) $P\{w(\cdot, \omega) \in C_N(1)\} > 1 - \varepsilon$,

(c) $|g(t, \omega)| < N$ if $y(\cdot, \omega) \in C_N(t)$,

(d) if $x(\cdot) \in C_N(s)$, $x(\cdot) \notin C_N(t)$, $t > s$, then there is a τ with $s < \tau < t$ such that

$$x(\cdot) \in C_N(u) \quad \text{for} \quad u < \tau, \qquad x(\cdot) \notin C_N(\tau).$$

Let

$$C_N(t) = \{x \in C : 1 + 2e^2(\kappa + \sup_{s \in [0,t]} |x(s)|^2) < N^2/\kappa\}.$$

It is easy to verify that properties (a), (b) and (d) obtain. To prove (c) consider that the growth condition on f gives

$$|w(t)|^2 \leqq 2|y(t)|^2 + 2\left(1 + \int_0^t |w(s)|^2\, ds\right),$$

$$\sup_{s\in[0,t]} |w(s)|^2 \leqq 2 \sup_{s\in[0,t]} |y(s)|^2 + 2\kappa\left(1 + \int_0^t \sup_{u\in[0,s]} |w(u)|^2\, ds\right)$$

$$\leqq 2\, e^{2\kappa}\left(\kappa + \sup_{s\in[0,t]} |y(s)|^2\right)$$

by Gronwall's inequality. Hence if $y(\cdot\,, \omega) \in C_N(t)$, then

$$\kappa(1 + |w(t)|^2) < N^2,$$

$$|g(t, \omega)| < N,$$

which proves (c).

LEMMA 1. *There exists a constant $\alpha > 1$ such that*

$$\sup_{g\in\mathscr{G}} E\, e^{\alpha\zeta(g)} < \infty.$$

Proof. Consider that

$$e^{\alpha\zeta(g)} = \exp\left\{\zeta(\alpha g) + \frac{\alpha^2 - \alpha}{2}\int_0^1 |g(t, \omega)|^2\, dt\right\}$$

$$\leqq \exp\left\{\zeta(\alpha g) + \frac{\alpha^2 - \alpha}{2}\kappa\int_0^1 (1 + |w(t)|^2)\, dt\right\}.$$

Since $E\, e^{\zeta(\alpha g)} = 1$, the density $e^{\zeta(\alpha g)}$ makes the function

$$x(t) = w(t) - \alpha\int_0^t g\, ds$$

a Wiener process. Also

$$|w(t)|^2 \leqq 2|x(t)|^2 + 2\alpha^2\kappa\left(1 + \int_0^t |w(u)|^2\, du\right),$$

for which Gronwall's inequality gives

$$|w(t)|^2 \leqq 2\left(\alpha^2\kappa + \sup_{0\leqq t\leqq 1} |x(t)|^2\right) \exp 2\alpha^2\kappa.$$

It follows that

$$E\, e^{\alpha\zeta(g)} \leqq e^{(\alpha^2-\alpha)\kappa/2} E \exp\left\{\zeta(\alpha g) + \kappa(\alpha^2 - \alpha)\left(\alpha^2\kappa + \sup_{0\leqq t\leqq 1} |x(t)|^2\right) \exp 2\alpha^2\kappa\right\}.$$

Since $x(\cdot)$ is a Wiener process under $\exp\zeta(\alpha g)\, dP$, the expectation on the right is of the form

$$h(\alpha) E \exp\left\{\kappa(\alpha^2 - \alpha)\, e^{2\alpha^2\kappa} \sup_{0\leqq t\leqq 1} |w(t)|^2\right\},$$

where $h(\cdot)$ is bounded near 1, and so is finite for $\alpha > 1$ small enough (Doob [loc. cit., p. 392]).

Lemma 1 implies that the set $\mathscr{D} = \{e^{\zeta(g)}, g \in \mathscr{A}\}$ of attainable densities is uniformly integrable.

LEMMA 2. *If $g_n, n \geqq 1$, is a sequence of admissible drifts, and*

$$\xi = \text{weak } \lim_n \exp \zeta(g_n)$$

in L_1 or L_2, then

$$E\{\xi|\mathscr{S}_t\} > 0 \quad a.s., \qquad 0 \leqq t \leqq 1.$$

Proof. If $E\{\xi|\mathscr{S}_t\} = 0$ on a set $A \in \mathscr{S}_t$, of positive measure, then

$$\begin{aligned} E\xi(t,\omega)\chi_A &= EE\{\xi|\mathscr{S}_t\}\chi_A \\ &= EE\{\xi\chi_A|\mathscr{S}_t\} \\ &= E\xi\chi_A = 0. \end{aligned}$$

Since $\xi \geqq 0$ a.e. we must have $\xi = 0$ a.e. on A. By weak convergence $\int_A e^{\zeta_0^1(g_n)}\,dP \to 0$. Hence some subsequence, relabeled $\{e^{\zeta(g_n)}, n \geqq 1\}$, converges to 0 a.e. on A, and so $\zeta_0^1(g_n) \to -\infty$ a.e. on A. But

$$\zeta_0^1(g_n) = \int_0^1 g_n\,dw - \frac{1}{2}\int_0^1 |g_n|^2\,dt.$$

The second term is bounded by $\kappa \int_0^1 [1 + |w(s)|^2]\,ds$, which is independent of n and finite a.e., so $\int_0^1 g_n\,dw \to -\infty$ a.e. on A. Thus

$$P\left\{A \cap \int_0^1 g_n\,dw > -N\right\} \to 0 \quad \text{for every} \quad N > 0.$$

Also though,

$$\begin{aligned} P\left\{A \cap \int_0^1 g_n\,dw \leqq -N\right\} &\leqq P\left\{\left|\int_0^1 g_n\,dw\right| \geqq N\right\} \\ &\leqq N^{-2}E\int_0^1 |g_n|^2\,dt, \\ &\leqq o(1), \end{aligned}$$

uniformly in n.

Now observe that

$$P\{A\} = P\left\{A \cap \int_0^1 g_n\,dw > -N\right\} + P\left\{A \cap \int_0^1 g_n\,dw \leqq -N\right\}.$$

Let $\varepsilon > 0$ be given. Pick first N so large that the second term is less than $\varepsilon/2$ uniformly in n; then pick $n = n(N)$ so large that the first term is less than $\varepsilon/2$. Thus $P\{A\} = 0$.

LEMMA 3 (H. P. McKean, Jr.). *If* $y(t) = \int_0^t \varphi(u, \omega)\, dw(u)$ *and*

$$T(t) = \int_0^t |\varphi(u, \omega)|^2\, du,$$

$$T^{-1}(x) = \inf t : T(t) = x \quad \textit{for} \quad x \leqq \int_0^1 |\varphi|^2\, du,$$

then

$$y(T^{-1}(t))$$

is a Wiener process terminated at $T(1)$.

Proof. Let $\chi_j(t)$ be the indicator of $T^{-1}(t_j) > t$, where $t_1, t_2, \cdots, t_n \in [0, \infty)$ and let

$$z(t) = \exp\left\{ i \sum_{j=1}^{n} \gamma_j \int_0^t \chi_j \varphi\, dw + \frac{1}{2} \sum_{j,k=1}^{n} \gamma_j \gamma_k \int_0^t \chi_j \chi_k |\varphi|^2\, du \right\}.$$

Note that on $T(1) > \max t_j$,

$$\int_0^1 \chi_j \chi_k |\varphi|^2\, du = \int_0^{\min\{T^{-1}(t_j), T^{-1}(t_k)\}} |\varphi|^2\, du$$

$$= \min(t_j, t_k)$$

and

$$\int_0^1 \chi_j \varphi\, dw = y(T^{-1}(t_j)).$$

The stochastic differential of $z(\cdot)$ is

$$dz(t) = iz(t) \sum_{j=1}^{n} \gamma_j \chi_j \varphi\, dw(t)$$

so that

$$z(t) = 1 + i \int_0^t z(u) \sum_{j=1}^{n} \gamma_j \chi_j(u) \varphi(u)\, dw(u).$$

Since

$$|z(t)| \leqq \exp \frac{1}{2} \sum_{j,k=1}^{n} |\gamma_j \gamma_k| t_j t_k,$$

we can multiply by $\chi_{T(1) > \max t_i}$ and take expectations at $t = 1$ to get

$$Ez(1) \chi_{T(1) > \max t_i} = \Pr\{T(1) > \max t_i\}$$

or

$$E \exp\left\{ i \sum_{j=1}^{n} \gamma_j y(T^{-1}(t_j)) \,\middle|\, T(1) > \max t_i \right\} = \exp \frac{1}{2} \sum_{j,k=1}^{n} \gamma_j \gamma_k \min(t_j, t_k).$$

This says that conditional on its still being defined at the maximum time $\max t_i$,

the process $y(T^{-1}(t))$ is Gaussian with a covariance appropriate to Brownian motion.

LEMMA 4. *For $w(\cdot)$ a Wiener process, and $y(\cdot)$ and $T(\cdot)$ as in Lemma 3,*

$$\Pr\{\sup_{0\leqq t\leqq 1}|y(t)|>a\}\leqq \Pr\{\sup_{0\leqq s\leqq t}|w(s)|>a\}+\Pr\{T(1)>t\}.$$

Proof. Let $t\in[0,1]$ so $T(t)$ is defined and $T^{-1}(T(t))\leqq t$. Then

$$y(T^{-1}(T(t)))=y(u)\quad\text{for some}\quad u\in[0,t]$$

and so

$$\sup_{0\leqq t\leqq 1}|y(t)|>a\Rightarrow\sup_{0\leqq u\leqq 1}|y(T^{-1}(T(u)))|>a.$$

Thus, since $y(T^{-1}(\cdot))$ is a Wiener process as long as it is defined,

$$\begin{aligned}\Pr\{\sup_{0\leqq t\leqq 1}|y(t)|>a\}&\leqq\Pr\{\sup_{0\leqq u\leqq 1}|y(T^{-1}(T(u)))|>a\}\\&\leqq\Pr\{\sup_{0\leqq s\leqq T(1)}|y(T^{-1}(s))|>a\}\\&\leqq\Pr\{\sup_{0\leqq s\leqq t}|w(s)|>a\}+\Pr\{T(1)>t\}.\end{aligned}$$

We shall also need a version [15] of an implicit function lemma due to McShane and Warfield [5]; this version allows simultaneous explicit as well as implicit dependence on the independent variable, provided that this dependence is measurable with respect to the same σ-algebra as is the desired function.

If $\mathcal{M}$ is a σ-algebra of subsets of a set M, and S is a topological space, we say that a function $g:N\to S$, $N\in\mathcal{M}$ (defined on N) is $\mathcal{M}$-measurable if and only if $g^{-1}(F)\in\mathcal{M}$ for closed $F\subseteq S$.

LEMMA 5. *Let $(M,\mathcal{M})$ be a measure space, A a separable metric space, and U a compact metric space. Let $k:M\times U\to A$ be continuous in its second argument for each value of the first, and $\mathcal{M}$-measurable in the first for each value of the second. Let $y:M\to A$ be $\mathcal{M}$-measurable, with*

$$y(x)\in k(x,U),\quad x\in M.$$

Then there exists an $\mathcal{M}$-measurable $u:M\to U$ such that

$$y(x)=k(x,u(x)).$$

From Lemma 5 we can obtain an implicit function lemma for limits.

LEMMA 6. *Let $(M,\mathcal{M})$, A and U be as in Lemma 5. Let $h:M\times U\to A$ be continuous in its second argument for each value of the first, and let $u_n:M\to U$ be a sequence of $\mathcal{M}$-measurable functions, and z a function such that pointwise*

$$h(x,u_n(x))\to z(x).$$

Then there exists an $\mathcal{M}$-measurable function $u:M\to U$ such that

$$z(x)=h(x,u(x)).$$

Proof. Since U is compact metric, there is a continuous map ψ of the Cantor set C onto U. Find, by Lemma 5, $\mathcal{M}$-measurable functions $\xi_n:M\to C$ such that $u_n(x)=\psi(\xi_n(x))$, $x\in M$. Define $\xi(x)=\limsup_{n\to\infty}\xi_n(x)$, $u(x)=\psi(\xi(x))$. Fix x

and find $n_i = n_i(x)$, $i = 1, 2, \cdots$, such that

$$\xi_{n_i(x)}(x) \to \xi(x).$$

Then $h(x, \psi(\xi_{n_i(x)}(x))) \to h(x, \psi(\xi)))$. Thus $z(x) = h(x, \psi(\xi(x)))$. u is $\mathcal{M}$-measurable because if 0 is an open set of U, then

$$u^{-1}(0) = \{x : \xi(x) \in \psi^{-1}(0)\} \in \mathcal{M},$$

since $\psi^{-1}(0)$ is open in C, because ψ is continuous.

6. Convexity. In deterministic control theory there are counterexamples which indicate that restrictions must be placed on the "right-hand side" of the constraining differential equation if optimal control laws are to exist. Conditions of linearity or convexity have been imposed for this purpose. The fundamental paper [16] of Markus and Lee postulated a form linear in the control $u(\cdot)$:

$$\dot{x}(t) = f(t, x(t), u(t)) = g(t, x(t)) + H(t, x(t))u(t), \tag{8}$$

with H a matrix and u a vector. Roxin [17] suggested that it was enough to assume convexity of $f(t, y, \Gamma)$ for each t, y, with Γ the space of control points. These conditions (of linearity or convexity) were used to show that a certain function obtained as a weak limit by a compactness argument was indeed an admissible "right-hand side", i.e., was $f(\cdot, x(\cdot), u(\cdot))$ for some measurable $u(\cdot)$ with values in Γ.

A similar, and in some respects a worse, situation holds for stochastic optimal control. In the deterministic case an admissible control has only to be measurable and to take values in the right set. As has been noted by Fleming [18, p. 79], in the stochastic case the concept of admissibility is much more complicated. This is because now control can depend with advantage on available information. As a result, difficulties arise in showing that functions obtained as weak limits are indeed of the desired form, e.g., admissible drifts ($g \in \mathcal{A}$) or attainable densities ($\xi \in \mathcal{D}$).

In our setup, the important and difficult problem seems to be that there are drawbacks to using each of the natural ways in (or places at) which to take convex combinations in order to turn weak convergence into strong by the Banach–Saks theorem. One can take them in $\mathcal{A}$, and try to show under convexity of $\mathcal{A}$ that from a minimizing sequence of admissible drifts one can obtain a convergent minimizing sequence by convexifying. Or one can take them in $\mathcal{D}$, provided $\mathcal{D}$ is convex, and then prove that $\mathcal{D}$ is closed. It turns out that $\mathcal{A}$ need not be convex even under Roxin's convexity condition, although special Roxin-type conditions suffice for certain forms of f in (8). And when $\mathcal{A}$ is convex, the property of being a minimizing sequence is not known to transfer from $\{g_n\} \subseteq \mathcal{A}$ to a sequence of convex combination of $\{g_n\}$. Further, we have managed to prove $\mathcal{D}$ convex only in the case of complete information about the past ($G_t = S_t$); counterexamples suggest that $\mathcal{D}$ is rarely convex. (Some of these results and counterresults are the meat of this section.) The trouble then is that it is difficult to show either that convex combinations of minimizing admissible drifts are minimizing, or that convex combinations of minimizing attainable densities are attainable; the second alternative is feasible when $G_t = S_t$, and most of our results concern this case.

We have described the structure of the available information by saying that for each time t there is a σ-algebra G_t on C representing information available at t, and that the control law at t must be measurable on G_t. However, the information in G_t may be very different from (usually it is much smaller than) that on which the system depends at t. In other words for some $u \in \Gamma$, $f(t, \cdot, u)$ may be nowhere near being G_t-measurable. As a result $\mathscr{A}$ may not be convex. In particular, Roxin's condition does not ensure that the set of admissible drifts is convex unless the system depends on no more than the controller knows, a situation atypical in practice.

We give two examples of the failure of Roxin's condition to guarantee convexity of $\mathscr{A}$. These examples are included mostly because they illustrate how the pattern of available information affects convexity, not because convexity of $\mathscr{A}$ is useful. Indeed the properties of the functional $\exp \zeta(\cdot)$ have precluded our finding a role for convexity of $\mathscr{A}$ so far.

Example 1. Let $\Gamma = [0, 1]$, $f(t, y, u) = \exp\{uy(t)\}$ for $y(\cdot) \in C$ scalar, and suppose that G_t is the four element algebra $\{\varnothing, A_t, A_t', C\}$, where A_t is some S_t-measurable set. This situation corresponds to knowing at t only whether $y(\cdot) \in A_t$ or not. We have

$$f(t, y, \Gamma) = [1, e^{y(t)}],$$

a closed convex set, so Roxin's condition is satisfied. Let u_1 and u_2 be two admissible control laws. Clearly $u_1(t, \cdot)$ and $u_2(t, \cdot)$ are constant on A_t. Suppose now that u_0 is an admissible control law such that

$$f(t, y, u_0(t, y)) = \lambda f(t, y, u_1(t, y)) + (1 - \lambda) f(t, u, u_2(t, y)).$$

Then, on A_t, dropping arguments on the u's, we have

$$e^{u_0 y(t)} = \lambda\, e^{u_1 y(t)} + (1 - \lambda)\, e^{u_2 y(t)},$$

$$u_0 = \frac{1}{y(t)} \log\{\lambda\, e^{u_1 y(t)} + (1 - \lambda)\, e^{u_2 y(t)}\}.$$

But u_0 must be constant on A_t, while the right-hand side above obviously depends on what function $y(\cdot)$ from A_t one has. Thus if A_t contains two functions assuming different values at time t, then u_0 cannot be constant on A_t, and so $\mathscr{A}$ cannot be convex.

Example 2. The drift in Example 1 is not at most linear in growth unless $u \equiv 0$. To get one that is, let Γ and G_t be as before, but take

$$f(t, y, u) = (u + y^2(t))^{1/2}.$$

Then $f(t, y, \Gamma) = [|y(t)|, (1 + y^2(t))^{1/2}]$, a convex set, so Roxin's condition is satisfied. Let now u_0, u_1, and u_2 be admissible control laws such that

$$(u_0 + y^2(t))^{1/2} = \lambda(u_1 + y^2(t))^{1/2} + (1 - \lambda)(u_2 + y^2(t))^{1/2},$$

with arguments on the u's omitted. Then

$$u_0 = (\lambda(u_1 + y^2(t))^{1/2} + (1 - \lambda)(u_2 + y^2(t))^{1/2})^2 - y^2(t)$$

must be independent of $y(\cdot)$ for $y(\cdot)$ in A_t. This is impossible unless $y(t)$ is the same number for all $y \in A_t$.

When the system depends on no more than the controller knows, then Roxin's condition implies the convexity of $\mathscr{A}$. Such is the content of the following theorem.

THEOREM 2. *If for each t, u $f(t, \cdot, u)$ is G_t-measurable and if for each t, y $f(t, y, \Gamma)$ is convex, then $\mathscr{A}$ is convex.*

Proof. $f(\cdot, \cdot, u)$ *is G-measurable for each* u, *and* $f(t, y, \cdot)$ is continuous for each t, y. By Roxin's condition, with u_1 and u_2 admissible,

$$\lambda f(t, y, u_1(t, y)) + (1 - \lambda) f(t, y, u_2(t, y)) \in f(t, y, \Gamma).$$

Then the result follows from the implicit function Lemma 5.

We turn now to consider the convexity of the set $\mathscr{D}$ of attainable densities. The principal result is the following theorem.

THEOREM 3. *If $G_t = S_t$, i.e., if the whole past is known, and if $f(t, y, \Gamma)$ is convex for $t, y \in [0, 1] \times C$, then $\mathscr{D}$ is convex.*

Proof. For $e^{\zeta(g_i)} \in \mathscr{D}$, $a_i \geqq 0$, $i = 1, \cdots, n$, and $\sum_{i=1}^{n} a_i = 1$, we are to find an admissible drift g such that

$$e^{\zeta(g)} = \sum_{i=1}^{n} a_i e^{\zeta(g_i)}. \tag{9}$$

Consider the process

$$g(t, \omega) = \frac{\sum_{i=1}^{n} a_i e^{\zeta_0^t(g_i)} g_i(t, \omega)}{\sum_{i=1}^{n} a_i e^{\zeta_0^t(g_i)}}.$$

We note that $g(t, \cdot)$ is $\mathscr{S}_t$-measurable, and that $g(\cdot, \omega)$ is Lebesgue measurable for almost all ω. Thus $g(\cdot, \cdot)$ differs from an $\mathscr{F}$-measurable function at most on a null set. By Roxin's condition, with λ = Lebesgue measure,

$$g(t, \omega) \in f(t, w(\omega), \Gamma) \quad \text{a.e.} \quad [\lambda \times P]. \tag{10}$$

By changing $g(t, \omega)$ on a set of measure zero, we can induce it to satisfy (10) everywhere. Thus by Lemma 5, there is an $\bar{\mathscr{F}}$-measurable function $\gamma : [0, 1] \times \Omega \to \Gamma$ such that

$$g(t, \omega) = f(t, w(\omega), \gamma(t, \omega)),$$

and further [15], there is a G-measurable function $u : [0, 1] \times C \to \Gamma$ such that a.e.

$$u(t, w(\omega)) = \gamma(t, \omega).$$

Hence $g(t, \omega) = f(t, w(\omega), u(t, w(\omega)))$, which shows that g is an admissible drift, $g \in \mathscr{A}$. To see that (9) holds we note that the stochastic differential of the ratio

$$e^{\zeta_0^t(g)} \Big/ \sum_{i=1}^{n} a_i e^{\zeta_0^t(g_i)}$$

is zero.

7. When is $\mathscr{D}$ a subset of L_2? Let $L_2 = L_2(\Omega, P, \mathscr{S}_1)$ be the real Hilbert space of functions measurable on $\mathscr{S}_1 = w^{-1}(S_1)$ and square-integrable with respect to P. When $\mathscr{D} \subseteq L_2$, L_2 provides a particularly convenient topology for $\mathscr{D}$. Unfortunately it turns out that $\mathscr{D}$ is demonstrably a subset of L_2 only when the "right-hand side" function $f(t, y, u)$ in the equation of interest,

$$dx = f(t, x, u(t, x))\, dt + dw,$$

increases with y either slower than linearly or linearly at a slow enough rate. While these conditions cover many cases of interest, they do not cover the general case of at most linear growth. Further, there are choices of f that are linear in y for which $\mathscr{D} \not\subseteq L_2$. In this section we prove some of these facts.

LEMMA 7. *If* $|f(t, y, u)|^2 \leqq \kappa(1 + |y(t)|^{2\alpha})$ *for some* $\alpha < 1$, *then* $\mathscr{D}$ *is a bounded set of* L_2.

Proof. Take $g \in \mathscr{A}$, $T = \int_0^1 |g|^2\, dt$. Then since $2\alpha < 2$,

$$\begin{aligned} T > t &\Rightarrow \int_0^1 |w(t)|^{2\alpha}\, dt > \frac{t}{\kappa} - 1 \\ &\Rightarrow \left(\int_0^1 |w(t)|^2\, dt\right)^{1/2\alpha} > \left(\frac{t}{\kappa} - 1\right)^{1/2\alpha} \\ &\Rightarrow \left(\int_0^1 |w(t)|^2\, dt\right)^{1/2} > \left(\frac{t}{\kappa} - 1\right)^{1/2\alpha}. \end{aligned}$$

Thus for any number $0 < \lambda < \pi^2/8$ ($\pi^2/8$ is the abscissa of convergence of $E \exp \lambda \|w\|^2$) [19],

$$\begin{aligned} P\{T > t\} &\leqq P\left\{\|w\|^2 > \left(\frac{t}{\kappa} - 1\right)^{1/\alpha}\right\} \qquad \left(\|w\|^2 = \int_0^1 |w|^2\, dt\right) \\ &\leqq e^{-\lambda(t/\kappa - 1)^{1/\alpha}}\, E \exp \lambda \|w\|^2. \end{aligned}$$

Lemma 4 now gives

$$\begin{aligned} P\{e^{\zeta(g)} > a\} &\leqq P\left\{\int_0^1 g\, dw > \log a\right\} \\ &\leqq \frac{1}{\log a}\sqrt{\frac{2t}{\pi}}\, e^{-(\log^2 a)/2t} + \text{const.}\, e^{-\lambda(t/\kappa - 1)^{1/\alpha}} \end{aligned}$$

Let $t = \log a/(4 + 2\varepsilon)$, so that

$$P\{e^{\zeta(g)} > a\} \leqq \sqrt{\frac{2 + \varepsilon}{\pi \log a}}\, e^{-(2+\varepsilon)\log a} + \text{const.}\, e^{-(t\lambda/\kappa)(t/\kappa - 1)^{1/\alpha - 1}}.$$

Since $(t/\kappa - 1)^{1/\alpha - 1}$ is eventually larger than κ/λ, the result follows.

To illustrate how $e^{\zeta(g)}$ can fail to belong to L_2 consider the scalar case ($d = 1$) $f(t, y, u) = \kappa y(t) + u$, take $\Gamma = [0, 1]$ together with the control law that

is identically 0, so that $g(t, \omega) = \kappa w(t)$, and

$$\zeta(g) = \kappa \int_0^1 w(t)\, dw(t) - \frac{\kappa^2}{2} \int_0^1 |w(t)|^2\, dt$$

$$= \frac{\kappa}{2} w^2(1) - \frac{\kappa}{2} - \frac{\kappa^2}{2} \int_0^1 |w(t)|^2\, dt.$$

If κ is negative, i.e., if feedback is negative, then $e^{\zeta(g)}$ is actually bounded by $\exp -\frac{1}{2}\kappa$. In general, we can use a result [19] of L. A. Shepp to find for what values of κ $e^{\zeta(g)}$ has what moments. He showed that for $m(\cdot)$ a measure on $[0, 1]$,

$$E \exp\left(-\frac{1}{2} \int_0^1 w^2(t)\, dm(t) \right) < \infty$$

if and only if the integral equation

$$g(t) = 1 - \int_t^1 (t - u) g(u)\, dm(u)$$

has a solution positive in $[0, 1]$; if it does, the value of the expectation is $g(0)^{-1/2}$.

Choosing $dm(u) = \lambda\kappa^2\, du - \lambda\kappa\delta(u - 1)\, du$, we have

$$g(t) = 1 - \lambda\kappa \int_t^1 (t - u) g(u)\, du + \lambda\kappa \int_t^1 (t - u) g(u) \delta(u - 1)\, du$$

$$= 1 - \lambda\kappa^2 \int_t^1 (t - u) g(u)\, du + \lambda\kappa(t - 1) g(1),$$

$$g'(t) = -\lambda\kappa^2 \int_t^1 g(u)\, du + \alpha\kappa g(1), \qquad g'(1) = \lambda\kappa g(1),$$

$$g''(t) = \lambda\kappa^2 g(t), \qquad g(1) = 1.$$

We find

$$2g(t) = (1 + \sqrt{\lambda})\, e^{\kappa\sqrt{\lambda}(t-1)} + (1 - \sqrt{\lambda})\, e^{-\kappa\sqrt{\lambda}(t-1)}, \qquad 0 \leqq t \leqq 1.$$

This has a zero at t if and only if

$$e^{2\kappa\sqrt{\lambda}t} = \frac{\sqrt{\lambda} - 1}{\sqrt{\lambda} + 1}\, e^{2\kappa\sqrt{\lambda}},$$

$$t = 1 + \frac{1}{2\kappa\sqrt{\lambda}} \log\left(\frac{\sqrt{\lambda} - 1}{\sqrt{\lambda} + 1} \right).$$

Thus the function on the right is out of the range $[0, 1]$ if and only if the expectation exists. In particular, if $\lambda = 2$ and

$$\kappa > -\frac{1}{2\sqrt{2}} \log \frac{\sqrt{2} - 1}{\sqrt{2} + 1},$$

then $E\, e^{2\zeta(g)} = +\infty$.

8. Closure and existence. We come finally to closure and existence theorems similar to those of deterministic control theory. When $\mathscr{D}$ is a bounded subset of L_2, these results are proved in a natural way using strong and weak L_2-topologies. When $\mathscr{D}$ is known only to be a subset of L_p, they are proved in a more laborious way by using stopping times $\tau_N \to 1$ as $N \to \infty$, for which

$$\exp \zeta_0^{\tau_N}(g)$$

is in L_2 for each N.

THEOREM 4. *$L_2 \cap \mathscr{D}$ is closed in L_2-norm topology.*

Proof. Let $\exp \zeta(g_n)$ converge strongly to $\xi \in L_2(P)$. By Itô's representation [20] of square-integrable functionals of Brownian motion we can write

$$\xi(\omega) = 1 + \int_0^1 \varphi(s, \omega)\, dw(s),$$

where φ is a nonanticipative functional with

$$E \int_0^1 |\varphi(t, \omega)|^2\, dt < \infty.$$

Let $\xi(t) = E\{\xi | \mathscr{S}_t\}$. Then

$$\begin{aligned}\xi(t) &= 1 + E\left\{\int_0^t \varphi(s, \omega)\, dw(s) | \mathscr{S}_t\right\} + E\left\{\int_t^1 \varphi(s, \omega)\, dw(s) | \mathscr{S}_t\right\} \\ &= 1 + \int_0^t \varphi(s, \omega)\, dw(s)\end{aligned}$$

because the first stochastic integral is $\mathscr{S}_t$-measurable, and the second conditional expectation is zero.

Jensen's inequality gives

$$\begin{aligned}|E\{\xi - e^{\zeta(g_n)} | \mathscr{S}_t\}| &\leqq E\{|\xi - e^{\zeta(g_n)}|\, | \mathscr{S}_t\}, \\ E|E\{\xi - e^{\zeta(g_n)} | \mathscr{S}_t\}| &\leqq E|\xi - e^{\zeta(g_n)}|.\end{aligned}$$

The right-hand side goes to zero, by Schwarz's inequality. This shows that for each t,

$$E|e^{\zeta_0^t(g_n)} - \xi(t)| \leqq E|\xi - e^{\zeta(g_n)}|^2 = o(1).$$

Hence $e^{\zeta_0^{\cdot}(g_n)} \to \xi(\cdot)$ in $L_1(\lambda \times P)$.

Now since

$$E\left|\int_0^1 e^{\zeta_0^s(g_n)} g_n(s)\, dw(s) - \int_0^1 \varphi(s)\, dw(s)\right|^2 = E\int_0^1 \left| e^{\zeta_0^s(g_n)} g_n(s) - \varphi(s) \right|^2 ds,$$

there exists a subsequence, assumed relabeled, such that

$$\begin{aligned}e^{\zeta_0^{\cdot}(g_n)} &\to \xi(\cdot), \\ e^{\zeta_0^{\cdot}(g_n)} g_n(\cdot) &\to \varphi(\cdot)\end{aligned}$$

a.s. $[\lambda \times P]$ (λ = Lebesgue measure).

Noting that

$$\xi(t)g_n(t) - \varphi(t) = [\xi(t) - e^{\zeta_0^t(g_n)}]g_n(t) + e^{\zeta_0^t g_n)}g_n(t) - \varphi(t)$$

and that $|g_n(t)|^2 \leqq \text{const.}\,[1 + |w(t,\omega)|^2]$, we can conclude from Lemma 2 that g_n converges almost surely to a function g. By Lemma 6, g is an admissible control, and

$$\xi(t) = 1 + \int_0^t \xi(s)g(s)\,dw(s).$$

Hence [14], $\xi(t) = \exp \zeta_0^t(g)$ with $g \in \mathscr{A}$. Thus $\xi \in \mathscr{D}$, and $\mathscr{D}$ is closed in norm topology.

Remark. If $\mathscr{D}$ is a bounded and weakly sequentially closed subset of L_2, then an optimal control law exists.

Proof. This is almost obvious. Let g_n be a minimizing sequence of admissible drifts. $\mathscr{D}$ is weakly sequentially compact, being bounded. Thus we may suppose that $\exp \zeta(g_n)$ converges weakly in L_2 to an L_2-function $\xi \in \mathscr{D}$. Hence there is an admissible drift g such that $\xi = \exp \zeta(g)$. The cost of using drift g_n is

$$Ew_0(1)\, e^{\zeta(g_n)} \to Ew_0(1)\, e^{\zeta(g)}.$$

Since g_n is minimizing, g is an optimal admissible drift.

Our basic L_2 existence result is the following theorem.

THEOREM 5. *If $G_t = S_t$, if $f(t, y, \Gamma)$ is convex, and if $\mathscr{D}$ is L_2-bounded, then an optimal control law exists.*

Proof. The first two hypotheses imply, by Theorem 3, that $\mathscr{D}$ is convex. Theorem 4 implies that $\mathscr{D}$ is strongly closed. Hence it is weakly closed, and existence of an optimal admissible drift follows from the preceding remark.

LEMMA 8. *If $\tau_N = \min\{1, \inf t : |w(t)| = N\}$, then*

$$\sup_{g \in \mathscr{A}} E \exp 2\zeta_0^{\tau_N}(g) < \infty.$$

Proof. If $\chi_N = \chi_{\{\sup_{0 \leqq s \leqq t} |w(s)| \leqq N\}}$, we have $\zeta_0^{\tau_N}(g) = \zeta_0^1(\chi_N g)$, and with $T(t) = \int_0^t |g|^2\,dt$, $T = T(1)$, $W(t) = \int_0^{T^{-1}(t)} g\,dw$, $T^{-1}(t) = \inf u : T(u) = t$, the argument in Lemma 4 gives

$$\chi_{\zeta(\chi_N g) > \log a} \leqq \chi_{\{T \leqq t,\ \sup_{0 \leqq s \leqq t} |W(s)| > \log a\}} + \chi_{\{T > t\}}.$$

Take $t = \kappa\{1 + N^2\}$ to get

$$\begin{aligned}
\chi_{\zeta(\chi_N g) > \log a} &\leqq \chi\{\sup_{0 \leqq s \leqq \kappa(1+N^2)} |W(s)| > \log a\}, \\
P\{e^{\zeta(\chi_N g)} > a\} &= P\{\zeta(\chi_N g) > \log a\} \\
&\leqq P\{\sup_{0 \leqq s \leqq \kappa(1+N^2)} |W(s)| > \log a\} \\
&\leqq \frac{1}{\log a}\sqrt{\frac{2}{\pi}}\sqrt{\kappa(1+N^2)} \exp - \frac{\log^2 a}{2\kappa(1+N^2)}
\end{aligned}$$

$$\leqq \frac{\text{const.}}{\log a}\, a^{-(\log a)/2\kappa(1+N^2)}$$

$$\leqq \text{const.}\, a^{-2-\varepsilon} \quad \text{for} \quad \frac{\log a}{2\kappa(1+N^2)} > 2+\varepsilon.$$

But

$$E\exp(2\zeta_0^{\tau^N}(g)) = \int_0^\infty P\{\exp\zeta(\chi_N g) > a\}a\,da$$

$$< \text{const. depending on } N \text{ but not on } g.$$

LEMMA 9. *Let ξ be a random variable measurable on $\mathscr{S}_1$, with $E|\xi|^p < \infty$ for some $p > 1$, and such that*

$$\xi(t) = E\{\xi|\mathscr{S}_t\}$$

is a continuous martingale. Let τ be a stopping time and let $\mathscr{S}_\tau$ be the σ-algebra of all $\mathscr{S}_1$ sets A such that $A \cap \{\tau \leqq t\} \in \mathscr{S}_t$. Then a.s.

$$\xi(\tau) = E\{\xi|\mathscr{S}_\tau\}.$$

Proof. We shall show that $\xi(\tau)$ is $\mathscr{S}_\tau$-measurable and that $B \in \mathscr{S}_\tau$ implies

$$\int_B \xi(\tau)\,dP = \int_B \xi\,dP.$$

To show that $\xi(\tau)$ is $\mathscr{S}_\tau$-measurable it is sufficient to show that every set of the form $\{\omega : \xi(\tau) \in A, \tau \leqq t\}$, A closed, belongs to $\mathscr{S}_t$. Let π_{mt} be a countable cover of $[0, t]$ by open sets of diameter $\leqq (\frac{1}{2})^m$. Then to within a set of measure zero [15],

$$\{\omega : \xi(\tau) \in A, \tau \leqq t\} = \bigcap_m \bigcup_{S \in \pi_{mt}} \{\tau \in S \text{ and } \xi^{-1}(A) \cap S \neq \varnothing\}.$$

Also, with D a countable set dense in S,

$$\{\omega : \xi^{-1}(A) \cap S = \varnothing\} = \bigcup_{s \in S} \{\omega : \xi(\omega, s) \in A\}$$

$$= \bigcup_{s \in D} \{\omega : \xi(\omega, s) \in A\}.$$

Thus $\xi(\tau)$ is $\mathscr{S}_\tau$-measurable. Moreover, with $B \in \mathscr{S}_\tau$, and $0 = t_0 < t_1 < \cdots < t_n = 1$,

$$\int_B \xi\,dP = \sum_{i=1}^n \int_{B \cap \{t_{i-1} < \tau \leqq t_i\}} \xi\,dP$$

$$= \sum_{i=1}^n \int_{B \cap \{t_{i-1} < \tau \leqq t_i\}} \xi(t_i)\,dP$$

$$= \int_B \xi(\tau)dP + \int_B \sum_{i=1}^n \chi_{\{t_{i-1} < \tau \leqq t_i\}}[\xi(t_i) - \xi(\tau)]\,dP.$$

With $\eta(\omega) = \xi(t_{i+1})$ on $\{t_i < \tau \leqq t_{i+1}\}$, we have

$$\int_B |\eta - \xi(\tau)|\,dP \leqq \left(\int_B |\eta - \xi(\tau)|^p\,dP\right)^{1/p} P^{p/(1-p)}(B).$$

Clearly $|\eta - \xi(\tau)|^p \leqq 2 \sup_{0 \leqq s \leqq 1} |\xi(s)|^p$, and by the martingale property,

$$E \sup_{0 \leqq s \leqq 1} |\xi(s)|^p \leqq \left(\frac{p}{p-1}\right)^p E|\xi(1)|^p.$$

Thus the integrand $|\eta - \xi(\tau)|^p$ approaches zero a.s. as $\max_{0<i\leqq n}|t_i - t_{i-1}| \to 0$ and is dominated by an integrable function; hence the second integral above goes to zero.

LEMMA 10. *If ξ is an integrable Brownian functional measurable on $\mathscr{S}_1$, with $\xi > 0$ a.s., then there is a functional ψ such that*

$$\xi(t) = E\{\xi|\mathscr{S}_t\} = e^{\zeta_0^t(\psi)}E\xi,$$

where $\int_0^1 |\psi|^2\, dt < \infty$ a.s. and $\psi(\cdot, \cdot)$ is nonanticipative.

Proof. It is easy to see that $\xi(t) > 0$ a.s. for each t. Now J. M. C. Clark has shown [21] that if ξ is an integrable $\mathscr{S}_1$-measurable function, then there is a nonanticipative function $\varphi(\cdot, \cdot)$ such that for $s < t$,

$$E\{\xi|\mathscr{S}_t\} - E\{\xi|\mathscr{S}_s\} = \int_s^t \varphi(u, \omega)\, dw(u),$$

$$P\left\{\int_0^1 |\varphi|^2\, du < \infty\right\} = 1,$$

$\mathscr{S}_0$ is trivial, so $\xi(0) = E\xi$. Let now

$$\psi(t, \omega) = \frac{\varphi(t, \omega)}{E\xi + \int_0^t \varphi\, dw} = \frac{\varphi(t, \omega)}{\xi(t, \omega)}.$$

For fixed ω, $\psi(\cdot, \omega)$ is a Lebesgue function of t; for fixed t, it is $\mathscr{S}_t$-measurable; so it is nonanticipative. Evidently

$$\xi(t) = E\xi + \int_0^t \xi(s)\psi(s)dw(s).$$

McKean [14] has shown that the only solution to this stochastic equation is

$$\xi(t) = \exp \zeta_0^t(\psi)E\xi.$$

It remains to prove that $\int_0^1 |\psi|^2\, dt < \infty$ a.s. This will follow from

$$P\left\{\inf_{0 \leqq t \leqq 1} \xi(t) \leqq 0\right\} = 0.$$

Since $\xi(\cdot)$ is a martingale, Jensen's inequality gives

$$E\{\log \xi(t)|\mathscr{S}_s\} \leqq \log E\{\xi(t)|\mathscr{S}_s\} = \log \xi(s),$$

so that $-\log \xi(\cdot)$ is a submartingale. Hence

$$P\{\inf_{0\leqq t\leqq 1} \xi(t) \geqq e^{-n}\} = P\{\sup_{0\leqq t\leqq 1} [-\log \xi(t)] \leqq n\}$$

$$\geqq 1 - n^{-1} \int_{\{\sup_{0\leqq t\leqq 1} [-\log \xi(t)]\geqq n\}} \xi \, dP,$$

$$P\{\inf_{0\leqq t\leqq 1} \xi(t) < e^{-n}\} \leqq n^{-1} E\xi.$$

Now

$$\{\omega: \inf_{0\leqq t\leqq 1} \xi(t) \leqq 0\} \subseteq \bigcap_{n\geqq 1} \{\omega: \inf_{0\leqq t\leqq 1} \xi(t) < e^{-n}\}$$

and the probability of the intersection on the right is zero.

THEOREM 6. *If $G_t = S_t$, and if $f(t, y, \Gamma)$ is convex for $t, y \in [0, 1] \times C$* (Roxin's condition), *then $\mathscr{D}$ is weakly sequentially closed in L_1.*

Proof. Let $e^{\zeta(g_n)} \in \mathscr{D}$ approach $\xi \in L_1$ weakly. By Lemma 10, the martingale $\xi(t) = E\{\xi|\mathscr{S}_t\}$ is representable as $\exp \zeta_0^t(\psi)$ with $\psi(t, \omega)$ nonanticipating, and square-integrable in t a.s. Thus $\xi(\cdot)$ has continuous sample paths almost surely. Since (Lemma 1) $\mathscr{D}$ is a bounded set in L_p for some $p > 1$ we can show that ξ belongs to the same L_p. The processes $\exp \zeta_0^t(g_n) = E\{e^{\zeta(g_n)}|\mathscr{S}_t\}$, $0 \leqq t \leqq 1$, are also continuous L_p-martingales. Introduce the stopping times

$$\tau_N = \min\{1, \inf t : |w(t)| = N\}.$$

Lemma 9 implies that a.s.

$$\xi(\tau_N) = E\{\xi|\mathscr{S}_{\tau_N}\},$$

$$\exp \zeta_0^{\tau_N}(g_n) = E\{e^{\zeta(g_n)}|\mathscr{S}_{\tau_N}\}.$$

Since $E\{\cdot|\mathscr{S}_{\tau_N}\}$ is "self-adjoint" we have for $h \in L_\infty$,

$$EhE\{e^{\zeta(g_n)}|\mathscr{S}_{\tau_N}\} = Ee^{\zeta(g_n)} E\{h|\mathscr{S}_{\tau_N}\},$$

$$EhE\{\xi|\mathscr{S}_{\tau_N}\} = E\xi E\{h|\mathscr{S}_{\tau_N}\}.$$

It follows that for each N, $\exp \zeta_0^{\tau_N}(g_n)$ approaches $\xi(\tau_N)$ in weak L_1. By Lemma 8, for each N, the sequence $\exp \zeta_0^{\tau_N}(g_n)$ is bounded in L_2. Hence the convergence above holds in weak L_2-topology, and $\xi(\tau_N) \in L_2$ for each N.

Let $\chi_N = \chi_N(t, \omega) = \chi_{\{\tau_N > t\}}$, so that $\xi(\tau_N) = \exp \zeta(\chi_N \psi)$ and $\zeta_0^{\tau_N}(g_n) = \zeta_0^1(\chi_N g_n)$. We have shown that for each N,

$$\exp \zeta(\chi_N g_n) \to \exp \zeta(\chi_N \psi) \quad \text{in weak } L_2.$$

Let N be fixed, and choose a subsequence of $\{g_n\}$, assumed relabeled, so that

$$\frac{1}{n} \sum_{i=1}^{n} \exp \zeta(\chi_N g_i) \to \exp \zeta(\chi_N \psi) \quad \text{in strong } L_2.$$

As in the convexity result Theorem 3, the function

$$g_n^*(t, \omega) = \frac{\sum_{i=1}^{n} e^{\zeta_0^t(g_i)} g_i(t)}{\sum_{i=1}^{n} e^{\zeta_0^t(g_i)}}$$

is an admissible drift such that

$$\frac{1}{n}\sum_{i=1}^{n}\exp\zeta(\chi_N g_i) = \exp\zeta(\chi_N g_n^*).$$

The strong L_2-convergence implies, as in Theorem 4, that $\chi_N\psi \in \chi_N f(t, w(\omega), \Gamma)$ a.s. It follows that

$$\psi(t,\omega)\in f(t, w(\omega), \Gamma) \quad \text{a.s.}$$

Thus by Lemma 5 and the elementary result in § 3, ψ is an admissible drift. This proves Theorem 6.

THEOREM 7. *If $\mathscr{D}$ is weakly sequentially closed in L_1, then an optimal control law exists.*

Proof. By Lemma 1, $\mathscr{D}$ is uniformly integrable. The inequality

$$\int_A e^{\zeta(g)}\,dP \leqq \int_{\zeta(g)>\log N} e^{\zeta(g)}\,dP + NP(A)$$

then shows that $\lim\int_A e^{\zeta(g)}\,dP = 0$ as $P(A)\to 0$, uniformly in $g\in\mathscr{A}$. $\mathscr{D}$ is clearly bounded, and so it is weakly sequentially compact in L_1. Thus we may assume as given a minimizing sequence $g_n, n \geqq 1$, of admissible drifts such that $e^{\zeta(g_n)}$ converges weakly in L_1 to a member $e^{\zeta(g)}$ of $\mathscr{D}$. The cost of using g_n is

$$Ex_0(1)e^{\zeta(g_n)} \to \inf_{u\in\mathscr{U}} k(u).$$

We next note that for some $p > 1, q = p/(p-1)$, and

$$\chi_N = \chi_{\{\sup_{0\leqq s\leqq 1}|w(s)|^2 > N\}},$$

we have for $h\in\mathscr{A}$,

$$|Ew_0(1)\chi_N e^{\zeta(h)}| \leqq E^{1/q}\chi_N|w_0(1)|^q \cdot E^{1/p}\,e^{p\zeta(h)}.$$

Lemma 1 shows that the second factor is bounded uniformly in $h\in\mathscr{A}$; the first factor vanishes as N increases. Let $\varepsilon > 0$ be given, and choose N so large that

$$|Ew_0(1)\chi_N e^{\zeta(h)}| < \frac{\varepsilon}{3}$$

for all $h\in\mathscr{A}$. Then by the weak L_1-convergence choose n_0 so large that $n > n_0$ implies

$$|Ew_0(1)(1-\chi_N)(e^{\zeta(g)} - e^{\zeta(g_n)})| < \frac{\varepsilon}{3}.$$

It will follow that $|k(g_n) - k(g)| < \varepsilon$ for $n > n_0$. Since ε was arbitrary, g is an optimal admissible drift.

Our final result removes from Theorem 5 the hypothesis that $\mathscr{D}$ be included in a ball of L_2.

THEOREM 8. *If $G_t = S_t$, if $f(t, y, \Gamma)$ is convex, and if $|f(t, y, u)|^2 \leqq$ const. $(1 + |y(t)|^2)$, then an optimal control law exists.*

Proof. Theorem 6 implies that $\mathscr{D}$ is weakly sequentially closed in L. The result follows from Theorem 7.

Acknowledgment. The author is indebted to S. R. S. Varadhan for calling his attention to Girsanov's work and for suggesting a martingale proof of Girsanov's theorem; to H. S. Witsenhausen for two years of suggesting, encouraging, and criticizing; to H. P. McKean, Jr. for the random time change arguments in Lemmas 3 and 4; to T. T. Kadota and L. A. Shepp for helpful discussions; to T. E. Duncan and P. Varaiya for clarifying the ideas underlying Lemma 0 by sending him a preprint [23] of work based on an earlier version of this paper.

REFERENCES

[1] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theor. Probability Appl., 5 (1960), pp. 285–301.

[2] W. H. FLEMING AND M. NISIO, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777–794.

[3] A. V. SKOROKHOD, *Studies in the Theory of Random Processes*, Kiev, 1961; English translation: Addison-Wesley, Reading, Massachusetts, 1965.

[4] R. H. CAMERON AND W. T. MARTIN, *Transformations of Wiener integrals under translations*, Ann. of Math., 45 (1944), pp. 386–96.

[5] E. J. MCSHANE AND R. B. WARFIELD, *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.

[6] H. J. KUSHNER, *On the existence of optimal stochastic controls*, this Journal, 3 (1965), pp. 463–474.

[7] W. H. FLEMING, *Optimal control of partially observable diffusions*, this Journal, 6 (1968), pp. 194–214.

[8] A. D. VENTCEL, *Additive functionals of multidimensional Wiener processes*, Dokl. Akad. Nauk SSSR, 139 (1961), pp. 13–16.

[9] R. E. MORTENSEN, *Optimal control of continuous-time stochastic systems*, Rep. ERL66-1, Electronic Research Laboratory, University of California, Berkeley, 1966.

[10] T. KAILATH, *A general likelihood-ratio formula for random signals in gaussian noise*, IEEE Trans. Information Theory, IT-15 (1969), pp. 350–361.

[11] T. E. DUNCAN, *Probability densities for diffusion processes with applications to nonlinear filtering theory and detection theory*, Rep. 7001–4 and 7050–12, Center for Systems Research, Stanford University, Stanford, Calif., 1967.

[12] G. KALLIANPUR AND C. STRIEBEL, *Estimation of stochastic systems; arbitrary system process with additive white noise observations error*, Ann. Math. Statist., 39 (1968), pp. 785–801.

[13] D. STROOCK AND S. R. S. VARADHAN, *Diffusion processes with continuous coefficients*, Comm. Pure Appl. Math., 22 (1969), pp. 345–400 and 479–530.

[14] H. P. MCKEAN, JR., *Stochastic Integrals*, Academic Press, New York and London, 1969.

[15] V. E. BENEŠ, *Existence of optimal strategies based on specified information, for a class of stochastic additive white noise observations error*, Ann. Math. Statist., 39 (1968), pp. 785–801.

[16] E. B. LEE AND L. MARKUS, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.

[17] E. ROXIN, *The existence of optimal controls*, Michigan Math., J., 9 (1962), pp. 109–119.

[18] W. H. FLEMING, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–509.

[19] L. A. SHEPP, *Radon–Nikodym derivatives of Gaussian measures*, Ann. Math. Statist. 37 (1966), pp. 321–354.

[20] K. ITÔ, *Multiple Wiener integral*, J. Math. Soc. Japan, 3 (1951), pp. 158–161.

[21] J. M. C. CLARK, *The representation of functionals of Brownian motion by stochastic integrals*, Ann. Math. Statist., 41 (1970), pp. 1282–1295.

[22] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[23] T. E. DUNCAN AND P. VARAIYA, *On the solutions of a stochastic control system*, to appear.

SIAM J. CONTROL
Vol. 9, No. 3, August 1971

STOCHASTIC CONTROL FOR SMALL NOISE INTENSITIES*

WENDELL H. FLEMING†

Abstract. This paper is concerned with the approximate solution of stochastic optimal control problems which arise by perturbing the system equations in the deterministic Pontryagin control model, through an additive white noise term with small coefficient. The system states are assumed completely observable. Mathematically the problem becomes one of singular perturbation of the Hamilton–Jacobi equation by a small second order term. Our main results concern expansions of solutions of the perturbed equation in powers $\varepsilon, \varepsilon^2, \varepsilon^3, \cdots$ of the noise variance coefficient ε. The results obtained hold in regions where the corresponding solution of the Hamilton–Jacobi equation is sufficiently well-behaved.

1. Introduction. This paper is concerned with the approximate solution of some stochastic optimal control problems. The models of control systems which we consider are perturbations of the deterministic optimal control model of Pontryagin [22]. We suppose that, in the stochastic problem, the system is perturbed by random disturbances which take the form of an additive white noise term in the system equations (1.1). If the noise coefficient σ in (1.1) is small, it is reasonable to seek an approximate solution of the stochastic problem in terms of quantities computable from the Pontryagin problem (with $\sigma = 0$). Some rather weak statements of this kind were proved in [7, pp. 269, 276]. Stronger results were obtained in [9, p. 527] for the randomly perturbed simplest problem in calculus of variations. In the present paper still sharper statements are proved; see the theorems in §§ 6 and 7.

Mathematically, the problem becomes one of singular perturbation of the Hamilton–Jacobi equation (1.8^0) by a small second order term (equation (1.8^ε)). Our main results concern the validity of the approximate formulas (1.11)–(1.12), in regions where the solution φ^0 of (1.8^0) is sufficiently well-behaved. Such regions are called *regions of strong regularity* (§ 3). A rather general first order equation $\varphi_s + F(s, x, \varphi_x) = 0$ can be regarded as the Hamilton–Jacobi equation for some control problem (indeed, for the simplest problem in calculus of variations). Hence our results apply to such equations. See Theorem 7.2.

We consider the same kind of stochastic optimal control model as in [7]. Let

$$\xi(t) = (\xi_1(t), \cdots, \xi_n(t))$$

denote the state of the system at time t. Thus $\xi(t)$ is a vector in n-dimensional space R^n. The state process ξ evolves according to stochastic differential equations, written in vector-matrix notation as

$$d\xi = f(t, \xi(t), u(t))\, dt + \sigma\, dw, \tag{1.1}$$

where w is an n-dimensional Brownian motion with $w(s) = 0$. At the initial time s, an initial state vector

$$x = \xi(s) \tag{1.2}$$

* Received by the editors May 26, 1970.

† Department of Mathematics, Brown University, Providence, Rhode Island 02912. This work was supported in part by the Air Force Office of Scientific Research under Grant AF-AFOSR 67-0693A and in part by the National Science Foundation under Grant GP6733.

is known. We require that $(s, x) \in Q$, where Q is a given open subset of R^{n+1}. The control $u(t)$ applied in (1.1) at time t is a vector subject to

$$u(t) \in K, \tag{1.3}$$

where the "control set" K is a closed convex subset of some R^k. The control u in (1.3), as well as ξ and w, is generally a stochastic process defined on some probability space Ω. Following custom, we simplify the notation by writing $u(t)$, $\xi(t)$, $w(t)$ instead of $u(t, \omega)$, $\xi(t, \omega)$, $w(t, \omega)$, $\omega \in \Omega$.

Let τ denote the first time $t \geqq s$ when $(t, \xi(t)) \notin Q$; τ is called the *exit time* from Q, and is sometimes denoted by τ_Q to indicate its dependence on Q. The criterion of performance of the system is

$$J = E \int_s^\tau L(t, \xi(t), u(t))\, dt, \tag{1.4}$$

where E denotes expected value.

If $\sigma = 0$ and u is a (deterministic) function of t, then (1.1) is a vector ordinary differential equation. Instead of (1.4) we have the performance criterion

$$J^0 = \int_s^\tau L(t, \xi(t), u(t))\, dt. \tag{1.4^0}$$

The problem of finding a control function u^0 such that J^0 is minimum is precisely that of Pontryagin. There is an extensive literature; see for instance [13], [19], [22].

When $\sigma \neq 0$ we must specify the information available to the controller. Let us assume that the state $\xi(t)$ is completely observable at each time t. This assumption is essential if the stochastic control problem is to be reduced to one in partial differential equations via dynamic programming. It is interesting to try to replace it by other observability assumptions. Wonham [26] proved a separation principle reducing certain partially observable problems to completely observable ones. Kushner [16] considered the case of random disturbances entering the system as impulses at discrete time instants, rather than as white noise as in (1.1). See also remarks about the open loop problem at the end of § 6.

Since the states are completely observable the controller may take

$$u(t) = Y(t, \xi(t)), \tag{1.5}$$

where Y is a *control policy* belonging to a certain class $\mathcal{Y}$ (§ 4). Sometimes Y is called a *closed loop*, or *Markov*, control policy, also a program [13]. Given a control policy Y and initial data (s, x), the states $\xi(t)$ evolve according to a Markov process. The policy Y controls the local drift coefficient f^Y, defined by

$$f^Y(s, x) = f(s, x, Y(s, x));$$

and (1.1) can be rewritten in the customary form [5], [12, Chap. 8]:

$$d\xi = f^Y\, dt + \sigma\, dw. \tag{1.1$'$}$$

The stochastic optimization problem is to find among all $Y \in \mathcal{Y}$ one for which J minimizes (1.4).

For simplicity let the noise coefficient σ be a diagonal matrix, with diagonal elements $(2\varepsilon)^{1/2}$. The components of $\sigma w(t)$ are then independent and have mean 0, variance $2\varepsilon(t - s)$. In (1.4), let us write $J = J^{\varepsilon}(Y; s, x)$, and

$$\varphi^{\varepsilon}(s, x) = \inf_{Y \in \mathcal{Y}} J^{\varepsilon}(Y; s, x). \tag{1.6$^{\varepsilon}$}$$

For each (s, x) and $p \in R^n$ let

$$H(s, x, p) = \min_{y \in K} [L(s, x, y) + pf(s, x, y)]. \tag{1.7}$$

Let ψ_s, ψ_{x_i} denote the partial derivatives of a function $\psi(s, x)$. Let

$$\psi_x = (\psi_{x_1}, \cdots, \psi_{x_n}), \qquad \Delta_x \psi = \sum_{i=1}^{n} \psi_{x_i x_i}$$

denote the gradient vector and Laplacean in the variables x. If Q belongs to a class $\mathcal{S}$ of regions suitable from the viewpoint of parabolic equations (§ 4), if K is compact, and if we impose suitable assumptions on f, L (§ 2), then φ^{ε} satisfies in Q the parabolic partial differential equation

$$\varphi^{\varepsilon}_s + \varepsilon \Delta_x \varphi^{\varepsilon} + H(s, x, \varphi^{\varepsilon}_x) = 0 \tag{1.8$^{\varepsilon}$}$$

together with the data

$$\varphi^{\varepsilon}(s, x) = 0 \quad \text{for } (s, x) \in \partial^* Q. \tag{1.9}$$

Here $\partial^* Q$ is the "essential" portion of the boundary ∂Q (§ 4). The optimal policy Y^{ε} has the property that

$$L(s, x, y) + \varphi^{\varepsilon}_x(s, x) f(s, x, y) = \min \quad \text{on } K \tag{1.10}$$

when $y = Y^{\varepsilon}(s, x)$.

Thus the completely observable optimal control problem is in principle reduced to solving the boundary problem (1.8$^{\varepsilon}$)–(1.9) for φ^{ε}, and then minimizing $L + \varphi^{\varepsilon}_x f$ over K for each $(s, x) \in Q$. This is usually difficult to do in practice, although some approximate methods have been proposed. See for instance [15], [25]. In the present paper we seek approximate formulas for φ^{ε} and φ^{ε}_x for small positive ε.

When $\varepsilon = 0$ we define φ^0 by

$$\varphi^0(s, x) = \inf_{u \in \mathcal{U}_s} J^0(u; s, x), \tag{1.6^0}$$

where J^0 is as in (1.4^0) and $\mathcal{U}_s$ is an appropriate class of (open loop) controls (§ 3). It can be shown that also

$$\varphi^0(s, x) = \inf_{Y \in \mathcal{Y}} J^0(Y; s, x).$$

Formally, φ^0 satisfies in Q the Hamilton–Jacobi equation

$$\varphi^0_s + H(s, x, \varphi^0_x) = 0 \tag{1.8^0}$$

with $\varphi^0 = 0$ on $\partial^* Q$; an optimal policy Y^0 is found formally by minimizing $L + \varphi^0_x f$ over K.

We wish to find $\varphi^{\varepsilon}, \varphi^{\varepsilon}_x$ (and hence also Y^{ε}) approximately in terms of quantities computable from φ^0. Two apparent difficulties are immediately seen. One is

mathematical, the other practical. First, the function φ^0 is not differentiable everywhere; one must usually interpret (1.8^0) in some generalized sense [7]. The policy Y^0 need not be continuous, and one does not know that the system equations $d\xi = f^{Y^0}\,dt$ can be solved for the optimal deterministic trajectories. The apparent practical difficulty is that the boundary problem (1.8^0)–(1.9) does not seem easier to solve numerically than (1.8^ε)–(1.9).

However, let us attempt to solve the stochastic problem, not in all of Q, but merely near some particular optimal trajectory γ^0. We suppose that γ^0 lies in some region N of strong regularity (§ 3). In N, φ^0 is rather smooth and satisfies (1.8^0) in the classical sense. We seek in N expansions of the type

$$\varphi^\varepsilon = \varphi^0 + \varepsilon\theta_1 + \varepsilon^2\theta_2 + \cdots + \varepsilon^l\theta_l + o(\varepsilon^l), \tag{1.11}$$

$$\varphi_x^\varepsilon = \varphi_x^0 + \varepsilon(\theta_1)_x + \varepsilon^2(\theta_2)_x + \cdots + \varepsilon^l(\theta_l)_x + o(\varepsilon^l). \tag{1.12}$$

The coefficients $\theta_1, \cdots, \theta_l$ satisfy the equations found by differentiating (1.8^ε) repeatedly with respect to ε and setting $\varepsilon = 0$. These equations involve the partial derivatives of H of corresponding orders and of φ^0. The validity of such expansions will depend on smoothness properties of H, discussed in § 2. For certain problems, including the linear time optimal problem with polyhedral K, we know only (1.11) with $l = 1$; in such problems, H is not even of class C^1. In other problems where a strong convexity condition holds we may take $l = 2$ in (1.11) and $l = 1$ in (1.12). See § 6. In § 7 we discuss problems without control constraints (thus of the type in calculus of variations). For such problems one can take l arbitrarily large, but finite. One cannot let $l \to \infty$ and obtain φ^ε as a convergent power series in ε; φ^ε is at best a C^∞, not real analytic, function of ε for $\varepsilon \geqq 0$. It turns out that the optimal trajectory γ^0 is a characteristic curve for the equations which θ_j, $(\theta_j)_x$ satisfy. Therefore the values $\theta_j(s, x)$, $\theta_{jx}(s, x)$ at the left endpoint (s, x) of γ^0 are integrals along γ^0 of expressions involving partial derivatives of φ^0 and H (§ 6). These partial derivatives can be calculated from the method of characteristics in a neighborhood of γ^0. See the Appendix. When there are no control constraints (§ 7) one needs for this purpose a solution to the secondary minimum problem in calculus of variations. For γ^0 to lie in a region of strong regularity it suffices in this case that:

(i) the initial point (s, x) of γ^0 be regular (meaning that γ^0 is the unique optimal deterministic trajectory with initial point (s, x)); and

(ii) the classical Jacobi condition holds (meaning that (s, x) is not a conjugate point).

One might expect that the optimal stochastic trajectories, with the same initial point (s, x) as γ^0, remain in N with probability very nearly 1. A sharp estimate of this kind is Corollary 5.7.

Our method also answers the following question. Suppose that the controller does not solve the stochastic optimization problem, but instead uses the optimal deterministic policy Y^0. (More precisely, suppose that he uses a policy $Y \in \mathscr{Y}$ agreeing with Y^0 in N.) How close to the optimum is the performance $J^\varepsilon(Y^0; s, x)$ in the stochastic problem? See the end of § 6.

The program of the paper is as follows. We begin with some properties of H (§ 2) and a discussion of the Pontryagin problem (§ 3). In §§ 4, 5 we obtain some

preliminary estimates for the convergence of φ^ε and φ_x^ε as $\varepsilon \to 0$, which are weaker than those eventually proved in §§ 6, 7. A crucial step is to establish convergence of φ_x^ε to φ_x^0 uniformly on compact subsets of N (Lemma 5.5).

In §§ 4–7 we impose a strong convexity condition (2.2′). This insures that the optimal policies Y^ε for $\varepsilon > 0$ belong to $\mathscr{Y}$. The standard Ito conditions then guarantee that the system equations (1.1) have a solution when $Y = Y^\varepsilon$. In § 8 we give some partial results when (2.2′) does not hold. The optimal Y^ε may then be discontinuous; however, the existence of solutions to (1.1) for $\varepsilon > 0$ is still guaranteed in the sense of [24]. An interesting open question is to prove precise formulas for the perturbation of switching surfaces for Y^ε, for example, in the linear time-optimal problem. Formal studies of such perturbation problems were made in [4] and [23].

In § 9 the autonomous problem is mentioned. Equation (1.8$^\varepsilon$) is then of elliptic type. In § 10 some examples are given, including one studied by E. Hopf [14] in which (1.8$^\varepsilon$) is equivalent to Burgers' equation. These examples show that our results are in many respects best possible.

2. Assumptions on f, L, K; Properties of H. If Γ is an open set, we write $g \in C^l(\Gamma)$ to mean that the function g together with its partial derivatives of orders $j = 1, \cdots, l$ are continuous on Γ. If Γ is not open, then $g \in C^l(\Gamma)$ means that g agrees on Γ with a function $g' \in C^l(\Gamma')$, where Γ' is open and $\Gamma \subset \Gamma'$. If Γ is clear from the context, we sometimes write merely $g \in C^l$.

Throughout, we assume that $f, L \in C^\infty(R^{n+k+1})$. However, we shall use their values only on $[T_0, T] \times R^n \times K$, where $[T_0, T]$ is a fixed (finite) time interval and the control set $K \subset R^k$ is closed and convex. (In § 9 we consider the autonomous case, with f, L defined on R^{n+k}.)

The following additional assumptions are made.

(2.1) f is linear in the control variables $y = (y_1, \cdots, y_k)$, namely, $f(s, x, y) = A(s, x) + B(s, x)y$. The functions A, B (respectively vector- and $n \times k$ matrix-valued) are bounded together with their first order partial derivatives.

$L(s, x, y)$ is convex in y, namely,

$$vL_{yy}v \geqq 0 \quad \text{for all vectors } v \in R^k. \tag{2.2}$$

(2.3) $L \geqq c_1 > 0$. Moreover, L and its gradient L_x in the variables $x = (x_1, \cdots, x_n)$ are bounded on $[T_0, T] \times R^n \times K_1$ for any compact $K_1 \subset K$.

In (2.2), as throughout the paper, we use vector-matrix notation. Thus L_{yy} denotes the matrix of second order partial derivatives, and

$$vL_{yy}v = \sum_{i,j=1}^{k} L_{y_i y_j} v_i v_j.$$

If in (1.4), $\tau \equiv T$, which means $Q = (T_0, T) \times R^n$, we can replace L by $L + \text{const.}$ In that case (2.3) is equivalent to the assumption that L is bounded below. (The boundedness assumptions on A, B, L can be replaced by suitable growth conditions as $|x| \to \infty$; see [10]. However, when we work in some bounded region N, the behavior of these functions for large $|x|$ is of interest only in that they insure the

global existence of the solutions φ^ε to (1.8$^\varepsilon$)–(1.9) and a uniform bound for φ^ε on N.)

To avoid some analytical difficulties, and to obtain stronger results, we replace (2.2) in §§ 4–7 by the stronger condition

$$vL_{yy}v \geqq c(|y|)|v|^2, \qquad c > 0, \tag{2.2'}$$

for all $v \in R^n$, where $c(r)$ is a positive, nonincreasing function of r. We also suppose in §§ 4–6 that K is compact. In that case $c(|y|) \geqq c_0 > 0$ for all $y \in K$.

Properties of H. If K is compact, then H is clearly well-defined by (1.7). If K is not compact (as in § 7), then we shall always assume (2.2') and the following.

(2.3') For each (s, x, p), $L + pf$ has a minimum on K at a unique $y = V(s, x, p)$. Moreover, $|p| \leqq v$ implies $|V| \leqq R(v)$ for some nondecreasing $R(v)$.

We are interested in continuity and differentiability properties of H. We begin with a known lemma. For completeness a proof is included.

LEMMA 2.1. *Let $\Gamma \subset R^m$ be open and $K \subset R^k$ compact. Let Φ and its gradient Φ_q in the variables q be continuous on $\Gamma \times K$; and let*

$$\Psi(q) = \min_{y \in K} \Phi(q, y). \tag{2.4}$$

Then:

(a) *Ψ satisfies a Lipschitz condition on any compact subset Γ_1 of Γ.*

(b) *If $\Phi(q_0, y)$ is minimum on K at y_0 and Ψ is differentiable at y_0, then*

$$\Psi_q(q_0) = \Phi_q(q_0, y_0). \tag{2.5}$$

(c) *Suppose that, for every $q \in \Gamma$, $\Phi(q, y)$ has a minimum on K at a unique point $y = V(q)$. Then $\Psi \in C^1(\Gamma)$.*

Proof. Since Φ_q is bounded on $\Gamma_2 \times K$, where Γ_2 is a neighborhood of Γ_1, a Lipschitz condition

$$|\Phi(q_1, y) - \Phi(q_0, y)| \leqq C|q_1 - q_0|$$

holds for $q_0, q_1 \in \Gamma_1$ and $y \in K$. (The constant C depends on Γ_1.) Let $\Phi(q_i, y)$ be minimum on K at y_i, $i = 0, 1$. Then

$$\begin{gathered} \Phi(q_1, y_1) - \Phi(q_0, y_1) \leqq \Psi(q_1) - \Psi(q_0) \leqq \Phi(q_1, y_0) - \Phi(q_0, y_0), \\ |\Psi(q_1) - \Psi(q_0)| \leqq C|q_1 - q_0|. \end{gathered} \tag{2.6}$$

This proves (a). To prove (b), we subtract $(q_1 - q_0)\Phi_q(q_0, y_0)$ from the middle and right-hand terms of (2.6) and divide by $|q_1 - q_0|$. This gives

$$\limsup_{q_1 \to q_0} |q_1 - q_0|^{-1}[\Psi(q_1) - \Psi(q_0) - (q_1 - q_0)\Phi_q(q_0, y_0)] \leqq 0. \tag{2.7}$$

Since Ψ is differentiable at q_0, this implies that $\Phi_q = \Psi_q$ (take $q_1 - q_0 = \pm\rho h$, where h is a fixed unit vector, $\rho \to 0$.) This proves (b).

To prove (c) by compactness of K, V is continuous. It suffices to verify that Ψ is differentiable at each $q_0 \in \Gamma$ and to use (b) with $V(q_0) = y_0$. Let us subtract $(q_1 - q_0)\Phi_q(q_1, y_1)$ from the middle and left-hand terms of (2.6), where $y_1 = V(q_1)$. By continuity of V and Φ_q we get

$$\liminf_{q_1 \to q_0} |q_1 - q_0|^{-1}[\Psi(q_1) - \Psi(q_0) - (q_1 - q_0)\Phi_q(q_0, y_0)] \geqq 0.$$

This together with (2.7) implies that Ψ is differentiable at q_0. This proves (c).

Note. In the older literature differentiability is called total differentiability. By Rademacher's theorem [6, p. 216], Ψ is differentiable almost everywhere in Γ if Ψ is Lipschitz on any compact $\Gamma_1 \subset \Gamma$.

Now let $q = (s, x, p)$, $\Phi = L + pf$, and $\Psi = H$. According to Lemma 2.1, we have the following lemma.

LEMMA 2.2. *Let K be compact. Then*:

(a) *H is Lipschitz on any compact set.*

(b) *If H is differentiable at (s, x, p), then*

$$H_s = L_s + pf_s, \quad H_x = L_x + pf_x, \quad H_p = f. \tag{2.8}$$

(c) *Let $\Gamma \subset R^{2n+1}$ be open. If $L + pf$ is minimum on K at a unique $y = V(s, x. p)$ for every $(s, x, p) \in \Gamma$, then $H \in C^1(\Gamma)$ and* (2.8) *holds there.*

In (2.8), $L_s, f_s, \cdots, f$ are evaluated at a point y, where $L + pf$ is minimum on K. In (c), $y = V$. The symbol f_x denotes the matrix of partial derivatives $\partial f_i/\partial x_j$, where $f = (f_1, \cdots, f_n)$.

Example 2.1. Let $f = y$, $L = 1$, $K = \{|y| \leqq 1\}$. Then $H(p) = 1 - |p|$ is not differentiable at $p = 0$. $V(p) = -|p|^{-1}p$, $p \neq 0$. This is a simple special case of the time-optimal problem. The autonomous version of (1.8^0) is just the eikonal equation $|\varphi_x^0| = 1$. See [2, p. 88].

LEMMA 2.3. *Assume* (2.2′), *and either K compact or* (2.3′). *Then*:

(a) $H \in C^1$.

(b) *H_s, H_x, H_p, V are Lipschitz on compact sets.*

(c) *If Γ is a set such that $V \in C^l(\Gamma)$, then $H \in C^{l+1}(\Gamma)$.*

Proof. First consider the case K compact. By reasoning in [7, p. 259], noting that (2.1) implies $\Phi_y = L_y + pB$, we find that V is Lipschitz on any compact set. The remaining assertions follow from (2.8), with $y = V(s, x, p)$.

If K is not compact, let $|p| \leqq v$. Then $|V(s, x, p)| \leqq r$ for $r = R(v)$, and we may replace K by the compact set $K_r = K \cap \{|y| \leqq r\}$.

Example 2.2. Let $f = y$, $L = \frac{1}{2}|y|^2$, $K = \{|y| \leqq 1\}$. Then

$$H(p) = \begin{cases} -\frac{1}{2}|p|^2, & \text{if } |p| \leqq 1, \\ \frac{1}{2} - |p|, & \text{if } |p| \geqq 1. \end{cases}$$

H_p is continuous, but H_{pp} is discontinuous when $|p| = 1$. $V(p) = -p$ if $|p| \leqq 1$, $V(p) = -|p|^{-1}p$ if $|p| \geqq 1$. V is Lipschitz, but not C^1 where $|p| = 1$.

If there are no control constraints, then more can be said.

LEMMA 2.4. *Assume* (2.2′), (2.3′) *and $K = R^k$. Then $H \in C^\infty$ and $L_y(s, x, V) = -pB(s, x)$.*

Proof. By calculus,

$$0 = \Phi_y = L_y + pB.$$

By (2.2′) and the implicit function theorem, $V \in C^\infty$. Hence $H \in C^\infty$.

Note. If $f = y$ and B is the identity matrix, the dual formulas $L_y = -p$, $H_p = y$ are the classical Legendre transformation and its inverse.

3. The Pontryagin problem; Regions of strong regularity. In this section we collect some definitions, and recall the classical formula for solution of a linear first order partial differential equation by characteristics. Then we prove a certain lemma, Lemma 3.1, which is similar to a lemma in [9, p. 523].

Given $s \in (T_0, T)$ let $\mathcal{U}_s$ denote the class of all bounded measurable functions u from $[s, T]$ into the control set K. (These u are sometimes called open loop controls.) As usual u and v are regarded as the same if $u(t) = v(t)$ almost everywhere. Let Q be open, with $Q \subset (T_0, T) \times R^n$. The Pontryagin problem is: given initial data $(s, x) \in Q$ minimize $J^0(s, x; u)$ in (1.4^0) among all $u \in \mathcal{U}_s$. Let us suppose that an optimal u^0 exists. (This is true if K is compact and (2.1)–(2.3) hold. For noncompact K additional assumptions are needed; a special case is treated in § 7.)

Let $\gamma^0 = \gamma^0(s, x)$ denote the optimal trajectory corresponding to u^0:

$$\gamma^0 = \{(t, \xi^0(t)) : s \leqq t \leqq \tau^0\}, \qquad \frac{d\xi^0}{dt} = f(t, \xi^0(t), u^0(t)), \tag{3.1}$$

with $\xi^0(s) = x$ and τ^0 the exit time of γ^0 from Q. We often write $z^0 = (\tau^0, \xi^0(\tau^0))$ for the exit place. If (2.2′) holds, then any optimal u^0 is continuous on $[s, \tau^0]$.

There may be several optimal control functions u^0 for the same initial point (s, x). Such points (s, x) are known to present difficulty from the viewpoint of constructing an optimal control policy Y^0. We call them *irregular points.* In the situation considered in § 7, the irregular points are exactly those where the function φ^0 in (1.6^0) fails to be differentiable [9, p. 520].

We shall need to assume that (s, x) belongs to a region N where stronger properties hold. Let us write $\bar{D} = D \cup \partial D$ for the closure of a set D.

DEFINITION 3.1. Let $N = \bar{Q} \cap \mathcal{O}_0$, where $\mathcal{O}_0$ is open. Then N is a *region of strong regularity* if there exist disjoint C^∞ hypersurfaces $\Sigma_1, \cdots, \Sigma_m$ and disjoint open sets $N_1, \cdots, N_m$ such that:

(a) $N - (\Sigma_1 \cup \cdots \cup \Sigma_m) = N_1 \cup \cdots \cup N_m$.

(b) $N \cap \Sigma_1 = N \cap \partial Q, \partial N_j \subset \Sigma_j \cup \Sigma_{j+1} \cup (\partial N - \partial Q)$ for $j = 1, \cdots, m-1$, $\partial N_m \subset \Sigma_m \cup (\partial N - \partial Q)$.

(c) For any $(s, x) \in N$, there is a unique optimal $u^0 \in \mathcal{U}_s$. The corresponding optimal trajectory $\gamma^0(s, x)$ is contained in N.

(d) For $(s, x) \in N_j$, $\gamma^0(s, x)$ meets Σ_i nontangentially at a single point (s_i, x_i) for $i \leqq j$, and $\gamma^0(s, x)$ does not meet Σ_i for $i > j$.

(e) $\varphi^0 \in C^1(\bar{N})$ and $\varphi^0 \in C^\infty(\bar{N}_j)$ for $j = 1, \cdots, m$.

(f) For $(s, x) \in N_j, j = 1, \cdots, m$, and $(s, x) \in \Sigma_1$, $L + \varphi^0_x f$ has a minimum on K at a unique point $y = Y^0(s, x)$, where $Y^0 \in C^\infty(\bar{N}_j)$.

(g) For $(s, x) \in N_j, j = 1, \cdots, m$, and $(s, x) \in \Sigma_1$, the function V in Lemma 2.3 belongs to $C^\infty(\Gamma)$ for some open set Γ containing $(s, x, \varphi^0_x(s, x))$.

See Fig. 1.

The classical technique for constructing regions of strong regularity is the method of characteristics, applied to the Hamilton–Jacobi equation. (In calculus of variations this is called constructing fields of extremals (see [13, Chap. 3]).) In the Appendix the method is outlined in the setting of the Pontryagin problem.

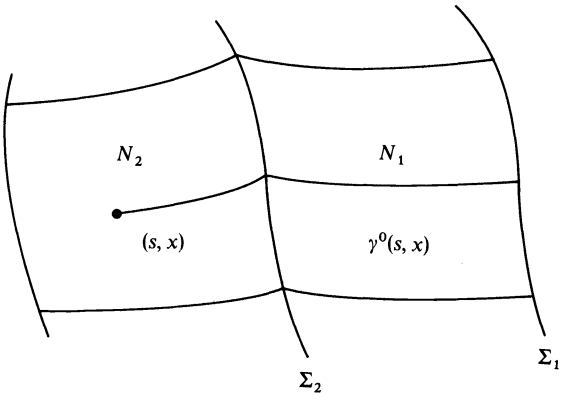


FIG. 1

The functions in (f), (g) are related by

$$Y^0(s, x) = V(s, x, \varphi_x^0(s, x)).$$

Since $\varphi^0 \in C^1(\bar{N})$ the Hamilton–Jacobi equation (1.8^0) holds in $\bar{N}$. Moreover,

$$u^0(t) = Y^0(t, \xi^0(t)) \tag{3.2}$$

except perhaps for $t = s_j$. For this reason we call Y^0 the *optimal control policy in* N. We do not require that the C^∞ extensions of Y^0 from N_{j-1} and N_j to $\bar{N}_{j-1}$ and $\bar{N}_j$ agree on Σ_j. When Y^0 is discontinuous across Σ_j, we call Σ_j a *switching* surface. (In (d) nontangency means from both sides of Σ_j.) If Y^0 is continuous across Σ_j but its first order partial derivatives are discontinuous there, then we call Σ_j a *transition surface.*

A switching or transition surface Σ_j is usually determined by an equation of the form $G_j(s, x, \varphi_x^0) = 0$. The hypersurface $G_j(s, x, p) = 0$ often separates two regions in R^{2n+1} such that V obeys some control constraint in one region but not in the other.

If the strong convexity condition (2.2′) holds, then there are no switching surfaces. If besides (2.2′) there are no control constraints (§ 7), then we may take $m = 1$. There are in that case neither switching nor transition surfaces.

For notational simplicity we set $f^0 = f^{Y^0}$. Thus

$$f^0(s, x) = f(s, x, Y^0(s, x)).$$

Consider a linear partial differential equation

$$\psi_s + f^0\psi_x + g(s, x) = 0 \quad \text{in } N \tag{3.3}$$

with the data $\psi = 0$ on Σ_1. Suppose that $g \in C^\infty(\bar{N}_j)$ for $j = 1, \cdots, m$. Let

$$\psi(s, x) = \int_{\gamma^0(s,x)} g = \int_s^{\tau^0} g(t, \xi^0(t))\, dt, \qquad (s, x) \in N. \tag{3.4}$$

Then ψ is continuous in N, $\psi \in C^\infty(\bar{N}_j \cap N)$ for $j = 1, \cdots, m$, and ψ is a solution to (3.3) with the data on Σ_1. This follows by the classical method of characteristics [2]. If Y^0 and g are continuous across Σ_j, then the first order partial derivatives of ψ are continuous there. Otherwise, the discontinuities in the first order partials of

ψ across Σ_j can be calculated. The derivatives of ψ in directions tangent to Σ_j are the same from both sides, while the derivatives tangent to the characteristic curves from (3.3) satisfy $d\psi/dt = -g$.

In particular, take

$$\psi = \varphi^0_{x_i}, \quad g = H^0_{x_i}, \quad \text{and} \quad \psi = \varphi^0_s, \quad g = H^0_s,$$

where $H^0_{x_i} = H_{x_i}(s, x, \varphi^0_{x_i})$, etc. Recall (2.8). We find that the second order partial derivatives of φ^0 are continuous across a transition surface Σ_j (however, the third order partial derivatives are generally discontinuous across Σ_j). Across a switching surface Σ_j, first order, but not second order, partials of φ_0 are continuous.

The following notations are used in the next lemma, and frequently in later sections:

(3.5) N denotes a region of strong regularity;
N' denotes a bounded set with $\bar{N}' \subset N$.

(3.6) $$d(z) = \text{distance}\,(z, \partial Q) \quad \text{for } z \in Q.$$

(3.7) $$\|\xi\|_t = \max_{s \leqq r \leqq t} |\xi(r)|.$$

In case $t = T$ and the meaning is clear, we write

$$\|\xi\|_T = \|\xi\|.$$

In the lemma it is convenient to denote the solution of the deterministic system equations by η (rather than ξ):

(3.8) $$\frac{d\eta}{dt} = f(t, \eta(t), u(t)), \qquad s \leqq t \leqq T,$$

with $\eta(s) = x$. Note that u, η are defined on $[s, T]$, although their values after the exit time τ from Q do not affect J^0. If $u = u^0$ is optimal, we write as above $\eta = \xi^0$ and τ^0, z^0 for the corresponding exit time and place.

LEMMA 3.1. *Let K be compact, and N', N as above. Given $a > 0$ there exists $\delta = \delta(a, N') > 0$ with the following property: If $(s, x) \in N'$, $d(t, \eta(t)) < \delta$, and*

$$\int_s^t L(r, \eta(r), u(r))\, dr < \varphi^0(s, x) + \delta,$$

then

$$|(t, \eta(t)) - z^0| < a, \quad \|\eta - \xi^0\|_t < a, \quad \int_s^t |u(r) - u^0(r)|\, dr < a.$$

Proof. The conclusions can be restated as follows. Suppose that $(s_m, x_m) \in N'$, $m = 1, 2, \cdots$, $(s_m, x_m) \to (s, x)$, $(t_m, \eta_m(t_m)) \to z^1$ as $m \to \infty$, and

$$\liminf_{m \to \infty} \int_s^{t_m} L(t, \eta_m, u_m)\, dt \leqq \lim_{m \to \infty} \varphi^0(s_m, x_m) = \varphi^0(s, x).$$

Here η_m satisfies (3.8) with $u = u_m$, $\eta_m(s_m) = x_m$, and $z^1 \in \partial Q$. We must show that $z^1 = z^0$, and as $m \to \infty$,

$$\|\eta_m - \xi^0\|_{t_m} \to 0, \qquad \int_s^{t_m} |u_m - u^0|\, dt \to 0.$$

Since K is compact we may assume (by taking subsequences) that u_m tends weak* in $L^\infty[s, T]$ to some $u^1 \in \mathcal{U}_s$. Then using (2.1), we have that η_m tends uniformly on $[s, T]$ to the solution η^1 of $\dot{\eta}^1 = f(t, u^1, \eta^1)$ with $\eta^1(s) = x$. By lower semicontinuity of J^0 with respect to weak* convergence (using (2.2)) and the strict positivity of L in (2.3), we conclude that u^1 is optimal. By uniqueness of u^0, (c) of the definition of strong regularity (Definition 3.1),

$$u^1(t) = u^0(t), \qquad \eta^1(t) = \xi^0(t) \quad \text{for } s \leqq t \leqq \tau^0,$$

and $z^1 = z^0$. Since $t_m \to \tau^0$, $\|\eta_m - \xi^0\|_{t_m} \to 0$. (So far we have used only the fact that each point in $\overline{N}'$ is regular.)

To complete the proof, let

$$h(t, y) = L(t, \xi^0(t), y) + p^0(t) f(t, \xi^0(t), y),$$
$$p^0(t) = \varphi_x^0(t, \xi^0(t)).$$

Then $h(t, y)$ has a minimum on K at the unique point $y = u^0(t)$, except perhaps for a finite number of $t = s_j$, $j = 2, \cdots, m$, as in (d) of Definition 3.1. Let I be any compact subset of $[s, \tau^0]$ not containing these s_j. Given $\alpha > 0$ there exists $\beta = \beta(\alpha)$ tending to 0 as $\alpha \to 0$ such that

$$h(t, y) - h(t, u^0(t)) < \alpha \Rightarrow |y - u^0(t)| < \beta, \tag{3.9}$$

provided $t \in I$. By linearity (2.1) of f in y and convergence of t_m to τ^0,

$$\lim_{m \to \infty} \int_s^{t_m} p^0(t) f(t, \xi^0, u_m)\, dt = \int_s^{\tau_0} p^0(t) f(t, \xi^0, u^0)\, dt.$$

By the first part of the proof,

$$\lim_{m \to \infty} \int_s^{t_m} L(t, \xi^0, u_m)\, dt = \lim_{m \to \infty} \int_s^{t_m} L(t, \xi_m, u_m)\, dt = \int_s^{\tau_0} L(t, \xi^0, u^0)\, dt.$$

Therefore,

$$\lim_{m \to \infty} \int_s^{t_m} h(t, u_m(t))\, dt = \int_s^{\tau_0} h(t, u^0(t))\, dt.$$

Since $h(t, u_m) \geqq h(t, u^0)$ and t_m tends to τ^0, $h(t, u_m) \to h(t, u^0)$ in measure on I. By (3.9), u_m tends to u^0 in measure on I. Since $u_m(t) \in K$, K is compact, and the measure of $[s, \tau^0] - I$ can be made arbitrarily small,

$$\lim_{m \to \infty} \int_s^{t_m} |u_m - u^0|\, dt = 0.$$

4. The stochastic problem; Preliminary estimates. Throughout §§ 4–6 we assume that K is compact, and that (2.1), (2.2′), (2.3) hold. The main results of the present section are the estimates contained in Lemmas 4.4 and 4.5.

We begin with some notation and preliminary remarks.

(a) *The class $\mathcal{Y}$ of control policies.* Let $\mathcal{Y}$ consist of all functions Y from $[T_0, T] \times R^n$ into K with the following property: for each $T' < T$ there exist

positive constants C, α (depending on Y and T') such that

$$|Y(s', x) - Y(s, x)| \leqq C|s' - s|^{\alpha},$$

$$|Y(s, x') - Y(s, x)| \leqq C|x' - x|$$

for every $x, x' \in R^n$ and $T_0 \leqq s, s' \leqq T'$.

(b) *The region* Q. To insure that the boundary problem (1.8$^\varepsilon$)–(1.9) has a well-defined solution, we take Q an open subset of $(T_0, T) \times R^n$ of a shape "suitable" from the viewpoint of parabolic equations. Thus we assume that

$$\partial Q = B_{T_0} \cup S \cup B_T,$$

where B_{T_0}, B_T are open subsets of the hyperplanes $\{T_0\} \times R^n$, $\{T\} \times R^n$ respectively, and the "lateral boundary" S is a compact subset of a C^∞ manifold S_0. Moreover, if $\nu = (\nu_0, \nu_1, \cdots, \nu_n)$ is normal to S_0 at a point of S, then $\nu_i \neq 0$ for some $i > 0$. Let $\mathscr{S}$ denote the class of all such Q.

The "essential" part of ∂Q is denoted by

$$\partial^* Q = S \cup B_T.$$

Let $C^{1,2}(Q)$ denote the class of functions ψ bounded and continuous on $\bar{Q}$ with $\psi_s, \psi_{x_i}, \psi_{x_i x_j}$ continuous on Q, $i, j = 1, \cdots, n$. If $Q \in \mathscr{S}$, then the function φ^ε in (1.6$^\varepsilon$) is the unique solution of (1.8$^\varepsilon$)–(1.9) in $C^{1,2}(Q)$. See [7, § 2], [11, Chap. 7] or [18]. Actually, φ^ε_x is Hölder continuous on $\bar{Q}$, and $\varphi^\varepsilon_s, \varphi^\varepsilon_{x_i x_j}$ are Hölder continuous on any compact subset of $\bar{Q} - (S \cap B_T)$, $i, j = 1, \cdots, n$. The optimal policy Y^ε defined by (1.10) belongs to $\mathscr{Y}$; more precisely, Y^ε has an extension for $(s, x) \notin Q$ which belongs to $\mathscr{Y}$. See [7, p. 261].

If S is empty, then $\partial^* Q$ is the hyperplane $\{T\} \times R^n$. The boundary data (1.9) are then Cauchy data.

We shall write

$$\Lambda^\varepsilon = \frac{\partial}{\partial s} + \varepsilon \Delta_x. \tag{4.1}$$

We also put $f^\varepsilon = f^{Y^\varepsilon}$, $L^\varepsilon = L^{Y^\varepsilon}$, where as in § 1

$$g^Y(s, x) = g(s, x, Y(s, x));$$

and we put $H^\varepsilon = H(s, x, \varphi^\varepsilon_x)$. Thus (1.8$^\varepsilon$) becomes

$$0 = \Lambda^\varepsilon \varphi^\varepsilon + H^\varepsilon = \Lambda^\varepsilon \varphi^\varepsilon + f^\varepsilon \varphi^\varepsilon_x + L^\varepsilon.$$

(c) *Some probabilistic notions.* Let Ω be the space of continuous functions ω from $[T_0, T]$ into R^n. We shall suppose that all processes which appear are defined on this sample space Ω. Fix $s \in [T_0, T]$. Let B_{st} denote the least σ-algebra containing all sets $\{\omega \in \Omega : \omega(r) \in A\}$, $s \leqq r \leqq t$, $A \subset R^n$ open. A process ξ on the time interval $[s, T]$ is called *nonanticipative* if $\xi(t)$ (meaning $\xi(t, \cdot)$) is B_{st} measurable for $s \leqq t \leqq T$. A random variable τ with $s \leqq \tau \leqq T$ is a (nonanticipative) *stopping* time if the event $\tau > t$ is B_{st} measurable, $s \leqq t \leqq T$. If τ and τ' are stopping times, then

$$\tau \wedge \tau' = \min(\tau, \tau')$$

is also a stopping time.

Let ξ be a continuous nonanticipative process on $[s, T]$, with values in R^n. Such a process is specified by a measure π on B_{sT}, if we set $\xi(t, \omega) = \omega(t)$, $\omega \in \Omega$. In particular, let w be an n-dimensional Brownian motion on $[s, T]$, with $w(s) = 0$; the associated π is a Wiener measure.

Let $D \subset [T_0, T] \times R^n$, with interior D_0 not empty. If $(s, x) \in D_0$, let

$$\tau_D = \inf \{t : (t, \xi(t)) \notin D_0\}.$$

We call τ_D the *exit time*; it is a stopping time. Thus in formula (1.4), $\tau = \tau_Q$. We also write $z_D = (\tau_D, \xi(\tau_D))$ for the exit place.

Given a control policy $Y \in \mathscr{Y}$, a continuous nonanticipative process ξ is determined by the system equations (1.1), the initial data (1.2), and (1.5). See [3, Chap. VI, § 3] or [12, Chap. 8]. The process ξ is in fact Markov and strongly Feller [5, Chaps. 11, 13], also [24].

If Y^ε is the optimal policy, then write ξ^ε for this process:

$$d\xi^\varepsilon = f^\varepsilon(t, \xi^\varepsilon(t))\, dt + (2\varepsilon)^{1/2}\, dw, \qquad s \leqq t \leqq T,$$

with $\xi^\varepsilon(s) = x$. The corresponding exit times and places are denoted by $\tau_D^\varepsilon, z_D^\varepsilon$.

(d) *Two estimates for maxima.* Suppose that

$$d\xi = f(t, \xi(t), u(t))\, dt + (2\varepsilon)^{1/2}\, dw,$$

$$d\eta = f(t, \eta(t), u(t))\, dt,$$

with $\xi(s) = \eta(s)$. By a standard estimate for ordinary differential equations,

$$\|\xi - \eta\|_t \leqq C\varepsilon^{1/2} \|w\|_t, \qquad s \leqq t \leqq T, \tag{4.2}$$

where C is a positive constant depending on $T - T_0$ and a bound for $|f_x|$. (This notation $\|\cdot\|_t$ was defined in § 3.)

Since w is a Brownian motion with $w(s) = 0$, the random variables $w_1(t), \cdots, w_n(t)$ are independent with mean 0 and variance $t - s$. For any $\lambda > 0$,

$$\Pr\{\|w\|_t \geqq \lambda\} \leqq 4n \Pr\left\{w_i(t) \geqq \frac{\lambda}{n}\right\} \leqq \frac{4n^2(t-s)^{1/2}}{(2\pi)^{1/2}\lambda} \exp\left\{-\frac{\lambda^2}{n^2(t-s)}\right\}.$$

See [3, p. 392]. Therefore for $\lambda = e\varepsilon^{-1/2}$ we get, after writing $\|w\|$ for $\|w\|_T$,

$$\Pr\{\varepsilon^{1/2}\|w\| \geqq e\} \leqq C\varepsilon^{1/2} \exp(-\beta/\varepsilon), \tag{4.3}$$

for suitable positive C, β depending on e.

(e) *A stochastic analogue of the fundamental theorem of calculus.* Suppose that ξ satisfies the stochastic differential equation

$$d\xi = b(t, \xi(t))\, dt + (2\varepsilon)^{1/2}\, dw, \qquad s \leqq t \leqq T,$$

with $\xi(s) = x$. Assume that b is continuous on $[T_0, T) \times R^n$, bounded, and satisfies a uniform Lipschitz condition in x on $[T_0, T'] \times R^n$ if $T' < T$. In particular, one can take $b = f^Y$, where $Y \in \mathscr{Y}$. When $Y = Y^\varepsilon$, $b = f^\varepsilon$; another useful choice will be $b = b^\varepsilon$, where b^ε is defined in § 6. Let $(s, x) \in D$, with D open, and τ' a stopping time such that $s \leqq \tau' \leqq \tau_D$. If $\psi \in C^{1,2}(D)$, with $\Lambda^\varepsilon\psi + b\psi_x$ bounded, then

$$\psi(s, x) = -E\int_s^{\tau'} (\Lambda^\varepsilon\psi + b\psi_x)\, dt + E\psi(z'), \tag{4.4}$$

where $z' = (\tau', \xi(\tau'))$. This is a consequence of the following well-known [12, p. 391] Ito formula for stochastic differentials of composites. Let $\tilde{\psi} \in C^{1,2}(R^{n+1})$. Then

$$d\tilde{\psi}(t, \xi(t)) = (\Lambda^\varepsilon \tilde{\psi} + b\tilde{\psi}_x)\, dt + (2\varepsilon)^{1/2}\tilde{\psi}_x\, dw. \tag{4.4'}$$

To get (4.4) from (4.4'), let $\bar{D}_1 \subset D$. Take $\tilde{\psi}$ such that $\tilde{\psi} = \psi$ on $\bar{D}_1$. If $\tau' \leqq \tau_{D_1}$, then we get (4.4) by integrating (4.4') between s and τ' and taking expected values. From properties of stochastic integrals,

$$E\int_s^{\tau'} \tilde{\psi}_x\, dw = E\int_s^T \chi\tilde{\psi}_x\, dw = 0,$$

where χ is the characteristic function of the interval $[s, \tau')$. Finally, D is the union of an expanding sequence $D_1, D_2, \cdots$ with $\bar{D}_j \subset D$; and τ' is the limit of the monotone sequence of stopping times $\tau_j = \tau' \wedge \tau_{D_j}$.

Note. Equation (4.4) is actually valid under the weaker assumption that ψ belongs to the class $\mathscr{F}(D)$ defined in § 8. This can be shown by an approximation argument, which we shall carry out under the following somewhat stronger assumption than $\psi \in \mathscr{F}(D)$. Suppose that $\psi \in C^1(\bar{D})$ with the second order partials $\psi_{x_i x_j}$ bounded and continuous except on a finite number of C^∞ hypersurfaces $\Sigma_2, \cdots, \Sigma_m$. (In the proofs of Theorems 6.4 and 6.5 we encounter such functions ψ.) Take $\tilde{\psi}_1, \tilde{\psi}_2, \cdots$ in $C^{1,2}(R^{n+1})$ such that, for each $i = 1, 2, \cdots$, $\tilde{\psi}_j$ tends to ψ uniformly on $\bar{D}_i$ as $j \to \infty$ while $\Lambda^\varepsilon\tilde{\psi}_j + b(\tilde{\psi}_j)_x$ is bounded on D_i and tends to $\Lambda^\varepsilon\psi + b\psi_x$ on $D_i - (\Sigma_2 \cup \cdots \cup \Sigma_m)$. For each $t > s$ the probability distribution of $\xi(t)$ is absolutely continuous with respect to Lebesgue measure. In fact, the probability density is a weak solution of the forward equation; see for instance [8]. Therefore,

$$\Pr\{\xi(t) \in \Sigma_2 \cup \cdots \cup \Sigma_m\} = 0$$

for almost all t, which implies

$$\lim_{j\to\infty} E\int_s^{\tau_i} [\Lambda^\varepsilon\tilde{\psi}_j + b(\tilde{\psi}_j)_x]\, dt = E\int_s^{\tau_i} [\Lambda^\varepsilon\psi + b\psi_x]\, dt.$$

From this we get (4.4) as before.

In the next lemma the notation is as follows. Let $\mathcal{O}$ be an open set and $D = \bar{Q} \cap \mathcal{O}$. We suppose that $D \subset N$, where N is a region of strong regularity; moreover, $f^0 = f^{Y^0}$ is assumed to be defined outside N so that $f^0 \in C^1(\mathcal{O} - \bar{D})$. See (f) of Definition 3.1. (We do not require outside N that $L + \varphi_x^0 f$ be minimum at Y^0.)

Let C_0 be such that

$$|f^0(s, x') - f^0(s, x)| \leqq C_0|x' - x|$$

for $(s, x), (s, x') \in \mathcal{O}$. Let

$$d\xi^0 = f^0(t, \xi^0(t))\, dt, \qquad s \leqq t \leqq t_1,$$

$$\xi^0(s) = x, \qquad (t, \xi^0(t)) \in D \quad \text{for } s \leqq t \leqq \tau^0,$$

$$(t, \xi^0(t)) \in \mathcal{O} - D \quad \text{for } \tau^0 < t \leqq t_1.$$

Let b and the process ξ be as in part (e) of this section, $\xi(s) = x$.

LEMMA 4.1. *Given $a > 0$ there exists $e > 0$ with the following property: If $|f^0 - b| < e$ in $\mathcal{O}$ and $\varepsilon^{1/2}\|w\|_{\tau_D} < e$, then:*

(i) $$|\tau_D - \tau^0| < a, \qquad (\tau_D, \xi(\tau_D)) \in \partial^* Q.$$

(ii) $$\|\xi - \xi^0\|_{\tau_D} < a.$$

Proof. Let $\tilde{w}(t) = w(t \wedge \tau_D)$,

$$\tilde{\xi}(t) = x + \int_s^t b(r, \tilde{\xi}(r))\, dr + (2\varepsilon)^{1/2}\tilde{w}(t).$$

Then $\|\tilde{w}\| = \|w\|_{\tau_D}$ and $\tilde{\xi}(t) = \xi(t)$ for $s \leqq t \leqq \tau_D$. By standard estimates for ordinary differential equations,

$$|\tilde{\xi}(t) - \xi^0(t)| \leqq 3e + C_0 \int_s^t |\tilde{\xi}(r) - \xi^0(r)|\, dr,$$

provided $\varepsilon^{1/2}\|\tilde{w}\| < e$ and $s \leqq t \leqq t_1 \wedge \tau_{\mathcal{O}}$. We may assume that $t_1 \geqq \tau^0 + a$. For sufficiently small e, the hypotheses of the lemma imply

$$\tau_{\mathcal{O}} > t_1, \qquad \|\tilde{\xi} - \xi^0\|_{t_1} < a,$$
$$(t, \xi(t)) \in D \quad \text{for } s \leqq t \leqq \tau^0 - a,$$
$$(t, \tilde{\xi}(t)) \in \mathcal{O} - D \quad \text{for } \tau^0 + a \leqq t \leqq t_1.$$

These inequalities imply (i) and (ii).

Note. Let γ^0 be the trajectory corresponding to ξ^0 on $[s, \tau^0]$, as in § 3. Then e depends on a, C_0, and distance $(\gamma^0, \partial\mathcal{O})$ but not on ε.

The following lemma gives an upper estimate for φ^ε near a boundary point z in a region of strong regularity. The method is to define locally a solution ψ of a linear equation (4.5) which serves as a "barrier" at z. In Lemma 5.1 a corresponding lower estimate for φ^ε near z will be proved.

By *relative neighborhood* of z we mean a set $M = \bar{Q} \cap \mathcal{O}$, where $\mathcal{O}$ is open and contains z.

LEMMA 4.2. *Let N be a region of strong regularity and $z \in N \cap \partial^* Q$. Then given $m_1 > 1$ there exist $\varepsilon_0 > 0$ and a relative neighborhood M' of z such that $\varphi^\varepsilon(s, x) \leqq m_1 \varphi^0(s, x)$ for all $(s, x) \in M'$ and $0 < \varepsilon < \varepsilon_0$.*

Proof. Let M be a relative neighborhood of z such that M is of class $\mathcal{S}$ and $\bar{M} \subset N_1$ (where N_1 is as in § 3). Let Y be some control policy such that $Y = Y^0$ in M, $Y \in \mathcal{Y}$, and $Y \in C^1(\mathcal{O})$ for some open $\mathcal{O}$ with $M = \bar{Q} \cap \mathcal{O}$. Let $\psi \in C^{1,2}(M)$ satisfy

$$\Lambda^\varepsilon \psi + f^Y \psi_x + L^Y = 0 \quad \text{in } M \tag{4.5}$$

with $\psi = 0$ on $\partial^* M$. Let ξ be the solution of (1.1) corresponding to Y with initial data $(s, x) \in M$. Since $Y = Y^0$ in M, by (4.4),

$$\psi(s, x) = E \int_s^{\tau_M} L^0\, dt, \qquad (s, x) \in M. \tag{4.5'}$$

Let

$$\tilde{\psi}(s, x) = J^\varepsilon(s, x; Y) = E \int_s^{\tau_Q} L^Y\, dt.$$

By the definition (1.6$^\varepsilon$), $\varphi^\varepsilon \leqq \tilde{\psi}$. By (4.4) and the fact that $\tilde{\psi}$ also satisfies (4.5),

$$\tilde{\psi}(s,x) = \psi(s,x) + E\tilde{\psi}(z_M), \qquad (s,x) \in M.$$

Let $\mu > 0$ and $\varphi = (1+\mu)\varphi^0$. Then in M,

$$\Lambda^\varepsilon \varphi + f^0 \varphi_x + L^0 = \varepsilon(1+\mu)\Delta_x \varphi^0 - \mu L^0. \tag{4.6}$$

Since $\Delta_x \varphi^0$ is continuous in $\bar{M}$ and $L^0 \geqq c_1 > 0$ by (2.3), the right side of (4.6) is negative for small ε. Moreover, $\varphi \geqq 0$ and $\psi = 0$ on $\partial^* M$. By the maximum principle for parabolic equations, $\psi \leqq \varphi$ on M. Moreover, $\tilde{\psi}$ is bounded on Q by $C = (T - T_0) \sup L$, and $\tilde{\psi} = 0$ on $\partial^* Q$. Thus

$$\tilde{\psi}(s,x) \leqq \varphi(s,x) + C \Pr\{z_M \in \partial M - \partial^* Q\}. \tag{4.7}$$

To estimate the probability on the right side, we use Lemma 4.1 with $b = f^0$. Let $a > 0$ and a relative neighborhood M' of z be chosen such that $\bar{M}' \subset M$ and, for all $(s,x) \in M'$,

$$\text{distance}\,(\gamma^0, \partial \mathcal{O}) \geqq a.$$

As usual $\gamma^0 = \gamma^0(s,x)$. Thus, for some $e > 0$,

$$\tilde{\psi}(s,x) \leqq \varphi(s,x) + C \Pr\{\varepsilon^{1/2} \|w\|_{\tau_M} \geqq e\}, \qquad (s,x) \in M'.$$

Since τ_M is a random (not fixed) time we cannot apply (4.3). However, let

$$\tilde{w}(t) = w(\tau_M \wedge t) = \int_s^t \chi(r)\, dw(r),$$

where χ is the characteristic function of $[s, \tau_M)$. By properties of stochastic integrals [12, p. 383],

$$\begin{aligned} \Pr\{\|w\|_{\tau_M} \geqq \lambda\} &= \Pr\{\|\tilde{w}\|_T \geqq \lambda\} \\ &\leqq \lambda^{-2} E \int_s^T \chi^2(r)\, dr = \lambda^{-2} E(\tau_M - s). \end{aligned}$$

Since $L^0 \geqq c_1 > 0$, by (4.5′)

$$E(\tau_M - s) \leqq c_1^{-1} \psi \leqq c_1^{-1} \tilde{\psi}.$$

Let $\lambda = e\varepsilon^{-1/2}$. For suitable C_1,

$$\Pr\{z_M \in \partial M - \partial^* Q\} \leqq C_1 \varepsilon e^{-2} \tilde{\psi}(s,x).$$

From this and (4.7),

$$(1 - C_2\varepsilon)\tilde{\psi} \leqq (1+\mu)\varphi^0.$$

Since $\varphi^\varepsilon \leqq \tilde{\psi}$ we get Lemma 4.2.

Note. If $A \subset N \cap \partial^* Q$ and A is compact, then the same kind of estimate in Lemma 4.2 holds in a relative neighborhood of A.

For making certain estimates it is a useful technical device to consider not only control policies, but also nonanticipative control processes. Such a process will be denoted by u. The system equations (1.1) with the initial data again have a well-defined solution ξ. The process ξ is nonanticipative, but not necessarily Markov.

Roughly speaking, a nonanticipative u may use past data in an arbitrary way, while a control policy Y uses at time t past data only through the current state $\xi(t)$. One cannot obtain better system performance using nonanticipative controls than with control policies. In fact, let Y^ε be the optimal control policy and ξ^ε as in part (c) of this section. Then

$$u^\varepsilon(t) = Y^\varepsilon(t, \xi^\varepsilon(t))$$

is an optimal nonanticipative control. See [7, p. 263]. This result is to be expected from the assumption of complete observability.

The following was proved in [7, p. 264].

LEMMA 4.3. *Let u be any nonanticipative control process and τ' any stopping time with $s \leqq \tau' \leqq \tau_Q$. Then*

$$\varphi^\varepsilon(s, x) \leqq E \int_s^{\tau'} L(t, \xi(t), u(t))\, dt + E\varphi^\varepsilon(\tau', \xi(\tau')).$$

Equality holds if $u = u^\varepsilon$, $\xi = \xi^\varepsilon$.

The proof of the next lemma follows that for a corresponding result [9, Theorem 4a] for the Cauchy problem ($\tau^0 = \tau^\varepsilon = T$). However, we need Lemmas 4.2 and 4.3 to deal with the case when exit may occur through the lateral part S of $\partial^* Q$. For the Cauchy problem the estimate for $\varphi^\varepsilon - \varphi^0$ in Lemma 4.4 is valid globally, not merely in N'.

LEMMA 4.4. *Let N, N' be regions of strong regularity with $\bar{N}' \subset N$. There exist positive C and ε_0 such that*

$$|\varphi^\varepsilon(s, x) - \varphi^0(s, x)| \leqq C\varepsilon^{1/2},$$
$$\Pr\{z_N^\varepsilon \in \partial N - \partial^* Q\} \leqq C\varepsilon^{1/2}$$

for all $(s, x) \in N'$ and $0 < \varepsilon < \varepsilon_0$.

Proof. Part 1. Let $(s, x) \in N'$. The optimal open loop control u^0 is suboptimal in the stochastic problem. The corresponding process ξ is defined by

$$d\xi = f(t, \xi(t), u^0(t))\, dt + (2\varepsilon)^{1/2}\, dw$$

with $\xi(s) = x$. Let τ_N be the exit time for that process, and

$$\tau' = \tau^0 \wedge \tau_N, \qquad z' = (\tau', \xi(\tau')).$$

By Lemma 4.3,

$$\varphi^\varepsilon(s, x) \leqq E \int_s^{\tau'} L(t, \xi, u^0)\, dt + E\varphi^\varepsilon(z').$$

Since $L > 0$ and $\tau' \leqq \tau^0$,

$$\int_s^{\tau'} L(t, \xi^0, u^0)\, dt \leqq \varphi^0(s, x),$$

$$\varphi^\varepsilon(s, x) \leqq \varphi^0(s, x) + E \int_s^{\tau'} L \Big|_{\xi^0, u^0}^{\xi, u^0} dt + E\varphi^\varepsilon(z').$$

From (4.2) with $u = u^0$, $\eta = \xi^0$, and boundedness of L_x the middle term on the right side is in absolute value $\leqq C\varepsilon^{1/2}$ (we use C for any large enough constant). By the note after Lemma 4.2 there is a relative neighborhood M of $A = \bar{N}' \cap \partial^* Q$ such that $\bar{M} \subset N$ and $\varphi^\varepsilon \leqq m_1 \varphi^0$ for all $(s, x) \in M$ and small ε. Since N' is a region of strong regularity, $\gamma^0(s, x) \subset N'$. Hence there exists $e > 0$ (not depending on the choice of $(s, x) \in N'$) such that $\varepsilon^{1/2}\|w\| < e$ implies $z' \in M$. Define $d(z)$ by (3.6). If $z' \in M$, then either

$$\tau' = \tau_N, \quad z' \in \partial^* Q, \quad d(z') = 0;$$

or else

$$\tau' = \tau^0, \qquad d(z') \leqq |\xi(\tau^0) - \xi^0(\tau^0)|.$$

In either case, $d(z') \leqq \|\xi - \xi^0\|_{\tau'}$. From Lemma 4.2 we have for such w,

$$\varphi^\varepsilon(z') \leqq m_1 \varphi^0(z') \leqq C d(z') \leqq C\varepsilon^{1/2}\|w\|;$$

$$E\{\varphi^\varepsilon(z'); \varepsilon^{1/2}\|w\| < e\} \leqq C\varepsilon^{1/2}.$$

On the other hand, φ^ε is bounded on N (by $(T - T_0) \sup L$). Hence from (4.3),

$$E\{\varphi^\varepsilon(z'); \varepsilon^{1/2}\|w\| \geqq e\} \leqq C\varepsilon^{1/2} \exp(-\beta/\varepsilon).$$

These two inequalities give

$$\varphi^\varepsilon(s, x) \leqq \varphi^0(s, x) + C\varepsilon^{1/2}. \tag{4.8}$$

Part 2. To get an inequality opposite to (4.8) we use the fact that the optimal stochastic u^ε is suboptimal in the deterministic Pontryagin problem with probability 1. Again, let $(s, x) \in N'$. From the definitions of ξ^ε, u^ε,

$$d\xi^\varepsilon = f(t, \xi^\varepsilon, u^\varepsilon)\, dt + (2\varepsilon)^{1/2}\, dw.$$

Define η^ε by

$$d\eta^\varepsilon = f(t, \eta^\varepsilon, u^\varepsilon)\, dt,$$

with $\eta^\varepsilon(s) = \xi^\varepsilon(s) = x$. Let $\tau^\varepsilon = \tau_Q^\varepsilon$, τ_1^ε = exit time from Q of $(t, \eta^\varepsilon(t))$, and

$$\tau = \tau^\varepsilon \wedge \tau_1^\varepsilon, \qquad z = (\tau, \eta^\varepsilon(\tau)).$$

With probability 1,

$$d(z) \leqq \|\xi^\varepsilon - \eta^\varepsilon\|_\tau \leqq C\varepsilon^{1/2}\|w\|, \tag{4.9}$$

$$\varphi^0(s, x) \leqq \int_s^\tau L(t, \eta^\varepsilon, u^\varepsilon)\, dt + \varphi^0(z). \tag{4.10}$$

Let $0 < a \leqq \frac{1}{2}$ distance $(N', \partial N - \partial^* Q)$, and δ be as in Lemma 3.1. Consider the three mutually exclusive events

$$B_1 = \{C\varepsilon^{1/2}\|w\| \geqq \delta\},$$

$$B_2 = \{C\varepsilon^{1/2}\|w\| < \delta, z_N^\varepsilon \in \partial N - M\},$$

$$B_3 = \{C\varepsilon^{1/2}\|w\| < \delta, z_N^\varepsilon \in M\},$$

where M is as in Part 1, C as in (4.9). Using (4.9) we also take δ small enough that

$\|\eta^\varepsilon - \xi^0\|_\tau > a$ on B_2. The event $\{z_N^\varepsilon \in \partial N - \partial^* Q\}$ is contained in $B_1 \cup B_2$. Thus

$$\Pr\{z_N^\varepsilon \in \partial N - \partial^* Q\} \leqq \Pr\{B_1\} + \Pr\{B_2\}. \tag{4.11}$$

By (4.3) and boundedness of φ^0,

$$E\{\varphi^0(z); B_1\} \leqq C\varepsilon^{1/2} \exp(-\beta/\varepsilon).$$

By Lemma 3.1 with $t = \tau, \eta = \eta^\varepsilon$,

$$\varphi^0(s,x) + \delta \leqq \int_s^\tau L(t, \eta^\varepsilon, u^\varepsilon)\, dt \quad \text{on } B_2. \tag{4.12}$$

By (4.9),

$$\varphi^0(z) \leqq C\varepsilon^{1/2}\|w\| \quad \text{on } B_3,$$
$$E\{\varphi^0(z); B_3\} \leqq C\varepsilon^{1/2}.$$

We now take expectations in (4.10), to obtain

$$\varphi^0(s,x) + \delta \Pr\{B_2\} \leqq E\int_s^\tau L(t, \eta^\varepsilon, u^\varepsilon)\, dt + C\varepsilon^{1/2}.$$

From $L > 0$, $\tau \leqq \tau^\varepsilon$, L_x bounded, and (4.2), we have

$$E\int_s^\tau L(t, \eta^\varepsilon, u^\varepsilon)\, dt \leqq E\int_s^\tau L(t, \xi^\varepsilon, u^\varepsilon)\, dt + C\varepsilon^{1/2}$$
$$\leqq E\int_s^{\tau^\varepsilon} L(t, \xi^\varepsilon, u^\varepsilon)\, dt + C\varepsilon^{1/2}.$$

The first term on the right side is $\varphi^\varepsilon(s,x)$. Thus,

$$\varphi^0(s,x) + \delta \Pr\{B_2\} \leqq \varphi^\varepsilon(s,x) + C\varepsilon^{1/2}. \tag{4.13}$$

This together with inequality (4.8) implies $|\varphi^\varepsilon - \varphi^0| \leqq C\varepsilon^{1/2}$ and $\Pr\{B_2\} \leqq C\varepsilon^{1/2}$. By the (sharper) estimate (4.3) for $\Pr\{B_1\}$ and (4.11), this completes the proof of Lemma 4.4.

In proving (4.13) we have also established

$$E\int_\tau^{\tau^\varepsilon} L(t, \xi^\varepsilon, u^\varepsilon)\, dt \leqq C\varepsilon^{1/2}. \tag{4.14}$$

This will be used in proving the next lemma.

LEMMA 4.5 *Let N, N' be as in Lemma 4.4. Given $a > 0$ there exist positive C and ε_0 such that*:

(i) $$\Pr\{|z_N^\varepsilon - z^0| \geqq a\} \leqq C\varepsilon^{1/2},$$

(ii) $$\Pr\{\|\xi^\varepsilon - \xi^0\|_{\tau^\varepsilon} \geqq a\} \leqq C\varepsilon^{1/2},$$

(iii) $$\Pr\left\{\int_s^{\tau^\varepsilon} |u^\varepsilon - u^0|\, dt \geqq a\right\} \leqq C\varepsilon^{1/2},$$

for all $(s,x) \in N'$ and $0 < \varepsilon < \varepsilon_0$. Here $\tau^\varepsilon = \tau_Q^\varepsilon$.

Proof. We may suppose that $0 < a <$ distance $(N', \partial N - \partial^* Q)$. Define B_1, B_2, B_3 as in Part 2 of the preceding proof, but with $\delta = \delta(a/2, N')$. Let

$$B_4 = B_3 \cap \left\{ \int_s^\tau L(t, \eta^\varepsilon, u^\varepsilon)\, dt < \varphi^0(s, x) + \delta \right\}.$$

The same proof as for (4.13) shows that $P\{B_3 - B_4\} \leqq C\varepsilon^{1/2}$. Since $P\{B_1\} + P\{B_2\} \leqq C\varepsilon^{1/2}$, it suffices to consider the case when B_4 holds. By Lemma 3.1, on B_4,

$$|z - z^0| < \frac{a}{2}, \quad \|\eta^\varepsilon - \xi^0\|_\tau < \frac{a}{2}, \quad \int_s^\tau |u^\varepsilon - u^0|\, dt < \frac{a}{2}.$$

From the definitions of z_N^ε, z,

$$|z_N^\varepsilon - z^0| \leqq |z_N^\varepsilon - z| + |z - z^0|,$$
$$|z_N^\varepsilon - z| \leqq |\tau_N^\varepsilon - \tau| + |\xi^\varepsilon(\tau_N^\varepsilon) - \eta^\varepsilon(\tau)|.$$

Moreover, $\tau_N^\varepsilon = \tau^\varepsilon$ on B_3 (hence on B_4) and $\tau \leqq \tau^\varepsilon$. Hence

$$|z_N^\varepsilon - z| \leqq \tau^\varepsilon - \tau + \|\xi^\varepsilon - \eta^\varepsilon\| + |\eta^\varepsilon(\tau^\varepsilon) - \eta^\varepsilon(\tau)|.$$

Since $d\eta^\varepsilon/dt$ is bounded, we get on B_4

$$|z_N^\varepsilon - z^0| \leqq C(\tau^\varepsilon - \tau) + C\varepsilon^{1/2}\|w\| + \tfrac{1}{2}a.$$

Similarly, on B_4,

$$\|\xi^\varepsilon - \xi^0\|_{\tau^\varepsilon} \leqq C(\tau^\varepsilon - \tau) + C\varepsilon^{1/2}\|w\| + \tfrac{1}{2}a,$$
$$\int_s^{\tau^\varepsilon} |u^\varepsilon - u^0|\, dt \leqq C(\tau^\varepsilon - \tau) + \tfrac{1}{2}a$$

for some positive C. Let $\mu < a/(2C) - \delta/C$. From (4.14) and (2.3),

$$\Pr\{\tau^\varepsilon - \tau \geqq \mu\} \leqq (\mu c_1)^{-1} E \int_\tau^{\tau^\varepsilon} L(t, \xi^\varepsilon, \mu^\varepsilon)\, dt \leqq C\varepsilon^{1/2}.$$

This together with the estimate (4.3) proves Lemma 4.5.

5. Convergence of φ_x^ε to φ_x^0. This is proved uniformly on compact subsets of a region N of strong regularity (Lemma 5.5). As a consequence of this and assumption (2.2′) the optimal control policy Y^ε tends in N to Y^0 (Corollary 5.6). Like the estimates in Lemma 4.4, Lemma 5.5 is a preliminary step toward more precise results in § 6.

We begin with a lower estimate for φ^ε near boundary points to complement the upper estimate in Lemma 4.2.

LEMMA 5.1. *Under the hypotheses of Lemma* 4.2, *given* $m_2 < 1$ *there exist* $\varepsilon_0 > 0$ *and a relative neighborhood* M' *of* z *such that* $\varphi^\varepsilon(s, x) \geqq m_2\varphi^0(s, x)$ *for all* $(s, x) \in M'$ *and* $0 < \varepsilon < \varepsilon_0$.

Proof. From equations (1.8^ε), (1.8^0), and (1.7),

$$\Lambda^\varepsilon\varphi^\varepsilon + f^\varepsilon\varphi_x^\varepsilon + L^\varepsilon = 0 \quad \text{in } Q,$$
$$\Lambda^\varepsilon\varphi^0 + f^\varepsilon\varphi_x^0 + L^\varepsilon \geqq \varepsilon\Delta_x\varphi^0 \quad \text{in } N.$$

(Recall § 4(b), (c) for notation.) Let $\psi = \varphi^\varepsilon - \varphi^0$. Then

$$\Lambda^\varepsilon \psi + f^\varepsilon \psi_x \leqq -\varepsilon \Delta_x \varphi^0 \quad \text{in } N.$$

Let M be a relative neighborhood of z with $\overline{M} \subset N$. Set $\tau' = \tau_M^\varepsilon$, $z' = z_M^\varepsilon$. By (4.4) with $\xi = \xi^\varepsilon$,

$$\psi(s, x) \geqq \varepsilon E \int_s^{\tau'} \Delta_x \varphi^0 \, dt + E\psi(z'), \qquad (s, x) \in M.$$

From Lemma 4.4, $|\psi(z')| \leqq C\varepsilon^{1/2}$; while by (1.9), $\psi(z') = 0$ if $z' \in \partial^* Q$. Thus

$$\psi(s, x) \geqq \varepsilon E \int_s^{\tau'} \Delta_x \varphi^0 \, dt - C\varepsilon^{1/2} \Pr\{z' \in \partial M - \partial^* Q\}. \tag{5.1}$$

Since $\tau' \leqq \tau_Q^\varepsilon$ and $L \geqq c_1$,

$$E \int_s^{\tau'} \Delta \varphi^0 \, dt \leqq CE(\tau_Q^\varepsilon - s) \leqq Cc_1^{-1} \varphi^\varepsilon(s, x). \tag{5.2}$$

From (1.1) we have the elementary estimate

$$\|\xi^\varepsilon - x\|_t \leqq c(t - s) + (2\varepsilon)^{1/2} \|w\|_t, \qquad s \leqq t \leqq T,$$

where $|f| \leqq c$. Take $a > 0$ and a relative neighborhood M' of z such that

$$\text{distance}\,((s, x), \partial M - \partial^* Q) \geqq 2a \quad \text{if } (s, x) \in M'.$$

Then $z' \in \partial M - \partial^* Q$ implies either $\tau' - s \geqq ac^{-1}$ or $(2\varepsilon)^{1/2} \|w\|_{\tau'} \geqq a$. But

$$\Pr\{\tau' - s \geqq ac^{-1}\} \leqq a^{-1} cE\{\tau' - s\} \leqq (ac_1)^{-1} c\varphi^\varepsilon(s, x),$$

$$\Pr\{(2\varepsilon)^{1/2} \|w\|_{\tau'} \geqq a\} \leqq C\varepsilon \varphi^\varepsilon(s, x)$$

by reasoning used to prove Lemma 4.2. Thus

$$\Pr\{z' \in \partial M - \partial^* Q\} \leqq C\varphi^\varepsilon(s, x). \tag{5.3}$$

From (5.1), (5.2), (5.3) and Lemma 4.2,

$$\psi(s, x) \geqq -C\varepsilon^{1/2} \varphi^0(s, x), \qquad (s, x) \in M'.$$

This proves Lemma 5.1.

LEMMA 5.2. *Let N be a region of strong regularity. Then $\varphi_x^\varepsilon \to \varphi_x^0$, $\varphi_s^\varepsilon \to \varphi_s^0$ as $\varepsilon \to 0$, uniformly on any compact set $A \subset N \cap \partial^* Q$.*

Proof. Since $\varphi^\varepsilon = \varphi^0 = 0$ on $N \cap \partial^* Q$, their derivatives in directions tangent to $\partial^* Q$ are 0 there. It suffices to show uniform convergence of the normal derivatives on A. Let ν be the interior unit normal at $z \in A$. Then

$$\frac{\partial \varphi^\varepsilon}{\partial \nu} - \frac{\partial \varphi^0}{\partial \nu} = \lim_{(s,x) \to z} \frac{\varphi^\varepsilon(s, x) - \varphi^0(s, x)}{d(s, x)} = \left(\frac{\partial \varphi^0}{\partial \nu}\right)^{-1} \lim_{(s,x) \to z} \frac{\varphi^\varepsilon(s, x) - \varphi^0(s, x)}{\varphi^0(s, x)}.$$

According to Lemmas 4.2 and 5.1 the latter limit approaches 0 as $\varepsilon \to 0$, uniformly on A. This proves Lemma 5.2.

To prove convergence of φ_x^ε to φ_x^0 at points of $N - \partial^* Q$ we need a formula expressing φ_x^ε in terms of the processes u^ε and ξ^ε. Let us write $f_x^\varepsilon(t, \xi) = f_x(t, \xi, u^\varepsilon)$.

Given s, define the fundamental matrices $W^\varepsilon = (W^\varepsilon_{ij})$, $i, j = 1, \cdots, n$, for each $\varepsilon \geqq 0$ by

$$dW^\varepsilon = f^\varepsilon_x(t, \xi^\varepsilon(t))W^\varepsilon\, dt, \qquad W^\varepsilon(s) = \text{identity}. \tag{5.4}$$

Suppose that $\tilde{\psi}_i \in C^{1,2}(R^{n+1})$ for $i = 1, \cdots, n$, and let $\zeta_i(t) = \tilde{\psi}_i(t, \xi^\varepsilon(t))$. The collection of processes $\xi^\varepsilon, \zeta, W^\varepsilon$ satisfies the stochastic differential equations (1.1) with $u = u^\varepsilon$, (5.4), and

$$d\zeta_i = (\Lambda^\varepsilon \tilde{\psi}_i + f^\varepsilon(\tilde{\psi}_i)_x)\, dt + (2\varepsilon)^{1/2}(\tilde{\psi}_i)_x\, dw. \tag{5.5}$$

Hence, we may compute the stochastic differential $d(\zeta_i W^\varepsilon_{ij})$ by the Ito rule, obtaining

$$d(\zeta_i W^\varepsilon_{ij}) = (d\zeta_i)W^\varepsilon_{ij} + \zeta_i\, dW^\varepsilon_{ij}. \tag{5.6}$$

The additional term which normally appears in the product rule for stochastic differentials is missing since no noise term appears in (5.4).

LEMMA 5.3. *Let τ' be any stopping time, $s \leqq \tau' \leqq \tau^\varepsilon_Q$, and let $z' = (\tau', \xi^\varepsilon(\tau'))$. Then for $\varepsilon > 0$ and $(s, x) \in Q$,*

$$\varphi^\varepsilon_x(s, x) = E\int_s^{\tau'} L^\varepsilon_x W^\varepsilon\, dt + E\{\varphi^\varepsilon_x(z')W^\varepsilon(\tau')\}, \tag{5.7}$$

where $L^\varepsilon_x = L_x(t, \xi^\varepsilon(t), u^\varepsilon(t))$.

Proof. Let $P_i = \varphi^\varepsilon_{x_i}$ and differentiate (1.8$^\varepsilon$) with respect to x_i, obtaining

$$\Lambda^\varepsilon P_i + H^\varepsilon_{x_i} + (P_i)_x H^\varepsilon_p = 0, \qquad i = 1, \cdots, n, \tag{5.8}$$

where $H^\varepsilon_x, H^\varepsilon_p$ are evaluated at (s, x, P). As noted in § 4(b), P_i is continuous on $\bar{Q}$. Using Lemma 2.3(b), we have that $H^\varepsilon_{x_i}$ and $(P_i)_x H^\varepsilon_p$ are Hölder continuous on any compact subset of Q; hence so is $\Lambda^\varepsilon P_i$. This implies that $P_i \in C^{1,2}(Q)$. See [11] or [18]. Using (2.8) we rewrite (5.8) as

$$\Lambda^\varepsilon P_i + f^\varepsilon(P_i)_x = -(L^\varepsilon_{x_i} + Pf^\varepsilon_{x_i}). \tag{5.8$'$}$$

For the moment suppose that $\tau' \leqq \tau_D$, where $\bar{D} \subset Q$. In (5.6) take $\tilde{\psi}_i \in C^{1,2}(R^{n+1})$ such that $\tilde{\psi}_i = P_i$ on $\bar{D}$. Then, for $(s, x) \in D$,

$$P_i(s, x)W^\varepsilon_{ij}(s) = -E\int_s^{\tau'} \{[\Lambda^\varepsilon P_i + f^\varepsilon(P_i)_x]W^\varepsilon_{ij}\, dt + P_i\, dW^\varepsilon_{ij}\} + E\{P_i(z')W^\varepsilon_{ij}(\tau')\}.$$

Let us sum on i, and use (5.4), (5.8$'$), $W(s) =$ identity. We get (5.7), in case $\tau' \leqq \tau_D$. By writing Q as the union of an expanding sequence $D_1, D_2, \cdots$ and τ' as the limit of the monotone sequence $\tau' \wedge \tau_{D_j}$, we then get Lemma 5.3.

LEMMA 5.4. *Let N be a region of strong regularity, and $\bar{N}' \subset N$. Then there exist positive C and ε_0 such that $|\varphi^\varepsilon_x| \leqq C$ for $(s, x) \in N'$ and $0 < \varepsilon < \varepsilon_0$.*

Proof. First of all, let $1 < \rho < \frac{3}{2}$. The following crude bound holds:

$$|\varphi^\varepsilon_x| \leqq C\varepsilon^{-\rho}, \qquad (s, x) \in \partial^* Q. \tag{5.9}$$

(This is rather similar to an estimate of Oleinik [21].) Clearly (5.9) holds on B_T since $\varphi^\varepsilon = 0$ there; see § 4(b) for notation. Suppose that $z_0 \in S$. If $v = (v_0, v_1, \cdots, v_n)$

is the interior unit normal to S at z_0, then $\nu_i \neq 0$ for some $i > 0$. Let $z_1 = z_0 - r\nu$, where r is chosen (simultaneously for all z_0) small enough that the sphere $|z - z_1| \leqq r$ meets $\bar{Q}$ only at z_0. Consider the function

$$\psi^\varepsilon(s, x) = 1 - \exp[-\varepsilon^{-\rho}(|z - z_1|^2 - r^2)],$$
$$\mathcal{O} = \{-\tfrac{1}{2}\varepsilon^\rho < |z - z_1|^2 - r^2 < \tfrac{1}{2}\varepsilon^\rho\}, \qquad M = \mathcal{O} \cap \bar{Q},$$

where $\varepsilon > 0$ is chosen small enough that $x \neq x_1$ if $z = (s, x) \in \mathcal{O}$, $z_1 = (s_1, x_1)$. An easy calculation shows that for small ε,

$$\Lambda^\varepsilon \psi^\varepsilon + f^\varepsilon \psi_x^\varepsilon + L^\varepsilon < 0 \quad \text{in } \mathcal{O},$$
$$\psi^\varepsilon = 1 - e^{-1/2} \quad \text{on } (\partial\mathcal{O}) \cap \bar{Q}.$$

Since

$$\Lambda^\varepsilon \varphi^\varepsilon + f^\varepsilon \varphi_x^\varepsilon + L^\varepsilon = 0 \quad \text{in } Q,$$
$$\Lambda^\varepsilon(\alpha\psi^\varepsilon - \varphi^\varepsilon) + f^\varepsilon(\alpha\psi^\varepsilon - \varphi^\varepsilon)_x < 0 \quad \text{in } M,$$

for any $\alpha \geqq 1$. On $\partial^* Q$, $\varphi^\varepsilon = 0$, $\psi^\varepsilon \geqq 0$; while on $\partial M - \partial^* Q$, $\alpha\psi^\varepsilon - \varphi^\varepsilon \geqq 0$ for suitable α. By the maximum principle, $\varphi^\varepsilon \leqq \alpha\psi^\varepsilon$ in M for such α. In Q, $\varphi^\varepsilon \geqq 0$; while $\varphi^\varepsilon = \psi^\varepsilon = 0$ at z_0. Hence

$$0 \leqq \frac{\partial\varphi^\varepsilon}{\partial\nu} \leqq \alpha\frac{\partial\psi^\varepsilon}{\partial\nu} \leqq C\varepsilon^{-\rho}$$

at z_0. Since the derivatives of φ^ε in directions tangent to S are 0 at z_0, this implies (5.9).

Now let D_1, D_2, D_3 be regions of strong regularity with

$$N' \subset D_1, \quad \bar{D}_1 \subset D_2, \quad \bar{D}_2 \subset D_3, \quad \bar{D}_3 \subset N.$$

Since $z_Q^\varepsilon \in \partial^* Q - \bar{D}_3$ implies $\tau_{D_3}^\varepsilon < \tau_Q^\varepsilon$, we have by Lemma 4.4,

$$\Pr\{z_Q^\varepsilon \in \partial^* Q - \bar{D}_3\} \leqq \Pr\{z_{D_3}^\varepsilon \in \partial D_3 - \partial^* Q\} \leqq C_1\varepsilon^{1/2}$$

for $(s, x) \in D_2$. By Lemma 4.2, φ_x^ε is bounded on $(\partial^* Q) \cap \bar{D}_3$. Moreover, L_x^ε and W^ε are bounded. By Lemma 5.3 with $\tau' = \tau_Q^\varepsilon$,

$$|\varphi_x^\varepsilon| \leqq C_2 + C_3 E\{|\varphi_x^\varepsilon(z_Q^\varepsilon)|\,;\, z_Q^\varepsilon \in \partial^* Q - \bar{D}_3\},$$
$$|\varphi_x^\varepsilon| \leqq C_4\varepsilon^{-\rho+1/2}$$

for $(s, x) \in \bar{D}_2$ and sufficiently small ε. We now apply Lemmas 4.2 and 5.3 again, with $\tau' = \tau_{D_2}^\varepsilon$, and use the estimate just obtained when $z' \in \partial D_2 - \partial^* Q$. Thus $|\varphi_x^\varepsilon| \leqq C\varepsilon^{-\rho+1}$ when $(s, x) \in \bar{D}_1$. Since $\rho < \frac{3}{2}$ another application of these lemmas gives $|\varphi_x^\varepsilon| \leqq C$, $(s, x) \in N'$.

LEMMA 5.5. *Let N, N' be as in Lemma 5.4. Then $\varphi_x^\varepsilon \to \varphi_x^0$ as $\varepsilon \to 0$, uniformly on N'.*

Proof. Let $(s, x) \in N'$. For $\varepsilon = 0$ formula (5.7) becomes

$$\varphi_x^0(s, x) = \int_s^{\tau^0} L_x^0 W^0\, dt + \varphi_x^0(z^0) W^0(\tau^0). \tag{5.7^0}$$

Let D be a region of strong regularity with $\bar{N}' \subset D$ and $\bar{D} \subset N$. Set $\tau' = \tau_D^\varepsilon$, $z' = z_D^\varepsilon$. From Lemma 4.5, $\|W^\varepsilon - W^0\|_{\tau'}$, $L_x^\varepsilon - L_x^0$, and $z' - z^0$ tend to 0 in probability as $\varepsilon \to 0$, uniformly with respect to $(s, x) \in N'$. Moreover, L_x^ε, W^ε are bounded, while $\varphi_x^\varepsilon(z')$ is bounded by Lemma 5.4. By Lemma 5.2,

$$\varphi_x^\varepsilon(z') - \varphi_x^0(z^0) = \varphi_x^\varepsilon(z') - \varphi_x^0(z') + \varphi_x^0(z') - \varphi_x^0(z^0)$$

tends uniformly to 0 in probability. Similarly,

$$|W^\varepsilon(\tau') - W^0(\tau^0)| \leqq \|W^\varepsilon - W^0\|_{\tau'} + |W^0(\tau') - W^0(\tau^0)|$$

tends uniformly to 0 in probability. Using (5.7), φ_x^ε tends uniformly on N' to the right side of (5.7^0) as $\varepsilon \to 0$. This proves Lemma 5.5.

The optimal control policy Y^ε is given by

$$Y^\varepsilon(s, x) = V(s, x, \varphi_x^\varepsilon(s, x)). \tag{5.10}$$

By Lemma 2.3, V is locally Lipschitz; thus Lemma 5.5 has the following corollary.

COROLLARY 5.6. *Y^ε tends to Y^0 uniformly on N'.*

In § 8 we shall see that Lemma 5.5 remains true in cases where Y^ε may have switching surfaces. However, Corollary 5.6 is not true in such cases.

From Corollary 5.6 there is a much sharper estimate than the one in Lemma 4.4 for the probability that the optimal stochastic trajectory leaves N before time τ_Q^ε.

COROLLARY 5.7. *There exist positive C, β, ε_0 such that*

$$\Pr\{z_N^\varepsilon \in \partial N - \partial^* Q\} \leqq C\varepsilon^{1/2} \exp(-\beta/\varepsilon)$$

for $(s, x) \in N'$ and $0 < \varepsilon < \varepsilon_0$.

Proof. Let D be as in the proof of Lemma 5.5, and $\mathcal{O}$ open with $D = \mathcal{O} \cap \bar{Q}$. By Corollary 5.6, Y^ε tends to Y^0 uniformly on $\bar{D}$. We can define Y^ε outside Q so that $Y^\varepsilon \in \mathcal{Y}$ and Y^ε tends to Y^0 uniformly on $\bar{\mathcal{O}}$. In Lemma 4.1, let $b = f^\varepsilon$. There exists $e > 0$ such that $(s, x) \in N'$, $|f^0 - f^\varepsilon| < e$ in $\mathcal{O}$, and $\varepsilon^{1/2}\|w\| < e$ imply

$$z_D^\varepsilon = z_N^\varepsilon = z_Q^\varepsilon \in \partial^* Q.$$

We apply (4.3).

6. Asymptotic formulas for φ^ε, φ_x^ε. We are now ready to state the main results, Theorems 6.2, 6.4, 6.5. At the end of the section we indicate how the same methods tell the goodness of the policy Y^0 in the stochastic problem.

As in §§ 4, 5 the control set K is assumed compact. We again assume (2.1), (2.3). Theorem 6.2 is true assuming the convexity condition (2.2). However, in Theorems 6.4, 6.5 we need the stronger condition (2.2′). The following observation will be used several times.

LEMMA 6.1. *Let $G(s, x, p)$ be a real-valued function on $N \times R^n$, satisfying for any v a Lipschitz condition on $N \times \{|p| \leqq v\}$. Let P, P^0 be bounded, measurable on N. Then:*

(a) *There exists a bounded Borel measurable b such that*

$$G(s, x, P) - G(s, x, P^0) = b(s, x) \cdot (P - P^0).$$

(b) *If G is C^1 in p, then we can take*

$$b(s, x) = \int_0^1 G_p(s, x, P^0 + \lambda(P - P^0))\, d\lambda.$$

Proof. In (a) let

$$b(s,x) = \begin{cases} \dfrac{G(s,x,P) - G(s,x,P^0)}{|P - P^0|^2}(P - P^0) & \text{if } P \neq P^0, \\ 0 & \text{if } P = P^0. \end{cases}$$

Part (b) is just the integral form of the mean value theorem.

We now consider formula (1.11) with $l = 1$. For brevity we set $\theta_1 = \theta$.

THEOREM 6.2. *Let N be a region of strong regularity, and N' such that $\bar{N}' \subset N$. Let θ be defined in N by*

$$\theta_s + f^0\theta_x + \Delta_x\varphi^0 = 0 \tag{6.1}$$

with $\theta = 0$ on $N \cap \partial^* Q$. Then

$$\varphi^\varepsilon = \varphi^0 + \varepsilon\theta + o(\varepsilon), \tag{6.2}$$

where $\varepsilon^{-1}o(\varepsilon) \to 0$ as $\varepsilon \to 0$, uniformly for $(s,x) \in N'$.

Proof. Let us assume (2.2′); changes in the proof needed when merely (2.2) holds will be indicated in § 8. By (3.4) with $g = \Delta_x\varphi^0$,

$$\theta(s,x) = \int_s^{\tau^0} \Delta_x\varphi^0(t, \xi^0(t))\,dt. \tag{6.3}$$

Since (2. 2′) holds, $\varphi^0 \in C^2(\bar{N})$ and $\theta \in C^1(\bar{N})$. Both φ^0 and θ are in $C^\infty(\bar{N}_j)$ for $j = 1, \cdots, m$. For $\varepsilon > 0$ let

$$\theta^\varepsilon = \varepsilon^{-1}(\varphi^\varepsilon - \varphi^0).$$

We need to show that $\theta^\varepsilon \to \theta$ uniformly on N' as $\varepsilon \to 0$. Define b^ε by Lemma 6.1(b) with $G = H$, $P = \varphi_x^\varepsilon$, $P^0 = \varphi_x^0$. Now $P^0 \in C^1(\bar{N})$, while P, P_x are continuous on $\bar{N}$ and H_p is Lipschitz on compact sets by Lemma 2.3(b). Hence b^ε is continuous on $\bar{N}$ and satisfies a uniform Lipschitz condition in x. We extend b^ε outside $\bar{N}$ to have these same properties. Then the Ito conditions hold, insuring that the process $\tilde{\xi}^\varepsilon$ satisfying $\tilde{\xi}^\varepsilon(s) = x$ and

$$d\tilde{\xi}^\varepsilon = b^\varepsilon(t, \tilde{\xi}^\varepsilon(t))\,dt + (2\varepsilon)^{1/2}\,dw, \qquad s \leqq t \leqq T,$$

is well-defined.

Let $\mathcal{O}, D$ be as in the proof of Corollary 5.7. By Lemma 5.5, $b^\varepsilon \to f^0$ uniformly on $\bar{D}$ as $\varepsilon \to 0$. We may arrange the above extension of b^ε outside $\bar{N}$ such that also $b^\varepsilon \to f^0$ uniformly on $\bar{\mathcal{O}}$. By definition of θ^ε and (1.8^ε), (1.8^0),

$$\Lambda^\varepsilon\theta^\varepsilon + b^\varepsilon\theta_x^\varepsilon + \Delta_x\varphi^0 = 0 \quad \text{in } N. \tag{6.1$^\varepsilon$}$$

From (4.4) we get, writing τ_D, z_D for the exit time and place for $(t, \tilde{\xi}^\varepsilon(t))$,

$$\theta^\varepsilon(s,x) = E\int_s^{\tau_D} \Delta_x\varphi^0(t, \tilde{\xi}^\varepsilon(t))\,dt + E\theta^\varepsilon(z_D). \tag{6.3$^\varepsilon$}$$

We apply Lemma 4.1 with $b = b^\varepsilon$. As $\varepsilon \to 0$, $\|\tilde{\xi}^\varepsilon - \xi^0\|_{\tau_D}$ and $\tau_D - \tau^0$ tend to 0 in probability, uniformly with respect to $(s,x) \in N'$. Moreover (as in the proof of Corollary 5.7),

$$\Pr\{z_D \in \partial D - \partial^* Q\} \leqq C\varepsilon^{1/2}\exp(-\beta/\varepsilon)$$

for small ε and $(s, x) \in N'$. By (1.9), $\theta^\varepsilon(z_D) = 0$ if $z_D \in \partial^* Q$. Since φ^ε, φ^0 are bounded,

$$|\theta^\varepsilon(s, x)| \leqq C\varepsilon^{-1}, \qquad (s, x) \in Q.$$

As $\varepsilon \to 0$, $E\theta^\varepsilon(z_D) \to 0$; while the middle term in (6.3$^\varepsilon$) tends to the right side of (6.3) uniformly on N'. This proves Theorem 6.2.

LEMMA 6.3. *Using the above notation* $\theta^\varepsilon_s \to \theta^0_s$, $\theta^\varepsilon_x \to \theta^0_x$ *as* $\varepsilon \to 0$ *uniformly on any compact set* $A \subset N \cap \partial^* Q$.

Proof. As in the proof of Lemma 5.2 it suffices to show that $(\varphi^0)^{-1}(\theta^\varepsilon - \theta^0)$ is uniformly small on a relative neighborhood M' of any $z \in A$ for small ε. Choose M, M' as in the proof of Lemma 4.2. By Lemma 4.1 there exists $e > 0$ such that $z_M \in \partial M - \partial^* Q$ implies $\varepsilon^{1/2}\|w\|_{\tau_M} \geqq e$ whenever $(s, x) \in M'$ and $|b^\varepsilon - f^0| < e$ on $\mathcal{O}$. Since b^ε tends to f^0 uniformly on $\mathcal{O}$, we have (see proof of Lemma 4.2)

$$\Pr\{z_M \in \partial M - \partial^* Q\} \leqq CE\{\tau_M - s\}, \qquad (s, x) \in M'. \tag{6.4}$$

Let $\omega^\varepsilon = \theta^\varepsilon - \theta$. By (6.1), (6.1$^\varepsilon$),

$$\Lambda^\varepsilon \omega^\varepsilon + b^\varepsilon \omega^\varepsilon_x + (b^\varepsilon - f^0)\theta_x + \varepsilon \Delta_x \theta = 0 \quad \text{in } N,$$

$$\omega^\varepsilon(s, x) = E\int_s^{\tau_M} [(b^\varepsilon - f^0)\cdot\theta_x + \varepsilon\Delta_x\theta]\, dt + E\omega^\varepsilon(z_M), \tag{6.5}$$

the integrand being evaluated at $(t, \tilde{\xi}^\varepsilon(t))$. On $N \cap \partial^* Q$, $\omega^\varepsilon = 0$. By Theorem 6.2, $\omega^\varepsilon(z_M)$ is uniformly small for small ε; moreover, $\omega^\varepsilon(z_M) = 0$ if $z_M \in \partial^* Q$. Since $\theta \in C^\infty(\overline{M})$, θ_x and $\Delta_x\theta$ are bounded there. By (6.4) and (6.5) it suffices to obtain for small ε an estimate

$$E\{\tau_M - s\} \leqq C\varphi^0(s, x), \qquad (s, x) \in M. \tag{6.6}$$

By (4.4), $E\{\tau_M - s\} = \psi(s, x)$, where ψ is the unique bounded solution in $C^{1,2}(M)$ of

$$\Lambda^\varepsilon \psi + b^\varepsilon \psi_x + 1 = 0$$

with $\psi = 0$ on $\partial^* M$. Let $\varphi = C\varphi^0$ with $C > c_1^{-1}$, c_1 as in (2.3). Then

$$\Lambda^\varepsilon(\varphi - \psi) + b^\varepsilon(\varphi - \psi)_x + (f^0 - b^\varepsilon)\varphi_x - \varepsilon\Delta_x\varphi \leqq -c < 0,$$

where $c = Cc_1 - 1$. For small ε, $|(f^0 - b^\varepsilon)\varphi_x - \varepsilon\Delta_x\varphi| < c$ in M. Since $\varphi - \psi \geqq 0$ on $\partial^* M$, the maximum principle implies $\varphi \geqq \psi$ in M for such ε. This establishes (6.6) and hence the lemma.

THEOREM 6.4. *Let* N, N' *be as in Theorem* 6.2. *Assume* (2.2′). *Then*

$$\varphi^\varepsilon_x = \varphi^0_x + \varepsilon\theta_x + o(\varepsilon), \tag{6.7}$$

where $\varepsilon^{-1}o(\varepsilon) \to 0$ *as* $\varepsilon \to 0$, *uniformly for* $(s, x) \in N'$.

Proof. In the notation above, (6.7) is equivalent to the statement $\theta^\varepsilon_x \to \theta^0_x$ as $\varepsilon \to 0$. For notational simplicity let

$$\pi_i = \theta^\varepsilon_{x_i} = \varepsilon^{-1}(\varphi^\varepsilon_{x_i} - \varphi^0_{x_i}).$$

Recall (see proof of Lemma 5.3) that $\varphi^\varepsilon_{x_i} \in C^{1,2}(\overline{N})$; while from § 3, $\varphi^0_{x_i} \in C^1(\overline{N})$ and $\varphi^0_{x_i} \in C^\infty(\overline{N}_j)$ for $j = 1, \cdots, m$. Using (5.8) we get

$$\Lambda^\varepsilon\pi_i + H^\varepsilon_p(\pi_i)_x + \varepsilon^{-1}(H^\varepsilon_{x_i} - H^0_{x_i}) + \varepsilon^{-1}(H^\varepsilon_p - H^0_p)\varphi^0_{xx_i} + (\Delta_x\varphi^0)_{x_i} = 0, \tag{6.8$^\varepsilon$}$$

where $H^\varepsilon_x = H_x(s, x, \varphi^\varepsilon_x)$, etc.

From Lemmas 6.1(a) and 2.3(b),

$$\varepsilon^{-1}(H_x^\varepsilon - H_x^0) = \pi B_1^\varepsilon, \qquad \varepsilon^{-1}(H_p^\varepsilon - H_p^0) = \pi B_2^\varepsilon,$$

where the matrices $B_i^\varepsilon(s, x)$ are bounded. Moreover, by Lemmas 6.1(b) and 2.3(c), $B_1^\varepsilon \to H_{px}^0$, $B_2^\varepsilon \to H_{pp}^0$ uniformly on any compact subset of $N - (\Sigma_2 \cup \cdots \cup \Sigma_m)$. Define fundamental matrices X^ε by

$$dX^\varepsilon = (B_1^\varepsilon + B_2^\varepsilon \varphi_{xx}^0)X^\varepsilon \, dt, \qquad X^\varepsilon(s) = \text{identity},$$

where B_i^ε, φ_{xx}^0 are evaluated at $(t, \xi^\varepsilon(t))$. Let D be as in the proof of Theorem 6.2 and $\tau' = \tau_D^\varepsilon$, $z' = z_D^\varepsilon$. By Lemma 2.3, $H_p^\varepsilon = f^\varepsilon$. As in the proof of Lemma 5.3 we get, from (6.8^ε),

$$\theta_x^\varepsilon(s, x) = E \int_s^{\tau'} (\Delta_x \varphi^0)_x X^\varepsilon \, dt + E\{\theta_x^\varepsilon(z')X^\varepsilon(\tau')\}$$

with $(\Delta_x \varphi^0)_x$ evaluated at $(t, \xi^\varepsilon(t))$.

From previous estimates,

$$\Pr\{z' \in \partial D - \partial^* Q\} \leqq C\varepsilon^{1/2} \exp(-\beta/\varepsilon)$$

for $(s, x) \in N'$. Since φ_x^ε, φ_x^0 are bounded on D, $|\theta_x^\varepsilon(z')| \leqq C\varepsilon^{-1}$. Using Lemma 4.1, with $b = f^\varepsilon$, Lemma 6.3, and previous reasoning we find that uniformly on N',

$$\lim_{\varepsilon \to 0} \theta_x^\varepsilon(s, x) = \int_s^{\tau^0} (\Delta_x \varphi^0)_x X^0 \, dt + \theta_x(z^0)X^0(\tau^0), \tag{6.9}$$

where $(\Delta_x \varphi^0)_x$ is now evaluated at $(t, \xi^0(t))$ and

$$\frac{dX^0}{dt} = (H_{px}^0 + H_{pp}^0 \varphi_{xx}^0)X^0, \qquad t \geqq s,$$

$$X^0(s) = \text{identity}.$$

Let us differentiate (6.1) with respect to x_i, $i = 1, \cdots, n$, recalling that $f^0 = H_p^0$. We get the linear system of partial differential equations

$$(\theta_{x_i})_s + f^0(\theta_{x_i})_x + \theta_x(H_{px_i}^0 + H_{pp}\varphi_{xx_i}^0) + (\Delta_x \varphi^0)_{x_i} = 0, \qquad i = 1, \cdots, n.$$

By the method of characteristics, $\theta_x(s, x)$ equals the right side of (6.9). This proves Theorem 6.4.

Using Theorem 6.4 we can sharpen Theorem 6.2 by including the second order term in the expansion (1.11). For brevity we write $\theta_2 = \chi$.

THEOREM 6.5. *Keep the same hypotheses as Theorem* 6.4. *Let* χ *be defined in* N *by*

$$\chi_s + f^0\chi_x + \tfrac{1}{2}\theta_x H_{pp}^0 \theta_x + \Delta_x \theta = 0 \tag{6.10}$$

with $\chi = 0$ *on* $N \cap \partial^* Q$. *Then*

$$\varphi^\varepsilon = \varphi^0 + \varepsilon\theta + \varepsilon^2\chi + o(\varepsilon^2), \tag{6.11}$$

where $\varepsilon^{-2}o(\varepsilon^2) \to 0$ *as* $\varepsilon \to 0$, *uniformly for* $(s, x) \in N'$.

Proof. Let

$$\chi^\varepsilon = \varepsilon^{-1}(\theta^\varepsilon - \theta) = \varepsilon^{-2}(\varphi^\varepsilon - \varphi^0 - \varepsilon\theta).$$

Then (6.11) states that $\chi^\varepsilon \to \chi$ as $\varepsilon \to 0$. Now

$$\Lambda^\varepsilon \chi^\varepsilon + b^\varepsilon \chi_x^\varepsilon + \varepsilon^{-1}(b^\varepsilon - f^0)\theta_x + \Delta_x \theta = 0.$$

Moreover,

$$\varepsilon^{-1}(b^\varepsilon - f^0) = \varepsilon^{-1} \int_0^1 (H'_p - H_p^0)\, d\lambda,$$

$$H'_p = H_p(s, x, \varphi_x^0 + \lambda(\varphi_x^\varepsilon - \varphi_x^0)).$$

Then $\varepsilon^{-1}(b^\varepsilon - f^0)$ is bounded on D (using D as above), and by Theorem 6.4,

$$\lim_{\varepsilon \to 0} \varepsilon^{-1}(b^\varepsilon - f^0) = \tfrac{1}{2}\theta_x H_{pp}^0$$

uniformly on any compact subset of $N - (\Sigma_1 \cup \cdots \cup \Sigma_m)$. By the same proof as for Theorem 6.2 we find that

$$\lim_{\varepsilon \to 0} \chi^\varepsilon(s, x) = \int_s^{\tau^0} (\tfrac{1}{2}\theta_x H_{pp}^0 \theta_x + \Delta_x \theta)\, dt \tag{6.12}$$

uniformly on N'. By (6.10) and (3.4) the right side is $\chi(s, x)$. This proves Theorem 6.5.

One may ask about the validity of formula (1.11) for larger values of l. Consider first the case when merely (2.2) holds, as in Theorem 6.2, and take $l = 2$. Then Y^0 may have switching surfaces. Across a switching surface, θ_x and $H_p^0 = f^0$ are generally discontinuous. It would seem that at a point of γ^0 on a switching surface, $\frac{1}{2}\theta_x H_{pp}^0 \theta_x + \Delta_x \theta$ should be interpreted in (6.12) as contributing a kind of Dirac delta function. (The author does not know which combination of right- and left-hand values is correct.) It, therefore, appears that χ will be discontinuous across a switching surface; see Example 10.4, where this occurs. Since χ^ε is continuous, it cannot happen that χ^ε tends to χ uniformly on N'. Thus Theorem 6.5 is not true without condition (2.2′).

Similarly, when (2.2′) holds consider $l = 3$. Then (1.11) is not correct in the sense of uniform convergence on N'. Let $\zeta = \theta_3$. The equation for ζ corresponding to (6.1), (6.10) is

$$\zeta_s + f^0 \zeta_x + \tfrac{1}{2}\theta_x H_{pp}^0 \chi_x + \tfrac{1}{6} H_{ppp}^0 \theta_x^3 + \Delta_x \chi = 0, \tag{6.13}$$

where

$$H_{ppp}^0 \theta_x^3 = \sum_{i,j,r=1}^{n} H_{p_i p_j p_r} \theta_{x_i} \theta_{x_j} \theta_{x_r}.$$

The difficulty now occurs across transition surfaces, where χ_x, H_{pp}^0 are discontinuous.

In § 7 we deal with a case where these difficulties do not arise.

Goodness of Y^0 in the stochastic problem. Suppose that, instead of seeking the optimal policy Y^ε for the stochastic problem, one simply uses the optimal deterministic policy Y^0. It is plausible that, for small ε, Y^0 should give approximately the optimum in the stochastic problem. The theorems just obtained, and their method of proof, put this rough statement on a quantitative basis.

Since Y^0 has been defined only in $\bar{N}$ (§ 3) we first extend Y^0 outside N so that $Y^0 \in \mathscr{Y}$. However, the values $Y^0(s, x)$ for $(s, x) \notin N$ have no effect on the coefficients $\theta, \tilde{\chi}$ in formula (6.14) below. Let

$$\Phi^\varepsilon(s, x) = J^\varepsilon(Y^0; s, x), \qquad (s, x) \in N.$$

In particular, $\Phi^0 = \varphi^0$. For $\varepsilon > 0$, $\Phi^\varepsilon(s, x) - \varphi^\varepsilon(s, x)$ represents how much Y^0 fails to be optimal in the stochastic problem. The function Φ^ε satisfies the linear parabolic equation

$$\Lambda^\varepsilon \Phi^\varepsilon + f^0 \Phi^\varepsilon_x + L^0 = 0 \quad \text{in } Q \tag{6.14$^\varepsilon$}$$

with $\Phi^\varepsilon = 0$ on $\partial^* Q$. The problem of expanding solutions of (6.14$^\varepsilon$) in powers of ε is easier than for the nonlinear equation (1.8$^\varepsilon$). Let us make the same assumptions as in Theorem 6.5. Then

$$\Phi^\varepsilon = \varphi^0 + \varepsilon\theta + \varepsilon^2 \tilde{\chi} + o(\varepsilon^2) \tag{6.15}$$

uniformly for $(s, x) \in N'$. The first order coefficient θ is the same as in (6.1). However, $\tilde{\chi}$ satisfies, instead of (6.10), the equation

$$\tilde{\chi}_s + f^0 \tilde{\chi}_x + \Delta_x \theta = 0 \quad \text{in } N \tag{6.16}$$

with $\tilde{\chi} = 0$ on $N \cap \partial^* Q$.

Let us indicate a proof of (6.15). Let

$$\tilde{\theta}^\varepsilon = \varepsilon^{-1}(\Phi^\varepsilon - \varphi^0), \qquad \tilde{\chi}^\varepsilon = \varepsilon^{-2}(\Phi^\varepsilon - \varphi^0 - \varepsilon\theta).$$

From (1.8^0), (6.1), (6.14$^\varepsilon$),

$$\Lambda^\varepsilon \tilde{\theta}^\varepsilon + f^0 \tilde{\theta}^\varepsilon_x + \Delta_x \varphi^0 = 0, \qquad \Lambda^\varepsilon \tilde{\chi}^\varepsilon + f^0 \tilde{\chi}^\varepsilon_x + \Delta_x \theta = 0.$$

Given $(s, x) \in N'$ consider the process defined by

$$d\xi = f^0(t, \xi(t))\, dt + (2\varepsilon)^{1/2}\, dw, \qquad s \leqq t \leqq T,$$

with $\xi(s) = x$. Corresponding to (6.3$^\varepsilon$) we have

$$\tilde{\theta}^\varepsilon(s, x) = E \int_s^{\tau_D} \Delta_x \varphi^0(t, \xi(t))\, dt + E\tilde{\theta}^\varepsilon(z_D),$$

$$\tilde{\chi}^\varepsilon(s, x) = E \int_s^{\tau_D} \Delta_x \theta(t, \xi(t))\, dt + E\tilde{\chi}^\varepsilon(z_D).$$

By a simplification of the arguments above (§§ 4–6) it can be shown that $\tilde{\theta}^\varepsilon \to \theta$, $\tilde{\chi}^\varepsilon \to \tilde{\chi}$ as $\varepsilon \to 0$, uniformly on N'. This gives (6.15).

By comparing (6.11) and (6.15) we find that

$$\Phi^\varepsilon(s, x) - \varphi^\varepsilon(s, x) = -\frac{\varepsilon^2}{2} \int_s^{\tau^0} \theta_x H^0_{pp} \theta_x\, dt + o(\varepsilon^2). \tag{6.17}$$

The integral is taken along the optimal trajectory $\gamma^0(s, x)$ with initial endpoint (s, x). Note that, from its definition (1.7), H is a concave function of p. Hence the integrand is nonpositive. Formula (6.17) shows that Y^0 gives *within order the square of the noise variance* 2ε *of the optimum if the initial data* (s, x) *belong to a region of strong regularity.*

If instead of (2.2′) we assume the weaker condition (2.2), then instead of (6.17) we can show the weaker estimate $0 \leqq \Phi^\varepsilon - \varphi^\varepsilon \leqq o(\varepsilon)$. This follows from (6.2) and the corresponding formula $\Phi^\varepsilon = \varphi^0 + \varepsilon\theta + o(\varepsilon)$.

An even simpler control choice than Y^0 in the stochastic problem would be the optimal open loop u^0 corresponding to particular initial data (s_0, x_0). Since u^0 ignores the observations of the states $\xi(t)$ obtained during system operation, u^0 may be expected to behave more poorly than Y^0 in the stochastic problem. To put this in quantitative terms, let

$$\Psi^\varepsilon(s, x) = J^\varepsilon(u^0; s, x);$$

then $\Psi^0(s_0, x_0) = \varphi^0(s_0, x_0)$, although generally $\Psi^0(s, x) > \varphi^0(s, x)$ when (s, x) is not on the optimal trajectory $\gamma^0(s_0, x_0)$. By the same proof as for (6.15) we find that

$$\Psi^\varepsilon = \Psi^0 + \varepsilon\Theta + o(\varepsilon) \tag{6.18}$$

uniformly for (s, x) in some neighborhood of $\gamma^0(s_0, x_0)$, provided $(s_0, x_0) \in N$. The function Θ is defined by

$$\Theta_s + f(s, x, u^0(s))\Theta_x + \Delta_x\Psi^0 = 0$$

with $\Theta = 0$ on $N \cap \partial^* Q$. (The second order coefficient in (6.18) can also be calculated in the same way as $\tilde{\chi}$, but is of no interest here.) Instead of (6.17) we get

$$\Psi^\varepsilon(s_0, x_0) - \varphi^\varepsilon(s_0, x_0) = \varepsilon \int_{s_0}^{\tau_0} \Delta_x(\Psi^0 - \varphi^0)\, dt + o(\varepsilon). \tag{6.19}$$

The integral is taken along $\gamma^0(s_0, x_0)$. Thus u^0 gives within order ε (rather than order ε^2) of the optimum in the stochastic problem.

We have considered the completely observable stochastic optimal control problem. It is an interesting open question to obtain approximate formulas for the optimal performance and control also for partially observable problems with small noise. In particular, one might study the open loop stochastic problem and the sampled data problem. For these problems necessary conditions for a minimum are known. See [17], [8].

7. Higher order approximations; The Cauchy problem. Let us now consider a situation where the approximations (1.11), (1.12) for φ^ε, φ_x^ε are valid for any finite l (Theorem 7.1). Later in the section this is interpreted as a result about the Cauchy problem in the theory of first order partial differential equations (Theorem 7.2).

The following assumptions are made:

$$K = R^k \qquad \text{(there are no control constraints)}; \tag{7.1}$$

$$Q = (T_0, T) \times R^n \qquad \text{(fixed stopping time } \tau = T); \tag{7.2}$$

$$f(s, x, y) = A(s, x) + B(s)y \quad \text{with } A_x \text{ bounded}; \tag{2.1′}$$

also (2.2′) and in addition to (2.3′):

$$L \geqq -c_0,$$

$$L \text{ is bounded on } Q \times \{|y| \leqq r\} \quad \text{for any } r; \tag{2.3″}$$

$$|L_x| \leqq c'L + c'' \quad \text{for suitable } c', c''.$$

By (7.1) the Pontryagin problem ($\varepsilon = 0$) is now of Lagrange type in calculus of variations. Condition (7.2) means that the parabolic equation (1.8^ε) is to be considered in the strip Q with Cauchy data $\varphi^\varepsilon(T, x) = 0$. Since the limits s and T of integration in (1.4) are now fixed, the problem is unchanged if L is replaced by $L + c_0$. Thus we may assume that $L \geqq 0$.

By Lemma 2.4, $H \in C^\infty$. We also have, if $\varepsilon > 0$ and N is a region of strong regularity,

$$\varphi^\varepsilon \in C^\infty(\bar{Q}), \quad Y^\varepsilon \in C^\infty(\bar{Q}), \quad \varphi^0 \in C^\infty(\bar{N}), \quad Y^0 \in C^\infty(\bar{N}).$$

(In the definition of § 3 we now take $m = 1$.)

Let us first verify the a priori bounds on Q:

$$|\varphi^\varepsilon| \leqq C, \qquad |\varphi^\varepsilon_x| \leqq C, \tag{7.3}$$

for suitable C. Let Y_1 be some bounded control policy (for instance, $Y_1(s, x) \equiv 0$). Since $L \geqq 0$ we have from (1.6^ε),

$$0 \leqq \varphi^\varepsilon(s, x) \leqq J^\varepsilon(s, x; Y_1),$$

and the right side is bounded. By Lemma 5.3 with $\tau' = T$,

$$\varphi^\varepsilon_x(s, x) = E \int_s^T L^\varepsilon_x W^\varepsilon \, dt,$$

where $\dot{W}^\varepsilon = A^\varepsilon_x W^\varepsilon$ by (2.1′). Hence $W^\varepsilon(t)$ is bounded. From the bound on L_x in (2.3″) and on φ^ε we get a bound on φ^ε_x. Thus (7.3) holds. (These estimates also appear in [7, § 7] and [9].) Now

$$Y^\varepsilon(s, x) = V(s, x, \varphi^\varepsilon_x(s, x)). \tag{7.4}$$

From (2.3′) and (7.3) this implies a bound $|Y^\varepsilon| \leqq r$ for some r. We get the same minimum for $J^\varepsilon(s, x; Y)$ using the compact control set $K = \{|y| \leqq r\}$ instead of R^k.

THEOREM 7.1. *Let N be a region of strong regularity, and N' with $\bar{N}' \subset N$. Assume* (7.1), (7.2), (2.1′)–(2.3′), (2.3″). *Then* (1.11), (1.12) *are valid for any l, uniformly for $(s, x) \in N'$. The coefficients θ_l, $l = 1, 2, \cdots$, in* (1.11) *satisfy linear first order partial differential equations, obtained by differentiating* (1.8^ε) l *times with respect to ε and formally setting $\varepsilon = 0$. The Cauchy data are $\theta_l(T, x) = 0$.*

This theorem can be proved by induction on l. If we let

$$\theta^\varepsilon_{l+1} = \varepsilon^{-1}(\theta^\varepsilon_l - \theta_l) = \varepsilon^{-l-1}(\varphi^\varepsilon - \varphi^0 - \varepsilon\theta_1 - \cdots - \varepsilon^l\theta_l),$$

then one shows first that $\theta^\varepsilon_{l+1} \to \theta_{l+1}$ and then $(\theta^\varepsilon_{l+1})_x \to (\theta_{l+1})_x$, in each case uniformly on N'. The reasoning is like that in § 6. The details are rather tedious, and we omit them. Since we are dealing with the Cauchy problem, the analogue of Lemma 6.3 is unnecessary. For all l,

$$\theta^\varepsilon_l = (\theta^\varepsilon_l)_x = \theta_l = (\theta_l)_x = 0 \quad \text{when } s = T.$$

Formula (1.11) gives the Taylor expansion with remainder of φ^ε about $\varepsilon = 0$. One cannot let $l \to \infty$ to obtain φ^ε as the sum of the corresponding power series in ε. The coefficients $\theta_l(s, x)$ are obtained by integrating along the optimal trajectory $\gamma^0(s, x)$ expressions computable from H and its partial derivatives. Thus $\theta_l(s, x)$ depends only on $H(s', x', p)$ for (s', x') in a neighborhood of $\gamma^0(s, x)$. However,

the solution φ^ε of the parabolic equation (1.8^ε) depends on $H(s', x', p)$ for (s', x') arbitrarily far from $\gamma^0(s, x)$.

From (1.12) we can get an approximate formula for the optimal control policy:

$$Y^\varepsilon = Y^0 + \varepsilon Z_1 + \varepsilon^2 Z_2 + \cdots + \varepsilon^l Z_l + o(\varepsilon^l) \tag{7.5}$$

uniformly on N'. The coefficients Z_l are obtained by repeatedly differentiating (7.4) with respect to ε and then setting $\varepsilon = 0$. For instance,

$$Z_1 = -L^0_{yy}\theta_x B. \tag{7.6}$$

Example 7.1 (linear regulator). Let $f = A(s)x + B(s)y$, and let L be quadratic in (x, y) with $L > 0$ unless $x = y = 0$. Then φ^0 is quadratic in x, $\theta_1 = \theta_1(s)$, $\theta_l = 0$ for $l \geqq 2$. The optimal policy Y^ε is a function linear in x, independent of ε (in particular, $Y^\varepsilon = Y^0$).

In the more general problem considered in this section, the optimal policy Y^0 can be found approximately in a neighborhood of an optimal trajectory $\gamma^0(s, x) \subset N$ by solving a linear regulator-type problem. This is done by linearizing f and expanding L to quadratic terms about $(t, \xi^0(t), u^0(t))$. A solution to the secondary minimum problem is needed for the computations. See [1].

In the linear regulator example the assumption L bounded on $Q \times \{|y| \leqq r\}$ in (2.3″) is violated. The optimal policies $Y^0(s, x)$ are linear in x, hence not bounded. However, these minor difficulties are inessential. It can be shown that Theorem 7.1 remains true if N' is assumed bounded and the boundedness assumption on L in question is replaced by certain polynomial-like growth assumptions [10, §§ 2, 5]. Under these growth assumptions, which hold in particular for the linear regulator, Y^ε satisfies a linear growth condition $|Y^\varepsilon(s, x)| \leqq C(1 + |x|)$.

We conjecture that Theorem 7.1 is true without assumption (7.2). A crucial step seems to be necessary to establish an a priori bound $|\varphi^\varepsilon_x| \leqq C$ in N, like the one in Lemma 5.4 for the case of compact control set K.

A result about partial differential equations. Consider a Cauchy problem of the type

$$\varphi^0_s + F(s, x, \varphi^0_x) = 0, \qquad T_0 < s < T, \tag{7.7^0}$$

$$\varphi^0(T, x) = 0, \tag{7.8}$$

and the corresponding problem when $\varepsilon > 0$,

$$\varphi^\varepsilon_s + \varepsilon\Delta_x\varphi^\varepsilon + F(s, x, \varphi^\varepsilon_x) = 0, \qquad T_0 < s < T, \tag{7.7$^\varepsilon$}$$

$$\varphi^\varepsilon(T, x) = 0. \tag{7.8$^\varepsilon$}$$

There is a considerable literature dealing with generalized solutions φ^0 of (7.7^0)–(7.8) and with the convergence of φ^ε to φ^0 as $\varepsilon \to 0$. See references in [20], [9]. However, to the author's knowledge, an expansion of the type (1.11) has not been proved for nonlinear F. We now obtain such an expansion by inventing a control

problem for which (7.7^0) is the Hamilton–Jacobi equation (1.8^0) and then applying Theorem 7.1. The control problem will in fact be a simple problem in calculus of variations.

Let us make the following assumptions about F (see [9, § 2]):

$$F \in C^\infty, \quad F(s,x,p) \leqq C, \quad |F(s,x,0)| \leqq C_0, \tag{7.9a}$$

$$\lim_{|p|\to\infty} \frac{F(s,x,p)}{|p|} = -\infty;$$

$$|F_x| \leqq c'(F - pF_p) + c''; \tag{7.9b}$$

(7.9c) $|F_p| \leqq R(|p|)$. For each $r > 0$ there exists v_r such that $|F_p| \leqq r$ implies $|p| \leqq v_r$.

$$C_1(|p|)|v|^2 \leqq -vF_{pp}v \leqq C_2(|p|)|v|^2 \quad \text{for all } v \in R^n, \tag{7.9d}$$

where $C_1(v)$, $C_2(v)$ are positive and respectively nonincreasing, nondecreasing in v.

For a control problem we take the simplest dynamical equations and no control constraints:

$$\frac{d\xi}{dt} = u(t), \qquad K = R^n.$$

Let L be dual to F in the sense of the classical canonical transformation:

$$L(s,x,y) = \max_{p\in R^n} (F - yp). \tag{7.10}$$

Then

$$F(s,x,p) = \min_{y\in R^n} (L + yp) \tag{7.11}$$

and (7.7$^\varepsilon$) becomes (1.8$^\varepsilon$) with $H = F$, for $\varepsilon \geqq 0$.

Assumption (2.1′) holds with $A = 0$, $B =$ identity. Now $L_{yy} = -(F_{pp})^{-1}$, where y and p are related by the (classical) formulas $y = F_p$, $p = -L_y$ according to (7.10) and (7.11). By assumptions (c) and (d), (2.2′) is satisfied with $c(r) = C_2(v_r)^{-1}$. Finally, assumptions (a), (b), (c) imply (2.3′), (2.3″).

We now take

$$\varphi^0(s,x) = \min \int_s^T L(t, \xi(t), \dot{\xi}(t))\, dt, \tag{7.12}$$

the minimum being taken among all Lipschitz ξ on $[s, T]$ with $\xi(s) = x$. It is known that the unique bounded solution $\varphi^\varepsilon \in C^{1,2}(\bar{Q})$ of (7.7$^\varepsilon$)–(7.8$^\varepsilon$) tends uniformly to φ^0 as $\varepsilon \to 0$ [9, p. 527]. Moreover, φ^0 satisfies (7.7^0) almost everywhere [7, p. 271].

Consider N such that:

$$N = \mathcal{O}_0 \cap \bar{Q}, \qquad \mathcal{O}_0 \text{ open}; \tag{7.13a}$$

$$\varphi^0 \in C^\infty(\bar{N}); \tag{7.13b}$$

$$(s,x) \in N \Rightarrow \gamma^0(s,x) \subset N. \tag{7.13c}$$

Then N is a region of strong regularity. The uniqueness of the optimal trajectory $\gamma^0(s, x)$ was proved in [9, p. 520] using an argument of Kuznetzov and Shishkin. The remaining conditions in the definition of § 3 automatically hold with $m = 1$, $N_1 = N$, Σ_1 the hypersurface $\{T\} \times R^n$.

A given optimal $\gamma^0(s, x)$ is contained in some such N provided (s, x) is neither a conjugate point nor an irregular point [9, p. 521]; also see the Appendix.

From Theorem 7.1 we then have the following theorem.

THEOREM 7.2. *Let F satisfy* (7.9), *N satisfy* (7.13), *and $\bar{N}' \subset N$. Then the conclusions of Theorem* 7.1 *hold (with $H = F$).*

In (7.13) we assumed that $\gamma^0(s, x)$ gives an absolute minimum, not merely a relative minimum. The method of characteristics is a classical method for constructing a smooth solution φ of (7.7^0) in some N. (The characteristic ground curves are solutions of Euler's equation.) In this construction the characteristic ground curve $\gamma(s, x)$ with initial point (s, x) gives a minimum among curves lying in N, but not always an absolute minimum. If the minimum is not absolute, then φ^ε tends in N not to φ but rather to φ^0.

8. Discontinuous control policies. Let us now impose the convexity condition (2.2) on L instead of the stronger condition (2.2′). For instance, (2.2) but not (2.2′) is satisfied in the time-optimal problem ($L \equiv 1$). One can no longer expect an optimal policy in the class $\mathscr{Y}$ defined in § 4. Simple examples show that the optimal Y^ε may have discontinuities.

The purpose of this section is to give a correct formulation of the control problem in this case, and to outline the modifications needed in the proof of Theorem 6.2. (As noted at the end of the proof of Theorem 6.5, that theorem is incorrect without assumption (2.2′).) As before we assume that the control set K is compact. Let $\mathscr{Y}'$ consist of all Borel measurable Y from $[T_0, T] \times R^n$ into K. For such Y, the standard (Ito) theory of stochastic differential equations no longer guarantees a solution ξ to the system equations

$$d\xi = f^Y\, dt + (2\varepsilon)^{1/2}\, dw, \tag{1.1}$$

since the drift coefficient f^Y is merely bounded and Borel measurable. However, for $\varepsilon > 0$ the noise coefficient matrix in (1.1) is nonsingular; and we may use instead of the standard theory the work of Stroock and Varadhan [24].

Let b be bounded, Borel measurable, with values in R^n; and take Ω as in § 4. According to [24], a continuous nonanticipative process ξ is a solution of the stochastic differential equations

$$d\xi = b(t, \xi(t))\, dt + (2\varepsilon)^{1/2}\, dw, \qquad s \leqq t \leqq T, \tag{8.1}$$

with $\xi(s) = x$ if, for each $\alpha \in R^n$,

$$\exp\left[\alpha\left(\xi(t) - x - \int_s^t b(r, \xi(r))dr\right) - |\alpha|^2\varepsilon(t - s)\right]$$

is a martingale with respect to the probability measure Π on Ω associated with the process ξ. For any $Y \in \mathscr{Y}'$, this process ξ exists and is unique in the sense that Π is unique [24, § 6]. Moreover, ξ is a Markov process with the strong Feller property.

The process

$$w(t) = (2\varepsilon)^{-1/2}\left[\xi(t) - x - \int_s^t b(r, \xi(r))\, dr\right]$$

is a Brownian motion [24, Corollary 3.2]. Thus ξ is a solution of (1.1) in the ordinary sense. In particular, we may take $b = f^Y$, $f \in \mathscr{Y}'$.

Any $Y \in \mathscr{Y}'$ can be approximated almost everywhere by a sequence $Y_1, Y_2, \cdots$ in $\mathscr{Y}$. It follows using facts about parabolic partial differential equations that

$$\lim_{j \to \infty} J^\varepsilon(s, x; Y_j) = J^\varepsilon(s, x; Y).$$

See [8, esp. Appendix]. Hence the φ^ε in (1.6$^\varepsilon$) also satisfies

$$\varphi^\varepsilon(s, x) = \min_{Y \in \mathscr{Y}'} J^\varepsilon(s, x; Y).$$

As before φ^ε is a solution in $C^{1,2}(Q)$ of (1.8$^\varepsilon$)–(1.9). An optimal $Y^\varepsilon \in \mathscr{Y}'$ is determined, for almost all $(s, x) \in Q$, by (1.10).

The following modifications in the proof of Theorem 6.2 are needed. First of all, the only results in §§ 4 and 5 which must be restated are Lemma 4.1 and Corollary 5.6. For the latter, we can only assert that Y^ε tends to Y^0 uniformly on any compact subsets of $N - \Sigma$, where Σ is the union of those Σ_j which are switching surfaces (§ 3). (It is an interesting open problem to prove statements about the perturbation of the optimal switching surfaces.) In applying Lemma 4.1 to prove Lemma 4.2, $f^0 \in C^\infty(\mathscr{O})$. However, in the proof of Theorem 6.2, f^0 is discontinuous at points of Σ. Lemma 4.1 must hence be modified to require $|f^0 - b| < e$ except at those (s, x) with distance $((s, x), \Sigma) < e$.

Let $\mathscr{F}(D)$ denote the class of all ψ such that ψ, ψ_x are continuous and bounded on $\bar{D}$ while $\psi_s, \psi_{x_i x_j}$, $i, j = 1, \cdots, n$, are square integrable on compact subsets of D. We need formula (4.4) when b is bounded, Borel measurable, $\psi \in \mathscr{F}(D)$ and $\Lambda^\varepsilon \psi + b\psi_x$ is bounded on D. As in § 4(e), this can be reduced to (4.4′) with $\tilde{\psi} \in \mathscr{F}(R^{n+1})$ and $G = \Lambda^\varepsilon \tilde{\psi} + b\tilde{\psi}_x$ bounded. By standard smoothing methods there exist $\tilde{\psi}^j \in C^{1,2}$ such that $\tilde{\psi}^j, \tilde{\psi}^j_x$ tend uniformly to $\tilde{\psi}, \tilde{\psi}_x$ as $j \to \infty$, while

$$\Lambda^\varepsilon \tilde{\psi}^j + b\tilde{\psi}^j_x = G^j$$

is uniformly bounded and tends almost everywhere to $\Lambda^\varepsilon \tilde{\psi} + b\tilde{\psi}_x = G$. By the Ito differential rule, (4.4′) is correct for each $\tilde{\psi}^j$ [12, p. 390]. From the fact that the random variable $\xi(r)$ has probability density absolutely continuous with respect to Lebesgue measure [24], $G^j(r, \xi(r))$ tends to $G(r, \xi(r))$ with probability 1 for almost all $r \in [s, T]$. It follows that (4.4′) holds for $\tilde{\psi}$.

We applied (4.4′) in proving Lemma 5.1, also in Lemma 5.3 ($\psi = P_i$) and Theorem 6.2 ($\psi = \theta^\varepsilon$). By differentiating (1.8$^\varepsilon$) with respect to x_i and using Lemma 2.2(a) together with results about parabolic equations [8, Appendix] or [18], $P_i \in \mathscr{F}(Q)$. Moreover, $\varphi^0 \in \mathscr{F}(N)$ since N is a region of strong regularity. Hence each of these $\psi \in \mathscr{F}(N)$. In the first two instances, $b = f^\varepsilon$; in the third, $b = b^\varepsilon$ is defined using Lemma 6.1(a) rather than Lemma 6.1(b).

9. Autonomous problems. Consider the autonomous system

$$d\xi = f(\xi, u)\,dt + (2\varepsilon)^{1/2}\,dw, \qquad t \geqq 0, \tag{9.1}$$

with initial data $\xi(0) = x$. We suppose that $x \in B$, where $B \subset R^n$ is open, bounded, and ∂B is a C^∞ manifold. Let $\mathscr{Y}_a$ be the class of Lipschitz functions Y from R^n into the control set K. Such Y are called autonomous control policies. Let

$$J^\varepsilon(Y; x) = E\int_0^{\tau_B} L[\xi(t), u(t)]\,dt,$$

$$u(t) = Y[\xi(t)],$$

$$\varphi^\varepsilon(x) = \inf_{Y \in \mathscr{Y}_a} J^\varepsilon(Y; x). \tag{9.2}$$

The autonomous version of (1.8^ε) is the elliptic equation

$$\Delta\varphi^\varepsilon + H(x, \varphi_x^\varepsilon) = 0 \quad \text{in } B \tag{9.3$^\varepsilon$}$$

with $\varphi^\varepsilon(x) = 0$ on ∂B.

We are concerned with the analogues of the theorems in § 6. By an *autonomous region of strong regularity* we mean a set $N = \bar{B} \cap \mathscr{O}$, where $\mathscr{O}$ is open, $\mathscr{O} \subset R^n$, and the autonomous versions of (a)–(g) in Definition 3.1 hold. Now $\Sigma_1, \cdots, \Sigma_m$ are hypersurfaces in R^n, $\gamma^0(x) = \{\xi^0(t)\colon 0 \leqq t \leqq \tau^0(x)\}$ is the unique optimal trajectory with initial point x for the problem $J^0(u; x) = \min$. Here $\tau^0 = \tau^0(x)$, $\xi^0(\tau^0)$ are the exit time and place from B for $\gamma^0(x)$.

Let $Y \in \mathscr{Y}_a$ be a policy which agrees on N with Y^0, and let $\tilde{\psi}(x) = J^\varepsilon(Y; x)$. We assume that Y can be chosen so that $\tilde{\psi}(x) \leqq C\varepsilon^{-1}$ for all $x \in B$. Then the autonomous versions of the theorems in § 6 hold. For instance, let N be an autonomous region of strong regularity, and assume (2.1), (2.2′), (2.3). Then

$$\varphi^\varepsilon(x) = \varphi^0(x) + \varepsilon\theta(x) + \varepsilon^2\chi(x) + o(\varepsilon^2) \tag{9.4}$$

uniformly on N', if $\bar{N}' \subset N$, where θ, χ are defined by the autonomous versions of (6.1), (6.10). If we assume only (2.2), then the autonomous version of Theorem 6.2 holds. As in § 8 the optimal policy Y^ε in that case belongs to the class of bounded measurable functions from R^n into K.

In an autonomous problem no bound is imposed on the exit time τ_B. This necessitates a few changes in the proofs. It may not be true that $\tilde{\psi}$ is bounded on B independent of ε. However, let $\bar{D} \subset N$. Then there is such a bound on $\bar{D}$ as follows. Since φ^0 is continuous on the compact set $\bar{N}$, φ^0 is bounded there. By (2.3), $\tau^0(x) \leqq c_1^{-1}\varphi^0(x)$. Take T large enough that

$$T > \tau^0(x) + 1 \quad \text{for all } x \in N.$$

Let $\tau' = \tau_N \wedge T$, where τ_N is the exit time for ξ satisfying $d\xi = f^Y\,dt + (2\varepsilon)^{1/2}\,dw$, $\xi(0) = x$. Then

$$\tilde{\psi}(x) = E\int_0^{\tau'} L^0\,dt + E\tilde{\psi}[\xi(\tau')], \qquad x \in N.$$

The first term on the right side is bounded. By assumption, $\tilde{\psi} \leqq C\varepsilon^{-1}$ in B. Since $\tilde{\psi} = 0$ on ∂B,

$$E\tilde{\psi}[\xi(\tau')] \leqq C\varepsilon^{-1}\,\Pr\{\xi(\tau') \notin N \cap \partial B\}.$$

There exists $e > 0$ such that $x \in \bar{D}$ and $\varepsilon^{1/2}\|w\|_T < e$ imply $\tau' = \tau_N < T$ and $\xi(\tau') \in N \cap \partial B$. From (4.3), $E\tilde{\psi}[\xi(\tau')]$ is then bounded, which implies $\tilde{\psi} \leqq C$ on $\bar{D}$. Since $0 \leqq \varphi^\varepsilon \leqq \tilde{\psi}$, this also bounds φ^ε on $\bar{D}$.

After that few further changes are needed. For Q we take the cylinder

$$Q = (0, T) \times B.$$

In Part 2 of the proof of Lemma 4.4 we now have

$$E \int_0^{\tau_Q^\varepsilon} L[\xi^\varepsilon(t), u^\varepsilon(t)]\, dt \leqq \varphi^\varepsilon(x)$$

rather than equality. In all proofs in §§ 5, 6 where some exit time τ_D appears we now take $\tau' = \tau_D \wedge T$ rather than $\tau' = \tau_D$. Thus all limiting arguments as $\varepsilon \to 0$ are carried out for τ' and t in the fixed interval $[0, T]$.

10. Examples. We have found rather good approximate formulas for $\varphi^\varepsilon(s, x)$ in terms of integrals along the optimal trajectory $\gamma^0(s, x)$, provided $\gamma^0(s, x)$ lies in some region of strong regularity N. It is an essentially classical result (stated in the Appendix) that such an N exists if (s, x) is regular and nonconjugate. Let us now show by examples that the approximate formulas (1.11), (1.12) are not generally correct near points which are irregular or conjugate. From formula (A.9) in the Appendix it is clear that $\Delta_x\varphi^0$ misbehaves at any conjugate point where $\partial p/\partial\alpha$ is nonsingular. We consider first an example of the simplest type in calculus of variations.

Example 10.1. Let

$$J = \int_s^0 \frac{1}{2}\dot{\xi}(t)^2\, dt + \Phi[\xi(0)],$$
$$\Phi(x) = -(x^2 + 1)^{1/2}.$$

This can be put in the form in § 7 with $n = k = 1$, $T = 0, f(y) = y$, $L = \frac{1}{2}y^2 + \Phi'(x)y$. The extremals are linear functions. Let

$$\xi(t, \alpha) = \alpha - t\Phi'(\alpha).$$

Let $s(\alpha) = -(\alpha^2 + 1)^{1/2}$, which implies $\xi(s(\alpha), \alpha) = 0$. The line segment

$$\gamma(\alpha) = \{(t, \xi(t, \alpha)) : s(\alpha) \leqq t \leqq 0\}$$

is an optimal trajectory. For any $\delta > 0$ the set

$$N = \{(s, x) : s \leqq 0, |x| > \delta \text{ or } -1 + \delta < s\}$$

is a region of strong regularity.

The point $(-1, 0)$ is the initial endpoint of the horizontal segment $\gamma(0)$. Since $\gamma(0)$ is the unique optimal trajectory with this initial endpoint, $(-1, 0)$ is a regular point. Since $\partial\xi/\partial\alpha = 0$ when $\alpha = 0$, $t = -1$, $(-1, 0)$ is a conjugate point. (For $\alpha \neq 0$ the solution $t(\alpha)$ of the equation $\partial\xi/\partial\alpha = 0$ also gives in the classical terminology a conjugate point. However, $t(\alpha) < s(\alpha)$ for $\alpha \neq 0$. Since we use only the part $\gamma(\alpha)$ of the line $\xi = \xi(t, \alpha)$ which minimizes (globally among all curves with the same initial point) these other conjugate points are of no interest here.) Let

$$p(t, \alpha) = \Phi'(\alpha) - \Phi'(\xi(t, \alpha)).$$

From (A.9), for $-1 < t \leqq 0$,

$$\varphi^0_{xx}(t, \xi(t, \alpha)) = \Phi''(\alpha)\left(\frac{\partial \xi}{\partial \alpha}\right)^{-1} - \Phi''(\xi(t, \alpha)),$$

$$\varphi^0_{xx}(t, 0) = -(1 + t)^{-1} + 1,$$

$$\theta(s, 0) = \int_s^0 [-(1 + t)^{-1} + 1]\, dt = \log(1 + s) - s,$$

$$\varphi^\varepsilon(s, 0) = \varphi^0(s, 0) + \varepsilon\theta(s, 0) + o(\varepsilon), \qquad -1 < s \leqq 0, \tag{10.1}$$

the last equation being just (6.3). By (10.1) no estimate of the type $|\varphi^\varepsilon - \varphi^0| \leqq C\varepsilon$ can hold uniformly in a neighborhood of $(-1, 0)$.

For $s < -1$, two minimizing trajectories $\gamma(\alpha), \gamma(-\alpha)$ issue from the point $(s, 0)$. The partial derivative φ^0_x has a jump discontinuity across the half-line consisting of such $(s, 0)$. This introduces difficulties with (1.11) and (1.12) which Example 10.2 will illustrate.

Example 10.1 is of a type studied by E. Hopf [14] in a pioneering paper on the small viscosity method for studying nonlinear conservation laws. The substitution $v^\varepsilon = \varphi^\varepsilon_x$ in the present instance turns (7.7^ε) into Burgers' equation with Cauchy data $v^\varepsilon = \Phi'$ when $s = 0$. The same substitution generally converts (7.7^ε) when $n = 1$ into

$$v^\varepsilon_s + \varepsilon v^\varepsilon_{xx} + F(s, x, v^\varepsilon)_x = 0.$$

For various results on the behavior of v^ε as $\varepsilon \to 0$ see [20]. Formula (1.12) gives an expansion of v^ε valid in any N such that the limit v^0 of v^ε as $\varepsilon \to 0$ belongs to $C^\infty(\bar{N})$ and

$$v^0(s, x) = \int_{\gamma^0(s,x)} F_x\, dt + v^0(T, \xi^0(T)), \qquad (s, x) \in N,$$

where the characteristic $\gamma^0(s, x) = \{(t, \xi^0(t)) : s \leqq t \leqq T\}$ is contained in N.

Example 10.2. Consider the following very simple autonomous, 1-dimensional, minimum time problem. Let

$$d\xi = u(t)\, dt + (2\varepsilon)^{1/2}\, dw,$$

with $\xi(0) = x$,

$$-1 \leqq x \leqq 1, \qquad -1 \leqq u(t) \leqq 1.$$

By straightforward calculations using (9.3^ε), the minimum expected time $\varphi^\varepsilon(x)$ to reach -1 or 1 from x is

$$\varphi^\varepsilon(x) = \varphi^0(x) + \varepsilon[\exp(-|x|/\varepsilon) - \exp(-1/\varepsilon)],$$

$$\varphi^{\varepsilon'}(x) = \varphi^{0'}(x) - \operatorname{sgn} x \exp(-|x|/\varepsilon),$$

where $\varphi^0(x) = 1 - |x|$ is the minimum time for the deterministic problem. All coefficients $\theta_j(x)$ in (1.11) are 0 for $x \neq 0$. Formally, $-(\varphi^0)''$ is a Dirac delta function, which according to (6.3) suggests the value $\theta_1(0) = 1$. This makes the formula $\varphi^\varepsilon(0) = \varphi^0(0) + \varepsilon\theta_1(0) + o(\varepsilon)$ correct. Of course, no formula $\varphi^\varepsilon(x) = \varphi^0(x) + \varepsilon\theta_1(x) + o(\varepsilon)$ holds uniformly in a neighborhood of 0.

Example 10.2′. Let us extend Example 10.2 by taking $x, \xi(t), u(t)$ vectors in R^n, $n \geqq 2$, with $|\xi(t)| \leqq 1$, $|u(t)| \leqq 1$. Then B is the spherical ball $|x| \leqq 1$ in R^n. It is easily seen that

$$Y^\varepsilon(x) = |x|^{-1}x \quad \text{if } x \neq 0,$$
$$\varphi^\varepsilon(x) = g^\varepsilon(|x|), \quad \text{all } x \in B,$$

where $g^\varepsilon(r)$ is decreasing on $0 \leqq r \leqq 1$, $g^\varepsilon(1) = 0$, and by (9.3^ε),

$$\varepsilon[(g^\varepsilon)'' + \frac{n-1}{r}(g^\varepsilon)'] + (g^\varepsilon)' + 1 = 0. \tag{10.2}$$

Since φ^ε_x is continuous at 0, $(g^\varepsilon)'(0) = 0$. For $0 < r_0 < 1$, the set

$$N = \{x : r_0 < |x| \leqq 1\}$$

is a region of strong regularity. The point 0 is both conjugate and irregular. In N the expansion (1.11) holds with

$$\varphi^0(x) = g^0(|x|) = 1 - |x|, \qquad \theta_j(x) = h_j(|x|),$$
$$h_j' + h_{j-1}'' + \frac{n-1}{r}h_{j-1}' = 0, \qquad h_j(1) = 0,$$
$$h_0(r) = g^0(r) = 1 - r,$$
$$h_1(r) = (n-1)\log r, \qquad h_2(r) = (n-1)(n-2)(r^{-1} - 1).$$

For $r = 0$ we proceed differently. Using the integrating factor $r^{n-1}e^{-r/\varepsilon}$ to solve (10.2), we get

$$g^\varepsilon(0) = \int_0^1 r^{1-n}e^{-r/\varepsilon}\,dr \int_0^r \rho^{n-1}e^{\rho/\varepsilon}\,d\rho.$$

By using integrations by parts on the inner integral we get the approximation

$$g^\varepsilon(0) = g^0(0) + (n-1)\varepsilon\log\varepsilon + c\varepsilon + o(\varepsilon)$$

for suitable c.

Example 10.3. In this example Theorem 6.4 gives some information about transition boundaries for the optimal control policy Y^ε. Let $n = k = 1$ (scalar state and control), $K = [-1, 1]$,

$$f = Ax + By, \quad B > 0, \qquad L = \tfrac{1}{2}Mx^2 + \tfrac{1}{2}y^2.$$

Take $Q = (T_0, T) \times R^n$, the Cauchy problem. Except for the control constraint $|u(t)| \leqq 1$ this would be a linear regulator problem. For $\varepsilon = 0$ the solution is as follows. There are transition curves $x = \pm G^0(s)$ such that:

$$Y^0(s,x) = \begin{cases} -1 & \text{for } x > G^0(s), \\ -g(s)x & \text{for } -G^0(s) < x < G^0(s), \\ 1 & \text{for } x < -G^0(s). \end{cases}$$

In the central region $-G^0(s) < x < G^0(s)$, $\varphi^0(s, x) = \frac{1}{2}g(s)x^2$ with $g(s) > 0$ for $s < T$. The solution coincides there with that of the linear regulator. Obviously $G^0(s) = g(s)^{-1}$. The second order, but not third order, partial derivatives of φ^0 are continuous across the transition curves. There are no conjugate or irregular points. The formulas (6.7), (6.11) hold everywhere in $\bar{Q}$.

In the stochastic problem,

$$Y^\varepsilon(s, x) = -\text{sat}\,[B\varphi_x^\varepsilon(s, x)],$$

where $\text{sat}\,a = a$ if $|a| \leqq 1$, $\text{sat}\,a = a|a|^{-1}$ if $|a| > 1$. From linearity of f, strict convexity of L in x, y and the optimality of Y^ε among nonanticipative controls, it can be seen that φ^ε is a strictly convex function of x. Thus φ_x^ε is a strictly increasing function of x; and φ_x^ε tends uniformly to φ_x^0 as $\varepsilon \to 0$. Therefore

$$\begin{aligned} Y^\varepsilon(s, x) &= -1 \quad \text{for } x > G^\varepsilon(s), \\ -1 < Y^\varepsilon(s, x) &< 1 \quad \text{for } -G^\varepsilon(s) < x < G^\varepsilon(s), \\ Y^\varepsilon(s, x) &= 1 \quad \text{for } x < -G^\varepsilon(s), \end{aligned}$$

where $G^\varepsilon(s) \to G^0(s)$ as $\varepsilon \to 0$. Since $\varphi_x^\varepsilon(s, G^\varepsilon(s)) = \varphi_x^0(s, G^0(s)) = B^{-1}$, we have

$$0 = \varepsilon^{-1}[\varphi_x^\varepsilon(s, G^\varepsilon) - \varphi_x^0(s, G^\varepsilon)] + \varepsilon^{-1}[\varphi_x^0(s, G^\varepsilon) - \varphi_x^0(s, G^0)].$$

Using Theorem 6.4 we get as $\varepsilon \to 0$,

$$\theta_x + \varphi_{xx}^0 \frac{\partial G^\varepsilon}{\partial \varepsilon}\bigg|_{\varepsilon = 0} = 0$$

at $(s, G^0(s))$. Since $\varphi_{xx}^0(s, G^0(s)) = g(s) > 0$, the derivative $\partial G^\varepsilon/\partial\varepsilon$ exists at $\varepsilon = 0$. However, $\theta = \theta(s)$ in the central region (see linear regulator example in § 7). Since θ is C^1 across the transition curves, $\theta_x(s, G^0(s)) = 0$. We have shown that in this example,

$$\frac{\partial G^\varepsilon}{\partial \varepsilon} = 0 \quad \text{when } \varepsilon = 0.$$

A multidimensional steady state version of this example was studied in [25], where a computational method for generating reasonably good suboptimal controls was found.

We do not know how to apply the method of Example 10.3 to study perturbation of optimal switching surfaces. Both θ_x and φ_{xx}^0 are discontinuous across a switching surface for Y^0. It is not clear what combination of right- and left-hand values is needed. A related problem is to determine when the switching surface for Y^ε perturbs toward the terminal surface Σ_1 and when away from Σ_1 (notation in § 3).

In the following one-dimensional autonomous example, the optimal switching point can be found by direct calculation. Some two-dimensional examples have been studied in [4], [23].

Example 10.4. Consider the autonomous system

$$d\xi = u(t)\,dt + (2\varepsilon)^{1/2}\,dw, \qquad \xi(0) = x,$$

with $0 < a \leqq u(t) \leqq 1$, $-b < x < 0$. The problem is to maximize (rather than minimize)

$$J = E \int_0^\tau [\xi(t) + l(u(t))]\, dt,$$

where τ is the exit time from $B = (-b, 0)$ for $\xi(t)$. Let $l_a = l(a), l_1 = l(1)$. We choose a, b, l such that $l(u) = c_1 u + c_2$ is linear and

$$0 < al_1 < l_a, \qquad -(1 - a)b < al_1 - l_a.$$

Equation (9.3$^\varepsilon$) becomes

$$\varepsilon(\varphi^\varepsilon)'' + \max_{a \leqq u \leqq 1} [l(u) + (\varphi^\varepsilon)'u] + x = 0$$

with $\varphi^\varepsilon(-b) = \varphi^\varepsilon(0) = 0$. (For $\varepsilon = 0$ no boundary condition is imposed at $-b$.) Define x^0 by

$$(1 - a)x^0 = al_1 - l_a.$$

Then x^0 is the optimal switching point in the deterministic problem:

$$Y^0(x) = \begin{cases} 1 & \text{if } x < x^0, \\ a & \text{if } x^0 < x \leqq 0. \end{cases}$$

Any interval $(x_1, 0]$ with $-b < x_1$ is a region of strong regularity. When $\varepsilon > 0$ the switching point x^ε is determined from the equation

$$l_a + (\varphi^\varepsilon)'a = l_1 + (\varphi^\varepsilon)'.$$

By straightforward calculations we find that

$$x^\varepsilon = x^0 + \varepsilon + O(\exp(-\beta/\varepsilon))$$

for some $\beta > 0$. In this example $\Sigma_1 = \{0\}, \Sigma_2 = \{x^0\}$; and x^ε perturbs toward Σ_1.

Appendix. We outline the method of characteristics for the Hamilton–Jacobi equation (1.8^0), and how quantities needed in the asymptotic formulas (1.11), (1.12) can be computed from it.

The characteristic equations for (1.8^0) form the following system of $2n$ ordinary differential equations for functions $\xi(t), p(t)$:

$$\frac{d\xi}{dt} = H_p, \qquad \frac{dp}{dt} = -H_x.$$

See [2, Chap. II]. Define the control function $u(\cdot)$ by

(A.1) $$L(t, \xi(t), y) + p(t)f(t, \xi(t), y) = \min \quad \text{on } K \quad \text{when } y = u(t).$$

By (2.8) the characteristic equations become

(A.2) $$\frac{d\xi}{dt} = f, \qquad \frac{dp}{dt} = -pf_x - L_x.$$

Equations (A.2) are the state and "costate" equations, and (A.1) the minimum condition in Pontryagin's principle [13], [22]. Suppose that (A.1), (A.2) hold for $\bar{\tau} \leqq t \leqq \tau$ with $z = (\tau, \xi(\tau)) \in \Sigma_1$, whereas in § 3, Σ_1 is a C^∞ hypersurface, $\Sigma_1 \subset \partial Q$.

We impose the transversality condition:

(A.3) For $t = \tau, (L + pf, -p)$ is a normal vector to Σ_1 at z.

The trajectory $\gamma = \{(t, \xi(t)) : \bar{\tau} \leqq t \leqq \tau\}$ is called a *characteristic ground curve.*

Suppose now that a triple $\xi(t, \alpha), p(t, \alpha), u(t, \alpha)$ has been constructed for each vector $\alpha = (\alpha_1, \cdots, \alpha_n)$ in the closure $\overline{\mathfrak{U}}$ of some open set $\mathfrak{U}$ and for $\bar{\tau}(\alpha) \leqq t \leqq \tau(\alpha)$. Moreover, suppose functions $\tau_j(\alpha)$, $\xi_j(\alpha)$ have been found for $j = 2, \cdots, m$, such that the following hold:

A1. $$\tau_1(\alpha) = \tau(\alpha), \qquad \tau_{m+1}(\alpha) = \bar{\tau}(\alpha),$$
$$\tau_j(\alpha) > \tau_{j+1}(\alpha), \qquad j = 1, \cdots, m, \quad \alpha \in \overline{\mathfrak{U}}.$$

A2. Let $\mathscr{B}_j = \{(t, \alpha) : \tau_{j+1}(\alpha) < t < \tau_j(\alpha), \alpha \in \mathfrak{U}\}$. Then $u \in C^\infty(\overline{\mathscr{B}}_j), j = 1, \cdots, m$. Moreover, (A.1)–(A.3) hold.

A3. If $\tau_{m+1}(\alpha) < s < \tau_1(\alpha), \alpha \in \mathfrak{U}$, then $u(\cdot, \alpha)$ is the unique optimal control for the initial data $(s, \xi(s, \alpha))$.

A4. There exist C^∞ hypersurfaces $\Sigma_2, \cdots, \Sigma_m$ such that the characteristic ground curve $\gamma(\alpha)$ meets Σ_j nontangentially at the single point $(\tau_j(\alpha), \xi_j(\alpha))$, $j = 1, \cdots, m, \alpha \in \overline{\mathfrak{U}}$.

A5. The mapping $(t, \alpha) \to (t, \xi(t, \alpha))$ is one-to-one and the matrices $\partial\xi/\partial\alpha$ are nonsingular for $\tau_{m+1}(\alpha) \leqq t \leqq \tau_1(\alpha), \alpha \in \overline{\mathfrak{U}}$.

A6. For $t \neq \tau_j(\alpha), j = 2, \cdots, m$:

(a) the minimum in (A.1) is attained only for $y = u(t, \alpha)$;

(b) the function V in Lemma 2.3 belongs to $C^\infty(\Gamma)$ for some open set Γ containing $(t, \xi(t, \alpha), p(t, \alpha))$.

We allow u (and hence $\dot{\xi} = \partial\xi/\partial t$) to be discontinuous when $t = \tau_j(\alpha)$, $j = 2, \cdots, m$. However, ξ and p are continuous there.

By the transversality condition (A.3) and (2.3), condition A4 holds when $t = \tau_1$. If α is a set of local coordinates for Σ_1, then condition A5 holds when $t = \tau_1$. Let

$$N = \{(t, \xi(t, \alpha)) : \tau_{m+1}(\alpha) < t \leqq \tau_1(\alpha), \alpha \in \mathfrak{U}\},$$
$$N_j = \{(t, \xi(t, \alpha)) : (t, \alpha) \in \mathscr{B}_j\}, \qquad j = 1, \cdots, m.$$

Then N is a region of strong regularity (Definition 3.1). We have

(A.4) $$\varphi^0(t, \xi(t, \alpha)) = \int_t^{\tau_1(\alpha)} L(r, \xi(r, \alpha), u(r, \alpha))\, dr,$$

(A.5) $$\varphi^0_x(t, \xi(t, \alpha)) = p(t, \alpha),$$

(A.6) $$V(t, \xi(t, \alpha), p(t, \alpha)) = u(t, \alpha).$$

The following result gives a sufficient condition that a given optimal trajectory lie in some region of strong regularity N. Suppose that u^0 is optimal for initial data (s, x) with corresponding $\xi^0, p^0, \gamma^0 = \gamma^0(s, x)$ such that the necessary conditions (A.1)–(A.3) for a minimum hold. Suppose also that a triple ξ, p, u satisfying A1, A2, A4, A6 has been constructed, coinciding with ξ^0, p^0, u^0 for $\alpha = \alpha^0$ and $s \leqq t \leqq \tau^0 = \tau_1(\alpha^0)$. Here we suppose that $\bar{\tau}(\alpha^0) < s < \tau_1(\alpha^0)$. Following classical terminology we say that $(t, \xi^0(t))$ is a *conjugate point* if the matrix $\partial\xi/\partial\alpha$ is singular at (t, α^0).

Besides the general assumptions in § 2, let us assume (as in previous sections) either that K is compact or the conditions in § 7 hold.

THEOREM A.1. *If (s, x) is regular and not conjugate, then there exists a region of strong regularity N containing $\gamma^0(s, x)$.*

The proof uses classical reasoning in calculus of variations together with Lemma 3.1. See [9, Theorem 2]. In fact, one can take

$$N = \{(t, \xi(t, \alpha)) : \tau < t \leqq \tau_1(\alpha), \alpha \in \mathfrak{U}_1\},$$

where $\mathfrak{U}_1$ is some neighborhood of α^0 and $\bar{\tau}(\alpha^0) < \tau < s$, τ sufficiently near s.

Frequently, the required triple ξ, p, u is constructed in a neighborhood of α^0 as follows. Let α be a set of local coordinates for Σ_1 near $z^0 = (\tau^0, \xi^0(\tau^0))$; points of Σ_1 near z^0 are of the form $(\tau_1(\alpha), \xi_1(\alpha))$. By Lemma 2.2, $(L_u + pf_u)V_p = 0$. Using this fact and the implicit function theorem, $u(\tau_1(\alpha), \alpha), p(\tau_1(\alpha), \alpha)$ are determined near α^0 by (A.3), (A.6) and $u^0(\tau^0)$. Equations (A.2), (A.6) then determine $\xi(t, \alpha)$, $p(t, \alpha)$, $u(t, \alpha)$. Let Δ denote a set of (s, x, p) such that $V \in C^\infty(R^{2n+1} - \Delta)$; recall the examples in §§ 2, 10. Let us suppose that $(t, \xi^0(t), p^0(t)) \in \Delta$ for $t = s_j, j = 2, \cdots, m$. Moreover, suppose that, in a neighborhood of $(s_j, \xi^0(s_j), p^0(s_j))$, Δ is described by an equation $U_j(s, x, p) = 0$ with $U_j \in C^\infty$ and

$$\frac{d}{dt} U_j(t, \xi^0(t), p^0(t)) \neq 0 \quad \text{for } t = s_j.$$

Then $\tau_j(\alpha)$ is determined for α near α^0 by

$$U_j(t, \xi(t, \alpha), p(t, \alpha)) = 0 \quad \text{for } t = \tau_j(\alpha),$$

with $s_j = \tau_j(\alpha^0)$. Let

$$\xi_j(\alpha) = \xi(\tau_j(\alpha), \alpha), \qquad j = 2, \cdots, m. \tag{A.7}$$

Assume that $\partial\xi/\partial\alpha$ is nonsingular for $\alpha = \alpha^0$, $s \leqq t \leqq \tau^0$ (no conjugate points). Then

$$\frac{\partial \xi_j}{\partial \alpha} = \xi \frac{\partial \tau_j}{\partial \alpha} + \frac{\partial \xi}{\partial \alpha}. \tag{A.8}$$

This implies that the vectors $(\partial\tau_j/\partial\alpha_\lambda, \partial\xi_j/\partial\alpha_\lambda)$, $\lambda = 1, \cdots, n$, are linearly independent when $\alpha = \alpha^0$. Hence, $(\tau_j(\alpha), \xi_j(\alpha))$ lies in a C^∞ manifold Σ_j of dimension n, for α near α^0; and the trajectory $\gamma(\alpha)$ is not tangent to Σ_j at $(\tau_j(\alpha), \xi_j(\alpha))$.

According to (6.3) to find the coefficient $\theta(s, x)$ in the approximation (6.2) we need to calculate $\Delta_x \varphi^0$ along $\gamma^0(s, x)$. Now $\Delta_x \varphi^0$ is the trace of the matrix φ^0_{xx}. By differentiating (A.5) with respect to α we get the matrix equation

$$\varphi^0_{xx} = \frac{\partial p}{\partial \alpha} \left(\frac{\partial \xi}{\partial \alpha} \right)^{-1}. \tag{A.9}$$

To find θ_x, which appears in the approximation (6.7), we can use (6.9). By differentiating (A.5) again with respect to α we get

$$(\varphi^0_{x_i})_{xx} \frac{\partial \xi}{\partial \alpha_\lambda} \frac{\partial \xi}{\partial \alpha_\mu} + (\varphi^0_{x_i})_x \frac{\partial^2 \xi}{\partial \alpha_\lambda \partial \alpha_\mu} = \frac{\partial^2 p_i}{\partial \alpha_\lambda \partial \alpha_\mu}, \qquad i, \lambda, \mu = 1, \cdots, n, \tag{A.10}$$

from which $(\Delta_x \varphi^0)_x$ is expressed using (A.9) in terms of first and second order α partial derivatives of ξ and p. These derivatives obey ordinary differential equations obtained by taking $\partial/\partial\alpha_\lambda$, $\partial^2/\partial\alpha_\lambda\partial\alpha_\mu$ in (A.2) and in the formula (A.6). From (6.3), with $x = \xi(s, \alpha)$ and $\tau^0 = \tau_1(\alpha)$, we have the data

$$\theta_x = (\Delta_x \varphi^0)\left(\frac{\partial \tau_1}{\partial \alpha}\right)\left(\frac{\partial \xi}{\partial \alpha}\right)^{-1} \quad \text{at } t = \tau_1(\alpha).$$

Let $p_j(\alpha) = p(\tau_j(\alpha), \alpha)$. Then

$$\frac{\partial p_j}{\partial \alpha} = \dot{p}\frac{\partial \tau_j}{\partial \alpha} + \frac{\partial p}{\partial \alpha}.$$

This, together with (A.2) and (A.8), gives a relation between right- and left-hand values of $\partial\xi/\partial\alpha$, $\partial p/\partial\alpha$ at $\tau_j(\alpha)$, $j = 2, \cdots, m$, if Σ_j is a switching surface. If Σ_j is a transition surface, then $\dot{\xi}, \dot{p}, \partial\xi/\partial\alpha, \partial p/\partial\alpha$ are continuous at $\tau_j(\alpha)$. Another differentiation in α gives a relation between right- and left-hand values of $\partial^2\xi/\partial\alpha_\lambda\partial\alpha_\mu$, $\partial^2 p/\partial\alpha_\lambda\partial\alpha_\mu$ at $\tau_j(\alpha)$. If α is taken as a set of local coordinates for Σ_1, then data for all of these quantities at time $\tau_1(\alpha)$ are determined from (A.3) and (A.6).

To calculate the coefficient χ in (6.11) from (6.12), one needs $\Delta_x\theta$ as well as θ_x along $\gamma^0(s, x)$. However,

$$\theta_{xx} = \frac{\partial q}{\partial \alpha}\left(\frac{\partial \xi}{\partial \alpha}\right)^{-1},$$

where

$$q(t, \alpha) = \theta_x(t, \xi(t, \alpha)),$$

and $\partial q/\partial\alpha$ can be found by using (6.9) and similar calculations to those above.

REFERENCES

[1] J. V. Breakwell, J. L. Speyer and A. E. Bryson, *Optimization and control of nonlinear systems using the second variation*, this Journal, 1 (1963), pp. 193–223.

[2] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, vol. 2, Interscience, New York, 1961.

[3] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[4] P. Dorato, Chang-Ming Hsieh and P. N. Robinson, *Optimal bang-bang control of linear stochastic systems with a small noise parameter*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 682–689.

[5] E. B. Dynkin, *Markov Processes*, Springer-Verlag, Berlin, 1965.

[6] H. Federer, *Geometric Measure Theory*, Springer-Verlag, Berlin, 1969.

[7] W. H. Fleming, *Duality and a priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., 16 (1966), pp. 254–279; Erratum, 19 (1969), p. 204.

[8] ———, *Optimal control of partially observable diffusions*, this Journal, 6 (1968), pp. 194–214.

[9] ———, *The Cauchy problem for a nonlinear first-order partial differential equation*, J. Differential Equations, 5 (1969), pp. 515–530.

[10] ———, *Controlled diffusions under polynomial growth conditions*, Control Theory and the Calculus of Variations, A. V. Balakrishnan, ed., Academic Press, 1969, pp. 209–234.

[11] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, N.J., 1964.

[12] I. I. Gikhman and A. V. Skorokhod, *Introduction to the Theory of Random Processes*, W. B. Saunders, Philadelphia, 1969.

[13] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[14] E. HOPF, *The partial differential equation* $u_t + u_{xx} = \mu u_{xx}$, Comm. Pure Appl. Math., 3 (1950), pp. 201–230.

[15] H. J. KUSHNER AND A. J. KLEINMAN, *Numerical methods for the solution of degenerate nonlinear elliptic equations arising in optimal stochastic control theory*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 344–353.

[16] H. J. KUSHNER, *Near optimal control in the presence of small stochastic perturbations*, Trans. ASME Ser. D, J. Basic Engrg., 87 (1965), pp. 103–108.

[17] ———, *On the stochastic maximum principle: fixed time of control*, J. Math. Anal. Appl., 11 (1965), pp. 78–92.

[18] O. A. LADYZHENSKAYA, V. A. SOLONNIKOV AND N. N. URAL'SEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, R.I., 1968.

[19] J. P. LASALLE AND H. HERMES, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[20] O. A. OLEINIK, *Discontinuous solutions of nonlinear differential equations*, Uspekhi Mat. Nauk, 12 (1957), no. 3, pp. 3–73; English transl., Amer. Math. Soc. Transl. (2), no. 26, pp. 95–172.

[21] ———, *On a problem of Fichera*, Dokl. Akad. Nauk SSSR, 157 (1964), pp. 1297–1301 = Soviet Math. Dokl., 5 (1964), pp. 1129–1133.

[22] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCKENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[23] R. L. STRATONOVICH, *On the theory of optimal control. An asymptotic method for solving the diffusive alternative equation*, Automat. Remote Control, 23 (1962), pp. 1352–1360.

[24] D. W. STROOCK AND S. R. S. VARADHAN, *Diffusion processes with continuous coefficients*, Comm. Pure Appl. Math., 22 (1969), pp. 345–400, 479–530.

[25] W. M. WONHAM AND W. F. CASHMAN, *A computational approach to optimal control of stochastic saturating systems*, Internat. J. Control, 10 (1969), pp. 77–98.

[26] W. M. WONHAM, *On the separation theorem of stochastic control*, this Journal, 6 (1968), pp. 312–326.

SIAM J. CONTROL
Vol. 9, No. 4, November 1971

WEAK SOLUTIONS OF A PARTIAL DIFFERENTIAL EQUATION OF DYNAMIC PROGRAMMING*

RAYMOND W. RISHEL†

Abstract. A formal dynamic programming argument relates the "value function of a stochastic optimal control problem" with the solution of a nonlinear parabolic partial differential equation. In cases in which the partial differential equation is degenerate, it may not have a classical solution but may have a weak solution in the sense of the theory of Schwartz distributions. It has been an open question as to whether a weak solution of the partial differential equation does equal the value function of the stochastic optimal control problem.

This paper shows that, roughly, whenever an associated uncontrolled system has an appropriately behaved density function, the equality holds. It is also shown that an optimal control law may be determined by minimizing a Hamiltonian formed in terms of the partial derivatives of the weak solution of the partial differential equation.

1. Introduction. In [3] Fleming studied the partial differential equation of dynamic programming connected with a stochastic optimal control problem. He showed, when the partial differential equation was uniformly parabolic, that there was a smooth solution and this solution was the value function for the stochastic optimal control problem. He called the case in which the partial differential equation was not uniformly parabolic, degenerate. Under mild assumptions he showed in the degenerate case that the partial differential equation had a weak solution. It was left as an open conjecture whether this weak solution equaled the value function of the stochastic control problem. Since in most applications the partial differential equation is degenerate, the degenerate case is especially important. For instance, for any stochastic control system governed by a single nth order differential equation with white noise input, the corresponding partial differential equation is degenerate.

This paper studies the relationship between a weak solution of the degenerate partial differential equation and the value function of the optimal control problem. Theorem 3 shows, under assumptions similar to those of [3], that if the partial derivatives of the weak solution are functions in appropriate L_p-spaces and solutions of an associated uncontrolled stochastic differential equation have a density which is in an appropriate L_p-space, then the weak solution equals the value function of the stochastic control problem. It also follows from Theorem 3 that there is a control function which minimizes a Hamiltonian function formed in terms of the partial derivatives of the weak solution, and this control is an optimal control for the stochastic control problem.

The question of obtaining precise conditions under which solutions of stochastic differential equations have densities is an interesting open question. It is known [6, Lemma 8.6] that solutions of nondegenerate stochastic differential equations have densities. Recent papers of Kushner [5] and Zakai [8] consider special cases of degenerate stochastic differential equations whose solutions have densities.

The optimal control problem is formulated for a class of controls which are merely bounded measurable functionals on the past of the processes involved.

* Received by the editors August 4, 1970, and in revised form March 19, 1971.

† Bell Telephone Laboratories, Whippany, New Jersey 07981.

A technique of Girsanov [4] is used to assure there are processes corresponding to this very general type of control. The author would like to express his appreciation to Dr. V. E. Beneš for pointing out to him the importance of Girsanov's technique. Beneš used this technique in [2] to obtain existence theorems for optimal stochastic controls.

2. The stochastic control problem.

2.1. Definition of notation. Let $f(t, x, y)$ be a bounded continuous function and $g(t, x, y, u)$ a bounded Borel measurable function:

$$f: E^1 \times E^n \times E^m \to E^n, \qquad g: E^1 \times E^n \times E^m \times E^k \to E^m. \tag{1}$$

Suppose that f satisfies

$$|f(t, x, y) - f(t, x', y')| \leqq K_1[|x - x'| + |y - y'|] \tag{2}$$

and that $g(t, x, y, u)$ is continuous as a function of u for fixed (t, x, y). Let M_m be the space of nonsingular $m \times m$ matrices with norm defined so that the square of the norm equals the sum of the squares of elements of the matrix. Let $\sigma(t, y)$ be a bounded continuous function,

$$\sigma: E^1 \times E^m \to M_m, \tag{3}$$

such that

$$|\sigma(t, y) - \sigma(t, y')| \leqq K_2|y - y'|. \tag{4}$$

Suppose, in addition, that the inverse matrix $\sigma(t, y)^{-1}$ of $\sigma(t, y)$ is a bounded continuous function. Let $C(t, x, y, u)$,

$$C: E^1 \times E^n \times E^m \times E^k \to E^1, \tag{5}$$

be a continuous real-valued function. Let (Ω, β, P) denote a triple of a probability space Ω, Borel field β, and probability measure P. Let $\xi(t)$ denote an m-dimensional Brownian motion process, with respect to (Ω, β, P), whose covariance matrix is given by

$$E\{\xi(t) \cdot \xi(s)\} = \min(t, s)I, \tag{6}$$

where I is the identity matrix.

Let $F[t_0, T]$ denote the space of all functions on the interval $[t_0, T]$ to E^{m+n}. Let $\mathcal{M}$ denote the Borel field on $F[t_0, T]$ generated by the cylinder sets. A function η,

$$\eta: E^1 \times F[t_0, T] \to E^k, \tag{7}$$

will be said to be nonanticipative if for each t and $x_1(s), x_2(s) \in F[t_0, T]$ such that $x_1(s) = x_2(s)$ if $t_0 \leqq s \leqq t$, then $\eta(t, x_1(\cdot)) = \eta(t, x_2(\cdot))$.

2.2. The optimal control problem. Let U be a compact subset of E^k. A function

$$\eta: E^1 \times F[t_0, T] \to U \tag{8}$$

will be called an admissible control if it is nonanticipative and Borel measurable.

Let B be a bounded closed region in E^{n+m} whose boundary is a twice continuously differentiable manifold. Let t_0 be an initial time and (x, y) an initial state satisfying respectively $0 \leqq t_0 \leqq T$; $(x, y) \in B$. Let trajectories of the control system be defined by the equations

$$x(t) = x + \int_{t_0}^{t} f(s, x(s), y(s))\, ds, \tag{9}$$

$$y(t) = y + \int_{t_0}^{t} g(s, x(s), y(s), u(s))\, ds + \int_{t_0}^{t} \sigma(s, y(s))\, d\xi(s), \tag{10}$$

$$u(s) = \eta(s, x(\cdot), y(\cdot)). \tag{11}$$

The last integral in (10) is as a stochastic Ito integral. Let τ denote the stopping time for the process which is the first time $(x(t), y(t))$ hits the boundary of B. Let $\tau_1 = \min[\tau, T]$.

The optimal control problem is to find a control η in the class of admissible controls so that

$$E\left\{\int_{t_0}^{\tau_1} C(s, x(s), y(s), u(s))\, ds\right\} \tag{12}$$

is a minimum.

3. Comments on the statement of the optimal control problem. Since the controls are only assumed to be Borel measurable functionals on the past of the process and the function $g(t, x, y, u)$ is only assumed to be Borel measurable, the question of whether solutions of (9)–(11) exist arises. The following procedure of I. V. Girsanov [4] can be used to show that there are solutions of (9)–(11). Under the assumptions above, classical theorems [7, p. 47] show that there is a unique solution to the equations

$$x(t) = x + \int_{t_0}^{t} f(s, x(s), y(s))\, ds, \tag{13}$$

$$y(t) = y + \int_{t_0}^{t} \sigma(s, y(s))\, d\xi(s). \tag{14}$$

Define $\zeta_{t_0}^{t}(g)$ by

$$\begin{aligned}\zeta_{t_0}^{t}(g) = {} & \int_{t_0}^{t} g(s, x(s), y(s), u(s))^T [\sigma(s, y(s))^T]^{-1}\, d\xi \\ & - \frac{1}{2}\int_{t_0}^{t} g(s, x(s), y(s), u(s))^T [\sigma(s, y(s))^T]^{-1} \\ & \cdot [\sigma(s, y(s))]^{-1} g(s, x(s), y(s), u(s))\, ds,\end{aligned} \tag{15}$$

where $u(s) = \eta(s, x(\cdot), y(\cdot))$ and $x(t)$, $y(t)$ are solutions of (13)–(14).

THEOREM 1 (Girsanov [4, p. 287 and p. 296]). *Define a measure* $\tilde{P}$ *by*

$$\tilde{P}(d\omega) = \exp[\zeta_{t_0}^{T}(g)] P(d\omega) \tag{16}$$

and a stochastic process $\tilde{\xi}(t)$ *for* $t_0 \leqq t \leqq T$ *by*

$$\tilde{\xi}(t) = \xi(t) - \int_{t_0}^{t} \sigma(s, y(s))^{-1} g(s, x(s), y(s), u(s))\, ds. \tag{17}$$

Then $\tilde{P}$ is a probability measure on (Ω, β).[1] *The process $\tilde{\xi}(t, \omega)$ is a Brownian motion process on $(\Omega, \beta, \tilde{P})$. The solution of (13)–(14), $x(t, \omega), y(t, \omega)$, when considered with respect to the probability triple $(\Omega, \beta, \tilde{P})$, is a solution of (9)–(11) with the Brownian motion $\tilde{\xi}(t)$ replacing $\xi(t)$.*

It will always be assumed that the solutions of (9)–(11) which are involved in the optimal control problem are constructed from solutions of (13)–(14) in the manner stated in the theorem. The question of whether (9)–(11) has a unique solution appears to be still open. In [6] Stroock and Varadhan prove in a more restricted case analogous to (9)–(11) that solutions of the equations always induce a unique probability measure on the space of continuous functions.

4. Intermediate results. The following lemma shows how densities of solutions of (13)–(14) are related to densities of solutions of (9)–(11) constructed according to Theorem 1. The proof of this lemma is nearly the same as that of Lemma 8.6 of [6].

LEMMA 1. *Let t_0, T, h and α be numbers such that $t_0 < t_0 + h < T$ and $\alpha > 1$. If for a solution $x(t)$, $y(t)$ of (13)–(14), the random vector $x(t)$, $y(t)$ has a probability density $p(t, x, y)$ such that*

$$\int_B \int_{t_0+h}^{T} p(t, x, y)^{\alpha}\, dt\, dx\, dy < \infty, \tag{18}$$

then for the solution $x(t)$, $y(t)$ of (9)–(11) constructed by Theorem 1, the random vector $x(t)$, $y(t)$ has a density $q(t, x, y)$ such that for $1 < \alpha' < \alpha$,

$$\int_B \int_{t_0+h}^{T} q(t, x, y)^{\alpha'}\, dt\, dx\, dy < \infty. \tag{19}$$

Since the solution of (9)–(11) will depend on which control $\eta(t, x(\cdot), y(\cdot))$ is used, the density $q(t, x, y)$ also will depend on which control η is being used. To save burdening the notation, this dependence will not be indicated in the formulas. However, the reader is cautioned to recall that these quantities will always depend on the control being considered.

The proof of Lemma 1 requires the following lemma of [4].

LEMMA 2 (Girsanov [4, p. 291]). *Let $\mathcal{F}_t$ denote the Borel field generated by the random vectors $x(s)$, $y(s)$ for $t_0 \leqq s \leqq t$. Let $\zeta_{t_0}^t(g)$ be defined by (15). Then for any F_t-measurable function $\eta(\omega)$,*

$$\int_{\Omega} \eta(\omega) \tilde{P}(d\omega) = \int_{\Omega} \eta(\omega) \exp\left[\zeta_{t_0}^t(g)\right] P(d\omega). \tag{20}$$

Proof of Lemma 1. For a real number $\rho > 1$ let

$$\int_{\Omega} \exp\left[\rho\, \zeta_{t_0}^t(g)\right] P(d\omega) = N_{\rho}. \tag{21}$$

The previous assumptions imply that there is a number k such that

$$\left|g(s, x(s), y(s), u(s))^T [\sigma(s, y(s))^T]^{-1} [\sigma(s, y(s))]^{-1} g(s, x(s), y(s), u(s))\right| < k. \tag{22}$$

[1] This first conclusion of Theorem 1 follows from the remark on p. 296 of [4]. The remaining conclusions are those of [4, Theorem 1, p. 287].

Using an abbreviated notation, we have

$$N_\rho = \int_\Omega \exp\left[\rho\int_{t_0}^t g^T\sigma^{-1^T}\sigma^{-1}\,dy - \frac{1}{2}\rho\int_{t_0}^t g^T\sigma^{-1^T}\sigma^{-1}g\,ds\right]P(d\omega)$$

$$\leqq \exp\left[\frac{1}{2}k(t-t_0)\rho(\rho-1)\right]\int_\Omega \exp\left[\int_{t_0}^t \rho g^T\sigma^{T^{-1}}\sigma^{-1}\,dy - \frac{1}{2}\int_{t_0}^t \rho g^T\sigma^{T^{-1}}\sigma^{-1}\rho g\,ds\right]P(d\omega) \tag{23}$$

$$\leqq \exp\tfrac{1}{2}k(t-t_0)\rho(\rho-1).$$

The last step follows because, by applying Lemma 2 with g replaced by ρg, the integral can be shown to equal one. By Lemma 2, if A is a set measurable with respect to the Borel field generated by $x(t)$, $y(t)$,

$$\int_A \tilde{P}(d\omega) = \int_A E\{\exp[\zeta_{t_0}^t(g)]|x(t), y(t)\}P(d\omega); \tag{24}$$

hence solutions of (9)–(11) have a density $q(t, x, y)$ given by

$$q(t,x,y) = E\{\exp[\zeta_{t_0}^t(g)]|x(t) = x, y(t) = y\}p(t,x,y).$$

Denote $E\{\exp[\zeta_{t_0}^t(g)|x(t) = x, y(t) = y\}$ more briefly by $r(t, x, y)$. Since conditional expectation does not increase the L_ρ-norm, definition (21) implies

$$\int_{E^{m+n}} |r(t,x,y)|^\rho p(t,x,y)\,dx\,dy \leqq N_\rho. \tag{25}$$

Consequently,

$$\int_{t_0+h}^T \int_{E^{m+n}} |r(t,x,y)|^\rho p(t,x,y)\,dx\,dy\,dt \leqq (T-t_0)\exp[\tfrac{1}{2}k(T-t_0)\rho(\rho-1)]. \tag{26}$$

This implies

$$\int_{t_0+h}^T \int_{E^{m+n}} |q(t,x,y)|^\rho |p(t,x,y)|^{1-\rho}\,dx\,dy\,dt \leqq (T-t_0)\exp[\tfrac{1}{2}k(T-t_0)\rho(\rho-1)]. \tag{27}$$

If $1 < \alpha' < \alpha$ and h, k, ρ and β are chosen so that $h > (\alpha-1)(\alpha-\alpha')^{-1}$, $h^{-1} + k^{-1} = 1$, $\rho = h\alpha'$, and $\beta = h^{-1}(\rho-1)$, then an elementary argument shows that $1 < \beta k < \alpha$. Therefore, writing $q^{\alpha'} = q^{\alpha'}p^{-\beta}p^\beta$, applying Hölder's inequality with parameters h and k, and using (27) and (18) gives, if $1 < \alpha' < \alpha$, that

$$\int_{t_0+h}^T \int_B |q(t,x,y)|^{\alpha'}\,dt\,dx\,dy < \infty, \tag{28}$$

which is the conclusion of Lemma 1.

Define the matrix $a(t, y)$ by $a(t, y) = \sigma(t, y)\sigma(t, y)^T$. For the stochastic optimal control problem define a value function as follows. Consider the optimal control

problem starting from a variable time t and point (x, y), where $0 \leqq t \leqq T$ and $(x, y) \in B$. Let $V(t, x, y)$ be the value of the infimum of the performance index over the class of admissible controls, given that the processes started at time t from the point (x, y). That is,

$$V(t, x, y) = \inf_{\eta} E\left\{\int_t^{\tau_1} C(s, x(s), y(s), u(s))\, ds \,\middle|\, x(t) = x, y(t) = y\right\}, \tag{29}$$

where $u(s) = \eta(s, x(\cdot), y(\cdot))$.

The method of dynamic programming gives a formal argument which connects the value function $V(t, x, y)$ with the solution $\phi(t, x, y)$ of

$$\begin{aligned} &\phi_t + \sum_{i=1}^{m} a_{ij}(t, y)\phi_{y_i y_j} + \sum_{i=1}^{n} f_i(t, x, y)\phi_{x_i} \\ &\qquad + \min_{u \in U}\left\{\sum_{i=1}^{m} g_i(t, x, y, u)\phi_{y_i} + C(t, x, y, u)\right\} = 0, \end{aligned} \tag{30}$$

which satisfies the boundary condition

$$\phi(t, x, y) = 0 \quad \text{if } (x, y) \in \text{boundary } B \quad \text{or} \quad t = T. \tag{31}$$

We shall say for $\beta > 1$ that a function ϕ is in class F_β if $\phi(t, x, y)$ is continuous on $[0, T] \times B$, and the Schwartz distribution theory second partial derivatives with respect to y variables and first partial derivatives with respect to t and x variables are functions which are in $L_\beta([0, T] \times B)$.

Define a norm on the functions of class F_β by

$$\begin{aligned} |\phi| = &\left[\sup_{0 \leqq t \leqq T, (x,y) \in B} \phi(t, x, y)\right] + \sum_{z = t, x_i, y_i} \left[\int_0^T \int_B |\phi_z|^\beta \, dx\, dy\, dt\right]^{1/\beta} \\ &+ \sum_{i,j=1}^{m} \left[\int_0^T \int_B |\phi_{y_i y_j}|^\beta \, dx\, dy\, dt\right]^{1/\beta}. \end{aligned} \tag{32}$$

A function ϕ in F_β will be said to be a weak solution of (30) if (30) is satisfied almost everywhere with respect to $(m + n + 1)$-dimensional Lebesgue measure on $[0, T] \times B$.

Denote by $\Lambda(t, x, y, u)$ the partial differential operator for which

$$\begin{aligned} \Lambda(t, x, y, u)[\phi] = &\phi_t + \sum_{i=1}^{m} a_{ij}(t, y)\phi_{y_i y_j} + \sum_{i=1}^{n} f_i(t, x, y)\phi_{x_i} \\ &+ \sum_{i=1}^{m} g_i(t, x, y, u)\phi_{y_i}. \end{aligned} \tag{33}$$

In cases in which the dependence of Λ on the variables and functions involved in its coefficients can be inferred from the context, the shortened notation $\Lambda[\phi]$ will be used for (33).

THEOREM 2. *Let the hypothesis of Lemma* 1 *be satisfied. Let* $\phi(t, x, y)$ *be a solution of* (30) *in* F_β, *where* $\beta = \alpha'(\alpha' - 1)^{-1}$. *Let* $x(t), y(t)$ *be solutions of* (9)–(11) *corresponding to the control* $\eta(t, x(\cdot), y(\cdot))$. *Let* τ *denote the first time* $x(t), y(t)$ *hits boundary* B *and* $\tau_1 = \min(\tau, T)$, $\tau_h = \min(\tau, t_0 + h)$. *Let* $\chi(t)$ *be the characteristic*

function of the set $\{\omega : \tau(\omega) \geqq t\}$. *Then*

$$E\{\phi(\tau_1, x(\tau_1), y(\tau_1))\} - E\{\phi(\tau_h, x(\tau_h), y(\tau_h))\}$$

$$= \int_{t_0+h}^{T} \int_B E\{\chi(s)\Lambda(s, x(s), y(s), \eta(s, x(\cdot), y(\cdot)))[\phi] | x(s) = x, y(s) = y\} \cdot q(s, x, y)\, dx\, dy\, ds. \tag{34}$$

Proof of Theorem 2. Let $\phi^n(t, x, y)$ be a sequence of twice continuously differentiable functions such that ϕ^n converges to ϕ in F_β. Since ϕ^n is twice continuously differentiable, Ito's formula [7] gives

$$E\{\phi^n(\tau_1, x(\tau_1), y(\tau_1))\} - E\{\phi^n(\tau_h, x(\tau_h), y(\tau_h))\}$$

$$= E\left\{\int_{t_0+h}^{T} \chi(s)\Lambda(s, x(s), y(s), \eta(s, x(\cdot), y(\cdot)))[\phi^n]\right\} ds. \tag{35}$$

Now by Fubini's theorem,

$$E\int_{t_0+h}^{T} \chi(s)\Lambda[\phi^n]\, ds = \int_{t_0+h}^{T} E\{\chi(s)\Lambda\phi^n\}\, ds$$

$$= \int_{t_0+h}^{T} \int_B E\{\chi(s)\Lambda\phi^n | x(s) = x, y(s) = y\} q(s, x, y)\, dx\, dy\, dt. \tag{36}$$

Looking at definition (33) we see that the equality

$$E\{\chi(s)g_i(s, y(s), \eta(s, x(\cdot), y(\cdot)))\phi^n_{y_i}(s, x(s), y(s)) | x(s) = x, y(s) = y\}$$

$$= \phi^n_{y_i}(s, x, y) E\{\chi(s)g_i(s, x(s), y(s), \eta(s, x(\cdot)y(\cdot))) | x(s) = x, y(s) = y\} \tag{37}$$

is typical for the terms of $E\{\chi(s)\Lambda(s)[\phi^n] | x(s) = x, y(s) = y\}$. A similar result holds when ϕ replaces ϕ^n in (37). Therefore, the expression

$$E\{\chi(s)(\Lambda(s)[\phi^n] - \Lambda(s)[\phi]) | x(s) = x, y(s) = y\} \tag{38}$$

can be written as a sum of bounded terms times terms which are converging to zero in $L_\beta([t_0, T] \times B)$. Therefore, we conclude that (38) converges to zero in $L_\beta([t_0, T] \times B)$.

By applying Hölder's inequality, involving (19) and (38), we see that the right side of (36) converges to the right side of (34). Since F_β convergence implies uniform convergence of ϕ_n to ϕ, the left side of (35) converges to the left side of (34), which completes the proof of Theorem 2.

If $\phi(t, x, y)$ is a weak solution of (30), an implicit function theorem of Beneš [1] asserts there is a measurable function $\gamma(t, x, y)$,

$$\gamma : [t_0, T] \times B \to U, \tag{39}$$

such that

$$\sum_{i=1}^{m} g_i(t, x, y, \gamma(t, x, y))\phi_{y_i} + C(t, x, y, \gamma(t, x, y))$$

$$= \min_{u \in U} \left\{ \sum_{i=1}^{m} g_i(t, x, y, u)\phi_{y_i} + C(t, x, y, u) \right\} \tag{40}$$

almost everywhere with respect to Lebesgue measure on $[t_0, T] \times B$. Extend γ in an arbitrary way to a Borel measurable function on $[t_0, T] \times E^{m+n}$ with values in U.

This function induces a measurable function on $[t_0, T] \times F[t_0, T]$ by the formula

$$\gamma(t, x(\cdot), y(\cdot)) = \gamma(t, x(t), y(t)). \tag{41}$$

5. Fundamental theorem.

THEOREM 3. *Let the solution of* (13)–(14) *have a density* $p(t, x, y)$ *such that for some* α, $\alpha > 1$,

$$\int_{t_0+h}^{T} p(t, x, y)^{\alpha} \, dx \, dy \, dt < \infty. \tag{42}$$

Let $1 < \alpha' < \alpha$, *and let* $\phi(t, x, y)$ *be a solution of* (30) *which is in* F_β *for* $\beta = \alpha'(\alpha' - 1)^{-1}$ *which satisfies the boundary condition* (31). *Let* $\gamma(t, x(\cdot), y(\cdot))$ *be the control defined in* (40)–(41).

Then if $x(t), y(t)$ *is a solution of* (9)–(11) *corresponding to the control* $u(t) = \gamma(t, x(\cdot), y(\cdot))$,

$$\phi(t_0, x, y) = E\left\{\int_{t_0}^{\tau_1} C(s, x(s), y(s), u(s)) \, ds\right\}. \tag{43}$$

If $x(t), y(t)$ *is a solution of* (9)–(11) *corresponding to an arbitrary control* $u(t) = \eta(t, x(\cdot), y(\cdot))$,

$$\phi(t_0, x, y) \leqq E\left\{\int_{t_0}^{\tau_1} C(s, x(s), y(s), u(s)) \, ds\right\}. \tag{44}$$

Notice that t_0, x, y could be any variables for which $0 < t_0 < T$ and $(x, y) \in B$. Thus Theorem 3 implies that $\phi(t, x, y) = V(t, x, y)$ on $[0, T] \times B$. That is, if there is a solution of (30) with boundary condition (31), it agrees with the value function of the stochastic control problem. Notice also that (43) and (44) imply that the control γ constructed from (40)–(41) is an optimal control law for the stochastic control problem.

Proof of Theorem 3. Since ϕ is a weak solution of (30), the definition of γ implies that there is a set N of $(m + n + 1)$-dimensional Lebesgue measure zero such that

$$-\Lambda(t, x, y, \gamma(t, x, y))[\phi] = C(t, x, y, \gamma(t, x, y)) \tag{45}$$

on $[t_0, T] \times B - N$.

Let $u(t) = \gamma(t, x(t), y(t))$ and $x(t)$, $y(t)$ be corresponding solutions of (8)–(11). Since $x(t)$, $y(t)$ has a density $q(t, x, y)$, (45) and Fubini's theorem imply that for almost every t,

$$-\chi(t)\Lambda(t, x(t), y(t), u(t))[\phi] = c(t, x(t), y(t), u(t)) \tag{46}$$

with probability one and

$$E\{-\chi(t)\Lambda(t, x(t), y(t), u(t))[\phi] | x(t) = x, y(t) = y\} \\ = -\Lambda(t, x, y, \gamma(t, x, y))[\phi] = C(t, x, y, y(t, x, y)) \tag{47}$$

except at a set of measure zero on B.

Substituting (47) in (34) gives

$$E\{\phi(\tau_h, x(\tau_h), y(\tau_h))\} - E\{\phi(\tau_1, x(\tau_1), y(\tau_1))\} \\ = \int_{t_0+h}^{T} E\{\chi(t)C(t, x(t), y(t), u(t))\}\, dt. \tag{48}$$

Since ϕ satisfies the boundary condition (31), $\phi(\tau_1, x(\tau_1), y(\tau_1)) = 0$. Since ϕ is bounded and continuous and $x(t)$, $y(t)$ has continuous paths, using Lebesgue's convergence theorem and passing to the limit as h approaches zero gives

$$\phi(t_0, x, y) = \lim_{h \to 0} E\{\phi(\tau_h, x(\tau_h), y(\tau_h))\} \\ = E\int_{t_0}^{\tau_1} C(t, x(t), y(t), u(t))\, dt. \tag{49}$$

To establish (44) notice that since ϕ is a weak solution of (30) that there is a set N of $(m + n + 1)$-dimensional Lebesgue measure zero, which does not depend on u, such that for all $u \in U$,

$$-\Lambda(t, x, y, u)[\phi] \leqq C(t, x, y, u) \tag{50}$$

on $[t_0, T] \times B - N$.

For any control $\eta(t, x(\cdot), y(\cdot))$, let $x(t)$, $y(t)$ be a solution of (8)–(11) corresponding to η. Since $x(t)$, $y(t)$ has a density $q(t, x, y)$, (50) implies

$$-\chi(t, \omega)\Lambda(t, x(t, \omega), y(t, \omega), u)[\phi] \leqq \chi(t, \omega)C(t, x(t, \omega)y(t, \omega), u) \tag{51}$$

for all $u \in U$ and $dt \times P$ almost every (t, ω) in $[t_0, T] \times \Omega$. Letting $u(t, \omega) = \eta(t, x(\cdot, \omega), y(\cdot, \omega))$, we have that inequality (51) implies

$$\chi(t, \omega)\Lambda(t, x(t, \omega), y(t, \omega), u(t, \omega))[\phi] \leqq \chi(t, \omega)C(t, x(t, \omega), y(t, \omega), u(t, \omega)), \tag{52}$$

$dt \times P$ almost everywhere on $[t_0, T) \times \Omega$.

By substituting (52) in (34) and proceeding in a manner similar to that of (48)–(49), it follows that (44) holds.

6. Concluding remarks. By combining Theorem 3 with results of Fleming [3], a fairly complete theory is available for the type of stochastic control problem discussed in this paper. Fleming [3, Theorem 5.1, p. 269] proves that there exists a weak solution of the partial differential equation (30) with boundary condition (31) which is in F_2. Then under the combined hypothesis of [3] and Theorem 3, where the exponent α in Theorem 3 is larger than two, there exists an optimal control, this optimal control minimizes the Hamiltonian as in (40), and the optimal value of the performance starting at time t from the point x, y is given by the solution of (30). It also follows from Theorem 3, under the hypothesis of the theorem, that a weak solution of the partial differential equation of dynamic programming (30) in F_β with boundary condition (31) is unique.

REFERENCES

[1] V. E. BENEŠ, *The existence of optimal strategies based on specified information for a class of stochastic decision problems*, this Journal, 8 (1970), pp. 179–188.

[2] ———, *The existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–472.

[3] W. H. FLEMING, *Duality and a priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., 16 (1966), pp. 254–279.

[4] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theor. Probability Appl., 5 (1960), pp. 285–301.

[5] H. J. KUSHNER, *The Cauchy problem for a class of degenerate parabolic equations and asymptotic properties of related diffusion processes*, J. Differential Equations, 6 (1969), pp. 209–231.

[6] D. W. STROOCK AND S. R. S. VARADHAN, *Diffusion processes with continuous coefficients. II*, Comm. Pure Appl. Math., 22 (1969), pp. 479–530.

[7] A. V. SKOROKHOD, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Mass., 1965.

[8] M. ZAKAI, *A Lyapunov criterion for the existence of stationary probability distributions for systems perturbed by noise*, this Journal, 7 (1969), pp. 390–397.

SIAM J. CONTROL
Vol. 9, No. 4, November 1971

GENERALIZED CUTTING PLANE ALGORITHMS*

B. CURTIS EAVES† AND W. I. ZANGWILL‡

Abstract. This paper introduces a master cutting plane algorithm for nonlinear programming that isolates the points it generates from one another until a solution is achieved. The master algorithm provides a foundation for the study of cutting plane algorithms and directs the way for development of procedures which permit deletion of old cuts.

Introduction. Since Kelley [5], Cheney and Goldstein [2] introduced the first cutting plane method for nonlinear programming over ten years ago, and through the developments of the Veinott [8] supporting hyperplane method and the Dantzig and Wolfe [3] dual cutting plane method, the following question has persisted. Is it possible to drop previously generated cutting planes and still guarantee convergence? All of the aforementioned methods require that every previous cut be retained. Then, since the linear programming subproblem must on every iteration increase in size, the subproblem soon becomes unwieldy and difficult, if not impossible, to solve. To circumvent this complication in practical application, one would simply drop old cuts and solve the resulting smaller subproblems. It was not known whether or not this procedure would converge.

One of the authors recently suggested a unifying theory for cutting plane algorithms [9, Chap. 14]. Although this theory did bring together the previous methods, it still did not permit old cuts to be dropped. By extending and generalizing this theory, this paper presents precise methods on how to drop cuts and still guarantee convergence.

The paper is divided into three parts. Section 1 introduces the master cutting plane algorithm (MCPA) and states important background material for convergence of sequences. In § 2 we derive four methods from the MCPA. One of these four will in § 3 further specialize into the Kelley, Cheney, Goldstein (KCG), the Veinott (VN), and the Dantzig-Wolfe methods (DW), while the other three will illustrate not only different approaches to cutting plane methods, but also how to drop old cuts.

The authors have recently been informed that Algorithm IV for the unique optimum case is similar to an approach developed independently and simultaneously by Professor Donald M. Topkis (see [6], [7]). Professor Topkis examined the uniformly strictly concave case and used an entirely different method of proof.

1. Fundamentals. Following the approach in [9, Chap. 14]—cutting plane methods generate a sequence of points z^k, $k = 1, 2, 3, \cdots$, in an effort to calculate a point in a set G, termed the goal set. The set G will be closed in a metric space and

* Received by the editors August 21, 1969, and in final revised form February 16, 1971. This work was supported in part by the U.S. Office of Naval Research under Grant Nonr-047-069, the U.S. Army Research Office—Durham under Contract DAHC-04-67-C-0028, and the National Science Foundation under Grant GK-5695.

† Department of Operations Research, Stanford University, Stanford, California 94305.

‡ Center for Research in Management Science, University of California, Berkeley, California. Now at U.S. Office of Education, Washington, D.C. 20202.

specially selected so that when the MCPA calculates a point in it, the nonlinear programming problem will be solved.

The points generated by the MCPA will be in a closed set X and satisfy the special property that if $z^k \notin G$, then z^{k+1} is separated from the points $z^1, z^2, \cdots, z^k$ by at least a certain distance. This distance will be determined by a separator function that we now define.

DEFINITION. Given the closed sets $\varnothing \neq G \subset X$ in a metric space, an extended real-valued function δ on $X \sim G$ is a *separator*

(a) if it is nonnegative

and

(b) if $z^k \to z^\infty$ and $\delta(z^k) \to 0$, imply $z^\infty \in G$.

An immediate consequence is that a separator is positive on $X \sim G$. For, suppose $z^* \in X \sim G$ and $\delta(z^*) = 0$. Then, letting $z^k = z^*$ for all k, we have

$$z^k \to z^*, \qquad \delta(z^k) \to 0$$

and by (b),

$$z^* \in G.$$

The contradiction is clear and δ must be positive on $X \sim G$.

The next lemma also follows immediately from the definition.

LEMMA 1. *Let δ be a real-valued function on $X \sim G$. If δ is positive and lower semicontinuous, it is a separator.*

An example of a separator is the distance $d(z, G)$ between a point $z \in X$ and the set G. It is easy to show that δ is a separator if and only if δ has a positive lower bound on each compact set in $X \sim G$. The minimum of a finite collection of separators is also a separator as is any positive multiple of a separator. Moreover, if δ is a separator and $\delta' \geqq \delta$ then so is δ'. A condition which is frequently imposed on a separator is $\delta \leqq d(\cdot, G)$, or in other words, that $\delta(z)$ not exceed the distance $d(z, G)$ from z to G for all $z \in X \sim G$.

We now pose the MCPA, a simple procedure that underlies the cutting plane methods.

MASTER CUTTING PLANE ALGORITHM. Let $\delta \leqq d(\cdot, G)$ be a separator on $X \sim G$ and suppose the points $z^1, \cdots, z^k$ in X have been generated. If $z^k \in G$, then terminate; otherwise calculate $z^{k+1} \in X$ such that

$$z^{k+1} \notin B(z^i, \delta(z^i)), \qquad i = 1, \cdots, k,$$

where $B(z, \delta(z))$ is the open ball of radius $\delta(z)$ at z.

THEOREM 2. *Given the closed sets $\varnothing \neq G \subset X$, let the sequence $z^1, z^2, \cdots$ in X be generated by the MCPA. Then any cluster point is in G.*

Proof. Suppose for some subsequence

$$z^k \to z^\infty.$$

By construction of the algorithm, this could occur only if

$$\delta(z^k) \to 0.$$

Then from the definition of a separator, $z^\infty \in G$.

Spaces of sets. To apply the MCPA we must first explore the suggestion in [9] that cutting plane algorithms operate by generating closed sets Z^k contained in X. Specifically, we must explore the space of all such sets Z^k in X.

Assume that X is a compact metric space, that $G \subset X$, and that G is compact. Then define $\mathscr{Z}$ to be the collection of all closed sets Z for which

$$G \subset Z \subset X.$$

$\mathscr{Z}$ can be topologized [4, pp. 166–172] to form a metric space. With this topology, $\lim Z^k = Z^\infty$ or equivalently $Z^k \to Z^\infty$ if and only if both (a) and (b) hold.

(a) If $z^k \in Z^k$ for $k = 1, 2, \cdots$ and $z^k \to z^\infty$, then $z^\infty \in Z^\infty$.

(b) If $z \in Z^\infty$, then there exists $z^k \in Z^k$ for $k = 1, 2, \cdots$ such that $z^k \to z$.

Intuitively, if $Z^k \to Z^\infty$, then as $k \to \infty$ the set Z^k becomes more and more like Z^∞. We summarize a key lemma [4] about this set space.

LEMMA 3. *Suppose X is a compact metric space and $G \subset X$ is a compact subset. Let*

$$\mathscr{Z} = \{Z | G \subset Z \subset X, Z \text{ closed}\}.$$

Then $\mathscr{Z}$ is compact (in the set topology).

The set space permits us to define cuts.

Cuts and maps. We know that cutting plane algorithms operate by taking cuts, yet the MCPA was not stated in terms of a cut but was phrased in terms of a separator. Now we must link the two concepts of a cut and a separator by defining a cut map. Recall that a point-to-set map from $X \sim G$ to $\mathscr{Z}$ takes points of $X \sim G$ into subsets of $\mathscr{Z}$.

DEFINITION. A point-to-set map $\varphi: X \sim G \to \mathscr{Z}$ is a *cut map* if there is a separator such that for each $x \in X \sim G$ and each $Z \in \varphi(x)$,

$$Z \cap B(x, \delta(x)) = \varnothing.$$

Note that any set $Z \in \mathscr{Z}$ can be a cut, but we shall only call Z a cut if $Z \in \bigcup \varphi(X \sim G)$, that is, if $Z \in \varphi(z)$ for some $z \in X \sim G$. Also notice that

$$G = \bigcap \bigcup \varphi(X \sim G)$$

which is to say that G is the intersection of all cuts. In particular, one can observe that if all cuts are convex, then G must also be convex. (In the computational procedures of § 3, cuts will be of the form $H \cap X$, where H is a half-space of E^n, n-dimensional Euclidean space, and X is a compact polyhedral convex set in E^n.)

In practice we may know that a map φ is a cut map without knowing an underlying separator. However, note that we can easily generate other separators from any cut map. Simply define δ on $X \sim G$ by setting $\delta(z) = d(z, Z)$ for some $Z \in \varphi(z)$.

LEMMA 4. *Let φ be a cut map, $z^k \to z^\infty \notin G$, $C^k \in \varphi(z^k)$, and $C^k \to C^\infty$. Then $z^\infty \notin C^\infty$.*

Proof. Let δ be a separator underlying φ and consider

$$\limsup \delta(z^k) \leqq \lim d(z^k, C^k) = d(z^\infty, C^\infty).$$

If $d(z^\infty, C^\infty) = 0$, then $\delta(z^\infty) \to 0$ and by definition $z^\infty \in G$.

Closed maps. In several connections we shall use the notion of a closed map. Let U and V be metric spaces and let $\beta: U \to V$ be a point-to-set map ($\beta(u) \subset V$). We say that β is a closed map if

$$u^i \to u^\infty, \quad v^i \to v^\infty, \quad v^i \in \beta(u^i) \quad \text{imply } v^\infty \in \beta(u^\infty)$$

for all such sequences and limit points.

The next lemma indicates a relationship between closed and cut maps.

LEMMA 5. *Let $\varnothing \neq G \subset X$ be compact sets and let $\varphi: X \sim G \to \mathscr{Y}$ be a closed map such that $z \notin Z$ for $Z \in \varphi(z)$. Then φ is a cut map.*

Proof. The result follows from the fact that

$$\delta(z) = \inf \{d(z, Z) | Z \in \varphi(z)\}$$

is a separator. Assume $z^k \to z^\infty$ and $\delta(z^k) \to 0$. Using closedness, choose $Z^k \in \varphi(z^k)$ such that $\delta(z^k) = d(z^k, Z^k)$. There is a subsequence for which $Z^k \to Z^\infty$ and we have $z^\infty \in Z^\infty$. If $z^\infty \notin G$, by closedness $Z^\infty \notin \varphi(z^\infty)$ and from the hypothesis $z^\infty \notin Z^\infty$, which is a contradiction.

Objective function. Our abstract algorithms of § 2 solve the following nonlinear programming problem:

$$\underset{y \in G}{\text{maximize}}\ r(y),$$

where r is a real-valued continuous function on X.

What is interesting is that the objective function r and a separator δ induce another separator δ'.

LEMMA 6. *Given a compact set X in a metric space, let δ be a separator on $X \sim G$ and r be a continuous function on X. Then the function δ' defined on $X \sim G$ by*

$$\delta'(z) = \inf \{d(z, x) | r(x) \leqq r(z) - \delta(z), x \in X\}$$

is a separator.

Proof. (We adopt the convention that $\inf \varnothing = +\infty$, and hence, $\delta'(z) = +\infty$ may occur.) If $\delta'(z) < +\infty$, then by compactness and continuity there is an $x \in X$ such that $\delta'(z) = d(z, x)$ and $r(x) \leqq r(z) - \delta(z)$.

Now suppose

$$z^k \to z^\infty \quad \text{and} \quad \delta'(z^k) \to 0.$$

Let

$$\delta'(z^k) = d(z^k, x^k)$$

and

$$r(x^k) \leqq r(z^k) - \delta(z^k).$$

Then $d(z^k, x^k) \to 0$, and it follows that $\delta(z^k) \to 0$. As δ is a separator, $z^\infty \in G$. This completes the proof.

Note that if z^k, $k = 1, 2, \cdots$, is a sequence in $X \sim G$ such that $z^k \to z^\infty$ and $r(z^{k+1}) \leqq r(z^k) - \delta(z^k)$, then $z^\infty \in G$. This fact follows from Lemma 6, but it also has a trivial proof, namely, $z^k \to z^\infty$ implies $\delta(z^k) \leqq r(z^k) - r(z^{k+1}) \to 0$ which implies $z^\infty \in G$.

We now adopt a function and a map that depend upon the nonlinear programming problem which facilitates further discussion.

Let r^* be the real-valued function on $\mathscr{Z}$ defined by

$$r^*(Z) = \max_{z \in Z} r(z).$$

In words, r^* is the optimal objective value as a function of the feasible region Z.

Next we define an "arg max" map on $\mathscr{Z}$ by

$$\Gamma(Z) = \{x \in Z | r(x) = r^*(Z)\}.$$

The map Γ specifies all the optimal solution points to the nonlinear programming problem.

The following lemma is proved in [1].

LEMMA 7. *Let r be a continuous real-valued function on X, a compact set in a metric space. Then the function $r^*: \mathscr{Z} \to E^1$ is continuous and the point-to-set map $\Gamma: \mathscr{Z} \to X$ is a closed map.*

2. Abstract cutting plane methods. In this section, we describe four abstract cutting plane methods, Algorithms I–IV, for solving the problem

$$\text{maximize}_{y \in G} r(y). \tag{1}$$

Each of the methods generates a sequence of sets Z^k and points $z^k \in \Gamma(Z^k)$, such that cluster points of the z^k sequence will be in $\Gamma(G)$. Consequently, any cluster point will solve the programming problem (1).

These four algorithms certainly do not exhaust the possibilities; nevertheless, an effort has been made to pick a representative selection from those of which we are aware.

The algorithms, I–IV, have been ordered so that, roughly, each succeeding one relies more upon improvement in objective value. Algorithm I retains all generated cuts. Algorithms II and III provide for the deletion of cuts, but just how many cuts will be dropped is problem dependent and could vary from none to numerous to almost all. Algorithm IV requires more stringent conditions than I–III, but it achieves the ideal with regard to the deletion of cuts; namely, all inactive cuts can be dropped (Z is inactive for z if z is interior to Z with respect to the space X). Algorithm II will probably prove to be the most useful.

We shall always assume that $\varnothing \neq G \subset X$ are compact sets in a metric space. Also

$$\mathscr{Z} = \{Z | G \subset Z \subset X, Z \text{ closed}\}$$

and r is continuous.

Algorithm 0 is the general algorithm that serves as a basis for the others. It assumes that a separator is given either implicitly or explicitly.

ALGORITHM 0. Given a separator $\delta \leqq d(\cdot, G)$, let the set $Z^1 \in \mathscr{Z}$ be selected arbitrarily and choose $z^1 \in \Gamma(Z^1)$. Assume that Z^i and z^i, $i = 1, \cdots, k$, have been generated. If $z^k \in G$, terminate; otherwise select Z^{k+1} and $z^{k+1} \in \Gamma(Z^{k+1})$ such that

$$z^{k+1} \notin B(z^i, \delta(z^i)) \quad \text{for } i = 1, \cdots, k.$$

Clearly, Algorithm 0 operates as the MCPA so that any cluster point z^∞ of the z^k sequence is in G. Moreover, by specification of $\mathscr{Z}$, $Z^k \supset G$, and consequently $r(z^k) \geqq r^*(G)$. From the continuity of r, $r(z^\infty) \geqq r^*(G)$, and we conclude that $z^\infty \in \Gamma(G)$. In other words, z^∞ solves (1).

ALGORITHM I. This algorithm, as will be seen in § 3 of the paper, is a generalization of the KCG, VN, and DW methods. It retains all the old cuts (except perhaps for the redundant ones). Given a cut map $\varphi: X \sim G \to \mathscr{Z}$, the algorithm is as follows.

Assume that $Z^1, \cdots, Z^k$ and $z^1, \cdots, z^k$ have been generated. If $z^k \in G$, terminate; otherwise select $C^k \in \varphi(z^k)$, form $C^k \cap Z^k$, and let $Z^{k+1} \in \mathscr{Z}$ be any subset of $C^k \cap Z^k$. Then calculate the point $z^{k+1} \in \Gamma(Z^{k+1})$.

To show that this operates as Algorithm 0, note that $z^{k+1} \in Z^{k+1}$ since $z^{k+1} \in \Gamma(Z^{k+1})$. This algorithm produces

$$Z^{k+1} \subset Z^k \cap C^k \subset Z^k \subset \cdots \subset Z^i \cap C^i \quad \text{for } 1 \leqq i \leqq k,$$

where $C^i \in \varphi(z^i)$. Thus

$$z^{k+1} \in C^i,$$

and by definition of the cut map, a separator exists such that

$$z^{k+1} \notin B(z^i, \delta(z^i)), \quad i = 1, \cdots, k.$$

Consequently, the algorithm is a special case of Algorithm 0, and any cluster point z^∞ solves (1).

ALGORITHM II. This algorithm drops cuts after r has made sufficient progress as judged by a separator δ. Briefly, if cut C^i is introduced to cut off the point z^i, then C^i can be dropped later on iteration $k+1$ if

$$r(z^{k+1}) \leqq r(z^i) - \delta(z^i)$$

and if the cuts which remain "maintain objective value." Let φ be a cut map and let δ be a separator on $X \sim G$. At each iteration k an index set I^k will specify the cuts that have not yet been dropped. The algorithm is as follows.

Initiate the algorithm with $I^1 = \varnothing$ and $z^1 \in \Gamma(X)$. Assume $I^i \subset \{1, \cdots, i-1\}$ and z^i for $i = 1, \cdots, k$ and C^i for $i = 1, \cdots, k-1$ have been generated. If $z^k \in G$, terminate; otherwise select $C^k \in \varphi(z^k)$, let

$$Y^{k+1} = \left(\bigcap_{i \in I^k} C^i \right) \cap C^k$$

and let $y^{k+1} \in \Gamma(Z^{k+1})$. Then let

$$Z^{k+1} = \bigcap_{i \in I^{k+1}} C^i$$

and $z^{k+1} \in \Gamma(Z^{k+1})$, where I^{k+1} is chosen so that:

(a) $I^{k+1} \subset I^k \cup \{k\}$,

(b) $\{i = 1, \cdots, k | r(y^{k+1}) > r(z^i) - \delta(z^i)\} \subset I^{k+1}$,

(c) $r(z^{k+1}) = r(y^{k+1})$.

This specifies the algorithm.

Let us clarify the selection of I^{k+1}. From (a) the indices in I^k and the index k are the only candidates. We may drop such a candidate index if from (b) there is sufficient decrease in objective value, i.e.,

$$r(y^{k+1}) \leqq r(z^i) - \delta(z^i),$$

and if from (c), dropping it does not cause the objective value to increase over its value at y^{k+1}. Note (b) and (c) must both be satisfied before i can be eliminated.

To prove convergence, let δ' be the separator corresponding to φ and let δ'' be the separator induced from δ and r as in Lemma 6. We shall use the separator

$$\delta^* = \min [\delta', \delta'']$$

to show that

$$z^{k+1} \notin B(z^i, \delta^*(z^i)), \qquad i = 1, \cdots, k.$$

Convergence will then follow from Algorithm 0.

If $i \notin I^{k+1}$, then by the definition of δ'' in Lemma 6,

$$z^{k+1} \notin B(z^i, \delta''(z^i)).$$

If $i \in I^{k+1}$, then since

$$z^{k+1} \in Z^{k+1} \subset C^i,$$

where $C^i \in \varphi(z^i)$,

$$z^{k+1} \notin B(z^i, \delta'(z^i)).$$

Therefore, $z^{k+1} \notin B(z^i, \delta^*(z^i))$ and convergence is proved.

ALGORITHM III. This algorithm uses features of both I and II and drops cuts en masse if sufficient improvement has been made in objective value. Let δ be a separator and φ a cut map, both on $X \sim G$.

To begin iteration (k, i) we have $z^1, \cdots, z^k = y_1^k, \cdots, y_i^k$, and Z_i^k where $y_i^k \in \Gamma(Z_i^k)$. If $y_i^k \in G$, terminate; otherwise:

(a) let $C \in \varphi(y_i^k)$;

(b) let $y \in \Gamma(Z_i^k \cap C)$:

(i) if $r(y) \leqq r(z^k) - \delta(z^k)$, then choose Z_1^{k+1} such that $r^*(Z_1^{k+1}) \leqq r(y)$ and let $y_1^{k+1} = z^{k+1} \in \Gamma(Z_1^{k+1})$;

(ii) otherwise, let $Z_{i+1}^k = Z_i^k \cap C$ and $y_{i+1}^k = y$.

Iteration (k, i) is complete.

If the sequence $z^1, z^2, \cdots$ generated is infinite, then

$$r(z^k) \leqq r(z^i) - \delta(z^i)$$

for $i < k$. Hence $z^k \notin B(z^i, \delta'(z^i))$, where δ' is the separator of Lemma 6 induced by r and δ, and convergence follows from Algorithm 0. If the generated sequence $y_1^k, y_2^k, \cdots$ is infinite, then the algorithm relapses into Algorithm I and any cluster point of $y_1^k, y_2^k, \cdots$ solves (1).

ALGORITHM IV. Let $\mathscr{Z}_c$ be a compact subset of $\mathscr{Z}$ which is closed under intersection and such that if $Z \in \mathscr{Z}_c$, then

$$z \in \Gamma(Z), \quad z \notin G \quad \text{imply} \quad \Gamma(Z) = \{z\}. \tag{2}$$

Let $\varphi: X \sim G \to \mathscr{Z}_c$ be a cut map. The algorithm is:

Assume $Z^i \in \mathscr{Z}_c$ and $z^i \in \Gamma(Z^i)$ for $i = 1, \cdots, k$ have been generated. Let $C^k \in \varphi(z^k)$ and select $Z^{k+1} \in \mathscr{Z}_c$ such that

$$r^*(Z^{k+1}) \leqq r^*(Z^k \cap C^k).$$

Let $z^{k+1} \in \Gamma(Z^{k+1})$.

To prove convergence, suppose $z^k \to z^\infty \notin G$ on some subsequence K. Without loss of generality and using Lemma 3, we can assume that $Z^k \to Z^\infty$ and $C^k \to C^\infty$ on the same subsequence K. By Lemma 7 and (2), we have $\{z^\infty\} = \Gamma(Z^\infty)$. By Lemma 4, $z^\infty \notin C^\infty$ and we have $r^*(Z^\infty \cap C^\infty) < r(z^\infty)$. But consider $r(z^\infty) = \lim r(z^k) = \lim r(z^{k+1}) \leqq \lim r^*(Z^k \cap C^k) \leqq r^*(Z^\infty \cap C^\infty) < r(z^\infty)$. This is a contradiction.

Observe that the bonafide example which meets the conditions of Algorithm IV is where G, X, and all cuts are convex and where r is strictly quasi-concave.

3. Computational procedures. The methods, Algorithms I–IV, of § 2 will now be further specialized into the specific computational algorithms KCG, VN, and DW. In each of these procedures, X will be a compact polyhedral convex subset of E^n, G will be a compact convex subset of X, and cuts will be of the form $X \cap H \supset G$, where H is a half-space of E^n.

The cuts for the procedures are generated from a common form. In detail, let us define the half-space

$$H(v) = \{x | a(v) + b(v)x \geqq 0\} \subset E^n$$

for given functions $a: Y \to E^1$, and $b: Y \to E^n$ where $Y \subset E^m$. Then using the point to set map $\alpha: X \sim G \to Y$, specify the point to set map $\varphi: X \sim G \to \mathscr{Z}$ by mapping z to

$$\{H(v) \cap X | v \in \alpha(z)\}.$$

From a point z, after calculating $v \in \alpha(z)$, we shall construct the cutting half space $H(v)$. The next proposition demonstrates that φ will be a cut map if a and b are continuous functions, the map α is closed, and if the half spaces "cut" off the point z.

LEMMA 8. *Let $\varnothing \neq G \subset X$ be compact convex sets in E^n, $Y \subset E^m$ be compact, and define the map $\varphi: X \sim G \to \mathscr{Z}$ by*

$$\varphi(z) = \{H(v) \cap X | v \in \alpha(z)\}$$

where

$$H(v) = \{x | a(v) + b(v)x \geqq 0\}.$$

Suppose:

(i) *the functions $a: Y \to E^1$ and $b: Y \to E^n$ are continuous and the map $\alpha: X \sim G \to Y$ is closed*;

(ii) *$z \in X \sim G$ and $H \cap X \in \varphi(z)$ imply $z \notin H$.*

Then φ is a cut map.

Proof. Define $\delta(z)$ for $z \in X \sim G$ to be

$$\inf \{d(z, y) | y \in H(v) \cap X, v \in \alpha(z)\}.$$

We shall show δ to be a separator and establish the result; assume that $z^k \to z^\infty$ and $\delta(z^k) \to 0$. By compactness of X and Y and by closedness of α we can select $v^k \in \alpha(z^k)$ and $y^k \in H(v^k) \cap X$ so that $\delta(z^k) = d(z^k, y^k)$. Then, again, by compactness and closedness, for appropriate subsequences $v^k \to v^\infty \in \alpha(z^\infty)$ and $y^k \to y^\infty$. By continuity of a and b, $y^\infty \in H(v^\infty) \cap X$. Since $d(z^k, y^k) \to 0$, $y^\infty = z^\infty \in H(v^\infty) \cap X$. But if $z^\infty \notin G$, then by hypothesis, $z^\infty \notin H(v^\infty)$, which is a contradiction, thus completing the proof.

The map α and functions a and b will differ for the various procedures KCG, VN, and DW. The sets Z^k will always have the form

$$X \cap (\bigcap_{i \in I} H^i),$$

where $H^i \cap X$ is a cut. The objective function r will be linear for Algorithms I, II, and III, and consequently, finding $z^k \in \Gamma(Z^k)$ will be a linear program. In Algorithm IV we shall let $\mathscr{Z}_c$ be the set of convex sets in $\mathscr{Z}$; however, a linear function would not necessarily meet the condition (2). Here, in general, r will not be linear.

To specify KCG, VN, and DW and their variants in terms of Algorithms I–IV, we need to define the goal set G, the polyhedral set X which encloses G, a cut map φ, an objective function r, and a separator δ. Algorithm I for KCG, VN, and DW corresponds to the original statements of these algorithms. Algorithms II and III are variants which permit deletion of cuts, but again the number of cuts which will be dropped is problem dependent, and at this time, is essentially unpredictable. These procedures, II and III, will probably be very useful for some problems and worthless for others, but in any case the effort to implement them seems comparatively small. Algorithm IV requires more conditions than the others, namely (2), but offers the advantage that all inactive cuts can be dropped.

The Kelley, Cheney, Goldstein concave cutting plane method (KCG). KCG is designed to solve the program

$$\begin{aligned} &\text{maximize} \quad qx = \sum q_i x_i \\ &\text{subject to} \quad g(x) \geqq 0, \end{aligned} \tag{3}$$

where g is real-valued, continuous, and concave. That the problem is equivalent to the general concave programming problem is discussed in [9]. We define G as the feasible set

$$G = \{x | g(x) \geqq 0\}$$

and assume that G is contained in the compact set X, where

$$X = \{x | Ax \geqq b\}.$$

Let $U(z)$ be the set of vectors u such that

$$g(x) \leqq g(z) + u(x - z)$$

for all $x \in X$. We assume that $U(z) \neq \varnothing$ for $z \in X$ and that

$$\bigcup_{z \in X} U(z)$$

is bounded. To define the cut map φ let α be the closed map

$$\alpha(z) = \{(z, u) \mid u \in U(z)\}$$

and let

$$\varphi(z) = \{H(z, u) \cap X \mid (z, u) \in \alpha(z)\},$$

where

$$H(z, u) = \{x \mid g(z) + u(x - z) \geqq 0\}.$$

Since α is closed, using Lemma 8 we see that φ is a cut map.

For Algorithms I, II, and III we let $r(x) = qx$. For Algorithm IV we assume that we have a continuous function p on X such that $r(x) = qx + p(x)$ satisfies (2) (where $\mathscr{Z}_c$ is the collection of convex sets in $\mathscr{Z}$) and that $r(x) = qx$ for $x \in G$ (for the procedure to be practicable, we would probably also need r to be quasi-concave on X). Possible candidates for p would be functions of type $\min(0, g(z))$ or $-(\min(0, g(z)))^2$.

For separators for Algorithms II and III one could use, for example, $\delta(z) = g^2(z)$ or $-g(z)$.

Each method, of course, generates sequences z^k whose cluster points solve (3). We now discuss the individual algorithms in detail.

Algorithm I for the KCG cut map. Algorithm I is the original KCG method. Let $Z^1 = X$; then given Z^k,

$$Z^{k+1} = Z^k \cap H^k,$$

where $H^k \in \varphi(z^k)$ and $z^k \in \Gamma(Z^k)$.

Explicitly, Z^k will have the form

$$Z^k = \{x \mid Ax \geqq b, g(z^i) + u^i(x - z^i) \geqq 0, i = 1, \cdots, k - 1\}.$$

Then, given Z^k, solve the problem

$$\underset{x \in Z^k}{\text{maximize}}\ qx$$

for the optimal point z^k. If $g(z^k) \geqq 0$, so that $z^k \in G$, terminate. Otherwise, obtain $u^k \in U(z^k)$, set

$$H^k = \{x \in X \mid g(z^k) + u^k(x - z^k) \geqq 0\},$$

and let

$$Z^{k+1} = Z^k \cap H^k.$$

Observe that the dimension of the linear programming subproblem increases by one each iteration.

Algorithm II with the KCG cut map. Use any separator δ. The set Z^k will be of the form

$$Z^k = X \cap \left(\bigcap_{i \in I^k} H^i\right) = \left\{ x \,\middle|\, \begin{array}{l} Ax \geqq b \\ g(z^i) + u^i(x - z^i) \geqq 0, i \in I^k \end{array} \right\},$$

and $z^k \in \Gamma(Z^k)$ will be given. Calculate

$$z^{k+1} \in \Gamma(Z^k \cap H^k),$$

where $H^k \in \varphi(z^k)$. Define $J \subset \{k\} \cap I^k$ to be the set of indices i such that *both*

$$g(z^i) + u^i(z^{k+1} - z^i) > 0$$

and

$$r(z^{k+1}) \leqq r(z^i) - \delta(z^i).$$

Then let

$$I^{k+1} = I^k \cup \{k\} \sim J.$$

Algorithm III with the KCG cut map. Given $z^1, \cdots, z^k = y_1^k, \cdots, y_i^k$, and Z_i^k, we iterate.

(a) Let $H \in \varphi(y_i^k)$.

(b) Let $y \in \Gamma(Z_i^k \cap H)$.

(i) If

$$r(y) \leqq r(z^k) - \delta(z^k),$$

set $y_1^{k+1} = z^{k+1} = y$ and let Z^{k+1} be the intersection of X with the active cuts at z^{k+1}.

(ii) Otherwise, let $Z_{i+1}^k = Z_i^k \cap H$ and $y_{i+1}^k = y$.

The iteration is complete and the next step is (a).

Briefly, at Z^k add cuts until the optimal objective function value of the subproblem drops by $\delta(z^k)$. Determine the active constraints at the corresponding optimal point and intersect these with X to form Z^{k+1}.

Algorithm IV and KCG. Assuming that $r(z) = qz + p(z)$ satisfies (2) and that $r(z) = qz$ for $z \in G$, the specifics of Algorithm IV are straightforward. Given Z^k and $z^k \in \Gamma(Z^k)$, compute $z^{k+1} \in \Gamma(Z^k \cap H^k)$, where $H^k \in \varphi(z^k)$. Then drop the inactive cuts at z^{k+1} so that Z^{k+1} will be X intersected with the active half-spaces at z^{k+1}.

Observe that at most n (the dimension of the space) cuts need to be retained (although it might be necessary to drop some active but redundant cuts). Of course, this procedure could also be applied to solve

$$\begin{aligned} &\text{maximize} \quad r(x) \\ &\text{subject to} \quad g(x) \geqq 0, \end{aligned}$$

where r is strictly quasi-concave on X.

The Veinott supporting hyperplane method (VN). VN is designed to solve

$$\begin{aligned} &\text{maximize} \quad qx \\ &\text{subject to} \quad g_i(x) \geqq 0, \qquad i = 1, \cdots, m, \end{aligned} \tag{4}$$

where the g_i are quasi-concave and continuously differentiable. As in KCG, we let

$$G = \{x | g_i(x) \geqq 0, i = 1, \cdots, m\}$$

and assume that

$$X = \{x | Ax \geqq b\}$$

is given and that $G \subset X$ are compact. It is assumed that $g_i(a) > 0$ for $i = 1, \cdots, m$ for some a and that $g_i(u) = 0$ implies $\nabla g_i(u) \neq 0$ for $i = 1, \cdots, m$ for each boundary point u of G.

Given any point $z \in X \sim G$, let $u(z)$ be the point on the boundary of G that is on the line segment between z and a. Note that u is continuous on $X \sim G$. Let $\alpha(z)$ be the set

$$\{(u, j) | g_j(u) = 0, j = 1, \cdots, m\}.$$

We define the cut map φ by assigning to z the cuts $H \cap X$, where

$$H = \{x | \nabla g_j(u)^t (x - u) \geqq 0\},$$

where $(u, j) \in \alpha(z)$; see Lemma 8.

The objective r for Algorithms I, II, III, and IV are chosen exactly as in KCG. Here a candidate for the function p is

$$-\sum_1^m (\min(0, g_i(z)))^2.$$

For a separator, Algorithms II and III can use

$$\delta(z) = \sum_1^m (g_i(z))^2.$$

Algorithm I for these G, X, φ, r is the original VN. Again, Algorithms II and III provide a vehicle for dropping cuts. Algorithm IV, if p is available, only requires that at most n constraints be retained; the subproblems, nevertheless, are nonlinear. Each of these procedures solves (4).

The Algorithm IV version of VN can also be used to solve the problem

$$\begin{aligned} &\text{maximize} && r(x) \\ &\text{subject to} && g_i(x) \geqq 0, && i = 1, \cdots, m, \end{aligned}$$

where r is strictly quasi-concave.

Dantzig–Wolfe method (DW). The DW method solves the concave programming problem

$$\begin{aligned} &\text{maximize} && f(t) \\ &\text{subject to} && g(t) \geqq 0, && t \in T, \end{aligned} \tag{5}$$

where $g = (g_1, \cdots, g_m)$, the functions f and g_i are concave and continuous, and the set T is convex and compact. It is further assumed that there is a given point $\bar{t} \in T$ for which $g(\bar{t}) > 0$.

As in [9] we shall deviate slightly from the Dantzig and Wolfe attitude in [3] by treating the dual of the problem, which is

$$\begin{aligned} &\text{minimize} && u \\ &\text{subject to} && (\lambda, u) \in H(t) && \text{for all } t \in T \end{aligned} \tag{6}$$

and

$$\lambda \geqq 0,$$

where $\lambda = (\lambda_1, \cdots, \lambda_m) \in E^m$, $u \in E^1$, and $H(t) = \{(\lambda, u) | f(t) + \lambda g(t) \leqq u\}$. Here the constraints range over the entire set T.

Observe that we can adjoin the additional constraint

$$u \leqq \bar{u},$$

where

$$\bar{u} \geqq \max \{f(t) | t \in T\}$$

without altering the solution set. Moreover, because $g(\bar{t}) > 0$, this additional constraint produces a compact feasible region, as is shown in [3] or [9].

We identify the dual problem with (1) by letting

$$X = \{(\lambda, u) | f(\bar{t}) + \lambda g(\bar{t}) \leqq u, u \leqq \bar{u}, \lambda \geqq 0\},$$

$$G = \{(\lambda, u) | (\lambda, u) \in H(t) \text{ for all } t \in T\} \cap X$$

and $r(\lambda, u) = -u$. Consequently, determining $z^* \in \Gamma(G)$ will solve the dual problem.

For each of the algorithms, the Z^k sets will be of the form $X \cap (\bigcap_{i \in I^k} H(t^i))$, where I^k is an appropriate index set. Thus, if we let $t^0 = \bar{t}$ and always maintain $0 \in I^k$ for $k = 0, 1, 2, \cdots$, then the subproblems, $\max \{r(z) : z \in Z^k\}$, are the linear programs

$$\begin{aligned} &\text{maximize} && -u \\ &\text{subject to} && f(t^i) + \lambda g(t^i) \leqq u, \qquad i \in I^k, \quad \lambda \geqq 0. \end{aligned}$$

Note we have dropped the constraint $u \leqq \bar{u}$ as it clearly does not alter the solution.

Given $(\lambda, u) \in X \sim G$ the map $\alpha(\lambda, u)$ determines the set of all $t^* \in T$ that solve

$$\underset{t \in T}{\text{maximize}} \quad f(t) + \lambda g(t).$$

Then for $t^* \in \alpha(\lambda, u)$, we specify H, where $H \cap X \in \varphi(\lambda, u)$, by

$$H = \{(\lambda, u) | f(t^*) + \lambda g(t^*) \leqq u\}.$$

This calculation of H determines the cut map φ for DW.

It follows easily that the procedures converge to a solution of the dual problem (6).

For the separator of Algorithms II and III use the one induced by the cut map; that is, for each $(\lambda, u) \in X \sim G$ let

$$\delta(\lambda, u) = \varepsilon(f(t^*) - \lambda g(t^*) - u)$$

for some $t^* \in \alpha(\lambda, u)$, where $0 < \varepsilon < 1$ is fixed (here a small ε seems best). Specifications of Algorithms I, II and III are precisely analogous to the KCG method except for the different cut map φ and separator δ. Algorithm I is the original DW method.

Although Algorithms I, II and III solve the dual problems, these procedures simultaneously solve the primal problem (5) as observed in [3] or more directly in [9].

Remark. This paper has developed cutting plane methods from the underlying principle of separators. However, it is evident that this principle is applicable not just to cutting plane methods but to nonlinear programming in general.

REFERENCES

[1] C. BERGE, *Topological Spaces*, Oliver and Boyd, Edinburgh and London, 1963.
[2] E. W. CHENEY AND A. A. GOLDSTEIN, *Newton's method of convex programming and Tchebycheff approximation*, Numer. Math., 1 (1959), pp. 253–268.
[3] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton University Press, Princeton, N.J., 1963.
[4] F. HAUSDORFF, *Set Theory*, 2nd ed., Chelsea, New York, 1962.
[5] J. E. KELLEY, *The cutting-plane method for solving convex programs*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 703–712.
[6] D. M. TOPKIS, *Cutting plane methods without nested constraint sets*, Operations Res., 18 (1970), pp. 404–413.
[7] ———, *A note on cutting plane methods without nested constraint sets*, Ibid., to appear.
[8] A. F. VEINOTT, JR., *The supporting hyperplane method for unimodal programming*, Ibid., 15 (1967), pp. 147–152.
[9] W. I. ZANGWILL, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1969.

SIAM J. CONTROL
Vol. 9, No. 4, November 1971

A NOTE ON COMPLETE CONTROLLABILITY*

INGE TROCH†

Abstract. For linear autonomous multivariable control systems the conditions are studied under which the system is always—that is, for every control matrix of appropriate rank—completely controllable, which means that for every initial state one can construct a control which steers the system to any given final state in finite time. The resultant necessary and sufficient condition says that the number of independent controls either must equal the system's dimension or must be smaller by one. In the latter case no eigenvalue of the system matrix may be real. These results are valid for continuous systems as well as for sampled-data systems and are derived by means of well-known criteria for complete controllability and some theorems of matrix theory. Analogous results which are valid for the question of complete observability are formulated too. As a special case, previous results of Hsin Chu are covered.

1. Introduction. In a previous work [1] Hsin Chu has shown that linear continuous multivariable control systems, which can be steered by a single control, are always completely controllable if and only if the dimension of the control system equals two and the eigenvalues of the system matrix are not real. In this work, for control systems of dimension n with m independent controls, conditions will be studied which guarantee that the control system is always completely controllable. By the principle of duality similar theorems for complete observability can be formulated. For the concepts mentioned above and terminology the reader is referred to [5] or [7].

2. Notation and formulation of the problem. In the following, linear multivariable control systems will be studied, which can be described by

$$\dot{x} = Ax + Bu, \qquad y = Cx + Du, \tag{1}$$

where x denotes the n-dimensional state vector ($x \in R^n$), $u \in R^m$ the control, and $y \in R^s$ the output vector; A, B, C, D are constant matrices of appropriate dimension. Moreover, λ denotes an eigenvalue of the matrix A, that is, a solution of the equation ($I = \operatorname{diag}(1, 1, \cdots, 1)$, unity matrix)

$$\det(A - \lambda I) = 0; \tag{2}$$

$(\cdot, \cdot)$ denotes the scalar product of two vectors; A^T denotes the transpose of the matrix A; and a_{ik} denotes the elements of the matrix A. As has been said, the system has m independent controls, which means that the matrix B is of rank m. Without loss of generality we may restrict ourselves to matrices B with m linearly independent column vectors $b_1, b_2, \cdots, b_m$. There is now the problem of deriving conditions on A and m so that for every matrix B with

$$\operatorname{rank} B = m, \tag{3}$$

the system (1) is completely controllable in the sense introduced by Kalman [5]. We note that, speaking intuitively, the system (1) is called completely controllable if, knowing the matrices A and B and the initial state, one can construct a control u which will bring the state to the zero state in finite time.

* Received by the editors November 3, 1970, and in revised form March 15, 1971.

† Institut für Mathematik der Technischen Hochschule in Wien, Vienna, Austria.

3. The condition on A. As is well known [7] a necessary and sufficient condition for complete controllability of (1) is

$$\text{rank } Q := \text{rank } (B, AB, \cdots, A^{n-1}B) = n. \tag{4}$$

As the formulated problem is quite trivial for $m = n$, this case will be omitted at first.

Assume now, that A has at least one real eigenvalue λ_1. Then there exists a matrix T over the real field, so that

$$A^{(1)} = TAT^{-1}$$

has only one nonzero element in its last line:

$$\begin{aligned} a_{nk}^{(1)} &= 0, \qquad k = 1, 2, \cdots, n-1, \\ a_{nn}^{(1)} &= \lambda_1. \end{aligned} \tag{5}$$

Now let B be chosen in such a way that for the elements of the corresponding matrix $B^{(1)} = TB$ the equations

$$b_{nk}^{(1)} = 0 \quad \text{for } k = 1, 2, \cdots, m$$

hold. Then the last line of Q (built with $A^{(1)}$ and $B^{(1)}$) has no nonzero element, and therefore Q cannot be of rank n. We have the first necessary condition.

FIRST NECESSARY CONDITION. *The system* (1) *can be completely controllable for every matrix B of rank $m < n$ only if no eigenvalue of the matrix A is real.*

4. The condition on the rank B. Note first that there exists a regular matrix T_1 over the real field such that A is similar to a generalized diagonal matrix

$$A^{(2)} = T_1 A T_1^{-1} = \text{diag } (P_1, \cdots, P_k),$$

where $P_j, j = 1, 2, \cdots, k$, corresponds to the jth elementary divisor (over the real field!) of A. If $\mu_j \pm i\nu_j (\nu_j \neq 0$ because of § 3) denotes the corresponding eigenvalues, then P_j is of the form

$$\begin{pmatrix} D_j & I & 0 & \cdots & 0 \\ & & & \cdots & \\ 0 & & & \cdots & D_j & I \\ 0 & & & & & D_j \end{pmatrix} \tag{6}$$

with

$$D_j = \begin{pmatrix} \mu_j & \nu_j \\ -\nu_j & \mu_j \end{pmatrix}, \qquad j = 1, \cdots, k.$$

For the proof consult [2, p. 106].

Assume now that $m \leqq n - 2$, and take a matrix B of rank m with ($B^{(2)} = T_1 B$)

$$b_{n-1,j}^{(2)} = b_{n,j}^{(2)} = 0 \quad \text{for } j = 1, 2, \cdots, m. \tag{7}$$

It then follows that the last two lines of the matrix $A^{(2)}B^{(2)}$—and also of $A^{(2)2}B^{(2)}, \cdots, A^{(2)n-1}B^{(2)}$—have no nonzero elements, and therefore the rank of Q (built with $A^{(2)}$ and $B^{(2)}$) is less than or equal to $n - 2$. This gives the second necessary condition.

SECOND NECESSARY CONDITION. *The system* (1) *can be completely controllable for every matrix B of rank m only if* $m \geqq n - 1$.

5. Necessary and sufficient condition. It shall now be shown that the two necessary conditions together are sufficient too, that is, the validity of the following theorem is shown.

THEOREM 1. *For every matrix B of rank m system* (1) *is completely controllable if and only if* (α) *or* (β) *holds*:

(α) $m = n$;

(β) $m = n - 1$, *and no eigenvalue of A is real.*

Proof. Necessity follows from §§ 3 and 4.

The proof of sufficiency is still left: Condition (4) is trivially fulfilled if (α) holds. Now let condition (β) hold and assume that there exists a matrix B of rank $n - 1$ such that

$$\text{rank } Q = n - 1. \tag{8}$$

Note first that there exists a vector $c \in R^n$, which is orthogonal to every vector b_j of B:

$$(c, b_j) = 0 \quad \text{for } j = 1, 2, \cdots, n - 1, \tag{9}$$

and which is unique except for multiplication by a scalar. From (8) it follows that every vector Ab_j, $j = 1, 2, \cdots, n - 1$, is linearly dependent on the vectors $b_1, \cdots, b_{n-1}$. But from this, (9) and the properties of the scalar product, we have also

$$(c, Ab_j) = 0 \quad \text{for every } j = 1, 2, \cdots, n - 1. \tag{10}$$

Now let S_{n-1} be the $(n - 1)$-dimensional subspace of R^n spanned by the vectors $b_1, \cdots, b_{n-1}$. Then R^n can be written as the direct sum (see e.g. [4])

$$R^n = S_{n-1} \oplus c.$$

Conditions (9) and (10) can be written as

$$B^T c = 0 \tag{9'}$$

and

$$(AB)^T c = 0, \tag{10'}$$

and the last equation is equivalent to

$$(B^T A^T)c = B^T(A^T c) = 0. \tag{11}$$

From (9'), (11) and the uniqueness of c (in the defined sense) it follows that

$$A^T c = \lambda c \quad \text{for some } \lambda \in R,$$

which means that A possesses c as a left eigenvector belonging to the real eigenvalue λ, in contradiction to assumption (β). This completes the proof.

6. The dual statement. We shall now make use of the second equation in (1) and state a theorem—dual to Theorem 1—which gives necessary and sufficient

conditions for the system (1) to be completely observable for every matrix C with appropriate rank. We just remind the reader that complete observability means that the knowledge of (1) and the output $y(t)$ over a finite interval is sufficient to determine uniquely the initial state of the system for every initial state. The precise definition can again be found for example in [7]. By the principle of duality we have the following theorem.

THEOREM 2. *System* (1) *is completely observable for every matrix* C *of rank* s *if and only if* (γ) *or* (δ) *holds*:

(γ) $s = n$;

(δ) $s = n - 1$, *and no eigenvalue of* A *is real.*

7. Sampled-data systems. As is well known, for the discrete sampled-data system

$$\begin{aligned} x((k+1)T) &= Ax(kT) + Bu(kT), \\ y(kT) &= Cx(kT) + Du(kT), \qquad k = 0, 1, 2, \cdots, \end{aligned} \tag{12}$$

with regular matrix A, equation (4) is a necessary and sufficient condition for complete controllability too. This means that Theorems 1 and 2 remain valid and unchanged for the multivariable control system described by (12).

REFERENCES

[1] HSIN CHU, *A remark on complete controllability*, this Journal, 3 (1966), pp. 439–442.

[2] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[3] F. R. GANTMACHER, *Matrizenrechnung*, Bd.I, Deutscher Verlag der Wissenschaften, Berlin, 1958.

[4] W. GRÖBNER, *Matrizenrechnung*, BI-Hochschultaschenbücher, Bd. 103/103a, Mannheim, 1966.

[5] R. E. KALMAN, *On the general theory of control systems*, Proc. First International Congress on Automatic Control, Eyre and Spottiswoode, London, 1961.

[6] R. E. KALMAN, Y. C. HO AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, vol. I, John Wiley, New York, 1961, pp. 189–213.

[7] L. A. ZADEH AND C. A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.

SIAM J. CONTROL
Vol. 9, No. 4, November 1971

ABSTRACT MODELS FOR THE SYNTHESIS OF OPTIMIZATION ALGORITHMS*

GERARD G. L. MEYER AND E. POLAK†

Abstract. In this paper, we present a systematic approach to the problem of synthesis of optimization algorithms. First, we develop abstract models for algorithms. These models guide the inventive process toward "conceptual" algorithms, namely, algorithms which may consist of operations that are inadmissible in a practical method (for example, a conceptual algorithm may require us to find the limit of an infinite sequence at each iteration). Once the abstract models are established, we present a set of methods for converting "conceptual" algorithms falling into the class defined by the abstract models, into "implementable" iterative procedures.

Introduction. The convergence of optimization algorithms has been studied extensively in recent years (see [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]). The approach generally adopted in these studies consisted of defining a class of algorithms and then giving convergence theorems which applied to every algorithm in this class. This approach has resulted in the development of general procedures which have considerably simplified the task of establishing whether an algorithm is convergent. However, the emphasis so far has been on *analysis*. Very few of the existing results provide guidelines for the *synthesis* of algorithms.

In this paper we present a systematic approach to the problem of synthesis of optimization algorithms. The development of an algorithm usually evolves through three phases. The first is a heuristic, or invention phase in which intuition plays an extremely important part. In the second phase one transforms one's intuitive ideas into a "conceptual" algorithm, i.e., an algorithm which may consist of operations that are inadmissible in a practical method. (For example, a conceptual algorithm may require us to find the limit of an infinite sequence at each iteration.) The last phase consists of converting the "conceptual" algorithm into an "implementable" algorithm. Our approach to the problem of synthesis consists of two parts. First we develop abstract models for algorithms. These models guide the inventive process towards "conceptual" algorithms which will later be easily made implementable. Once the abstract models are established, we present a set of methods for converting "conceptual" algorithms, falling into the class defined by the abstract models, into "implementable" iterative procedures.

One of the most frequently occurring difficulties in the implementation of a conceptual algorithm is the requirement that an implicit relation be solved at each iteration, e.g., minimize a function along a line, maximize a linear function in a convex set, etc. Generally, we only have methods for constructing a sequence whose limit point satisfies such an implicit relation. Now it is well known that no limit point of an infinite sequence can be determined on a digital computer in a finite time. Consequently, the task most frequently encountered in the design of a transition from a "conceptual" algorithm to an "implementable" one is that of finding methods for avoiding the need to construct limit points.

* Received by the editors January 19, 1970, and in final revised form February 12, 1971.

† Department of Electrical Engineering and Computer Sciences, Electronics Research Laboratory, University of California, Berkeley, California 94720. The first author is now at Department of Electrical Engineering, University of Southern California, Los Angeles, California 90007. This work was sponsored by the National Aeronautics and Space Administration under Grant NGL-05-003-016(Sup.6).

In this paper, we propose two methods for obviating the need for constructing infinite sequences in subprocedures. The first is a truncation procedure and is presented in § 2, the second is an ε-approximation procedure which is presented in § 4.

The scope, to which a paper must be held, does not permit us to copiously illustrate the applicability of the ideas presented. It is our hope, however, that the two examples given in § 3 will convince the reader of the great usefulness of the approach described.

1. Abstract models for a class of iterative procedures. Throughout this paper, we shall assume that we are given a closed and bounded subset T of R^n in which we wish to find points with a specific property π. We shall call points in T with the property π *desirable.*

The simplest algorithms for finding desirable points in T are composed of a map $\xi(\cdot)$ from T into R^1 and of a map $A(\cdot)$ from T into all the subsets of T, having the following form.

ALGORITHM 1.

Step 0. Compute a point z_0 in T and set $i = 0$.

Step 1. Compute a point z_{i+1} in $A(z_i)$.

Step 2. If $\xi(z_{i+1}) < \xi(z_i)$, set $i = i + 1$ and go to Step 1; otherwise stop.

DEFINITION 1. We shall say that an iterative procedure of the form of Algorithm 1 is *convergent* if any sequence of points $\{z_i\}$, generated by it, satisfies one of the following conditions:

(i) If the sequence $\{z_i\}$ is finite, i.e., $\{z_i\} = \{z_0, z_1, \cdots, z_k\}$, then z_{k-1} is desirable.

(ii) If the sequence $\{z_i\}$ is infinite, then any of its cluster points is desirable.

It is quite easy to show that Algorithm 1 is convergent under the following assumption (see [11]).

HYPOTHESIS 1 [Zangwill].

(i) z in T is desirable if there exists at least one point a in $A(z)$ such that $\xi(a) - \xi(z) \geqq 0$;[1]

(ii) $\xi(\cdot)$ is continuous on T;

(iii) $A(\cdot)$ is upper semicontinuous on T.[2]

Hypothesis 1 is not the only assumption which ensures that Algorithm 1 is convergent. We now state another assumption which ensures that Algorithm 1 is convergent (see [9]).

HYPOTHESIS 2 [Polak].

(i) z in T is desirable if there exists at least one point a in $A(z)$ such that $\xi(a) - \xi(z) \geqq 0$.

(ii) $\xi(\cdot)$ is bounded from below on T.

(iii) If z in T satisfies $\xi(a) - \xi(z) < 0$ for all a in $A(z)$, then there exist an $\varepsilon > 0$ and a $\delta > 0$ such that $\xi(a') - \xi(z') \leqq -\delta < 0$ for all z' in T such that $\|z' - z\| \leqq \varepsilon$, for all a' in $A(z')$.

[1] Note: Part (i) of Hypothesis 1 simply states that the set of points z in T, for which there exists at least one point a in $A(z)$ satisfying $\xi(a) - \xi(z) \geqq 0$, is a subset of the set of desirable points.

[2] Note: Let $A(\cdot)$ be a map from T into all the subsets of T. If for any sequence $\{y_i\}$ converging to y^* and for any sequence $\{a_i\}$ converging to a^*, with a_i in $A(y_i)$, a^* belongs to $A(y^*)$, then we say that the map $A(\cdot)$ is upper semicontinuous on T.

Remark. It can be shown that Hypothesis 2 is weaker than Hypothesis 1 (see [1]).

In this paper we differentiate between explicit and implicit algorithms. This differentiation is largely *heuristic* but is extremely important in the construction of computationally efficient algorithms.

By an *explicit* algorithm we shall mean an algorithm of the form of Algorithm 1 in which the computation of a point y in $A(z)$ for z in T can be carried out in a reasonably straightforward manner.

Explicit algorithms do not lead to computational difficulties and therefore we shall say no more about them. Implicit algorithms on the other hand cannot be readily implemented on a digital computer, as we shall shortly show, and must therefore be regarded as "conceptual" rather than as "practical" algorithms. The following sections of this paper will be devoted to developing methods for modifying convergent implicit algorithms in such a way as to produce convergent algorithms in explicit form.

We shall consider two abstract models of implicit algorithms. The first one, which is defined below, uses a map $A(\cdot)$ such that to compute a point z in $A(z)$, we must solve an implicit equation. To obtain some motivation for the specific decomposition of $A(\cdot)$ in Definition 2, below, the reader should digress for a moment and examine Problem 1, the steepest descent Algorithm 8 and Definition 5, in § 3.1.

DEFINITION 2. Let $U(\cdot)$ be a map from T into all the subsets of T and let $\alpha(\cdot)$ and $\gamma(\cdot)$ be maps from T into R^1. Then, for every z in T, we define the set $A(z)$ as consisting of all y in $U(z)$ such that $\gamma(y) = \alpha(z)$.

The following assumption ensures that the set $A(z)$ is not empty.

HYPOTHESIS 3. Given any point z in T, there exists a point y in $U(z)$ such that $\alpha(z) = \gamma(y)$.

In this case, Algorithm 1 takes on the following expanded form.

ALGORITHM 2.

Step 0. Compute a point z_0 in T and set $i = 0$.

Step 1. Compute a point z_{i+1} in $U(z_i)$, satisfying $\gamma(z_{i+1}) = \alpha(z_i)$.

Step 2. If $\xi(z_{i+1}) < \xi(z_i)$, set $i = i + 1$ and go to Step 1; otherwise stop.

HYPOTHESIS 4.

(i) A point z in T is desirable if there exists at least one point y in $U(z)$ such that $\alpha(z) = \gamma(y)$ and $\xi(y) - \xi(z) \geqq 0$.

(ii) The maps $\alpha(\cdot)$, $\gamma(\cdot)$ and $\xi(\cdot)$ are continuous on T.

(iii) The map $U(\cdot)$ is upper semicontinuous on T.

The proof of the following proposition is easy and has been omitted.

PROPOSITION 1. *If the maps* $\alpha(\cdot)$, $\gamma(\cdot)$, $\xi(\cdot)$, *and* $U(\cdot)$ *satisfy Hypotheses* 3 *and* 4, *then the map* $A(\cdot)$ *given by Definition* 2 *is upper semicontinuous on* T *and Algorithm* 2 *is convergent.*

We now present the second specific form of the map $A(\cdot)$ that we wish to consider. This form is characterized by the fact that, to find points in $A(z)$, we must compute intermediate points. To obtain some motivation for the specific decomposition of the map $A(\cdot)$ in Definition 3, below, the reader should digress for a moment and examine Problem 2, the Frank and Wolfe algorithm (Algorithm 11) and Definition 7, in § 3.2.

DEFINITION 3. Let U be a closed and bounded subset of R^n, let $\beta(\cdot,\cdot)$ be a map from $T \times U$ into R, let $b(\cdot,\cdot)$ be a map from $T \times U$ into T and let $\alpha(\cdot)$ be a map from T into R. Then for every z in T, we define the set $A(z)$ as follows:

$$A(z) = \{y | y = b(z, w),\ w \in U \text{ such that } \beta(z, w) = \alpha(z)\}.$$

The following assumption ensures that for every z in T the set $A(z)$ is nonempty.

HYPOTHESIS 5. Given any z in T, there exists a point w in U such that $\beta(z, w) = \alpha(z)$.

With the map $A(\cdot)$ defined as above, Algorithm 1 expands as follows.

ALGORITHM 3.

Step 0. Compute a point z_0 in T and set $i = 0$.

Step 1. Compute a point w_i in U, satisfying $\alpha(z_i) = \beta(z_i, w_i)$.

Step 2. Set $z_{i+1} = b(z_i, w_i)$.

Step 3. If $\xi(z_{i+1}) < \xi(z_i)$, set $i = i + 1$ and go to Step 1; otherwise stop.

HYPOTHESIS 6.

(i) A point z in T is desirable if there exists at least one point w in U satisfying $\alpha(z) = \beta(z, w)$ such that $\xi(b(z, w)) - \xi(z) \geqq 0$.

(ii) The maps $\alpha(\cdot)$ and $\xi(\cdot)$ are continuous on T.

(iii) The maps $\beta(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are jointly continuous on $T \times U$.

PROPOSITION 2. *If the maps* $\alpha(\cdot), \beta(\cdot,\cdot), b(\cdot,\cdot)$ *and* $\xi(\cdot)$ *satisfy Hypotheses* 5 *and* 6, *then the map* $A(\cdot)$ *given by Definition* 3 *is upper semicontinuous on* T *and Algorithm* 3 *is convergent.*

The transformation of an implicit algorithm into an explicit one usually depends upon the specific conceptual method one has in mind for finding points in the set $A(z)$. We shall now introduce two infinite subprocedures for calculating points in $A(z)$, and in the next section we shall show how these subprocedures can be truncated to produce convergent explicit algorithms.

We begin by considering a subprocedure for computing points in the set $A(z)$ when the map $A(\cdot)$ is defined as in Definition 2 (we suppose, of course, that we are unaware of a more straightforward method for calculating points in $A(z)$). Let us denote by N the set of positive integers and suppose that we have a mapping $m(\cdot,\cdot,\cdot)$ from $T \times T \times N$ into T which satisfies the following assumption.

HYPOTHESIS 7.

(i) $m(z, y, j)$ belongs to $U(z)$ for all z in T, y in $U(z)$ and j in N.

(ii) The sequence $\{\gamma(m(z, y, j))\}_{j=0}^{\infty}$ converges to $\alpha(z)$ for all z in T and y in $U(z)$.[3]

Clearly, for any z in T and y in $U(z)$, every cluster point of the sequence $\{m(z, y, j)\}_{j=0}^{\infty}$ is in $A(z)$ given by Definition 2. When the map $m(\cdot,\cdot,\cdot)$ is introduced, Algorithm 2 assumes the following specific form.

ALGORITHM 4.

Step 0. Compute a point z_0 in T and set $i = 0$.

Step 1. Compute a point y_i in $U(z_i)$ and let z_{i+1} be any cluster point of the sequence $\{m(z_i, y_i, j)\}_{j=0}^{\infty}$.

[3] Note that part (ii) of Hypothesis 7 does not imply that the sequence $m(z, y, j)$ converges.

Step 2. If $\xi(z_{i+1}) < \xi(z_i)$, set $i = i + 1$ and go to Step 1; otherwise stop.

In view of Proposition 1, the following proposition is obvious.

PROPOSITION 3. *If the maps* $\alpha(\cdot), \gamma(\cdot), \xi(\cdot), U(\cdot)$ *satisfy Hypotheses* 3 *and* 4, *and the map* $m(\cdot,\cdot,\cdot)$ *satisfies Hypothesis* 7, *then Algorithm* 4 *is convergent.*

Remark. Obviously, Algorithm 4 cannot be implemented on a digital computer since it would inevitably jam up in Step 1.

We now consider a subprocedure for computing points in $A(z)$ when the map $A(\cdot)$ is defined as in Definition 3. Thus, suppose that we have a map $m(\cdot,\cdot,\cdot)$ from $T \times U \times N$ into T which satisfies the following assumption.

HYPOTHESIS 8.

(i) $m(z, y, j)$ belongs to U for all z in T, y in U and j in N.

(ii) The sequence $\{\beta(z, m(z, y, j))\}_{j=0}^{\infty}$ converges to $\alpha(z)$ for all z in T and y in U.

When such a map is introduced, Algorithm 3 assumes the following specific form.

ALGORITHM 5.

Step 0. Compute a point z_0 in T and set $i = 0$.

Step 1. Compute a point y_i in U and let w_i be any cluster point of the sequence $\{m(z_i, y_i, j)\}_{j=0}^{\infty}$.

Step 2. Set $z_{i+1} = b(z_i, w_i)$.

Step 3. If $\xi(z_{i+1}) < \xi(z_i)$, set $i = i + 1$ and go to Step 1; otherwise stop.

In view of Proposition 2, the following proposition is obvious.

PROPOSITION 4. *If the maps* $\alpha(\cdot)$, $\beta(\cdot,\cdot)$, $\xi(\cdot)$ *and* $b(\cdot,\cdot)$ *satisfy Hypotheses* 5 *and* 6 *and the map* $m(\cdot,\cdot,\cdot)$ *satisfies Hypothesis* 8, *then Algorithm* 5 *is convergent.*

Again it is clear that if implemented on a digital computer, Algorithm 5 would inevitably jam up in Step 1.

2. Truncation methods. As we have just seen, Algorithms 4 and 5 will inevitably jam up in Step 1 since it is impossible to compute cluster points of infinite sequences in a finite time by means of a digital computer. Even if one relies on the finite word length of a digital computer to stop calculations after a finite time, this finite time will usually be prohibitively long. Consequently, some sort of truncation procedure must be used in converting these algorithms into a more realistic form.

We begin by defining a class of maps from N into N (the set of all positive integers) which will be called truncation functions.

DEFINITION 4. We shall say that a map $\ell(\cdot)$, from N into N, is a *truncation function* if $\ell(i) \geqq 1$ for all i in N, and given any m in N, there exists a k in N such that $\ell(i) \geqq m$ for all $i \geqq k$, i in N.

We first use truncation functions in Algorithm 4 which then takes on the following form.

ALGORITHM 6. Let $\ell(\cdot)$ be a given truncation function.

Step 0. Compute a point z_0 in T and set $i = 0$.

Step 1. Compute a point y_i in $U(z_i)$ and set $z_{i+1} = m(z_i, y_i, \ell(i))$.

Step 2. If $\xi(z_{i+1}) < \xi(z_i)$, set $i = i + 1$ and go to Step 1; otherwise stop.

Remark. In Step 1 of Algorithm 6, it is required that a point y_i be found in $U(z_i)$. We suppose that this task is easy. In fact, in almost all applications, z belongs to $U(z)$, and a natural choice for y_i in $U(z_i)$ consists in letting $y_i = z_i$. In order to

ensure that cluster points of infinite sequences generated by Algorithm 6 are desirable the following assumption must be made.

HYPOTHESIS 9.

(i) The sequence $\{\gamma(m(z, y, j))\}_{j=0}^{\infty}$ is monotonically decreasing for all z in T and y in $U(z)$.

(ii) Given any z in T, y in $U(z)$, and $\delta > 0$, there exists an $\varepsilon > 0$ and a k in N, depending on z, y and δ, such that $\gamma(m(z', y', j)) - \gamma(m(z, y, j)) \leqq \delta$ for all z' in T such that $\|z' - z\| \leqq \varepsilon$, for all y' in $U(z')$ such that $\|y' - y\| \leqq \varepsilon$ and for all $j \geqq k$.

(iii) If $\xi(m(z, y, j)) \geqq \xi(z)$ for some y in $U(z)$ and for some $j \geqq 1$, then z is desirable.

PROPOSITION 5. *If the maps* $\alpha(\cdot), \gamma(\cdot), \xi(\cdot), U(\cdot)$ *satisfy Hypotheses* 3 *and* 4, *the map* $m(\cdot, \cdot, \cdot)$ *satisfies Hypotheses* 7 *and* 9 *and the map* $\ell(\cdot)$ *is a truncation function, then Algorithm* 6 *is convergent.*

Proof. Because of part (iii) of Hypothesis 9, the case of finite sequences is trivial.

Now consider an infinite sequence $\{z_i\}$ generated by Algorithm 6, and let z^* be a cluster point of this sequence, i.e., let K_2 be a subset of the integers such that the subsequence $\{z_i\}_{K_2}$ converges to z^*. Consider the sequences $\{z_{i+1}\}_{K_2}$ and $\{y_i\}_{K_2}$ in T. The compactness of T implies that there exists K_1, a subset of K_2, such that the subsequence $\{z_{i+1}\}_{K_1}$ converges to z^{**} in T, and the subsequence $\{y_i\}_{K_1}$ converges to y^* in T. The property of convergent sequences implies that the subsequence $\{z_i\}_{K_1}$ also converges to z^*.

In order to show that z^* is desirable it is enough to establish the three following facts:

(i) z^{**} belongs to $U(z^*)$;

(ii) $\alpha(z^*) = \gamma(z^{**})$;

(iii) $\xi(z^{**}) - \xi(z^*) \geqq 0$.

By construction, $z_{i+1} = m(z_i, y_i, \ell(i))$ and y_i is in $U(z_i)$. It now follows from (i) of Hypothesis 7 that z_{i+1} is in $U(z_i)$. Next, since the map $U(\cdot)$ is upper semi-continuous, we must have z^{**} and y^* in $U(z^*)$.

Let $\delta > 0$; then part (ii) of Hypothesis 7 implies that there exists k_4 in N such that $\gamma(m(z^*, y^*, j)) - \delta/2 \leqq \alpha(z^*)$ for all $j \geqq k_4$. From Hypothesis 9 it follows that there exist an $\varepsilon > 0$ and k_3 such that $\gamma(m(z', y', j)) - \gamma(m(z^*, y^*, j)) \leqq \delta/2$ for all z' in T such that $\|z' - z^*\| \leqq \varepsilon$, for all y' in $U(z')$ such that $\|y' - y^*\| \leqq \varepsilon$ and for all $j \geqq k_3$. Let $k_2 = \max(k_3, k_4)$; then there exists k_1 in N such that $\gamma(m(z_i, y_i, k_2)) - \delta \leqq \alpha(z^*)$ for all $i \geqq k_1$, i in K_1. Since the map $\ell(\cdot)$ is a truncation function, there exists a k_0 in N such that $\ell(i) \geqq k_2$ for all $i \geqq k_0$, i in K_1. It follows from part (i) of Hypothesis 9 that $\gamma(m(z_i, y_i, \ell(i))) - \delta \leqq \alpha(z^*)$ for all $i \geqq k$, i in K_1, where $k = \max(k_0, k_1)$. But $\gamma(\cdot)$ is continuous and therefore $\gamma(z^{**}) - \delta \leqq \alpha(z^*)$. Since this is true for any $\delta > 0$, it follows that $\gamma(z^{**}) \leqq \alpha(z^*)$.

Now, given any z_i and y_i in $U(z_i)$, part (i) of Hypothesis 9 implies that $\gamma(m(z_i, y_i, j)) \geqq \alpha(z_i)$ for all j in N, i.e., $\gamma(m(z_i, y_i, \ell(i))) \geqq \alpha(z_i)$. It follows that $\gamma(z^{**}) \geqq \alpha(z^*)$, and therefore, $\gamma(z^{**}) = \alpha(z^*)$.

Now suppose that $\xi(z^{**}) - \xi(z^*) < 0$. The continuity of $\xi(\cdot)$ implies that there exist a $\delta > 0$ and an $\varepsilon > 0$ such that

$$\xi(z'') - \xi(z') \leqq -\delta < 0$$

for all z'' in T such that $\|z'' - z^{**}\| \leqq \varepsilon$, for all z' in T such that $\|z' - z^*\| \leqq \varepsilon$. It follows that there exists a k such that $\xi(z_{i+1}) - \xi(z_i) \leqq -\delta$ for all $i \geqq k$, i in K_1. The sequence $\{\xi(z_i)\}$ is monotonically decreasing, the set T is bounded, and therefore $\{\xi(z_i)\}$ converges to ξ^*, which contradicts the fact that $\xi(z_{i+1}) - \xi(z_i) \leqq -\delta$ for all $i \geqq k$, i in K_1. It follows that $\xi(z^{**}) - \xi(z^*) \geqq 0$ and z^* is desirable.

We now use truncation functions in Algorithm 5 to construct the following form.

ALGORITHM 7. Let $\ell(\cdot)$ be a given truncation function.

Step 0. Compute a point z_0 in T and set $i = 0$.

Step 1. Compute a point y_i in U and set $w_i = m(z_i, y_i, \ell(i))$.

Step 2. Set $z_{i+1} = b(z_i, w_i)$.

Step 3. If $\xi(z_{i+1}) < \xi(z_i)$, set $i = i + 1$ and go to Step 1; otherwise stop.

Remark. In Step 1 of Algorithm 7, it is required that a point y_i be found in the set U. We suppose that this task is easy. In fact, in almost all applications, z belongs to U, and a natural choice for y_i consists in letting $y_i = z_i$.

In order to ensure that cluster points of infinite sequences generated by Algorithm 7 are desirable, the following assumption must be made.

HYPOTHESIS 10.

(i) The sequence $\{\beta(z, m(z, y, j))\}_{j=0}^{\infty}$ is monotonically decreasing for all z in T and y in U.

(ii) Given any z in T, y in U and $\delta > 0$, there exist an $\varepsilon > 0$ and a k in N, possibly depending on z, y, and δ, such that

$$\beta(z', m(z', y', j)) - \beta(z, m(z, y, j)) \leqq \delta$$

for all z' in T such that $\|z' - z\| \leqq \varepsilon$, for all y' in U such that $\|y' - y\| \leqq \varepsilon$ and for all $j \geqq k$.

(iii) If $\xi(b(z, m(z, y, j))) \geqq \xi(z)$ for some y in U and for some $j \geqq 1$, then z is desirable.

The following proposition can be proved easily, following the same type of argument as in the proof of Proposition 5.

PROPOSITION 6. *If the maps* $\alpha(\cdot)$, $\beta(\cdot,\cdot)$, $\xi(\cdot)$, $b(\cdot,\cdot)$ *satisfy Hypotheses* 5 *and* 6, *the map* $m(\cdot,\cdot,\cdot)$ *satisfies Hypotheses* 8 *and* 10 *and the map* $\ell(\cdot)$ *is a truncation function, then Algorithm* 7 *is convergent.*

3. Applications. In order to clarify the concepts and methods exposed in the preceding sections we are now going to examine two specific problems and the algorithms usually used to solve them.

3.1. Unconstrained minimization problems. In this subsection we shall examine the following classical problem.

PROBLEM 1. Find a $\hat{z}$ in R^n such that

$$f(\hat{z}) \leqq f(z) \quad \text{for all } z \text{ in } R^n,$$

where $f(\cdot)$, a convex map from R^n into R^1, is continuously differentiable, with the property that the set $\{z | f(z) \leqq \alpha\}$ is bounded for every α in R^1.

Suppose that we have a point z_0 in R^n. We define T as

$$T = \{y \in R^n | f(y) \leqq f(z_0)\}.$$

In this case, a point $\hat{z}$ in T is desirable if and only if it minimizes $f(z)$ over T, i.e., $f(\hat{z}) \leqq f(z)$ for all z in T. Since $f(\cdot)$ is convex and continuously differentiable, we recognize $\hat{z}$ to be desirable if and only if $\nabla f(\hat{z}) = 0$.

We propose to use the steepest descent method for solving Problem 1, i.e., the following algorithm.

ALGORITHM 8.

Step 0. Let z_0 be a point in R^n and set $i = 0$.

Step 1. Let z_{i+1} be any point satisfying:

(i) z_{i+1} belongs to $U(z_i)$;

(ii)

$$f(z_{i+1}) = \min\{f(y) | y \in U(z_i)\},$$

where

$$U(z_i) = \{y \in R^n | y = z_i + v\nabla f(z_i), v \in [-1, 0]\}.$$

Step 2. If $f(z_{i+1}) < f(z_i)$, set $i = i + 1$ and go to Step 1; otherwise stop.

Algorithm 8 can be seen to be of the form of Algorithm 2, with the maps $\alpha(\cdot), \gamma(\cdot), \xi(\cdot), U(\cdot)$ in Algorithm 2 defined as follows.

DEFINITION 5. Let the maps $\alpha(\cdot), \gamma(\cdot), \xi(\cdot)$ from T into R^1 and the map $U(\cdot)$ from T into all the subsets of T be defined as:

(i) $U(z) = \{y \in T | y = z + v\nabla f(z), v \in [-1, 0]\}$;

(ii) $\alpha(z) = \min\{f(y) | y \in U(z)\}$;

(iii) $\gamma(z) = f(z)$;

(iv) $\xi(z) = f(z)$.

Since by inspection the maps $\alpha(\cdot), \gamma(\cdot), \xi(\cdot)$ and $U(\cdot)$ satisfy Hypotheses 3 and 4, Algorithm 8 is convergent when applied to Problem 1.

At this point, we see the first advantage of using abstract models of algorithms as a tool for proving convergence of algorithms. On the one hand they provide us with a pattern to follow, while on the other, we find that the proof of convergence of a specific algorithm becomes decomposed into fairly simple and independent parts.

In Algorithm 8, the computation of z_{i+1} from z_i is not explicit. We therefore proceed as indicated in § 2 in order to produce truncations in the calculation of z_{i+1} from z_i.

Making use of Proposition 3, we now define a subalgorithm which we shall use in order to modify Algorithm 8.

For every z in R^n, y in R^n and j a positive integer, consider the following subprocedure.

ALGORITHM 9. Let $\bar{\alpha} \in (0, 1)$ be given.

Step 0. Set $y_0 = y, i = 0$ and $v = 0$.

Step 1. If $\langle \nabla f(z), \nabla f(y_i) \rangle = 0$, go to Step 6; otherwise set

$$v = -\operatorname{sgn} \langle \nabla f(z), \nabla f(y_i) \rangle$$

and go to Step 2.

Step 2. Set $\tilde{y} = y_i + v\nabla f(z)$.

Step 3. Compute θ defined by $\theta = f(\tilde{y}) - f(y_i) - \bar{\alpha} v \langle \nabla f(z), \nabla f(y_i) \rangle$.

Step 4. If $\theta \leqq 0$, set $y_{i+1} = \tilde{y}$ and go to Step 5; otherwise set $v = v/2$ and go to Step 2.

Step 5. If $i < j$, set $i = i + 1$ and go to Step 1; otherwise go to Step 6.

Step 6. Set $\tilde{v} = \text{sat } v$,[4] set $y_j = y_0 + \tilde{v}\nabla f(z)$ and stop.

Remark. We choose this subalgorithm because it has a nontrivial amount of structure rather than because it is the best computationally. The structure of this subalgorithm should serve the purpose of illustrating the complexity that can be found in a subalgorithm for computing the values of the map $m(\cdot, \cdot, \cdot)$.

DEFINITION 6. Let $m(\cdot, \cdot, \cdot)$ be the map from $T \times T \times N$ into T defined by $m(z, y, j) = y_j$, where y_j is given by Algorithm 9.

It can be verified (see Appendix), that the map $m(\cdot, \cdot, \cdot)$, given by Definition 6, satisfies Hypotheses 7 and 9. Using the map $m(\cdot, \cdot, \cdot)$ and a truncation function $\ell(\cdot)$, we obtain from Algorithm 8 the following "explicit" algorithm.

ALGORITHM 10. Let $\ell(\cdot)$ be a given truncation function.

Step 0. Compute a z_0 in R^n and set $i = 0$.

Step 1. Set $z_{i+1} = m(z_i, z_i, \ell(i))$.

Step 2. If $f(z_{i+1}) < f(z_i)$, set $i = i + 1$ and go to Step 1; otherwise stop.

In view of Proposition 5, the following is obvious.

PROPOSITION 7. *Algorithm* 10 *is convergent for Problem* 1.

3.2. Constrained minimization problem. In this subsection we shall examine the following classical problem.

PROBLEM 2. Given T a closed, bounded convex subset of R^n, and t, a point in R^n but not in T, find $\hat{z}$ in T such that

$$\|\hat{z} - t\| \leqq \|z - t\| \quad \text{for all } z \text{ in } T.$$

We shall suppose that T is of the form

$$T = \{z \in R^n | f^i(z) \leqq 0,\ i = 1, 2, \cdots, m\},$$

where the maps $f^i(\cdot)$ from R^n into R^1 are continuously differentiable.

Suppose that we try to solve Problem 2 by means of the Frank–Wolfe algorithm.

ALGORITHM 11 (Frank–Wolfe).

Step 0. Compute a point z_0 in T and set $i = 0$.

Step 1. Compute a point w_i in T satisfying $\langle z_i - t, w_i \rangle \leqq \langle z_i - t, w \rangle$ for all w *in* T.

Step 2. Let z_{i+1} be the point in $[z_i, w_i]$ satisfying $\|z_{i+1} - t\| \leqq \|z - t\|$ for all z in $[z_i, w_i]$.[5]

Step 3. If $\|z_{i+1} - t\| < \|z_i - t\|$, set $i = i + 1$ and go to Step 1; otherwise stop.

---

[4] Note: The function sat $(\cdot)$: $R^1 \to R^1$ is defined by

$$\text{sat}(v) = \begin{cases} v & \text{if } |v| \leqq 1, \\ 1 & \text{if } v > 1, \\ -1 & \text{if } v < -1. \end{cases}$$

[5] Note: Given two points x_1 and x_2 in R^n, the set $\{y \in R^n | y = vx_1 + (1 - v)x_2, 0 \leqq v \leqq 1\}$ is denoted by $[x_1, x_2]$.

Algorithm 11 can be seen to be of the form of Algorithm 3 by defining the following maps.

DEFINITION 7. Let the set $U = T$ and the maps $\alpha(\cdot)$, $\xi(\cdot)$ from T into R^1, $\beta(\cdot,\cdot)$ from $T \times U$ into R^1, $b(\cdot,\cdot)$ from $T \times U$ into T, be defined as follows:

(i) $\alpha(z) = \min\{\alpha | \alpha = \langle z - t, w\rangle, w \in T\}$;

(ii) $\xi(z) = \|z - t\|$;

(iii) $\beta(z, w) = \langle z - t, w\rangle$;

(iv) $b(z, w)$ is defined by:

(a) $b(z, w)$ belongs to $[z, w]$;

(b) $\|b(z, w) - t\| \leqq \|y - t\|$ for all y in $[z, w]$.

In order to show that Algorithm 11 is convergent for Problem 2, it suffices to show that the mappings $\alpha(\cdot)$, $\xi(\cdot)$, $\beta(\cdot,\cdot)$ and $b(\cdot,\cdot)$ satisfy Hypotheses 5 and 6, It is easily verified that this is indeed so and we therefore conclude that Algorithm 11 is convergent for Problem 2.

Once again we see the advantage of using an abstract model in proving the convergence of a specific algorithm.

In Algorithm 11, both the point w_i and the point $b(z_i, w_i)$ are defined by implicit relations. However, since the computation of $b(z_i, w_i)$ from z_i and w_i is extremely simple, we shall consider that $z_{i+1} = b(z_i, w_i)$ is an "explicit" function of z_i and w_i. Thus we shall consider that the only real difficulty lies in the computation of w_i.

To obtain from Algorithm 11 an "explicit" algorithm of the form of Algorithm 7, we must introduce a map $m(\cdot,\cdot,\cdot)$ satisfying Hypotheses 8 and 10. For example, one can use a method of feasible directions to define such a map.

We now state a method of feasible directions in the required truncated form.

For every z in T, y in T and positive integer j, consider the following algorithm.

ALGORITHM 12. Suppose that S, a compact neighborhood of the origin in R^n, and ε, a positive scalar, are given. Let $f^0(\cdot)$ be the map from R^n into R^1 defined by $f^0(x) = \langle z - t, x\rangle$ for all x in R^n.

Step 0. Set $y_0 = y$ and $i = 0$.

Step 1. Set $\varepsilon_i = \varepsilon$.

Step 2. Compute ϕ_{ε_i} and h_{ε_i} by solving the following:

$$\phi_{\varepsilon_i} = \min_{h \in S} \max_{j \in J_{\varepsilon_i}} \langle \nabla f^j(y_i), h\rangle.$$

h_{ε_i} is any vector in S such that

$$\phi_{\varepsilon_i} = \max_{j \in J_{\varepsilon_i}} \langle \nabla f^j(y_i), h_{\varepsilon_i}\rangle,$$

where

$$J_{\varepsilon_i} = \{j \in \{1, 2, \cdots, m\} | f^j(y_i) + \varepsilon_i \geqq 0\} \cup \{0\}.$$

Step 3. If $\phi_{\varepsilon_i} < -\varepsilon_i$, set $h_i = h_{\varepsilon_i}$ and go to Step 5; otherwise, compute ϕ_0 defined by

$$\phi_0 = \min_{h \in S} \max_{j \in J_0} \langle \nabla f^j(y_i), h\rangle,$$

where

$$J_0 = \{j \in \{1, 2, \cdots, m\} | f^i(y_i) = 0\} \cup \{0\},$$

and go to Step 4.

Step 4. If $\phi_0 = 0$, set $y_{i+1} = y_i$ and go to Step 7; otherwise, set $\varepsilon_i = \varepsilon_i/2$ and go to Step 2.

Step 5. Compute $\lambda_i \geqq 0$ such that

$$\lambda_i = \max \{\lambda | f^j(y_i + \lambda h_i) \leqq 0 \text{ for all } j = 1, 2, \cdots, m\}.$$

Step 6. Set $y_{i+1} = y_i + \lambda_i h_i$.

Step 7. If $i < j$, set $i = i + 1$ and go to Step 1; otherwise stop.

DEFINITION 8. Let $m(\cdot, \cdot, \cdot)$ be the map from $T \times T \times N$ into T defined by $m(z, y, j) = y_j$, where y_j is given by Algorithm 12.

It can be verified that the map $m(\cdot, \cdot, \cdot)$ given by Definition 8 satisfies Hypotheses 8 and 10 (see [9]). Using the map $m(\cdot, \cdot, \cdot)$ and a truncation function $\ell(\cdot)$ in Algorithm 11, we obtain the following "explicit" method.

ALGORITHM 13. Let $\ell(\cdot)$ be a given truncation function.

Step 0. Compute a point z_0 in T, and set $i = 0$.

Step 1. Set $w_i = m(z_i, z_i, \ell(i))$.

Step 2. Let z_{i+1} be the point in $[z_i, w_i]$ satisfying

$$\|z_{i+1} - t\| \leqq \|z - t\| \quad \text{for all } z \text{ in } [z_i, w_i].$$

Step 3. If $\|z_{i+1} - t\| < \|z_i - t\|$, set $i = i + 1$, and go to Step 1; otherwise stop.

In view of Proposition 6, the following is obvious.

PROPOSITION 8. *Algorithm* 13 *is convergent for Problem* 2.

The use of the approach defined in this paper may show relations between different well-known algorithms. For example, if a function $\ell(\cdot)$ from N into N defined by $\ell(i) = 1$ for all i is used in Algorithm 13 instead of a truncation function, then Algorithm 13 is a method of feasible directions. On the other hand, if a "function" $\ell(\cdot)$ from N into N "defined" by $\ell(i) = \infty$ for all i is used in Algorithm 13 instead of a truncation function, then Algorithm 13 is the Frank–Wolfe algorithm. The use of a truncation function in Algorithm 13 thus produces algorithms which are "between" a method of feasible directions and the Frank–Wolfe algorithm.

4. ε-approximations for a class of iterative procedures. The approach to the synthesis of algorithms described in §§ 1 and 2 is by no means the only one possible. In this section we shall show that an alternative approach exists and gives extremely interesting results.

Throughout this section we shall consider maps $A(\cdot)$ and $\xi(\cdot)$ satisfying Hypothesis 1 in § 1 and we shall suppose that it is impossible to use them in iterative procedures of the form of Algorithm 1 due to the fact that the computation of points z_{i+1} in $A(z_i)$ is impossible (or that we are not aware of ways of doing it).

The main idea developed in this section consists in using a map $B(\cdot, \cdot)$ from $R^1_+ \times T$ into all the subsets of T such that the set $B(0, z)$ is identical to the set $A(z)$ for all z.

We shall suppose that the task of finding a y in $B(\varepsilon, z)$ for $\varepsilon > 0$ and z in T is relatively easy.

Consider the following algorithm.

ALGORITHM 14. Let $\bar{\varepsilon} > 0$ be given.

Step 0. Compute z_0 in T and set $i = 0$.

Step 1. Set $\varepsilon = \bar{\varepsilon}$.

Step 2. Find a point y in $B(\varepsilon, z_i)$.

Step 3. If $\xi(y) - \xi(z_i) \leqq -\varepsilon$, set $\varepsilon_i = \varepsilon$, $z_{i+1} = y$, $i = i + 1$ and go to Step 1; otherwise set $\varepsilon = \varepsilon/2$ and go to Step 2.

We remark that if we let $\bar{\varepsilon} = 0$, then Algorithm 14 is almost of the form of Algorithm 1. The following assumption on $B(\cdot, \cdot)$ ensures that cluster points of infinite sequences generated by Algorithm 14 are desirable.

HYPOTHESIS 11.

(i) $B(0, z) = A(z)$ for all z in T.

(ii) $B(\cdot, \cdot)$ is jointly upper semicontinuous on T; i.e., for any sequence $\{\varepsilon_i\}$ converging to ε^*, for any sequence $\{z_i\}$ converging to z^*, for any sequence $\{y_i\}$ converging to y^*, with y_i in $B(\varepsilon_i, z_i)$, y^* belongs to $\beta(\varepsilon^*, z^*)$.

THEOREM 1. *If the maps $A(\cdot)$ and $\xi(\cdot)$ satisfy parts* (i) *and* (ii) *of Hypothesis* 1, *and the map $B(\cdot, \cdot)$ satisfies Hypothesis* 11, *then every cluster point of an infinite sequence generated by Algorithm* 14 *is desirable.*

Proof. Let $\{z_i\}$ be an infinite sequence generated by Algorithm 14, and let z^* be a cluster point of this sequence. Thus, for some subset K_1 of the integers, the subsequence $\{z_i\}_{K_1}$ converges to z^*. Consider the sequence $\{z_{i+1}\}_{i+1 \in K_1}$ in T and the sequence $\{\varepsilon_i\}_{i \in K_1}$ in $[0, \bar{\varepsilon}]$. The boundedness of T and of the interval $[0, \bar{\varepsilon}]$ ensures that there exists K, an infinite subset of K_1, such that the subsequences $\{z_{i+1}\}_{i+1 \in K}$ and $\{\varepsilon_i\}_K$ converge to points z^{**} and ε^* in T and $[0, \bar{\varepsilon}]$, respectively. The properties of convergent sequences ensure that the subsequence $\{z_i\}_K$ also converges to z.

Now suppose that $\varepsilon^* > 0$; then the form of Algorithm 14 implies that there exists an integer k such that $\xi(z_{i+1}) - \xi(z_i) \leqq \varepsilon^*/2$ for all $i \geqq k$, i in K, and this contradicts part (ii) of Hypotheses 1. Consequently $\varepsilon^* = 0$.

The map $B(\cdot, \cdot)$ is jointly upper semicontinuous on T, and z_{i+1} belongs to $B(\varepsilon_i, z_i)$ for all i. It follows that z^{**} is in $B(\varepsilon^*, z^*)$, i.e., in $A(z^*)$. Consequently z^* is desirable.

HYPOTHESIS 12.

(i) $B(0, z) = A(z)$ for all z in T.

(ii) If $\xi(a) - \xi(z) < 0$ for all a in $A(z)$, then there exist $\varepsilon > 0$, $\delta > 0$ and $\gamma > 0$, possibly depending on z, such that

$$\xi(b') - \xi(z') \leqq -\gamma < 0$$

for all z' in T such that $\|z' - z\| \leqq \varepsilon$ and for all b' in $B(v, z')$, $0 \leqq v \leqq \delta$.

The proof of the following theorem can be carried out by using the same types of arguments as were used to prove Theorem 1, and it is therefore omitted.

THEOREM 2. *If the maps $A(\cdot)$ and $\xi(\cdot)$ satisfy parts* (i) *and* (ii) *of Hypothesis* 2, *and the map $B(\cdot, \cdot)$ satisfies Hypothesis* 12, *then every cluster point of an infinite sequence generated by Algorithm* 14 *is desirable.*

Remark. It can be shown that if maps $A(\cdot)$, $\xi(\cdot)$ and $B(\cdot, \cdot)$ satisfy Hypotheses 1 and 11, they satisfy Hypotheses 2 and 12 (see [1]).

For examples of how the ε-procedure is used in the synthesis of algorithms, see E. Polak [14].

5. Conclusion. To conclude, we would like to highlight the two most important aspects of the theory we have presented in this paper. The first is that by using models one can separate out the essential properties of an algorithm from the nonessential ones. Thus, for example, in a gradient method one need not specify in advance exactly which procedure one will use to search along the direction of steepest descent, one only has to specify that the search procedure will have certain properties. The second point that we wish to emphasize is that given a "conceptual" algorithm in which one has to perform in sequence several operations each of which requires an infinite number of iterations, one can obtain an "implementable" algorithm by "shuttling" between these infinite operations, combining them into a single infinite operation. Thus, our method of obtaining an "implementable" algorithm from a plurally infinitely iterative "conceptual" algorithm consists in "parallelizing" the infinite operations of the conceptual algorithm.

Obviously, this paper does not exhaust the study of models for computational methods or of the possibilities of constructing "implementable" algorithms from "conceptual" ones. We hope that this paper will lead to further work, in particular in the study of algorithms with infinite memory.

Appendix. We shall assume, throughout this Appendix, that the maps $f(\cdot), \alpha(\cdot), \gamma(\cdot), \xi(\cdot), m(\cdot,\cdot,\cdot)$ and $U(\cdot)$, and the set T, are as defined in § 3.1. The fact that the map $m(\cdot,\cdot,\cdot)$ satisfies part (i) of Hypothesis 7 and part (i) of Hypothesis 9 is an obvious consequence of its definition, and the fact that the map $m(\cdot,\cdot,\cdot)$ satisfies part (ii) of Hypothesis 7 and part (ii) of Hypothesis 9 is a direct consequence of the following theorem.

THEOREM A. *Given any z in T and $\delta > 0$, there exist $\varepsilon > 0$ and k in N, depending on z and δ, such that $f(m(z', y', j)) \leqq \alpha(z') + \delta$, for all z' in T, satisfying $\|z' - z\| \leqq \varepsilon$, for all y' in $U(z')$, for all $j \geqq k$.*

The proof of Theorem A consists of several steps which we shall state as lemmas. These lemmas being quasi-trivial, their proofs have been deleted.

LEMMA A.1. *Given any z in T, satisfying $\|\nabla f(z)\| > 0$, and $\delta > 0$, there exist $\varepsilon > 0$ and $\rho > 0$, depending on z and δ, such that $|\langle \nabla f(z'), \nabla f(y')\rangle| \geqq \rho > 0$, for all z' in T such that $\|z' - z\| \leqq \varepsilon$, for all y' in $U(z')$ such that $f(y') \geqq \alpha(z') + \delta$.*

LEMMA A.2. *Given any z in T, y in $U(z)$, j in N and $\rho > 0$ satisfying $|\nabla f(m(z, y, j)), \nabla f(z)| \geqq \rho > 0$, then there exists $v > 0$, depending on ρ only, such that $f(m(z, y, j + 1)) - f(m(z, y, j)) \leqq -v$.*

LEMMA A.3. *Given any z in T, satisfying $\|\nabla f(z)\| > 0$, and $\delta > 0$, there exist $\varepsilon > 0$, $\rho > 0$ and $v > 0$, depending on z and δ, such that $f(m(z', y', j + 1)) - f(m(z', y', j)) \leqq -v$, for all z' in T, such that $\|z' - z\| \leqq \varepsilon$, for all y' in $U(z')$ such that $f(m(z', y', j)) \geqq \alpha(z') + \delta$.*

LEMMA A.4. *Given any z in T, satisfying $\|\nabla f(z)\| > 0$, and $\delta > 0$, there exist $\varepsilon > 0$ and k in N, depending on z and δ, such that $f(m(z', y', j)) \leqq \alpha(z') + \delta$, for all z' in T, such that $\|z' - z\| \leqq \varepsilon$, for all y' in $U(z')$.*

LEMMA A.5. *Given any z in T, satisfying $\|\nabla f(z)\| = 0$, and $\delta > 0$, there exists $\varepsilon > 0$, depending on δ only, such that $f(m(z', y', j)) \leqq \alpha(z') + \delta$, for all z' in T satisfying $\|z' - z\| \leqq \varepsilon$, for all y' in $U(z')$, for all j in N.*

REFERENCES

[1] E. MICHAEL, *Topologies on spaces of subsets*, Trans. Amer. Math. Soc., 71 (1951), pp. 152–182.

[2] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, U.S. Nav. Res. Logist. Quart., 3 (1956), pp. 95–110.

[3] G. ZOUTENDIJK, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.

[4] B. T. POLYAK, *Gradient methods for the minimization of functionals*, U.S.S.R. Comput. Math. and Math. Phys., 3 (1963), no. 4, pp. 864–878 = Zh. Vychisl. Mat. i Mat. Fiz., 3 (1963), no. 4, pp. 643–653.

[5] E. S. LEVITIN AND B. T. POLYAK, *Constrained minimization methods*, U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), no. 5, pp. 1–50 = Zh. Vychisl. Mat. i Mat. Fiz., 6 (1966), no. 5, pp. 787–823.

[6] W. I. ZANGWILL, *Convergence conditions for nonlinear programming algorithms*, Working Paper 197, Center for Research in Management Science, Univ. of California, Berkeley, 1966.

[7] D. M. TOPKIS AND A. VEINOTT, *On the convergence of some feasible directions algorithms for nonlinear programming*, this Journal, 5 (1967), pp. 268–279.

[8] E. POLAK, *On primal and dual methods for solving discrete optimal control problems*, Proc. Second International Conference on Computing Methods in Optimization Problems (San Remo, Italy, 1968), Academic Press, New York, 1969.

[9] ———, *On the convergence of optimization algorithms*, Rev. Francaise Informatique et Recherche Operationelle, Ser. Rouge, 16 (1969), pp. 17–34.

[10] ———, *Computational methods in discrete optimal control and nonlinear programming: A unified approach*, Memo. ERL-M261, Electronics Research Laboratory, College of Engineering, Univ. of California, Berkeley, 1969.

[11] W. I. ZANGWILL, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1969.

[12] G. MEYER AND E. POLAK, *Abstract models for the synthesis of optimization algorithms*, Memo. ERL-268, Electronics Research Laboratory, College of Engineering, Univ. of California, Berkeley, 1969.

[13] R. MEYER, *The validity of a family of optimization methods*, this Journal, 8 (1970), pp. 41–54.

[14] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1970.

SIAM J. CONTROL
Vol. 9, No. 4, November 1971

A GENERALIZED LAGRANGE MULTIPLIER RULE FOR EQUALITY CONSTRAINTS IN NORMED LINEAR SPACES*

D. O. NORRIS†

Abstract. It is shown that a necessary and sufficient condition for a differentiable function to have a critical point on an affine constraint set is that a Lagrange multiplier rule hold. The rule applies to problems which are not covered by the standard multiplier theorem. The results are extended to the determination of necessary conditions for a convex function to have a minimum on a nonaffine constraint set.

1. Introduction. Let X and Y denote B-spaces and let $f: X \to R$ and $g: X \to Y$ denote Frechét differentiable functions. Consider the problem of minimizing f on the set of points for which $g(x) = 0$. The problem was apparently first treated in this form by H. H. Goldstine [5] in 1940. He obtained the classical multiplier rule under the hypothesis that the range of $Dg(x_0)$, the derivative of g at x_0, be closed, f and g have continuous derivatives, and any one of four other conditions be satisfied. A proof of the Lagrange multiplier rule which requires that the range of $Dg(x_0)$ be closed and that f and g have continuous derivatives may be found in [9, p. 380] or [8, pp. 243–244]. Specifically, we have the following theorem.

THEOREM 1.1. *Let X and Y denote B-spaces. Let $f: X \to R$ and $g: X \to Y$ each have a continuous Frechét derivative in a neighborhood of x_0. Let f have a local minimum on the set $N(g) = \{x: g(x) = 0\}$ at x_0 and suppose that the range of $Dg(x_0)$ is closed. Then there is a nonzero $(\eta^*, y^*) \in R \times Y^*$ such that*

$$\eta^* Df(x_0)h + y^* Dg(x_0)h = 0$$

for all $h \in X$.

If $Dg(x_0)$ is surjective, then both η^* and y^* may be chosen to be nonzero, whereas if $Dg(x_0)$ is not surjective, then $\eta^* = 0^*$ and y^* nonzero (for some y^*) is always a possible solution. The latter, however, is often unsatisfactory since a solution for which both η^* and y^* are nonzero may exist. In fact, such will be the case if $Df(x_0) \in R(Dg(x_0)^*)$, the range of the adjoint of $Dg(x_0)$.

In this paper the question of relaxing the requirement that the range of $Dg(x_0)$ be closed will be considered. It will be shown that for certain classes of problems, necessary and sufficient conditions for the existence of a Lagrange multiplier can be obtained without assuming that the range of $Dg(x_0)$ is closed. Before proceeding with the development of the theory we present an example for which Theorem 1.1 does not apply, yet a nonzero Lagrange multiplier exists.

The problem of minimizing a function subject to constraints (set constraints, inequality constraints, and equality constraints) has been treated extensively and the question arises as to how some of these other treatments relate to the work considered here. Varaiya [13] and Guignard [6] treat the problem: minimize $f: X \to R$ on $A \cap g^{-1}(B)$, where $g: X \to Y$, $A \subset X$ and $B \subset Y$. If $B = \{0\}$ and $A = X$, then their problem reduces to the one considered here. If Varaiya's weak constraint qualification is substituted for the requirement that f and g have

* Received by the editors September 15, 1970, and in final revised form March 25, 1971.

† Department of Mathematics, Ohio University, Athens, Ohio 45701.

continuous derivatives, then his major result reduces to Theorem 1.1 for the problem under discussion. However, the weak constraint qualification may be difficult to verify, and in order to get the multiplier rule in terms of $Df(x_0)$ and $Dg(x_0)$ he must assume $Dg(x_0)$ has a closed range. Similarly, Guignard's major result for the problem treated here requires that $Dg(x_0)$ have a closed range. Russell [12], Pshenichniy [11], Dubovitskii and Milyutin [1], and Neustadt [10] have treated problems where the inequality constraints and equality constraints are more explicitly apparent. For example, Russell treats the problem of minimizing a function f subject to an arbitrary number of inequalities and equality constraints given by real-valued functions. His definition of "linearly approximable" is a regularity condition (see Definition 4.3 of this paper) and his multiplier rule is given in set theoretic terms rather than the differential form given in Theorem 1.1. For the problem treated in this paper, Russell's necessary condition for a minimum is that $Df(x_0)$ be a member of the weak* closure of the range of $Dg(x_0)^*$. Thus, unless $Df(x_0)$ is actually a member of the range of $Dg(x_0)^*$, Russell's result does not apply to the question of solving the equation $Df(x_0) + Dg(x_0)^*y^* = 0^*$, and therefore characterization of x_0 may be difficult. Neustadt has developed a very general theory of extremals which accounts for equality constraints, inequality constraints, and set constraints. For the problem discussed here (i.e., equality constraints only, together with differentiability), it is usually possible to recast Neustadt's canonical optimization problem into a problem in which the function g defining the equality constraints has its range in a finite-dimensional space. In this case, Theorem 1.1 always holds since the range of $Dg(x_0)$ is closed, being a finite-dimensional subspace. The work of Pshenichniy and Dubovitskii and Milyutin bear a good deal of similarity. Dubovitskii and Milyutin base their treatment of optimization problems on a result they refer to as Euler's equation (Theorem 2.1). For the problem treated in this paper, in which there are no inequality constraints, their result may be stated as follows:

Let $\Omega_0 \subset X$ denote an open convex cone with vertex at zero and let $\Omega \subset X$ denote a closed convex cone with vertex at zero. $\Omega \cap \Omega_0 = \varnothing$ if and only if there are $x_0^* \in \Omega_0^*$ (the dual cone of Ω) and $x^* \in \Omega^*$ such that $x_0^* + x^* = 0$.

For a differentiable $f: X \to R$, $\Omega_0 = \{h: Df(x_0)h < 0\}$, where x_0 is the desired minimum. If $g: X \to Y$ is affine or satisfies a regularity condition, $\Omega = \{h: Dg(x_0)h = 0\}$. For this situation Euler's equation may be expressed as $-Df(x_0) + x^* = 0$ for some $x^* \in \Omega^*$. For the problems treated by Dubovitskii and Milyutin it is usually possible to show that the range of $Dg(x_0)^*$ is closed in which case x^* can be chosen in the range of $Dg(x_0)^*$; i.e., Theorem 1.1 holds.

Thus, for the papers discussed above and except for [13] and [6], the multiplier rules are not given in the form of Theorem 1.1 and the question of characterizing the range of $Dg(x_0)$ does not arise until applications are made.

2. Example. Represent an element $x \in l_p$ as $x = (\xi_1, \xi_2, \cdots)$. Define $f: l_2 \to R$ by $f(x) = \|x\|^2$ and let $g: l_2 \to l_1$ such that $g(x) = (2^{-1}\xi_2 - 2^{-2}, 3^{-1}\xi_3 - 3^{-2}, \cdots)$. Clearly, f and g have continuous derivatives. Denote the range and null set of g by $R(g)$ and $N(g)$, respectively. (Henceforth, this notation will be used for all functions.) We assert the following facts:

(i) $x \in N(g)$ if and only if $x = (\xi_1, \frac{1}{2}, \frac{1}{3}, \cdots)$, where ξ_1 is arbitrary.

(ii) $Df(x) = 2x$.

(iii) $Dg(x)$ may be represented by

$$Dg(x) = \begin{bmatrix} 0 & \frac{1}{2} & & & 0 \\ & 0 & \frac{1}{3} & & \\ & & 0 & \frac{1}{4} & \\ & & & \ddots & \\ 0 & & & & \end{bmatrix}.$$

(iv) $R(Dg(x_0)) \neq l_1$ since, for example, there is no $h \in l_2$ such that $Dg(x_0)h = (2^{-3/2}, 3^{-3/2}, \cdots) \in l_1$.

(v) $R(Dg(x_0))$ is not closed and $\overline{R(Dg(x_0))} = l_1$. For, given $(\eta_1, \eta_2, \cdots) \in l_1$ define $\xi_{n+1} = (n+1)\eta_n$, $h_n = (0, \xi_2, \cdots, \xi_{n-1}, 0, \cdots)$. Thus, $\lim_{n\to\infty} Dg(x_0)h_n = (\eta_1, \eta_2, \cdots)$.

(vi) All the hypotheses of Theorem 1.1 are satisfied except that $R(Dg(x_0))$ is not closed. Hence, the theorem does not apply. However, it is easy to check that with $\eta^* = 1$ and y^* represented by $(-2, -2, \cdots) \in l_\infty$, (η^*, y^*) is the desired nonzero Lagrange multiplier. It is worth noting that had the range of g been chosen as l_2, then since $(-2, -2, \cdots) \notin l_2$ the existence of a continuous linear functional as a multiplier could not have been asserted. However, the multiplier exists as a linear functional (not continuous).

3. Problems with affine constraints. We now propose to present results which make it possible to assert the existence of Lagrange multipliers in cases illustrated by the example in § 2. That is, the problem of minimizing a differentiable function subject to an affine constraint will be treated.

First, however, a characterization of the range of the adjoint of a linear operator will be given. This result is needed in the sequel and is less restrictive than the well-known result that if $R(T)$ is closed, where T is a linear operator, then $R(T^*) = \{x^* : N(T) \subset N(x^*)\}$ (e.g., see [2, p. 487]).

THEOREM 3.1. *Let X and Y denote normed linear spaces and let $T : X \to Y$ denote a continuous linear function. For each nonzero $b^* \in X^*$, where X^* is the topological dual of X, the equation $T^*y^* = b^*$ has a solution in Y^* if and only if there is an $h_0 \in X$ such that*

(i) $h_0 \notin N(b^*)$,

(ii) $Th_0 \notin \overline{T(N(b^*))}$.

Proof. To show the necessity, assume that $T^*y^* = b^*$ has a solution $y_0^* \in Y^*$. $b^* \neq 0^*$, so $N(b^*)$ has codimension equal to one and consequently there is an $h_0 \notin N(b^*)$ such that $X = \text{span}\{h_0, N(b^*)\}$. Now suppose for every $h_0 \notin N(b^*)$ it happens that $Th_0 \in \overline{T(N(b^*))}$. Since $X = \text{span}\{h_0, N(b^*)\}$, every $h \in X$ may be written as $h = \alpha h_0 + n$ for some real α and some $n \in N(b^*)$. By hypothesis, $b^*h - y_0^*Th = 0$ for all $h \in X$, so $b^*\alpha h_0 - y_0^*T(\alpha h_0 + n) = 0$ for all real α and all $n \in N(b^*)$. For $\alpha = 1$, it follows that $b^*h_0 - y_0^*Th_0 = y_0^*Tn$ for all $n \in N(b^*)$. Thus, $y_0^*Tn = 0$ for all $n \in N(b^*)$. Now, note that if $Th_0 \in \overline{T(N(b^*))}$, then there is a sequence $\{T(x_n)\} \subset T(N(b^*))$ such that $Tx_n \to Th_0$. y_0^* is continuous so $y_0^*Th_0 = 0$. Therefore, $b^*h_0 = 0$ contrary to the assumption that $h_0 \notin N(b^*)$.

Conversely, if (i) and (ii) hold, then $Th_0 \neq 0$, and by the Hahn–Banach theorem there is a nonzero $y_0^* \in Y^*$ such that $y_0^*y = 0$ for all $y \in \overline{T(N(b^*))}$ and $y_0^*Th_0 = b^*h_0$. Then $b^*h = b^*(\alpha h_0 + n) = \alpha y_0^*Th_0 = y_0^*(T(\alpha h_0 + n)) = y_0^*Th$ for all $h \in X$, and thus $T^*y_0^* = b^*$.

COROLLARY 3.1.1. *Let X and Y denote linear spaces and let $T: X \to Y$ denote a linear function. For each nonzero $b' \in X'$, where X' is the algebraic dual of X, the equation $T'y' = b'$ has a solution in Y' if and only if there is an $h_0 \in X$ such that*

(i) $h_0 \notin N(b')$,

(ii) $Th_0 \notin T(N(b'))$.

Proof. The proof is almost the same as that given in Theorem 3.1.

Theorem 3.1 and its corollary will now be connected with the existence of a critical point of f on an affine subspace. It is a well-known result that if f has an extremum at x_0 on an affine subspace A, then $Df(x_0)h = 0$ for all $h \in A - x_0$, i.e., f has a critical point on A at x_0. To be precise we have the following definition.

DEFINITION 3.2. Let X denote a normed linear space and suppose that A is an affine subspace in X (i.e., the translate of a subspace). Let $f: X \to R$ denote a differentiable function. f has a *critical point* at $x_0 \in A$ if and only if $Df(x_0)h = 0$ for all $h \in A - x_0$.

THEOREM 3.3. *Let $g: X \to Y$ denote a continuous affine transformation, $g(x) = T(x - x_1)$, where T is linear and continuous. Let f have a nonzero derivative at x_0.*

(a) *If there is an $h_0 \in X$ such that*

(i) $h_0 \notin N(Df(x_0))$ *and*

(ii) $Th_0 \notin \overline{T(N(Df(x_0)))}$,

then f has a critical point on $N(g)$ at x_0.

(b) *Conversely, if f has a critical point on $N(g)$ at x_0, then there is an $h_0 \in X$ such that*

(iii) $h_0 \notin N(Df(x_0))$ *and*

(iv) $Th_0 \notin T(N(Df(x_0)))$.

Proof. Part (a). If x_0 is not a critical point, then there is an $h_1 \in N(T)$ such that $Df(x_0)h_1 \neq 0$. Then $Df(x_0)h_1 + y^*Th_1 \neq 0$ for all $y^* \in Y^*$. In view of conditions (i) and (ii), this contradicts Theorem 3.1.

Part (b). If for every $h_0 \notin N(Df(x_0))$ it happens that $Th_0 \in T(N(Df(x_0)))$, then $T(X) \subset T(N(Df(x_0)))$ and we have $T(X) = T(N(Df(x_0)))$. Since $Df(x_0) \neq 0^*$, there is an $h_2 \notin N(Df(x_0))$ such that $X = \text{span}\{h_2, N(Df(x_0))\}$. But, $T(X) = T(N(Df(x_0)))$ so that there is an $h_3 \in N(Df(x_0))$ such that $Th_3 = Th_2$. Then, $T(h_3 - h_2) = 0$ and since x_0 is a critical point it follows that $Df(x_0)(h_3 - h_2) = 0$. But, $Df(x_0)h_3 = 0$, so $Df(x_0)h_2 = 0$ contrary to the assumption.

The connection with Lagrange multipliers is now easily established.

THEOREM 3.4. *Let $g: X \to Y$ denote a continuous affine transformation, $g(x) = T(x - x_1)$, where T is linear and continuous. Let f have a nonzero derivative at x_0.*

(a) *If f has a critical point on $N(g)$ at x_0, then there is a nonzero $y' \in Y'$ such that*

$$Df(x_0)h + y'Dg(x_0)h = 0$$

for all $h \in X$.

(b) *If there is a nonzero $y^* \in Y^*$ such that*

$$Df(x_0)h + y^*Dg(x_0)h = 0$$

for all $h \in X$, then f has a critical point on $N(g)$ at x_0.

Proof. Part (a). If f has a critical point on $N(g)$ at x_0, then the conclusion of Theorem 3.3(b) holds. But $Dg(x_0) = T$ so from Corollary 3.1.1 it can be concluded that there is a nonzero $y' \in Y'$ such that $Dg(x_0)'y' = -Df(x_0)$ which is equivalent to the desired result.

Part (b) follows by a similar argument with the aid of Theorem 3.3(a) and Theorem 3.1.

It can readily be seen that the example of § 2 is covered by Theorem 3.4. Furthermore, the conclusion of part (a) cannot be strengthened to make y' continuous without the addition of more hypotheses. This is illustrated by the example of § 2 when the range of g is taken to be l_2 instead of l_1 (see (vi) in § 2).

Theorem 3.4 may be applied to linear distributed parameter control problems. For example, suppose a system is described by the differential equation

$$x'(t) = Ax(t) + u(t), \qquad x(0) = x_0,$$

where A is the infinitesimal generator of a strongly continuous semigroup of operators, $T(t)$. Suppose it is desired to find u which transfers x_0 to 0 at time t_1 such that $f: U \to R$ is minimized, where f is a differentiable function. If $u(t) \in X$, X a B-space, and if $I = [0, t_1]$, then it may be desirable to choose U as $L_2(I, X)$, where

$$\|u\|_U = \left[\int_0^{t_1} \|u(t)\|_X^2 \, dt\right]^{1/2}.$$

Define $g: U \to X$ such that

$$g(u) = T(t_1)x_0 + \int_0^{t_1} T(t_1 - s)u(s)\, ds.$$

The problem may now be reformulated as follows:

$$\text{minimize } f$$

subject to the affine equality constraint

$$g(u) = 0.$$

Theorem 3.4 clearly applies. Furthermore, in distributed parameter control problems, X is not finite-dimensional, and consequently g may not have a closed range. In such cases, the classical Lagrange multiplier theorem (Theorem 1.1) will not apply.

4. Convex functions and nonaffine constraints. We now propose to extend some of the results of the last section to cover the problem of minimizing a convex, differentiable function subject to a nonaffine equality constraint.

DEFINITION 4.1. Let X denote a normed linear space and suppose $A \subset X$. Let $x_0 \in A$. The *tangent cone*, $\mathrm{TC}(A, x_0)$, for A at x_0 is the set of vectors $x \in X$ such that:

(i) there exists a sequence $\{y_n\} \subset A$ for which $y_n \to x_0$, and

(ii) there exists a sequence $\{\lambda_n : \lambda_n \geqq 0\}$ for which $\lambda_n(y_n - x_0) \to x$.

Varaiya [13] gives an equivalent definition and calls $\mathrm{TC}(A, x_0)$ the local closed cone of A at x_0. Other authors (e.g., see [3], [7]) define the unit vectors of

the tangent cone and require, accordingly, that $\lambda_n = \|y_n - x_0\|^{-1}$. The tangent cone is the cone generated by these unit vectors. It is easy to see that these definitions are equivalent.

In the next theorem it is established that when f is a continuous, convex function and has a minimum on $A \subset X$ at x_0, then f has a minimum on $x_0 + \mathrm{TC}(A, x_0)$ at x_0. A similar result where $A = N(g)$ has been proved by Flett [4, Theorem 1]. If, in addition, x_0 satisfies a regularity condition such as $N(Dg(x_0)) = \mathrm{TC}(N(g), x_0)$, then the problem of minimizing the convex function f subject to a differentiable equality constraint can be reduced to the problem of minimizing f subject to an affine constraint. When f is differentiable the results of the last section apply.

THEOREM 4.2. *Let X denote a normed linear space. Let $A \subset X$ and suppose that $f: X \to R$ is a convex, continuous function. If f has a minimum on A at x_0, then f has a minimum on $x_0 + \mathrm{TC}(A, x_0)$ at x_0.*

Proof. If $\mathrm{TC}(A, x_0) = \{0\}$, the result is trivially true. Thus, assume that $\mathrm{TC}(A, x_0) \neq \{0\}$. Suppose there is an element $x_1 \in x_0 + \mathrm{TC}(A, x_0)$ such that $f(x_1) < f(x_0)$. Then there is a sequence $\{y_n\} \subset A$ and a sequence $\{\lambda_n : \lambda_n \geqq 0\}$ such that $y_n \to x_0$ and $\lambda_n(y_n - x_0) \to x_1 - x_0$. f is continuous, so for n sufficiently large

$$f(\lambda_n(y_n - x_0) + x_0) < f(x_0). \tag{1}$$

Now $\lambda_n \to \infty$ as $n \to \infty$ so for $\lambda_n > 1$ note that

$$y_n = (1 - \lambda_n^{-1})x_0 + \lambda_n^{-1}(x_0 + \lambda_n(y_n - x_0)).$$

Use (1) and the convexity of f to conclude that

$$\begin{aligned} f(y_n) &\leqq (1 - \lambda_n^{-1})f(x_0) + \lambda_n^{-1}f(x_0 + \lambda_n(y_n - x_0)) \\ &< (1 - \lambda_n^{-1})f(x_0) + \lambda_n^{-1}f(x_0) = f(x_0). \end{aligned}$$

This contradicts the assumption that f achieves its minimum on A to x_0.

COROLLARY 4.2.1. *Under the hypothesis of Theorem 4.2, let Y denote a vector space, $g: X \to Y$, and $A = N(g)$. Then f has a minimum on $x_0 + \mathrm{TC}(N(g), x_0)$.*

DEFINITION 4.3. Let X and Y denote normed linear spaces and let $g: X \to Y$ denote a differentiable function. $x_0 \in N(g)$ is a *regular point* if and only if $\mathrm{TC}(N(g), x_0) = N(Dg(x_0))$.

Some authors (e.g., see [5], [8], [9]) define regularity by the requirement that $Dg(x_0)$ be surjective. However, Flett [3, Theorem 3] has shown that if g has a continuous derivative and $Dg(x_0)$ is surjective, then $\mathrm{TC}(N(g), x_0) = N(Dg(x_0))$. Definition 4.3 has been given by Hestenes [7, p. 29] in the case that g has a finite-dimensional range.

THEOREM 4.4. *Let X and Y denote normed linear spaces. Suppose $f: X \to R$ is a convex differentiable function and $g: X \to Y$ is differentiable. If f has a minimum on $N(g)$ at x_0 and x_0 is a regular point, then there is a nonzero $y' \in Y'$ such that*

$$Df(x_0)h + y'Dg(x_0)h = 0$$

for all $h \in X$.

Proof. By Corollary 4.2.1 it follows that f has a minimum on $x_0 + \mathrm{TC}(N(g), x_0)$ at x_0, and from regularity it follows that f has a minimum on $x_0 + N(Dg(x_0))$ at

x_0, i.e., an affine subspace. But, then f has a critical point on $x_0 + N(Dg(x_0))$ at x_0, and by Theorem 3.4(a) the desired conclusion follows.

REFERENCES

[1] A. Ya. Dubovitskii and A. A. Milyutin, *Extremum problems in the presence of restrictions*, U.S.S.R. Comput. Math. and Math. Phys., 5 (1965), no. 3, pp. 1–80.

[2] N. Dunford and J. T. Schwartz, *Linear Operators, Part I*, Interscience, New York, 1964.

[3] T. M. Flett, *On differentiation in normed vector spaces*, J. London Math. Soc., 42 (1967), pp. 523–533.

[4] ———, *Points of minimum or maximum norm on smooth surfaces in Banach spaces*, Ibid., 44 (1969), pp. 583–586.

[5] H. H. Goldstine, *Minimum problems in the functional calculus*, Bull. Amer. Math. Soc., 46 (1940), pp. 142–149.

[6] M. Guignard, *Generalized Kuhn–Tucker conditions for mathematical programming problems in a Banach space*, this Journal, 7 (1969), pp. 232–241.

[7] M. R. Hestenes, *Calculus of Variations and Optimal Control*, John Wiley, New York, 1966.

[8] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[9] L. A. Lusternik and V. J. Sobolev, *Elements of Functional Analysis*, Frederick Ungar, New York, 1961.

[10] L. W. Neustadt, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.

[11] B. N. Pshenichniy, *Convex programming in a normed space*, Kibernetika, 5 (1965), pp. 46–54.

[12] D. L. Russell, *The Kuhn–Tucker conditions in Banach space with an application to control theory*, J. Math. Anal. Appl., 15 (1966), pp. 200–212.

[13] P. P. Varaiya, *Nonlinear programming in Banach space*, SIAM J. Appl. Math., 15 (1967), pp. 284–293.

SIAM J. CONTROL
Vol. 9, No. 4, November 1971

JUSTIFICATION OF THE DESCRIBING FUNCTION METHOD*

A. R. BERGEN† AND R. L. FRANKS‡

Abstract. Explicit conditions are given under which the use of the describing function method to investigate the nature of oscillations in autonomous nonlinear feedback systems is justified. When these conditions are satisfied, bounds are given for the frequency, fundamental magnitude, and higher harmonics of the oscillation based on the describing function approximation.

The feedback systems considered are those which can be decomposed into a linear time-invariant subsystem, not necessarily causal, stable, or finite-dimensional, and a nonlinear frequency independent subsystem, possibly containing hysteresis.

The approach taken is to consider the describing function equation as an approximation to a determining equation for periodic solutions of the autonomous system's operator equation, and to use local degree theory to guarantee the existence of a solution to the determining equation.

1. Introduction. The describing function method was introduced in this country by R. J. Kochenburger [6] in 1950. It is used by engineers to investigate the nature of oscillations in nonlinear feedback systems. Frequently it is used as an aid in design to avoid or attenuate such periodic responses. This paper is concerned with justifying the use of the describing function method to investigate periodic responses in autonomous nonlinear feedback systems, and gives error bounds for those responses.

Many authors have discussed the use of the describing function. A good discussion is given in Gelb and Vander Velde [4]. The general approach is as follows:

Consider the autonomous feedback system in Fig. 1. $\mathcal{N}$ is usually a frequency independent nonlinear subsystem whose characteristic has odd symmetry and $\mathcal{G}$ is a time-invariant linear subsystem. The input to $\mathcal{N}$ is taken to be $x = \mathrm{Re}\,(re^{j\omega t})$. The steady state output of $\mathcal{N}$ and the input to $\mathcal{G}$ is then

$$y = \mathrm{Re}\,\Big(\sum_{n=1,\,\mathrm{odd}}^{\infty} A_n e^{jn\omega t}\Big).$$

The system connection requires $x = -z$; i.e., $r = -G(j\omega)A_1$ and $G(jn\omega)A_n = 0$ for n greater than or equal to 3. The coefficient A_1 ordinarily depends only on r so it is usually written as $A_1 = rN(r)$. $N(r)$ is called the describing function for $\mathcal{N}$ and is its steady state first harmonic gain. The first harmonic equation can then be written as

$$1 + G(j\omega)N(r) = 0 \tag{1}$$

and is called the describing function equation. This equation is usually solved graphically by plotting the Nyquist locus $G(j\omega):\omega \in \mathbb{R}^+$ and the critical locus,

* Received by the editors June 11, 1970, and in revised form January 4, 1971.

† Department of Electrical Engineering and Computer Sciences, Electronic Research Laboratory, University of California, Berkeley, California 94720. This author's work was supported by the National Science Foundation under Grant GK-10656X.

‡ Department of Electrical Engineering and Computer Sciences, Electronics Research Laboratory, University of California, Berkeley, California. Now at Bell Telephone Laboratories, Whippany, New Jersey 07981. This author's work was supported by a NASA Traineeship.

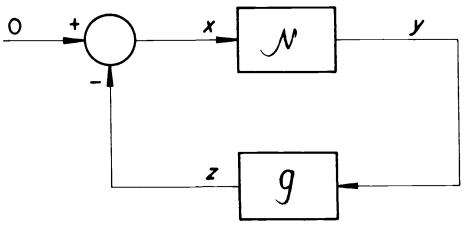


FIG. 1. *System* (S)

$-1/N(r): r \in \mathbb{R}^+$. An intersection of these loci gives values of ω and r which satisfy the describing function equation. The basic appeal of the method is the simplicity of this graphical solution.

The condition that $G(jn\omega)A_n = 0$ for n greater than or equal to 3 can never be satisfied, but it is ignored. If there is a pair (ω, r) which satisfies the describing function equation, then it is felt that the system has a periodic solution near $x(t) = r \sin \omega t$; the heuristic justification is that if $\mathcal{N}$ is not too nonlinear and $\mathcal{G}$ is a good low-pass filter, then the higher harmonics $G(jn\omega)A_n$ are approximately zero and can be ignored. Notice the importance of the assumption of an odd nonlinearity in this connection.

Clearly this approach requires mathematical justification, and error bounds for the describing function approximation would be desirable.[1] Bass [1] considered the problem of justifying the describing function method and gave conditions under which the method was justified. Unfortunately, as he points out, one cannot tell whether or not the conditions are satisfied except in very simple cases.

2. Outline of method. The system of interest is shown in Fig. 1, where $\mathcal{N}$ is a frequency-independent nonlinear subsystem and $\mathcal{G}$ is a linear time-invariant subsystem. The problem is to show that under certain conditions related to the describing function method, there is a periodic function $\hat{x}: \mathbb{R} \to \mathbb{R}$ such that

$$\hat{x} = -\mathcal{G}\mathcal{N}\hat{x}. \tag{2}$$

The approach, motivated by Cesari [3], is first to time scale the problem and then to separate it into two, more tractable problems.

Using time scaling, it is sufficient to consider functions of period 2π. If $\hat{x}$ is a periodic solution of period T, define the time-scaled function x by $x(\omega t) = \hat{x}(t)$, where $\omega T = 2\pi$. Then x has period 2π. The operator $\mathcal{G}$ is not invariant under time scaling, that is, $[\mathcal{G}\hat{x}](t) \neq [\mathcal{G}x](\omega t)$. This means that time scaling requires the introduction of new operators, $\mathcal{G}_\omega: \omega \in [\omega_*, \omega^*]$, such that $[\mathcal{G}\hat{x}](t) = [\mathcal{G}_\omega x](\omega t)$.

In terms of time-scaled functions and operators, the problem is to find a function x of period 2π and an $\omega > 0$ such that

$$x = -\mathcal{G}_\omega \mathcal{N} x. \tag{3}$$

Note that if $\mathcal{N}$ were not frequency independent, an appropriate operator $\mathcal{N}_\omega$ would be required.

[1] With the paper virtually complete the authors learned of a similar effort by Kudrewicz [7] reported in the 1969 IFAC Conference in Warsaw. The system considered by us includes hysteresis nonlinearities, and the criterion for justification is easier to apply and somewhat less restrictive.

To separate the problem into two, more tractable problems, let $\mathscr{P}x$ be the projection of a 2π periodic function x onto its first M harmonics and $\mathscr{P}^*x$ be the projection of x onto its remaining harmonics, so that $x = \mathscr{P}x + \mathscr{P}^*x$. Then $x = -\mathscr{G}_\omega \mathscr{N} x$ if and only if both

$$\mathscr{P}^*x = -\mathscr{P}^*\mathscr{G}_\omega \mathscr{N} x, \tag{4}$$

$$\mathscr{P}x = -\mathscr{P}\mathscr{G}_\omega \mathscr{N} x. \tag{5}$$

Under certain conditions, for any $\omega \in [\omega_*, \omega^*]$ and any set $\mathbf{r}$ of Fourier coefficients for the first M harmonics of x, there is a function $\tilde{x}(\mathbf{r})$ with those first M harmonics which satisfies (4). Substituting $\tilde{x}(\mathbf{r})$ into (5) gives

$$\mathscr{P}\tilde{x}(\mathbf{r}) = -\mathscr{P}\mathscr{G}_\omega \mathscr{N} \tilde{x}(\mathbf{r}). \tag{6}$$

Equation (6) is a determining equation for (3) in the sense that if there are an ω and a set $\mathbf{r}$ of Fourier coefficients such that (6) is satisfied, then $x = \tilde{x}(\mathbf{r})$ is a solution to (3) for that ω. Defining $\hat{x}(t) = x(\omega t)$ implies $\hat{x}$ is the required periodic response of the system.

If $M = 1$, equation (6) corresponds to an operator $V: \mathbb{R}^2 \to \mathbb{R}^2$. The describing function equation corresponds to an operator $U: \mathbb{R}^2 \to \mathbb{R}^2$. U is an approximation of V and has local degree one on a particular set. Under certain conditions V is homotopically equivalent to U on that set, so V has degree one on that set. Therefore (6) has a solution for some (ω, r) in that set. This means that (3) has a periodic solution.

This approach is carried out in the next three sections. It is recommended that the reader who is unfamiliar with local degree refer to Appendix II at this time. Theorems II.2 and II.3 in that appendix are fundamental to the motivation of this paper.

3. Basic result. We are interested in the solution of (2) in a space of periodic functions (more precisely periodic functions with only odd harmonics). This will be done by solving (3) in the space

$$\mathscr{L}_\pi^2 = \{x: \mathbb{R} \to \mathbb{R} | x \text{ has period } 2\pi \text{ and only odd harmonics}\},$$

$\mathscr{L}_\pi^2$ is a Hilbert space with the inner product

$$\langle x, y \rangle = \frac{1}{\pi} \int_0^{2\pi} x(\theta) y(\theta)\, d\theta.$$

The corresponding norm will be referred to as $\|x\|$. The elements of $\mathscr{L}_\pi^2$ are half-wave symmetric periodic functions. We shall refer to them as π-symmetric functions.

In order to avoid proving some later theorems twice we shall introduce new operators $\mathscr{A}_\omega$ and $\mathscr{F}$, and new equations

$$x = -\mathscr{A}_\omega \mathscr{F} x, \tag{7}$$

$$\mathscr{P}^*x = -\mathscr{P}^*\mathscr{A}_\omega \mathscr{F} x, \tag{8}$$

$$\mathscr{P}x = -\mathscr{P}\mathscr{A}_\omega \mathscr{F} x. \tag{9}$$

In § 5, $\mathscr{A}_\omega$ will be related to $\mathscr{G}_\omega$, and $\mathscr{F}$ will be related to $\mathscr{N}$.

The basic result of this section is Theorem 1 which gives sufficient conditions for the existence of a solution to (8) and (9) in the space $\mathscr{L}_\pi^2$.

Let

$$\xi_k(\theta) = \begin{cases} \sin k\theta, & k \text{ odd}, \\ \cos(k-1)\theta, & k \text{ even}; \end{cases}$$

then $\{\xi_k\}$ is a complete orthonormal basis for $\mathscr{L}_\pi^2$. For a fixed positive integer κ, define:

1. $\mathscr{P}: \mathscr{L}_\pi^2 \to \mathscr{L}_\pi^2$ such that $\mathscr{P}x \triangleq \sum_{k=1}^{\kappa} \langle x, \xi_k \rangle \xi_k$,

2. $\mathscr{P}^*: \mathscr{L}_\pi^2 \to \mathscr{L}_\pi^2$ such that $\mathscr{P}^* x = x - \mathscr{P}x$,

3. $y: \mathbb{R}^{\kappa-1} \to \mathscr{P}\mathscr{L}_\pi^2$ such that $y(\mathbf{r}) \triangleq \sum_{k=1}^{\kappa-1} r_k \xi_k$, where

   $\mathbf{r} = (r_1, r_2, \cdots, r_{\kappa-1}) \in \mathbb{R}^{\kappa-1}$,

4. the usual Lipschitz norm

$$\|\mathscr{F}\| \triangleq \sup \left\{ \frac{\|\mathscr{F}x - \mathscr{F}y\|}{\|x - y\|} : x, y \in \mathscr{L}_\pi^2, x \neq y \right\}.$$

For the remainder of this section the following hypothesis is required.

HYPOTHESIS A.

1. There is an ω_* and ω^* such that for each $\omega \in [\omega_*, \omega^*]$, $\mathscr{A}_\omega: \mathscr{L}_\pi^2 \to \mathscr{L}_\pi^2$ is a bounded linear operator such that $\|\mathscr{A}_\omega - \mathscr{A}_\nu\| \to 0$ as $\omega \to \nu$ for all $\nu \in [\omega_*, \omega^*]$.

2. $\mathscr{F}: \mathscr{L}_\pi^2 \to \mathscr{L}_\pi^2$ is a bounded continuous operator with $\|\mathscr{F}\| < \infty$.

3. There exists a function $\rho: [\omega_*, \omega^*] \to \mathbb{R}$ such that $\|\mathscr{F}\| \|\mathscr{P}^* \mathscr{A}_\omega\| \leqq \rho_\omega$ for all $\omega \in [\omega_*, \omega^*]$ and $\sup \{\rho_\omega : \omega \in [\omega_*, \omega^*]\} < 1$.

Theorem 1 can most conveniently be stated and proved by first proving two lemmas. Lemma 1 gives conditions under which an equation similar to (6) has a solution, and gives important properties of that solution.

LEMMA 1. *Given Hypothesis* A, *for each* $\omega \in [\omega_*, \omega^*]$ *and each* $\mathbf{r} \in \mathbb{R}^{\kappa-1}$ *the equation* $x = -\mathscr{P}^* \mathscr{A}_\omega \mathscr{F}[y(\mathbf{r}) + x]$ *has a unique solution* $x^*(\omega, \mathbf{r})$. *Define* $x(\omega, \mathbf{r}) = y(\mathbf{r}) + x^*(\omega, \mathbf{r})$. *Then*:

a. $x: [\omega_*, \omega^*] \times \mathbb{R}^{\kappa-1} \to \mathscr{L}_\pi^2$ *is continuous*,

b. $\mathscr{P}x(\omega, \mathbf{r}) = y(\mathbf{r})$,

c. $\mathscr{P}^* x(\omega, \mathbf{r}) = -\mathscr{P}^* \mathscr{A}_\omega \mathscr{F} x(\omega, \mathbf{r})$,

d. $\|\mathscr{P}^* x(\omega, \mathbf{r})\| \leqq \dfrac{\|\mathscr{P}^* \mathscr{A}_\omega \mathscr{F} y(\mathbf{r})\|}{1 - \rho_\omega}$.

The proofs of all lemmas and theorems are given in Appendix I.

Before giving conditions under which (6) has a solution it is necessary to define functions $U: [\omega_*, \omega^*] \times \mathbb{R}^{\kappa-1} \to \mathbb{R}^\kappa$ and $V: [\omega_*, \omega^*] \times \mathbb{R}^{\kappa-1} \to \mathbb{R}^\kappa$ whose

components $k = 1, 2, \cdots, \kappa$ satisfy

$$U_k(\omega, \mathbf{r}) = r_k + \langle \mathscr{A}_\omega \mathscr{F} y(\mathbf{r}), \xi_k \rangle, \tag{10}$$

$$V_k(\omega, \mathbf{r}) = r_k + \langle \mathscr{A}_\omega, \mathscr{F} x(\omega, \mathbf{r}), \xi_k \rangle, \tag{11}$$

where r_k is the kth component of $\mathbf{r} \in \mathbb{R}^{\kappa-1}$ for $k \leqq \kappa - 1$ and $r_\kappa = 0$.

In Theorem 2, the condition $U(\omega, \mathbf{r}) = 0$ will be identified with the describing function problem. U is an approximation for V in the sense of the following lemma.

LEMMA 2. *Given Hypothesis* A, *U and V are continuous and*

$$|U(\omega, \mathbf{r}) - V(\omega, \mathbf{r})|_2 \leqq \frac{\|\mathscr{P}\mathscr{A}_\omega\| \cdot \|\mathscr{F}\| \cdot \|\mathscr{P}^*\mathscr{A}_\omega \mathscr{F} y(\mathbf{r})\|}{1 - \rho_\omega}.$$

($|\cdot|_2$ is the Euclidean norm in $\mathbb{R}^\kappa$.)

The main result of this section can now be given in terms of U, V, and $d[U, 0, \Omega]$, the local degree of U with respect to 0 and the set Ω. For the reader's convenience, Appendix II contains a definition of local degree and two related theorems.

THEOREM 1. *Given Hypothesis* A, *if there exists an open bounded set $\Omega \subset \mathbb{R}^\kappa$ such that*:

1. $\Omega \subset (\omega_*, \omega^*) \times \mathbb{R}^{\kappa-1}$,
2. $d[U, 0, \Omega] \neq 0$,
3. $|U(\omega, \mathbf{r}) - V(\omega, \mathbf{r})|_2 < |U(\omega, \mathbf{r})|_2$ *for each* $(\omega, \mathbf{r}) \in \partial\Omega$, *then there exists a point* $(\omega, \mathbf{r}) \in \Omega$ *such that*:
   a. $\mathscr{P}x(\omega, \mathbf{r}) = -\mathscr{P}\mathscr{A}_\omega \mathscr{F} x(\omega, \mathbf{r})$,
   b. $\mathscr{P}^*x(\omega, \mathbf{r}) = -\mathscr{P}^*\mathscr{A}_\omega \mathscr{F} x(\omega, \mathbf{r})$,
   c. $\|\mathscr{P}^*x(\omega, \mathbf{r})\| \leqq \dfrac{\|\mathscr{P}^*\mathscr{A}_\omega \mathscr{F} y(\mathbf{r})\|}{1 - \rho_\omega}$.

4. Application to autonomous systems. In order to use Theorem 1 to investigate the periodic oscillations of the autonomous feedback system in Fig. 1, it is necessary to make some assumptions about $\mathscr{N}$ and $\mathscr{G}$.

HYPOTHESIS B1. The system (S) shown in Fig. 1 satisfies:

1. $\mathscr{N}: \mathscr{L}_\pi^2 \to \mathscr{L}_\pi^2$.
2. If $x \in \mathscr{L}_\pi^2$ and $\tilde{x}(t) = x(\omega t)$ for all $t \in \mathbb{R}$ and some $\omega > 0$, then $(\mathscr{N}x)(\omega t) = (\mathscr{N}\tilde{x})(t)$ for all $t \in \mathbb{R}$.
3. $\mathscr{G}$ is a linear time-invariant system with frequency response function $G(j\omega)$.
4. For any odd integer k, $G(j\omega k)$ is continuous at each $\omega \in [\omega_*, \omega^*]$.
5. $|G(j\omega k)| \to 0$ as $k \overset{\text{odd}}{\to} \infty$, uniformly in $\omega \in [\omega_*, \omega^*]$.

Hypothesis B1.1 means π-symmetric inputs to $\mathscr{N}$ result in π-symmetric outputs. Hypothesis B1.2 means $\mathscr{N}$ is frequency independent for half-wave symmetric periodic inputs. Hypotheses B1.3, B1.4 and B1.5 imply $\mathscr{G}$ is a low-pass filter, at least for half-wave symmetric periodic inputs which have their fundamental frequencies in a given band.

Rather than consider the operation of $\mathscr{G}$ on periodic functions of all periods, we shall restrict our attention to an equivalent operation on functions of period 2π. The response to $\mathscr{G}$ is frequency-dependent so it is necessary to consider a continuum of operators, $\{\mathscr{G}_\omega : \omega \in [\omega_*, \omega^*]\}$, to obtain this equivalent operation.

DEFINITION (the operator $\mathscr{G}_\omega$). For each $\omega \in [\omega_*, \omega^*]$, $\mathscr{G}_\omega : \mathscr{L}_\pi^2 \to \mathscr{L}_\pi^2$ is such that if the Fourier series of z is $\sum_{n=-\infty}^{\infty} c_n e^{jn\theta}$, then the Fourier series of $\mathscr{G}_\omega z$ is

$$\sum_{n=-\infty}^{\infty} c_n G(jn\omega) e^{jn\theta}.$$

The equivalence of the operation of the linear time-invariant system $\mathscr{G}$ and the operators $\mathscr{G}_\omega$ is given by the following lemma.

LEMMA 3. *If* $x \in \mathscr{L}_\pi^2$ *is such that* $x = -\mathscr{G}_\omega \mathscr{N} x$, *then* $\tilde{x} = -\mathscr{G}\mathscr{N}\tilde{x}$, *where* $x(t) \triangleq \tilde{x}(\omega t)$.

Properties of the operators $\mathscr{G}_\omega$ are given in the next lemma.

LEMMA 4. *Given the system* (S) *and an even positive integer* κ:

a. $\|\mathscr{G}_\omega\| = \max\{|G(j\omega k)| : k \text{ an odd integer}\} < \infty$,

b. $\|\mathscr{P}\mathscr{G}_\omega\| = \max\{|G(j\omega k)| : k \text{ an odd integer}, k < \kappa\} < \infty$,

c. $\|\mathscr{P}^*\mathscr{G}_\omega\| = \max\{|G(j\omega k)| : k \text{ an odd integer}, k > \kappa\} < \infty$,

d. $\|\mathscr{P}^*(\mathscr{G}_\omega - \mathscr{G}_\nu)\| \leqq \|\mathscr{G}_\omega - \mathscr{G}_\nu\| \to 0$ *as* $\omega \to \nu$ *for all* $\nu \in [\omega_*, \omega^*]$.

In the most straightforward application of Theorem 1, $\mathscr{A}_\omega = \mathscr{G}_\omega$ and $\mathscr{F} = \mathscr{N}$. In this case the components of the function $U : [\omega_*, \omega^*] \times \mathbb{R}^{\kappa-1} \to \mathbb{R}^\kappa$ are given by the following lemma.

LEMMA 5. *Given Hypothesis* B1, *if* $\mathscr{A}_\omega = \mathscr{G}_\omega$, $\mathscr{F} = \mathscr{N}$, *and* U *is defined as in* (10), *then*

$$U_k(\omega, \mathbf{r}) = \begin{cases} \operatorname{Re}[r_k + G(j\omega k)\eta_k], & k \text{ odd}, \\ r_k + \operatorname{Im} G[j\omega(k-1)]\eta_{k-1}, & k \text{ even}, \end{cases}$$

where

$$\eta_k \triangleq \frac{j}{\pi}\int_0^{2\pi} \mathscr{N}[y(\mathbf{r})](\theta) e^{-jk\theta}\, d\theta.$$

If $\|\mathscr{N}\| < \infty$, then parts 1 and 2 of Hypothesis A are satisfied with $\mathscr{A}_\omega \triangleq \mathscr{G}_\omega$ and $\mathscr{F} \triangleq \mathscr{N}$. Defining $\rho_\omega = \|\mathscr{N}\| \max\{|G(j\omega k)| : k \text{ an odd integer}, k > \kappa\}$ for some even positive integer κ gives $\rho_\omega \geqq \|\mathscr{N}\| \|\mathscr{P}^*\mathscr{G}_\omega\|$ for all $\omega \in [\omega_*, \omega^*]$. Hypothesis A will be completely satisfied if $\sup\{\rho_\omega : \omega \in [\omega_*, \omega^*]\} < 1$, and this can always be accomplished by choosing κ large enough since $|G(j\omega k)| \to 0$ as $k \to \infty$, uniformly in $\omega \in [\omega_*, \omega^*]$.

Therefore Theorem 1 can be applied to any system (S) which satisfies the relatively weak requirements of Hypothesis B1, if $\|\mathscr{N}\| < \infty$, and if a set Ω can be found satisfying the theorem's hypothesis.

Searching for an appropriate set Ω is greatly simplified if it is possible to take $\kappa = 2$. In this case all computations are in $\mathbb{R}^2$ and can be performed conveniently in the complex plane. In fact, the conditions are closely related to the loci used in the describing function method. The case $\kappa = 2$ will be considered exclusively henceforth.

5. Justification and error bounds for the describing function method. Hypothesis B2 and Lemma 6 give conditions such that $d[\omega, 0, \Omega] \neq 0$ for a function W which will later be identified with the function U in Theorem 1. The function W is introduced to allow application of Theorem 1 to systems which do not satisfy $\|\mathscr{N}\| < \infty$. With $\kappa = 2$, $\mathbf{r}$ is a scalar and will henceforth be written as r.

HYPOTHESIS B2.

1. N has a continuous derivative on $[r_*, r^*]$, where $0 < r_* < r^*$ and
$$N(r) = \frac{j}{\pi r}\int_0^{2\pi} \mathcal{N}(r\xi_1)(\theta)e^{-j\theta}\,d\theta.$$
2. G has a continuous derivative on $[\omega_*, \omega^*]$, where $0 < \omega_* < \omega^*$ and G is abbreviated notation for $G[j(\cdot)]$.
3. $1 + N(r)G(j\omega) = 0$ with $(\omega, r) \in [\omega_*, \omega^*] \times [r_*, r^*]$ has a unique solution (ω_0, r_0) and
   a. $(\omega_0, r_0) \in (\omega_*, \omega^*) \times (r_*, r^*)$,
   b. $dN/dr|_{r_0} \neq 0$,
   c. $dG/d\omega|_{\omega_0} \neq 0$,
   d. the loci of $-1/N$ and G are not tangent at (ω_0, r_0).

Usually inspection of N, G and the loci of $-1/N$ and G is enough to check this hypothesis.

LEMMA 6. *Given Hypotheses* B1 *and* B2, *define* $W:[\omega_*, \omega^*] \times [r_*, r^*] \to \mathbb{R}^2$ *by*
$$W(\omega, r) \triangleq (\operatorname{Re}[r + rN(r)G(j\omega)], \operatorname{Im}[r + rN(r)G(j\omega)]);$$
then $d[W, 0, \Omega] \neq 0$ *for any open bounded set* Ω *in* $\mathbb{R}^2$ *such that* $(\omega_0, r_0) \in \Omega \subset (\omega_*, \omega^*) \times (r_*, r^*)$.

Hypotheses B3 and B3′ are given below in order to facilitate the statement of Theorem 2, the main result of this paper. Hypothesis B3 covers the case where $\|\mathcal{N}\| < \infty$.

HYPOTHESIS B3.

1. There is a real number M such that $\|\mathcal{N}\| \leqq M$.
2. $\rho_\omega = M \max\{|G(j\omega k)| : k = 3, 5, \cdots\} < 1$ for all $\omega \in [\omega_*, \omega^*]$.

A large class of nonlinear systems which exhibit hysteresis do not satisfy $\|\mathcal{N}\| < \infty$. This case is dealt with in Hypothesis B3′. To state it we need another definition.

DEFINITION (the space $\mathcal{L}_\pi^\infty$).
$$\mathcal{L}_\pi^\infty = \{x : \mathbb{R} \to \mathbb{R} | x \text{ is continuous, } x(\theta) = -x(\theta + \pi)\},$$
where $\|x\|_\infty = \max_{\theta \in \mathbb{R}} |x(\theta)|$.

Notice that any function in $\mathcal{L}_\pi^\infty$ is also in $\mathcal{L}_\pi^2$.

HYPOTHESIS B3′.

1. There is a real number M' such that $\|\mathcal{N}x - \mathcal{N}y\| \leqq M'\|x - y\|_\infty$ for all $x, y \in \mathcal{L}_\pi^\infty$.
2. $k|G(j\omega k)| \to 0$ as $k \overset{\text{odd}}{\to} \infty$, uniformly in $\omega \in [\omega_*, \omega^*]$.
3. $\rho'_\omega = (M'\pi/\sqrt{8}) \max\{k|G(j\omega k)| : k = 3, 5, \cdots\} < 1$ for all $\omega \in [\omega_*, \omega^*]$.

The main result of this paper can now be stated.

THEOREM 2. *Given Hypotheses* B1, B2, *and either* B3 *or* B3′, *if there exists an open bounded set* $\Omega \subset \mathbb{R}^2$ *such that*:

1. $(\omega_0, r_0) \in \Omega \subset (\omega_*, \omega^*) \times (r_*, r^*)$,
2. *for all* $(\omega, r) \in \partial\Omega$,
$$B(\omega)T(r) < \left|\frac{1}{N(r)} + G(j\omega)\right|,$$

where

$$T(r) \triangleq \left(\frac{\|\mathcal{N}(r\xi_1)\|^2}{|rN(r)|^2} - 1 \right)^{1/2}$$

and

$$B(\omega) = \begin{cases} \dfrac{|G(j\omega)|\rho_\omega}{1 - \rho_\omega} & \text{if B3 is satisfied,} \\ \dfrac{|G(j\omega)|\rho'_\omega}{1 - \rho'_\omega} & \text{if B3' is satisfied;} \end{cases}$$

then for some $(\omega, \mathbf{r}) \in \Omega$ *there exists an* $x \in \mathcal{L}^2_\pi$ *such that*:

a. $x = -\mathcal{G}_\omega \mathcal{N} x$,

b. $(\mathcal{P}x)(\theta) = r\xi_1(\theta) = r \sin \theta$,

c. $\|\mathcal{P}^* x\| \leqq |rN(r)|T(r)S(\omega)$,

where

$$S(\omega) = \begin{cases} \dfrac{\rho_\omega}{(1 - \rho_\omega)M} & \text{if B3 is satisfied,} \\ \dfrac{\rho'_\omega}{(1 - \rho'_\omega)M'} \dfrac{\sqrt{8}}{\pi} & \text{if B3' is satisfied;} \end{cases}$$

i.e., there exists a periodic function $\tilde{x}$ *with fundamental harmonic* $h(t) = r \sin \omega t$ *such that* $\tilde{x} = -\mathcal{G}\mathcal{N}\tilde{x}$ *for some initial state of* $\mathcal{G}$ *and*

$$\frac{\omega}{2\pi} \int_0^{\omega/(2\pi)} |\tilde{x}(t) - h(t)|^2 \, dt \leqq |rN(r)|T(r)S(\omega).$$

Theorem 2 gives not only sufficient conditions for the existence of a periodic response of system (S) but gives error bounds for that response in terms of error bounds on both its fundamental frequency and amplitude, and also on the $\mathcal{L}^2$-norm of its higher harmonics.

The heuristic justification of the describing function method stated in § 1 may now be rendered more precise. The statement "$\mathcal{N}$ is not too nonlinear" is replaced by the requirement of a small $T(r)$ in Theorem 2; notice that $T(r) = 0$ if $\mathcal{N}$ is linear. The statement "$\mathcal{G}$ is a sufficiently good low-pass filter" is replaced by Hypotheses B1.5, B3.2 (or B3.3'), and the requirement of a small $B(\omega)$ in Theorem 2. Hypothesis B2 specifies the nature of the intersection of the loci of $-1/N$ and G and makes precise the notion that the intersection should be "solid."

As a further connection with the describing function method we note that the key inequality in Theorem 2 has as its right-hand side a quantity which may be obtained immediately from the graphs of $G(j\omega)$ and $-1/N(r)$.

6. Examples. Theorem 2 can be applied to the autonomous nonlinear feedback system shown in Fig. 1 whether or not it contains hysteresis. This section contains an example of the use of Theorem 2 in either case.

The approach taken here is to use a computer to calculate $-1/N(r)$, $G(j\omega)$, and $|1/N(r) + G(j\omega)|/(B(\omega)T(r))$ as functions of ω and r. Plotting the loci of $-1/N$ and G gives the values of ω_0 and r_0 at their intersection. With the printout of

$|1/N(r) + G(j\omega)|/(B(\omega)T(r))$ in a rectangular array, it is a simple matter to find a rectangle in that array with sides $\omega = \omega_*$, $\omega = \omega^*$, $r = r_*$, and $r = r^*$ such that

$$1 < \frac{|1/N(r) + G(j\omega)|}{B(\omega)T(r)}$$

on its sides, and which contains (ω_0, r_0).

After finding such a rectangle it is a simple matter to check the remaining hypotheses of Theorem 2.

6.1. Without hysteresis. Consider the system shown in Fig. 1 with $\mathcal{G}$ having the transfer function

$$G(s) = \frac{1000(s^2 + 3s + 2)}{s^5 + 31s^4 + 259s^3 + 1319s^2 + 1289s + 990}$$

and $\mathcal{N}$ such that

$$(\mathcal{N}x)(\theta) = \begin{cases} 16 & \text{if } x(\theta) > 1, \\ 16x(\theta) & \text{if } |x(\theta)| \leqq 1, \\ -16 & \text{if } x(\theta) < -1. \end{cases}$$

For this $\mathcal{N}$ the describing function is [4, p. 59]

$$N(r) = \begin{cases} \dfrac{32}{\pi}\left[\sin^{-1}\dfrac{1}{r} + \dfrac{1}{r}\sqrt{1 - \dfrac{1}{r^2}}\right], & r > 1, \\ 16, & 0 \leqq r \leqq 1. \end{cases}$$

Hypothesis B1 is clearly satisfied for this system.

N is continuously differentiable for $r > 1$ and G is continuously differentiable on the imaginary axis. The loci $-1/N(r)$ and $G(j\omega)$ are shown in Fig. 2. They intersect at $(\omega_0, r_0) \approx (13.5, 4.6)$ and are clearly not tangent there. A simple computation, or closer examination of the loci, shows that both dN/dr and $dG/d\omega$ are not zero at the intersection. Therefore Hypothesis B2 is satisfied.

Hypothesis B3 is satisfied with $M = 16$, $\omega_* = 10$ and $\omega^* = 21$.

Table 1 gives $|1/N(r) + G(j\omega)|/(B(\omega)T(r))$ for several values of ω and r. This ratio is greater than one for the points tabulated on the boundary of the set

$$\Omega = \{(\omega, r): 10.5 < \omega < 14.5,\ 3.5 < r < 7.5\}.$$

It is in fact greater than one for every point on the boundary of this set. Therefore the system has a periodic response with fundamental frequency $\omega \in (10.5, 14.5)$ and with fundamental magnitude $r \in (3.5, 7.5)$.

6.2. With hysteresis. Consider the system shown in Fig. 1 with $\mathcal{G}$ having the transfer function

$$G(s) = 15{,}000e^{-s/3}/(s + 1)^4$$

and with $\mathcal{N}$ the backlash nonlinearity shown in Fig. 3. For this $\mathcal{N}$ the describing

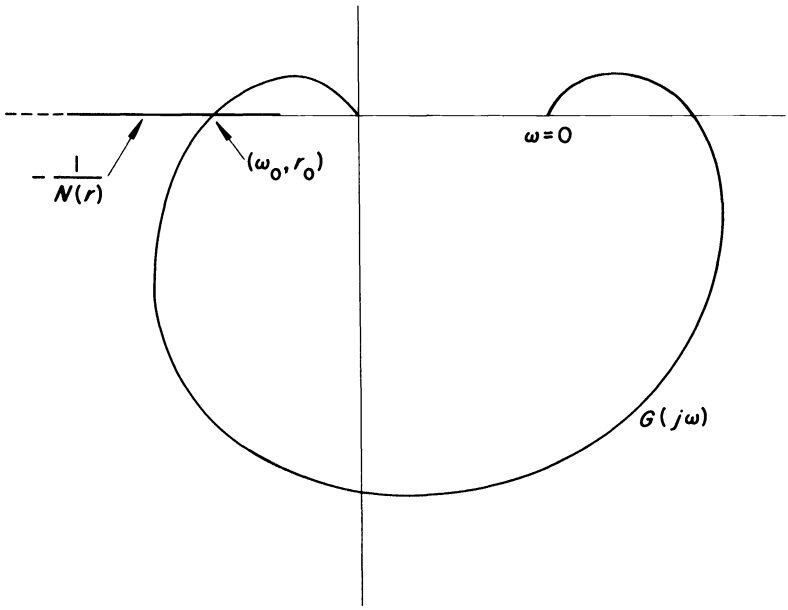


FIG. 2. *Loci for the example in* § 6.1 (*not to scale*)

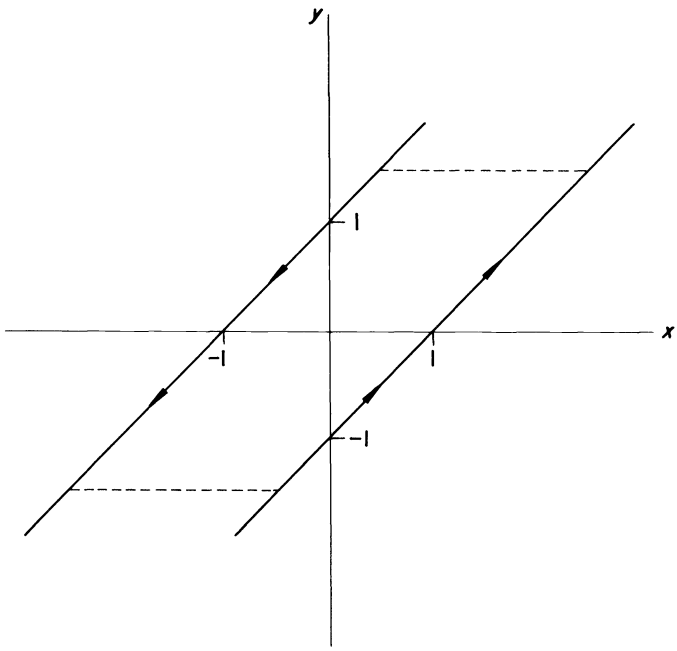


FIG. 3. *Backlash nonlinearity for the example in* § 6.2

TABLE 1

$|1/N(r) + G(j\omega)|/B(\omega)T(r)$ *versus* (ω, r) *used to find the set* Ω *for the example in* § 6.1

$r = 8.0$	.95	1.00	1.08	1.37	1.99	3.01	4.45	6.35	8.77	11.78	15.46	19.91	25.21
7.5	.99	**1.01**	1.01	1.14	1.59	2.45	3.74	5.47	7.69	**10.47**	13.88	18.01	22.95
7.0	1.04	1.06	1.00	.98	1.22	1.90	3.02	4.57	6.59	9.14	12.29	16.11	20.69
6.5	1.11	1.14	1.07	.94	.92	1.35	2.28	3.66	5.48	7.80	10.68	14.19	18.41
6.0	1.21	1.27	1.21	1.04	.81	.85	1.54	2.73	4.36	6.45	9.05	12.25	16.12
5.5	1.32	1.42	1.41	1.26	.97	.60	.80	1.78	3.20	5.07	7.41	10.30	13.80
5.0	1.46	1.60	1.65	1.57	1.33	.90	.32	.79	2.03	3.68	5.76	8.34	11.47
4.5	1.61	1.87	1.94	1.94	1.80	1.49	.98	.26	.90	2.30	4.10	6.37	9.14
4.0	1.80	2.08	2.28	2.38	2.36	2.20	1.87	1.36	.85	1.23	2.59	4.47	6.86
3.5	2.03	**2.39**	2.68	2.90	3.02	3.02	2.88	2.59	2.22	**1.79**	1.96	3.03	4.84
3.0	2.33	2.79	3.19	3.55	3.83	4.01	4.07	4.01	3.82	3.54	3.27	3.29	3.97
$\omega = 10.0$	**10.5**	11.0	11.5	12.0	12.5	13.0	13.5	14.0	**14.5**	15.0	15.5	16.0	

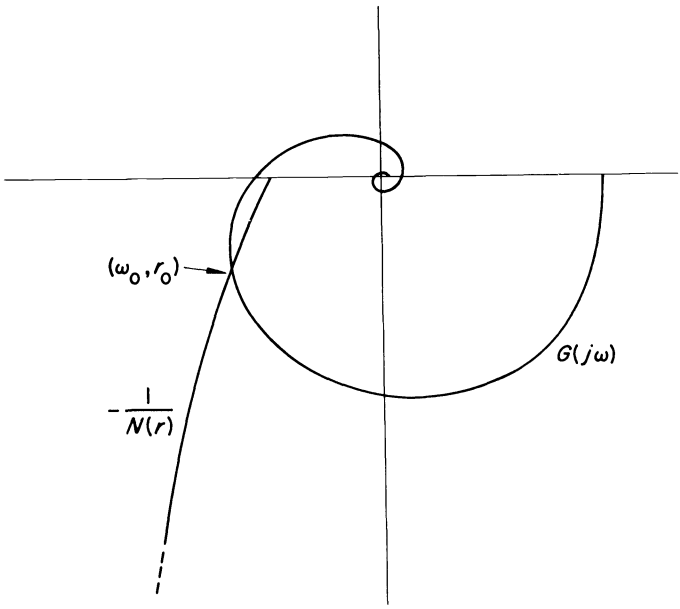


FIG. 4. *Loci for the example in* § 6.2 *(not to scale)*

function is [4, p. 69]

$$N(r) = \frac{1}{2} - \frac{1}{\pi}\sin^{-1}\left(1 - \frac{2}{r}\right) + \frac{1}{\pi}\left(1 - \frac{2}{r}\right)\sqrt{\frac{4}{r} - \frac{4}{r^2}} - j\frac{4}{\pi}\left(\frac{1}{r} - \frac{1}{r^2}\right), \quad r \geqq 1.$$

Hypothesis B1 is clearly satisfied for this system.

N is continuously differentiable for $r > 1$ and G is continuously differentiable on the imaginary axis. The loci of $G(j\omega)$ and $-1/N(r)$ are shown in Fig. 4. They intersect at $(\omega_0, r_0) \approx (1.35, 8.2)$ and are clearly not tangent there. Both dN/dr and $dG/d\omega$ are not zero at the intersection. Therefore Hypothesis B2 is satisfied.

Hypothesis B3′ is satisfied with $M' = \sqrt{2}$, $\omega_* = 7.5$ and $\omega^* = 11$.

Finally the set

$$\Omega = \{(\omega, r): 8.0 < \omega < 8.4,\ 1.3 < r < 1.4\}$$

satisfies the remaining hypothesis of Theorem 2. Therefore the system has a periodic response with fundamental frequency $\omega \in (8.0, 8.4)$ and fundamental magnitude $r \in (1.3, 1.4)$.

7. Conclusion. Sufficient conditions for the existence of half-wave symmetric oscillations are given and error bounds for the approximate solution found by the use of the describing function method are provided. The heuristic justifications concerning the describing function method are validated and rendered more precise.

Appendix I. This appendix consists of the proofs of all theorems and lemmas in this paper. Appendix II contains the definition of local degree and statements of theorems which are used in the proofs here.

Proof of Lemma 1. Let

$$F(\omega, \mathbf{r})x \triangleq -\mathscr{P}^*\mathscr{A}_\omega\mathscr{F}[y(\mathbf{r}) + x]$$

for $(\omega, \mathbf{r}) \in [\omega_*, \omega^*] \times \mathbb{R}^{\kappa-1}$.

Given $(\omega, \mathbf{r}) \in [\omega_*, \omega^*] \times \mathbb{R}^{\kappa-1}$ and x and $z \in \mathscr{P}^*\mathscr{L}_\pi^2$, the linearity of $\mathscr{A}_\omega$ and the bounds on the operator norms assumed in Hypothesis A yield

$$\begin{aligned}\|F(\omega, \mathbf{r})x - F(\omega, \mathbf{r})z\| &= \|\mathscr{P}^*\mathscr{A}_\omega\mathscr{F}[y(\mathbf{r}) + x] - \mathscr{P}^*\mathscr{A}_\omega\mathscr{F}[y(\mathbf{r}) + z]\| \\ &= \|\mathscr{P}^*\mathscr{A}_\omega[\mathscr{F}[y(\mathbf{r}) + x] - \mathscr{F}[y(\mathbf{r}) + z]\| \\ &\leqq \|\mathscr{P}^*\mathscr{A}_\omega\| \cdot \|\mathscr{F}\| \cdot \|x - z\| \\ &\leqq \rho_\omega\|x - z\|.\end{aligned}$$

$F(\omega, \mathbf{r}): \mathscr{P}^*\mathscr{L}_\pi^2 \to \mathscr{P}^*\mathscr{L}_\pi^2$ is a contraction for each $(\omega, \mathbf{r}) \in [\omega_*, \omega^*] \times \mathbb{R}^{\kappa-1}$ since $\rho_\omega < 1$ for each $\omega \in [\omega_*, \omega^*]$

By the contraction theorem, Theorem II.1 in Appendix II, there is a unique $x^*(\omega, \mathbf{r}) \in \mathscr{P}^*\mathscr{L}_\pi^2$ such that

$$x^*(\omega, \mathbf{r}) = -\mathscr{P}^*\mathscr{A}_\omega\mathscr{F}[y(\mathbf{r}) + x^*(\omega, \mathbf{r})] \tag{I.1}$$

and

$$\|x^*(\omega, \mathbf{r})\| \leqq \frac{\|\mathscr{P}^*\mathscr{A}_\omega\mathscr{F}y(\mathbf{r})\|}{1 - \rho_\omega}. \tag{I.2}$$

$x: [\omega_*, \omega^*] \times \mathbb{R}^{\kappa-1} \to \mathscr{L}_\pi^2$ is continuous by Theorem II.1, since $F(\omega, \mathbf{r})$ is continuous.

The remainder of the lemma may be established by noting that $x(\omega, \mathbf{r}) = y(\mathbf{r}) + x^*(\omega, \mathbf{r})$ implies $\mathscr{P}x(\omega, \mathbf{r}) = y(\mathbf{r})$, which is part b.

It also implies

$$\begin{aligned}\mathscr{P}^*x(\omega, \mathbf{r}) &= x^*(\omega, \mathbf{r}) \\ &= -\mathscr{P}^*\mathscr{A}_\omega\mathscr{F}[y(\mathbf{r}) + x^*(\omega, \mathbf{r})] \quad \text{(by (I.1))} \\ &= -\mathscr{P}^*\mathscr{A}_\omega\mathscr{F}x(\omega, \mathbf{r}).\end{aligned} \tag{I.3}$$

This completes part c.

Equations (I.2) and (I.3) imply

$$\|\mathscr{P}^* x(\omega, \mathbf{r})\| \leqq \frac{\|\mathscr{P}^* \mathscr{A}_\omega \mathscr{F} y(\mathbf{r})\|}{1 - \rho_\omega}.$$

Proof of Lemma 2. (i) By Lemma 1, $\mathscr{A}_\omega \mathscr{F} x(\omega, \mathbf{r})$ is continuous in $(\omega, \mathbf{r})$. Therefore $\langle \mathscr{A}_\omega \mathscr{F} x(\omega, \mathbf{r}), \xi_k \rangle$ is continuous in $(\omega, \mathbf{r})$, and as a result of (11), V is continuous. Similarly U is continuous.

(ii)

$$\begin{aligned} |U_k(\omega, \mathbf{r}) - V_k(\omega, \mathbf{r})|^2 &= |\langle \mathscr{A}_\omega \mathscr{F} y(\mathbf{r}), \xi_k \rangle - \langle \mathscr{A}_\omega \mathscr{F} x(\omega, \mathbf{r}), \xi_k \rangle|^2 \\ &= |\langle \mathscr{A}_\omega [\mathscr{F} y(\mathbf{r}) - \mathscr{F} x(\omega, \mathbf{r})], \xi_k \rangle|^2 \\ &= \langle \mathscr{A}_\omega [\mathscr{F} y(\mathbf{r}) - \mathscr{F} x(\omega, \mathbf{r})], \xi_k \rangle^2. \end{aligned}$$

Therefore,

$$\begin{aligned} |U(\omega, \mathbf{r}) - V(\omega, \mathbf{r})|_2^2 &= \sum_{k=1}^{\kappa} \langle \mathscr{A}_\omega [\mathscr{F} y(\mathbf{r}) - \mathscr{F} x(\omega, \mathbf{r})], \xi_k \rangle^2 \\ &= \|\mathscr{P} \mathscr{A}_\omega [\mathscr{F} y(\mathbf{r}) - \mathscr{F} x(\omega, \mathbf{r})]\|^2. \end{aligned}$$

Therefore,

$$\begin{aligned} |U(\omega, \mathbf{r}) - V(\omega, \mathbf{r})|_2 &= \|\mathscr{P} \mathscr{A}_\omega [\mathscr{F} y(\mathbf{r}) - \mathscr{F} x(\omega, \mathbf{r})]\| \\ &\leqq \|\mathscr{P} \mathscr{A}_\omega\| \cdot \|\mathscr{F}\| \cdot \|y(\mathbf{r}) - x(\omega, \mathbf{r})\| \\ &\leqq \|\mathscr{P} \mathscr{A}_\omega\| \cdot \|\mathscr{F}\| \cdot \|x^*(\omega, \mathbf{r})\| \\ &\leqq \|\mathscr{P} \mathscr{A}_\omega\| \cdot \|\mathscr{F}\| \cdot \| \frac{\|\mathscr{P}^* \mathscr{A}_\omega \mathscr{F} y(\mathbf{r})\|}{1 - \rho_\omega}. \end{aligned}$$

Proof of Theorem 1. (i) In light of Theorem II.2 in Appendix II, define

$$\Phi(\omega, \mathbf{r}, \mu) = U(\omega, \mathbf{r}) - \mu[U(\omega, \mathbf{r}) - V(\omega, \mathbf{r})] \quad \text{for } \mu \in [0, 1] \quad \text{and} \quad (\omega, \mathbf{r}) \in \bar{\Omega}.$$

Then $\Phi : \bar{\Omega} \times [0, 1] \to \mathbb{R}^\kappa$ is continuous since U and V are continuous.

If $(\omega, \mathbf{r}) \in \partial\Omega$,

$$\begin{aligned} |\Phi(\omega, \mathbf{r}, \mu)|_2 &= |U(\omega, \mathbf{r}) - \mu[U(\omega, \mathbf{r}) - V(\omega, \mathbf{r})]|_2 \\ &\geqq |U(\omega, \mathbf{r})|_2 - \mu |U(\omega, \mathbf{r}) - V(\omega, \mathbf{r})|_2 \\ &\geqq |U(\omega, \mathbf{r})| - |U(\omega, \mathbf{r}) - V(\omega, \mathbf{r})|_2 \\ &> 0 \quad \text{(by assumption 3 of Theorem 1).} \end{aligned}$$

Therefore $(\omega, \mathbf{r}) \in \partial\Omega$ implies $\Phi(\omega, \mathbf{r}, \mu) \neq 0$ for all $\mu \in [0, 1]$.

By Theorem II.2, $d[\Phi(\cdot, \cdot, 1), 0, \Omega] = d[\Phi(\cdot, \cdot, 0), 0, \Omega]$. Now $\Phi(\cdot, \cdot, 1) = V$ and $\Phi(\cdot, \cdot, 0) = U$ so by assumption 2 of Theorem 1,

$$d[V, 0, \Omega] = d[U, 0, \Omega] \neq 0.$$

Therefore, by Theorem II.3 there is a point $(\omega, \mathbf{r}) \in \Omega$ such that $V(\omega, \mathbf{r}) = 0$; i.e., using (11), we have

$$-r_k = \langle \mathscr{A}_\omega \mathscr{F} x(\omega, \mathbf{r}), \xi_k \rangle \quad \text{for } k = 1, 2, \cdots, \kappa.$$

Therefore,

$$\mathscr{P}\mathscr{A}_\omega\mathscr{F}x(\omega, \mathbf{r}) = -\sum_{k=1}^{\kappa-1} r_k\xi_k$$

$$= -y(\mathbf{r}) = -\mathscr{P}x(\omega, \mathbf{r}).$$

(ii) By Lemma 1,

$$\mathscr{P}^*\mathscr{A}_\omega\mathscr{F}x(\omega, \mathbf{r}) = -\mathscr{P}^*x(\omega, \mathbf{r}).$$

(iii) Also by Lemma 1,

$$\|\mathscr{P}^*x(\omega, \mathbf{r})\| \leqq \frac{\|\mathscr{P}^*\mathscr{A}_\omega\mathscr{F}y(\mathbf{r})\|}{1 - \rho_\omega},$$

which proves the theorem.

Proof of Lemma 3. Suppose $x \in \mathscr{L}_\pi^2, x = -\mathscr{G}_\omega\mathscr{N}x, \omega \in [\omega_*, \omega^*], \omega > 0$, and $\tilde{x}(t) = x(\omega t)$ for all $t \in \mathbb{R}$. Then

$$\tilde{x}(t) = x(\omega t) = -[\mathscr{G}_\omega\mathscr{N}x](\omega t)$$

$$= -\left[\sum_{-\infty}^{\infty} G(j\omega k)\eta_k e^{jk\theta}\right]\Bigg|_{\theta=\omega t}$$

$$= -\sum_{-\infty}^{\infty} G(j\omega k)\eta_k e^{jk\omega t},$$

where η_k is the kth complex Fourier coefficient of $\mathscr{N}x$; i.e.,

$$\eta_k = \frac{1}{\pi}\int_0^{2\pi} [\mathscr{N}x](\theta)e^{-jk\theta}\,d\theta$$

$$= \frac{2}{T}\int_0^T [\mathscr{N}x](\omega t)e^{-jk\omega t}\,dt.$$

Therefore η_k is the kth complex Fourier coefficient of $\mathscr{N}\tilde{x}$ as well. For the input $\mathscr{N}x: \mathbb{R} \to \mathbb{R}$ the output of $\mathscr{G}$ is

$$[\mathscr{G}\mathscr{N}\tilde{x}](t) = \sum_{k=-\infty}^{\infty} G(j\omega k)\eta_k e^{j\omega kt}, \quad t \in \mathbb{R}.$$

Therefore $\tilde{x} = -\mathscr{G}\mathscr{N}\tilde{x}$.

Proof of Lemma 4. To prove parts a, b and c, pick any $x \in \mathscr{L}_\pi^2$, expand it in a Fourier series, and use the usual manipulations. This same approach gives

$$\|\mathscr{G}_\omega - \mathscr{G}_\nu\| = \max\{|G(j\omega k) - G(j\nu k)| : k \text{ odd}\}.$$

By Hypotheses B1.4 and B1.5, $\|\mathscr{G}_\omega - \mathscr{G}_\nu\| \to 0$ as $\omega \to \nu$.

Proof of Lemma 5.

$$U_k(\omega, r) = r_k + \langle \mathscr{G}_\omega\mathscr{N}y(\mathbf{r}), \xi_k \rangle.$$

The complex Fourier coefficients of $\mathscr{N}y(\mathbf{r})$ for $l = \pm 1, \pm 3, \pm 5, \cdots$ are

$$\frac{1}{2\pi}\int_0^{2\pi} \mathscr{N}[y(\mathbf{r})](\theta)e^{-jl\theta}\,d\theta = \frac{1}{2j}\eta_l.$$

Therefore $\mathscr{G}_\omega \mathscr{N} y(\mathbf{r})$ has Fourier series $(1/(2j)) \sum_{-\infty}^{\infty} G(j\omega l)\eta_l e^{jl\theta}$. Now for $n = \pm 1, \pm 3, \cdots$,

$$\frac{1}{\pi}\int_0^{2\pi}\left[\frac{1}{2j}\sum_{-\infty}^{\infty} G(j\omega l)\eta_l e^{jl\theta}\right]e^{+jn\theta}\,d\theta = \frac{1}{j}G(j\omega n)\eta_{-n}.$$

Therefore if k is odd, $\xi_k(\theta) = \sin k\theta$ and

$$\begin{aligned}\langle \mathscr{G}_\omega \mathscr{N} y(\mathbf{r}), \xi_k\rangle &= \frac{1}{\pi}\int_0^{2\pi}\left[\frac{1}{2j}\sum_{-\infty}^{\infty} G(j\omega l)\eta_l e^{jl\theta}\right]\left[\frac{e^{j\theta k} - e^{-j\theta k}}{2j}\right]d\theta \\ &= -\tfrac{1}{2}G(-j\omega k)\eta_{-k} + \tfrac{1}{2}G(j\omega k)\eta_k \\ &= \tfrac{1}{2}[\bar{G}(j\omega k)\bar{\eta}_k + G(j\omega k)\eta_k] \\ &= \operatorname{Re} G(j\omega k)\eta_k\end{aligned}$$

since

$$\eta_{-k} = \frac{j}{\pi}\int_0^{2\pi} \mathscr{N}(y(\mathbf{r}))(\theta)e^{jk\theta}\,d\theta = -\bar{\eta}_k,$$

the conjugate of $-\eta_k$. Therefore

$$\begin{aligned}U_k(\omega, \mathbf{r}) &= r_k + \langle \mathscr{G}_\omega \mathscr{N} y(\mathbf{r}), \xi_k\rangle \\ &= \operatorname{Re}[r_k + G(j\omega k)\eta_k], \qquad k \text{ odd}.\end{aligned}$$

If k is even, $\xi_k(\theta) = \cos(k-1)\theta$ and

$$\begin{aligned}\langle \mathscr{G}_\omega \mathscr{N} y(\mathbf{r}), \xi_k\rangle &= \frac{1}{\pi}\int_0^{2\pi}\left[\frac{1}{2j}\sum_{-\infty}^{\infty} G(j\omega l)\eta_k e^{jl\theta}\right]\left[\frac{e^{j\theta(k-1)} + e^{-j\theta(k-1)}}{2}\right] \\ &= \frac{1}{2j}G(-j\omega(k-1))\eta_{-(k-1)} + \frac{1}{2j}G(j\omega(k-1))\eta_{k-1} \\ &= \frac{1}{2j}[-\bar{G}(j\omega(k-1))\bar{\eta}_{k-1} + G(j\omega(k-1))\eta_{k-1}] \\ &= \frac{1}{2j}[2j \operatorname{Im} G(j\omega(k-1))\eta_{k-1}] \\ &= \operatorname{Im} G(j\omega(k-1))\eta_{k-1}.\end{aligned}$$

Therefore,

$$\begin{aligned}U_k(\omega, \mathbf{r}) &= r_k + \langle \mathscr{G}_\omega \mathscr{N} y(\mathbf{r}), \xi_k\rangle \\ &= r_k + \operatorname{Im}[G(j\omega(k-1))\eta_{k-1}], \qquad k \text{ even}.\end{aligned}$$

Proof of Lemma 6. From assumption 3 of Hypothesis B2, for all $(\omega, r) \in \Omega \subset [\omega_*, \omega^*] \times [r_*, r^*]$,

$$W(\omega, r) \triangleq (\operatorname{Re}[r + rN(r)G(j\omega)], \operatorname{Im}[r + rN(r)G(j\omega)]) = (0, 0)$$

if and only if $(\omega, r) = (\omega_0, r_0) \in \Omega$. By assumptions 1 and 2 of Hypothesis B2, $W(\omega, r)$ is $C^{(1)}$ on $\bar{\Omega}$. Using the analytic definition in Appendix II, we may therefore calculate the local degree of the map W relative to the point 0 and the set Ω,

$$d[W, 0, \Omega] = \operatorname{sgn}\det J_W(\omega_0, r_0),$$

where

$$J_W(\omega_0, r_0) = \begin{bmatrix} \dfrac{\partial W_1}{\partial \omega} & \dfrac{\partial W_1}{\partial r} \\ \dfrac{\partial W_2}{\partial \omega} & \dfrac{\partial W_2}{\partial r} \end{bmatrix}\Bigg|_{(\omega_0, r_0)}.$$

We next use the remaining assumptions of Hypothesis B2 to show that $\det J_W(\omega_0, r_0) \neq 0$ and therefore $d[W, 0, \Omega] \neq 0$. By direct calculation,

$$J_W(\omega_0, r_0) = r_0 \begin{bmatrix} \operatorname{Re} N(r_0)G'(j\omega_0) & \operatorname{Re} N'(r_0)G(j\omega_0) \\ \operatorname{Im} N(r_0)G'(j\omega_0) & \operatorname{Im} N'(r_0)G(j\omega_0) \end{bmatrix},$$

where the prime indicates the derivative and use has been made of the assumption in B2 that $1 + N(r_0)G(j\omega_0) = 0$.

Now if the columns of J_W are linearly independent, $\det [J_W(\omega_0, r_0)] \neq 0$. It is natural to reintroduce the complex plane; $\det [J_W(\omega_0, r_0)] \neq 0$ if the complex numbers $z_1 = N(r_0)G'(j\omega_0)$ and $z_2 = N'(r_0)G(j\omega_0)$ are linearly independent. Now $N(r_0)G(j\omega_0) = -1$ implies $N(r_0)$ and $G(j\omega_0) \neq 0$. Assumption 3 of B2 specifies $G'(j\omega_0)$ and $N'(r_0) \neq 0$. Therefore neither of the complex numbers z_1 or z_2 is zero.

We now use the relation

$$\frac{d}{dr}\left[-\frac{1}{N(r)}\right]\Bigg|_{r_0} = \frac{N'(r_0)}{N^2(r_0)}$$

and find

$$z_2 = N'(r_0)G(j\omega_0) = -N(r_0)\frac{d}{dr}\left[-\frac{1}{N(r)}\right]\Bigg|_{r_0}.$$

It is clear that the two (nonzero) complex numbers z_1 and z_2 are linearly independent if $G'(j\omega_0)$ and $(d/dr)[-1/N(r)]|_{r_0}$ are linearly independent. Geometrically, $G'(j\omega_0)$ is a vector tangent to the Nyquist locus at the point ω_0 while $(d/dr)[-1/N(r)]|_{r_0}$ is a vector tangent to the critical locus at r_0. Since by assumption neither vector is zero, the linear independence is assured if the two tangents to the loci at their point of intersection (ω_0, r_0) differ. This is assured by part d in assumption 3 of Hypothesis B2.

Therefore $d[\omega, 0, \Omega] \neq 0$.

Proof of Theorem 2. It is useful first to derive expressions (I.4) and (I.5) below, and then to consider cases B and B′ separately. Set $\kappa = 2$ for defining $\mathscr{P}$ and $\mathscr{P}^*$. Then:

(i)

$$\begin{aligned} rN(r) &= \frac{j}{\pi}\int_0^{2\pi} [\mathscr{N}(r\xi_1)](\theta)e^{-j\theta}\,d\theta \\ &= \frac{j}{\pi}\int_0^{2\pi} [\mathscr{N}(r\xi_1)](\theta)[\cos\theta - j\sin\theta]\,d\theta \\ &= \langle \mathscr{N}(r\xi_1), \xi_1\rangle + j\langle \mathscr{N}(r\xi_1), \xi_2\rangle, \end{aligned}$$

so

$$[\mathscr{P}\mathscr{N}(r\xi_1)](\theta) \triangleq \langle \mathscr{N}(r\xi_1), \xi_1 \rangle \sin\theta + \langle \mathscr{N}(r\xi_1), \xi_2 \rangle \cos\theta,$$

$$[\mathscr{P}\mathscr{N}(r\xi_1)](\theta) = |rN(r)| \sin[\theta + \angle N(r)] \quad \text{for } r > 0. \tag{I.4}$$

(ii)

$$\begin{aligned} \|\mathscr{P}^*\mathscr{N}(r\xi_1)\|^2 &= \|\mathscr{N}(r\xi_1)\|^2 - \|\mathscr{P}\mathscr{N}(r\xi_1)\|^2 \\ &= \|\mathscr{N}(r\xi_1)\|^2 - |rN(r)|^2 \end{aligned}$$

so

$$\begin{aligned} T(r) &= \sqrt{\frac{\|\mathscr{N}(r\xi_1)\|^2}{|rN(r)|^2} - 1} = \sqrt{\frac{\|\mathscr{N}(r\xi_1)\|^2 - |rN(r)|^2}{|rN(r)|^2}} \\ &= \frac{\|\mathscr{P}^*\mathscr{N}(r\xi_1)\|}{|rN(r)|}. \end{aligned}$$

Therefore $B(\omega)T(r) < |1/N(r) + G(j\omega)|$ implies

$$B(\omega)\|\mathscr{P}^*\mathscr{N}(r\xi_1)\| < |r + rN(r)G(j\omega)| = |W(\omega, r)|_2. \tag{I.5}$$

Case B3. (i) Hypothesis A is satisfied with $\mathscr{A}_\omega \triangleq \mathscr{G}_\omega$, $\mathscr{F} \triangleq \mathscr{N}$, $\kappa \triangleq 2$ and $y(r) = r\xi_1$.

Proof. Clearly part 1 is satisfied by Hypothesis B1 and Lemma 4. Parts 2 and 3 are satisfied by Hypothesis B3 and Lemma 4.

(ii)

$$\begin{aligned} \langle \mathscr{G}_\omega\mathscr{N}(r\xi_1), \xi_1 \rangle &= \langle \mathscr{G}_\omega\mathscr{P}\mathscr{N}(r\xi_1), \xi_1 \rangle \\ &= \frac{1}{\pi}\int_0^{2\pi} |G(j\omega)|\,|rN(r)| \sin(\theta + \angle N + \angle G)\sin\theta\, d\theta \\ &= |G(j\omega)| \cdot |rN(r)| \cos(\angle N + \angle G) \\ &= \operatorname{Re}[rN(r)G(j\omega)], \qquad r \geqq r_*. \end{aligned}$$

Similarly,

$$\langle \mathscr{G}_\omega\mathscr{N}(r\xi_1), \xi_2 \rangle = \operatorname{Im}[rN(r)G(j\omega)].$$

Therefore by (7),

$$\begin{aligned} U(\omega, r) &= (r + \operatorname{Re}[rN(r)G(j\omega)], \operatorname{Im}[rN(r)G(j\omega)]) \\ &= (\operatorname{Re}[r + rN(r)G(j\omega)], \operatorname{Im}[r + rN(r)G(j\omega)]) \\ &= W(\omega, r). \end{aligned}$$

So $d[U, 0, \Omega] \neq 0$ by Lemma 4.

(iii)

$$|U(\omega, r) - V(\omega, r)|_2 \leqq \frac{\|\mathscr{P}\mathscr{G}_\omega\| \cdot \|\mathscr{N}\| \cdot \|\mathscr{P}^*\mathscr{G}_\omega\mathscr{N}(r\xi_1)\|}{1 - \rho_\omega}$$

$$\leqq \frac{\|\mathscr{P}\mathscr{G}_\omega\| \cdot \|\mathscr{N}\| \cdot \|\mathscr{P}^*\mathscr{G}_\omega\| \cdot \|\mathscr{P}^*\mathscr{N}(r\xi_1)\|}{1 - \rho_\omega}$$

$$= \frac{|G(j\omega)|\rho_\omega}{1 - \rho_\omega}\|\mathscr{P}^*\mathscr{N}(r\xi_1)\|$$

$$= B(\omega)\|\mathscr{P}^*\mathscr{N}(r\xi_1)\|$$

$$< |W(\omega, r)|_2 = |U(\omega, r)|_2,$$

by (I.5) for all $(\omega, r) \in \partial\Omega$. Therefore,

$$|U(\omega, r) - V(\omega, r)|_2 < |U(\omega, r)| \quad \text{for all } (\omega, r) \in \partial\Omega.$$

By Theorem 1, there is an $(\omega, r) \in \Omega$ such that $x = -\mathscr{G}_\omega\mathscr{N}x$, where $x = x(\omega, r)$. Also for $x \in \mathscr{L}_\pi^2$,

$$\mathscr{P}x = r\xi_1 \quad \text{and} \quad \|\mathscr{P}^*x\| = \|x^*(\omega, r)\| \leqq \frac{\|\mathscr{P}^*\mathscr{G}_\omega\mathscr{N}(r\xi_1)\|}{1 - \rho_\omega}$$

$$\leqq \frac{\|\mathscr{P}^*\mathscr{G}_\omega\| \cdot \|\mathscr{P}^*\mathscr{N}(r\xi_1)\|}{1 - \rho_\omega}$$

$$= \left[\frac{\rho_\omega}{1 - \rho_\omega}\|\mathscr{P}^*\mathscr{N}(r\xi_1)\|\right]\frac{1}{M}$$

$$= \frac{\rho_\omega}{1 - \rho_\omega}\frac{|rN(r)|}{M}T(r).$$

Case B3′: The proof in this case is very similar to the preceding one. The essential difference is that hypothesis B3.1 is not satisfied. This problem is avoided by defining an operator $\mathscr{C}: \mathscr{L}_\pi^2 \to \mathscr{L}_\pi^2$ such that y is the derivative of $\mathscr{C}y$, and by defining an operator $\mathscr{D}: \mathscr{G}_\omega\mathscr{L}_\pi^2 \to \mathscr{L}_\pi^2$ such that $\mathscr{D}y$ is the derivative of y, i.e., $\mathscr{D}\mathscr{C}y = y$.

The proof now proceeds in essentially the same way as before but with $\mathscr{A}_\omega = \mathscr{D}\mathscr{G}_\omega$ and $\mathscr{F} = \mathscr{N}\mathscr{C}$.

(i) Define $\mathscr{C}: \mathscr{L}_\pi^2 \to \mathscr{L}_\pi^2$ such that

$$(\mathscr{C}x)(\theta) = \begin{cases} \frac{1}{2}\int_0^\pi \operatorname{sgn}(\theta - \tau)x(\tau)\,d\tau, & 0 \leqq \theta < \pi, \\ -(\mathscr{C}x)(\theta - \pi), & \pi \leqq \theta < 2\pi. \end{cases}$$

Note 1.

$$(\mathscr{C}x)(\theta) = \int_0^\theta x(\tau)\,d\tau - \frac{1}{2}\int_0^\pi x(\tau)\,d\tau, \qquad 0 \leqq \theta < \pi,$$

so $(\mathscr{C}x)(\theta) = -(\mathscr{C}x)(\pi) = (\mathscr{C}x)(2\pi)$ and $\mathscr{C}x$ is continuous for all $x \in \mathscr{L}^2_\pi$.

Note 2.

$$\mathscr{C}[e^{jn(\cdot)}] = \frac{1}{jn} e^{jn(\cdot)}, \quad n \text{ odd}.$$

Note 3.

$$\begin{aligned} |(\mathscr{C}x)(\theta)| &= \frac{1}{2}\left| \int_0^\pi \operatorname{sgn}(\theta - \tau) x(\tau)\, d\tau \right| \quad (0 \leqq \theta < 2\pi) \\ &\leqq \frac{1}{2}\int_0^\pi |x(\tau)|\, d\tau \\ &\leqq \frac{1}{2}\left(\int_0^\pi 1^2\, d\tau\right)^{1/2}\left(\int_0^\pi |x(\tau)|^2\, d\tau\right)^{1/2} \\ &= \frac{\pi}{2\sqrt{2}}\|x\|. \end{aligned}$$

Therefore,

$$\|\mathscr{C}x\|_\infty \leqq \frac{\pi}{2\sqrt{2}}\|x\| \quad \text{for all } x \in \mathscr{L}^2_\pi.$$

Now

$$\begin{aligned} \|\mathscr{N}\mathscr{C}x - \mathscr{N}\mathscr{C}y\| &\leqq M'\|\mathscr{C}x - \mathscr{C}y\|_\infty \\ &\leqq M'\frac{\pi}{2\sqrt{2}}\|x - y\| \quad \text{for all } x, y \in \mathscr{L}^2_\pi, \end{aligned}$$

so $\|\mathscr{N}\mathscr{C}\| \leqq \pi M'/(2\sqrt{2})$.

(ii) Define $\mathscr{D}$ such that $\mathscr{D}[e^{jn(\cdot)}] = jne^{jn(\cdot)}$ and

$$\mathscr{D}\sum_{k=1}^{\infty} a_k\xi_k = \sum_{k=1}^{\infty} a_k\mathscr{D}\xi_k.$$

Now $\mathscr{D}$ does not map $\mathscr{L}^2_\pi$ into $\mathscr{L}^2_\pi$, but $\mathscr{D}\mathscr{G}_\omega$ does. $[\mathscr{D}\mathscr{G}_\omega][e^{jn(\cdot)}] = jnG(jn\omega) \cdot e^{jn(\cdot)}$, and $[jnG(jn\omega)]$ satisfies Hypothesis B1, by Hypothesis B3′, so by Lemma 4,

$$\begin{aligned} &\mathscr{D}\mathscr{G}_\omega : \mathscr{L}^2_\pi \to \mathscr{L}^2_\pi, \\ &\|\mathscr{D}\mathscr{G}_\omega\| = \max\{n|G(jn\omega)| : n \text{ odd}\}, \\ &\|\mathscr{P}^*\mathscr{D}\mathscr{G}_\omega\| = \max\{k|G(jk\omega)| : k = 3, 5, \cdots\}, \\ &\|\mathscr{P}\mathscr{D}\mathscr{G}_\omega\| = |G(j\omega)|, \\ &\|\mathscr{P}^*(\mathscr{D}\mathscr{G}_\omega - \mathscr{D}\mathscr{G}_\nu)\| \leqq \|\mathscr{D}\mathscr{G}_\omega - \mathscr{D}\mathscr{G}_\nu\| \underset{\nu\to\omega}{\to} 0 \quad \text{for all } \omega \in [\omega_*, \omega^*]. \end{aligned}$$

(iii)

$$\begin{aligned} \|\mathscr{P}^*\mathscr{D}\mathscr{G}_\omega\| \cdot \|\mathscr{N}\mathscr{C}\| &\leqq \frac{\pi M'}{2\sqrt{2}} \max\{k|G(jk\omega)| : k = 3, 5, \cdots\} \\ &\triangleq \rho'_\omega < 1 \quad \text{for all } \omega \in [\omega_*, \omega^*] \end{aligned}$$

by Hypothesis B3′. Therefore Hypothesis A is satisfied with $\mathscr{A}_\omega \mathscr{D}\mathscr{G}_\omega$, $\mathscr{F} = \mathscr{N}\mathscr{C}$, and $\kappa = 2$.

(iv) Hypothesis A is independent of the basis chosen for $\mathscr{L}_\pi^2$ so define a new orthonormal basis for $\mathscr{L}_\pi^2$, $\{z_k\}_1^\infty$, such that

$$z_1 = \xi_2, \qquad z_2 = -\xi_1, \qquad z_k = \xi_k \quad \text{for } k > 2.$$

Notice that $\mathscr{D}\xi_1 = z_1, \mathscr{D}\xi_2 = z_2$, $\mathscr{C}z_1 = \xi_1$ and $\mathscr{C}z_2 = \xi_2$. So,

$$\begin{aligned}\langle \mathscr{D}y, z_i\rangle &= \langle \mathscr{D}[\langle y, \xi_1\rangle\xi_1 + \langle y, \xi_i\rangle\xi_2], z_i\rangle \\ &= \langle [\langle y, \xi_1\rangle z_1 + \langle y, \xi_2\rangle z_2], z_i\rangle \\ &= \langle y, \xi_i\rangle \quad (i = 1 \text{ and } 2) \quad \text{for all } y \in \mathscr{L}_\pi^2,\end{aligned}$$

and

$$\begin{aligned}\langle \mathscr{D}\mathscr{G}_\omega \mathscr{N}\mathscr{C}y(r), z_i\rangle &= \langle \mathscr{G}_\omega \mathscr{N}\mathscr{C}y(r), \xi_i\rangle \\ &= \langle \mathscr{G}_\omega \mathscr{N}\mathscr{C}(rz_1), \xi_i\rangle \\ &= \langle \mathscr{G}_\omega \mathscr{N}(r\xi_1), \xi_i\rangle \quad (i = 1 \text{ and } 2).\end{aligned}$$

Therefore,

$$\begin{aligned}U(\omega, r) &\triangleq (r + \langle \mathscr{D}\mathscr{G}_\omega \mathscr{N}\mathscr{C}y(r), z_1\rangle, \langle \mathscr{D}\mathscr{G}_\omega \mathscr{N}\mathscr{C}y(r), z_2\rangle) \\ &= (r + \langle \mathscr{G}_\omega \mathscr{N}(r\xi_1), \xi_1\rangle, \langle \mathscr{G}_\omega \mathscr{N}(r\xi_1), \xi_2\rangle) \\ &= W(\omega, r)\end{aligned}$$

as in Case B3. Therefore $d[u, 0, \Omega] \neq 0$ by Lemma 4.

(v)

$$\begin{aligned}|U(\omega, r) - V(\omega, r)|_2 &\leqq \frac{\|\mathscr{P}\mathscr{D}\mathscr{G}_\omega\| \cdot \|\mathscr{N}\mathscr{C}\| \cdot \|\mathscr{P}^*\mathscr{D}\mathscr{G}_\omega \mathscr{N}\mathscr{C}(r\xi_1)\|}{1 - \rho_\omega} \quad \text{(by Lemma 2).} \\ &\leqq \frac{\|\mathscr{P}\mathscr{D}\mathscr{G}_\omega\| \cdot \|\mathscr{N}\mathscr{C}\| \cdot \|\mathscr{P}^*\mathscr{D}\mathscr{G}_\omega\| \cdot \|\mathscr{P}^*\mathscr{N}(r\xi_1)\|}{1 - \rho_\omega} \\ &\leqq \frac{|G(j\omega)|\rho_\omega}{1 - \rho_\omega}\|\mathscr{P}^*\mathscr{N}(r\xi_1)\| \\ &= B(\omega)\|\mathscr{P}^*\mathscr{N}(r\xi_1)\|,\end{aligned}$$

so

$$|U(\omega, r) - V(\omega, r)|_2 < B(\omega)\|\mathscr{P}^*\mathscr{N}(r\xi_1)\| \quad \text{for all } (\omega, r) \in \bar{\Omega}.$$

By assumption 2 of Theorem 2, $(\omega, r) \in \partial\Omega$ implies $B(\omega)T(r) < |1/N(r) + G(j\omega)|$, and that implies

$$B(\omega)\|\mathscr{P}^*\mathscr{N}(r\xi_1)\| < |W(\omega, r)|_2 = |U(\omega, r)|_2$$

by the comments at the beginning of the proof. Therefore

$$|U(\omega, r) - V(\omega, r)|_2 < |U(\omega, r)|_2 \quad \text{for all } (\omega, r) \in \partial\Omega.$$

Therefore the hypotheses of Theorem 1 are satisfied, so there exists $(\omega, r) \in \Omega$ such that

$$x(\omega, r) = -\mathscr{D}\mathscr{G}_\omega \mathscr{N}\, \mathscr{C}x(\omega, r), \qquad \mathscr{C}x(\omega, r) = -\mathscr{G}_\omega \mathscr{N}\, \mathscr{C}x(\omega, r).$$

Let $z \triangleq \mathscr{C}x(\omega, r)$. Then $z = -\mathscr{G}_\omega \mathscr{N} z$, which proves part a.
Also

$$\mathscr{P}z = \mathscr{P}\,\mathscr{C}x(\omega, r) = \mathscr{C}\mathscr{P}x(\omega, r) = \mathscr{C}y(r) = \mathscr{C}rz_1 = r\xi_1,$$

which proves part b.

$$\begin{aligned}
\|\mathscr{P}^* z\| &= \|\mathscr{P}^*\,\mathscr{C}x(\omega, r)\| = \|\,\mathscr{C}\mathscr{P}^* x(\omega, r)\| \leqq \|\mathscr{P}^* x(\omega, r)\| \\
&\leqq \frac{\|\mathscr{P}^*\mathscr{D}\mathscr{G}_\omega \mathscr{N}\, \mathscr{C}y(r)\|}{1 - \rho'_\omega} \quad \text{(by Theorem 1)} \\
&\leqq \frac{\|\mathscr{P}^*\mathscr{D}\mathscr{G}_\omega\| \cdot \|\mathscr{P}^*\mathscr{N}\mathscr{C}y(r)\|}{1 - \rho'_\omega} = \frac{\rho'_\omega}{1 - \rho'_\omega} \frac{\sqrt{8}}{(\pi M')} \|\mathscr{P}^*\mathscr{N}(r\xi_1)\| \\
&= \frac{\rho'_\omega \sqrt{8}\, T(r)|rN(r)|}{(1 - \rho'_\omega)\pi M'}.
\end{aligned}$$

Appendix II. This appendix is included for the reader's convenience. It contains a statement of the contraction theorem, a definition of local degree, and two theorems related to local degree which are used in the proof of Theorem 1. The material related to contraction is taken from [5] and material related to local degree is taken from [2].

THEOREM II.1 (fixed points by contraction). *Let $\mathscr{L}$ be a Banach space, let S be a set $\mathbb{R}^n$, and for each $s \in S$, let $T_s : \mathscr{L} \to \mathscr{L}$ be continuous and such that $\|T_s x - T_s y\| \leqq \rho_s \|x - y\|$ for all $x, y \in \mathscr{L}$, where $\rho_s < 1$. Then:*

a. *There exists a unique $x_s^* \in \mathscr{L}$ such that $T_s x_s^* = x_s^*$.*
b. *$\|x_s^* - x\| \leqq \|T_s x - x\|/(1 - \rho_s)$ for all $x \in \mathscr{L}$.*
   In particular $\|x_s^*\| \leqq \|T\mathbf{0}\|/(1 - \rho_s)$, where $\mathbf{0}$ is the zero of $\mathscr{L}$.
c. *If in addition, $s \to s_0$ implies $\|T_s - T_{s_0}\| \to 0$, then $s \to s_0$ implies $\|x_s - x_{s_0}\| \to 0$.*

DEFINITION (local degree for a special case). Let:

1. Ω be an open bounded set in $\mathbb{R}^n$,
2. $f : \bar{\Omega} \to \mathbb{R}^n$ have continuous first partial derivatives,
3. $f(x) \neq p$ for all $x \in \partial\Omega$, the boundary of Ω,
4. $\det J_f(x) \triangleq \det(\partial f_i/\partial x_j) \neq 0$ for all $x \in f^{-1}(p)$.

Then the *degree of f relative to p and Ω* is

$$d[f, p, \Omega] \triangleq \sum_{x \in f^{-1}(p) \cap \Omega} \operatorname{sgn}[\det J_f(x)].$$

Berger and Berger continue the definition to the case where f is continuous but not differentiable. This part of the definition has been omitted as it is not needed here.

THEOREM II.2 (homotopic invariance of d). *If $\Phi : \bar{\Omega} \times [0, 1] \to \mathbb{R}^n$ is continuous and $\Phi(x, \mu) \neq p$ for all $x \in \partial\Omega$ and all $\mu \in [0, 1]$, then $d[\Phi(\cdot, \mu), p, \Omega]$ is the same for all $\mu \in [0, 1]$.*

THEOREM II.3 (existence of preimages). *If* $d[f, p, \Omega] \neq 0$, *then there is an* $x \in \Omega$ *such that* $f(x) = p$.

REFERENCES

[1] R. W. BASS, *Mathematical legitimacy of equivalent linearization by describing functions*, Proc. 1960 International Federation of Automatic Control Congress, Butterworths, London, pp. 2074–2084.

[2] M. BERGER AND M. BERGER, *Perspectives in Nonlinearity*, W. A. Benjamin, New York, 1968.

[3] L. CESARI, *Functional Analysis and Periodic Solutions of Nonlinear Differential Equations*, Contributions to Differential Equations, vol. 1, Interscience, New York, 1962.

[4] A. GELB AND W. E. VANDER VELDE, *Multiple-Input Describing Functions and Nonlinear System Design*, McGraw-Hill, New York, 1968.

[5] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis in Normal Spaces*, Pergamon Press, Oxford, 1964.

[6] R. J. KOCHENBURGER, *A frequency response method for analyzing and synthesizing contractor servomechanisms*, *Part* 1, Trans. AIEE, 69 (1950), pp. 270–284.

[7] J. KUDREWICZ, *Theorems on the existence of periodic vibrations based upon the describing function method*, Proc. International Federation on Automatic Control, Warsaw, Poland, 1969.

SIAM J. CONTROL
Vol. 9, No. 4, November 1971

AN EXISTENCE THEOREM FOR LAGRANGE PROBLEMS WITH UNBOUNDED CONTROLS AND A SLENDER SET OF EXCEPTIONAL POINTS*

L. CESARI,† J. R. LA PALM‡ AND D. A. SANCHEZ§

Abstract. In a number of existence theorems for Lagrange problems with unbounded controls certain growth conditions are required, which generally can be omitted for bounded controls. In the present paper the authors show that even for unbounded controls the growth conditions can be dispensed with in an arbitrary subset in the tx-space E_{1+n}, provided such a subset is "slender" according to suitable definitions. Any set contained in finitely many or countably many smooth curves $x = \phi(t)$, $a \leqq t \leqq b$, is certainly slender. Many examples are given.

1. We deal here with Lagrange problems of control concerning the minimum of an integral of the form

$$I[x, u] = \int_{t_1}^{t_2} f_0(t, x(t), u(t))\, dt \tag{1}$$

in classes Ω of admissible pairs $x(t) = (x^1, \cdots, x^n), u(t) = (u^1, \cdots, u^m), t_1 \leqq t \leqq t_2$, x absolutely continuous (AC), u measurable, satisfying a system of ordinary differential equations

$$dx/dt = f(t, x(t), u(t)), \qquad t \in [t_1, t_2] \quad \text{a.e.}, \tag{2}$$

$f = (f_1, \cdots, f_n)$, constraints of the forms

$$(t, x(t)) \in A \subset E_{n+1}, \qquad t \in [t_1, t_2], \tag{3}$$

$$u(t) \in U(t, x(t)) \subset E_m, \qquad t \in [t_1, t_2] \quad \text{a.e.}, \tag{4}$$

and boundary conditions usually written in the McShane form

$$(t_1, x(t_1), t_2, x(t_2)) \in B \subset E_{2n+1}, \tag{5}$$

where A, B are given fixed sets and $U(t, x)$ is a given set which may depend on t and x.

We do not exclude that A may coincide with the whole space E_{n+1} and that U may coincide with the space E_m. For $m = n$, $f = u$, $U = E_m$ these problems reduce to free problems of the calculus of variations, that is, problems concerning the minimum of a functional

$$I[x] = \int_{t_1}^{t_2} f_0(t, x(t), x'(t))\, dt \tag{6}$$

* Received by the editors February 11, 1970, and in final revised form February 4, 1971.

† Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48104. This author's research was supported in part by AFOSR Research Grant 69-1662, and in part by NONR 233 (76).

‡ Department of Mathematics, Wayne State University, Detroit, Michigan 48202. This author's research was completed in the frame of AFOSR Research Project 942-65.

§ Department of Mathematics, University of California, Los Angeles, California 90024. This author's research was supported in part by the National Science Foundation under Research Grant GP-9658, and in part by NONR 233 (76).

in classes Ω of AC vector-valued functions $x(t) = (x^1, \cdots, x^n)$, $t_1 \leqq t \leqq t_2$, satisfying constraint and boundary conditions

$$(t, x(t)) \in A \subset E_{n+1}, \qquad t \in [t_1, t_2], \tag{7}$$

$$(t_1, x(t_1), t_2, x(t_2)) \in B \subset E_{2n+1}. \tag{8}$$

In previous papers Cesari [1ab] has given existence theorems for optimal solutions with unbounded controls, based on convexity properties of the relevant sets, Helly's selection theorem, recent Filippov's type implicit function theorems, and ensuing lower closure theorems. In such an analysis Ascoli's selection theorem is still used on the trajectories. When the sets $U(t, x)$ are unbounded, or are not uniformly bounded, in order to guarantee the compactness of a minimizing sequence of trajectories, growth conditions concerning f_0 and f are usually requested. Cesari [1b] and Cesari, La Palm and Nishiura [2] proposed the following growth condition (ε): Given $\varepsilon > 0$ there is some locally L-integrable scalar function $\psi_\varepsilon(t) \geqq 0$ such that $|f(t, x, u)| \leqq \psi_\varepsilon(t) + \varepsilon f_0(t, x, u)$ for all t, x, u. As shown by Cesari [1b] and La Palm [3b] this condition is quite general, and contains growth conditions previously proposed, in particular, those of Tonelli [8] and McShane [4ab] for free problems.

Tonelli [8ab], McShane [4abd], and Turner [9] proved that for free problems growth conditions can be dispensed with in an arbitrary set G of points (t, x) in A, for which the only requirement is to be "slender" in the terminology of Turner [9]. For instance, Tonelli [8ab] and McShane [4ab] required G to be either finite, or countable, or contained in a countable set of straight lines parallel to the t-axis, or of curves each of the type $\Gamma: x = x(t), t' \leqq t \leqq t''$, x being AC on $[t', t'']$, and other analogous cases. McShane [4d] and Turner [9] proposed general definitions of "slender" sets, which contain all these cases.

The purpose of this paper is to show that also for general Lagrange problems growth conditions can be dispensed with in arbitrary slender subsets of A. We shall use the same definition of "slender" sets that Turner used for free problems. In § 2 we state an existence theorem for Lagrange problems (1)–(5) with a slender set $G \subset A$ of exceptional sets. In § 3 we prove the theorem, and in § 4 we state and prove a slight variant of the same existence theorem. In § 5 we summarize a number of points concerning the detailed conditions which we are using (upper semicontinuity of sets, property (Q), growth conditions). In § 6 we derive a number of existence theorems for free problems (6)–(8) as corollaries of Cesari's previous existence theorem, the present existence theorem, and the previous analysis concerning growth conditions. In particular, we prove that Tonelli's existence theorems I–V of [8a] can be obtained as corollaries of analogous results for Lagrange problems. In § 7 we list a few examples of Lagrange and free problems satisfying the main hypotheses of the present existence theorem.

2. Statement of the theorem. In the formulation of the Lagrange problem (1)–(5) above we assume that A and B are given subsets of the tx-space E_{n+1} and of the $t_1x_1t_2x_2$-space E_{2n+1} respectively, and that for every $(t, x) \in A$ a given subset $U(t, x)$ of E_m is given (control space). We denote by M the set of all $(t, x, u) \in E_{n+m+1}$ with $(t, x) \in A, u \in U(t, x)$, and we assume that f_0 and $f = (f_1, \cdots, f_n)$ are given real functions on M. A pair $x(t) = (x^1, \cdots, x^n)$, $u(t) = (u^1, \cdots, u^m)$,

$t_1 \leqq t \leqq t_2$, x AC, u measurable, is said to be admissible provided (2), (3), (4), (5) hold, and, in addition, $f_0(t, x(t), u(t))$ is L-integrable in $[t_1, t_2]$.

Let G be a fixed subset of A in the tx-space E_{n+1}, $G \subset A \subset E_{n+1}$. For every set S of the t-axis and $i = 1, \cdots, n$, we shall denote by $G^i(S)$ the set of all real ξ such that for some (t, x) we have $(t, x) \in G$, $t \in S$, $x^i = \xi$. We shall say that $G^i(S)$ is the "image" of S on the x^i-axis by means of the set G. Note that, when G is the graph of a curve $x = g(t)$, $t_1 \leqq t \leqq t_2$, in the tx-space, or $x^i = g_i(t)$, $t_1 \leqq t \leqq t_2$, $i = 1, \cdots, n$, then $G^i(S) = g_i(S)$ is exactly the image of S on the x^i-axis by means of the component g_i of g. A set of $G \subset A$ is said to be "slender" if the following property holds:

(s) For every set S of measure zero on the t-axis, the sets $G^i(S)$ also have measure zero on the x^i-axis, $i = 1, \cdots, n$.

In other words, G is slender provided $|S| = 0$ implies $|G^i(S)| = 0$, $i = 1, \cdots, n$.

As usual we denote by M the subset of all $(t, x, u) \in E_{n+m+1}$ with $(t, x) \in A$, $u \in U(t, x)$ (graph of $U(t, x)$). For every $(\bar t, \bar x) \in A$ and $\delta > 0$ we denote by $N_\delta(\bar t, \bar x)$, or δ-neighborhood of $(\bar t, \bar x)$ in A, the set of all $(t, x) \in A$ at a distance $\leqq \delta$ from $(\bar t, \bar x)$. Also, we denote by $U(\bar t, \bar x; \delta)$ the union of all $U(t, x)$ with $(t, x) \in N_\delta(\bar t, \bar x)$, and then the usual Kuratowski upper semicontinuity property of the sets $U(t, x)$ at $(\bar t, \bar x)$ is expressed by the following requirement, or property (U) at $(\bar t, \bar x)$:

$$U(\bar t, \bar x) = \bigcap_\delta \operatorname{cl} U(\bar t, \bar x; \delta).$$

It is well known that, if A is closed, then $U(t, x)$ has property (U) at every point (t, x) of A if and only if M is closed.

As usual, we denote by $\tilde{\tilde Q}(t, x)$ the set of all $\bar z = (z^0, z) \in E_{n+1}$ with $z^0 \geqq f_0(t, x, u)$, $z = f(t, x, u)$, $u \in U(t, x)$. We then say that the sets $\tilde{\tilde Q}(t, x)$ satisfy property (Q) at a point $(\bar t, \bar x)$ of A provided:

$$\tilde{\tilde Q}(\bar t, \bar x) = \bigcap_\delta \operatorname{cl} \operatorname{co} \tilde{\tilde Q}(\bar t, \bar x; \delta).$$

Cesari showed in [1c] that this condition can be thought of as an extension for Lagrange problems of Tonelli's and McShane's property of seminormality for free problems, and we shall summarize the main concepts in § 4 below. Property (Q) has been used extensively by Cesari, Olech, Lasota, La Palm, Baum, Angell, and others. (For the comparison with a different interpretation of this property in Olech [7ab] see Cesari [1d].)

EXISTENCE THEOREM 2.1 (For Lagrange problems (1)–(5) with unbounded controls and an exceptional slender set). *Let A be compact, and B and M closed. Let $f_0(t, x, u)$, $f(t, x, u) = (f_1, \cdots, f_n)$ be continuous on M, and let us assume that the sets $\tilde{\tilde Q}(t, x)$ are all convex, closed, and satisfy property* (Q) *at all points $(t, x) \in A$ with exception perhaps of a set of points whose t-coordinate lies on a set of measure zero on the t-axis.*

Let us assume that: (Λ) *For every point $(\bar t, \bar x) \in A$ there are a neighborhood $N_\delta(\bar t, \bar x)$, real constants r, $b = (b_1, \cdots, b_n)$, and $\nu > 0$ such that $(t, x) \in N_\delta(\bar t, \bar x)$, $u \in U(t, x)$ implies $f_0(t, x, u) \geqq r + b \cdot f(t, x, u) + \nu |f(t, x, u)|$. Let us assume that a closed set $G \subset A$ is given satisfying property* (s): $|G^i(S)| = 0$, $i = 1, \cdots, n$, *for every set S of measure zero on the t-axis. Let us assume that the following growth condition holds*: (ε) *For every point $(\bar t, \bar x) \in A - G$ there is a neighborhood $N_\delta(t, x)$ such that given $\varepsilon > 0$ we can determine an L-integrable function $\psi_\varepsilon(t) \geqq 0$, $\bar t - \delta \leqq t \leqq \bar t + \delta$,*

such that $(t, x) \in N_\delta(\bar{t}, \bar{x})$, $u \in U(t, x)$ *implies* $|f(t, x, u)| \leqq \psi_\varepsilon(t) + \varepsilon f_0(t, x, u)$. *If* Ω *is any closed nonempty family of admissible pairs, then functional* (1) *has an absolute minimum in* Ω.

In § 6 we shall show that growth condition (ε) is a generalization of the growth condition used by Tonelli and McShane for free problems. In view of the terminology in common use for free problems, condition (Λ) can be called a "condition of normality" for the pair f_0, f. Details concerning these concepts will be given in § 5. Finally, in § 6, we shall show that Tonelli's existence theorems I–V of [8a] can be derived from the existence theorem above for Lagrange problems, from Cesari's previous existence theorems [1abcd], and results in [3ab] and [2]. A growth condition similar to condition (ε) above has been recently proposed also by C. Olech [7a].

3. Proof of Existence Theorem 2.1. We shall denote by $|x|$ the usual absolute value for real numbers x, and $|x| = \max[|x^i|,\ i = 1, \cdots, n]$ for vectors $x = (x^1, \cdots, x^n)$. If X denotes the set of all continuous vector functions $x(t) = (x^1, \cdots, x^n)$, defined in arbitrary intervals $a \leqq t \leqq b$, we shall make X a metric space by means of the ρ-metric defined as follows. For any two elements $x(t) = (x^1, \cdots, x^n)$, $a \leqq t \leqq b$, and $y(t) = (y^1, \cdots, y^n)$, $c \leqq t \leqq d$, let $\rho(x, y)$ be the distance function defined by

$$\rho(x, y) = |a - c| + |b - d| + \max |x(t) - y(t)|,$$

where max is taken for $-\infty < t < \infty$ and x and y are extended by continuity and constancy outside the original intervals (a, b) and (c, d).

From hypothesis (Λ), to each point $(\bar{t}, \bar{x}) \in A$ we may associate an expression $\xi = r + b \cdot f + v|f|$ and a constant $\bar{\delta} > 0$ such that $f_0(t, x, u) \geqq \xi$ for all $(t, x, u) \in M$ with $|t - \bar{t}| \leqq 5\delta$, $|x - \bar{x}| \leqq 5\delta$. A finite number of the cubes $\{(t, \bar{x}) \,|\, |t - \bar{t}| \leqq \bar{\delta}, |x - \bar{x}| \leqq \bar{\delta}\}$ cover A. Let δ be the minimum of the numbers $\bar{\delta}$ associated with the members of such a finite covering. Divide E_{n+1} into cubes S whose sides have length δ by means of the hyperplanes $t = h\delta$, $x^i = l_i\delta$ for $h, l_i = 0, \pm 1, \pm 2, \cdots$, $i = 1, 2, \cdots, n$. Each cube S which intersects A intersects one of the cubes of the covering. Let us associate with this cube S the expression $\xi = r + b \cdot f + v|f|$ which is associated with the cube of the finite covering. In this manner a set of cubes S_{hl} of the form $\{(t, x) | (h - 1)\delta \leqq t \leqq h\delta,\ (l_i - 1)\delta \leqq x^i \leqq l_i\delta,\ i = 1, \cdots, n\}$ are obtained which cover A, and with each cube S_{hl} we have associated an expression

$$\xi_{hl} = r_{hl} + b_{hl} \cdot f(t, x, u) + v_{hl}|f(t, x, u)|, \qquad l = (l_1, \cdots, l_n),$$

with $f_0(t, x, u) \geqq \xi_{hl}$ for all $(t, x) \in A \cap S_{hl}$, $u \in U(t, x)$, and this is true even if we replace S_{hl} by any one of the $3^{n+1} - 1$ adjacent cubes. The t-coordinates of the vertices of these cubes define a partition of a section of the t-axis; thus, there are two integers p, q such that each vertex of the cubes has t-coordinate of the form $s\delta$ with $p \leqq s \leqq q$.

For every $t_0 \in [p\delta, q\delta]$, the set $G^i(\{t_0\})$ has measure zero. Given η, $0 < \eta < \delta$, for every $i = 1, \cdots, n$, the set $G^i(\{t_0\})$ can be covered by an open set of measure $< \eta$. The Cartesian product F_0 of the open sets is open in the hyperplane $H(t_0): t = t_0$. Now the set $(H(t_0) - F_0) \cap A$ is a compact subset of A free of points of the exceptional closed set G. By hypothesis (ε), given $N > 0$ integer and any point

$(t_0, x_0) \in (H(t_0) - F_0) \cap A$, there is a number $\rho_0 > 0$ and an L-integrable function $\psi_{0N}(t)$, $t_0 - \rho_0 \leqq t \leqq t_0 + \rho_0$, such that $|f(t, x, u)| \leqq \psi_{0N}(t) + N^{-1} \cdot f_0(t, x, u)$ when $(t, x) \in A$, $|t - t_0| \leqq \rho_0$, $|x - x_0| \leqq \rho_0$, $u \in U(t, x)$.

The compact set $(H(t_0) - F) \cap A$ may be covered by a finite number of these hypercubes. Let $\bar{\rho} > 0$ be less than δ and less than the minimum of the numbers ρ_0 of the given finite covering. Thus, there is an interval $(t_0 - \bar{\rho}, t_0 + \bar{\rho})$ such that $|f(t, x, u)| \leqq \psi_N(t) + N^{-1} f_0(t, x, u)$ when $(t, x) \in A$, (t, x) is in the union K of these open hypercubes, and $|t - t_0| < \bar{\rho}$. Here $\psi_N(t) = \max\{\psi_{0N}(t)\}$, this maximum being taken on the finitely many functions $\psi_{0N}(t)$ associated with the chosen finite covering. The complement H of K in the section $(t_0 - \bar{\rho}, t_0 + \bar{\rho})$ has the property that each of its projections on the coordinate axes x^i has measure $< \eta$.

In this manner we have associated an open interval of the form $(t - \rho, t + \rho)$ with each $t \in [p\delta, q\delta]$, thereby obtaining an open covering of $[p\delta, q\delta]$. By a suitable contraction these intervals may be used to define a partition $P: p\delta = t_0 < t_1 < \cdots < t_R = q\delta$ of $[p\delta, q\delta]$, and it may be assumed without loss of generality that the points $s\delta$ for $p \leqq s \leqq q$ are used in this partition. Now let us refine the previous partition of A into parts S_{ml} by means of the hyperplanes $t = t_j$, $j = 1, \cdots, R$. The new parts are intervals, say Q_{jl}; we shall still call them cubes, for the sake of simplicity. These parts Q_{jl} are of the form

$$Q_{jl} = \{(t, x) | t_{j-1} \leqq t \leqq t_j, (l_i - 1)\delta \leqq x^i \leqq l_i\delta, i = 1, \cdots, n\},$$
$$j = 1, \cdots, R, \quad t_j - t_{j-1} \leqq \delta, \quad l = (l_1, \cdots, l_n).$$

Let $\xi_{jl} = r_{jl} + b_{jl} \cdot f + v_{jl}|f|$ denote the function associated with the cube of the former partition which contains Q_{jl}.

Summarizing, the following type of partition may be obtained: Given $\eta > 0$ and $N > 0$, there are expressions ξ_{jl} as above, and a partition of A into cubes Q_{jl} as above, whose edges in the x^i direction have length $\delta > 0$ independent of η and N, such that $f_0(t, x, u) \geqq \xi_{jl}$ for all $(t, x) \in Q_{jl} \cap A$, or $(t, x) \in A$ in any of the $3^{n+1} - 1$ cubes adjacent to Q_{jl}, and $|f(t, x, u)| \leqq \psi_N(t) + N^{-1} f_0(t, x, u)$ for all (t, x) with $t_{j-1} \leqq t \leqq t_j$, except for a set H_j (made up of cubes Q_{jl}) whose projections on the x^i-axes have measures $< \eta$, $i = 1, \cdots, n$. Moreover, the constants r_{jl} and b_{jl} are independent of η and N.

Let $r = \max|r_{jl}|$, $b = \max|b_{jl}|$, $v = \min v_{jl}$, and take $\eta < \delta/2$, $N > 2b \cdot (1 + 4\sqrt{n + 1})$.

Let Ω^* denote the class of all "admissible pairs" $x(t), u(t), a \leqq t \leqq b$, defined on an arbitrary interval $[a, b]$, in the sense that we require x to be AC in $[a, b]$, u to be measurable in $[a, b]$, we require $(t, x(t)) \in A$ for all $t \in [a, b]$, $u(t) \in U(t, x(t))$ and $x'(t) = f(t, x(t), u(t))$ a.e. in $[a, b]$, and we require $f_0(t, x(t), u(t))$ to be L-integrable in $[a, b]$. Thus, we do not require any boundary condition for inclusion in Ω^*, and thus $\Omega^* \supset \Omega$. Let $C_j: x = x(t)$, $t_{j-1} \leqq t \leqq t_j$, be the part of the trajectory x (if any) defined on $[t_{j-1}, t_j]$. Divide C_j into more subarcs $C_{j1}, \cdots, C_{jT_j}$ as follows: The first endpoint of C_{1j} is $x(t_{j-1})$ (or $x(a)$ if $t_{j-1} < a < t_j$); the second endpoint is either the first point where C_j leaves one of the $3^n - 1$ cubes in the section $\{(t, x) | (t, x) \in A, t_{j-1} \leqq t \leqq t_j\}$ adjacent to the cube containing $x(t_{j-1})$, or $x(t_j)$ if C_j does not leave these $3^n - 1$ cubes (or $x(b)$ if $t_{j-1} < b < t_j$). The first point of C_{j2} is the endpoint of C_{j1}, and the endpoint of C_{j2} is either the first point of C_j

which leaves the $3^n - 1$ cubes adjacent to the cube containing the first endpoint of C_{j2} or $x(t_j)$ if C_j does not leave these cubes (or $x(b)$ if $t_{j-1} < b < t_j$). Continuing in this manner, C_j is broken up into arcs $C_{j\sigma}$, $\sigma = 1, \cdots, T_j$. This process must terminate after a finite number of steps, since each arc $C_{j\sigma}$ except the last has length $> \delta$.

Let $\Lambda_{j\sigma}$ be the set of all t in the domain of $C_{j\sigma}$, where $x(t) \in H_j$; let $\Lambda'_{j\sigma}$ be the complement of $\Lambda_{j\sigma}$ in this domain. Let $\lambda_{j\sigma} = \int_{\Lambda_{j\sigma}} |x'(t)|\, dt$, and $\lambda'_{j\sigma} = \int_{\Lambda'_{j\sigma}} |x'(t)|\, dt$. Let the initial point of $C_{j\sigma}$ be in Q_{jl_k}. Then

$$I[x,u] \geqq \sum_{j=1}^{R} \sum_{\sigma=1}^{T_j} \left\{ \int_{\Lambda_{j\sigma}} [r_{jl_\sigma} + b_{jl_\sigma} \cdot x'(t) + v_{j\sigma}|x'(t)|]\, dt + \int_{\Lambda'_{j\sigma}} f_0(t, x(t), u(t))\, dt \right\}$$

$$\geqq \sum_{j=1}^{R} \sum_{\sigma=1}^{T_j} \left\{ -r|\Lambda_{j\sigma}| - b \left| \int_{\Lambda_{j\sigma}} x'(t)\, dt \right| + v_{j\sigma}\lambda_{j\sigma} + \int_{\Lambda'_{j\sigma}} f_0(t, x(t), u(t))\, dt \right\}.$$

For $t \in \Lambda'_{j\sigma}$ we have $|f(t, x, u)| \leqq \psi_N(t) + N^{-1} f_0(t, x, u)$ for all $u \in U(t, x(t))$. Then

$$I[x,u] \geqq -r(b-a) + \sum_{j=1}^{R} \sum_{\sigma=1}^{T_j} \left\{ -b \left| \int_{\Lambda_{j\sigma}} x'(t)\, dt \right| + v\lambda_{j\sigma} + \int_{\Lambda'_{j\sigma}} [f_0(t,\ x(t), u(t)) + N\psi_N(t)]\, dt - \int_{\Lambda'_{j\sigma}} N\psi_N(t)\, dt \right\}$$

$$\geqq -r(b-a) - \int_a^b N\psi_N(t)\, dt + \sum_{j=1}^{R} \sum_{\sigma=1}^{T_j} \left\{ -b \left| \int_{\Lambda_{j\sigma}} x'(t)\, dt \right| + v\lambda_{j\sigma} + N\lambda'_{j\sigma} \right\}.$$

Now $|(\int_{\Lambda_{j\sigma}} + \int_{\Lambda'_{j\sigma}}) x'(t)\, dt| \leqq 2\delta\sqrt{n+1}$, and thus

$$\left| \int_{\Lambda_{j\sigma}} x'(t)\, dt \right| - \left| \int_{\Lambda'_{j\sigma}} x'(t)\, dt \right| \leqq 2\delta\sqrt{n+1},$$

$$\left| \int_{\Lambda_{j\sigma}} x'(t)\, dt \right| \leqq 2\delta\sqrt{n+1} + \lambda'_{j\sigma} \qquad \text{for all } j \text{ and } \sigma.$$

Also, $\lambda'_{j\sigma} > \delta - \eta > \delta - \delta/2 = \delta/2$ for $\sigma = 1, \cdots, T_j - 1$. Let

$$\lambda = \sum_{j=1}^{R} \sum_{\sigma=1}^{T_j} \lambda_{j\sigma}, \qquad \lambda' = \sum_{j=1}^{R} \sum_{\sigma=1}^{T_j} \lambda'_{j\sigma},$$

and let D be the diameter of A. Then

$$I[x,u] > -r(b-a) - \int_a^b N\psi_N(t)\,dt + v\lambda + N\lambda' + \sum_{j=1}^{R}\sum_{\sigma=1}^{T_j}\{-b(\lambda'_{j\sigma} + 2\delta\sqrt{n+1})\}$$

$$\geqq -rD - \int_{p\delta}^{q\delta} N\psi_N(t)\,dt + v\lambda + N\lambda' - b\sum_{j=1}^{R}\sum_{\sigma=1}^{T_j-1}(\lambda'_{j\sigma} + 4\sqrt{n+1}\lambda'_{j\sigma})$$

$$-b\sum_{j=1}^{R}(\lambda'_{jT_j} + 2\delta\sqrt{n+1})$$

$$\geqq -rD - \int_{p\delta}^{q\delta} N\psi_N(t)\,dt - 2\delta bR\sqrt{n+1} + v\lambda + [N - b(1 + 4\sqrt{n+1})]\lambda'$$

$$\geqq -rD - \int_{p\delta}^{q\delta} N\psi_N(t)\,dt - 2\delta bR\sqrt{n+1} + v\lambda + 2^{-1}N\lambda'.$$

Let $V = \min(v, N/2) > 0$. Then

$$I[x;u] > -rD - \int_{p\delta}^{q\delta} N\psi_N(t)\,dt - 2\delta bR\sqrt{n+1} + V\int_a^b |x'(t)|\,dt.$$

Thus, given any constant $M_0 > 0$, for any admissible pair x, u with $I[x,u] \leqq M_0$, the trajectory x has uniformly bounded total variation, and then uniformly bounded length. Moreover, this inequality shows that

$$Z = \inf I[x,u] \geqq -rD - \int_{p\delta}^{q\delta} N\psi_N(t)\,dt - 2\delta bR\sqrt{n+1},$$

where inf is taken in the class $\Omega^* \supset \Omega$.

It will now be proved that for the same admissible pairs $x(t), u(t)$, $a \leqq t \leqq b$, with $I[x,u] \leqq M_0$, the trajectories x are equicontinuous.

If they are not, then there is an $\varepsilon > 0$ such that for every positive integer k there is an admissible pair $x_k(t), u_k(t)$, $a_k \leqq t \leqq b_k$, and two points $t_{k1}, t_{k2} \in [a_k, b_k]$, such that $0 < t_{k2} - t_{k1} < k^{-1}$, $I[x_k, u_k] \leqq M_0$, and $|x_k(t_{k2}) - x_k(t_{k1})| > \varepsilon$. Suppose without loss of generality that $t_{k1} \to t_0$, $t_{k2} \to t_0$, $x_k(t_{k1}) \to x_1$, $x_k(t_{k2}) \to x_2$ as $k \to \infty$. Then $|x_2 - x_1| \geqq \varepsilon$. Let L be a bound for the lengths of the trajectories x_k.

The sets $G^i(\{t_0\})$ have measure zero. Hence, they may be covered by open sets F_i of measure $< \eta$, $0 < \eta \leqq \varepsilon/4$. Let F denote the set $F = \{(t_0, x) | x^i \in F_i,\ i = 1, \cdots, n\}$. Then F is open in the hyperplane $H(t_0): t = t_0$. Let $N = (4/\varepsilon)[M_0 + |2Z - 2r\delta - bL - 1|]$, where r, δ and b are the constants defined above. The set $(H(t_0) - F) \cap A$ is compact, and for every point $(t_0, \bar{x}) \in (H(t_0) - F) \cap A$ there is a $\bar{\rho} > 0$ and an L-integrable function $\psi_N(t)$ such that $|f(t,x,u)| \leqq \psi_N(t) + N^{-1}f_0(t,x,u)$ when $|t - t_0| < \bar{\rho}$. A finite number of these intervals $\{(t,x) | |t - t_0| < \bar{\rho}, |x - \bar{x}| < \bar{\rho}\}$ covers $(H(t_0) - F) \cap A$. Let ρ be the minimum $\bar{\rho}$ for such a covering.

Divide the curve $C_k: x = x_k(t)$, $a_k \leqq t \leqq b_k$, into three parts: C_{k1}, C_{k2}, C_{k3} according as $a_k \leqq t \leqq t_{k1}$, $t_{k1} \leqq t \leqq t_{k2}$, $t_{k2} \leqq t \leqq b_k$, respectively. Divide the interval $[t_{k1}, t_{k2}]$ into two subsets, say $E_2 = [t | x(t) \in F]$, $E_1 = [t_{k1}, t_{k2}] - E_2$.

Then, for some k_0 and all $k > k_0$ we have $|t_{k1} - t_0| < \rho$, $|t_{k2} - t_0| < \rho$, and $|x_k(t_{k1}) - x_k(t_{k2})| > \varepsilon/2$. Then for $k > k_0$ we have also

$$\begin{aligned} I[x_k, u_k] &= I_{k1} + I_{k2} + I_{k3} \geqq 2Z + I_{k2} \\ &\geqq 2Z + \left(\int_{E_1} + \int_{E_2} \right) f_0(t, x_k(t), u_k(t))\, dt \\ &\geqq 2Z + \int_{E_1} (-N\psi_N(t) + N|f(t, x_k(t), u_k(t))|)\, dt \\ &\quad + \int_{E_2} [-r - b|x_k'(t)|]\, dt \\ &\geqq 2Z - 2r\delta - bL + N \int_{E_1} |x_k'(t)|\, dt - \int_{t_{1k}}^{t_{2k}} N\psi_N(t)\, dt. \end{aligned}$$

Now

$$\int_{E_2} |x_k'(t)|\, dt < \eta,$$

$$\begin{aligned} \int_{E_1} |x_k'(t)|\, dt &= \int_{t_{1k}}^{t_{2k}} |x_k'(t)|\, dt - \int_{E_2} |x_k'(t)|\, dt \\ &\geqq (\varepsilon/2) - \eta \geqq \varepsilon/4. \end{aligned}$$

On the other hand, for k sufficiently large, the integral of $N\psi_N(t)$ over $[t_{1k}, t_{2k}]$ is certainly <1 since $t_{2k} - t_{1k} \to 0$. Finally,

$$\begin{aligned} I[x_k, u_k] &> 2Z - 2r\delta - bL - 1 + N\varepsilon/4 \\ &= 2Z - 2r\delta - bL - 1 + [M_0 + |2Z - 2r\delta - bL - 1|] \geqq M_0, \end{aligned}$$

a contradiction. Thus, for admissible pairs x, u with $I[x, u] \leqq M$, the trajectories x are equicontinuous.

Let i be the infimum of $I[x, u]$ in the original family Ω of admissible pairs $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$. Thus, for these pairs, the trajectories satisfy the required boundary conditions $(t_1, x(t_1), t_2, x(t_2)) \in B$, together with $(t, x(t)) \in A$ for all $t \in [t_1, t_2]$. Since $\Omega \subset \Omega^*$, we have $i \geqq Z > -\infty$; thus, i is finite. Let $x_k(t)$, $u_k(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, be a minimizing sequence of admissible pairs, all in Ω, with $I[x_k, u_k] < i + k^{-1}$, $k = 1, 2, \cdots$. Then the trajectories x_k are equibounded and equicontinuous. By a suitable extraction we obtain a subsequence, which we still call $[k]$, such that x_k converges in the ρ-metric toward a continuous vector function $x(t)$, $t_1 \leqq t \leqq t_2$, and $(t, x(t)) \in A$ for all $t_1 \leqq t \leqq t_2$, and $(t_1, x(t_1), t_2, x(t_2)) \in B$. Also, the curve $C: x = x(t), t_1 \leqq t \leqq t_2$, has finite length $l(C)$, since $l(C) \leqq \liminf l(C_k) \leqq L$ as $k \to \infty$.

Let us prove that x is AC. Suppose that this is not the case. Let s, $0 \leqq s \leqq l = l(C)$, denote the usual arc length parameter along C, thought of as a path curve in the tx-space E_{n+1}. Note that given any measurable set $E \subset [a, b]$, the usual Lebesgue measure $|E|$ of E is the infimum of the numbers $\sum_i (\beta_i - \alpha_i)$ for any countable covering (α_i, β_i), $i = 1, 2, \cdots$, of E. Analogously, we can define another measure, or length measure $l(E)$ of E, by taking the infimum of the numbers $\sum_i (s(\beta_i) - s(\alpha_i))$ for all the same open coverings of E. Obviously, $|E| \leqq l(E)$. If x

is not AC then there is some set E of Lebesgue measure zero on $[a, b]$ which has positive length measure, or $|E| = 0$, $l(E) = \lambda > 0$. Now the n sets $G^i(E)$, $i = 1, \cdots, n$, have all zero Lebesgue measure. Thus, if $P = \{(t, x) | t \in E,\ x = x(t)\}$, then $P \cap G$ has projection of zero Lebesgue measure on each coordinate axis. As a consequence, there is some closed set $E' \subset E$ such that $|E'| = 0$, $l(E') > \lambda/2$, and $(t, x(t)) \notin G$ for $t \in E'$.

Let $N = (2/\lambda)(|i| + 2 + rD + bL)$, where L is a bound on the lengths of the trajectories x in Ω. Then, there is some $\rho > 0$ and an L-integrable function $\psi_N(t) \geqq 0$ such that $|f(t, x, u)| \leqq \psi_N(t) + N^{-1} f_0(t, x, u)$ when (t, x) has a distance $< \rho$ from the compact set $P' = \{(t, x) | t \in E', x = x(t)\}$. Since E' is compact, $|E'| = 0$, it may be covered by a finite set of open intervals (α_j, β_j), $j = 1, \cdots, R$, such that if $F = \bigcup_{j=1}^{R} (\alpha_j, \beta_j)$, we have $\int_F N\psi_N(t)\, dt < 1$, and x maps F into the ρ-neighborhood of P'. Let k_0 be such that $\int_F |x_k'(t)|\, dt > \lambda/2$ for all $k \geqq k_0$. Finally,

$$
\begin{aligned}
I[x_k, u_k] &= \int_{t_{1k}}^{t_{2k}} f_0(t, x_k(t), u_k(t))\, dt \\
&\geqq -rD - bL + \int_F f_0(t, x_k(t), u_k(t))\, dt \\
&\geqq -rD - bL + \int_F (-N\psi_N(t) + N|x_k'(t)|)\, dt \\
&\geqq -rD - bL - 1 + N\lambda/2 \\
&= -rD - bL - 1 + (|i| + 2 + rD + bL) = |i| + 1 \geqq i + 1,
\end{aligned}
$$

a contradiction. We have proved that x is AC. By the lower closure theorem (see, for instance, Cesari [1d]) we know that there is a measureable function $u(t)$, $t_1 \leqq t \leqq t_2$, such that the pair $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, is admissible, and

$$
I[x, u] \leqq \liminf I[x_k, u_k] = i.
$$

Here (x, u) belongs to Ω since Ω is closed; thus $I[x, u] \geqq i$, and hence $I[x, u] = i$. This proves Existence Theorem 2.1.

Remark 1. In the proof of Existence Theorem 2.1, we actually prove that there is a positive constant m_0 and a real number m_1 such that $I[x, u] \geqq m_0 BV(x) + m_1$, where $BV(x)$ is the total variation of x. Hence, the trajectories of a minimizing sequence for I in Ω have uniformly bounded variation. The proof of this inequality is similar to the one given by Turner [9] for free problems in E_n. Also, McShane [4b] noticed an analogous inequality for free problems for $n = 1$.

Remark 2. Condition (Λ) in Existence Theorem 2.1 can be replaced by the following one: $(\bar{\Lambda})$ For every point $(\bar{t}, \bar{x}) \in A$ there are a neighborhood $N_\delta(\bar{t}, \bar{x})$, real constants $r, b = (b_1, \cdots, b_n)$, and $v > 0$, and a point $\bar{z} \in E_n$ such that $(t, x) \in N_\delta(\bar{t}, \bar{x})$, $u \in U(t, x)$ implies $f_0(t, x, u) \geqq r + b \cdot f(t, x, u) + v|f(t, x, u) - \bar{z}|$. The proof is a simple modification of the proof given above. In particular, we can take $\bar{z} = f(\bar{t}, \bar{x}, \bar{u})$ for some $\bar{u} \in U(\bar{t}, \bar{x})$.

4. A variant of Existence Theorem 2.1.

EXISTENCE THEOREM 4.1. *This is the same as Theorem* 2.1 *with condition* (Λ) *satisfied only at the points* $(\bar{t}, \bar{x})$ *of the closed set* G.

Proof. Let us perform the initial subdivision in cubes Q_{hl} as at the beginning of § 3 by taking $\delta > 0$ so small that the following holds:

(a) If $Q_{hl} \cap G \neq \varnothing$, then Q_{hl} and the $3^{n+1} - 1$ adjacent cubes Q form a set F with $F \cap A \subset N_\delta$, where N_δ denotes one of the neighborhoods of property (Λ). (b) If Q_{hl} is any one of the cubes Q with $Q_{hl} \cap A \neq \varnothing$ and not covered by any of the sets F above, then Q_{hl} and the $3^{n+1} - 1$ adjacent cubes Q form a set F with $F \cap A \subset N_\delta$, where N_δ denotes one of the neighborhoods of property (ε).

We may denote by $Q_1, \cdots, Q_\nu$, the cubes Q_{hl} of the first kind, and by $Q'_1, \cdots, Q'_\mu$, the cubes Q_{hl} of the second kind. Finally, for every curve $C: x = x(t)$, $a \leqq t \leqq b$, with $(t, x(t)) \in A$ there is a subdivision into finitely many arcs $C_1, \cdots, C_\nu$, $C'_1, \cdots, C'_\mu$, each $C_1, \cdots, C_\nu$ contained in a cube of the first kind, and each $C'_1, \cdots, C'_\mu$ in a cube of the second kind. Each of these arcs has diameter $> \delta$. We shall now apply to the arcs $C_1, \cdots, C_\nu$ the arguments of § 3, and to the arcs $C'_1, \cdots, C'_\mu$, the usual arguments of [1a, pp. 391–395], and [1b, Remark 10, p. 539]. These arguments show that for $I[x, u] \leqq M_0$ the curves $C: x = x(t)$, $t_1 \leqq t \leqq t_2$, have total length $l(C) \leqq L$ for some fixed L, and this yields bounds for the numbers ν and μ. The concluding argument of the proof is then the same as in § 3.

Remark. In Existence Theorem 4.1, condition (Λ) at the points of G can be replaced by a condition more geometric in character which we shall state in § 5 after Theorem 5.5.

5. Remarks. (a) Given any two functions $g(t, x, u)$, $f_0(t, x, u)$, f_0 scalar, defined and continuous on the same set M as in § 2, we say that $g(t, x, u)$ is "of slower growth than $f_0(t, x, u)$ with respect to u" in a neighborhood $N_\delta(\bar{t}, \bar{x})$ of $(\bar{t}, \bar{x}) \in A$, provided, given $\varepsilon > 0$, there is some $\bar{u} = \bar{u}(\delta, \varepsilon) \geqq 0$ such that $(t, x, u) \in M$, $(t, x) \in N_\delta(t, x)$, $|u| \geqq \bar{u}$, implies $|g(t, x, u)| \leqq \varepsilon f_0(t, x, u)$. We proved in [1b, (2.1)(i), (2.2) (ii)] and in [1c, (2.i)] the following criterion for property (Q).

CRITERION 5.1. *If* 1 *and* $f(t, x, u)$ *are of slower growth than* $f_0(t, x, u)$ *with respect to* u *in a neighborhood* $N_\delta(\bar{t}, \bar{x})$ *of* $(\bar{t}, \bar{x})$ *in* A, *then the sets* $\tilde{\tilde{Q}}(t, x)$, *if convex, satisfy property* (Q) *at* $(\bar{t}, \bar{x})$.

(b) Let $f(t, x, u) = (f_1, \cdots, f_n)$ and $f_0(t, x, u)$ be continuous and defined on M as in § 2. As usual, we denote by $Q(t, x) \subset E_n$ the sets defined by $Q(t, x) = f(t, x, U(t, x)) = [z | z = f(t, x, u),\ u \in U(t, x)] \subset E_n$. As usual we denote by R the linear manifold of E_n of minimum dimension r, $0 \leqq r \leqq n$, containing $Q(t, x)$, and we denote by $R\text{int}\, Q(t, x)$ the set of all points $z \in Q(t, x)$ which are interior to $Q(t, x)$ with respect to R. We say that $f(t, x, u) = (f_1, \cdots, f_n)$ and $f_n(t, x, u)$ satisfy property (α) at a point $(\bar{t}, \bar{x}) \in A$ provided:

$$(\alpha) \quad \text{If } (z^0, z) \in \bigcap_\delta \text{cl co } \tilde{\tilde{Q}}(t, \bar{x}; \delta), \quad \text{then } z \in Q(\bar{t}, \bar{x}).$$

As we noticed in [1c] this condition is necessary for the sets $\tilde{\tilde{Q}}(t, x)$ to satisfy property (Q) at $(\bar{t}, \bar{x})$.

For free problems (that is, when $m = n$, $f = u$, $U = E_n$; see § 5 below), we have $Q(t, x) = E_n$ for all $(t, x) \in A$, and then condition (α) is trivially satisfied. We

consider again continuous functions $f(t, x, u) = (f_1, \cdots, f_n)$ and $f_0(t, x, u)$ scalar as in § 2. We say that f and f_0 satisfy property (X) at the point $(\bar{t}, \bar{x}) \in A$ provided, for every $\bar{z} \in Q(\bar{t}, \bar{x})$ there is at least one point $\bar{u} \in U(\bar{t}, \bar{x})$ such that (i) $\bar{z} = f(\bar{t}, \bar{x}, \bar{u})$, and (ii) given $\varepsilon > 0$ there are numbers $\delta > 0$ and $r, b = (b_1, \cdots, b_n)$ such that

$$(X') \quad f_0(t, x, u) \geqq r + b \cdot f(t, x, u) \quad \text{for all } (t, x) \in N_\delta(\bar{t}, \bar{x}) \quad \text{and } u \in U(t, x);$$

$$(X'') \quad f_0(\bar{t}, \bar{x}, \bar{u}) \leqq r + b \cdot f(\bar{t}, \bar{x}, \bar{u}) + \varepsilon.$$

We say that f and f_0 satisfy property (X*) at the point $(\bar{t}, \bar{x}) \in A$ provided, for every $\bar{z} \in Q(\bar{t}, \bar{x})$, there is at least one point $\bar{u} \in U(\bar{t}, \bar{x})$ such that (i) $\bar{z} = f(\bar{t}, \bar{x}, \bar{u})$, and (ii) given $\varepsilon > 0$ there are numbers $\delta > 0$, $\nu > 0$, and $r, b = (b_1, \cdots, b_n)$ real such that

$$(X'^*) \quad f_0(t, x, u) \geqq r + b \cdot f(t, x, u) + \nu |f(t, x, u) - f(\bar{t}, \bar{x}, \bar{u})|$$
$$\text{for all } (t, x) \in N_\delta(\bar{t}, \bar{x}) \quad \text{and } u \in U(t, x);$$

$$(X'') \quad f_0(\bar{t}, \bar{x}, \bar{u}) \leqq r + b \cdot \bar{f}(\bar{t}, \bar{x}, \bar{u}) + \varepsilon.$$

Here, by $b \cdot f$ we mean as usual $b_1 f_1 + \cdots + b_n f_n$.

For free problems, that is, $m = n$, $f = u$, $U = E_n$, condition (X) reduces to the following condition which concerns only the function f_0, or condition (X_f): For every $\bar{u} = (\bar{u}^1, \cdots, \bar{u}^n) \in E_n$ and $\varepsilon > 0$ there are numbers $\delta > 0$ and $r, b = (b_1, \cdots, b_n)$ real such that

$$(X'_f) \quad f_0(t, x, u) \geqq r + b \cdot u \quad \text{for all } (t, x) \in N_\delta(\bar{t}, \bar{x}) \quad \text{and all } u \in E_n,$$

$$(X''_f) \quad f_0(\bar{t}, \bar{x}, \bar{u}) \leqq r + b \cdot \bar{u} + \varepsilon.$$

When this condition is satisfied, we say that $f_0(t, x, u)$, $(t, x) \in A \times E_n$, is *seminormal* at $(\bar{t}, \bar{x}) \in A$. This is an equivalent formulation of definitions due to Tonelli [8ab] and McShane [4]. Again, for free problems, condition (X*) reduces to the following condition concerning f_0, or condition (SN): For every $\bar{u} = (\bar{u}^1, \cdots, \bar{u}^n) \in E_n$ and $\varepsilon > 0$, there are numbers $\delta > 0$, $\nu > 0$, and r, $b = (b_1, \cdots, b_n)$ real such that

$$(SN') \quad f_0(t, x, u) \geqq r + b \cdot u + \nu |u - \bar{u}| \quad \text{for all } (t, x) \in N_\delta(\bar{t}, \bar{x}) \quad \text{and } u \in E_n;$$

$$(SN'') \quad f_0(\bar{t}, \bar{x}, \bar{u}) \leqq r + b \cdot \bar{u} + \varepsilon.$$

When this condition is satisfied, we say that $f_0(t, x, u)$, $(t, x) \in A \times E_n$, is *normal* at $(\bar{t}, \bar{x}) \in A$. This is an equivalent formulation of definitions due to Tonelli [8ab] and McShane [4b].

We see, therefore, that conditions (X) and (X*) above appear as natural extensions to Lagrange problems of the usual conditions of seminormality and normality respectively for free problems.

We proved in [1c, (4.i)] the following second criterion for property (Q) of the sets $\tilde{\tilde{Q}}(t, x)$.

CRITERION 5.2. *If conditions* (α) *and* (X) *hold at the point* $(\bar{t}, \bar{x}) \in$ A, *then the set* $\tilde{\tilde{Q}}(\bar{t}, \bar{x})$ *is closed and convex, and the sets* $\tilde{\tilde{Q}}(t, x)$ *satisfy property* (Q) *at* $(\bar{t}, \bar{x})$.

(c) The case in which f is linear in u, or $f(t, x, u) = B(t, x)u + C(t, x)$ (B, C matrices of dimensions $n \times m$, $n \times 1$ with continuous entries), is of interest. The

free problems are of this kind, with $m = n, f = u$, B the identity matrix of order n, and $C = 0$. We proved in [1c, (7.i)] the following criterion for property (Q).

CRITERION 5.3. *If A is closed, $U = E_m$, $M = A \times E_m$, f_0 is continuous on M, convex in u, and seminormal in u, at a point $(\bar{t}, \bar{x}) \in A$, and if $f(t, x, u) = B(t, x)u + C(t, x)$, then the sets $\tilde{\tilde{Q}}(t, x)$ satisfy property* (Q) *at* $(\bar{t}, \bar{x})$.

(d) Given $f(t, x, u) = (f_1, \cdots, f_n)$ and $f_0(t, x, u)$ continuous on M as usual, and for any point $z \in Q(t, x) = f(t, x, U(t, x)) \subset E_n$, let us define $T(z; t, x)$ by taking

$$T(z; t, x) = \inf [z^0 | (z^0, z) \in \tilde{\tilde{Q}}(t, x)]$$

$$= \inf [z^0 | z^0 \geqq f_0(t, x, u), z = f(t, x, u) \text{ for some } u \in U(t, x)].$$

Thus, for every $(\bar{t}, \bar{x}) \in A$, the scalar function $T(z; \bar{t}, \bar{x})$, $z \in Q(\bar{t}, \bar{x})$, is defined, with $-\infty \leqq T(z; \bar{t}, \bar{x}) < +\infty$. It was proved in [1c, (8.i)] that, if $\tilde{\tilde{Q}}(t, x)$ is convex, then either $T(z; \bar{t}, \bar{x}) = -\infty$ for all $z \in R\text{int } Q(\bar{t}, \bar{x})$, or $T(z; \bar{t}, \bar{x}) > -\infty$ for all $z \in Q(\bar{t}, \bar{x})$. In the latter case, $T(z; \bar{t}, \bar{x})$ is finite everywhere and a convex function of z in $Q(\bar{t}, \bar{x})$, $T(z; \bar{t}, \bar{x})$ is bounded below on every bounded subset of $Q(\bar{t}, \bar{x})$, and $T(z; \bar{t}, \bar{x})$ is continuous in the convex set $R\text{int } Q(\bar{t}, \bar{x})$. Finally, if $\tilde{\tilde{Q}}(\bar{t}, \bar{x})$ is convex and closed, and $T(z; \bar{t}, \bar{x}) > -\infty$ in $Q(\bar{t}, \bar{x})$, then $T(z; \bar{t}, \bar{x})$ is a lower semicontinuous function of z at every point $z \in Q(\bar{t}, \bar{x}) - R\text{int } Q(\bar{t}, \bar{x})$. We proved in [1c, (9.i)] and [1e, (4.vii)] the following statements.

THEOREM 5.4. *If $T(z; \bar{t}, \bar{x}) > -\infty$ in $Q(\bar{t}, \bar{x})$, then the sets $\tilde{\tilde{Q}}(t, x)$ have property* (Q) *at* $(\bar{t}, \bar{x})$ *if and only if properties* (α) *and* (X) *hold at the point* $(\bar{t}, \bar{x})$.

THEOREM 5.5. *If $T(z; \bar{t}, \bar{x}) > -\infty$ in $Q(\bar{t}, \bar{x})$ and if the sets $\tilde{\tilde{Q}}(t, x$ contain no straight line, then the sets $\tilde{\tilde{Q}}(t, x)$ have property* (Q) *at* $(\bar{t}, \bar{x})$ *if and only if properties* (α) *and* (X*) *hold.*

As a consequence of Theorem 5.5 we can now replace condition (Λ) in Existence Theorem 4.1 by the following condition more geometric in character: (Λ') At every point $(\bar{t}, \bar{x}) \in G$ the sets $\tilde{\tilde{Q}}(\bar{t}, \bar{x})$ satisfy property (Q), $T(z; \bar{t}, \bar{x}) > -\infty$, and $\tilde{\tilde{Q}}(\bar{t}, \bar{x})$ contains no straight line.

Indeed, condition (Λ') implies that both conditions (α) and (X^*) hold at the same points, and the latter implies (Λ) for $\bar{z} = 0$.

Theorem 5.5 reduces for free problems ($m = n, f = u, U = E_n$) to the following well-known statement:

THEOREM 5.6. *If $f_0(t, x, u)$ is continuous in $A \times E_n$, then f_0 is normal in u at $(\bar{t}, \bar{x})$ if and only if $f_0(\bar{t}, \bar{x}, u)$ is convex in u, and for no $\bar{u}$, $u_1 \in E_n$, $u_1 \neq 0$, it occurs that $f_0(\bar{t}, \bar{x}, \bar{u}) = 2^{-1}[f_0(\bar{t}, \bar{x}, \bar{u} + \lambda u_1) + f_0(\bar{t}, \bar{x}, \bar{u} - \lambda u_1)]$ for all $\lambda \geqq 0$.*

This last statement was proved by Tonelli [8a] under smoothness hypotheses, and then extended by Turner [9] under the sole continuity hypotheses above.

6. Free problems. We consider now free problems, that is, Lagrange problems with $m = n, f = u, U = E_n$, as stated in § 1. Thus, we are interested in problems concerning the minimum of an integral

$$I[x] = \int_{t_1}^{t_2} f_0(t, x(t), x'(t))\, dt \tag{6}$$

in classes Ω of admissible AC vector-valued functions $x(t) = (x^1, \cdots, x^n)$,

$t_1 \leqq t \leqq t_2$, satisfying constraint and boundary conditions

$$(t, x(t)) \in A \subset E_{n+1}, \qquad t \in [t_1, t_2], \tag{7}$$

$$(t_1, x(t_1), t_2, x(t_2)) \in B \subset E_{2n+1}. \tag{8}$$

The only requirement for admissibility (§ 3), which has not yet been restated, is now that $f_0(t, x(t), x'(t))$ be L-integrable in $[t_1, t_2]$. We shall show in this section that existence theorems I to V of the memoir [8a] of Tonelli for free problems can be derived as particular cases from the existence theorems above for Lagrange problems, from the previous Cesari existence theorems [1abcd], and results in [3ab] and [2].

THEOREM 6.1. *Let A be compact, B closed, $f_0(t, x, x')$ continuous in $A \times E_n$ and convex in u for every $(t, x) \in A$, and let us assume that*: (ϕ) *There exists a continuous real-valued function $\phi(\xi)$, $0 \leqq \xi < +\infty$, with $f_0(t, x, x') \geqq \phi(|x'|)$, $\phi(\xi)/\xi \to +\infty$ as $\xi \to +\infty$. Then, integral* (6) *has an absolute minimum in every closed nonempty class of admissible AC curves $x = x(t)$, $t_1 \leqq t \leqq t_2$, satisfying* (7) *and* (8).

Indeed, here $f = u$, $m = n$, $U = E_n$, and by force of (ϕ) both 1 and f are of slower growth than f_0 in A, so that the convex sets $\tilde{\tilde{Q}}(t, x) = [(z^0, u) | z^0 \geqq f_0(t, x, u), u \in E_n]$ satisfy property (Q) at every $(\bar{t}, \bar{x}) \in A$ by force of Criterion 5.1. Theorem 6.1 is now a corollary of Cesari's Existence Theorem 1 for Lagrange problems in [1a, p. 390]. Of course, we could use Theorem 2.1 instead. Indeed, by force of (ϕ), first $\phi(\xi) \geqq c + d\xi$ for suitable constants c real and $d > 0$, and then condition (Λ) is trivially satisfied with $r = c$, $b = 0$, $v = d$. On the other hand, condition (ϕ) certainly implies condition (ε). Thus, Existence Theorem 2.1 applies with $G = \phi$. Note that Theorem 6.1 is essentially Tonelli's Theorem I of [8a, p. 208] for any $n \geqq 1$ instead of $n = 1$ only.

THEOREM 6.2. *This is the same as Theorem* 6.1 *with* (ϕ) *replaced by*

$$(l) \quad \lim_{|u| \to \infty} f_0(t, x, u)/|u| = +\infty \quad \textit{for every } (t, x) \in A.$$

Indeed, for f_0 convex, conditions (ϕ) and (l) are equivalent (see [8a]). Note that Criterion 5.2 is essentially Tonelli's Theorem II of [8a, p. 211] for any $n \geqq 1$ instead of $n = 1$ only.

THEOREM 6.3. *This is the same as Theorem* 6.1 *with* (ϕ) *replaced by*: (ε_0) *Given $\varepsilon > 0$ there is a locally integrable function $\psi_\varepsilon(t)$ which may depend on ε, such that $|u| \leqq \psi_\varepsilon(t) + \varepsilon f_0(t, x, u)$ for all $(t, x, u) \in A \times E_n$.*

Indeed, condition (ε_0) implies that, for all $(\bar{t}, \bar{x}) \in A$, with exception perhaps of a set of points whose t-coordinate lies in a set of measure zero on the t-axis, the figurative $z = f_0(\bar{t}, \bar{x}, u)$, $u \in E_n$, does not contain any straight line, and hence, because of Theorem 5.6, f_0 is normal in u at $(\bar{t}, \bar{x})$. This in turn implies, because of Criterion 5.3, that the sets $\tilde{\tilde{Q}}(t, x) = [(z^0, u) | z^0 \geqq f_0(t, x, u), u \in E_n)$ satisfy condition (Q) at the same points $(\bar{t}, \bar{x}) \in A$. We are now in a position to apply Cesari's existence theorem for Lagrange problems [1b, Cor. 2, pp. 545–546, and Remark 10, p. 539]. Theorem 6.3 is thereby proved.

Concerning Tonelli's Theorem III of [8a, p. 213], let us assume that there is a closed subset E of A such that f_0 satisfies condition (l) at every point of A—E, and the condition (β) stated by Tonelli in [8a, p. 213] and restated in [3b, p. 380]

under the name of condition (T). As proved by J. R. La Palm in [3b, pp. 380–382], f_0 then satisfies condition (ε_0) in A. and Theorem 6.3 applies.

THEOREM 6.4. *Let A be compact, B closed, f_0 continuous in $A \times E_n$, convex in u for every $(t, x) \in A$, and normal. Let G be a closed subset of A contained in a countable family of AC curves $\Gamma: x = x(t)$, $t' \leqq t \leqq t''$, or in a collection of straight lines $x = \bar{x} = (\bar{x}^1, \cdots, \bar{x}^n)$ such that each set $\{\bar{x}^i\}$, $i = 1, \cdots, n$, of real numbers is of measure zero on the real line. Let us assume that at every point $(t, x) \in A - G$ we have: $(l) f_0(t, x, u)/|u| \to +\infty$ as $|u| \to +\infty$. Then, the integral (6) has an absolute minimum in every closed nonempty class Ω of AC curves $x = x(t)$, $t_1 \leqq t \leqq t_2$, satisfying (7) and (8).*

Since f_0 is normal, f_0 satisfies conditions (SN′), (SN″) of § 5, part (b). In particular, for $\bar{u} = 0$ we obtain $f_0(t, x, u) \geqq r + b \cdot u + \nu|u|$ for all $(t, x) \in N_\delta(\bar{t}, \bar{x})$, and all $u \in E_n$. In other words, f_0 satisfies property (Λ) of Theorem 2.1 with $f = u$, $m = n$, $U = E_n$. Furthermore, again by force of the normality, and because of Criterion 5.3, the sets

$$\tilde{\tilde{Q}}(t, x) = [(z^0, u) | z^0 \geqq f_0(t, x, u),\ u \in E_n]$$

satisfy property (Q) for every $(t, x) \in A$. Finally, for every $(\bar{t}, \bar{x}) \in A - G$, there is a neighborhood $N_\delta(\bar{t}, \bar{x})$ at all points of which condition (l) holds; hence, by Tonelli's argument, f_0 satisfies a condition (ϕ) in $N_\delta(\bar{t}, \bar{x})$, and, a fortiori, a condition (ε) as in Theorem 2.1 with $f = u$, $m = n$, $U = E_n$. Theorem 2.1 applies and Theorem 6.4 is proved. Note that Theorem 6.4 is essentially Tonelli's Theorem IV with $n \geqq 1$ instead of $n = 1$.

THEOREM 6.5. *Let A be compact, B closed, f_0 continuous in $A \times E_n$, and convex in u for every $(t, x) \in A$. Let G be any subset of A as described in Theorem* 6.4. *Let us assume that* (i) *at every point $(\bar{t}, \bar{x}) \in G$ the figurative $Q(\bar{t}, \bar{x}) = [(z^0, u), z^0 = f_0(\bar{t}, \bar{x}, u),\ u \in E_n]$ contains no straight line*; (ii) *for every $(\bar{t}, \bar{x}) \in A - G$ there is a neighborhood $N_\delta(\bar{t}, \bar{x})$ of $(\bar{t}, \bar{x})$ in A and, for every $\varepsilon > 0$, a locally integrable nonnegative function $\psi_\varepsilon(t)$ such that $|u| \leqq \psi_\varepsilon(t) + \varepsilon f_0(t, x, u)$ for all $(t, x) \in N_\delta(\bar{t}, \bar{x})$ and $u \in E_n$. Then, the same conclusion holds as in Theorem* 6.4.

For any point $(\bar{t}, \bar{x}) \in G$ the figurative $Q(\bar{t}, \bar{x})$ contains no straight line, and then the same holds at every point (t, x) of some neighborhood $N_\delta(\bar{t}, \bar{x})$ of $(\bar{t}, \bar{x})$ in A. Thus, f_0 is normal in $N_\delta(\bar{t}, \bar{x})$, and consequently satisfies property (Λ) in $N_\delta(\bar{t}, \bar{x})$. Also, the sets $\tilde{\tilde{Q}}(t, x) = [(z^0, u) | z^0 \geqq f_0(t, x, u),\ u \in E_n]$ satisfy property (Q) in all of $N_\delta(\bar{t}, \bar{x})$ by force of Criterion 5.3. Note that for every $(\bar{t}, \bar{x}) \in A - G$ the growth condition described in hypothesis (ii) holds in a neighborhood $N_\delta(\bar{t}, \bar{x})$ of $(\bar{t}, \bar{x})$ in A. As a consequence, property (l) holds at all $(t, x) \in N_\delta(\bar{t}, \bar{x})$ with exception perhaps of a set of points whose t-coordinate lies in a set of measure zero on the t-axis. Hence, the figurative $Q(t, x)$ contains no straight line and f_0 is normal at all $(t, x) \in N_\delta(\bar{t}, \bar{x})$ with perhaps the same exception as above.

By force of Criterion 5.3, property (Q) holds at all points of A–G with the same possible exception. Thus, property (Q) holds in all of A, with the same possible exception, as required in Theorem 2.1. Conditions (s) and (ε) of Theorem 2.1 hold, and condition (Λ) of Theorem 4.1 also holds. Thus, Theorem 4.1 applies and Theorem 6.5 is proved.

Concerning Tonelli's Theorem V of [8a, p. 225], let us assume that G is given as in Theorem 6.5 and that hypothesis (i) is satisfied. Also, let us assume that there

is a closed set E of A such that f_0 satisfies condition (l) at every point of A—G—E, and that condition (β) stated by Tonelli in [8a, p. 213] holds. This condition was restated in [3b, p. 380] under the name of Tonelli's condition (T). As proved by J. R. La Palm in [3b, pp. 380–382], f_0 then satisfies hypothesis (ii) in A—G, and Theorem 6.5 applies. The reader may have noticed that the terminology used in the present paper is slightly different from the one used in [1c]. This change was made in order to be closer to Tonelli's and McShane's terminology.

7. Examples. (a) For free problems the functions

1. $f_0 = |x|u^2 + (1 + u^2)^{1/2}$, $m = n = 1$, $U = E_1$, $f = u$;
2. $f_0 = |x|\,|u|^{1+\alpha} + (1 + u^2)^{1/2}$, $\alpha > 0$, $m = n = 1$, $U = E_1$, $f = u$;
3. $f_0 = (x^2 + y^2)(u^2 + v^2) + (1 + u^2 + v^2)^{1/2}$, $m = n = 2$, $U = E_2$, $f_1 = u$, $f_2 = v$;
4. $f_0 = (x - t)^2u^2 + (1 + u^2)^{1/2}$, $n = m = 1$, $U = E_1$, $f = u$;
5. $f_0 = (x \sin 1/x)^2u^2 + (1 + u^2)^{1/2}$, $m = n = 1$, $U = E_1$, $f = u$;
6. $f_0 = ((x^2 - t^2)^2 + y^2)(u^2 + v^2) + (1 + u^2 + v^2)^{1/2}$, $m = n = 2$, $U = E_2$, $f_1 = u$, $f_2 = v$;
7. $f_0 = |x|u^2 + (1 + u^2)^{1/2} - 2u$, $m = n = 1$, $U = E_1$, $f = u$;
8. $f_0 = (x^2 + y^2)(u^2 + v^2) + (1 + u^2 + v^2)^{1/2} - 3u - 5v - 1$, $m = n = 2$, $U = E_2$, $f_1 = u$, $f_2 = v$,

all satisfy the condition of the Existence Theorem 2.1. The exceptional set is respectively, the single straight line $x = 0$ in the first, second, and seventh examples, the single straight line $x = y = 0$ in the third and eighth, the straight line $x = t$ in the fourth, the countable family of straight lines $x = 0$ and $x = \pm(k\pi)^{-1}$, $k = 1, 2, \cdots$, in the fifth, and the two straight lines, $x = \pm t$, $y = 0$ in the sixth. In the seventh and eighth examples condition (Λ) is satisfied by taking respectively $b = 2$, and $b = (3, 5)$ in the whole of A.

(b) For general Lagrange problems the following functions:

1. $f_0 = |x|u^4 + (1 + u^4)^{1/2}$, $f = u^2$, $n = 1$, $m = 1$, $U = E_1$;
2. $f_0 = |x|u^4 + (1 + u^4)^{1/2} - 3u^2$, $f = u^2$, $n = 1$, $m = 1$, $U = E_1$;
3. $f_0 = x^3u^2$, $f = x^2u$, $n = 1$, $m = 1$, $x \geqq 0$, $U = E_1$;
4. $f_0 = x^3u^2 - x^2u$, $f = x^2u$, $n = 1$, $m = 1$, $x \geqq 0$, $U = E_1$,

satisfy the conditions of the Existence Theorem 2.1.

REFERENCES

[1a] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints. I, II*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412, 413–429.

[1b] ———, *Existence theorems for optimal controls of the Mayer type*, this Journal, 6 (1968), pp. 517–552.
[1c] ———, *Seminormality and upper semicontinuity in optimal control*, J. Optimization Theory Appl., 6 (1970), pp. 114–137.
[1d] ———, *Closure, lower closure, and semicontinuity theorems in problems of optimal control*, this Journal, 9 (1971), pp. 287–301.
[1e] ———, *Lagrange problems of optimal control and convex sets not containing any straight line*, J. Optimization Theory Appl., to appear.
[2] L. CESARI, J. R. LA PALM AND T. NISHIURA, *Remarks on some existence theorems for optimal control*, Ibid., 3 (1969), pp. 296–305.
[3a] J. R. LA PALM, *Remarks on certain growth conditions in problems of optimal control*, Ibid., 4 (1969), pp. 321–329.
[3b] ———, *A recent significant growth condition in optimal control theory*, Ibid., 4 (1969), pp. 378–385.
[4a] E. J. MCSHANE, *On the semicontinuity of integrals in the calculus of variations*, Ann. of Math., 33 (1932), pp. 460–484.
[4b] ———, *Existence theorems for ordinary problems of the calculus of variations*, Ann. Scuola Norm. Sup. Pisa (2), 3 (1934), pp. 181–211, 287–315.
[4c] ———, *Semicontinuity of integrals in the calculus of variations*, Duke Math. J., 2 (1936), pp. 597–616.
[4d] ———, *Some existence theorems for problems in the calculus of variations*, Ibid., 4 (1938), pp. 132–156.
[4e] ———, *A navigation problem in the calculus of variations*, Amer. Math. J., 59 (1937), pp. 327–334.
[4f] ———, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.
[5] E. J. MCSHANE AND R. B. WARFIELD, *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1969), pp. 41–47.
[6] T. NISHIURA, *On an existence theorem for optimal control*, this Journal, 5 (1967), pp. 532–544.
[7a] C. OLECH, *Existence theorems for optimal problems with vector valued cost functions*, Trans. Amer. Math. Soc., 136 (1967), pp. 157–180.
[7b] ———, *Existence theorems for optimal control problems involving multiple integrals*, J. Differential Equations, 6 (1969), pp. 512–526.
[8a] L. TONELLI, *Sugli integrali del calcolo delle variazioni in forma ordinaria*, Ann. Scuola Norm. Sup. Pisa (2), 3 (1934), pp. 401–450 = Opere Scelte, Edizioni Cremonese, Roma, 3 (1962), pp. 192–254.
[8b] ———, *Fondamenti di Calcolo delle Variazioni*, 2 vols., Zanichelli, Bologna, 1921–23.
[9] L. TURNER, *The direct method in the calculus of variations*, Doctoral thesis, Purdue University, Lafayette, Ind., 1957.

SIAM J. CONTROL
Vol. 9, No. 4, November 1971

NONLINEAR PROGRAMMING IN COMPLEX SPACE: NECESSARY CONDITIONS*

ROBERT A. ABRAMS† AND ADI BEN-ISRAEL‡

Abstract. Necessary conditions of the Kuhn–Tucker type are given for two classes of nonlinear programming problems over polyhedral cones in finite-dimensional complex space. The first class consists of problems of the form:

$$\text{Minimize } \operatorname{Re} f(z) \quad \text{subject to } g(z) \in S,$$

where S is a polyhedral cone in C^m and $f: C^n \to C, g: C^n \to C^m$ are analytic functions. A necessary condition for a feasible point z^0 to be optimal is that there exist a vector $u \in S^*$ such that $\overline{\nabla f(z^0)} = [D_z^H g(z^0)]u$ and $\operatorname{Re}(g(z^0), u) = 0$. The second class consists of problems of the form:

$$\text{Minimize } \operatorname{Re} f(z, \bar{z}) \quad \text{subject to } g(z, \bar{z}) \in S,$$

where $f: C^{2n} \to C, g: C^{2n} \to C^m$ are analytic. A necessary condition for a feasible point z^0 to be optimal is that there exist a $u \in S^*$ such that $\overline{\nabla_z f(z^0, \overline{z^0})} + \nabla_{\bar{z}} f(z^0, \overline{z^0}) = [D_z^H g(z^0, \overline{z^0})]u + [D_{\bar{z}}^T g(z^0, \overline{z^0})]\bar{u}$ and $\operatorname{Re}(g(z^0, \overline{z^0}), u) = 0$. The derivation of necessary conditions for problems of the first class is analogous to that used in the real case. For problems of the second class, necessary conditions are obtained by considering an equivalent problem in the form of the first class.

1. Introduction. In this paper first order necessary conditions of the Kuhn–Tucker type are given for nonlinear programming problems over polyhedral cones in finite-dimensional complex space. Two classes of problems are considered. In the first class, both the objective and constraint functions are assumed to be analytic functions of n complex variables, i.e., the problems are of the form:

$$\text{Minimize } \operatorname{Re} f(z) \quad \text{subject to } g(z) \in S,$$

where S is a polyhedral cone in C^m and the functions $f: C^n \to C$ and $g: C^n \to C^m$ are analytic, at least in a neighborhood of a feasible point z^0 being tested for optimality. A necessary condition for optimality of z^0, given in Theorem 1, is that there exist a u in a subcone of S^* such that

$$\overline{\nabla_z f(z^0)} = [D_z^H g(z^0)]u \quad \text{and} \quad \operatorname{Re}(g(z^0), u) = 0.$$

In the second, more general, class of problems, both objective and constraint functions may involve z and $\bar{z}$ and are assumed to be analytic functions of $2n$ complex variables in a neighborhood of $(z^0, \overline{z^0})$, where z^0 is a feasible point being tested for optimality. Problems in this class are of the form:

$$\text{Minimize } \operatorname{Re} f(z, \bar{z}) \quad \text{subject to } g(z, \bar{z}) \in S,$$

where $f: C^{2n} \to C$ and $g: C^{2n} \to C^m$ are analytic. A necessary condition for the optimality of z^0 is, by Theorem 3, that there exist a vector u in a certain subcone

* Received by the editors February 23, 1970, and in revised form October 8, 1970.

† Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60201.

‡ Departments of Engineering Science and of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60201, and Department of Applied Mathematics, Technion-Israel Institute of Technology, Haifa, Israel. The work of this author was supported in part by the National Science Foundation under Project GP 13546.

of S^* such that

$$\overline{\nabla_z f(z^0, \overline{z^0})} + \nabla_{\bar{z}} f(z^0, \overline{z^0}) = [D_z^H g(z^0, \overline{z^0})]u + [D_{\bar{z}}^T g(z^0, \overline{z^0})]\bar{u}$$

and $\operatorname{Re}(g(z^0, \overline{z^0}), u) = 0$.

Sufficient conditions for optimality, and duality theorems based on the complex analogues of convexity, are given in [1], [2] and [11].

Complex linear programming was studied in [13] (where complex mathematical programming was introduced), [5] and [3] (where the natural extension to quadratic programming is made). Other references are given in the survey article [4], which also contains some applications of complex programming, and in which some of the results of the present paper and of [2] are announced without proof.

This paper contains five sections and an Appendix. The preliminaries of § 2 are used to derive the necessary conditions for problems of the first and second classes in § 3 and § 4 respectively. An example is given in § 5. Some results on derivatives of complex functions which are used here and in [2] are derived in the Appendix.

2. Preliminaries.

NOTATIONS 2.1.

C^n [R^n] denotes the *n-dimensional complex* [*real*] *vector space*;

$C^{m\times n}$ [$R^{m\times n}$], the $m \times n$ *complex* [*real*] *matrices*;

$R^n_+ \equiv \{x \in R^n : x_i \geqq 0,\ i = 1, \cdots, n\}$, the *nonnegative orthant of* R^n;

$x \geqq y$ denotes $x - y \in R^n_+$ for $x, y \in R^n$.

For $A = (a_{ij}) \in C^{m\times n}$,

$\bar{A} \equiv (\bar{a}_{ij})$ denotes the *conjugate*,

$A^T \equiv (a_{ji})$ denotes the *transpose*,

$A^H \equiv \bar{A}^T$ denotes the *conjugate transpose*.

For $x = (x_i) \in C^n$, $y \in C^n$,

$(x, y) \equiv y^H x$ denotes the *inner product* of x and y,

$\bar{x} \equiv (\bar{x}_i)$ denotes the *conjugate*,

$\operatorname{Re} x \equiv (\operatorname{Re} x_i) \in R^n$ denotes the *real part*,

$\operatorname{Im} x \equiv (\operatorname{Im} x_i) \in R^n$ denotes the *imaginary part*,

$\arg x \equiv (\arg x_i)$ denotes the *argument* of x.

For a subspace $L \subset C^n$, $L^\perp \equiv \{y \in C^n : l \in L \Rightarrow (y, l) = 0\}$ denotes the *orthogonal complement* of L.

For a nonempty set $S \subset C^n$, $S^* \equiv \{y \in C^n : x \in S \Rightarrow \operatorname{Re}(y, x) \geqq 0\}$ denotes the *dual* (also *polar*) of S.

For a nonempty set $S \subset R^n$,

$$S^* = \{y \in R^n : x \in S \Rightarrow (x, y) \geqq 0\}.$$

For an analytic function $f : C^n \to C$ and a point $z^0 \in C^n$, $\nabla_z f(z^0) \equiv ((\partial f/\partial z_i)(z^0))$, $i = 1, \cdots, n$, denotes the *gradient* of f at z^0.

For a complex function $f(w^1, w^2)$ analytic in the $2n$ variables (w^1, w^2) at the point $(z^0, \overline{z^0}) \in C^n \times C^n$,

$$\nabla_z f(z^0, \overline{z^0}) \equiv \left(\frac{\partial f}{\partial w_i^1}(z^0, \overline{z^0})\right), \qquad i = 1, \cdots, n,$$

and

$$\nabla_{\bar{z}} f(z^0, \overline{z^0}) \equiv \left(\frac{\partial f}{\partial w_i^2}(z^0, \overline{z^0}) \right), \qquad i = 1, \cdots, n.$$

For an analytic function $g: C^n \to C^m$,

$$D_z g(z^0) \equiv \left(\frac{\partial g_i}{\partial z_j}(z^0) \right), \qquad i = 1, \cdots, m, \quad j = 1, \cdots, n.$$

Similarly, for a function $g: C^n \times C^n \to C^m$ analytic in the $2n$ variables (w^1, w^2) at $(z^0, \overline{z^0}) \in C^n \times C^n$,

$$D_z g(z^0, \overline{z^0}) = \left(\frac{\partial g_i}{\partial w_j^1}(z^0, \overline{z^0}) \right), \qquad i = 1, \cdots, m, \quad j = 1, \cdots, n,$$

and

$$D_{\bar{z}} g(z^0, \overline{z^0}) = \left(\frac{\partial g_i}{\partial w_j^2}(z^0, \overline{z^0}) \right), \qquad i = 1, \cdots, m, \quad j = 1, \cdots, n.$$

Also

$$D_z^T g(z^0, \overline{z^0}) = (D_z g(z^0, \overline{z^0}))^T,$$
$$D_z^H g(z^0, \overline{z^0}) = (D_z g(z^0, \overline{z^0}))^H.$$

DEFINITIONS 2.2. A nonempty set $S \subset C^n$ is:

(a) *convex* if $0 \leqq \lambda \leqq 1 \Rightarrow \lambda S + (1 - \lambda)S \subset S$,
(b) a *cone* if $0 \leqq \lambda \Rightarrow \lambda S \subset S$,
(c) a *polyhedral cone* if for some positive integer k and $A \in C^{n \times k}$,

$$S = AR_+^k = \{Ax : x \in R_+^k\};$$

i.e., S is generated by finitely many vectors (the columns of A).

The following results are needed in the sequel.

Result 2.3. A polyhedral cone in C^n is a closed convex cone.

Result 2.4. A nonempty set $S \subset C^n$ is a closed convex cone if and only if $S = S^{**}$ (for proof see, e.g., [5, Theorem 1.5]).

Result 2.5. If S, T are polyhedral cones, then $S \times T$ is a polyhedral cone.

Result 2.6. For any nonempty sets S, T: $(S \times T)^* = S^* \times T^*$.

Result 2.7. Let $A \in C^{m \times n}$, $b \in C^m$ and $S \subset C^n$ be a polyhedral cone. Then the following are equivalent:

(a) $Ax = b$, $x \in S$ is consistent.
(b) $A^H y \in S^* \Rightarrow \operatorname{Re}(b, y) \geqq 0$. (See [5, Theorem 3.5].)

Result 2.8. The nonnegative orthant R_+^n is a self-dual set in R^n: $(R_+^n)^* = R_+^n$.

Result 2.9. Let S be a polyhedral cone in C^n. Then S is the intersection of finitely many closed half-spaces, each including the origin in its boundary:

$$S = \bigcap_{k=1}^{p} H_{u_k},$$

where $H_{u_k} = \{z \in C^n : \operatorname{Re}(z, u_k) \geqq 0\}$. (It is proved similarly to the real case, e.g., [14].)

Result 2.10. Let $S = \bigcap_{k=1}^{p} H_{u_k}$ be a polyhedral cone in C^n or R^n and let $z^0 \in S$. Then $S(z^0)$ is defined to be the intersection of those closed half-spaces H_{u_k} which include z^0 in their boundaries, i.e.,

$$S(z^0) = \bigcap_{k \in B(z^0)} H_{u_k},$$

where $B(z^0) \equiv \{k : \operatorname{Re}(z^0, u_k) = 0\}$. If z^0 is in the interior of S, then $S(z^0) = C^n$.

Result 2.11. Let $\varnothing \neq S \subset T \subset C^n$. Then $T^* \subset S^*$.

Result 2.12. Let $\{S_i : i = 1, \cdots, p\}$ be closed convex cones in C^n. Then

$$\left(\bigcap_{i=1}^{p} S_i \right)^* = \operatorname{cl} \sum_{i=1}^{p} S_i^*$$

(which follows from [5, Corollary 1.7]).

3. Necessary conditions for problems of the first class. Necessary conditions for optimal points of analytic nonlinear programming problems will now be derived. In the real case the necessary conditions are obtained by using the Farkas theorem to show that the gradient of the objective function lies in a cone generated by the gradients of the constraint functions, e.g., [10, § 2.1]. The generalized Farkas theorem, Result 2.7, is used here in a similar manner.

Let $f : C^n \to C$ and $g : C^n \to C^m$ both be analytic in a neighborhood of a point z^0. ($g : C^n \to C^m$ is analytic if each of its components $g_i : C^n \to C$, $i = 1, \cdots, m$, is analytic.) Let S be a polyhedral cone in C^m. The complex nonlinear programming problem of the first class is:

$$\text{Minimize } \operatorname{Re} f(z) \quad \text{subject to } g(z) \in S. \tag{1}$$

An analogue of the Kuhn–Tucker constraint qualification provides a regularity condition which the constraint is assumed to satisfy at a point being checked for optimality. Let z^0 be a feasible point of (1). Then the (Kuhn–Tucker) *constraint qualification* holds at z^0, or the point z^0 is *qualified*, if every $z \in C^n$ such that $[D_z g(z^0)]z \in S(g(z^0))$ is tangent to a once differentiable arc $a(\theta)$ beginning at z^0 and leading into the feasible region, i.e., $a(0) = z^0$, $g(a(\theta)) \in S$ for $0 \leqq \theta < \varepsilon$ and $a'(0) = kz$ for some $k > 0$, $\varepsilon > 0$.

THEOREM 1. *Let S be a polyhedral cone in C^m. Let $f : C^n \to C$ and $g : C^n \to C^m$ both be analytic in a neighborhood of a qualified point z^0. Then a necessary condition for z^0 to be a local minimum of the problem:*

$$\textit{Minimize } \operatorname{Re} f(z) \quad \textit{subject to } g(z) \in S,$$

is that there exist a vector $u \in [S(g(z^0))]^* \subset S^*$ such that

$$\overline{\nabla_z f(z^0)} = [D_z^H g(z^0)]u \tag{2}$$

and

$$\operatorname{Re}(g(z^0), u) = 0. \tag{3}$$

Proof. Define the set

$$Z = \{z \in C^n : [D_z g(z^0)]z \in S(g(z^0)), \operatorname{Re}(\overline{\nabla_z f(z^0)}, z) < 0\}. \tag{4}$$

If Z is empty, then

$$[D_z g(z^0)]z \in S(g(z^0)) \Rightarrow \operatorname{Re}(\overline{\nabla_z f(z^0)}, z) \geqq 0. \tag{5}$$

The cone $S(g(z^0))$ is polyhedral and hence closed and convex. Therefore by (2.4), $S(g(z^0))$ may be replaced by $([S(g(z^0))]^*)^*$ in (5). By Result 2.7, the complex Farkas theorem, (5) is equivalent to the existence of a $u \in [S(g(z^0))]^*$ such that

$$\overline{\nabla_z f(z^0)} = [D_z^H g(z^0)]u.$$

Thus in order to prove (2), it is sufficient to show that Z is empty.

Consider any z such that $[D_z g(z^0)]z \in S(g(z^0))$. Since z^0 is a qualified point, there is a feasible arc $a(\theta)$ with $a(0) = z^0$ and $a'(0) = kz$ for some $k > 0$. By assumption, z^0 is a local minimum of $\operatorname{Re} f(z)$ along $a(\theta)$. Therefore

$$\frac{d}{d\theta} \operatorname{Re} [f(a(\theta))]|_{\theta=0} \geqq 0 \tag{6}$$

which is equivalent to

$$\operatorname{Re} \frac{d}{d\theta} f(a(\theta))|_{\theta=0} \geqq 0. \tag{7}$$

Using the chain rule, we find that (7) becomes $\operatorname{Re} [\nabla_z^T f(z^0)a'(0)] \geqq 0$, or with $a'(0) = kz$, $\operatorname{Re} (\overline{\nabla f(z)}, z) \geqq 0$, which proves that Z is empty.

To prove (3), note that $u \in [S(g(z^0))]^*$. Then

$$\begin{aligned}[S(g(z^0))]^* &= \left(\bigcap_{k \in B(g(z^0))} H_{u_k} \right)^* = \operatorname{cl} \sum_{k \in B(g(z^0))} (H_{u_k})^* \\ &= \operatorname{cl} \sum_{k \in B(g(z^0))} \{\alpha_k u_k : \alpha_k \geqq 0\} = \sum_{k \in B(g(z^0))} \{\alpha_k u_k : \alpha_k \geqq 0\},\end{aligned}$$

where the second equality follows from (2.12) and the last follows from the fact that $\sum_{k \in B(g(z^0))} \{\alpha_k u_k : \alpha_k \geqq 0\}$ is a polyhedral cone and hence closed (Result 2.3). Therefore $u \in [S(g(z^0))]^*$ implies

$$u = \sum_{i \in B(g(z^0))} \beta_i u_i \quad \text{with } \beta_i \geqq 0. \tag{8}$$

By the definition of $B(g(z^0))$, $\operatorname{Re} (u_i, g(z^0)) = 0$ for $i \in B(g(z^0))$, and hence (3) follows from (8). This completes the proof.

For comparison with the real case, a real version of Theorem 1, i.e., the Kuhn–Tucker theorem for constraints over polyhedral cones, is stated now. The proof follows from the proof of Theorem 1 by replacing complex spaces with real spaces and replacing analytic functions with differentiable functions. The definition of a qualified point is analogous to that given above for the complex case.

THEOREM 2. *Let S be a polyhedral cone in R^m. Let $f: R^n \to R$ and $g: R^n \to R^m$ both be (Fréchet) differentiable in a neighborhood of a qualified point x^0. Then a necessary condition for x^0 to be a local minimum of the problem*:

$$\textit{Minimize} \quad f(x) \quad \textit{subject to } g(x) \in S,$$

is that there exist a $u \in [S(g(x^0))]^*$ such that

$$\nabla_x f(x^0) = [D_x^T g(x^0)]u$$

and

$$(g(x^0), u) = 0.$$

The usual form of the Kuhn–Tucker theorem for l inequality constraints and p equality constraints is obtained from the above theorem by taking

$$S = R^l_+ \times \mathbf{0}^{m-l},$$

where $\mathbf{0}^{m-l}$ is the zero element of the $(m-l)$-dimensional real vector space. The representation of S as an intersection of half-spaces is

$$S = \bigcap_{k=1}^{m} H_{e_k} \bigcap_{k=l+1}^{m} H_{-e_k},$$

where e_k is the kth unit vector in R^m. Suppose that at the point x^0 the first r inequality constraints are binding and that $l - r$ are not, i.e.,

$$(g(x^0), e_k) = 0, \quad k = 1, \cdots, r, \quad \text{and} \quad (g(x^0), e_k) > 0, \quad k = r+1, \cdots, l.$$

Then

$$S(g(x^0)) = \bigcap_{k=1}^{r} H_{e_k} \bigcap_{k=l+1}^{m} H_{e_k} \bigcap_{k=l+1}^{m} H_{-e_k}$$

and

$$[S(g(x^0))]^* = \sum_{k=1}^{r} \{\alpha_k e_k : \alpha_k \geqq 0\} + \sum_{k=l+1}^{m} \{(\alpha_k - \alpha_k')e_k : \alpha_k \geqq 0, \alpha_k' \geqq 0\}.$$

Thus $u \in [S(g(x^0))]^*$ implies

$$u = \sum_{k=1}^{r} \beta_k e_k + \sum_{k=l+1}^{m} \gamma_k e_k,$$

where $\beta_k \geqq 0$, $k = 1, \cdots, r$, and γ_k is unrestricted, $k = l+1, \cdots, m$. The β_k and the γ_k are multipliers of the inequality constraints and the equality constraints respectively.

4. Necessary conditions for problems of the second class. The class of problems under consideration is now extended to those in which f and g are analytic functions of the $2n$ variables $(w^1, w^2) \in C^{2n}$ and in which the additional constraint $w^2 = \overline{w^1}$ is added. Thus let $f : C^{2n} \to C$ and $g : C^{2n} \to C^m$ both be analytic in a neighborhood of a point $(z^0, \overline{z^0}) \in C^{2n}$ and let S be a polyhedral cone in C^m. Consider the problem:

(9) Minimize $\operatorname{Re} f(w^1, w^2)$ subject to $g(w^1, w^2) \in S$ and $w^2 = \overline{w^1}$,

which is rewritten in the more suggestive form:

(10) Minimize $\operatorname{Re} f(z, \bar{z})$ subject to $g(z, \bar{z}) \in S$.

From (A.20) it follows that a function explicitly involving the variable $\bar{z}$, such as $f(z, \bar{z})$ and $g(z, \bar{z})$ in (10), cannot be an analytic function of the variable z. Therefore Theorem 1 is not directly applicable to problems of the form (10). However, it is possible to recast (10) in the form (1), and thus optimality conditions for (10) can be obtained from Theorem 1.

Define the set Q by

(11) $$Q \equiv \left\{ \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \in C^{2n} : w^2 = \overline{w^1} \right\}.$$

Q is a polyhedral cone; it is generated by the set of vectors

$$\left\{ \bigcup_{j=1}^{n} \begin{pmatrix} e_j \\ e_j \end{pmatrix} \bigcup_{j=1}^{n} \begin{pmatrix} -e_j \\ -e_j \end{pmatrix} \bigcup_{j=1}^{n} \begin{pmatrix} ie_j \\ -ie_j \end{pmatrix} \bigcup_{j=1}^{n} \begin{pmatrix} -ie_j \\ ie_j \end{pmatrix} \right\}, \tag{12}$$

where e_j is the jth unit vector in R^n. The dual of Q is easily seen to be the polyhedral cone

$$Q^* = \left\{ \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \in C^{2n} : w^2 = -\overline{w^1} \right\}. \tag{13}$$

Since S and Q are polyhedral cones, $S \times Q$ is also (Result 2.5), and therefore (10) may be written in the form (1) as:

$$\begin{aligned} &\text{Minimize} \quad \operatorname{Re} f(w^1, w^2) \\ &\text{subject to} \quad \begin{pmatrix} g(w^1, w^2) \\ \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \end{pmatrix} \in S \times Q. \end{aligned} \tag{14}$$

A qualified point for (10) is defined using (14) and the definition of a qualified point for (1). Thus (v^1, v^2) is a qualified point for (14) if, for every $(w^1, w^2) \in C^{2n}$ such that

$$D_{\binom{w^1}{w^2}} \begin{pmatrix} g(v^1, v^2) \\ \begin{pmatrix} v^1 \\ v^2 \end{pmatrix} \end{pmatrix} \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \in [S \times Q] \begin{pmatrix} g(v^1, v^2) \\ (v^1, v^2) \end{pmatrix}, \tag{15}$$

there exists an arc $a(\theta) \subset C^{2n}$ such that $a(0) = (v^1, v^2), a'(0) = k \begin{pmatrix} w^1 \\ w^2 \end{pmatrix}, k > 0,$ and

$$\begin{pmatrix} g(a(\theta)) \\ a(\theta) \end{pmatrix} \in S \times Q, \qquad 0 \leqq \theta < \varepsilon.$$

Condition (15) is equivalent to

$$\begin{pmatrix} D_{\binom{w^1}{w^2}} g(v^1, v^2) \\ I \end{pmatrix} \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \in S(g(v^1, v^2)) \times Q(v^1, v^2). \tag{16}$$

For any point (v^1, v^2) in the cone Q, $Q(v^1, v^2) = Q$, and therefore (16) and (15) become

$$D_{\binom{w^1}{w^2}} g(v^1, v^2) \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \in S(g(v^1, v^2)) \quad \text{and} \quad \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \in Q.$$

Thus we may say that a point $(z^0, \overline{z^0})$ is qualified for (10) if for all $(z, \bar{z})$ such that

$$[D_{\binom{z}{\bar{z}}} g(z^0, \overline{z^0})] \begin{pmatrix} z \\ \bar{z} \end{pmatrix} \in S(g(z^0, \overline{z^0})),$$

there exists an arc $(\beta(\theta), \overline{\beta(\theta)}) \subset C^{2n}$ such that $\beta(0) = z^0, \beta'(0) = kz$ for some $k > 0$ and $g(\beta(\theta), \overline{\beta(\theta)}) \in S$ for $0 \leqq \theta < \varepsilon$.

THEOREM 3. *Let S be a polyhedral cone in C^m. Let $f: C^{2n} \to C$ and $g: C^{2n} \to C^m$ both be analytic in a neighborhood of a qualified point $(z^0, \overline{z^0})$. A necessary condition for $z^0, \overline{z^0}$ to be a local minimum of the problem:*

$$\text{Minimize} \quad \operatorname{Re} f(z, \bar{z}) \quad \text{subject to } g(z, \bar{z}) \in S,$$

is that there exist a $u \in [S(g(z^0, \overline{z^0}))]^* \subset S^*$ such that

$$\overline{\nabla_z f(z^0, \overline{z^0})} + \nabla_{\bar{z}} f(z^0, \overline{z^0}) = [D_z^H g(z^0, \overline{z^0})]u + [D_{\bar{z}}^T g(z^0, \overline{z^0})]\bar{u} \tag{17}$$

and

$$\operatorname{Re}(g(z^0, \overline{z^0}), u) = 0. \tag{18}$$

Proof. The problem under consideration has been shown to be equivalent to (14) which is of the form (1). We apply Theorem 1 to (14). Thus a necessary condition for qualified point (w^1, w^2) to be an optimal point of (14) is that there exist a

$$\begin{pmatrix} u \\ \lambda \end{pmatrix} \in \left\{ [S \times Q] \begin{pmatrix} g(w^1, w^2) \\ \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \end{pmatrix} \right\}^* = [S(g(w^1, w^2))]^* \times Q^*$$

such that

$$\overline{\nabla_{\binom{w^1}{w^2}} f(w^1, w^2)} = D^H_{\binom{w^1}{w^2}} \begin{pmatrix} g(w^1, w^2) \\ \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} \tag{19}$$

and

$$\operatorname{Re} \left\langle \begin{pmatrix} g(w^1, w^2) \\ \begin{pmatrix} w^1 \\ w^2 \end{pmatrix} \end{pmatrix}, \begin{pmatrix} u \\ \lambda \end{pmatrix} \right\rangle = 0. \tag{20}$$

From (13), $\lambda \in Q^*$ implies $\lambda = \begin{pmatrix} \pi \\ -\bar{\pi} \end{pmatrix}$. Therefore (19) becomes

$$\begin{aligned} \overline{\nabla_{w^1} f(w^1, w^2)} &= D^H_{w^1} g(w^1, w^2)u + \pi, \\ \overline{\nabla_{w^2} f(w^1, w^2)} &= D^H_{w^2} g(w^1, w^2)u - \bar{\pi}. \end{aligned} \tag{21}$$

Conjugating the second equation of (21) and adding to the first yields

$$\overline{\nabla_{w^1} f(w^1, w^2)} + \nabla_{w^2} f(w^1, w^2) = D^H_{w^1} g(w^1, w^2)u + D^T_{w^2} g(w^1, w^2)\bar{u}$$

which is equivalent to (17) since for any feasible point of (14), $w^2 = \overline{w^1}$.

Substituting $\lambda = \begin{pmatrix} \pi \\ -\bar{\pi} \end{pmatrix}$, and $w^2 = \overline{w^1}$ in (20) yields

$$\operatorname{Re}(g(w^1, \overline{w^1}), u) + \operatorname{Re}[{w^1}^H \pi - {w^1}^T \bar{\pi}] = 0. \tag{22}$$

The second term of (22) is zero, and therefore (23) follows. This completes the proof.

COROLLARY. *In addition to the hypothesis of Theorem* 3, *assume that* $g(z, \bar{z})$ *is real-valued for all* $z \in C^n$. *Then* S *may be assumed to be real, and a necessary condition for a qualified point* $(z^0, \overline{z^0})$ *to be a local minimum of* (10) *is that there exist a* $v \in [S(g(z^0, \overline{z^0}))]^*$, *where* $S(g(z^0, \overline{z^0}))$ *and the dual cone are calculated in* R^m, *such that*

$$\overline{\nabla_z f(z^0, \overline{z^0})} + \nabla_{\bar{z}} f(z^0, \overline{z^0}) = D_z^H g(z^0, \overline{z^0}) v \tag{23}$$

and

$$(g(z^0, \overline{z^0}), v) = 0. \tag{24}$$

Proof. If S contains any complex vectors, it may be replaced by the polyhedral cone $\hat{S} \equiv \{z : z \in S, \operatorname{Im} z = 0\}$ with no loss in generality. Since $g(z, \bar{z})$ is real-valued, (A.33) implies $D_{\bar{z}} g(z^0, \overline{z^0}) = D_z g(z^0, \overline{z^0})$. Therefore (17) yields the existence of a vector u in the complex dual of $[\hat{S}(g(z^0, \overline{z^0}))]^*$ such that

$$\overline{\nabla_z f(z^0, \overline{z^0})} + \nabla_{\bar{z}} f(z^0, \overline{z^0}) = D_z^H g(z^0, \overline{z^0})(u + \bar{u}). \tag{25}$$

The cone $\hat{S} \subset C^m$ may be represented as the intersection of S with the cone

$$K = \bigcap_{j=1}^{m} H_{ie_j} \bigcap_{j=1}^{m} H_{-ie_j}, \tag{26}$$

where e_j is the jth unit vector in R^m. Therefore

$$\hat{S}(g(z^0, \overline{z^0})) = (S \cap K)(g(z^0, \overline{z^0})) = S(g(z^0, \overline{z^0})) \cap K(g(z^0, \overline{z^0})) = S(g(z^0, \overline{z^0})) \cap K$$

is a real polyhedral cone. The cones $\hat{S} \subset C^m$ and $\hat{S}(g(z^0, \overline{z^0})) \subset C^m$ which contain only real vectors will be identified with their isomorphic images in R^m. The same notation, $\hat{S}$ and $\hat{S}(g(z^0, \overline{z^0}))$, will be used in either case, but the meaning will be clear from the context.

If $\hat{S} \subset C^m$ is written as an intersection of half-spaces; i.e., $\hat{S} = \bigcap_{k=1}^{p} H_{u_k}$, then in R_m we have $\hat{S} = \bigcap_{k=1}^{p} H_{\operatorname{Re} u_k}$. For any point $g(z^0, \overline{z^0}) \in \hat{S}$, $\operatorname{Re}(u_k, g(z^0, \overline{z^0})) = 0$ if and only if $(\operatorname{Re} u_k, g(z^0, \overline{z^0})) = 0$. Thus $\hat{S}(g(z^0, \overline{z^0})) = \bigcap_{k \in B(g(z^0, \overline{z^0}))} H_{u_k}$ is the same as $\bigcap_{k \in B(g(z^0, \overline{z^0}))} H_{\operatorname{Re} u_k}$, where the $H_{\operatorname{Re} u_k}$ are calculated in R^m.

It follows easily that if $u \in [\hat{S}(g(z^0, \overline{z^0}))]^*$ calculated in C^m, then $v = 2 \operatorname{Re} u$ is an element of $[\hat{S}(g(z^0, \overline{z^0}))]^*$ calculated in R^m. Thus (23) holds. Since both v and $g(z^0, \overline{z^0})$ are real, (24) follows from (18). This completes the proof.

If the objective function $f(z, \bar{z})$ is real-valued for $z \in C^n$, then using (A.33) we may replace the sum $\overline{\nabla_z f(z^0, \overline{z^0})} + \nabla_{\bar{z}} f(z^0, \overline{z^0})$ in (17) or (23) by either of its terms. If both $f(z, \bar{z})$ and $g(z, \bar{z})$ were obtained from analytic functions $f(z)$ and $g(z)$ by substituting $z = x + iy$ and then $x = (z + \bar{z})/2$ and $y = (z - \bar{z})/(2i)$, then (A.20) implies that $\nabla_{\bar{z}} f(z, \bar{z}) = D_{\bar{z}} g(z, \bar{z}) = 0$. Using (A.22) and (A.34) we obtain $\nabla_z f(z, \bar{z}) = \nabla_z f(z)$ and $D_z g(z, \bar{z}) = D_z g(z)$, and thus, in this case, (17) reduces to (2).

5. Example. A simple example, with $n = m = 1$, is used to illustrate the above results:

$$\text{Minimize } \operatorname{Re} z^2 \quad \text{subject to } |z| \leqq 1. \tag{27}$$

To put (27) in the form (1), note that $|z| \leqq 1$ may be rewritten:

$$\log z \equiv \log |z| + i \arg z \in S \equiv \text{left half-plane.} \tag{28}$$

However, in (28) the point $z = 0$ has been mapped to the point at infinity and must therefore be considered separately. By Theorem 1, a necessary condition for the problem:

$$\text{Minimize } \operatorname{Re} z^2 \quad \text{subject to } \log z \in S,$$

to have a minimum at a point z is that there exist a $u \in S^*$ such that

$$\overline{2z} = \frac{1}{\bar{z}} u \quad \text{and} \quad \operatorname{Re}(\log z, u) = 0. \tag{29}$$

In this case S^* is the negative real axis and thus (29) becomes

$$\bar{z}^2 = \frac{u}{2} \leqq 0 \quad \text{and} \quad (\log |z|, u) = 0. \tag{30}$$

The second condition implies that either $|z| = 1$ or $u = 0$. If $|z| = 1$, the first condition implies that $\bar{z}^2$ is negative and therefore $\operatorname{Re} z = 0$ and $\operatorname{Im} z = +1$, which are clearly optimal points of the problem. If $u = 0$, then the first condition implies $z = 0$, which is a saddle point of the problem.

By writing $|z| \leqq 1$ as $1 - z\bar{z} \geqq 0$, (27) may alternatively be put in the form of (10) Using the corollary to Theorem 3 and the following paragraph, the necessary conditions are that there exist a $u \in [S(g(z, \bar{z}))]^* \subset S^* = R_+$ such that

$$\bar{z} = -zu \quad \text{and} \quad (1 - z\bar{z}, u) = 0. \tag{31}$$

If $u = 0$, then $z = 0$. If $u > 0$, then the second constraint gives $|z| = 1$ and the first implies

$$\bar{z}^2 = -z\bar{z}u = -u < 0,$$

and hence (31) is equivalent to (30).

Appendix. Identities for differentiation of complex functions. The definition of analyticity of a function of several variables is usually given in terms of a power series or by means of the Cauchy–Riemann equation in each variable, e.g., [6], [7] and [8]. The power series definition will be used in this paper.

DEFINITION. Let $f: C^n \to C$ and let D be a domain in C^n. Then $f(z)$ will be called *analytic* in D if in some neighborhood of every point of D it may be represented by an absolutely convergent power series about that point in the n complex variables.

If $f(z)$ is analytic in n variables, then it is analytic in each variable separately when the others are held fixed, and hence the Cauchy–Riemann equations hold in each variable, i.e.,

$$\nabla_x F^R(x, y) = \nabla_y F^I(x, y), \qquad \nabla_x F^I(x, y) = -\nabla_y F^R(x, y), \tag{A.1}$$

where $F^R(x, y)$ and $F^I(x, y)$ are the real and imaginary parts of

$$F(x, y) \equiv f(z(x, y)). \tag{A.2}$$

When $f(z)$ is analytic it also follows that all partials of all orders of $F^R(x, y)$ and $F^I(x, y)$ exist and are continuous. Therefore both $F^R(x, y)$ and $F^I(x, y)$ have Fréchet derivatives at points (x, y), where $z = x + iy$ is a point at which $f(z)$ is analytic.

Applying the chain rule to (A.2) gives

$$\frac{\partial F(x, y)}{\partial x_k} = \frac{\partial f(z)}{\partial z_k}\frac{\partial z_k}{\partial x_k} = \frac{\partial f(z)}{\partial z_k}, \qquad k = 1, \cdots, n, \tag{A.3a}$$

or

$$\nabla_x F(x, y) = \nabla_z f(z) \tag{A.3b}$$

and

$$\frac{\partial F(x, y)}{\partial y_k} = \frac{\partial f(z)}{\partial z_k}\frac{\partial z_k}{\partial y_k} = i\frac{\partial f(z)}{\partial z_k}, \qquad k = 1, \cdots, n, \tag{A.4a}$$

$$\nabla_y F(x, y) = i\nabla_z f(z). \tag{A.4b}$$

If in the functions $F^R(x, y)$ and $F^I(x, y)$, $(z + \bar{z})/2$ and $(z - \bar{z})/(2i)$ are formally substituted for x and y respectively, "functions" of the "variables" z and $\bar{z}$ result. It will be convenient at times to formally differentiate these "functions" with respect to z or $\bar{z}$. The following lemma guarantees that these functions are analytic.

LEMMA A.1. *Let $f: C^n \to C$ be analytic in D. Then there exist functions $f^R: C^{2n} \to C$ and $f^I: C^{2n} \to C$, both analytic in a neighborhood of $(z, \bar{z})$ for any $z \in D$, such that on the manifold $\{(w^1, w^2) \in C^{2n}: w^2 = \overline{w^1}\}$ both functions are real-valued, and*

$$f(z) = f^R(z, \bar{z}) + if^I(z, \bar{z}). \tag{A.5}$$

Proof. Let $z^0 \in D$. Then $f(z)$ may be represented as a power series about z^0 which is absolutely convergent in some polycylinder $P = \{z \in C^n: |z_k - z_k^0| < r_k, k = 1, \cdots, n\}$. Define $h: C^{2n} \to C^n$ by $h(u_1, \cdots, u_n, v_1, \cdots, v_n) = (u_1 + iv_1, \cdots, u_n + iv_n)$. The function $h(u, v)$ is clearly analytic and therefore the composition of $f(z)$ and $h(u, v)$ is analytic in $h^{-1}(P)$, i.e.,

$$E(u, v) \equiv f(h(u, v)), \quad \text{where } (u, v) \in C^{2n}, \tag{A.6}$$

is analytic in $h^{-1}(P)$. The point $(u^0, v^0) \in C^{2n}$, where $u^0 = \operatorname{Re} z^0$ and $v^0 = \operatorname{Im} z^0$, is an element of $h^{-1}(P)$, and hence $E(u, v)$ may be expanded in an absolutely convergent power series about this point; i.e.,

$$E(u, v) = \sum_j a_j(u_1 - u_1^0)^{j_1} \cdots (u_n - u_n^0)^{j_n}(v_1 - v_1^0)^{j_{n+1}} \cdots (v_n - v_n^0)^{j_{2n}}, \tag{A.7}$$

where $j = (j_1, \cdots, j_{2n})$ and each j_k is a positive integer, converges in a neighborhood of the (real) point (u^0, v^0) contained in $h^{-1}(P)$.

The analytic functions $E^R(u, v)$ and $E^I(u, v)$ are defined by replacing a_j by $\operatorname{Re} a_j$ and $\operatorname{Im} a_j$ respectively in (A.7), i.e.,

$$E^R(u, v) = \sum_j \operatorname{Re}(a_j)(u_1 - u_1^0)^{j_1} \cdots (v_n - v_n^0)^{j_{2n}} \tag{A.8}$$

and

$$E^I(u, v) = \sum_j \operatorname{Im}(a_j)(u_1 - u_1^0)^{j_1} \cdots (v_n - v_n^0)^{j_{2n}}. \tag{A.9}$$

The series (A.8) and (A.9) converge absolutely at any point at which (A.7) converges absolutely, and hence they represent analytic functions at these points. The function $p: C^{2n} \to C^{2n}$ defined by

$$p(z^1, z^2) = \left(\frac{z^1 + z^2}{2}, \frac{z^1 - z^2}{2i} \right) \tag{A.10}$$

is now composed with (A.8) and (A.9) to give

$$f^R(z^1, z^2) \equiv E^R(p(z^1, z^2)), \tag{A.11}$$

$$f^I(z^1, z^2) \equiv E^I(p(z^1, z^2)). \tag{A.12}$$

Since $p(z^1, z^2)$ is analytic in each component on the whole plane, $f^R(z^1, z^2)$ and $f^I(z^1, z^2)$ are analytic in the inverse image under p of the domain of analyticity of $E^R(u, v)$ and $E^I(u, v)$ which as noted above contains the domain of analyticity of $E(u, v)$. Thus if $f(z)$ is analytic in a neighborhood of z^0, $E^R(u, v)$ and $E^i(u, v)$ are analytic in a neighborhood of $(\operatorname{Re} z^0, \operatorname{Im} z^0) \in C^{2n}$. The inverse image of this point under the map p is the point $(z^0, \overline{z^0}) \in C^{2n}$, and therefore $f^R(z^1, z^2)$ and $f^I(z^1, z^2)$ are analytic in a neighborhood of this point. Since $p(z, \bar{z})$ is real-valued, $f^R(z, \bar{z})$ and $f^I(z, \bar{z})$ are both real-valued. Also $h(p(z, \bar{z})) = z$. Therefore

$$\begin{aligned} f^R(z, \bar{z}) + if^I(z, \bar{z}) &= E^R(p(z, \bar{z})) + iE^I(p(z, \bar{z})) \\ &= E(p(z, \bar{z})) = f[h(p(z, \bar{z}))] = f(z). \end{aligned} \tag{A.13}$$

This completes the proof.

A number of identities will now be derived for the analytic function $f(z)$ and the related functions defined above. With $f^R(z^1, z^2)$ and $f^I(z^1, z^2)$ defined by (A.11) and (A.12), let

$$f(z^1, z^2) \equiv f^R(z^1, z^2) + if^I(z^1, z^2). \tag{A.14}$$

Derivatives of left- and right-hand terms of (A.14) will usually be evaluated on the manifold $\{(z^1, z^2): z^2 = \overline{z^1}\}$. Thus as in Notations 2.1, $\partial f(z, \bar{z})/\partial z_k^1$ and $\partial f(z, \bar{z})/\partial z_k^2$ will be denoted by $\partial f(z, \bar{z})/\partial z_k$ and $\partial f(z, \bar{z})/\partial \bar{z}_k$ respectively; i.e., derivatives will be written as though z and $\bar{z}$ were independent variables. Derivatives of $f^R(z^1, z^2)$ and $f^I(z^1, z^2)$ are written similarly. From (A.11) and (A.12) we have

$$f^R(z, \bar{z}) = E^R(p(z, \bar{z})), \quad f^I(z, \bar{z}) = E^I(p(z, \bar{z})), \quad f(z, \bar{z}) = E(p(z, \bar{z})). \tag{A.15}$$

Denote the components of $p(z^1, z^2)$ (defined in (A.10)) by $(p_1^1(z^1, z^2) \cdots p_n^1(z^1, z^2)$, $p_1^2(z^1, z^2) \cdots p_n^2(z^1, z^2))$. Differentiating the last of equations (A.15) using the chain rule (which may be applied to the composition of analytic functions) gives

$$\frac{\partial f(z, \bar{z})}{\partial z_k} = \frac{\partial E}{\partial u_k} \frac{\partial p_k^1}{\partial z_k} + \frac{\partial E}{\partial v_k^2} \frac{\partial p_k^2}{\partial z_k}, \qquad k = 1, \cdots, n, \tag{A.16}$$

or

$$\nabla_z f(z, \bar{z}) = \frac{1}{2} \nabla_u E(x, y) - \frac{i}{2} \nabla_v E(x, y). \tag{A.17}$$

The functions $E(u, v)$, $E^R(u, v)$ and $E^I(u, v)$ are analytic continuations of $F(x, y)$, $F^R(x, y)$ and $F^I(x, y)$ respectively. Therefore $\nabla_u E(x, y) = \nabla_x F(x, y)$ and $\nabla_v E(x, y) = \nabla_y F(x, y)$ with similar equations holding for $E^R(u, v)$ and $E^I(u, v)$. Thus (A.17) becomes

$$\nabla_z f(z, \bar{z}) = \frac{1}{2}\nabla_x F(x, y) - \frac{i}{2}\nabla_y F(x, y). \tag{A.18}$$

Taking partials with respect to z^2 instead of z^1 as was done above yields

$$\nabla_{\bar{z}} f(z, \bar{z}) = \frac{1}{2}\nabla_x F(x, y) + \frac{i}{2}\nabla_y F(x, y). \tag{A.19}$$

Expanding the right-hand side of (A.19) gives, as an alternate form of the Cauchy–Riemann equations,

$$\begin{aligned} \nabla_{\bar{z}} f(z, \bar{z}) &= \tfrac{1}{2}[\nabla_x F^R(x, y) + i\nabla_x F^I(x, y) + i\nabla_y F^R(x, y) - \nabla_y F^I(x, y)] \\ &= 0 \quad \text{(by (A.1))}. \end{aligned} \tag{A.20}$$

A similar application of the chain rule to $f^R(z, \bar{z})$ instead of $f(z, \bar{z})$ gives

$$2\nabla_z f^R(z, \bar{z}) = \nabla_x F^R(x, y) - i\nabla_y F(x, y), \tag{A.21}$$

which, with (A.1) and (A.3), yields the interesting formula

$$2\nabla_z f^R(z, \bar{z}) = \nabla_z f(z). \tag{A.22}$$

For the applications of this paper the domain of interest of a function of $2n$ complex variables will be the manifold $\{(w^1, w^2): w^2 = \overline{w^1}\}$. The following lemma for functions of $2n$ complex variables is analogous to Lemma A.1.

LEMMA A.2. *Let $f: C^{2n} \to C$ be analytic in a domain D containing a point $(z^0, \overline{z^0}) \in C^{2n}$. Then there exist functions $f^R: C^{2n} \to C$ and $f^I: C^{2n} \to C$ analytic in a neighborhood of $(z^0, \overline{z^0})$ such that on the manifold $\{(w^1, w^2): w^2 = \overline{w^1}\}$, $f^R(w^1, w^2)$ and $f^I(w^1, w^2)$ are real-valued and*

$$f(w^1, w^2) = f^R(w^1, w^2) + if^I(w^1, w^2). \tag{A.23}$$

Outline of proof. Expand $f(w^1, w^2)$ in an absolutely convergent series about $(z^0, \overline{z^0})$. Define $\bar{f}(w^1, w^2)$ by

$$\bar{f}(w^1, w^2) = \overline{f(\overline{w^2}, \overline{w^1})}. \tag{A.24}$$

Then the series $\bar{f}(w^1, w^2)$ also converges, and hence $\bar{f}(w^1, w^2)$ is analytic in a neighborhood of $(z^0, \overline{z^0})$. The required functions are then defined by

$$f^R(w^1, w^2) = \frac{f(w^1, w^2) + \bar{f}(w^1, w^2)}{2} \tag{A.25}$$

and

$$f^I(w^1, w^2) = \frac{f(w^1, w^2) - \bar{f}(w^1, w^2)}{2i}. \tag{A.26}$$

They are analytic and, since $\overline{f(z, \bar{z})} = \bar{f}(z, \bar{z})$, real-valued on $\{(w^1, w^2): w^2 = \overline{w^1}\}$. This completes the proof.

Identities for analytic functions of $2n$ variables are now obtained. Let $f: C^{2n} \to C$ be analytic in a neighborhood of $(z, \bar{z})$. Define $h: R^{2n} \to C^n$ by $h(u, v) = u + iv$ and $H: R^{4n} \to C$ by $H(x^1, y^1, x^2, y^2) = f(h(x^1, y^1), h(x^2, y^2))$. In addition, define $g_1(x) = x$ and $g_2(x) = -x$ for $x \in R^n$. Then, what is written informally as $F(x, y) = f(z(x, y), \bar{z}(x, y))$ may be defined more precisely as

$$F(x, y) \equiv H(x, y, g_1(x), g_2(y)) = f(h(x, y), h(g_1(x), g_2(y))). \tag{A.27}$$

$F(x, y)$ is the composition of a complex function of complex variables and a complex function of real variables both of which are differentiable. Hence the chain rule may be applied to (A.27) to give

$$\begin{aligned} \frac{\partial F(x, y)}{\partial x_k} &= \frac{\partial f(z, \bar{z})}{\partial w_k^1} \frac{\partial h}{\partial u_k} + \frac{\partial f(z, \bar{z})}{\partial w_k^2} \frac{\partial h}{\partial u_k} \frac{\partial g}{\partial x_k} \\ &= \frac{\partial f(z, \bar{z})}{\partial z} + \frac{\partial f(z, \bar{z})}{\partial \bar{z}}, \qquad k = 1, \cdots, n, \end{aligned} \tag{A.28}$$

where $\partial f(z, \bar{z})/\partial z_k$ and $\partial f(z, \bar{z})/\partial \bar{z}_k$ have been used to denote $\partial f(z, \bar{z})/\partial z_k^1$ and $\partial f(z, \bar{z})/\partial z_k^2$ respectively.

In a similar manner we obtain

$$\frac{\partial F(x, y)}{\partial y_k} = i \frac{\partial f(z, \bar{z})}{\partial z_k} - i \frac{\partial f(z, \bar{z})}{\partial \bar{z}_k}, \qquad k = 1, \cdots, n, \tag{A.29}$$

or rewriting (A.28) and (A.29) in gradient notation,

$$\nabla_x F(x, y) = \nabla_z f(z, \bar{z}) + \nabla_{\bar{z}} f(z, \bar{z}), \tag{A.30a}$$

$$\nabla_y F(x, y) = i \nabla_z f(z, \bar{z}) - i \nabla_{\bar{z}} f(z, \bar{z}). \tag{A.30b}$$

Muliplying (A.30a) by i and adding to (A.30b) gives

$$\nabla_z f(z, \bar{z}) = \frac{1}{2} \nabla_x F(x, y) - \frac{i}{2} \nabla_y F(x, y), \tag{A.31}$$

while multiplying (A.30b) by i and adding to (A.30a) gives

$$\nabla_{\bar{z}} f(z, \bar{z}) = \frac{1}{2} \nabla_x F(x, y) + \frac{i}{2} \nabla_y F(x, y). \tag{A.32}$$

Since (A.30)–(A.32) hold for any analytic function of $2n$ variables, they hold for the $f^R(w^1, w^2)$ and $f^I(w^1, w^2)$ defined in Lemma A.1, and hence (A.31) and (A.32) agree with (A.18) and (A.19). If the function is real-valued, e.g., $f^R(z, \bar{z})$ or $f^I(z, \bar{z})$ as defined in Lemmas A.1 or A.2, then $F(x, y)$ is real-valued; and therefore, (A.31) and (A.32) imply

$$\overline{\nabla_z f(z, \bar{z})} = \nabla_{\bar{z}} f(z, \bar{z}). \tag{A.33}$$

A final identity is now obtained. First differentiate (A.23) with respect to w^1 and then with respect to w_2 to give

$$\nabla_z f(z, \bar{z}) = \nabla_z f^R(z, \bar{z}) + i \nabla_z f^I(z, \bar{z}) \tag{a}$$

and

$$\text{(b)} \qquad \nabla_{\bar{z}} f(z, \bar{z}) = \nabla_{\bar{z}} f^R(z, \bar{z}) + i \nabla_{\bar{z}} f^I(z, \bar{z}).$$

Conjugating (a), adding (b) and using (A.33) yields

$$\text{(A.34)} \qquad \overline{\nabla_z f(z, \bar{z})} + \nabla_{\bar{z}} f(z, \bar{z}) = 2 \nabla_z f^R(z, \bar{z}).$$

REFERENCES

[1] R. A. ABRAMS, *Nonlinear programming in complex space*, Doctoral thesis in Applied Mathematics, Northwestern Univ., Evanston, Ill., 1969.

[2] ———, *Nonlinear programming in complex space: Sufficient conditions and duality*, J. Math. Anal. Appl., to appear.

[3] R. A. ABRAMS AND A. BEN-ISRAEL, *A duality theorem for complex quadratic programming*, J. Optimization Theory Appl., 4 (1969), pp. 244–252.

[4] ———, *Complex mathematical programming*, Developments in Operations Research, B. Avi-Itzhak, ed., Gordon and Breach, New York, 1971.

[5] A. BEN-ISRAEL, *Linear equations and inequalities on finite dimensional, real or complex, vector spaces: A unified theory*, J. Math. Anal. Appl., 27 (1969), pp. 367–389.

[6] S. BOCHNER AND W. T. MARTIN, *Several Complex Variables*, Princeton University Press, Princeton, N.J., 1948.

[7] H. J. BREMERMAN, *Several complex variables*, Studies in Real and Complex Analysis, I. Hirschman, ed., The Mathematical Association of America, Washington, D.C., 1965.

[8] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[9] J. FARKAS, *Über die Theorie der einfachen Ungleichungen*, J. Reine Angew. Math., 124 (1902), pp. 1–24.

[10] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[11] M. A. HANSON AND B. MOND, *Duality for nonlinear programming in complex space*, J. Math. Anal. Appl., 28 (1969), pp. 52–58.

[12] H. W. KUHN AND A. W. TUCKER, *Non-linear programming*, Proc. Second Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., University of California Press, Berkeley, Calif., 1951, pp. 481–493.

[13] N. LEVINSON, *Linear programming in complex space*, J. Math. Anal. Appl., 14 (1966), pp. 44–62.

[14] H. WEYL, *The elementary theory of convex polyhedra*, Contributions to the Theory of Games, vol. I, Annals of Mathematics Studies No. 24, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, N.J., 1950, pp. 3–18.

SIAM J. CONTROL
Vol. 9, No. 4, November 1971

REPRESENTATION OF MARTINGALES, QUADRATIC VARIATION AND APPLICATIONS*

E. WONG†

Abstract. In this paper, we present two related results. First, we shall obtain a sufficient condition under which a second order sample-continuous martingale can be represented as a stochastic integral in terms of a Brownian motion. Secondly, we shall show that if X and Y are sample-continuous local martingales (not necessarily with respect to the same family of σ-algebras) and if either $X + Y$ or $X - Y$ is almost surely of bounded variation, then the quadratic variations of the two martingales are equal. This rather simple result has some surprising consequences.

1. Introduction. Let $\{X_t, t \geqq 0\}$ be a sample-continuous second order martingale. Then $\{X_t^2, t \geqq 0\}$ is a sample-continuous first order submartingale and the conditions for Meyer's decomposition [1] are always satisfied; thus we can write

$$X_t^2 = M_t + A_t, \qquad t \geqq 0, \tag{1}$$

where M is a martingale, A is an increasing process, and both are sample-continuous. The decomposition is unique if we set $M_0 = X_0^2$. If X is a local martingale [2], (1) remains valid except now M is a local martingale. Following Kunita and Watanabe [2], we shall adopt the suggestive notation $\langle X \rangle_t$ for the increasing process A_t.

In this paper, we present two related results. First, we shall obtain a sufficient condition under which a second order sample-continuous martingale can be represented as a stochastic integral in terms of a Brownian motion. Second, we shall show that if X and Y are sample-continuous local martingales (not necessarily with respect to the same family of σ-algebras) and if either $X + Y$ or $X - Y$ is almost surely of bounded variation, then $\langle X \rangle_t = \langle Y \rangle_t$. This rather simple result has some surprising consequences.

2. Martingales and stochastic integrals. Let $(\Omega, \mathscr{A}, \mathscr{P})$ be a probability space, and let $\{\mathscr{A}_t, t \geqq 0\}$ be an increasing family of sub-σ-algebras. A process $\{X_t, t \geqq 0\}$ is said to be *adapted* to $\{\mathscr{A}_t\}$ if, for each t, X_t is $\mathscr{A}_t$-measurable. We say that $\{X_t, \mathscr{A}_t\}$ is a *martingale* if X is adapted to $\{\mathscr{A}_t\}$ and for every $t > s$,

$$E^{\mathscr{A}_s} X_t = X_s \tag{2}$$

almost surely. If $\{X_t, \mathscr{A}_t\}$ is a sample-continuous second order martingale, then the increasing process $\langle X \rangle_t$ introduced earlier is well-defined and $E\langle X \rangle_t < \infty$.

A nonnegative random variable τ is said to be a *stopping time* of $\{\mathscr{A}_t\}$ if $\{\omega : \tau(\omega) \leqq t\} \in \mathscr{A}_t$ for every t. A process $\{X_t, \mathscr{A}_t\}$ is said to be a *local martingale* if there exists an increasing sequence of stopping times $\{\tau_n\}$ such that $\tau_n \uparrow \infty$

* Received by the editors September 15, 1970, and in revised form January 20, 1971.

† Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720. This work was supported by the United States Army Research Office, Durham, under Contract DAHC04-67-0046 and the National Science Foundation under Grant GK-1065X.

almost surely and $\{X_{\min(t,\tau_n)}, \mathscr{A}_t\}$ is a second order martingale. Let $t \wedge s$ denote $\min(t, s)$ and set

$$X_{n,t} = X_{t \wedge \tau_n}.$$

Kunita and Watanabe have proved [2] that if $\{X_t, \mathscr{A}_t\}$ is a sample-continuous local martingale, then there exists a sample-continuous increasing process $\langle X \rangle_t$ such that

$$\langle X \rangle_{t \wedge \tau_n} = \langle X_n \rangle_t.$$

Of course, now we need not have $E\langle X \rangle_t < \infty$.

If $\{W_t, \mathscr{A}_t\}$ is a sample-continuous second order martingale such that for $t > s$,

$$E^{\mathscr{A}_s}(W_t - W_s)^2 = t - s, \tag{3}$$

then W is necessarily a Brownian motion and for each s, $\{W_t - W_s, t \geqq s\}$ is independent of $\mathscr{A}_s$ [3, p. 384]. We describe this situation by saying that $\{W_t, \mathscr{A}_t\}$ is a Brownian motion. Let $\{W_t, \mathscr{A}_t\}$ be a Brownian motion and let $\{\phi_t, t \geqq 0\}$ be a measurable process adapted to $\{\mathscr{A}_t\}$ such that

$$\int_0^t E\phi_s^2 \, ds < \infty \tag{4}$$

for each t. The stochastic integral $\int_0^t \phi_s \, dW_s$ is well-defined as the quadratic limit of a sequence of sums $\sum_v \phi_{t_v^{(n)}}[W_{t_{v+1}^{(n)}} - W_{t_v^{(n)}}]$, where $\{t_v^{(n)}\}$ is a sequence of partitions of $[0, t]$ such that

$$\max_v (t_{v+1}^{(n)} - t_v^{(n)}) \xrightarrow[n \to \infty]{} 0.$$

If we define

$$X_t = \int_0^t \phi_s \, dW_s \tag{5}$$

and choose a separable version for X, then $\{X_t, \mathscr{A}_t\}$ is a second order sample-continuous martingale, with

$$E^{\mathscr{A}_s}(X_t - X_s)^2 = \int_s^t E^{\mathscr{A}_s}\phi_\tau^2 \, d\tau, \qquad 0 \leqq s \leqq t. \tag{6}$$

If, instead of (4), ϕ merely satisfies

$$\int_0^t \phi_s^2 \, ds < \infty, \qquad \text{a.s.}, \tag{7}$$

then $\int_0^t \phi_s \, dW_s$ can be defined as follows: Let $\tau_n(\omega)$ be defined by

$$\tau_n(\omega) = \begin{cases} \inf\left\{t : \int_0^t \phi_s^2(\omega) \, ds \geqq n\right\} \\ \infty \quad \text{if } \int_0^t \phi_s^2(\omega) \, ds < n \quad \text{for all } t, \end{cases} \tag{8}$$

and set

$$\phi_{n,s}(\omega) = \begin{cases} \phi_s(\omega), & s \leqq \tau_n(\omega), \\ 0, & s > \tau_n(\omega). \end{cases} \tag{9}$$

For each n, $\int_0^t \phi_{n,s}\, dW_s$ is well-defined. It can be shown that $\int_0^t \phi_{n,s}\, dW_s$ converges in probability as $n \to \infty$, and we define $\int_0^t \phi_s\, dW_s$ as this limit. Now, the process $X_t = \int_0^t \phi_s\, dW_s$ need no longer be second order or a martingale, but it is still sample-continuous if a separable version is chosen. Moreover, for each n, $\{X_{t \wedge \tau_n}, \mathscr{A}_t\}$ is a second order martingale. By definition, X is a local martingale.

If X is a stochastic integral of the form

$$X_t = X_0 + \int_0^t \phi_s\, dW_s \tag{10}$$

and f is a twice continuously differentiable function of a real variable, then Ito's differentiation formula [4] yields

$$f(X_t) = f(X_0) + \int_0^t f'(X_s)\phi_s\, dW_s + \frac{1}{2}\int_0^t f''(X_s)\phi_s^2\, ds. \tag{11}$$

In particular,

$$X_t^2 - X_0^2 = 2\int_0^t X_s\phi_s\, dW_s + \int_0^t \phi_s^2\, ds. \tag{12}$$

We can now identify $X_0^2 + 2\int_0^t X_s\phi_s\, dW_s$ as the local martingale term in the decomposition (1) of X_t^2, and thus the increasing process is given by

$$\langle X \rangle_t = \int_0^t \phi_s^2\, ds. \tag{13}$$

If X is of the form (10), then we can define the stochastic integral $\int_0^t \psi_s\, dX_s$ by

$$\int_0^t \psi_s\, dX_s = \int_0^t \psi_s\phi_s\, dW_s,$$

provided that $\int_0^t \psi_s^2\phi_s^2\, ds < \infty$ almost surely. If X is a local martingale, not necessarily of the form (10), the stochastic integral $\int_0^t \phi_s\, dX_s$ can still be defined, provided that $\int_0^t \phi_s^2\, d\langle X \rangle_s < \infty$ almost surely [2]. Finally, if $X = Y + Z$ where Y is a local martingale and $\{Z_t, t \geqq 0\}$ is a process with sample functions almost surely of bounded variation, then we can define $\int_0^t \psi_s\, dX_s$ by

$$\int_0^t \psi_s\, dX_s = \int_0^t \psi_s\, dY_s + \int_0^t \psi_s\, dZ_s,$$

provided that the first integral exists as a stochastic integral and the second as a Stieltjes integral. For the case where Y and Z are sample continuous and f is a

twice continuously differentiable function, Kunita and Watanabe has extended Ito's differentiation rule to read

$$f(X_t) = f(X_0) + \int_0^t f'(X_s)\,dX_s + \frac{1}{2}\int_0^t f''(X_s)\,d\langle Y\rangle_s.$$

In particular, we note that

$$X_t^2 - X_0^2 - 2\int_0^t X_s\,dX_s = \langle Y\rangle_t$$

is independent of Z and serves to define $\langle X\rangle_t$.

3. Representation of martingales. Not every sample-continuous second order martingale can be represented as a stochastic integral in the form of (10). It is clear from (13) that for such a representation to be possible the increasing process $\langle X\rangle(\omega, t)$ must be an absolutely continuous function of t (w.r.t. the Lebesgue measure) for almost all ω. As Fisk has observed [5], this condition is also sufficient by virtue of a theorem of Doob [3, p. 449], but it may be necessary to enlarge the underlying probability space by the adjunction of a Brownian motion. Specifically, Doob proved the following theorem.

THEOREM 3.1 (Doob). *Let $\{X_t, \mathscr{A}_t, 0 \leqq t \leqq T\}$ be a sample-continuous second order martingale. Suppose that there exists a nonnegative measureable process $\{\psi_t, 0 \leqq t \leqq T\}$ adapted to $\{\mathscr{A}_t\}$ such that for $t > s$,*

$$E^{\mathscr{A}_s}(X_t - X_s)^2 = \int_s^t E^{\mathscr{A}_s}\psi_\tau\,d\tau. \tag{14}$$

If the set $\{(\omega, t): \psi(\omega, t) = 0\}$ has zero $d\mathscr{P}\,dt$ measure, then there exists a Brownian motion $\{W_t, \mathscr{A}_t, 0 \leqq t \leqq T\}$ such that

$$X_t = X_0 + \int_0^t \psi_s^{1/2}\,dW_s \tag{15}$$

with probability 1. *Without the hypothesis that ψ vanishes almost nowhere, representation* (15) *is still valid with the adjunction of a Brownian motion to the probability space.*

The condition that $\langle X\rangle$ be almost surely continuous with respect to the Lebesgue measure is both somewhat stringent and difficult to verify. Perhaps, it is more natural to consider representations of a more general form

$$X_t(\omega) = X_0(\omega) + \int_0^t \phi_s(\omega)\,dW_{F(s)}(\omega), \tag{16}$$

where W is a Brownian motion and F is a continuous increasing function.

THEOREM 3.2. *Let $\{X_t, \mathscr{A}_t, 0 \leqq t \leqq T\}$ be a sample-continuous local martingale. We assume that $\{\mathscr{A}_t\}$ is right-continuous (i.e. $\bigcap_{s>t}\mathscr{A}_s = \mathscr{A}_t$) and each $\mathscr{A}_t$ is completed. A representation of the form* (16) *exists if and only if $\langle X\rangle$ is absolutely continuous with respect to F with probability* 1.

Proof. We only need to prove the theorem for the case $F(t) = t$, because $\{X_t, \mathscr{A}_t\}$ can be transformed into a sample-continuous local martingale $\{\tilde{X}_t, \tilde{\mathscr{A}}_t\}$ with $E\tilde{X}_t^2 = t$ by defining

$$F^{-1}(t) = \inf\{s : F(s) = t\},$$
$$\tilde{X}_t = X_{F^{-1}(t)},$$
$$\tilde{\mathscr{A}}_t = \mathscr{A}_{F^{-1}(t)}.$$

Even though F^{-1} may be discontinuous, $\{\tilde{\mathscr{A}}_t\}$ is right-continuous and $\tilde{X}$ is still sample-continuous, because $F(t) = F(s)$ implies $X_t = X_s$ almost surely. Since $X_t = \tilde{X}_{F(t)}$ with probability 1, a representation

$$\tilde{X}_t = \tilde{X}_0 + \int_0^t \tilde{\phi}_s \, dW_s,$$

where $\{W_t, \tilde{\mathscr{A}}_t\}$ is a Brownian motion, implies a representation

$$X_t = \tilde{X}_{F(t)} = X_0 + \int_0^{F(t)} \tilde{\phi}_s \, dW_s$$
$$= X_0 + \int_0^t \phi_s \, dW_{F(s)},$$

which is just (16).

To prove Theorem 3.2 for the case $F(t) = t$, we first note that necessity follows from (13). To prove sufficiency, we assume that $\langle X \rangle$ is absolutely continuous with probability 1 and write

$$\langle X \rangle_t(\omega) = \int_0^t \psi_s(\omega) \, ds, \tag{17}$$

where ψ can always be chosen to be a measurable process because $\langle X \rangle$ is a measurable process. For each t, ψ_t is measurable with respect to $\bigcap_{s>t} \mathscr{A}_s$, which is equal to $\mathscr{A}_t$ by assumption.

We now follow Doob [3, p. 449] and define

$$g_s(\omega) = \begin{cases} \psi_s^{-1/2}(\omega) & \text{if } \psi_s(\omega) > 0, \\ 0 & \text{if } \psi_s(\omega) = 0. \end{cases} \tag{18}$$

Since $\int_0^t g_s^2 \, d\langle X \rangle_s \leqq t < \infty$, the integral $\int_0^t g_s \, dX_s = W_t$ is well-defined as a local martingale. If $\psi_s(\omega) > 0$ for almost all (ω, s), then $\langle W \rangle_t = t$ and it follows from Theorem 2.3 of [2] that W is a Brownian motion. If not, we adjoin an independent Brownian motion B to the underlying probability space and define

$$W_t = \int_0^t g_s \, dX_s + \int_0^t [1 - g_s \sqrt{\psi_s}] \, dB_s.$$

Then, W is a Brownian motion. In either case, we have

$$X_t = X_0 + \int_0^t \sqrt{\psi_s} \, dW_s$$

and the proof is complete.

Theorem 3.2 is basically the same as Theorem 2.1 of [5], except for the introduction of the increasing function F and the generalization to local martingales. We now come to the main result of the section, namely, a sufficient condition for the representation (16) that can be verified in terms of two-dimensional distributions of a martingale X.

THEOREM 3.3. *Let $\{X_t, \mathscr{A}_t, 0 \leqq t \leqq T\}$ be a sample-continuous second order martingale and let $F(t) = E(X_t - X_0)^2 = E\langle X\rangle_t$. Suppose that there exist finite positive constants α and β such that*

$$\sup_{0 < F(t) - F(s) \leqq \beta} \frac{E|X_t - X_s|^{2+2\alpha}}{[F(t) - F(s)]^{1+\alpha}} < \infty. \tag{19}$$

Then $\langle X\rangle$ is almost surely absolutely continuous with respect to the F-measure, and X has a representation of the form of (16).

Proof. By virtue of the Lebesgue decomposition, we can always write

$$\langle X\rangle_t(\omega) = \int_0^t \psi_s(\omega)\, dF(s) + \mu_t(\omega), \tag{20}$$

where μ is almost surely singular with respect to F. Now, $E\psi_s = 1$ implies

$$E\mu_t = -\int_0^t dF(t) + E\langle X\rangle_t = 0,$$

which in turn implies that $\mu \equiv 0$ almost surely since μ is nonnegative and sample-continuous. Therefore, we only need to prove that (19) implies $E\psi_s = 1, 0 \leqq s \leqq T$.

Let $T_n = \{t_\nu^{(n)}, \nu = 0, 1, \cdots, n\}$ be a sequence of nested (i.e., $T_{n+1} \supset T_n$) partitions of the interval $[0, T]$ such that

$$\max_\nu [F(t_{\nu+1}^{(n)}) - F(t_\nu^{(n)})] \xrightarrow[n\to\infty]{} 0.$$

Define $\psi_{n,t}$, $0 \leqq t \leqq T$, as follows:

$$\psi_{n,t} = \frac{\langle X\rangle_{t_{\nu+1}^{(n)}} - \langle X\rangle_{t_\nu^{(n)}}}{F(t_{\nu+1}^{(n)}) - F(t_\nu^{(n)})}, \qquad t_\nu^{(n)} \leqq t < t_{\nu+1}^{(n)}. \tag{21}$$

It is well known (see, e.g., [3, pp. 346–347]) that for each ω, $\psi_{n,t}$ converges for almost all t (F-measure) to the Radon-Nikodym derivative of the absolutely continuous component of $\langle X\rangle$ with respect to F. That is, $\psi_{n,t} \to \psi_t$ for almost all (ω, t). Since it is obvious that $E\psi_{n,t} = 1$, the desired result $E\psi_t = 1$ will follow if for each t, $\{\psi_{n,t}\}$ is a uniformly integrable family of random variables.

Now, it is known [6] that for any $p > 1/2$ there exists a constant κ_p such that

$$E|\langle X\rangle_t - \langle X\rangle_s|^p \leqq \kappa_p E|X_t - X_s|^{2p}.$$

Therefore, if we let N be the smallest n such that

$$\max_\nu [F(t_{\nu+1}^{(n)}) - F(t_\nu^{(n)})] \leqq \beta,$$

then

$$\sup_{n \geqq N} E\psi_{n,t}^{1+\alpha} \leqq \kappa_{1+\alpha} \sup_{0 < F(t) - F(s) \leqq \beta} \left\{ \frac{E|X_t - X_s|^{2+2\alpha}}{[F(t) - F(s)]^{1+\alpha}} \right\} < \infty,$$

so that $\{\psi_{n,t}\}$ is a uniformly integrable family of random variables. This, together with Theorem 3.2, completes the proof.

Theorem 3.2 is reminiscent of Kolmogorov's condition for sample continuity and has similar advantages, the primary one being that it can be verified in terms of the two-dimensional distributions of X.

4. Quadratic variation. Let $T_n = \{t_\nu^{(n)}\}$ be a nested sequence of partitions of $[0, T]$ such that $\max_\nu (t_\nu^{(n)} - t_{\nu+1}^{(n)}) \xrightarrow[n\to\infty]{} 0$. Let X be a sample-continuous second order martingale, and let $t \wedge s$ denote $\min(t, s)$. Fisk [5] has shown that the sequence of sums

$$Q_n(X, t) = \sum_\nu [X_{t \wedge t_{\nu+1}^{(n)}} - X_{t \wedge t_\nu^{(n)}}]^2 \tag{22}$$

converges to $\langle X\rangle_t$ in L^1-mean, i.e.,

$$E|Q_n(X, t) - \langle X\rangle_t| \xrightarrow[n\to\infty]{} 0. \tag{23}$$

For this reason, $\langle X\rangle_t$ is said to be the quadratic variation of X on $[0, t]$. Now, suppose that $\{Z_t, 0 \leqq t \leqq T\}$ is a sample-continuous process, the sample functions of which are almost surely of bounded variation. Then there exists an almost surely finite random variable A such that

$$\sup_n \sum_\nu |Z_{t_{\nu+1}^{(n)}} - Z_{t_\nu^{(n)}}| \leqq A.$$

Therefore,

$$Q_n(Z, T) = \sum_\nu |Z_{t_{\nu+1}^{(n)}} - Z_{t_\nu^{(n)}}|^2 \leqq A \max_\nu |Z_{t_{\nu+1}^{(n)}} - Z_{t_\nu^{(n)}}|$$

$$\xrightarrow[n\to\infty]{\text{a.s.}} 0$$

so that Z has zero quadratic variation on $[0, T]$.

THEOREM 4.1. *Let $\{\mathscr{A}_t\}$ and $\{\tilde{\mathscr{A}}_t\}$ be two increasing families of σ-algebras and let $\{X_t, \mathscr{A}_t\}$ and $\{\tilde{X}_t, \tilde{\mathscr{A}}_t\}$ be sample-continuous local martingales. If $X + \tilde{X}$ or $X - \tilde{X}$ is of bounded variation, then $\langle X\rangle_t = \langle \tilde{X}\rangle_t$ almost surely for every t.*

Proof. First, suppose that X and $\tilde{X}$ are second order martingales. Let B denote the process of bounded variation given by $X - \tilde{X}$ or $X + \tilde{X}$. From the inequality

$$\left(\sqrt{\sum_k a_k^2} - \sqrt{\sum_k b_k^2}\right)^2 \leqq \sum_k (a_k \pm b_k)^2 \leqq \left(\sqrt{\sum_k a_k^2} + \sqrt{\sum_k b_k^2}\right)^2,$$

we have

$$\left(\sqrt{Q_n(\tilde{X}, t)} - \sqrt{Q_n(B, t)}\right)^2 \leqq Q_n(X, t) \leqq \left(\sqrt{Q_n(\tilde{X}, t)} + \sqrt{Q_n(B, t)}\right)^2.$$

Since $Q_n(\tilde{X}, t)$, $Q_n(X, t)$ and $Q_n(B, t)$ converge respectively to $\langle \tilde{X}\rangle_t$, $\langle X\rangle_t$ and 0, we have $\langle X\rangle_t = \langle \tilde{X}\rangle_t$ almost surely for all $t \in [0, T]$.

If X and $\tilde{X}$ are local martingales, then there exist sequences of stopping times $\{\tau_m\}$ and $\{\tilde{\tau}_m\}$ increasing to ∞, so that $X_{m,t} = X_{\tau_m \wedge t}$ and $\tilde{X}_{m,t} = \tilde{X}_{\tau_m \wedge t}$ are second order martingales for each m. Let

$$\Omega_{m,t} = \{\omega : t < \tau_m \wedge \tilde{\tau}_m\}.$$

Since $\tau_m \uparrow \infty$ and $\tilde{B}_m \uparrow \infty$, we have $\Omega_{m,t} \uparrow \Omega$. For $\omega \in \Omega_{m,t}$, we have

$$X_{m,s}(\omega) = X_s(\omega), \qquad 0 \leqq s \leqq t,$$
$$\tilde{X}_{m,s}(\omega) = \tilde{X}_s(\omega), \qquad 0 \leqq s \leqq t.$$

Therefore, for $\omega \in \Omega_{m,t}$, either $X_m(\omega) + \tilde{X}_m(\omega)$ or $X_m(\omega) - \tilde{X}_m(\omega)$ is of bounded variation on $[0, t]$, and it follows from the same inequality as before that

$$\langle X_m \rangle_t(\omega) = \langle \tilde{X}_m \rangle_t(\omega), \qquad \omega \in \Omega_{m,t},$$

or

$$\langle X \rangle_t(\omega) = \langle \tilde{X} \rangle_t(\omega), \qquad \omega \in \Omega_{m,t}.$$

The proof is completed by letting $m \uparrow \infty$.

We should note that in Theorem 4.1 we do not assume that X and $\tilde{X}$ are local martingales with respect to the same family of σ-algebras. If they are, then $X + \tilde{X}$ and $X - \tilde{X}$ are both local martingales. If one of them is also of bounded variation, say $X + X$, then $X_t + \tilde{X}_t = X_0 + \tilde{X}_0$ with probability 1 for all $t \in [0, T]$, in which case the result of Theorem 4.1 trivially follows. The more interesting cases arise when neither $X + \tilde{X}$ nor $X - \tilde{X}$ is a local martingale.

An interesting application of Theorem 4.1 is in connection with quasi-martingales [7]. A process $\{X_t, 0 \leqq t \leqq T\}$ is said to be a quasi-martingale with respect to $\{\mathscr{A}_t\}$ if there exist $\{B_t, 0 \leqq t \leqq T\}$ and $\{M_t, 0 \leqq t \leqq T\}$ both adapted to $\{\mathscr{A}_t\}$ such that $X = M + B$, B is of bounded variation, and $\{M_t, \mathscr{A}_t\}$ is a martingale. We shall be interested only in those cases where both M and B are sample-continuous and where the total variation of B has a finite expectation. Under these assumptions, if $\{X_t, \mathscr{A}_t\}$ is a quasi-martingale, then $\{X_t, \mathscr{A}_{xt}\}$ is always a quasi-martingale, where $\mathscr{A}_{xt}$ denotes the σ-algebra generated by $\{X_s, 0 \leqq s \leqq t\}$. More generally, let $\{\mathscr{A}_t\}$ be any increasing family of σ-algebras such that $\mathscr{A}_{xt} \subseteq \tilde{\mathscr{A}}_t \subseteq \mathscr{A}_t$ for every t. Then, $\{X_t, \tilde{\mathscr{A}}_t\}$ is a quasi-martingale.[1] That is, there exist $\tilde{B}$ and $\tilde{M}$ adapted to $\{\mathscr{A}_t\}$ such that B is of bounded variation. $\{\tilde{M}_t, \tilde{\mathscr{A}}_t\}$ is a martingale, both $\tilde{M}$ and $\tilde{B}$ are sample-continuous, and the total variation of $\tilde{B}$ has finite expectation. It follows from Theorem 4.1 that $\langle M \rangle_t = \langle \tilde{M} \rangle_t$ for every t.

An important class of quasi-martingales is made up of Ito processes, which are processes having the representation

$$X_t = X_0 + \int_0^t \psi_s \, ds + \int_0^t \phi_s \, dW_s, \qquad 0 \leqq t \leqq T, \tag{24}$$

where ψ and ϕ are measurable processes adapted to an increasing family of σ-algebras $\{\mathscr{A}_t\}$, $\{W_t, \mathscr{A}_t\}$ is a Brownian motion, and X_0 is $\mathscr{A}_0$-measurable. In addition, we assume

$$\int_0^T E|\psi_s| \, ds < \infty, \tag{25}$$

$$\int_0^T E\phi_s^2 \, ds < \infty. \tag{26}$$

[1] This fact is easily proved by verifying the two conditions of Theorem 3.3 in [7].

It is clear that $\{X_t, \mathscr{A}_t\}$ is a quasi-martingale. Thus, $\{X_t, \mathscr{A}_{xt}\}$ is also a quasi-martingale and the representation of X as a quasi-martingale with respect to $\{\mathscr{A}_{xt}\}$ is given in the following theorem.

THEOREM 4.2. *Let* X_t, $0 \leqq t \leqq T$, *be an Ito process satisfying* (24), (25) *and*

$$\int_0^T \phi_s^2 \, ds < \infty \qquad \text{almost surely.} \tag{27}$$

Then there exists a representation of X *in the form*

$$X_t = X_0 + \int_0^t E^{\tilde{\mathscr{A}}_s} \psi_s \, ds + \int_0^t |\psi_s| \, d\tilde{W}_s, \tag{28}$$

where $\{\tilde{\mathscr{A}}_t\}$ *is any increasing family of* σ*-algebras satisfying*

$$\mathscr{A}_{xt} \subseteq \tilde{\mathscr{A}}_t \subseteq \mathscr{A}_t \quad \text{for each } t \tag{29}$$

and $\tilde{W}$ *is a Brownian motion.*

Remark. If we take $\tilde{\mathscr{A}}_t = \mathscr{A}_{xt}$, then we have

$$X_t = X_0 + \int_0^t E^{\mathscr{A}_{xs}} \psi_s \, ds + \int_0^t |\phi_s| \, d\tilde{W}_s,$$

where both integrands $E^{\mathscr{A}_{xs}} \psi_s$ and $|\phi_s|$ are $\mathscr{A}_{xs}$-measurable functions for almost all s. The latter stems from the fact that

$$|\phi_t| = \left[\frac{d}{dt} \langle X \rangle_t \right]^{1/2}, \tag{30}$$

where d/dt stands for the Radon-Nikodym derivative and $\langle X \rangle_t$ is defined by

$$\langle X \rangle_t = X_t^2 - X_0^2 - 2 \int_0^t X_s \, dX_s. \tag{31}$$

Proof. First, suppose that, in place of (27), the stronger condition (26) is satisfied. Then, $\{X_t, \tilde{\mathscr{A}}_t\}$ is a quasi-martingale and we can write

$$X_t = X_0 + \tilde{B}_t + \tilde{\mathscr{A}}_t, \qquad 0 \leqq t \leqq T,$$

where $\tilde{B}$ is of bounded variation and $\{\tilde{M}_t, \tilde{\mathscr{A}}_t\}$ is a martingale. Since

$$\tilde{M}_t - \int_0^t \phi_s \, dW_s = \int_0^t \psi_s \, ds - \tilde{B}_t$$

and $M_t = \int_0^t \phi_s \, dW_s$ is a martingale under condition (26), we have from Theorem 4.1:

$$\langle \tilde{M} \rangle_t = \langle M \rangle_t = \int_0^t \phi_s^2 \, ds, \qquad 0 \leqq t \leqq T;$$

Theorem 3.1 yields the representation

$$\langle \tilde{M} \rangle_t = \int_0^t |\phi_s| \, d\tilde{W}_s,$$

where $\tilde{W}$ is a Brownian motion.

To get an integral representation for $\tilde{B}$, let $T_n = \{t_\nu^{(n)}\}$, $n = 1, 2, \cdots$, be a nested sequence of partitions of $[0, T]$ such that $\max_\nu (t_{\nu+1}^{(n)} - t_\nu^{(n)}) \xrightarrow[n\to\infty]{} 0$. Define

$$\tilde{B}_{n,t} = \sum_\nu E[X_{t\wedge t_{\nu+1}^{(n)}} - X_{t\wedge t_\nu^{(n)}} | \tilde{\mathscr{A}}_{t_\nu^{(n)}}].$$

Fisk [7] has shown that

$$E|\tilde{B}_{n,t} - \tilde{B}_t| \xrightarrow[n\to\infty]{} 0.$$

From (24) we get

$$E[X_{t_{\nu+1}^{(n)}} - X_{t_\nu^{(n)}} | \tilde{\mathscr{A}}_{t_\nu^{(n)}}] = \int_{t_\nu^{(n)}}^{t_{\nu+1}^{(n)}} E(\psi_s | \tilde{\mathscr{A}}_{t_\nu^{(n)}})\, ds.$$

If we denote $f_s = E^{\tilde{\mathscr{A}}_s}\psi_s$, then

$$\begin{aligned}
&\sup_{0\leqq t\leqq T} E\left|\int_0^t E^{\tilde{\mathscr{A}}_s}\psi_s\, ds - \tilde{B}_t\right| \\
&\leqq \lim_{n\to\infty}\left\{E\left|\sum_\nu \int_{t_\nu^{(n)}}^{t_{\nu+1}^{(n)}} [f_s - f_{t_\nu^{(n)}}]\, ds\right|\right. \\
&\qquad \left. + E\left|\sum_\nu \int_{T_\nu^{(n)}}^{t_{\nu+1}^{(n)}} E[\psi_s - \psi_{t_\nu^{(n)}} | \tilde{\mathscr{A}}_{t_\nu^{(n)}}]\, ds\right|\right\} \\
&\leqq \lim_{n\to\infty} E\sum_\nu \int_{t_\nu^{(n)}}^{t_{\nu+1}^{(n)}} [|f_s - f_{t_\nu^{(n)}}| + |\psi_s - \psi_{t_\nu^{(n)}}|]\, ds.
\end{aligned}$$

For a suitable sequence of partitions $\{t_\nu^{(n)}\}$, the last limit can always be made zero (Doob [3, pp. 63–65]). Since the first expression is independent of $\{t_\nu^{(n)}\}$, we must have

$$\sup_{0\leqq t\leqq T} E\left|\int_0^t E^{\tilde{\mathscr{A}}_s}\psi_s\, ds - \tilde{B}_t\right| = 0.$$

If ϕ satisfies (27) and not (26), we define $\{\tau_n\}$ by (8). Since $\langle X\rangle_t = \int_0^t \psi_s^2\, ds$, $\{\tau_n\}$ is a sequence of stopping times not only for $\{\mathscr{A}_t\}$ but also for $\{\tilde{\mathscr{A}}_t\}$. Therefore, for each n, $\{X_{t\wedge\tau_n}, \tilde{\mathscr{A}}_t\}$ is a quasi-martingale with a representation

$$X_{t\wedge\tau_n} = X_0 + \int_0^{t\wedge\tau_n} E^{\tilde{\mathscr{A}}_s}\psi_s\, ds + \tilde{M}_{t\wedge\tau_n}.$$

Hence, $\tilde{M}_t = X_t - X_0 - \int_0^t E^{\tilde{\mathscr{A}}_s}\psi_s\, ds$ is a local martingale. It follows from Theorem 4.1 and (24) that

$$\langle \tilde{M}\rangle_t = \int_0^t \phi_s^2\, ds,$$

and the proof is completed by using Theorem 3.2.

Theorem 4.2 can be generalized to a vector Ito process practically without change. Let X be an n-vector-valued process satisfying

$$X_t = X_0 + \int_0^t \psi_s\, ds + \int_0^t \phi_s\, dW_s, \tag{32}$$

where W is an m-vector process, the components of which are independent Brownian motions and ϕ is an $n \times m$ matrix. Instead of (25) and (27), we now assume

$$\int_0^T E\|\psi_s\|\, ds < \infty \tag{33}$$

and

$$\int_0^T \|\phi_s\|^2\, ds < \infty \qquad \text{almost surely,} \tag{34}$$

where $\|\cdot\|$ denotes the Euclidean norm.

Let α be an n-vector and let prime denote transpose. For every $\alpha \in R^n$, $\{\alpha' X_t\}$ is a scalar-valued Ito process of the form (24), and we can write

$$\alpha' X_t = \int_0^t \alpha' E^{\tilde{\mathscr{A}}_s}\psi_s\, ds + M_{\alpha,t},$$

where $\{M_{\alpha,t}, \tilde{\mathscr{A}}_t\}$ is a local martingale for any $\{\tilde{\mathscr{A}}_t\}$ satisfying (29) and

$$\langle M_\alpha \rangle_t = \int_0^t \alpha'\phi_s\phi_s'\alpha\, ds.$$

It follows that $\int_0^t \phi_s\phi_s'\, ds$ is $\mathscr{A}_{xt}$-measurable for every t, and the positive semi-definite matrix $\phi_s\phi_s'$ is $\mathscr{A}_{xs}$-measurable for almost all s. Let $\phi_s\phi_s'$ be diagonalized so that

$$\phi_s\phi_s' = A_s\Lambda_s A_s', \tag{35}$$

where Λ_s is diagonal and the orthogonal matrix A_s is $\mathscr{A}_{xt}$-measurable for almost all s. Now, if we define

$$Y_t = \int_0^t A_s'\, dX_s$$

and apply Theorem 4.2 to the components of Y, we get the following theorem.

THEOREM 4.3. *Let X be an n-vector Ito process of the form* (32) *satisfying* (33) *and* (34). *Let A_s and Λ_s be matrices defined by the diagonalization* (35). *Then, for any increasing family of σ-algebras satisfying* (29), *there exists an n-vector Brownian motion $\tilde{W}$ so that*

$$X_t = X_0 + \int_0^t E^{\mathscr{A}_{xs}}\psi_s\, ds + \int_0^t A_s\Lambda_s^{1/2}\, d\tilde{W}_s. \tag{36}$$

5. Applications. Equation (32) is widely used to model a dynamical system disturbed by Gaussian white noise. Theorem 4.3 has some interesting and surprising consequences for such models. For example, in filtering problems we often interpret X in (32) as the observed process, ψ as the process to be estimated, and the stochastic integral term as the noise. If we identify $\hat{\psi}_t = E^{\mathscr{A}_{xt}}\psi_t$ as the estimator, then Theorem 4.3 implies that

$$\int_0^t (\psi_s - \hat{\psi}_s)\, ds = -\int_0^t \phi_s\, dW_s + \int_0^t A_s\Lambda_s^{1/2}\, d\tilde{W}_s.$$

The stochastic integral $-\int_0^t \phi_s\, dW_s$ can always be re-represented as $\int_0^t A_s \Lambda_s^{1/2}\, dV_s$, where V is now an n-dimensional Brownian motion, so that

$$\int_0^t (\psi_s - \hat{\psi}_s)\, ds = \int_0^t A_s \Lambda_s^{1/2} (dV_s + d\tilde{W}_s).$$

Formally, the estimation error can be expressed as

$$\psi_t - \hat{\psi}_t = A_t \Lambda_t^{1/2} (\dot{V}_t + \dot{\tilde{W}}_t).$$

One may be tempted to say that the estimation error is "white," but that would be misleading. For the special case $\phi(\omega, t) = K$, a constant, we have

$$\psi_t - \hat{\psi}_t = A \Lambda^{1/2} (\dot{V}_t + \dot{\tilde{W}}_t),$$

which is the sum of Gaussian white noise (albeit dependent) processes. This was originally observed by Wonham [8] in a more limited context and generalized by Kailath [9]. We note that the same observation is true if $\phi\phi'$ is a constant even if ϕ is not.

Another observation worthy of note is that for almost all t, $\phi_t \phi_t'$ is $\mathscr{A}_{xt}$-measurable. Thus, for example, if ψ_t is a function only of t and $\{\phi_s \phi_s', 0 \leqq s \leqq t\}$, then the estimation error is necessarily zero for almost all t. Such a result would be difficult to prove outside of the context of quadratic variations. A similar observation can be made with respect to singular detection [10].

Finally, as a simple example of possible applications to control problems, consider a scalar equation

$$dX_t = f(X_0^t, U_0^t, t)\, dk + U_t\, dW_t,$$

where X represents the state and U the control. Suppose that $U_t \geqq 0$ for all t. Then Theorem 4.2 implies that for any optimization problem whatever, the optimizing control can always be implemented in state feedback form. This means that even if we observe the past of the noise process W and use it in constructing the control, performance cannot be improved. Surely, this is an unexpected result.

Acknowledgment. I am grateful to Professor P. P. Varaiya for many valuable suggestions. In particular, it was at his persuasion and with his help that I extended the result of § 4 to local martingales and to the multidimensional case.

REFERENCES

[1] P. A. Meyer, *A decomposition theorem for supermartingales*, Illinois J. Math., 6 (1962), pp. 193–205.

[2] H. Kunita and S. Watanabe, *On square integrable martingales*, Nagoya Math. J., 30 (1967), pp. 209–245.

[3] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[4] K. Ito, *On a formula concerning stochastic differentials*, Nagoya Math. J., 3 (1951), pp. 55–65.

[5] D. L. Fisk, *Sample quadratic variation of sample continuous second order martingales*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 6 (1966), pp. 273–278.

[6] P. A. Meyer, *Une majoration du processus croissant naturel associé à une surmartingale*, Seminaire de probabilités, II, Springer-Verlag, Berlin, 1967.

[7] D. L. Fisk, *Quasi-martingales*, Trans. Amer. Math. Soc., 120 (1965), pp. 369–389.

[8] W. M. Wonham, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. 2, A. T. Bharucha-Reid, ed., Academic Press, New York, 1970.

[9] T. KAILATH, *An innovation approach to least-squares estimation. Part I: Linear filtering in additive white noise,* IEEE Trans. Automatic Control, AC-13 (1968), pp. 646–655.

[10] E. WONG AND M. ZAKAI, *The oscillation of stochastic integrals*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 4 (1965), pp. 103–112.