SIAM J. CONTROL
Volume 12, No. 1, February 1974

CONTROLLABILITY OF NONLINEAR SYSTEMS ON COMPACT MANIFOLDS*

C. LOBRY†

Abstract. It is proved that a nonlinear conservative control system on a compact manifold is controllable if (and only if in the analytical case) a certain condition expressed in terms of the "Taylor expansion of the system" at each point is satisfied.

1. Introduction. This paper is a self-contained version of a result in the author's thesis [8]. The general background for this work is developed in papers [3], [6], [10] and [12], and applications to optimal control problems are proposed in [2], [3], [6], [7], [10] and [12].

Denote by $\mathbb{N}$ the set of natural numbers, $\mathbb{R}$ the set of real numbers and by $\mathbb{R}^+$ the set of positive real numbers.

We define a *dynamical polysystem* (abbreviated DP) as a collection of complete vector fields on a smooth, connected, paracompact manifold of dimension n. We denote this collection by $(X^i)_{i\in I}$ and by

$$(x, t) \to X^i_t(x),$$

the dynamical system generated by X^i. The *orbit* of a point x under the action of the DP is the collection of all the points of the form $X^{i_p}_{t_p} \circ \cdots \circ X^{i_j}_{t_j} \circ \cdots \circ X^{i_1}_{t_1}(x)$, with $i_j \in I$, $t_j \in \mathbb{R}$, $p \in \mathbb{N}$. We denote this orbit by $(X^i)_{i\in I}x$. The *positive orbit* of a point x is defined in the same way as the orbit, but we are restricted to positive t_j; we denote

$$(X^i)^+_{i\in I}x = \{X^{i_p}_{t_p} \circ \cdots \circ X^{i_j}_{t_j} \circ \cdots \circ X^{i_1}_{t_1}(x),\ i_j \in I, t_j \in \mathbb{R}^+, p \in \mathbb{N}\}.$$

DEFINITION. A DP is *controllable* if for every point x in the manifold M, the positive orbit of x is the whole manifold.

For general results on orbits, positive orbits and controllability, one can see the references listed above. In a recent paper, V. Jurdjevic [4] proved necessary and sufficient conditions for controllability of an analytical, right invariant dynamical polysystem defined on a compact Lie group. He asked to what extent this result is true for a general compact manifold. The proposition and example of the present paper are partial answers.

2. Definitions and preliminary results. Consider the smallest family of vector fields on M containing the DP $(X^i)_{i\in I}$ closed under the Jacobi bracket operation. At each point of M, the values of the elements of this family are vectors in the tangent space to M which generate a certain linear subspace; we define the rank at a given point of the DP as the dimension of this subspace. The following is proven in [6] as a slight generalization of Chow's result [1] and R. Hermann's ideas [3].

* Received by the editors February 22, 1973.

† U.E.R. de Mathématiques et Informatique, Université de Bordeaux I, Talence, France.

THEOREM 1. *If (and only if in the analytical case) the rank of the* DP *is equal to n (n is the dimension of manifold M) at every point of M, then for every point x in M, the orbit $(X^i)_{i\in I}x$ of x is the whole manifold M.*

The following theorem is proven in [8] (see also [5] for a shorter proof and other related topics) and by H. Sussmann and V. Jurdjevic (see [11]) for the analytical case.

THEOREM 2. *If (and only if in the analytical case) the rank of the* DP *is equal to n at every point of the manifold M, then for every point x in M, the interior points of the positive orbit $(X^i)^+_{i\in I}x$ are dense in the positive orbit.*

Now, let M be a compact Riemannian manifold; a conservative vector field is a vector field for which the natural measure in M is invariant under the action of the dynamical system associated to it. The well-known theorem below is the main trick in the proof of Proposition 1.

THEOREM (*Poincaré*). *The Poisson stable points of a dynamical system generated by a conservative vector field are dense.*

Recall that a point is Poisson stable if and only if for every neighborhood $\mathscr{V}_x$ of x and every positive T, there exist t_1 and t_2, greater than T, such that $X_{t_1}(x)$ and $X_{-t_2}(x)$ are in $\mathscr{V}_x$.

We say that the $\mathrm{DP}(X^i)_{i\in I}$ is conservative if every X^i is a conservative vector field on M.

3. A controllability result.

PROPOSITION 1. *Let $(X^i)_{i\in I}$ be a conservative* DP *on a compact Riemannian manifold M of dimension n. If (and only if in the analytical case) the rank of $(X^i)_{i\in I}$ is n at every point of m, the* DP *is controllable.*

Proof. The "only if" in the analytical case is a trivial consequence of the fact that the positive orbit is included in the orbit and of Theorem 1. Let us prove now the "if." Let x and y be two points in M. We have to prove that there exist some $i_1, i_2, \cdots, i_j, \cdots, i_p$ and $t_1, t_2, \cdots, t_j, \cdots, t_p$ in $\mathbb{R}^+$ such that

$$y = X^{i_p}_{t_p} \circ \cdots \circ X^{i_j}_{t_j} \circ \cdots \circ X^{i_1}_{t_1}(x).$$

Theorem 2 is true if we replace $(X^i)^+_{i\in I}x$ by $(X^i)^-_{i\in I}x$ (with obvious notation); choose a point $\bar{y}$ in the interior of $(X^i)^-_{i\in I}y$ which is not empty by Theorem 2. Let $\mathscr{V}_{\bar{y}}$ be an open neighborhood of $\bar{y}$ contained in $(X^i)^-_{i\in I}y$. By Theorem 1 there exist i_j and τ_j, not necessarily positive, such that

$$\bar{y} = X^{i_q}_{\tau_q} \circ \cdots \circ X^{i_j}_{\tau_j} \circ \cdots \circ X^{i_1}_{\tau_1}(x).$$

Suppose for simplicity (the general case goes by induction exactly in the same way) that

$$\bar{y} = X^{i_2}_{\tau_2} \circ X^{i_1}_{\tau_1}(x),$$

where τ_1 is positive and τ_2 is negative. Let $p = X^{i_1}_{\tau_1}(x)$; then the set

$$\mathscr{V}_p = X^{i_2}_{-\tau_2}(\mathscr{V}_{\bar{y}})$$

is an open neighborhood of p, and it has a nonempty intersection with $(X^i)^+_{i\in I}x$ (using, again, Theorem 2); thus it contains a point z which is Poisson stable for X^{i_2}. Let $\mathscr{V}_z$ be a neighborhood of z contained in $\mathscr{V}_p \cap (X^i)^+_{i\in I}x$; by the definition of

Poisson stable, there exists a real number t, greater than $|\tau_2|$, such that the point $X_t^{i_2}(x)$ belongs to $\mathscr{V}_z$. Then the following hold.

(i) The number $t + \tau_2$ is positive.

(ii) The point $X^{i_2}_{t+\tau_2}(z)$ belongs to $\mathscr{V}_{\bar{y}}$.

The point y is in the positive orbit of any point in $(X^i)^-_{i\in I}(y)$, hence it belongs to

$$(X^i)^+_{i\in I} X^{i_2}_{t+\tau_2}(z);$$

thus the proposition is proved.

4. Final comments.

Example. On the sphere S^2, consider the DP generated by the two vector fields X^1 and X^2 whose trajectories are "meridian" and "longitude" defined for two different axes (see Fig. 1). These two vector fields are analytic; one of them is conservative. From Theorem 1, we see that the DP has rank 2 at every point because its orbits from every point are the whole S^2, as one sees easily on the picture. Conversely, one can see that the positive orbit from the north pole is certainly included in the north hemisphere.

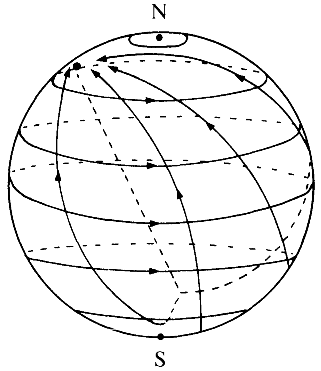

Fig. 1

Remark 1. This example shows that the assumption on the conservativeness of the system is unremovable (we need at least that two vector fields of the family are conservative and satisfy the rank assumption). On the other hand, by a slight modification of the proofs in [6] (see also [7]), one can see that the rank assumption is generic, i.e., the set of conservative DP for which the rank assumption is satisfied is an open dense subset of the set of conservative DP for a reasonable topology.

Remark 2. The proof of the proposition of this paper also works for right invariant vector fields on a compact Lie group as soon as we remark that trajectories of such vector fields are certainly Poisson stable.

Acknowledgment. The result in this paper was suggested to me by Prof. L. W. Markus; the proof is inspired by his paper [9]. I also thank him for many helpful encouragements.

REFERENCES

[1] W. L. Chow, *Über Systeme von linearen partiellen differentialgshungen erster ordnung*, Math. Ann., 117 (1939), pp. 98–105.

[2] G. W. HAYNES AND H. HERMES, *Nonlinear controllability via Lie theory*, this Journal, 8 (1970), pp. 450–460.

[3] R. HERMANN, *On the accessibility problem in control theory*, International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963.

[4] V. JURDJEVIC, *Certain controllability property of analytic control systems*, this Journal, 10 (1972), pp. 354–360.

[5] A. KRENER, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems*, to appear.

[6] C. LOBRY, *Contrôlabilité des systèmes non linéaires*, this Journal, 8 (1970), pp. 573–605.

[7] ———, *Une propriété générique des couples de champs de vecteurs*, Czechoslovak Math. J., 22 (1972), pp. 230–237.

[8] ———, *Quelques aspects qualitatifs de la théorie de la commande*, Thèse Sciences Math., Grenoble, 1972.

[9] L. MARKUS, *Control dynamical system*, Math. Systems Theory, vol. 3, No. 2, pp. 179–185.

[10] H. SUSSMANN, *The bang bang problem for certain control systems in $GL(N, \mathbb{R})$*, this Journal, 10 (1971), pp. 470–476.

[11] ———, *Orbits of families of vector fields and integrability of distributions*, to appear.

[12] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

SIAM J. CONTROL
Vol. 12, No. 1, February 1974

REGULATION AND INTERNAL STABILIZATION IN LINEAR MULTIVARIABLE SYSTEMS*

W. M. WONHAM† AND J. B. PEARSON‡

Abstract. For the multivariable control system described by

$$\dot{x} = Ax + Bu, \quad y = Cx, \quad z = Dx,$$

constructive necessary and sufficient conditions are given for the existence of state feedback $u = Fx$ such that (i) $\operatorname{Ker} F \supset \operatorname{Ker} C$ (observability constraint), (ii) $D \exp[t(A + BF)] \to 0$ as $t \to \infty$ (output regulation), and (iii) any unstable modes of $A + BF$ are either uncontrollable or unobservable at y (internal stability). It is assumed that Ker C is A-invariant, or equivalently that an observer or dynamic compensator is utilized. A common application is treated, and sensitivity is considered for a simple example.

1. Introduction. In this paper we continue the discussion in [1] on output regulation for the system

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{1}$$

$$\begin{aligned} y(t) &= Cx(t), \\ z(t) &= Dx(t). \end{aligned} \tag{2}$$

We refer to [1] for notation and a general description of the problem. As in [1], we regard $y(\cdot)$ as the observed variable and $z(\cdot)$ as the variable to be regulated. We assume that either a dynamic observer as in [1], or a dynamic compensator as in [2], is utilized; equivalently, writing $\mathscr{N} = \operatorname{Ker} C$, we have $A\mathscr{N} \subset \mathscr{N}$.

For regulation of $z(\cdot)$, it is required to find a feedback map $F : \mathscr{X} \to \mathscr{U}$ such that

$$\mathscr{X}^+(A + BF) \subset \operatorname{Ker} D. \tag{3}$$

To respect the observability constraint, F must satisfy the condition

$$\operatorname{Ker} F \supset \mathscr{N}. \tag{4}$$

Necessary and sufficient conditions for the existence of F subject to (3) and (4) were given in [1].

In this paper we impose the additional requirement that F stabilize all the unstable modes of A which are both controllable and observable. Precisely, regard $\mathscr{X}/\mathscr{N}$ as the state space of the system (1), (2) made observable by reduction mod $\mathscr{N}$. The controllable, observable subspace is then $(\langle A|\mathscr{B}\rangle + \mathscr{N})/\mathscr{N}$. We require that the map induced on $(\langle A|\mathscr{B}\rangle + \mathscr{N})/\mathscr{N}$ by the closed loop system map $A + BF$ be stable. Equivalently, any observable, unstable modes of $A + BF$

* Received by the editors August 11, 1972, and in revised form March 13, 1973.

† Department of Electrical Engineering, University of Toronto, Toronto, Canada, M5S 1A4. This research was supported in part by the National Research Council of Canada under Grant A-7399.

‡ Department of Electrical Engineering, Rice University, Houston, Texas 77001.

must be uncontrollable; that is,

$$\frac{\mathscr{X}^+(A + BF) + \mathscr{N}}{\mathscr{N}} \cap \frac{\langle A|\mathscr{B}\rangle + \mathscr{N}}{\mathscr{N}} = 0. \tag{5}$$

It is natural to call a system in which F has been chosen to satisfy (5) *internally stable*. In this way we are led to formulate the following.

REGULATOR PROBLEM WITH INTERNAL STABILIZATION (RPIS). *Given the maps $A:\mathscr{X} \to \mathscr{X}$, $B:\mathscr{U} \to \mathscr{X}$, $D:\mathscr{X} \to \mathscr{Z}$, and a subspace $\mathscr{N} \subset \mathscr{X}$ with $A\mathscr{N} \subset \mathscr{N}$, find $F:\mathscr{X} \to \mathscr{U}$ such that*

$$\operatorname{Ker} F \supset \mathscr{N}, \tag{6}$$

$$\mathscr{X}^+(A + BF) \cap (\langle A|\mathscr{B}\rangle + \mathscr{N}) \subset \mathscr{N}, \tag{7}$$

and

$$\mathscr{X}^+(A + BF) \subset \operatorname{Ker} D. \tag{8}$$

Here it is easily checked that (7) is equivalent to (5).

The "restricted regulator problem" (RRP) was defined in [1] and is identical to RPIS except that the internal stability requirement (7) is dropped. We remark that RRP may be solvable when RPIS is not: that is, internal stability need not be compatible with output regulation, as shown by examples in [1] and [2].

In § 2 we provide a first set of necessary and sufficient conditions that RPIS be solvable. While not constructive, they are exploited to show that in this problem, dynamic compensation in the sense of [1] is redundant. In § 3 we give constructive necessary and sufficient conditions in the case $\mathscr{N} = 0$, and in § 4 extend them to the general case. An application is discussed in § 5. Finally, in § 6 we indicate how to deal with the sensitivity problem which may arise when the map A is subject to small perturbations.

2. Solution of RPIS: general considerations.

THEOREM 1. *RPIS is solvable if and only if there exists a subspace $\mathscr{V} \subset \mathscr{X}$ such that*

$$\mathscr{V} \subset \operatorname{Ker} D \cap A^{-1}(\mathscr{V} + \mathscr{B}), \tag{9}$$

$$\mathscr{X}^+(A) \cap \mathscr{N} + A(\mathscr{V} \cap \mathscr{N}) \subset \mathscr{V}, \tag{10}$$

$$\mathscr{V} \cap (\langle A|\mathscr{B}\rangle + \mathscr{N}) \subset \mathscr{N}, \tag{11}$$

and

$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle + \mathscr{V}. \tag{12}$$

We observe that conditions (9), (10) and (12) are equivalent to solvability of RRP, as shown by Theorem 1 of [1]; only condition (11) is new. The proof follows exactly the same lines as in the theorem cited, and so need only be sketched.

Proof. Suppose RPIS is solvable and put $\mathscr{V} = \mathscr{X}^+(A + BF)$. Since $A\mathscr{N} \subset \mathscr{N}$ and $\operatorname{Ker} F \supset \mathscr{N}$, we have by Lemma 2 of [1],

$$\mathscr{X}^+(A + BF) \cap \mathscr{N} = \mathscr{X}^+(A) \cap \mathscr{N}.$$

Then (9)–(11) follow immediately from (6)–(8), and (12) follows from the general identity [1, Lemma 4],

$$\langle A|\mathscr{B}\rangle + \mathscr{X}^+(A) = \langle A|\mathscr{B}\rangle + \mathscr{X}^+(A + BF). \tag{13}$$

Conversely, if (9) and (10) are true, there exists $F_0:\mathscr{X} \to \mathscr{U}$ such that $\operatorname{Ker} F_0 \supset \mathscr{N}$ and $(A + BF_0)\mathscr{V} \subset \mathscr{V}$. Write $A_0 = A + BF_0$. By (12) and (13),

$$\mathscr{X}^+(A_0) \subset \langle A_0|\mathscr{B}\rangle + \mathscr{V}. \tag{14}$$

Now

$$\mathscr{X}^+(A_0) \cap \mathscr{N} = \mathscr{X}^+(A) \cap \mathscr{N} \subset \mathscr{V}. \tag{15}$$

Just as in the proof of Theorem 1 of [1], (14) and (15) imply the existence of $F_1:\mathscr{X} \to \mathscr{U}$ such that $\operatorname{Ker} F_1 \supset \mathscr{N}$ and $\mathscr{X}^+(A_0 + BF_1) \subset \mathscr{V}$. Then $F = F_0 + F_1$ has all the properties required.

While Theorem 1 does not indicate how to find a suitable $\mathscr{V}$ if one exists, it is well suited to showing that if RPIS is not solvable, then no solution can be obtained by broadening the assumptions to include the possibility of state space extension, that is, dynamic compensation. We may interpret this result as a deterministic "separation theorem" which asserts that, after insertion of a dynamic observer, no further *dynamic* signal processing is required to achieve the stated design objectives, if these objectives can be achieved at all.[1]

To be precise, introduce extended spaces and maps exactly as in [1, § 3]. In the notation used there, we now define the "extended regulator problem with internal stabilization" (ERPIS) as that of finding suitable $\mathscr{X}_a$ (that is, $d(\mathscr{X}_a)$) and then $F_e:\mathscr{X}_e \to \mathscr{U}_e$, such that

$$\operatorname{Ker} F_e \supset \mathscr{N},$$

$$\mathscr{X}_e^+(A_e + B_eF_e) \cap (\langle A_e|\mathscr{B}_e\rangle + \mathscr{N}) \subset \mathscr{N},$$

and

$$\mathscr{X}_e^+(A_e + B_eF_e) \subset \operatorname{Ker} D \oplus \mathscr{X}_a.$$

THEOREM 2. *ERPIS is solvable only if RPIS is solvable.*

Proof. If ERPIS is solvable, Theorem 1 implies the existence of $\mathscr{V}_e \subset \mathscr{X}_e$ such that

$$\mathscr{V}_e \subset (\operatorname{Ker} D \oplus \mathscr{X}_a) \cap A_e^{-1}(\mathscr{V}_e + \mathscr{B} + \mathscr{B}_a), \tag{16}$$

$$\mathscr{X}_e^+(A_e) \cap \mathscr{N} + A_e(\mathscr{V}_e \cap \mathscr{N}) \subset \mathscr{V}_e, \tag{17}$$

$$\mathscr{V}_e \cap (\langle A|\mathscr{B}\rangle + \mathscr{X}_a + \mathscr{N}) \subset \mathscr{N}, \tag{18}$$

$$\mathscr{X}_e^+(A_e) \subset \langle A|\mathscr{B}\rangle + \mathscr{X}_a + \mathscr{V}_e. \tag{19}$$

Here we have used the facts (cf. [1, § 3]) that $\mathscr{N}_e = \mathscr{N}$, $\operatorname{Ker} D_e = \operatorname{Ker} D \oplus \mathscr{X}_a$, and $\langle A_e|\mathscr{B}_e\rangle = \langle A|\mathscr{B}\rangle \oplus \mathscr{X}_a$.

Let $P:\mathscr{X}_e \to \mathscr{X}_e$ be the projection on $\mathscr{X}$ along $\mathscr{X}_a$ and define $\mathscr{V} = P\mathscr{V}_e$. It is enough to show that $\mathscr{V}$ has the properties (9)–(12), and this requires only the

[1] It should, however, be borne in mind that in the present problem formulation, no explicit account is taken of the sensitivity of the synthesis to parameter perturbations; and if this is done, additional dynamic elements may sometimes be used to advantage (§ 6).

application of P to both sides of the relations (16)–(19). By definition of P and A_e, $PA_e = A_eP$ and $A_e|\mathscr{X} = A$. Using these facts and rewriting (16) as

$$\mathscr{V}_e \subset \operatorname{Ker} D \oplus \mathscr{X}_a, \quad A_e\mathscr{V}_e \subset \mathscr{V}_e + \mathscr{B} + \mathscr{B}_a,$$

there follows

$$\mathscr{V} \subset \operatorname{Ker} D, \quad A\mathscr{V} \subset \mathscr{V} + \mathscr{B},$$

which is equivalent to (9). Next, the obvious relation

$$\mathscr{X}_e^+(A_e) = \mathscr{X}^+(A) \oplus \mathscr{X}_a, \tag{20}$$

together with (19), establishes (12). To verify (11) from (18) we use the following general result for a map P and subspaces $\mathscr{R}$, $\mathscr{S}$ [3, Proposition A.5]:

$$P(\mathscr{R} \cap \mathscr{S}) = (P\mathscr{R}) \cap (P\mathscr{S}) \tag{21}$$

if and only if

$$(\mathscr{R} + \mathscr{S}) \cap \operatorname{Ker} P = \mathscr{R} \cap \operatorname{Ker} P + \mathscr{S} \cap \operatorname{Ker} P. \tag{22}$$

With $\operatorname{Ker} P = \mathscr{X}_a$, $\mathscr{R} = \mathscr{V}$ and $\mathscr{S} = \langle A|\mathscr{B}\rangle + \mathscr{X}_a + \mathscr{N}$, (22) follows at once, and then (21) applied to (18) yields (11). It remains to check (10) from (17). By (20),

$$\mathscr{X}^+(A_e) \cap \mathscr{N} = \mathscr{X}^+(A) \cap \mathscr{N},$$

and so

$$\mathscr{X}^+(A) \cap \mathscr{N} \subset P\mathscr{V}_e = \mathscr{V}. \tag{23}$$

Also, by (21), (22) we shall have

$$P(\mathscr{V}_e \cap \mathscr{N}) = \mathscr{V} \cap \mathscr{N}, \tag{24}$$

provided

$$(\mathscr{V}_e + \mathscr{N}) \cap \mathscr{X}_a = \mathscr{V}_e \cap \mathscr{X}_a + \mathscr{N} \cap \mathscr{X}_a. \tag{25}$$

As for (25), let

$$x_a = v_e + n \in (\mathscr{V}_e + \mathscr{N}) \cap \mathscr{X}_a,$$

with $v_e \in \mathscr{V}_e$ and $n \in \mathscr{N}$. Then

$$v_e = x_a - n \in \mathscr{X}_a + \mathscr{N},$$

and by (18), $v_e \in \mathscr{N}$. Therefore $x_a \in \mathscr{N} \cap \mathscr{X}_a = 0$, that is,

$$\mathscr{X}_a \cap (\mathscr{V}_e + \mathscr{N}) = 0,$$

proving (25). Then (24) is true, and (17) yields

$$\mathscr{V} \supset PA_e(\mathscr{V}_e \cap \mathscr{N}) = A(\mathscr{V} \cap \mathscr{N}). \tag{26}$$

Finally, (10) results from (23) and (26).

3. Constructive solution of RPIS when $\mathscr{N} = 0$. We first recall from [3] and [4] certain properties of (A, B)-invariant subspaces, that is, subspaces $\mathscr{V} \subset \mathscr{X}$ such that $A\mathscr{V} \subset \mathscr{V} + \mathscr{B}$. Let $\underline{F}(\mathscr{V})$ be the set of maps $F : \mathscr{X} \to \mathscr{U}$ such that

$(A + BF)\mathscr{V} \subset \mathscr{V}$. Then $\underline{F}(\mathscr{V}) \neq \varnothing$ if and only if $\mathscr{V}$ is (A, B)-invariant. We denote by $\mathscr{V}^*$ the largest (in the sense of inclusion) (A, B)-invariant subspace $\mathscr{V} \subset \operatorname{Ker} D$,

$$\mathscr{V}^* = \sup\{\mathscr{V} : \mathscr{V} \subset \operatorname{Ker} D \cap A^{-1}(\mathscr{V} + \mathscr{B})\};$$

and by $\mathscr{R}^*$ the largest controllability subspace [4] contained in $\operatorname{Ker} D$. Then $\mathscr{R}^* \subset \mathscr{V}^*$ and $\underline{F}(\mathscr{V}^*) \subset \underline{F}(\mathscr{R}^*)$. If $F \in \underline{F}(\mathscr{V}^*)$ and $A_F = A + BF$, let $\bar{A}_F$ be the map induced in $\mathscr{X}/\mathscr{R}^*$ by A_F. Then the restriction $\bar{A}_F|(\mathscr{V}^*/\mathscr{R}^*)$ is independent of the choice of $F \in \underline{F}(\mathscr{V}^*)$.

Let $A\mathscr{T} \subset \mathscr{T}$ and $A\mathscr{R} \subset \mathscr{R} \subset \mathscr{T}$. The subspace $\mathscr{R}$ *decomposes* $\mathscr{T}$ *relative to* A if there exists a subspace $\mathscr{S}$ such that $A\mathscr{S} \subset \mathscr{S}$ and $\mathscr{R} \oplus \mathscr{S} = \mathscr{T}$. A constructive necessary and sufficient condition that $\mathscr{R}$ decompose $\mathscr{T}$ is given in the Appendix; it amounts to the well-known fact that decomposability is equivalent to the existence of a solution to a simple linear matrix equation.

THEOREM 3. *Let* $\mathscr{N} = 0$. *Then RPIS is solvable if and only if*

$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle + \mathscr{V}^*, \tag{27}$$

and in $\mathscr{X}/\mathscr{R}^*$, *with* $F \in \underline{F}(\mathscr{V}^*)$, *the subspace*

$$(\mathscr{V}^* \cap \mathscr{X}^+(A_F) \cap \langle A|\mathscr{B}\rangle + \mathscr{R}^*)/\mathscr{R}^*$$

decomposes the subspace

$$(\mathscr{V}^* \cap \mathscr{X}^+(A_F) + \mathscr{R}^*)/\mathscr{R}^*$$

relative to the map induced by A_F *in* $\mathscr{V}^*/\mathscr{R}^*$.

It is clear that this solvability criterion is constructive, as $\mathscr{V}^*$, $\mathscr{R}^*$ are computable by simple algorithms [4]; an (arbitrary) $F \in \underline{F}(\mathscr{V}^*)$ is readily constructed [4] and decomposability is verifiable by a transformation of basis and rank checks (Appendix).

Proof. (If). Let $F \in \underline{F}(\mathscr{V}^*)$. By (27) and Lemma 4 of [1],

$$\mathscr{X}^+(A_F) \subset \langle A|\mathscr{B}\rangle + \mathscr{V}^*.$$

As the subspace on the right is A_F-invariant, we have

$$\mathscr{X}^+(A_F) = \langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A_F) + \mathscr{V}^* \cap \mathscr{X}^+(A_F). \tag{28}$$

Also, by the assumption of decomposability, there exists a subspace $\mathscr{W} \subset \mathscr{X}$ such that

$$A_F\mathscr{W} \subset \mathscr{W}, \tag{29}$$

$$\mathscr{R}^* \subset \mathscr{W} \subset \mathscr{V}^* \cap \mathscr{X}^+(A_F) + \mathscr{R}^*, \tag{30}$$

and

$$\frac{\mathscr{V}^* \cap \mathscr{X}^+(A_F) + \mathscr{R}^*}{\mathscr{R}^*} = \frac{\mathscr{V}^* \cap \langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A_F) + \mathscr{R}^*}{\mathscr{R}^*} \oplus \frac{\mathscr{W}}{\mathscr{R}^*}. \tag{31}$$

We remark that with $\mathscr{W}$ fixed, (29)–(31) hold for all $F \in \underline{F}(\mathscr{V}^*)$. By (31),

$$\mathscr{V}^* \cap \mathscr{X}^+(A_F) \subset \mathscr{V}^* \cap \langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A_F) + \mathscr{W}, \tag{32}$$

and by (28) and (32),

$$\mathscr{X}^+(A_F) \subset \langle A|\mathscr{B}\rangle + \mathscr{V}^* \cap \mathscr{X}^+(A_F) \subset \langle A|\mathscr{B}\rangle + \mathscr{W}. \tag{33}$$

Clearly,

$$\mathscr{W} \subset \mathscr{V}^*. \tag{34}$$

Also, by (31),

$$\mathscr{W}/\mathscr{R}^* \cap \langle A|\mathscr{B}\rangle/\mathscr{R}^* = 0,$$

so that

$$\mathscr{W} \cap \langle A|\mathscr{B}\rangle \subset \mathscr{R}^*. \tag{35}$$

Finally, let $\bar{A}_F$ denote the map induced by A_F in $\mathscr{X}/\mathscr{R}^*$, and choose $F \in \underline{F}(\mathscr{V}^*)$ such that

$$\sigma(A_F|\mathscr{R}^*) \cap \sigma(\bar{A}_F|(\mathscr{W}/\mathscr{R}^*)) = \varnothing.$$

Then $\mathscr{R}^*$ decomposes $\mathscr{W}$ relative to A_F; that is, there exists $\mathscr{V} \subset \mathscr{W}$ (depending on F) such that

$$A_F\mathscr{V} \subset \mathscr{V} \tag{36}$$

and

$$\mathscr{R}^* \oplus \mathscr{V} = \mathscr{W}. \tag{37}$$

From (33)–(37), we conclude that

$$\mathscr{V} \subset \operatorname{Ker} D \cap A^{-1}(\mathscr{V} + \mathscr{B}),$$
$$\mathscr{V} \cap \langle A|\mathscr{B}\rangle = 0,$$
$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle + \mathscr{V},$$

and it follows by Theorem 1 that RPIS is solvable.

(Only if). If RPIS is solvable, Theorem 1 supplies a subspace $\mathscr{V}$ such that

$$\mathscr{V} \subset \operatorname{Ker} D \cap A^{-1}(\mathscr{V} + \mathscr{B})$$

and

$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle \oplus \mathscr{V}. \tag{38}$$

Since $\mathscr{V} \cap \langle A|\mathscr{B}\rangle = 0$ we have $\mathscr{V} \cap \mathscr{R}^* = 0$, so that $\mathscr{V}^* \supset \mathscr{R}^* \oplus \mathscr{V}$. From this it is clear that $F \in \underline{F}(\mathscr{V}) \cap \underline{F}(\mathscr{V}^*)$ can be chosen such that $A_F|\mathscr{R}^*$ is stable. By (38) and Lemma 4 of [1] we have

$$\mathscr{X}^+(A_F) \subset \langle A|\mathscr{B}\rangle \oplus \mathscr{V} \subset \langle A|\mathscr{B}\rangle + \mathscr{V}^*.$$

Since all the subspaces here are A_F-invariant there follows

$$\begin{aligned}\mathscr{X}^+(A_F) &\subset \langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A_F) \oplus \mathscr{V} \cap \mathscr{X}^+(A_F)\\ &\subset \langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A_F) + \mathscr{V}^* \cap \mathscr{X}^+(A_F) \subset \mathscr{X}^+(A_F),\end{aligned}$$

and therefore

$$\mathscr{V}^* \cap \mathscr{X}^+(A_F) \subset \langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A_F) \oplus \mathscr{V} \cap \mathscr{X}^+(A_F).$$

Intersecting both sides with $\mathscr{V}^*$ and using $\mathscr{V}^* \supset \mathscr{V}$, we obtain

$$\begin{aligned}\mathscr{V}^* \cap \mathscr{X}^+(A_F) &= \langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A_F) \cap \mathscr{V}^* \oplus \mathscr{V} \cap \mathscr{X}^+(A_F)\\ &= \mathscr{S} \oplus \mathscr{T}, \text{ say.}\end{aligned} \tag{39}$$

Let $P:\mathscr{X} \to \mathscr{X}/\mathscr{R}^*$ be the canonical projection. By the stability of $A_F|\mathscr{R}^*$, we have

$$\begin{aligned}(\mathscr{S} \oplus \mathscr{T}) \cap \operatorname{Ker} P &= (\mathscr{S} \oplus \mathscr{T}) \cap \mathscr{R}^*\\ &\subset \mathscr{X}^+(A_F) \cap \mathscr{R}^*\\ &= 0\\ &= \mathscr{S} \cap \operatorname{Ker} P \oplus \mathscr{T} \cap \operatorname{Ker} P,\end{aligned}$$

and therefore

$$(P\mathscr{S}) \cap (P\mathscr{T}) = P(\mathscr{S} \cap \mathscr{T}) = 0. \tag{40}$$

By (39) and (40) we have, finally,

$$P[\mathscr{V}^* \cap \mathscr{X}^+(A_F)] = P[\langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A_F) \cap \mathscr{V}^*] \oplus P[\mathscr{V} \cap \mathscr{X}^+(A_F)],$$

a decomposition of the type required.

Remark. The foregoing proof of sufficiency made no essential use of the fact that $\mathscr{V}^*$ is actually the supremal element of the family of subspaces

$$\underline{\mathscr{V}} = \{\tilde{\mathscr{V}} : \tilde{\mathscr{V}} \subset \operatorname{Ker} D \cap A^{-1}(\tilde{\mathscr{V}} + \mathscr{B})\}.$$

The sole reason for stating Theorem 3 in terms of $\mathscr{V}^*$ is that this element of $\underline{\mathscr{V}}$ is readily computable algorithmically, and so the obtained conditions are constructive. It is clear from the proof that the conclusion of Theorem 3 is valid provided the stated conditions hold for *some* element $\tilde{\mathscr{V}} \in \underline{\mathscr{V}}$, with $\mathscr{R}^*$ replaced by the largest controllability subspace (c.s.) $\tilde{\mathscr{R}} \subset \tilde{\mathscr{V}}$.

4. Constructive solution of RPIS for arbitrary $\mathscr{N}$. It is not difficult to extend Theorem 3 to the general case. Suppose first that RPIS is solvable with the map F. Since $\operatorname{Ker} F \supset \mathscr{N}$, we have by Lemma 2 of [11] that

$$\mathscr{X}^+(A) \cap \mathscr{N} = \mathscr{X}^+(A + BF) \cap \mathscr{N} \tag{41}$$

is $(A + BF)$-invariant. Let

$$P:\mathscr{X} \to \bar{\mathscr{X}} = \mathscr{X}/[\mathscr{X}^+(A) \cap \mathscr{N}]$$

be the canonical projection, and let bars designate the maps induced in $\bar{\mathscr{X}}$. As $\operatorname{Ker} F \supset \operatorname{Ker} P$, $\bar F:\bar{\mathscr{X}} \to \mathscr{U}$ exists uniquely such that $\bar F P = F$, and it is easily seen that $P \operatorname{Ker} F = \operatorname{Ker} \bar F$. Similarly, by (8) and (41), $\bar D:\bar{\mathscr{X}} \to \mathscr{Z}$ exists uniquely such that $\bar D P = D$, and $P \operatorname{Ker} D = \operatorname{Ker} \bar D$. Finally, define $\bar B:\mathscr{U} \to \bar{\mathscr{X}}$ by $\bar B = PB$.

Now $\overline{A + BF} = \bar A + \bar B \bar F$ so (by [5, Lemma 2]),

$$P\mathscr{X}^+(A + BF) = \bar{\mathscr{X}}^+(\bar A + \bar B \bar F).$$

Also

$$\begin{aligned}[\mathscr{X}^+(A + BF) + \langle A|\mathscr{B}\rangle + \mathscr{N}] \cap \operatorname{Ker} P \\ &= \mathscr{X}^+(A) \cap \mathscr{N} \\ &= \mathscr{X}^+(A + BF) \cap \operatorname{Ker} P + (\langle A|\mathscr{B}\rangle + \mathscr{N}) \cap \operatorname{Ker} P.\end{aligned}$$

With these observations, we may project both sides of (6)–(8) to obtain

$$\operatorname{Ker} \bar{F} \supset \bar{\mathscr{N}}, \tag{42}$$

$$\bar{\mathscr{X}}^+(\bar{A} + \bar{B}\bar{F}) \cap (\langle \bar{A}|\bar{\mathscr{B}}\rangle + \bar{\mathscr{N}}) \subset \bar{\mathscr{N}}, \tag{43}$$

$$\bar{\mathscr{X}}^+(\bar{A} + \bar{B}\bar{F}) \subset \operatorname{Ker} \bar{D}. \tag{44}$$

Automatically

$$\bar{\mathscr{X}}^+(\bar{A}) \cap \bar{\mathscr{N}} = \bar{0},$$

or equivalently,

$$\bar{\mathscr{N}} \subset \bar{\mathscr{X}}^-(\bar{A}). \tag{45}$$

We have shown that if RPIS is solvable, so is the reduced problem (42)–(44) in $\bar{\mathscr{X}}$, and (45) is true as well. Conversely, suppose

$$\mathscr{X}^+(A) \cap \mathscr{N} \subset \operatorname{Ker} D, \tag{46}$$

and that $\bar{F} : \bar{\mathscr{X}} \to \mathscr{U}$ exists such that (42)–(44) are true. Define $F = \bar{F}P$. By reversing the steps which led to (42)–(44), one can routinely verify that (6)–(8) are true, i.e., that RPIS is solvable. We therefore have the following.

LEMMA 1. *RPIS is solvable if and only if the reduced problem* (42)–(44) *is solvable under assumption* (46).

Next we show that in (43) we may set $\bar{\mathscr{N}} = \bar{0}$.

LEMMA 2. *If* (42)–(46) *are true, then*

$$\bar{\mathscr{X}}^+(\bar{A} + \bar{B}\bar{F}) \cap \langle \bar{A}|\bar{\mathscr{B}}\rangle = \bar{0}. \tag{47}$$

Conversely, if (46) *and* (47) *hold, so does* (43).

Proof. By (43),

$$\begin{aligned}\bar{\mathscr{X}}^+(\bar{A} + \bar{B}\bar{F}) \cap (\langle \bar{A}|\bar{\mathscr{B}}\rangle + \bar{\mathscr{N}}) &\subset \bar{\mathscr{N}} \cap \bar{\mathscr{X}}^+(\bar{A} + \bar{B}\bar{F}) \\ &= \bar{\mathscr{N}} \cap \bar{\mathscr{X}}^+(\bar{A}) \qquad \text{by (42)} \\ &= \bar{0}.\end{aligned}$$

Conversely, the left side of (43) can be written

$$\begin{aligned}\bar{\mathscr{X}}^+(\bar{A} + \bar{B}\bar{F}) \cap [\langle \bar{A}|\bar{\mathscr{B}}\rangle \cap \bar{\mathscr{X}}^+(\bar{A} + \bar{B}\bar{F}) + \langle \bar{A}|\bar{\mathscr{B}}\rangle \cap \bar{\mathscr{X}}^-(\bar{A} + \bar{B}\bar{F}) + \bar{\mathscr{N}} \cap \bar{\mathscr{X}}^-(\bar{A})] \\ &= \bar{\mathscr{X}}^+\ (\bar{A} + \bar{B}\bar{F}) \cap \langle \bar{A}|\bar{\mathscr{B}}\rangle \\ &= \bar{0}.\end{aligned}$$

By Lemmas 1 and 2, the solvability of RPIS is equivalent to solvability of the reduced problem (42), (44), (47) under assumption (46). Our next result implies that condition (42) is redundant. For simplicity of notation, we temporarily drop bars.

LEMMA 3. *Let $F_0 : \mathscr{X} \to \mathscr{U}$ be such that*

$$\mathscr{X}^+(A + BF_0) \cap \langle A|\mathscr{B}\rangle = 0.$$

There exists $F_1 : \mathscr{X} \to \mathscr{U}$ such that

$$\operatorname{Ker} F_1 \supset \mathscr{X}^-(A), \tag{48}$$

$$\mathscr{X}^+(A + BF_1) = \mathscr{X}^+(A + BF_0). \tag{49}$$

Proof. In this proof, primes are used as indices. The lemma will be proved in three steps. First, let $P^+ : \langle A|\mathscr{B}\rangle \to \langle A|\mathscr{B}\rangle$ be the projection on $\langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A)$ along $\langle A|\mathscr{B}\rangle \cap \mathscr{X}^-(A)$, and write $A^+ = A|[\langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A)], B^+ = P^+B$. Since

$$\langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A) = P^+\langle A|\mathscr{B}\rangle = \langle A^+|\mathscr{B}^+\rangle,$$

we have that (A^+, B^+) is controllable, so there exists $F^+ : \mathscr{X}^+(A) \to \mathscr{U}$ such that $A^+ + B^+F^+$ is stable. Choose $\mathscr{S}^-$ such that

$$\langle A|\mathscr{B}\rangle \cap \mathscr{X}^-(A) \oplus \mathscr{S}^- = \mathscr{X}^-(A),$$

and then $\mathscr{S}$ such that $\mathscr{S} \supset \mathscr{S}^-$ and $\langle A|\mathscr{B}\rangle \oplus \mathscr{S} = \mathscr{X}$. Now define $F_0' : \mathscr{X} \to \mathscr{U}$ according to

$$F_0'|[\langle A|\mathscr{B}\rangle \cap \mathscr{X}^+(A)] = F^+,$$

$$F_0'|[\langle A|\mathscr{B}\rangle \cap \mathscr{X}^-(A) \oplus \mathscr{S}] = 0.$$

Write $A_0' = A + BF_0'$ and $F_0'' = F_0 - F_0'$. It is then clear that

$$\operatorname{Ker} F_0' \supset \mathscr{X}^-(A), \tag{50}$$

$$\mathscr{X}^+(A_0') \cap \langle A|\mathscr{B}\rangle = 0,$$

$$\mathscr{X}^+(A_0' + BF_0'') \cap \langle A|\mathscr{B}\rangle = 0.$$

As the second step, we claim there exists an $F_1' : \mathscr{X} \to \mathscr{U}$ such that

$$\operatorname{Ker} F_1' \supset \langle A|\mathscr{B}\rangle, \tag{51}$$

$$\mathscr{X}^+(A_0' + BF_1') = \mathscr{X}^+(A_0' + BF_0'') \quad (= \mathscr{X}^+(A + BF_0)). \tag{52}$$

For this let

$$F_1'|\langle A|\mathscr{B}\rangle = 0,$$

$$F_1'|\mathscr{X}^+(A_0' + BF_0'') = F_0''|\mathscr{X}^+(A_0' + BF_0''),$$

and let $F_1'|\mathscr{T}$ be defined arbitrarily on some complement $\mathscr{T}$ of $\mathscr{X}^+(A_0' + BF_0'') + \langle A|\mathscr{B}\rangle$ $(= \mathscr{X}^+(A) + \langle A|\mathscr{B}\rangle)$ in $\mathscr{X}$. Write

$$A_0'' = A_0' + BF_0'', \quad A_1' = A_0' + BF_1'.$$

Since $\mathscr{X}^+(A_0'')$ is A_0''-invariant, and

$$A_1'|\mathscr{X}^+(A_0'') = A_0''|\mathscr{X}^+(A_0''),$$

there follows by Lemma 2 of [1],

$$\mathscr{X}^+(A_1') \cap \mathscr{X}^+(A_0'') = \mathscr{X}^+(A_0''),$$

so that

$$\mathscr{X}^+(A_0'') \subset \mathscr{X}^+(A_1'). \tag{53}$$

Similarly, we have

$$\mathscr{X}^+(A_1') \cap \langle A|\mathscr{B}\rangle = \mathscr{X}^+(A_0') \cap \langle A|\mathscr{B}\rangle = 0.$$

By Lemma 4 of [1], there results

$$\langle A|\mathscr{B}\rangle \oplus \mathscr{X}^+(A_0'') = \langle A|\mathscr{B}\rangle \oplus \mathscr{X}^+(A_1'), \tag{54}$$

and (52) follows at once from (53) and (54).

As the last step, we prove the existence of $F_1'':\mathscr{X} \to \mathscr{U}$ such that

$$\operatorname{Ker} F_1'' \supset \mathscr{X}^-(A) + \langle A|\mathscr{B}\rangle, \tag{55}$$

$$\mathscr{X}^+(A_0' + BF_1'') = \mathscr{X}^+(A_0' + BF_1'). \tag{56}$$

A bar will denote a subspace or induced map in $\bar{\mathscr{X}} = \mathscr{X}/\langle A|\mathscr{B}\rangle$. Let $P:\mathscr{X} \to \bar{\mathscr{X}}$ be the canonical projection. We have $\bar{\mathscr{X}} = \bar{\mathscr{X}}^+(\bar{A}) \oplus \bar{\mathscr{X}}^-(\bar{A})$, and by (51), $F_1' = \bar{F}_1'P$ for some $\bar{F}_1':\bar{\mathscr{X}} \to \mathscr{U}$. Define $F_1'' = \bar{F}_1''P$, where

$$\bar{F}_1''|\bar{\mathscr{X}}^+(\bar{A}) = \bar{F}_1'|\bar{\mathscr{X}}^+(\bar{A}),$$
$$\bar{F}_1''|\bar{\mathscr{X}}^-(\bar{A}) = 0.$$

Then

$$P \operatorname{Ker} F_1'' = \operatorname{Ker} \bar{F}_1'' \supset \bar{\mathscr{X}}^-(\bar{A}),$$

so that

$$\mathscr{X}^-(A) \subset \operatorname{Ker} F_1'' + \langle A|\mathscr{B}\rangle = \operatorname{Ker} F_1'',$$

and (55) is true. Also if $x \in \mathscr{X}^+(A_0' + BF_1')$ then $Px \in \bar{\mathscr{X}}^+(\bar{A})$, so

$$F_1''x = \bar{F}_1''Px = \bar{F}_1'Px = F_1'x;$$

therefore

$$(A_0' + BF_1'')|\mathscr{X}^+(A_0' + BF_1') = (A_0' + BF_1')|\mathscr{X}^+(A_0' + BF_1'),$$

and there follows

$$\mathscr{X}^+(A_0' + BF_1'') \supset \mathscr{X}^+(A_0' + BF_1'). \tag{57}$$

Similarly,

$$(A_0' + BF_1'')|\langle A|\mathscr{B}\rangle = (A_0' + BF_1')|\langle A|\mathscr{B}\rangle,$$

and so

$$\mathscr{X}^+(A_0' + BF_1'') \cap \langle A|\mathscr{B}\rangle = \mathscr{X}^+(A_0' + BF_1') \cap \langle A|\mathscr{B}\rangle = 0.$$

This means

$$\begin{aligned}\mathscr{X}^+(A_0' + BF_1'') &\approx \bar{\mathscr{X}}^+(\overline{A_0' + BF_1''})\\ &= \bar{\mathscr{X}}^+(\overline{A_0' + BF_1'})\\ &\approx \mathscr{X}^+(A_0' + BF_1'),\end{aligned} \tag{58}$$

and (56) follows by (57) and (58).

It remains only to define

$$F_1 = F_0' + F_1''.$$

Then (48) follows by (50) and (55); statement (49) follows by (52) and (56).

It is now easy to prove our main result. For this we revert to the notation introduced at the beginning of this section.

THEOREM 4. *In the general case* $\mathscr{N} \neq 0$, *RPIS is solvable if and only if* (i)

$$\mathscr{X}^+(A) \cap \mathscr{N} \subset \operatorname{Ker} D, \tag{59}$$

and (ii),

in the factor space $\bar{\mathscr{X}} = \mathscr{X}/[\mathscr{X}^+(A) \cap \mathscr{N}]$, *the reduced problem is solvable, that is, there exists* $\bar{F}_0 : \bar{\mathscr{X}} \to \mathscr{U}$ *such that*

$$\bar{\mathscr{X}}^+(\bar{A} + \bar{B}\bar{F}_0) \subset \operatorname{Ker} \bar{D} \tag{60}$$

and

$$\bar{\mathscr{X}}^+(\bar{A} + \bar{B}\bar{F}_0) \cap \langle \bar{A}|\bar{\mathscr{B}}\rangle = \bar{0}. \tag{61}$$

Of course, the reduced problem (ii) is formally identical to that of Theorem 3.

Proof. (If). Suppose the reduced problem (RP) defined by (60) and (61) is solvable. Lemma 3 applied to RP yields a map $\bar{F} : \bar{\mathscr{X}} \to \mathscr{U}$ such that

$$\operatorname{Ker} \bar{F} \supset \bar{\mathscr{X}}^-(\bar{A}),$$

and $\bar{F}$ satisfies (44) and (47). Since $\bar{\mathscr{X}}^-(\bar{A}) \supset \mathscr{N}$ we are assured that (42) is true as well. As already noted, Lemmas 1 and 2 now imply that RPIS is solvable.

(Only if). The necessity of (59) is immediate from (9) and (10); the necessity of (60) and (61) follows by Lemmas 1 and 2.

5. Application: regulation in the presence of step disturbances. As a simple application of Theorem 3, consider the system

$$\begin{aligned}\dot{x}_1 &= A_1x_1 + A_3x_2 + B_1u,\\ \dot{x}_2 &= 0,\\ z &= D_1x_1 + D_2x_2.\end{aligned}$$

We assume that $y = x$ and (A_1, B_1) is controllable. The equations represent a controllable plant subjected to step disturbances which enter both dynamically and directly at the regulated output, a situation common in industrial process control.

In basis-free terms, our assumptions amount to the following:

$$\mathscr{N} = 0, \tag{62}$$

$$\operatorname{Im} A \subset \langle A|\mathscr{B}\rangle. \tag{63}$$

We now have our next theorem.

THEOREM 5. *Subject to the assumptitons* (62) *and* (63) *RPIS is solvable if and only if*

$$\langle A|\mathscr{B}\rangle + \operatorname{Ker} D \cap A^{-1}\mathscr{B} = \mathscr{X}. \tag{64}$$

Proof. (If). Exploiting the remark after the proof of Theorem 3, let

$$\tilde{\mathscr{V}} = \operatorname{Ker} D \cap A^{-1}\mathscr{B}. \tag{65}$$

From (64) it is clear, first, that

$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle + \tilde{\mathscr{V}}.$$

Also, as $A\tilde{\mathscr{V}} \subset \mathscr{B}$, there exists an $F \in \underline{F}(\tilde{\mathscr{V}})$ such that $A_F\tilde{\mathscr{V}} = 0$, where $A_F = A + BF$. Then

$$\tilde{\mathscr{V}} \subset \operatorname{Ker} A_F \subset \mathscr{X}^+(A_F),$$

so

$$\tilde{\mathscr{V}} \cap \mathscr{X}^+(A_F) \cap \langle A|\mathscr{B}\rangle = \tilde{\mathscr{V}} \cap \langle A|\mathscr{B}\rangle.$$

According to Theorem 4.3 of [4], the supremal c.s. $\tilde{\mathscr{R}}$ in $\tilde{\mathscr{V}}$ is given by

$$\tilde{\mathscr{R}} = \langle A_F|\mathscr{B} \cap \tilde{\mathscr{V}}\rangle = \mathscr{B} \cap \tilde{\mathscr{V}}.$$

The second condition of Theorem 3 (with $\tilde{\mathscr{V}}$, $\tilde{\mathscr{R}}$ in place of $\mathscr{V}^*$, $\mathscr{R}^*$) will thus be satisfied if $(\tilde{\mathscr{V}} \cap \langle A|\mathscr{B}\rangle)/(\mathscr{B} \cap \tilde{\mathscr{V}})$ decomposes $\tilde{\mathscr{V}}/(\mathscr{B} \cap \tilde{\mathscr{V}})$ relative to the map induced by A_F in $\tilde{\mathscr{V}}/(\mathscr{B} \cap \tilde{\mathscr{V}})$. Since $A_F|\tilde{\mathscr{V}} = 0$ this is trivial, and the result follows.

(Only if). Let $\bar{\mathscr{X}} = \mathscr{X}/\langle A|\mathscr{B}\rangle$, and now use bars for subspaces and induced maps in $\bar{\mathscr{X}}$. By (63), $\bar{A} = 0$. Suppose RPIS is solvable by the map F. Since $\bar{A}_F = \bar{A}$ for all F, and since $\mathscr{X}^+(A_F) \cap \langle A|\mathscr{B}\rangle = 0$, we know that $\mathscr{X}^+(A_F) = \operatorname{Ker} A_F$. Now $\operatorname{Ker} A_F \subset A^{-1}\mathscr{B}$ for any F, so

$$\mathscr{X}^+(A_F) \subset \operatorname{Ker} D \cap A^{-1}\mathscr{B}.$$

By Lemma 3 of [1], there results

$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle + \operatorname{Ker} D \cap A^{-1}\mathscr{B},$$

and therefore

$$\begin{aligned} \mathscr{X} &= \langle A|\mathscr{B}\rangle + \mathscr{X}^+(A) \\ &\subset \langle A|\mathscr{B}\rangle + \operatorname{Ker} D \cap A^{-1}\mathscr{B} \\ &\subset \mathscr{X}. \end{aligned}$$

6. Dynamic compensation and sensitivity. Even if RPIS is solved, it need not be true that output regulation is maintained if internal parameters deviate slightly from their nominal values. However, it is often possible to achieve insensitivity by application of the theory to a suitably augmented version of the original problem. We illustrate the method by a trivial example, deferring a general treatment to a future article. Consider

$$\begin{aligned} \dot{x}_1 &= ax_1 + u, & \dot{x}_2 &= 0, \\ y &= (x_1, x_2)', & z &= x_1 - x_2, \end{aligned}$$

where x_1, x_2 are scalars. RPIS is solvable with $u = f_1x_1 + f_2x_2$, where

$$f_1 + a < 0, \qquad f_2 = -(a + f_1).$$

If a changes to $a + \varepsilon$, with (f_1, f_2) fixed, the condition $f_1 + a + \varepsilon < 0$ for internal stability is preserved for small ε, but

$$z(\infty) = -(a + \varepsilon + f_1)^{-1}\varepsilon x_2(0+),$$

so output regulation fails.

Heuristically, replace εx_1 by a new variable x_3 with $\dot{x}_3 = 0$. That is, x_3 should satisfy the unstable differential equation induced in $\mathscr{X}/\langle A|\mathscr{B}\rangle$; since only $A|\langle A|\mathscr{B}\rangle$ is assumed to change, this equation does not depend on ε. Now consider RPIS for the augmented system

$$\dot{x}_1 = ax_1 + x_3 + u,$$
$$\dot{x}_2 = \dot{x}_3 = 0,$$
$$y = (x_1, x_2)', \qquad z = x_1 - x_2.$$

Here the pair (C, A) is observable; an observer for $w = x_1 - x_3$ is given by

(66) $$\dot{w} = -w + (a + 1)y_1 + u.$$

Also, RPIS is solvable; one solution is $u = f'x$, with

$$f_1 = -1 - a, \qquad f_2 = 1, \qquad f_3 = -1;$$

this can be implemented as

(67) $$u = -(2 + a)y_1 + y_2 + w.$$

Returning to the original perturbed system,

$$\dot{x}_1 = (a + \varepsilon)x_1 + u,$$

utilize (66) and (67) to find the transfer function

$$\hat{z}(s)/\hat{x}_2(s) = -s(s + 1 - \varepsilon)/[(s + 1)^2 - \varepsilon s].$$

With $\hat{x}_2(s) = x_2(0+)/s$, clearly $z(t) \to 0$ as $t \to \infty$ for all sufficiently small ε.

To summarize, parametric insensitivity is achieved by additional integrators, associated with an observer for suitably chosen fictitious external disturbances.

Appendix: on decomposability relative to a given subspace.[2] We assume that subspaces $\mathscr{R}, \mathscr{T} \subset \mathscr{X}$ are given, with $A\mathscr{T} \subset \mathscr{T}$ and $A\mathscr{R} \subset \mathscr{R} \subset \mathscr{T}$. By restricting A to $\mathscr{T}$, we can and do assume that $\mathscr{T} = \mathscr{X}$.

Let $J : \mathscr{R} \to \mathscr{X}$ be the natural injection, $1_{\mathscr{R}}$ the identity on $\mathscr{R}$, and $A_1 = A|\mathscr{R}$. It is a standard fact that $\mathscr{R}$ decomposes $\mathscr{X}$ relative to A if and only if there exists a map $Q : \mathscr{X} \to \mathscr{R}$ such that

(A.1) $$QJ = 1_{\mathscr{R}},$$

(A.2) $$QA = A_1Q.$$

[2] While no originality is claimed for the following discussion, it seems not to be explicit in most textbooks on linear algebra.

Indeed, if (A.1) and (A.2) hold, set $\mathscr{S} = \operatorname{Ker} Q$. Then if $x \in \mathscr{X}, x = JQx + (1 - JQ)x$; since $Q(1 - JQ)x = 0$, we have $x \in \mathscr{R} + \mathscr{S}$, so that $\mathscr{R} + \mathscr{S} = \mathscr{X}$. Also, $x \in \mathscr{R} \cap \mathscr{S}$ implies $x = 1_{\mathscr{R}}x = QJx = 0$, hence $\mathscr{R} \cap \mathscr{S} = 0$. Finally, $Qx = 0$ implies $QAx = A_1Qx = 0$, so $A\mathscr{S} \subset \mathscr{S}$. Conversely if $\mathscr{R} \oplus \mathscr{S} = \mathscr{X}$ with $A\mathscr{S} \subset \mathscr{S}$, let Q be the natural projection on $\mathscr{R}$ along $\mathscr{S}$.

Now let $\mathscr{R} \oplus \tilde{\mathscr{S}} = \mathscr{X}$, where $\tilde{\mathscr{S}}$ is an arbitrary complement of $\mathscr{R}$ in $\mathscr{X}$. In a compatible basis, A and J have matrices

$$A \sim \begin{bmatrix} A_1 & A_3 \\ 0 & A_2 \end{bmatrix}, \qquad J \sim \begin{bmatrix} I \\ 0 \end{bmatrix}. \tag{A.3}$$

By (A.3), the relations (A.1) and (A.2) are equivalent to

$$Q \sim [I \quad Q_2]$$

and

$$A_1Q_2 - Q_2A_2 - A_3 = 0. \tag{A.4}$$

Thus to check whether $\mathscr{R}$ decomposes $\mathscr{X}$, it is enough to verify whether the linear matrix equation (A.4) has a solution Q_2, a computational problem which in principle is straightforward.

Of greater theoretical interest is the following result, which can be inferred from [6, p. 199, Theorem 13] together with [7, p. 97, Ex. 6].

PROPOSITION. *$\mathscr{R}$ decomposes $\mathscr{X}$ if and only if the elementary divisors of $A|\mathscr{R}$, together with those of the induced map $\bar{A}$ in $\mathscr{X}/\mathscr{R}$, give all the elementary divisors of A.*

In (A.3), A_1 is the matrix of $A|\mathscr{R}$ and A_2 that of $\bar{A}$. The proposition thus solves the existence problem for (A.4). As a well-known special case, (A.4) has a solution (which is even unique) if the spectra of A_1 and A_2 are disjoint.

Acknowledgment. It is a pleasure to acknowledge stimulating discussions with E. J. Davison and S-H Wang, and with our graduate students, O. A. Sebakhy and T. L. Bratton.

REFERENCES

[1] W. M. WONHAM, *Tracking and regulation in linear multivariable systems*, this Journal, 11 (1973), pp. 424–437.

[2] S. P. BHATTACHARYYA AND J. B. PEARSON, *Error systems and the servomechanism problem*, Princeton Conference on Information and Systems Sciences, Princeton Univ., March, 1971.

[3] A. S. MORSE AND W. M. WONHAM, *Decoupling and pole assignment by dynamic compensation*, this Journal, 8 (1970), pp. 317–337.

[4] ———, *Decoupling and pole assignment in linear multivariable systems: a geometric approach*, this Journal, 8 (1970), pp. 1–18.

[5] S. P. BHATTACHARYYA, J. B. PEARSON AND W. M. WONHAM, *On zeroing the output of a linear system*, Information and Control, 20 (1972), pp. 135–142.

[6] F. R. GANTMACHER, *The Theory of Matrices*, vol. 1, Chelsea, New York, 1959.

[7] N. JACOBSON, *Lectures in Abstract Algebra, vol. 2: Linear Algebra*, Van Nostrand, Princeton, N.J., 1953.

SIAM J. CONTROL
Vol. 12, No. 1, February 1974

FINDING THE POINT OF A POLYHEDRON CLOSEST TO THE ORIGIN*

B. F. MITCHELL, V. F. DEM'YANOV AND V. N. MALOZEMOV

Abstract. An algorithm is given for finding the point of a convex polyhedron in an n-dimensional Euclidean space which is closest to the origin. It is assumed that the convex polyhedron is defined as the convex hull of a given finite set of points. This problem arises when one wishes to determine the direction of steepest descent for certain minimax problems.

1. Let a finite set of points $H = \{z_i\}_{i=1}^s$ be given in an n-dimensional Euclidean space E_n. We denote by L the convex hull of the points z_i:

$$L = \left\{z = \sum_{i=1}^{s} \alpha_i z_i | \alpha_i \geqq 0, \sum_{i=1}^{s} \alpha_i = 1\right\}.$$

Obviously, L is a bounded closed convex set. We shall denote by z^* the point of L which is closest to the origin

$$(z^*, z^*) = \min_{z \in L} (z, z).$$

Our goal is to describe a new method of successive approximations for finding the point z^*.

2. It is not difficult to show that the point z^* exists and is unique. Moreover, the following inequality holds for any $z \in L$ (see, e.g., [1]):

$$(1) \qquad (z, z^*) \geqq (z^*, z^*).$$

We set

$$\delta(z) = (z, z) - \min_{i \in [1:s]} (z_i, z).$$

Since

$$(2) \qquad (v, z) \geqq \min_{i \in [1:s]} (z_i, z)$$

for any $v, z \in L$, we have $\delta(z) \geqq 0$ if $z \in L$.

The following lemma also follows immediately from (1) and (2).

LEMMA 1. *The inequality*

$$(3) \qquad \|z - z^*\| \leqq \min \{\sqrt{\delta(z)}, \|z\|\}$$

holds for any $z \in L$.

COROLLARY 1. *If a sequence of points* $v_k \in L$, $k = 0, 1, 2, \cdots$, *is such that* $\delta(v_k) \xrightarrow[k \to \infty]{} 0$, *then*

$$v_k \xrightarrow[k \to \infty]{} z^*.$$

* Originally published in *Vestnik Leningrad Univ.*, 19 (1971), pp. 38–45. Submitted December 30, 1969. This translation into English has been prepared by K. Makowski.

COROLLARY 2. *If a sequence of points* $v_k \in L$, $k = 0, 1, 2, \cdots$, *is such that* $\|v_{k+1}\| \leqq \|v_k\|$, *and if there exists a subsequence* $\{v_{k_j}\}$ *for which* $\delta(v_{k_j}) \xrightarrow[j\to\infty]{} 0$, *then* $v_k \xrightarrow[k\to\infty]{} z^*$.

The following theorem holds.

THEOREM 1. *For a point* $\bar{z} \in L$ *to be the point of* L *closest to the origin, it is necessary and sufficient that* $\delta(\bar{z}) = 0$.

Proof. The *sufficiency* follows from (3). The *necessity*. Let $\bar{z} = z^*$. Then, first, $\delta(z^*) \geqq 0$. On the other hand, we have by virtue of (1), $(z_i, z^*) \geqq (z^*, z^*)$ for any $i \in [1:s]$, because $z_i \in H \subset L$. Hence, $\min_{i\in[1:s]}(z_i, z^*) \geqq (z^*, z^*)$ or, which is the same, $\delta(z^*) \leqq 0$. Therefore, $\delta(z^*) = 0$. The theorem has been proved.

3. We denote by Ξ the set of vectors A of the form

$$A = (\alpha_1, \cdots, \alpha_s), \qquad \alpha_i \geqq 0, \quad \sum_{i=1}^{s} \alpha_i = 1.$$

We set

$$z(A) = \sum_{i=1}^{s} \alpha_i z_i;$$

$$\Delta(A) = \max_{\{i|\alpha_i>0\}} (z_i, z(A)) - \min_{i\in[1:s]} (z_i, z(A)). \tag{4}$$

We denote by $i' = i'(A)$ a subscript at which the maximum in the right-hand side of (4) is attained (if there are several such subscripts, then we take any one of them). Thus, $\alpha_{i'} > 0$ and $(z_{i'}, z(A)) = \max_{\{i|\alpha_i>0\}}(z_i, z(A))$.

LEMMA 2. *The inequalities* $\alpha_{i'}\Delta(A) \leqq \delta(v) \leqq \Delta(A)$ *hold for any vector* $v = z(A)$, $A \in \Xi$.

Proof. We note that

$$(v, v) = \sum_{i=1}^{s} \alpha_i(z_i, z(A)) \leqq \max_{\{i|\alpha_i>0\}} (z_i, z(A)).$$

Hence, the inequality $\delta(v) \leqq \Delta(A)$ follows. We denote by $z_{i''}$, $i'' = i''(A)$, the point of the set H for which

$$(z_{i''}, z(A)) = \min_{i\in[1:s]} (z_i, z(A)).$$

In this case

$$\Delta(A) = (z_{i'} - z_{i''}, z(A)). \tag{5}$$

We set $\bar{A} = \{\bar{\alpha}_1, \cdots, \bar{\alpha}_s\} \in \Xi$, where

$$\bar{\alpha}_i = \begin{cases} \alpha_i & \text{for } i \neq i', i'', \\ 0 & \text{for } i = i', \\ \alpha_{i'} + \alpha_{i''} & \text{for } i = i''. \end{cases}$$

Obviously,

$$z(\bar{A}) = z(A) + \alpha_{i'}(z_{i''} - z_{i'}). \tag{6}$$

Since $z(\bar{A}) \in L$, we have by virtue of (2)

$$(z(\bar{A}), z(A)) \geqq \min_{i\in[1:s]} (z_i, z(A)). \tag{7}$$

Taking into account (7), (6), and (5), we obtain $\delta(z(A)) \geqq \alpha_{i'}\Delta(A)$. The lemma has been proved.

THEOREM 2. *For a point $v = z(A)$, $A \in \Xi$, to be the point of L closest to the origin, it is necessary and sufficient that $\Delta(A) = 0$.*

The proof follows in an obvious way from Lemma 2 and Theorem 1.

4. We shall now describe the method of successive approximations for finding the point z^*. We choose a vector $A_0 \in \Xi$ in an arbitrary way, and we set $v_0 = z(A_0)$. Assume that the kth approximation $v_k \in L : v_k = z(A_k)$, $A_k = (\alpha_1^{(k)}, \cdots, \alpha_s^{(k)}) \in \Xi$, has already been found. We describe the construction of v_{k+1}.

First of all, we find vectors $z_{i'_k}$ and $z_{i''_k}$ of H such that

$$(z_{i'_k}, v_k) = \max_{\{i|\alpha_i^{(k)} > 0\}} (z_i, z(A_k)),$$

$$(z_{i''_k}, v_k) = \min_{i \in [1:s]} (z_i, z(A_k)).$$

In this case,

$$\Delta_k \overset{\text{def}}{=} \Delta(A_k) = (z_{i'_k} - z_{i''_k}, v_k). \tag{8}$$

We consider the interval

$$\dot{v}_k(t) = v_k + t\alpha_{i'_k}^{(k)}(z_{i''_k} - z_{i'_k}), \qquad 0 \leqq t \leqq 1. \tag{9}$$

Let t_k with $0 \leqq t_k \leqq 1$ be determined by the relation

$$(v_k(t_k), v_k(t_k)) = \min_{0 \leqq t \leqq 1} (v_k(t), v_k(t)).$$

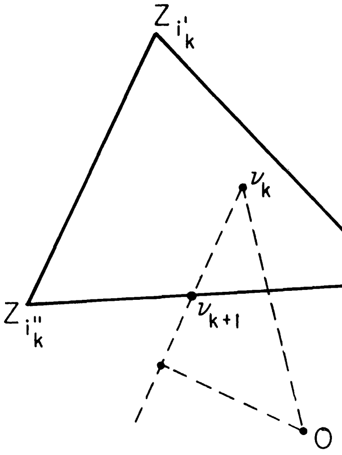

FIG. 1

We set $v_{k+1} = v_k(t_k)$ (see Fig. 1). It is not difficult to verify that $v_{k+1} = z(A_{k+1})$, where

$$A_{k+1} = (\alpha_1^{(k+1)}, \cdots, \alpha_s^{(k+1)}) \in \Xi,$$

$$\alpha_i^{(k+1)} = \begin{cases} \alpha_i^{(k)} & \text{for } i \neq i'_k, i''_k, \\ a_{i'_k}^{(k)} - t_k \alpha_{i'_k}^{(k)} & \text{for } i = i'_k, \\ \alpha_{i''_k}^{(k)} + t_k \alpha_{i'_k}^{(k)} & \text{for } i = i''_k. \end{cases}$$

For the sake of simplicity, we shall subsequently make use of the following notation: $\alpha'_k = \alpha_{i'_k}^{(k)}$, $z'_k = z_{i'_k}$, $\bar{z}_k = z_{i''_k}$.

Continuing the process described, we obtain the sequence of points $v_k \in L$, $k = 0, 1, 2, \cdots$, with

$$\|v_{k+1}\| \leqq \|v_k\|. \tag{10}$$

LEMMA 3. *The following limit relation holds*:

$$\lim_{k \to \infty} \alpha'_k \Delta_k = 0. \tag{11}$$

Proof. First, we note that by virtue of (8) and (9),

$$(v_k(t), v_k(t)) = (v_k, v_k) - 2t\alpha'_k\Delta_k + t^2(\alpha'_k\|\bar{z}_k - z'_k\|)^2. \tag{12}$$

Assume that the assertion of the lemma is false. Then there exists a subsequence $\{v_{k_j}\}$ for which $\alpha'_{k_j}\Delta_{k_j} \geqq \varepsilon > 0$. By virtue of (12) we have, for all $t \in [0, 1]$ and uniformly with respect to k_j,

$$(v_{k_j}(t), v_{k_j}(t)) \leqq (v_{k_j}, v_{k_j}) - 2t\varepsilon + t^2 d^2,$$

where $d = \max_{l,p \in [1:s]} \|z_l - z_p\| > 0$. Hence, it follows that the following inequality holds for $t_0 = \min\{\varepsilon/d^2, 1\}$ (obviously, $0 < t_0 \leqq 1$):

$$(v_{k_j}(t_0), v_{k_j}(t_0)) \leqq (v_{k_j}, v_{k_j}) - t_0\varepsilon.$$

Taking into account that, by definition, $(v_{k_j+1}, v_{k_j+1}) \leqq (v_{k_j}(t_0), v_{k_j}(t_0))$, we obtain

$$(v_{k_j+1}, v_{k_j+1}) \leqq (v_{k_j}, v_{k_j}) - t_0\varepsilon$$

uniformly with respect to k_j.

The number of such reductions in the monotonically nonincreasing sequence (v_k, v_k) is infinite, which contradicts the fact that all the (v_k, v_k) are nonnegative. The lemma has been proved.

LEMMA 4. *The limit relation*

$$\varliminf_{k \to \infty} \Delta_k = 0 \tag{13}$$

holds.

Proof. Assume the contrary: $\varliminf_{k \to \infty} \Delta_k = \Delta' > 0$. Then we have

$$\Delta_k \geqq \Delta'/2 \tag{14}$$

for numbers $k \geqq k_0$ sufficiently large. Taking into account (11), we conclude that

$$\alpha'_k \xrightarrow[k \to \infty]{} 0. \tag{15}$$

We also note that, by virtue of (12) and (14),

$$\|v_{k+1}\| < \|v_k\| \tag{16}$$

for $k \geqq k_0$. We denote by $\bar{t}_k$ the point at which $(v_k(t), v_k(t))$ attains its global minimum. Obviously (see (12)),

$$\bar{t}_k = \frac{\Delta_k}{\alpha'_k \|\bar{z}_k - z'_k\|^2}.$$

By virtue of (14) and (15), $\bar{t}_k \xrightarrow[k\to\infty]{} \infty$. Hence, it follows that for numbers $k \geqq k_1 \geqq k_0$ sufficiently large, the minimum of $(v_k(t), v_k(t))$ on the interval $0 \leqq t \leqq 1$ is attained for $t_k = 1$. Therefore, for these k,

$$v_{k+1} = v_k + \alpha'_k(\bar{z}_k - z'_k). \tag{17}$$

However, the sequence of points $v_{k_1}, v_{k_1+1}, v_{k_1+2}, \cdots$, which are connected by relation (17), can contain only a finite number of mutually distinct elements,[1] which contradicts (16). The lemma has been proved.

THEOREM 3. *The sequence $\{v_k\}$ constructed above converges to the point z^*.*

The proof follows from Lemmas 4 and 2 and from Corollary 2 to Lemma 1 in an obvious way.

Remark. If it turns out that, for some k, $\Delta_k = 0$, i.e., that $v_k = z^*$, then $v_{k+1} = z^*$ for all $j = 1, 2, \cdots$. This fact follows from (12).

5. We note certain peculiarities of the method of successive approximations described in the preceding section. We introduce the hyperplane

$$G = \{z | (z, z^*) = (z^*, z^*)\}.$$

THEOREM 4. *If $z^* \neq \mathbf{0}$, i.e., if the origin does not belong to L, then, beginning with some number, $v_k \in G$.*

Proof. We note that, by virtue of Theorem 1,

$$\min_{i\in[1:s]} (z_i, z^*) = (z^*, z^*).$$

We set

$$H_1 = \{z_i \in H | (z_i, z^*) = (z^*, z^*)\}; \quad H_2 = H \diagdown H_1 = \{z_i \in H | (z_i, z^*) > (z^*, z^*)\}.$$

If H_2 is an empty set, then $v_k \in G$ for all $k = 0, 1, 2, \cdots$. Therefore, we henceforth assume that H_2 is a nonempty set. We introduce the notation

$$\tau = \min_{z_i \in H_2} (z_i, z^*) - (z^*, z^*) > 0.$$

Since $v_k \xrightarrow[k\to\infty]{} z^*$, we have

$$\max_{i\in[1:s]} |(z_i, v_k) - (z_i, z^*)| < \tau/4$$

for numbers $k \geqq k_0$ sufficiently large. It is not difficult to show that the following relations hold for the same numbers $k \geqq k_0$:

$$(z_i, v_k) \leqq (z^*, z^*) + \tau/4 \quad \text{if } z_i \in H_1;$$

$$(z_i, v_k) \geqq (z^*, z^*) + 3\tau/4 \quad \text{if } z_i \in H_2.$$

[1] This remark is due to M. S. Al'tmark.

Hence, it follows that

$$\min_{i\in[1:s]}(z_i, v_k) = \min_{z_i\in H_1}(z_i, v_k). \tag{18}$$

Further, if a point $z_i \in H_2$ enters the representation of v_k, $k \geqq k_0$, with a nonzero coefficient, then

$$\Delta_k \geqq \tau/2, \tag{19}$$

where

$$\|v_{k+1}\| < \|v_k\|. \tag{20}$$

Let

$$v_k = \sum_{\{i|z_i\in H_1\}} \alpha_i^{(k)} z_i + \sum_{\{i|z_i\in H_2\}} \alpha_i^{(k)} z_i.$$

By virtue of the definitions of H_1 and τ,

$$(v_k - z^*, z^*) = \sum_{\{i|z_i\in H_2\}} \alpha_i^{(k)}(z_i - z^*, z^*) \geqq \tau \sum_{\{i|z_i\in H_2\}} \alpha_i^{(k)}.$$

Since the left-hand side of this inequality tends to zero as $k \to \infty$,

$$\sum_{\{i|z_i\in H_2\}} \alpha_i^{(k)} \xrightarrow[k\to\infty]{} 0.$$

We choose a large $k_1 \geqq k_0$ such that the following inequality holds for $k \geqq k_1$:

$$\sum_{\{i|z_i\in H_2\}} \alpha_i^{(k)} \leqq \frac{\tau}{2d^2}, \tag{21}$$

where $d = \max_{z_i\in H_1, z_j\in H_2} \|z_i - z_j\| > 0$. We denote by $\bar{t}_k$ the point at which $(v_k(t), v_k(t))$ attains its global minimum. If $z_i \in H_2$ enters the representation of v_k, $k \geqq k_1$, with a nonzero coefficient, then we obtain by virtue of (19) and (21)

$$\bar{t}_k = \frac{\Delta_k}{\alpha_k' \|\bar{z}_k - z_k'\|^2} \geqq \frac{\tau}{2\left(\sum_{\{i|z_i\in H_2\}} \alpha_i^{(k)}\right) d^2} \geqq 1.$$

Hence, it follows that

$$v_{k+1} = v_k + \alpha_k'(\bar{z}_k - z_k'). \tag{22}$$

We assume that points $z_i \in H_2$ enter the representations of all vectors

$$v_{k_1}, v_{k_1+1}, v_{k_1+2}, \cdots \tag{23}$$

with nonzero coefficients. By virtue of (22), sequence (23) contains only a finite number of mutually distinct elements. However, this contradicts (20). Therefore, there exists a point $v_{\bar{k}}$, $\bar{k} \geqq k_1$, which has the following representation:

$$v_{\bar{k}} = \sum_{\{i|z_i\in H_1\}} \alpha_i^{(\bar{k})} z_i; \quad \alpha_i^{(\bar{k})} \geqq 0, \quad \sum_{\{i|z_i\in H_1\}} \alpha_i^{(\bar{k})} = 1.$$

By virtue of (18), all the v_k with $k \geqq \bar{k}$ have similar representations. In particular, we have by definition of H_1, for $k \geqq \bar{k}$, that $(v_k, z^*) = (z^*, z^*)$, i.e., $v_k \in G$. The theorem has been proved.

THEOREM 5. *The limit relation*

$$\lim_{k\to\infty} \Delta_k = 0$$

holds.

Proof. If $z^* = \mathbf{0}$, then the assertion of the theorem follows from the definition of Δ_k and from the fact that $\|v_k\| \xrightarrow[k\to\infty]{} 0$. Therefore, we assume that $z^* \neq \mathbf{0}$. By virtue of Theorem 4, we have for $k \geqq \bar{k}$,

$$\Delta_k \leqq \max_{\{i|z_i\in H_1\}} (z_i, v_k) - \min_{\{i|z_i\in H_1\}} (z_i, v_k).$$

According to the definition of H_1, the right-hand side of this inequality tends to zero as $k \to \infty$. Therefore, $\Delta_k \xrightarrow[k\to\infty]{} 0$ also, since $\Delta_k \geqq 0$. The theorem has been proved.

6. We set

$$\tilde{\Delta}_k = \Delta_k/\|v_k\|^2, \qquad k = 0, 1, 2, \cdots .$$

If $v_k = \mathbf{0}$, then we set by definition $\tilde{\Delta}_k = \infty$.

THEOREM 6. *For the origin to belong to the set L, it is necessary and sufficient that the following inequality hold for all* $k = 0, 1, 2, \cdots$:

$$(24) \qquad \tilde{\Delta}_k \geqq 1.$$

Proof. The necessity. We have $z^* = \mathbf{0}$. Taking into account that $v_k \in L$, we obtain on the basis of Lemmas 1 and 2

$$\tilde{\Delta}_k = \frac{\Delta_k}{\|v_k\|^2} \geqq \frac{\delta(v_k)}{\|v_k\|^2} \geqq 1.$$

The sufficiency. Assume that $z^* \neq \mathbf{0}$. Then $\tilde{\Delta}_k \leqq \Delta_k/\|z^*\|^2$. By virtue of Theorem 5, we obtain $\tilde{\Delta}_k \xrightarrow[k\to\infty]{} 0$, which contradicts (24). The theorem has been proved.

Thus, if $z^* = \mathbf{0}$, then $\|v_k\| \xrightarrow[k\to\infty]{} 0$ and, for all $k = 0, 1, 2, \cdots$, the inequality $\tilde{\Delta}_k \geqq 1$ holds. If $z^* \neq \mathbf{0}$, then $\|v_k\| \geqq \|z^*\|$ and $\tilde{\Delta}_k \xrightarrow[k\to\infty]{} 0$.

If the inequality $\tilde{\Delta}_k < 1$ holds for some k, then, by virtue of Theorem 6, the origin does not belong to the set L. Moreover, it is not difficult to prove that, in this case, the hyperplane $(v_k, z) - (v_k, \bar{z}_k) = 0$ strictly separates the origin from L.

7. We remind the reader that $v_k = z(A_k)$, $A_k \in \Xi$. We set $I_k = \{i|\alpha_i^{(k)} > 0\}$ and introduce the set

$$B_k = \left\{ z = z_{i_0} + \sum_{\substack{i\in I_k\\ i\neq i_0}} \alpha_i(z_i - z_{i_0}) | \alpha_i \in (-\infty, \infty) \right\}.$$

Here, i_0 is an arbitrary subscript of I_k. We denote by $\tilde{v}_k$ the vector of B_k with the smallest norm: $\|\tilde{v}_k\| = \min_{z\in B_k} \|z\|$. We note that the point $\tilde{v}_k \in B_k$ is unique, although its representation in the form

$$\tilde{v}_k = z_{i_0} + \sum_{\substack{i\in I_k\\ i\neq i_0}} \bar{\alpha}_i(z_i - z_{i_0})$$

may not be unique.

It is not difficult to show that the numbers $\bar{\alpha}_l$ constitute the solution of the following linear system:

$$\left(z_{i_0} + \sum_{\substack{i \in I_k \\ i \neq i_0}} \alpha_i(z_i - z_{i_0}), z_j - z_{i_0}\right) = 0, \qquad j \in I_k, \quad j \neq i_0. \tag{25}$$

THEOREM 7. *There exists an infinite subsequence of vectors* $\{\tilde{v}_{k_j}\}$ *such that* $\tilde{v}_{k_j} = z^*$ *for all* k_j.

Proof. We shall assume that $z^* \neq \mathbf{0}$ (if $z^* = \mathbf{0}$, then the proof is only simplified). First, we separate a subsequence $\{v_{k_j}\}$ such that

(i) $\alpha_i^{(k_j)} \xrightarrow[j \to \infty]{} \alpha_i^*$, $i \in [1:s]$, in this case,

$$v_{k_j} \xrightarrow[j \to \infty]{} z^* = \sum_{i=1}^{s} \alpha_i^* z_i;$$

(ii) $v_{k_j} \in G$ (see Theorem 4).

We set $I^* = \{i | \alpha_i^* > 0\}$. Obviously, we have for k_j sufficiently large

$$I^* \subset I_{k_j}. \tag{26}$$

Henceforth, we consider only such k_j. We denote by L^* the convex hull of the points z_i, $i \in I^*$. Obviously, $z^* \in L^*$. We denote by L_{k_j} the convex hull of the points z_i, $i \in I_{k_j}$. By virtue of (ii), (26), and of the definition of the set B_{k_j}, we have

$$L^* \subset L_{k_j} \subset B_{k_j} \subset G.$$

Further, $\|z^*\| = \min_{z \in G} \|z\| \leqq \min_{z \in B_{k_j}} \|z\|$. Since $z^* \in L^*$, $z^* \in B_{k_j}$. Therefore, $\|z^*\| = \min_{z \in B_{k_j}} \|z\|$.

Taking into account that the point of B_{k_j} with the smallest norm is unique, we obtain $\tilde{v}_{k_j} = z^*$. The theorem has been proved.

On the basis of this theorem, one can assert that finding z^* reduces to solving a finite number of systems of linear equations of form (25). We note that it is purposeful to solve these systems only for the k for which either $\|v_k\|$ or $\tilde{\Delta}_k$ is sufficiently small.

Regarding other methods of finding the point of a polyhedron which is closest to the origin, see [2]–[4].

REFERENCES

[1] S. KARLIN, *Mathematical Methods and Theory in Games, Programming and Economics*, Addison-Wesley, Reading, Mass., 1959.

[2] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[3] V. F. DEM'YANOV AND A. M. RUBINOV, *Minimization of a smooth, convex functional on a convex set*, Vestnik Leningrad Univ., 19 (1964), pp. 5–17.

[4] B. N. KOZINETZ, *On a learning algorithm for a linear perceptron*, Computational Technique and Programming Problems, Leningrad Univ. Press, Leningrad, 1964, pp. 80–83.

[5] *ALGOL-procedures, Methodic Materials on Computer Software*, vol. 9, Leningrad Univ. Press, Leningrad, 1971.

SIAM J. CONTROL
Vol. 12, No. 1, February 1974

EXISTENCE AND LOWER CLOSURE THEOREMS FOR ABSTRACT CONTROL PROBLEMS*

LEONARD D. BERKOVITZ†

Abstract. A general control problem that includes, among others, the distributed control problem and the optimal control problem for systems governed by ordinary differential equations is considered. Two lower closure theorems and an existence theorem with hypotheses corresponding to those of the lower closure theorems are proved. The first lower closure theorem was first proved by Cesari. The present proof is different and simpler. The second lower closure theorem dispenses with Cesari's property (Q), but requires a generalized Lipschitz condition to hold.

1. Introduction. In [7] L. Cesari formulated a general control problem and proved lower closure theorems and existence theorems for this problem. Cesari's earlier results [4] on the existence of optimal controls in distributed control problems are special cases of these theorems, as are certain semicontinuity and existence theorems of Morrey [15] and Fichera [10]. Existence theorems for many other control problems and problems in the calculus of variations can also be obtained as special cases of these theorems. For details see § 9 of [7].

In § 5 of this paper we shall present a different and simpler proof of Cesari's results. Although our results will be slightly more general than Cesari's in one small point, the principal contribution of this work is in the simplification of the proof. Our method is also applicable to problems in which Cesari's property (Q) is not assumed. In § 6 we prove a lower closure theorem in which property (Q) is replaced by a generalized Lipschitz condition. The usual convexity assumptions, of course, are retained. The methods and results presented in this paper were developed in [1] and [2] for control problems governed by systems of ordinary differential equations.

2. Notation and formulation of problem. We shall use single letters to denote vectors, we shall use subscripts to distinguish vectors, and we shall use superscripts to denote components of vectors. The letter t will denote a vector $(t^1, \cdots, t^\nu)$, in real Euclidean space R^ν, $\nu \geqq 1$, the letter x will denote a vector $(x^1, \cdots, x^n)$ in real Euclidean space R^n, $n \geqq 1$, and the letter w a vector in R^m, $m \geqq 1$. The Euclidean norm of a vector x will be denoted by $|x|$. The inner product of two vectors x_1 and x_2 will be written as $\langle x_1, x_2 \rangle$. Thus $|x| = \langle x, x \rangle^{1/2}$.

Let $f^0 : (t, x, w) \to f^0(t, x, w)$ be a real-valued function defined on $R^\nu \times R^n \times R^m$ and let $f: (t, x, w) \to f(t, x, w)$ be a vector-valued function defined on $R^\nu \times R^n \times R^m$ with range in R^r. Let G be a bounded region of the t-space R^ν, and let X be a region of the x-space R^n. Let $\mathscr{R}$ denote the Cartesian product $\bar{G} \times \bar{X}$, where $\bar{G}$ denotes the closure of G and $\bar{X}$ denotes the closure of X. Let Ω be a mapping that assigns to each point (t, x) in $\mathscr{R}$ a subset $\Omega(t, x)$ of the w-space R^m.

* Received by the editors July 17, 1972, and in revised form January 15, 1973.

† Division of Mathematical Sciences, Purdue University, Lafayette, Indiana 47907. This research was supported by the National Science Foundation under Grant GP-3351X.

As usual, let $L_{p_i}(G)$, $p_i \geqq 1$, denote the Banach space of real-valued measurable functions z^i defined on G such that $\int_G |z^i|^{p_i}\,dt < \infty$, and let

$$\|z^i\|_{p_i} = \left\{ \int_G |z^i|^{p_i}\,dt \right\}^{1/p_i}, \qquad p_i \geqq 1.$$

Let $\mathscr{Y}$ and $\mathscr{Z}$ denote the following Banach spaces:

$$\mathscr{Y} = \prod_{i=1}^{r} L_{q_i}(G), \qquad \mathscr{Z} = \prod_{i=1}^{n} L_{p_i}(G) \qquad q_i, p_i \geqq 1.$$

The norm of an element $z = (z^1, \cdots, z^n)$ in $\mathscr{Z}$ is given by

$$\|z\| = \left\{ \sum_{i=1}^{n} \|z^i\|_{p_i}^2 \right\}^{1/2}.$$

A similar formula gives the norm of an element $y = (y^1, \cdots, y^r)$ in $\mathscr{Y}$.

Let $\mathscr{F}$ be a Banach space and let the norm of an element ϕ in $\mathscr{F}$ be denoted by $\|\phi\|$. Let M be a mapping from $\mathscr{F}$ to $\mathscr{Z}$ and let N be a mapping from $\mathscr{F}$ to $\mathscr{Y}$. Thus the image under N of an element ϕ in $\mathscr{F}$ is an element $N\phi = y = (y^1, \cdots, y^r)$ in $\mathscr{Y}$, where each y^i is in $L_{q_i}(G)$. Similarly the image under M of an element ϕ in $\mathscr{F}$ is an element $M\phi = z = (z^1, \cdots, z^n)$ in $\mathscr{Z}$, where each z^i is in $L_{p_i}(G)$. Note that N and M need not be linear.

Let $\mathscr{M}$ denote the set of all measurable functions u defined on G with range in R^m. Thus if $u \in \mathscr{M}$, then $u = (u^1, \cdots, u^m)$, u^i real-valued and measurable in G.

To motivate the formulation of the abstract optimal control problem to be given below we recall the formulation of the optimal control problem for systems governed by ordinary differential equations and the formulation of the optimal control problem for distributed parameter systems.

Let t be a scalar ($v = 1$) and let $G = (a, b)$. One of the important optimal control problems for systems governed by ordinary differential equations can be stated somewhat imprecisely as follows. Minimize

$$\int_a^b f^0(t, \phi(t), u(t))\,dt$$

subject to $\phi'(t) = f(t, \phi(t), u(t))$ and $u(t) \in \Omega(t, \phi(t))$. For the problem to make sense, the functions ϕ must be absolutely continuous. Since $v = 1$, the functions ϕ that are under consideration can be considered to be elements of the Sobolev space $H_1^1(G)$. Let $\mathscr{F} = H_1^1(G)$. For any set $\mathscr{L}$, let $[\mathscr{L}]^k$ denote the k-fold Cartesian product of $\mathscr{L}$ with itself. Since the functions ϕ' are in $[L_1(G)]^n$, the differentiation operator, which we henceforth write as N, is a mapping from $\mathscr{F}$ into $[L_1(G)]^n$. The function ϕ, in addition to being an element of $\mathscr{F}$, is also an element of $[L_1(G)]^n$. Let M denote the mapping that imbeds $H_1^1(G)$ in $[L_1(G)]^n$; thus $(M\phi)(t) = \phi(t)$, with ϕ in $H_1^1(G)$ and $M\phi \in [L_1(G)]^n$. The control problem can now be stated. Minimize

$$\int_a^b f^0(t, (M\phi)(t), u(t))\,dt$$

subject to $(N\phi)(t) = f(t, (M\phi)(t), u(t))$ and $u(t) \in \Omega(t, (M\phi)(t))$.

The use of Sobolev spaces to formulate the optimal control problem for systems governed by ordinary differential equations may seem somewhat pretentious. For distributed parameter systems ($\nu > 1$), however, Sobolev spaces must be introduced. As we shall see, the Sobolev space formulation of the ordinary control problem carries over to the distributed parameter systems.

Let $\alpha = (\alpha^{i_1}, \cdots, \alpha^{i_n})$ denote a generic multi-index and let $|\alpha| = \sum \alpha^{i_j}$. Let

$$D^\alpha = \partial^{|\alpha|}/\partial x_n^{\alpha^{i_n}} \cdots \alpha x_1^{\alpha^{i_1}}.$$

One of the optimal control problems for distributed parameter systems is the following. Minimize

$$\int_G f^0(t, (D^\beta \phi)(t), \phi(t), u(t))\, dt$$

subject to

$$(D^\alpha \phi)(t) = f(t, (D^\beta \phi)(t), \phi(t), u(t))$$

and $u(t) \in \Omega(t, \phi(t))$, where α ranges over multi-indices with $|\alpha| = s$ and β ranges over multi-indices with $|\beta| < s$. From experience with partial differential equations and multidimensional variational problems without side conditions (e.g., minimize $\int_G f^0(t, \Phi(t), (\nabla\phi)(t))\, dt$ we know that to establish a reasonably general existence theorem we must take the functions ϕ to be elements of an appropriate Sobolev space. Since the highest order derivative appearing is s, we take ϕ to be in $H_p^s(G)$ for some $p \geqq 1$. Let $\mathscr{F} = H_p^s(G)$. We can write $N = D^\alpha$. The differential operator D^β maps ϕ into $D^\beta\phi$, an element of $[L_p(G)]^\gamma$, for appropriate γ. The function ϕ itself also belongs to $[L_p(G)]^n$. Therefore, we may define a mapping M from $\mathscr{F}$ to $[L_p(G)]^{\gamma+n}$ as follows: $M : \phi \to (D^\beta\phi, \phi)$. In this notation the optimal control problem for distributed parameter systems becomes: Minimize

$$\int_G f^0(t, (M\phi)(t), u(t))\, dt$$

subject to $(N\phi)(t) = f(t, (M\phi)(t), u(t))$ and $u(t) \in \Omega(t, (M\phi)(t))$. Formally, the distributed parameter problem now has the same format as the problem in which the state is governed by a system of ordinary differential equations.

The motivation for the following definition and problem formulation should now be clear.

DEFINITION 2.1. An element ϕ in $\mathscr{F}$ is said to be an *admissible trajectory* if there is a function u in $\mathscr{M}$ such that the following hold:

(2.1)

(i) $(t, (M\phi)(t)) \in \mathscr{R}$ a.e. in $\bar{G}$,

(ii) $(N\phi)(t) = f(t, (M\phi)(t), u(t))$ a.e. in G,

(iii) $u(t) \in \Omega(t, (M\phi)(t))$ a.e. in G,

(iv) the mapping $t \to f^0(t, (M\phi)(t), u(t))$ is in $L_1(G)$.

The function u is said to be an *admissible control* and the pair (ϕ, u) is said to be an *admissible pair*. The set of all admissible pairs will be denoted by $\mathscr{A}$.

The *optimal control problem* is to minimize the functional

$$J(\phi, u) = \int_G f^0(t, (M\phi)(t), u(t))\, dt \tag{2.2}$$

in a given class $\mathscr{A}_0 \subseteq \mathscr{A}$ of admissible pairs. That is, we are to find a (ϕ^*, u^*) in $\mathscr{A}_0$ such that $J(\phi^*, u^*) \leqq J(\phi, u)$ for all (ϕ, u) in $\mathscr{A}_0$. Such a pair (ϕ^*, u^*) is called an *optimal pair.* The element ϕ^* is called an *optimal trajectory*; the element u^* is called an *optimal control.*

We have already shown how a distributed control problem and a control problem governed by ordinary differential equations can be obtained as special cases of the general problem just formulated. To obtain the classical multidimensional variational problem we take $\mathscr{F} = H_1^1(G)$, $N = \nabla, f = w$, and take M to be the identity map that assigns to ϕ in $H_1^1(G)$ the same function considered as an element in $[L_1(G)]^n$. For other problems covered by the general formulation, the reader is again referred to § 9 of [7].

3. Outline of proof of existence. The proof of the existence of an optimal pair in a class $\mathscr{A}_0 \subseteq \mathscr{A}$ proceeds in general as follows.

Let

$$\mu = \inf \{J(\phi, u) : (\phi, u) \in \mathscr{A}_0\}. \tag{3.1}$$

If $\mu = +\infty$, there is nothing to prove. If $\mu < \infty$, then there exists a minimizing sequence such that $J(\phi_k, u_k) \to \mu$. Conditions are then placed on the problem to ensure that $\mu > -\infty$. Additional conditions are placed on the problem to ensure that the sequence of trajectories $\{\phi_k\}$ is conditionally compact in some sense. From experience in less abstract problems we know that conditional weak compactness of the minimizing sequence will suffice.

Having guaranteed the weak compactness of $\{\phi_k\}$ we select a subsequence, again labeled $\{\phi_k\}$, that converges weakly to an element ϕ^* in $\mathscr{F}$. Any boundary conditions or other conditions on the trajectories ϕ that are used to define the subclass $\mathscr{A}_0$ must be such that if $\phi_k \to \phi^*$ weakly, then ϕ^* must satisfy these conditions.

At this point a "lower closure" theorem is used. This theorem states that, under appropriate hypotheses, if $\phi_k^* \to \phi$ weakly in $\mathscr{F}$, then there is a u^* in $\mathscr{M}$ such that (ϕ^*, u^*) is admissible and

$$J(\phi^*, u^*) \leqq \liminf J(\phi_k, u_k). \tag{3.2}$$

If we can also show that $(\phi^*, u^*) \in \mathscr{A}_0$, then $J(\phi^*, u^*) \geqq \mu$. On the other hand for the minimizing sequence, and hence for the subsequence with weakly convergent trajectories, we have

$$\liminf J(\phi_k, u_k) = \lim J(\phi_k, u_k) = \mu.$$

Combining this with (3.2) and $J(\phi^*, u^*) \geqq \mu$ gives $J(\phi^*, u^*) = \mu$. Hence (ϕ^*, u^*) is an optimal pair.

The difficult step in the existence proof is the proof of the lower closure theorem. The next three sections will be devoted to two lower closure theorems. We shall return to the existence question in § 7.

4. Assumptions. In this section we discuss a set of assumptions that will be in force for all of our theorems. For ease of reference we shall label this set, Assumption A.

One of the statements in A requires the introduction of a function Q^+ that assigns to each point in $\mathscr{R} = \bar{G} \times \bar{X}$ a subset $Q^+(t, x)$ of R^{r+1} as follows:

$$Q^+(t, x) = \{(\eta, \xi) : \eta \geqq f^0(t, x, w),\ \xi = f(t, x, w),\ w \in \Omega(t, x)\},$$

where η is a scalar and ξ is an r-vector.

Assumption A. (i) If $\{\phi_k\}$ is a sequence of elements in $\mathscr{F}$ that converges weakly to an element ϕ in $\mathscr{F}$, then $N\phi_k \to N\phi$ weakly in $\mathscr{Y}$ and $M\phi_k \to M\phi$ strongly in $\mathscr{X}$. (ii) For each (t, x) in $\mathscr{R}$ the set

$$\mathscr{D} = \{(t, x, w) : (t, x) \in \mathscr{R},\ w \in \Omega(t, x)\}$$

is closed. (iii) For each (t, x) in $\mathscr{R}$ the set $Q^+(t, x)$ is closed and convex. (iv) There exists a real-valued function ψ in $L_1(G)$ such that $\psi(t) \geqq 0$ on G, and a constant r-vector b such that

$$f^0(t, x, w) - \langle b, f(t, x, w)\rangle \geqq -\psi(t)$$

for all (t, x, w) in $\mathscr{D}$.

Assumption A(i) is fulfilled in many important problems. In the control problem governed by ordinary differential equations, $v = 1$ and $\bar{G} = [a, b]$. Admissible trajectories are in the Sobolev space $H_1^1(\bar{G})$, which we take to be $\mathscr{F}$ in the present case. The weak convergence of a sequence $\{\phi_k\}$ in $H_1^1(\bar{G})$ implies the convergence in $[L_1(\bar{G})]^n$ of $\{\phi_k\} = \{M\phi_k\}$ and the weak convergence in $[L_1(\bar{G})]^n$ of the sequence $\{N\phi_k\} = \{\phi_k'\}$. The weak convergence in L_1 of the $\{\phi_k'\}$ implies the equiabsolute continuity of the $\{\phi_k\}$. Since $v = 1$ and $\bar{G} = [a, b]$, we even have the stronger result that the sequence $\{\phi_k\}$ is uniformly convergent (i.e., converges in $C[a, b]$). Conversely, a uniformly convergent sequence $\{\phi_k\}$ that is equiabsolutely continuous converges weakly in $\mathscr{F} = H_1^1(\bar{G})$.

In distributed control problems $(v > 1)$, the functions ϕ are in a Sobolev space $H_p^s(G)$ or in the Cartesian product of Sobolev spaces $H_{p_i}^{s_i}(G)$ (see [5], [7]). It is a standard result [16, Thm. 3.4.4, p. 75] that if $\phi_n \to \phi$ weakly in $H_p^s(G)$, then $\phi_n \to \phi$ strongly in $H_p^{s-1}(G)$, provided the boundary of G is sufficiently regular. Thus, since N is a differential operator of order s and M is defined by the relation $M\phi = (D^\beta\phi, \phi)$ with $|\beta| < s$, Assumption 1 is fulfilled, provided we assume that the boundary of G is sufficiently regular.

Assumption A(ii) is equivalent to the assumption that Ω is upper semicontinuous on $\mathscr{R}$ in the sense of Kuratowski [13, pp. 32–34]. Assumption A(iii) was introduced by Cesari in [3].

In [7] Cesari assumes that for every point t' in $\bar{G}$ there is a neighborhood $N_\delta(t')$ of t', a nonnegative function ψ in $L_1(N_\delta(t'))$ and a vector b such that the inequality in Assumption A(iv) holds for all t in $N_\delta(t')$, all x in $\bar{X}$, and $w \in \Omega(t, x)$. Note that ψ and b depend on the neighborhood. It is easy to see that as a consequence of the compactness of $\bar{G}$ it suffices to consider Assumption A(iv) as stated here.

5. First lower closure theorem. One of the assumptions in the first lower closure theorem involves a slight generalization of Cesari's property (Q). Criteria for property (Q) and a discussion of the relationship of property (Q) to various growth conditions and other assumptions used in the calculus of variations and in optimal control theory are given by Cesari in § 6 of [8].

Let $\delta > 0$, let $(t_0, x_0) \in \mathscr{R}$ and let $N_x(t_0, x_0, \delta)$ denote the set of points (t_0, x) in $\mathscr{R}$ such that $|x - x_0| < \delta$. Let

$$Q^+(N_x(t_0, x_0, \delta)) = \bigcup\{Q^+(t_0, x) : (t_0, x) \in N_x(t_0, x_0, \delta)\}.$$

For a set A let cl co A denote the closure of the convex hull of A. The mapping Q^+ is said to satisfy the weak Cesari property (Q*) at (t_0, x_0) if

$$Q^+(t_0, x_0) = \bigcap_{\delta > 0} \operatorname{cl} \operatorname{co} Q^+(N_x(t_0, x_0, \delta)).$$

It is readily verified that if property (Q) holds at a point then so does property (Q*). Note that if the sets $Q^+(t, x)$ are independent of x, then the weak property (Q*) holds.

THEOREM 1. *Let Assumption* A *hold. Let the function f^0 be lower semicontinuous and let the function f be continuous. Let the mapping Q^+ satisfy property* (Q*) *at all points of $\mathscr{R}$ with the possible exception of a set of points whose t-coordinates lie in a set of measure zero in G. Let $\{(\phi_k, u_k)\}$ be a sequence of admissible pairs such that the sequences of trajectories $\{\phi_k\}$ converges weakly in $\mathscr{F}$ to an element ϕ. Let* $\liminf J(\phi_k, u_k) < +\infty$. *Then there exists a function u in $\mathscr{M}$ such that (ϕ, u) is admissible and*

$$J(\phi, u) \leqq \liminf J(\phi_k, u_k). \tag{5.1}$$

This theorem, with the assumption that property (Q) holds instead of (Q*), is due to Cesari [5], [7].

Remark. Theorem 1 remains true if we replace the assumptions on f^0 and f by the following weaker assumption.

Assumption B. For each t in G, the function $F = (f^0, f)$ is a continuous function of (x, w) on R^{n+m}, and for each (x, w) in R^{n+m} the function F is measurable with respect to t in G.

At the appropriate place in the proof of Theorem 1, we shall indicate the modifications that must be made to accommodate the weakened hypotheses.

Our proof of Theorem 1 is different from Cesari's and will exploit Mazur's theorem which states that a strongly closed convex set in a Banach space is weakly closed. Thus, our proof essentially exploits the Hahn–Banach theorem. In the proof we shall select subsequences of various sequences. Unless stated otherwise, we shall relabel the subsequence with the labeling of the original sequence.

Let $\{\phi_k\}$ and ϕ be as in the hypotheses of the theorem. Let

$$y_k = N\phi_k, \qquad y = N\phi,$$
$$z_k = M\phi_k, \qquad z = M\phi.$$

For the value of y_k at t we write $y_k(t)$; thus $y_k(t) = (N\phi_k)(t)$. Similar notation will hold for y, z_k and z.

We break the proof up into several steps.

Step 1. Let

$$\gamma = \liminf J(\phi_k, u_k). \tag{5.2}$$

Then γ is finite.

Since by hypothesis $\gamma < +\infty$, we need only show that $\gamma > -\infty$.

From Assumption A(iv) and (2.1) (ii) we get that for a.e. t in G,

$$\begin{aligned} f^0(t, z_k(t), u_k(t)) &\geqq -\psi(t) + \langle b, f(t, z_k(t), u_k(t))\rangle \\ &= -\psi(t) + \langle b, y_k(t)\rangle. \end{aligned} \tag{5.3}$$

From A(i) and the hypothesis that $\phi_k \to \phi$ weakly, we get that y_k converges weakly to $y = N\phi$ in $\mathscr{Y}$. Hence $\{\|y_k\|\}$ is bounded, say by a constant $B > 0$. From this and from the inequality

$$\left|\int_G \langle b, y_k(t)\rangle\, dt\right| \leqq C\|y_k\|,$$

where C is an appropriate constant, it follows that the integrals $\int_G \langle b, y_k(t)\rangle\, dt$ are bounded. Integration of (5.3) now shows that the sequence $J(\phi_k, u_k)$ is bounded below. Hence $\gamma > -\infty$.

Before presenting the rest of the proof we outline the idea of the proof. Since $y_k \to y$ weakly in $\mathscr{Y}$, there is a sequence $\{\psi_j\}$ of convex combinations of the y_k that converges strongly to y in $\mathscr{Y}$. Hence there is a subsequence of the $\{\psi_j\}$ that converges a.e. to y. We then use the same convex combinations of the functions $t \to f^0(t, y_k(t), u_k(t))$ as were used to define the ψ_j to define a sequence $\{\lambda_j\}$. We then show that the function $\lambda = \liminf \lambda_j$ is integrable and that $\int_G \lambda\, dt \leqq \gamma$. Property (Q*) enables us to also show that for a.e. t in G, $(\lambda(t), y(t)) \in Q^+(t, z(t))$. Thus there is a function $v: G \to R^m$ such that $y(t) = f(t, z(t), v(t))$ and $\lambda(t) \geqq f^0(t, z(t), v(t))$. We then use the McShane–Warfield extension of Filippov's lemma to show that we can replace v by a measurable function u. Thus, (ϕ, u) is admissible and $\lambda(t) \geqq f^0(t, z(t), u(t))$. Integration of the last inequality and the previously established relation $\int_G \lambda\, dt \leqq \gamma$ then give the theorem.

Step 2. There exists a real-valued function λ that is integrable on G such that

$$\int_G \lambda\, dt \leqq \gamma \tag{5.4}$$

and such that $(\lambda(t), y(t)) \in Q^+(t, z(t))$ a.e. in G.

We first select a subsequence $\{(\phi_k, u_k)\}$ such that

$$\lim J(\phi_k, u_k) = \gamma. \tag{5.5}$$

For this subsequence we still have $\phi_k \to \phi$ weakly in $\mathscr{F}$. Therefore, $\{z_k\} = \{M\phi_k\}$ converges strongly in $\mathscr{Z}$. From the definition of $\mathscr{Z}$ in § 2, it follows that the sequences of components $\{z_k^i\}, i = 1, \cdots, n$, converge in $L_{p_i}(G)$ to $z^i = (M\phi)^i$. Hence there is a subsequence of $\{(\phi_k, u_k)\}$ such that $z_k(t) \to z(t)$ a.e. in G. We now select this subsequence.

Since we still have $\phi_k \to \phi$ weakly in $\mathscr{F}$, we have that $y_k \to y$ weakly in $\mathscr{Y}$. Therefore, from Mazur's theorem (e.g., [11, Cor., Thm. 2.9.3, p. 36]) we obtain the following statement. For each integer j there exists an integer n_j, a set of integers $i = 1, \cdots, k$, where $k = k(j)$ depends on j, and a set of numbers $\alpha_{ij}, \cdots, \alpha_{kj}$ satisfying

$$\alpha_{ij} \geqq 0, \quad i = 1, \cdots, k, \quad \sum_{i=1}^{k} \alpha_{ij} = 1, \tag{5.6}$$

such that $n_{j+1} > n_j + k(j)$ and

$$\left\| y - \sum_{i=1}^{k} \alpha_{ij} y_{n_j+i} \right\| \leqq \frac{1}{j}. \tag{5.7}$$

Let

$$\psi_j = \sum_{i=1}^{k} \alpha_{ij} y_{n_j+i}.$$

From (2.1) (ii) we have that for a.e. t in G,

$$\psi_j(t) = \sum_{i=1}^{k} \alpha_{ij} f(t, z_{n_j+i}(t), u_{n_j+i}(t)). \tag{5.8}$$

In terms of ψ_j, (5.7) says that $\psi_j \to y$ in the norm topology of $\mathscr{Y}$. From the definition of $\mathscr{Y}$ in § 2, it follows that each component ψ_j^i, $i = 1, \cdots, r$, of ψ_j converges to y^i in $L_{p_i}(G)$. Hence there exists a subsequence $\{\psi_j\}$ such that

$$\psi_j(t) \to y(t) \quad \text{a.e. in } G. \tag{5.9}$$

We note for future reference that since G is bounded, this subsequence also converges to y in $[L_1(G)]^n$.

We now suppose that (5.8) is the subsequence for which (5.9) holds. Corresponding to (5.8) we define a sequence $\{\lambda_j\}$ as follows:

$$\lambda_j(t) = \sum_{i=1}^{k} \alpha_{ij} f^0(t, z_{n_j+i}(t), u_{n_j+i}(t)), \tag{5.10}$$

where for each j the numbers α_{ij}, the indices n_{j+i} and the functions z_{n_j+i} and u_{n_j+i} are as in (5.8).

Define

$$\lambda(t) = \liminf \lambda_j(t). \tag{5.11}$$

From (5.8), (5.10), Assumption A(iv) and (5.6) we get

$$\begin{aligned} &\lambda_j(t) - \langle b, \psi_j(t) \rangle \\ &\quad = \sum_{i=1}^{k} \alpha_{ij}(f^0(t, z_{n_j+i}(t), u_{n_j+i}(t)) - \langle b, f(t, z_{n_j+i}(t), u_{n_j+i}(t)) \rangle) \\ &\quad \geqq \sum_{i=1}^{k} \alpha_{ij}(-\psi(t)) = -\psi(t). \end{aligned}$$

Hence we may apply Fatou's lemma and (5.9) and (5.11) to obtain

$$\begin{aligned} \liminf \int_G (\lambda_j - \langle b, \psi_j \rangle)\, dt &\geqq \int_G \liminf (\lambda_j - \langle b, \psi_j \rangle)\, dt \\ &= \int_G (\lambda - \langle b, y \rangle)\, dt \end{aligned} \tag{5.12}$$

and

$$\lambda(t) \geqq -\psi(t) + \langle b, y(t) \rangle \quad \text{a.e.} \tag{5.13}$$

We noted in the sentence after (5.9) that $\psi_j \to y$ in $[L_1(G)]^n$. From this observation and from (5.12) we therefore get

$$\int_G \lambda \, dt \leqq \liminf \int_G \lambda_j \, dt. \tag{5.14}$$

But from (5.10), (5.5) and (5.6) we get

$$\liminf \int_G \lambda_j \, dt = \liminf \sum_{i=1}^{k} \alpha_{ij} \int_G f^0(t, z_{n_j+i}(t), u_{n_j+i}(t)) \, dt$$

$$= \liminf \sum_{i=1}^{k} \alpha_{ij} J(\phi_{n_j+i}, u_{n_j+i}) = \gamma.$$

If we combine this result with (5.14) and note (5.13) which shows that λ is bounded from below by an integrable function, we see that λ is integrable and that (5.4) holds. Note that λ must be finite almost everywhere in G.

We next show that $(\lambda(t), y(t)) \in Q^+(t, z(t))$. Let G_1 denote the set of points in G at which $\lambda(t)$ is finite, $\psi_j(t) \to y(t)$, $z_k(t) \to z(t)$, and at which property (Q*) holds. Then meas $G_1 =$ meas G. For each integer k define a set E_k as follows: $E_k = \{t \in G : u_k(t) \notin \Omega(t, z_k(t))\}$. Then by (2.1)(iii), meas $E_k = 0$. Let E denote the union of the sets E_k and let G_2 denote the points in G that are not in E. Let $G' = G_1 \cap G_2$. Clearly, meas $G' =$ meas G.

Let t be a point in G'. There exists a subsequence $\{\lambda_j(t)\}$, which depends on t, such that $\lambda_j(t) \to \lambda(t)$. For the corresponding subsequence $\psi_j(t)$ we have, by (5.9), $\psi_j(t) \to y(t)$. Since $z_k(t) \to z(t)$ it follows that for each $\delta > 0$ there exists an integer k_0 depending on δ such that if $k > k_0$ then $|z_k(t) - z(t)| < \delta$. Hence for $k > k_0$, $(t, z_k(t)) \in N_x(t, z(t), \delta)$. Therefore, for j sufficiently large,

$$(y^0_{n_j+i}(t), y_{n_j+i}(t)) \in Q^+(N_x(t, z(t), \delta)),$$

where $y^0_{n_j+i}(t) = f^0(t, z_{n_j+i}(t), u_{n_j+i}(t))$ and $y_{n_j+i}(t) = f(t, z_{n_j+i}(t), u_{n_j+i}(t))$. Therefore, by (5.8), (5.10) and (5.6),

$$(\lambda_j(t), \psi_j(t)) \in \operatorname{co} Q^+(N_x(t, z(t), \delta)),$$

where co A denotes the convex hull of A. Since $\lambda_j(t) \to \lambda(t)$ and $\psi_j(t) \to y(t)$, we have

$$(\lambda(t), y(t)) \in \operatorname{cl} \operatorname{co} Q^+(N_x(t, z(t), \delta)).$$

Since $\delta > 0$ is arbitrary, $(\lambda(t), y(t))$ is in $\operatorname{cl} \operatorname{co} Q^+(N_x(t, z(t), \delta))$ for every $\delta > 0$, and hence in the intersection of these sets. Therefore, by property (Q*), we get that $(\lambda(t), y(t)) \in Q^+(t, z(t))$. Since t was an arbitrary point in G', the proof of Step 2 is complete.

Step 3. There exists a measurable function $u = (u_1^1, \cdots, u^m)$ defined on G such that for almost all t in G: (i) $y(t) = f(t, z(t), u(t))$; (ii) $u(t) \in \Omega(t, z(t))$; (iii) $\lambda(t) \geqq f^0(t, z(t), u(t))$.

The existence of a function v satisfying the conclusion of Step 3 is a restatement of $(\lambda(t), y(t)) \in Q^+(t, z(t))$. We show that there is a measurable function u with this property. Let $T = \{t : (\lambda(t), y(t)) \in Q^+(t, z(t))\}$, let $Z = R^\nu \times R^n \times R^m \times R^1$ and let $D = \{(t, x, w, \eta) : (t, x, w) \in \mathscr{D}, \ \eta \geqq f^0(t, x, w)\}$. The functions y and λ are measurable. Clearly, T is measurable and Z is Hausdorff. Since by Assumption A(ii), $\mathscr{D}$ is closed, and since f^0 is lower semicontinuous, the set D is closed and hence can be written as the union of a countable number of compact sets. Let $\Phi: t \rightarrow (t, z(t), y(t), \lambda(t))$; thus Φ is a measurable map from T to Z. Let $\Gamma: (t, x, w, \eta) \rightarrow (t, x, f(t, x, w), \eta)$. Then Γ is a continuous map from D to Z and $\Phi(T) \subset \Gamma(D)$. Thus, the hypotheses of the McShane–Warfield extension of Filippov's lemma (see [14]) are satisfied. Hence there exists a measurable mapping $\mu: T \rightarrow D$, such that $\mu: t \rightarrow (\tau(t), x(t), u(t), \eta(t))$ and such that $\Gamma(\mu(t)) = \Phi(t)$. Hence,

$$(\tau(t), x(t), f(\tau(t), x(t), u(t)), \eta(t)) = (t, z(t), y(t), \lambda(t))$$

for all t in T. From this, Step 3 follows.

If we replace the continuity assumptions on F by Assumption B, then by a well-known theorem ([19, Thm. 18.2, p. 142] or [18]) there exists for each $\varepsilon > 0$ an open set $E \subset G$ such that $\operatorname{meas}(E) < \varepsilon$ and such that F is continuous on $(G-E) \times R^n \times R^m$. We proceed as above to obtain a measurbale u on $T - (E \cap T)$. Since ε is arbitrary we obtain the desired result. For details see [12] or [17].

Step 4. *Completion of proof.*

We first show that (ϕ, u) is admissible, where u is the function obtained in Step 3.

Statements (i) and (ii) of Step 3 assert that (ϕ, u) satisfies (2.1)(ii) and (iii). For almost all t in G, $z(t) = \lim z_k(t) = \lim (M\phi_k)(t)$. Since $\mathscr{R} = \bar{G} \times \bar{X}$ is closed and since for each k, $(t, z_k(t)) \in \mathscr{R}$ for almost all t in $\bar{G}$, it follows that $(t, z(t)) \equiv (t, (M\phi)(t)) \in \mathscr{R}$, a.e. in $\bar{G}$. Thus (ϕ, u) satisfies (i) of (2.1). Since z and u are measurable and f^0 is either lower semicontinuous or satisfies Assumption B, it follows that the function y^0 defined by $y^0(t) = f^0(t, z(t), u(t))$ is measurable. From (iii) of Step 3 we see that y^0 is bounded above by an integrable function. From Assumption A(iv) we have

$$y^0(t) \geqq -\psi(t) + \langle b, f(t, z(t), u(t)) \rangle,$$

and so y^0 is bounded below by an integrable function. Hence y^0 is integrable and (2.1)(iv) holds. Thus (ϕ, u) is admissible.

From (iii) of Step 3, (5.4) and (5.2), we obtain (5.1) and thereby complete the proof of Theorem 1.

6. Second lower closure theorem. Let p_i, $i = 1, \cdots, n$, be as in the third paragraph of § 2. Let $p = \min \{p_i : i = 1, \cdots, n\}$. Since G is bounded, all components of functions $z = M\phi$, where ϕ is an element of $\mathscr{Z}$, are in L_p. Moreover any sequence $\{z_k\}$ that converges strongly in $\mathscr{Z}$ converges in L_p. If $p > 1$, let q be defined by the relationship $1/p + 1/q = 1$; if $p = 1$, let $q = +\infty$.

We next list a set of hypotheses about the data of the problem that will be in effect for Theorem 2.

Assumption C. (i) The function $F = (f^0, f)$ is continuous. (ii) The sets $\Omega(t, x)$ are independent of x; i.e., for a given t in $\bar{G}$, $\Omega(t, x) = \Omega(t, x')$ for all x and x' in $\bar{X}$. (iii) There exist a nondecreasing function μ defined on $[0, \infty)$ and a nonnegative function H defined on $\bar{G} \times R^m$ such that the following hold. (a) $\lim_{\delta \rightarrow 0} \mu(\delta) = 0$. (b) There exists a $\delta_0 > 0$ such that for $\delta > \delta_0$, $\mu(\delta) \leqq \delta$. (c) For all (t, x, w) and

(t, x', w) in $\mathscr{D}$,

$$|F(t, x, w) - F(t, x', w)| \leqq \mu(|x - x'|)\{H(t, w) + K|x - x'|^{p-1}\}, \tag{6.1}$$

where K is a nonnegative constant.

Note that if F is uniformly continuous on $\mathscr{D}$, which occurs if $\mathscr{D}$ is compact, then (6.1) holds with $H \equiv 1$ and μ the modulus of continuity, suitably defined for large values of δ. If F is Lipschitz in x, then (6.1) holds with $\mu(\delta) = \delta, p = 1$ and H equal to the Lipschitz constant.

In the linear plant with quadratic criterion problem, we have

$$f(t, x, w) = A(t)x + B(t)w$$

and

$$f^0(t, x, w) = \langle x, Q(t)x\rangle + \langle x, P(t)w\rangle + \langle w, R(t)w\rangle,$$

where the matrix functions A and B have entries in $L_2(G)$ and the matrix functions Q, P, R have entries that are bounded and measurable. The matrices $Q(t)$ and $R(t)$ are also symmetric. It is easily verified that (6.1) holds in this problem.

THEOREM 2. *Let Assumptions* A *and* C *hold. Let* $\{(\phi_k, u_k)\}$ *be a sequence of admissible pairs such that the sequence of trajectories* $\{\phi_k\}$ *converges weakly in* $\mathscr{F}$ *to an element* ϕ. *Let* $\liminf J(\phi_k, u_k) < \infty$. *Let there exist a constant* A *such that for all* k

$$\|H_k\|_q < A, \tag{6.2}$$

where $H_k(t) = H(t, u_k(t))$ *and* $\| \cdots \|_q$ *denotes the* L_q*-norm*, $1 \leqq q \leqq \infty$. *Then there exists a function* u *in* $\mathscr{M}$ *such that* (ϕ, u) *is admissible and* (5.1) *holds.*

The special case in which $v = 1$ and the control system is governed by ordinary differential equations was treated by us in [2]. In [2] we also gave an example wherein Theorem 2 was applicable, while Theorem 1 was not.

Note that if F is Lipschitz in x, or if $\mathscr{D}$ is compact, then (6.2) always holds. In the linear plant quadratic criterion problem, if all of the controls u_k lie in a ball in $[L_2(G)]^m$, then (6.2) holds. Under suitable hypotheses on the matrices P, Q, R, this will be true whenever $\{(\phi_k, u_k)\}$ is a minimizing sequence. One such set of hypotheses is that P and Q are positive semidefinite on $\bar{G}$ and that R is continuous and positive definite on $\bar{G}$.

The proof of Theorem 2 proceeds just as the proof of Theorem 1 does up to and including the definition of λ_j in (5.10). The argument in Step 2 following (5.10) is different from that of Step 2 in Theorem 1. The reader is cautioned to keep in mind the order in which various subsequences are chosen.

Define functions σ_j and ω_j corresponding to ψ_j and λ_j as follows:

$$\begin{aligned} \sigma_j(t) &= \sum_{i=1}^{k} \alpha_{ij} f(t, z(t), u_{n_j+i}(t)), \\ \omega_j(t) &= \sum_{i=1}^{k} \alpha_{ij} f^0(t, z(t), u_{n_j+i}(t)). \end{aligned} \tag{6.3}$$

The functions σ_j and ω_j are measurable. Let $\hat{f}_s(t) = f(t, z(t), u_s(t))$ and $f_s(t)$

$= f(t, z_s(t), u_s(t))$. Then using (6.3), (5.8) and (6.1) we may write

$$\begin{aligned}\int_G |\sigma_j - \psi_j|\, dt &\leqq \sum_{i=1}^{k} \alpha_{ij} \int_G |\hat{f}_{n_j+i} - f_{n_j+i}|\, dt \\ &\leqq \sum_{i=1}^{k} \alpha_{ij} \int_G \mu(|z_{n_j+i}(t) - z(t)|) \\ &\qquad \cdot \{H_{n_j+i}(t) + K|z_{n_j+i}(t) - z(t)|^{p-1}\}\, dt.\end{aligned} \tag{6.4}$$

We next show that there exist subsequences such that

$$\sigma_j(t) - \psi_j(t) \to 0, \qquad \omega_j(t) - \lambda_j(t) \to 0 \quad \text{a.e.} \tag{6.5}$$

We henceforth take the functions in (5.8), (5.10), and (6.3) to be the functions in these subsequences.

Let $\varepsilon > 0$ be given. Since $\|H_k\|_q \leqq A$ for all k, it follows from Hölder's inequality when $q < \infty$ and from the definition of $\|\cdot\|_\infty$ when $q = \infty$, that there exists an $\eta > 0$ and $< \varepsilon$ such that for every measurable subset $G' \subset G$ with meas $G' < \eta$,

$$\int_{G'} H_k\, dt < \varepsilon. \tag{6.6}$$

Since $z_k(t) \to z(t)$ almost everywhere in G, it follows from Egoroff's theorem that there is a measurable set $G_1 \subset G$ with meas $G_1 < \eta$ such that $z_k \to \dot{z}$ uniformly on $G_2 \equiv G - G_1$. Since $\mu(\delta) \to 0$ as $\delta \to 0$, it follows that for k sufficiently large and all t in G_2,

$$\mu(|z_k(t) - z(t)|) < \varepsilon, \tag{6.7}$$

where ε is as in the preceding paragraph.

Let $G_3 = \{t : t \in G_1, |z_k(t) - z(t)| \leqq \delta_0\}$ and let $G_4 = \{t : t \in G_1, |z_k(t) - z(t)| > \delta_0\}$, where δ_0 is as in Assumption C(iii)(b). Then by Assumption C(iii),

$$\begin{aligned}\mu(|z_k(t) - z(t)|) &\leqq \mu(\delta_0) \quad \text{for } t \in G_3, \\ \mu(|z_k(t) - z(t)|) &\leqq |z_k(t) - z(t)| \quad \text{for } t \in G_4.\end{aligned} \tag{6.8}$$

Since G is the union of the sets G_2, G_3, G_4 and these sets are pairwise disjoint, it follows from (6.6)–(6.8), and the convergence of z_k to z in L_p that for j sufficiently large,

$$\begin{aligned}&\int_G H_{n_j+i}(t)\mu(|z_{n_j+i}(t) - z(t)|)\, dt \\ &\leqq \varepsilon \int_{G_2} H_{n_j+i}\, dt + \mu(\delta_0) \int_{G_3} H_{n_j+i}\, dt + \int_{G_4} H_{n_j+i}|z_{n_j+i} - z|\, dt \\ &\leqq \varepsilon\{\|H_{n_j+i}\|_q(\text{meas } G)^{1/p} + \mu(\delta_0)\} + \|z_{n_j+i} - z\|_p \|H_{n_j+i}\|_q \\ &\leqq \varepsilon[A(1 + (\text{meas } G)^{1/p}) + \mu(\delta_0)].\end{aligned} \tag{6.9}$$

It also follows that for j sufficiently large,

$$\begin{aligned}
&\int_G \mu(|z_{n_j+i} - z|)|z_{n_j+i} - z|^{p-1}\,dt \\
&\leqq \varepsilon \int_{G_2} |z_{n_j+i} - z|^{p-1}\,dt + \mu(\delta_0)\int_{G_3} |z_{n_j+i} - z|^{p-1}\,dt \\
(6.10)\qquad &\quad + \int_{G_4} |z_{n_j+i} - z|^p\,dt \\
&\leqq \varepsilon\{\varepsilon^{p-1}\operatorname{meas}(G) + \mu(\delta_0)\delta_0^{p-1}\} + (\|z_{n_j+i} - z\|_p)^p \\
&\leqq \varepsilon[\varepsilon^{p-1}\operatorname{meas}(G) + \mu(\delta_0)\delta_0^{p-1} + 1],
\end{aligned}$$

where we recall that meas $(G_3) < \eta < \varepsilon$.

Combining (6.9) and (6.10) with (6.4) we get that for j sufficiently large $\sigma_j - \psi_j$ is in $L_1(G)$ and that $\sigma_j - \psi_j \to 0$ in $L_1(G)$. We apply similar arguments to $\omega_j - \lambda_j$ and obtain the existence of subsequences for which (6.5) holds.

We now define a function λ by (5.11) as in the proof of Theorem 1. The same argument that was used in the proof of Theorem 1 shows that λ is integrable, that (5.4) holds, and that λ is finite almost everywhere.

We now show that $(\lambda(t), y(t)) \in Q^+(t, z(t))$ for a.e. t. Let the set G' be as in the proof of Theorem 1. Let G'' denote the set of points in G' at which (6.5) holds. Clearly, meas G'' = meas G.

Let t be an arbitrary point in G''. Since $\psi_j(t) \to y(t)$, it follows from (6.5) that $\sigma_j(t) \to y(t)$. From the definition of λ it follows that there is a subsequence $\{\lambda_j(t)\}$ which will in general depend on t such that $\lambda_j(t) \to \lambda(t)$. By virtue of (6.5), $\omega_j(t) \to \lambda(t)$. For the corresponding subsequence $\{\sigma_j(t)\}$ we still have $\sigma_j(t) \to y(t)$. From the definition of G'', it follows that for all j and i,

$$u_{n_j+i}(t) \in \Omega(t, z_{n_j+i}(t)) = \Omega(t, z(t)),$$

the last equality being a consequence of Assumption C(ii). Hence,

$$(f^0(t, z(t), u_{n_j+i}(t)), f(t, z(t), u_{n_j+i}(t))) \in Q^+(t, z(t)).$$

Since $Q^+(t, z(t))$ is convex, the points $(\omega_j(t), \sigma_j(t))$ belong to $Q^+(t, z(t))$. Since $Q^+(t, z(t))$ is closed and $(\omega_j(t), \sigma_j(t)) \to (\lambda(t), y(t))$, we get that $(\lambda(t), y(t)) \in Q^+(t, z(t))$. Since t was an arbitrary point of G'', it follows that $(\lambda(t), y(t)) \in Q^+(t, z(t))$ a.e.

From this point onward, the proof of Theorem 2 is the same as the proof of Theorem 1.

We note that if we have $z_k \to z$ in L_∞, then the argument to establish the existence of subsequences for which (6.5) holds can be simplified as follows.

If $z_k \to z$ in L_∞, then since μ is nondecreasing we can continue the chain of inequalities in (6.4) and get

$$\int_G |\sigma_j - \psi_j|\,dt \leqq \sum_{i=1}^{k} \alpha_{ij}\mu(\|z_{n_j+i} - z\|_\infty)\{A + K \operatorname{meas}(G)\|z_{n_j+i} - z\|_\infty^{p-1}\}.$$

Thus $\sigma_j - \psi_j$ is in $L_1(G)$. Since $\|z_{n_j+i} - z\|_\infty \to 0$ and $\mu(\delta) \to 0$ as $\delta \to 0$ we get that $\sigma_j - \psi_j \to 0$ in $L_1(G)$. A similar argument shows that $\omega_j - \lambda_j$ is in $L_1(G)$ and that $\omega_j - \lambda_j \to 0$ in $L_1(G)$. Hence there exist subsequences such that (6.5) holds.

7. The existence theorem. From the discussion in § 3, it is apparent that in order to formulate an existence theorem, hypotheses must be made to ensure that a lower closure theorem holds, that minimizing sequences are conditionally weakly compact, and that if $\phi_k \to \phi$ weakly in $\mathscr{F}$, where $\{(\phi_k, u_k)\} \in \mathscr{A}_0$, then (ϕ, u) is in $\mathscr{A}_0$ as well as in $\mathscr{A}$.

The specification of a subclass $\mathscr{A}_0$ involves conditions in addition to (2.1), such as boundary conditions. Because of the variety of conditions used to define subclasses $\mathscr{A}_0$ and the variety of special problems and classes of problems, it is not feasible to formulate one set of hypotheses that will guarantee that (ϕ, u) is always in $\mathscr{A}_0$. It is also not advisable to formulate a different theorem for each individual subclass $\mathscr{A}_0$. Instead, the notion of a closed subclass $\mathscr{A}_0$ is used and the general existence theorem is stated in terms of this notion, which will be defined presently. The verification in a particular problem that $\mathscr{A}_0$ is a closed subclass is to be carried out when the general theorem is applied to that problem.

DEFINITION. A sequence of admissible pairs $\{(\phi_k, u_k)\}$ is said to be *weakly lower closed in $\mathscr{F}$ with respect to J*, or simply *weakly lower closed*, whenever the following holds. If $\phi_k \to \phi$ weakly in $\mathscr{F}$, then there exists a measurable function $u: G \to R^m$ such that (ϕ, u) is admissible and

$$\liminf_{k \to \infty} J(\phi_k, u_k) \geqq J(\phi, u), \tag{7.1}$$

where J is defined by (2.2).

Theorems 1 and 2 give sufficient conditions for a sequence to be weakly lower closed with respect to J.

DEFINITION. A subclass $\mathscr{A}_0$ of admissible pairs is said to be *closed* whenever the following holds. If $\{(\phi_k, u_k)\}$ is a sequence of admissible pairs in $\mathscr{A}_0$ that is weakly lower closed and $\phi_k \to \phi$ weakly in $\mathscr{F}$, then for at least one u such that (ϕ, u) is admissible and (7.1) holds it is also true that $(\phi, u) \in \mathscr{A}_0$.

In control problems governed by systems of ordinary differential equations, the subclasses $\mathscr{A}_0$ are often determined by imposing conditions on the trajectories such as the following. The graphs of all trajectories are required to lie in a certain compact set. All trajectories are required to have at least one point in a preassigned compact set. The endpoints of all trajectories are required to lie in a given compact set. As noted in the discussion following Assumption A, the weak convergence in $\mathscr{F}$ of $\{\phi_k\}$ in this case implies the uniform convergence of the ϕ_k in $[a, b]$. Thus if $\mathscr{A}_0$ is defined by any of the conditions listed above, then $\mathscr{A}_0$ will be a closed subclass.

For different special classes of problems, different Banach spaces $\mathscr{F}$ are appropriate. Therefore, in stating the general existence theorem it is not possible to impose a usable condition that will ensure conditional weak compactness of minimizing sequences. Specific conditions will be left to specific applications. We will, however, discuss briefly two important special problems after the statement of the theorem.

If $\mathscr{A}_0$ is a subclass of admissible pairs (ϕ, u), then $\mathscr{A}_{OT}$ will designate the set of trajectories ϕ such that $(\phi, u) \in \mathscr{A}_0$ for some u.

We now state our general existence theorem.

THEOREM 3. *Let $\mathscr{A}_0$ be a closed subclass of admissible pairs such that $\mathscr{A}_{OT}$ is weakly sequentially compact in $\mathscr{F}$. Let any minimizing sequence in $\mathscr{A}_0$ with weakly* *convergent trajectories be weakly lower closed with respect to J. Then J attains its minimum on* $\mathscr{A}_0$.

Proof. The proof of Theorem 3 proceeds as outlined in § 3. Note that for a minimizing sequence, the number μ defined in (3.1) is equal to γ defined in (5.2), so that μ is finite.

Theorem 3 is a "do it yourself theorem" in the sense that we have not specified conditions for $\mathscr{A}_0$ to be a closed subclass or for $\mathscr{A}_{0T}$ to be weakly sequentially compact. As already noted, the present problem is too general for this to be done in a sensible manner. These conditions must therefore be supplied when problems that are more specialized are attacked. Theorems 1 and 2, however, do give usable and verifiable sufficient conditions for minimizing sequences with weakly convergent trajectories to be weakly lower closed. These theorems constitute the principal contribution of this paper.

We now discuss criteria for weak sequential compactness of $\mathscr{A}_{0T}$ in problems governed by systems of ordinary differential equations and in distributed parameter systems.

First consider control problems governed by systems of ordinary differential equations. Here $v = 1$ and the appropriate space $\mathscr{F}$ is $H_1^1(\bar{G}) = H_1^1([a, b])$, where we take each equivalence class of functions in $H_1^1(G)$ to be represented by its absolutely continuous member. As usually stated, existence theorems for such problems involve a "growth condition" that ensures the conditional compactness in $C([a, b])$ of any minimizing sequence of trajectories $\{\phi_k\}$. Actually, the growth conditions imply the equiabsolute continuity of the sequence of integrals $\{\int_E \phi_k' \, dt\}$. Since $v = 1$, this implies that the sequence $\{\phi_k\}$ is conditionally compact in $C([a, b])$. The equiabsolute continuity of the integrals $\int_E \phi_k' \, dt$, however, also implies the conditional weak compactness in $H_1^1([a, b])$ of the sequence $\{\phi_k\}$. In [6] Cesari introduced the following condition guaranteeing the equiabsolute continuity of the sequence $\{\phi_k\}$. It is assumed that $f^0 \geqq 0$. For every $\varepsilon > 0$ there exists a nonnegative integrable function ψ_ε on $[a, b]$ such that $|f(t, x, w)| \leqq \psi_\varepsilon(t) + f^0(t, x, w)$ for all (t, x, w) in $\mathscr{D}$. For generalizations and further discussions of this condition see [6] and [9].

In distributed control problems it is appropriate to consider functions ϕ in $H_p^s(G)$ or functions whose components ϕ^i are in $H_{p_i}^{s_i}(G)$, $i = 1, \cdots, j$. Thus $\mathscr{F} = H_p^s(G)$ or $\mathscr{F} = \prod_{i=1}^j H_{p_i}^{s_i}(G)$. If all $p_i > 1$, the conditional weak compactness in $\mathscr{F}$ of $\{\phi_k\}$ is ensured merely by requiring that the sequence $\{\|\phi_k\|\}$ be bounded. (Recall that $\|\cdot\|$ denotes the norm in $\mathscr{F}$.) If for some index i we have $p_i = 1$, then to ensure conditional weak compactness we need to assume that the various integrals

$$\int_E D^\alpha \phi_k^i \, dt, \qquad 0 \leqq |\alpha| \leqq s, \quad E \subset G$$

are equiabsolutely continuous. For those indices i such that $p_i > 1$, the boundedness of the sequence of norms $\{\|\phi_k^i\|_i\}$ is required, where $\|\cdot\|_i$ denotes the norm in $H_{p_i}^{s_i}(G)$. For further details see [5] and [7].

Throughout this paper we have assumed that the sets $Q^+(t, x)$ are convex. If the sets $Q^+(t, x)$ are not convex, one can replace the original problem with a "relaxed problem" in which the vectors $(N\phi)(t)$ lie in co $Q^+(t, x)$. The relaxed problem is cast as a new control problem in which the set that plays the role of $Q^+(t, x)$ is convex. An existence theorem for the relaxed problem is then easily obtained. For details see [6] and [7].

REFERENCES

[1] L. D. BERKOVITZ, *Existence theorems in problems of optimal control*, Studia Math., 44 (1972), pp. 275–285.

[2] ———, *Existence theorems in problems of optimal control without property* (Q), Techniques of Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972, pp. 197–209.

[3] L. CESARI, *Existence theorems for optimal solutions in Pontryagin and Lagrange problems*, this Journal, 3 (1966), pp. 475–498.

[4] ———, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412.

[5] ———, *Existence theorems for multidimensional Lagrange problems*, J. Optimization Theory Appl., 1 (1967), pp. 87–111.

[6] ———, *Existence theorems for optimal controls of the Mayer type*, this Journal, 6 (1968), pp. 517–552.

[7] ———, *Existence theorems for abstract multidimensional control problems*, J. Optimization Theory Appl., 6 (1970), pp. 210–236.

[8] ———, *Closure, lower closure, and semicontinuity theorems in optimal control*, this Journal, 9 (1971), pp. 287–315.

[9] L. CESARI, J. R. LAPALM AND T. NISHIURA, *Remarks on some existence theorems for optimal control*, J. Optimization Theory Appl., 3 (1969), pp. 296–305.

[10] G. FICHERA, *Semicontinuity of multiple integrals in ordinary form*, Arch. Rational Mech. Anal., 17 (1964), pp. 339–352.

[11] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-Groups*, rev. ed., American Mathematical Society, Providence, R.I., 1957.

[12] M. Q. JACOBS, *Attainable sets in systems with unbounded controls*, J. Differential Equations, 4 (1968), pp. 408–423.

[13] K. KURATOWSKI, *Topologie II*, 3rd ed., Pańrstwowe Wydawnictwo Naukowe, Warsaw, 1961.

[14] E. J. MCSHANE AND R. B. WARFIELD, JR., *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.

[15] C. B. MORREY, JR., *Multiple integral problems in the calculus of variations and related topics*, Univ. of California Publ. in Math. new ser., 1 (1943), pp. 1–130.

[16] ———, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, 1966.

[17] C. OLECH, *Existence theorems for optimal control problems with vector valued cost function*, Trans. Amer. Math. Soc., 136 (1969), pp. 159–180.

[18] G. SCORZA-DRAGONI, *Un teorema sulle funzioni continue rispetto ad una e misurabile rispetto ad un'altra variabile*, Rend. Sem. Mat. Univ. Padova, 17 (1948), pp. 102–106.

[19] M. M. VAINBERG, *Variational Methods for the Study of Nonlinear Operators*, Holden-Day, San Francisco, 1964.

SIAM J. CONTROL
Vol. 12, No. 1, February 1974

A GENERALIZATION OF CHOW'S THEOREM AND THE BANG-BANG THEOREM TO NONLINEAR CONTROL PROBLEMS*

ARTHUR J. KRENER†

Abstract. The main results of this paper are two-fold. The first, Theorem 1, is a generalization of the work of Chow and others concerning the set of locally accessible points of a nonlinear control system. It is shown that under quite general conditions, this set lies on a surface in state space and has a nonempty interior in the relative topology of that surface.

The second result, Theorem 3, generalizes the bang-bang theorem to nonlinear control systems using higher order control variations as developed by Kelley and others. As a corollary we obtain Halkin's bang-bang theorem for a linear piecewise analytic control system.

1. Introduction.

Consider the control system

$$\dot{x} = f(x(t), u(t)), \qquad x(0) = x^0, \qquad u(t) \in \Omega, \tag{1}$$

where $x = (x_1, \cdots, x_n)$ are coordinates of the state space, M is a paracompact n-dimensional manifold, $u = (u_1, \cdots, u_k)$ is the control, $x^0 = (x_1^0, \cdots, x_n^0)$ is the initial state, $\Omega \subseteq \mathbb{R}^k$ is the set of admissible controls, and f is an n-vector-valued function. We assume that $x_1 = t$ so that the first coordinate of f is identically 1; also we assume that f is C^∞ with respect to $x_2, \cdots, x_n, u_1, \cdots, u_k$ and piecewise C^∞ with respect to x_1. We require that $u(t)$ be a piecewise C^∞-function of $t = x_1$. The requirement of C^∞ differentiability is not essential, it is only to avoid counting the degree of differentiability required in any argument. The tangent space to M at x is denoted by M_x. A control $u(t)$ defines a vector field, $f_u(x) = f(x, u(x_1))$ on M; and given two controls $u(t)$, $v(t)$, we can define a new vector field by means of the Lie bracket,

$$[f_u, f_v](x) = \frac{\partial f_v}{\partial x}(x) f_u(x) - \frac{\partial f_u}{\partial x}(x) f_v(x), \tag{2}$$

where $(\partial f_v/\partial x)(x)$ is an $n \times n$ matrix of partial derivatives at x.

A slight problem arises since f_u, f_v are only piecewise C^∞-functions of x_1, but at those values of x_1 we can consider (2) as either undefined or as double-valued by taking left and right limits. Since the difficulties that arise because of this can be dealt with by simple but lengthy arguments, we shall ignore them.

2. Integrability and semi-integrability.

The set, $V(M)$, of all C^∞-vector fields on M is a module over the ring, $C(M)$, of all C^∞-real-valued functions with domain M, with addition and multiplication defined pointwise. With the definition of the bracket (2), $V(M)$ becomes a Lie algebra of infinite dimension over the field, $\mathbb{R}$. Suppose H is a submodule of $V(M)$. We define $H_x = \{f(x) : f \in H\}$. Let U be an open subset of M and L a submanifold of U. L is an *integral manifold* of H in U if L is connected and $H_x = L_x$ for all $x \in L$ (L_x is the tangent space to L at x). An integral manifold of H in U is always contained in a maximal integral manifold of H in U. H is *integrable on* U if there exists a partition of U by maximal integral

* Received by the editors March 20, 1972, and in revised form December 30, 1972.

† Department of Mathematics, University of California–Davis, Davis, California 95616. This research was supported in part by the Office of Naval Research.

manifolds of H in U. For H to be integrable on U a necessary condition is that H restricted to U be a subalgebra of $V(U)$. If, in addition, H satisfies one of the following then H is integrable on U;

(i) Frobenius. The dimension of H_x is constant for all $x \in U$.

(ii) Hermann [9]. H is *locally finitely generated*, that is, for all $x \in U$, there exists a neighborhood $V \subseteq U$ of x such that H restricted to V is the $C(V)$ span of a finite number of vector fields of H. (Lobry [14] has a slightly weaker form of this condition.)

(iii) Nagano [16]. M is a real analytic manifold and H is a subalgebra of the Lie algebra of real analytic vector fields on U.

If H is a submodule but not a subalgebra, then there exists a smallest subalgebra containing H, which we denote by DH. We can construct DH as follows. We define $D^0H = H$ and $D^kH = D^{k-1}H + [H, D^{k-1}H]$. For example, D^1H is the submodule of all linear combinations of vector fields of H and Lie brackets of vector fields of H with coefficients from $C(M)$. DH is the union of this ascending sequence of submodules.

Suppose U is an open neighborhood of $x^0 = (x_1^0, \cdots, x_n^0)$. Then we split U into two open halves,

$$U^+ = \{x \in U : x_1 > x_1^0\} \quad \text{and} \quad U^- = \{x \in U : x_1 < x_1^0\}.$$

The control system (1) is *locally semi-integrable* if for all $x^0 \in M$, there exists an open neighborhood U of x^0 and submodules H^+, H^- of $V(M)$ such that

(i) $H_x^+ = \text{span}\,\{f(x, u) : u \in \Omega\} \subseteq M_x$ for all $x \in U^+$,
$H_x^- = \text{span}\,\{f(x, u) : u \in \Omega\} \subseteq M_x$ for all $x \in U^-$.

(ii) DH^+ and DH^- are integrable on U with maximal integrable manifolds L^+ and L^- in U containing x^0.

Suppose $u(t)$ is an admissible control and $\gamma_u(s)x$ is the family of integral curves of the vector field $f_u(x)$, that is, $\gamma_u(0)x = x$ and $(d/ds)\gamma_u(s)x = f_u(\gamma_u(s)x)$. We define the *set* $\mathscr{A}(x^0, U)$ *of all points accessible from* x^0 *in* U as $\{\gamma_u(s)x^0 : s \geqq 0, u(t) \in \Omega$, and $\gamma_u(r)x^0 \in U$ for all $r \in [0, s]\}$ and the *set* $\mathscr{C}(x^0, U)$ *of all points controllable to* x^0 *in* U as $\{\gamma_u(s)x^0 : s \leqq 0, u(t) \in \Omega$, and $\gamma_u(r)x^0 \in U$ for all $r \in [s, 0]\}$.

If (1) is locally semi-integrable in U, then it is easily shown that $\mathscr{A}(x^0, U) \subseteq L^+$ and $\mathscr{C}(x^0, U) \subseteq L^-$.

We now raise the question whether $\mathscr{A}(x^0, U)$ is "thick" in L^+, i.e., whether $\mathscr{A}(x^0, U)$ has any interior as a subset of L^+. The answer is affirmative as the following generalization of the work of Chow [2], Lobry [14], and Sussmann and Jurdjevic [19] shows.

THEOREM 1. *Assume* (1) *is a locally semi-integrable control system and* x^0, U, L^+ *and* L^- *are as above. Then the* L^+*-interior of* $\mathscr{A}(x^0, U)$ *and the* L^-*-interior of* $\mathscr{C}(x^0, U)$ *are nonempty.*

Proof. In theorems of this type we shall only prove one assertion since the proof of the other is identical. We construct inductively a sequence of maps $\varphi_j : V^j \to \mathscr{A}(x^0, U) \subseteq L^+$ defined on a sequence of open sets $V^j \subseteq \mathbb{R}^j$ such that the image $N^j = \varphi_j(V^j)$ is a submanifold of dimension j. We continue until j equals the dimension of L^+.

Choose any control, say $u^1(t) = (u_1^1(t), \cdots, u_k^1(t))$, and let $f_1(x)$ be the vector field $f_1(x) = f(x, u^1(x_1))$. Let $\delta > 0$ such that the integral curve $s_1 \mapsto \gamma_1(s_1)x^0$

of f_1 is C^∞ for $s_1 \in (0, \delta)$. Let $V^1 = (0, \delta)$ and $\varphi_1(s_1) = \gamma_1(s_1)x^0$. Since the first coordinate of f_1 is identically 1, the image $N^1 = \varphi_1(V^1)$ is a one-dimensional submanifold of L^+.

Suppose we have constructed $N^{j-1} = \varphi_{j-1}(V^{j-1})$ and $j \leqq$ dimension of L^+. Choose $x \in N^{j-1}$ and a control u^j such that $f_j(x) = f(x, u^j(x_1)) \notin N_x^{j-1}$, the tangent space to N^{j-1} at x. This can always be done, for if not, then for all $x \in N^{j-1}$ and for all $u \in \Omega$, $f_u(x) \in N_x^{j-1}$. This implies that $H_x \subseteq N_x^{j-1}$ for all $x \in N^{j-1}$, and the set of vector fields on N^{j-1}, $V(N^{j-1})$, is an algebra; therefore DH restricted to N^{j-1} is contained in $V(N^{j-1})$. But this implies that $j \leqq$ dimension of L^+ = dimension $DH_x \leqq$ dimension of $N_x^{j-1} = j - 1$.

By passing to a smaller V^{j-1} and N^{j-1} we can assume that $f_j(x) \notin N_x^{j-1}$ for all $x \in N^{j-1}$ and also for some $\delta > 0$, the integral curve of f_j starting at x satisfies $\gamma_j(s_j)x \in U$ for all $x \in N^{j-1}$ and $0 < s_j < \delta$. We define $\varphi_j(s_1, \cdots, s_j) = \gamma_j(s_j)\varphi_{j-1}(s_1, \cdots, s_{j-1})$, $V^j = V^{j-1} \times (0, \delta)$ and $N^j = \varphi_j(V^j) \subseteq \mathscr{A}(x^0, U)$. The Jacobian $(\partial\varphi_j/\partial s)(s_1, \cdots, s_{j-1}, 0)$ is nonsingular for every $(s_1, \cdots, s_{j-1}) \in V^{j-1}$ and hence for δ sufficiently small $\varphi_j : V^j \to N^j$ is a diffeomorphism. Q.E.D.

Example 1. Suppose $M = \mathbb{R}^2$ and consider the control system $\dot{x}_1 = 1$, $\dot{x}_2 = u \cdot g(x_1)$, $|u| \leqq 1$, where $g(x_1)$ is a C^∞-(or pwC^∞)-function satisfying $g(x_1) = 0$ if $x_1 \leqq 0$ and $g(x_1) > 0$ if $x_1 > 0$. The system is locally semi-integrable; for example, if $x^0 = (0, 0)$, then we take $U = M$, $H^+ = DH^+ = V(M)$, $L^+ = M$,

$$\mathscr{A}(x^0, U) = \left\{(x_1, x_2) : x_1 \geqq 0, |x_2| \leqq \int_0^{x_1} g(x_1)\, dx_1\right\},$$

$$H^- = DH^- = \left\{\begin{pmatrix} h(x) \\ 0 \end{pmatrix} : h(x) \in C(M)\right\},$$

$$L^- = \{(x_1, 0) : x_1 \in \mathbb{R}\} \quad \text{and} \quad \mathscr{C}(x^0, U) = \{(x_1, 0) : x_1 \leqq 0\}.$$

The system $\dot{x}_1 = 1$, $\dot{x}_2 = u \cdot g(x_2)$ is not locally semi-integrable.

Example 2. Suppose $M = \mathbb{R}^2$ and $\dot{x}_1 = 1$, $\dot{x}_2 = ux_2$, $|u| \leqq 1$. The submodule $H = \{(h_1(x), x_2h_2(x)) : h_i(x) \in C(M)\}$ is an integrable subalgebra and carries the system everywhere; that is, for each $x \in M$, $H_x = \operatorname{span}\{f(x, u) : |x| \leqq 1\}$. It partitions M into 3 integral manifolds:

$$L^1 = \{(x_1, x_2) : x_2 > 0\}, \qquad L^2 = \{(x_1, x_2) : x_2 = 0\}$$

and

$$L^3 = \{(x_1, x_2) : x_2 < 0\}.$$

Suppose $x^0 = (0, 1) \in L^1$. Then $U = M$, $H^+ = DH^+ = H^- = DH^- = H$, $L^+ = L^- = L^1$, $\mathscr{A}(x^0, U) = \{(x_1, x_2) : x_1 \geqq 0, e^{-x_1} \leqq x_2 \leqq e^{x_1}\}$ and $\mathscr{C}(x^0, U) = \{(x_1, x_2) : x_1 \leqq 0,\ e^{x_1} \leqq x_2 \leqq e^{-x_1}\}$. If $x^0 = (0, 0) \in L^2$, then $U = M$, $H^+ = DH^+ = H^- = DH^- = H$, $L^+ = L^- = L^2$, $\mathscr{A}(x^0, U) = \{(x_1, 0) : x_1 \geqq 0\}$ and $C(x^0, U) = \{(x_1, 0) : x_1 \leqq 0\}$.

Example 3. Suppose $M = \mathbb{R}^3$ and $\dot{x}_1 = 1$, $\dot{x}_2 = u$, $\dot{x}_3 = ux_1$, $|u| \leqq 1$. Let $f_1(x), f_{-1}(x)$ be the vector fields corresponding to the constant controls $u = \pm 1$.

Let H be the submodule which is the $C(M)$ span of f_1 and f_{-1}. There is one linearly independent bracket

$$[f_1, f_{-1}](x) = \begin{pmatrix} 0 \\ 0 \\ -2 \end{pmatrix},$$

so DH is spanned by f_1, f_{-1} and $[f_1, f_{-1}]$. The integrable manifold of DH through any point x^0 is M. The sets $\mathscr{A}(x^0, M) = \{\gamma_1(s_3)\gamma_{-1}(s_2)\gamma_1(s_1)x^0 : s_i \geqq 0\}$ and $\mathscr{C}(x^0, M) = \{\gamma_1(s_3)\gamma_{-1}(s_2)\gamma_1(s_1)x^0 : s_i \leqq 0\}$ both have nonempty interior.

3. The bang-bang theorem.

Henceforth we shall consider the system

$$(3) \qquad \begin{gathered} \dot{x} = \sum_{i=0}^{k} u_i(t)a_i(x), \\ x(0) = x^0, \quad u_i(t) \geqq 0, \quad \sum u_i(t) = 1, \end{gathered}$$

where $a_0, \cdots, a_k$ are vector-valued functions C^∞ with respect to $x_2, \cdots, x_n$ and piecewise C^∞ with respect to $x_1 = t$. The controls, $u(t)$, are piecewise C^∞-functions of $t = x_1$, lying in the compact convex set $\Omega = \{u : u_i \geqq 0, \sum u_i = 1\}$. We let E denote the set of extreme points of Ω. E is the set of unit vectors, $(0, \cdots, 0, 1, 0, \cdots, 0)$, in $\mathbb{R}^k$. We call Ω the set of *admissible* controls and E the set of *bang-bang* controls. We alter our notation to distinguish between the set of points, $\mathscr{A}(x^0, U, \Omega)$, accessible in U from x^0 by admissible controls, and the set of points, $\mathscr{A}(x^0, U, E)$, accessible in U from x^0 by bang-bang controls. We adopt a similar convention regarding $\mathscr{C}(x^0, U, \Omega)$ and $\mathscr{C}(x^0, U, E)$. The bang-bang question is, under what conditions is it true that $\mathscr{A}(x^0, U, E) = \mathscr{A}(x^0, U, \Omega)$ and $\mathscr{C}(x^0, U, E) = \mathscr{C}(x^0, U, \Omega)$. It is well known that $\mathscr{A}(x^0, U, E) \subseteq \mathscr{A}(x^0, U, \Omega) \subseteq$ closure $\mathscr{A}(x^0, U, E)$ and $\mathscr{C}(x^0, U, E) \subseteq \mathscr{C}(x^0, u, \Omega) \subseteq$ closure $\mathscr{C}(x^0, U, E)$.

THEOREM 2. *Suppose* (3) *is locally semi-integrable and* U, L^+, L^- *are as above. Then* L^+-interior $\mathscr{A}(x^0, U, E) = L^+$-interior $\mathscr{A}(x^0, U, \Omega)$ *and* L^--interior $\mathscr{C}(x^0, U, E) = L^-$-interior $\mathscr{C}(x^0, U, \Omega)$.

Proof. To simplify the proof we restrict (3) to a control system on the manifold L^+, in other words we take $M = L^+$. Clearly interior $\mathscr{A}(x^0, U, E) \subseteq$ interior $\mathscr{A}(x^0, U, \Omega)$. To show the opposite inclusion we let $x \in$ interior $\mathscr{A}(x^0, U, \Omega)$. We choose an open connected neighborhood V of x such that $V \subseteq$ interior $\mathscr{A}(x^0, U, \Omega)$. The set of vector fields $\{f_u : u \in \Omega\}$ and $\{f_u : u \in E\}$ generate the same submodule H and hence by Theorem 1, $\mathscr{C}(x, V, E)$ has a nonempty interior. Let $y \in$ interior $\mathscr{C}(x, V, E) \subseteq V \subseteq \mathscr{A}(x^0, U, \Omega) \subseteq$ closure $\mathscr{A}(x^0, U, E)$. Then there is a sequence $y^m \in \mathscr{A}(x^0, U, E)$, such that y^m converges to y. For m sufficiently large, $y^m \in$ interior $\mathscr{C}(x, V, E)$, so y^m is bang-bang accessible from x^0 and bang-bang controllable to x. This implies $x \in \mathscr{A}(x^0, U, E)$. Q.E.D.

From Theorem 2 it is clear that $\mathscr{A}(x^0, U, E)$ will equal $\mathscr{A}(x^0, U, \Omega)$ if every admissible trajectory which does not come from a bang-bang control goes to an interior point of $\mathscr{A}(x^0, U, \Omega)$. To decide when this will happen we study the effect of control variations.

Let $u^j(x_1)$ be an admissible control, $f_j(x) = \sum u_i^j(x_1)a_i(x)$ and $\gamma_j(s)x$ be the family of integral curves of $f_j(x)$ for $j = 0, 1$. Suppose as we approach $x = \gamma_0(s)x^0 \in \mathscr{A}(x^0, U, \Omega)$ using the control u^0, we replace u^0 with u^1 for r units of time. The result is a trajectory whose endpoint is $\gamma_1(r)\gamma_0(s - r)x^0 = \gamma_1(r)\gamma_0(-r)x$. If we vary r through small nonnegative values, we obtain a C^∞-curve $q(r) = \gamma_1(r)\gamma_0(-r)x$ satisfying $q(0) = x$. To compute the derivative from the right at 0, we define $q(r_0, r_1) = \gamma_1(r_1)\gamma_0(-r_0)x$. Then

$$\frac{dq(0_+)}{dr} = \frac{\partial q(0)}{\partial r_1} - \frac{\partial q(0)}{\partial r_0} = f_1(x) - f_0(x).$$

If we continue to $x^1 = \gamma_0(s_1)x^0 \in \mathscr{A}(x^0, U, \Omega)$ using the control u^0, we can define a new curve $q(r) = \gamma_0(s_1 - s)\gamma_1(r)\gamma_0(-r)x$. This is also C^∞ for small nonnegative r and $q(0) = x^1$. The derivative from the right at 0 is

$$\text{(4)} \qquad \frac{dq(0_+)}{dr} = \gamma_0(s_1 - s)_*(f_1(x) - f_0(x)),$$

where $\gamma_0(s_1 - s)_*$ is the tangent space map induced by the map $x \mapsto \gamma_0(s_1 - s)x$. If $f_0(x)$ and $f_1(x)$ are C^∞ in a neighborhood of the trajectory joining x and x^1, then (4) can be expressed in a Taylor series,

$$\text{(5)} \qquad \frac{dq(0_+)}{dr} = \sum_{m=0}^{h} \frac{(s - s_1)^m}{m!} ad^m(f_0)(f_1 - f_0)(x^1) + \mathcal{O}(s - s_1)^{h+1},$$

where

$$ad^0(f_0)(f_1 - f_0)(x^1) = f_1(x^1) - f_0(x^1),$$

$$ad^m(f_0)(f_1 - f_0)(x^1) = [f_0, ad^{m-1}(f_0)(f_1 - f_0)](x^1)$$

and $\mathcal{O}(s - s_1)^{h+1}$ is an error term of order $(s - s_1)^{h+1}$.

The second type of control variation is similar to the one introduced by Kelley [11].

Suppose u^0, u^1, u^2, u^3 are admissible controls such that $u^0 = (2u^1 + u^2 + u^3)/4$. Then $f_0(x) = (2f_1(x) + f_2(x) + f_3(x))/4$. For ease of notation we introduce another control $u^4 = u^1$ so $u^0 = (\sum_{i=1}^4 u_i)/4$, $f_0(x) = (\sum_{i=1}^4 f_i(x))/4$. Consider the control modification $p(r)$ made at $x \pm \gamma(s)x^0$, where $p(r) = \gamma_4(r) \cdot \gamma_3(r)\gamma_2(r)\gamma_1(r)\gamma_0(-4r)x$.

To compute the first two derivatives of this curve, we introduce new variables $r_0 = -4r, r_1 = r_2 = r_3 = r_4 = r$ and use the chain rule

$$\frac{dp(0_+)}{dr} = \sum_{i=0}^{4} \frac{dr_i}{dr}\frac{\partial p(0)}{\partial r_i} = f_1(x) + f_2(x) + f_3(x) + f_4(x) - 4f_0(x) = 0,$$

$$\begin{aligned} \frac{d^2p(0_+)}{dr^2} &= \sum_{i=0}^{4} \left(\frac{dr_i}{dr}\right)^2 \frac{\partial^2 p(0)}{\partial r_i^2} + 2 \sum_{0 \leq i < j \leq 4} \frac{dr_i}{dr}\frac{dr_j}{dr}\frac{\partial^2 p(0)}{\partial r_i \partial r_j} \\ &= \sum_{i=0}^{4} \left(\frac{dr_i}{dr}\right)^2 \frac{\partial f_i(x)}{\partial x} f_i(x) + 2 \sum_{0 \leq i < j \leq 4} \frac{dr_i}{dr}\frac{dr_j}{dr}\frac{\partial f_j(x)}{\partial x} f_i(x) \\ &= [f_2, f_3](x). \end{aligned}$$

Since $dp(0_+)/dr = 0$, the curve $q(r)$ defined for small nonnegative r by $q(r^2/2) = p(r)$ is C^1 and $dq(0_+)/dr = d^2p(0_+)/dr^2$. We can pull this control modification along to $x^1 = \gamma_0(s_1)x^0$ as before and obtain

$$(6)\quad \gamma_0(s_1 - s)_*[f_2, f_3](x) = \sum_{m=0}^{h} \frac{(s - s_1)^m}{m!} ad^m(f_0)[f_2, f_3](x^1) + \mathcal{O}(s - s_1)^{h+1}.$$

Notice that if we reverse u^2 and u^3 in defining $p(r)$, we obtain $\gamma_0(s_1 - s)_*[f_3, f_2](x) = -\gamma_0(s_1 - s)_*[f_2, f_3](x)$.

The last type of control modification which we consider is to stop short of x^1 or continue on past x^1. These lead to curves $q(r) = \gamma_0(x_1 \pm r)x^0 = \gamma_0(\pm r)x^1$, whose derivatives are

$$(7)\qquad \frac{dq(0_+)}{dr} = \pm f_0(x^1).$$

Let K_{x^1} be the convex cone in $L^+_{x^1}$ generated by the vectors of the form (4), (6) and (7), for all $0 < s \leqq s_1$ and admissible controls $u^i(t)$, $i = 1, 2, 3$. We say the trajectory of u^0 between x^0 and x^1 is *singular* if K_{x^1} is a proper subset of $L^+_{x^1}$. This definition is different from the usual one stated in terms of the maximal principle (see Gabasov and Kirillova [6] and Hermes [20]). Since $t = x_1$, the usual one is equivalent to the following: the trajectory is singular if the cone generated by the vectors of the form (4) and (7) is a proper subset of M_{x^1}. There are of course less singular controls under our definition. It can be shown, using the standard methods (implicit function or fixed point theorem), that if $K_{x^1} = L^+_{x^1}$, then $x^2 \in L^+$-interior $\mathscr{A}(x^0, U, \Omega)$ and so is bang-bang accessible. It follows then that $\mathscr{A}(x^0, U, E)$ will equal $\mathscr{A}(x^0, U, \Omega)$ if the only singular trajectories are bang-bang. Consider the following examples.

Example 4. Let $M = \mathbb{R}^3$ and $\dot{x} = ua_1 + (1 - u)a_2$, $0 \leqq u \leqq 1$, where

$$a_1(x) = \begin{pmatrix} 1 \\ 1 \\ x_2 \end{pmatrix} \quad \text{and} \quad a_2(x) = \begin{pmatrix} 1 \\ -1 \\ -x_2 \end{pmatrix}.$$

H is the $C(M)$ span of a_1 and a_2 and since $[a_1, a_2] = 0$, $DH = H$. The integral manifold of H through $x^0 = (0, 0, 0)$ is $L = \{(x_1, x_2, x_3) : x_3 = (x_2)^2\}$. Let $u^0(t) = 1/2$ and $x^1 = \gamma_0(s_1)x^0 = (s_1, 0, 0)$. The cone K_{x^1}, generated by $\pm(1/2)(a_1(x^1) + a_2(x^1))$, $(a_1(x^1) - a_0(x^1))$, and $(a_2(x^1) - a_0(x^1)) = (a_0(x^1) - a_1(x^1))$ equals L_{x^1} so the trajectory is not singular in our sense. However, K_{x^1} is a proper subset of M_{x^1} and so the trajectory is singular in the usual sense. Notice that x^1 is bang-bang accessible, $x^1 = \gamma_2(s_1/2)\gamma_1(s_1/2)x^0$ or any other bang-bang trajectory that uses a_1 and a_2 each a total of $s_1/2$ units of time.

Example 5. Let $M = \mathbb{R}^4$ and $\dot{x} = u_1a_1(x) + u_2a_2(x) + u_3a_3(x)$, $u_i \geqq 0$, $\sum u_i = 1$, where

$$a_1(x) = \begin{pmatrix} 1 \\ -1/2 \\ -1/2 \\ -x_2/2 \end{pmatrix}, \quad a_2(x) = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad a_3(x) = \begin{pmatrix} 1 \\ 0 \\ 1 \\ x_2 \end{pmatrix},$$

$$[a_1, a_2](x) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1/2 \end{pmatrix}, \quad [a_1, a_3](x) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1/2 \end{pmatrix}, \quad [a_2, a_3](x) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

and all other brackets are zero.

DH is of dimension 4 everywhere so the integral manifold of DH through $x^0 = (0, 0, 0, 0)$ is exactly M. The control $u^0 = (1/2, 1/4, 1/4)$ gives rise to the vector field

$$f_0(x) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

and $[f_0, a_i] = 0$. If $x^1 = \gamma_0(s_1)x^0 = (s_1, 0, 0, 0)$, then the cone generated by control variations of type (4) and (7) is a linear space of dimension 3, since the trajectory is singular in the usual sense. If we add the variations of type (6), we see $K_{x^1} = L_{x^1} = M_{x^1}$ and so is not singular in our sense. Notice that x^1 is bang-bang accessible,

$$x^1 = \gamma_3(s_1/8)\gamma_2(s_1/8)\gamma_1(s_1/2)\gamma_2(s_1/8)\gamma_3(s_1/8)x^0.$$

A *subsystem* of (3) is a system obtained by restricting the control $u(t)$ to lie on one of the faces of Ω, that is, if I is a subset of $\{1, \cdots, k\}$ the subsystem specified by I is given by requiring $u_i(t) = 0$ for $i \notin I$. We consider Ω a face of Ω, so that (3) is a subsystem of itself.

THEOREM 3. *Suppose for every subset I of $\{1, \cdots, k\}$, the subsystem specified by I is locally semi-integrable. Let U be a neighborhood of x^0 and H^+ and L^+ be the submodule and integral manifold which carry the subsystem specified by I on U^+. If there exists $h > 0$ such that*

(i) *$D^hH_x^+ = DH_x^+$ for all $x \in L^+$,*

(ii) *given any $i_j \in I, j = 1, \cdots, 4$, and any m, $1 \leqq m < h$, there exists a function $\mu(x) \geqq 0$ such that for all $x \in L^+$ either*

$$ad^m(a_{i_3})[a_{i_1}, a_{i_2}](x) \equiv \mu(x)ad^m(a_{i_4})[a_{i_1}, a_{i_2}](x) \bmod D^mH_x^+$$

or

$$ad^m(a_{i_3})[a_{i_1}, a_{i_2}](x) \equiv \mu(x)ad^m(a_{i_4})[a_{i_1}, a_{i_2}](x) \bmod D^mH_x^+,$$

then $\mathscr{A}(x^0, U, E) = \mathscr{A}(x^0, U, \Omega)$.

Proof. Let $I = \{1, \cdots, k\}$ and $u^1(t), \cdots, u^m(t)$, $v^1(t), \cdots, v^m(t)$ be controls lying in the interior of Ω, that is, $0 < u_i^j(t) < 1$ and $0 < v_i^j(t) < 1$ for $j = 1, \cdots, m$, $i = 1, \cdots, k$. Let $f_j(x) = \sum_{i=1}^k u_i^j(x_1)a_i(x)$ and $g_j(x) = \sum_{i=1}^k v_i^j(x_1)a_i(x)$. By induction on $m < h$, we show there exists a $\lambda(x) > 0$ such that

$$[f_1 \cdots [f_m[a_{i_1}, a_{i_2}]] \cdots](x) \equiv \lambda(x)[g_1 \cdots [g_m[a_{i_1}, a_{i_2}]] \cdots](x) \bmod D^mH_x^+.$$

It is trivial for $m = 0$ and it follows immediately from (ii) for $1 \leqq m < h$.

Therefore if $u^0(t)$ lies in the interior of Ω in some neighborhood of $x^1 = \gamma_0(s_1)x^0$, then $D^hH_{x^1}^+$ is spanned by the vectors

$$a_i(x^1), [a_i, a_j](x^1), ad^1(f_0)[a_i, a_j](x^1), \cdots, ad^{h-1}(f_0)[a_i, a_j](x^1),$$

for $1 \leqq i, j \leqq k$. But the cone K_{x^1} contains $\pm f_0(x^1)$, $a_i(x^1) - f_0(x^1)$, and

$$\sum_{m=0}^{h-1} \frac{(s - s_1)^m}{m!} ad^m(f_0)[a_i, a_j](x^1) + \mathcal{O}(s - s_1)^h$$

for all $1 \leqq i, j \leqq k$ and small $s - s_1 \leqq 0$. Hence, K_{x^1} equals $D^hH_{x^1}^+ = DH_{x^1}^+$. This implies $x^1 \in \mathscr{A}(x^0, U, E)$.

If $u^0(t)$ is not interior to Ω at x^1 but is interior to some face of Ω of dimension $\geqq 1$, then we repeat the above argument for the subsystem generated by that face. The controls that lie on faces of dimension 0 are bang-bang controls. Q.E.D.

There is a bang-bang controllability version of Theorem 3 that assumes the same hypothesis except H^- and L^- replace H^+ and L^+ in (i) and (ii). Together they yield a global result.

COROLLARY 4. *Suppose for all $x \in M$, there exists a neighborhood U of x such that $\mathscr{A}(x, U, E) = \mathscr{A}(x, U, \Omega)$ and $\mathscr{C}(x, U, E) = \mathscr{C}(x, U, \Omega)$. Then $\mathscr{A}(x^0, M, E) = \mathscr{A}(x^0, M, \Omega)$ and $\mathscr{C}(x^0, M, E) = \mathscr{C}(x^0, M, \Omega)$.*

Proof. Suppose $u^0(t)$ is an admissible control. We must show $\gamma_0(s)x^0 \in \mathscr{A}(x^0, M, E)$ for all $s \geqq 0$. Let $s = \inf\{r \geqq 0 : \gamma_0(s)x^0 \notin \mathscr{A}(x^0, M, E)\}$ and $x = \gamma_0(s)x^0$. If $x \in \mathscr{A}(x^0, M, E)$, then by hypothesis there exists an $\varepsilon > 0$ such that for all $r \in [0, \varepsilon)$, $\gamma_0(r)x = \gamma_0(r + s)x^0 \in \mathscr{A}(x, M, E) \subseteq \mathscr{A}(x^0, M, E)$. This contradicts the definition of s. If $x \notin \mathscr{A}(x^0, M, E)$, then there exists $\varepsilon > 0$ such that for all $r \in (-\varepsilon, 0]$, $\gamma_0(r)x = \gamma_0(r + s)x^0 \in \mathscr{C}(x, M, E)$. By the definition of s, for small r, $\gamma_0(r + s)x^0 \in \mathscr{A}(x^0, M, E)$ so $x \in \mathscr{A}(x^0, M, E)$. This is a contradiction.

COROLLARY 5 (Halkin–Levinson). *Consider the linear control system defined on $M = \mathbb{R}^n$ by*

$$\dot{x} = F(t)x + G(t)v + h(t), \tag{8}$$

where $F(t)$, $G(t)$ are matrices, $h(t)$ is a vector of piecewise analytic functions and the control $v(t) = (v_1(t), \cdots, v_n(t))$ is a piecewise analytic function satisfying $|v_i(t)| \leqq 1$. If x is accessible from x^0 by an admissible control, then x is accessible from x^0 by a piecewise analytic bang-bang control $v(t)$, where $|v_i(t)| = 1$.

Proof. Let $a_1(x), \cdots, a_k(x)$ be the right side of (8) for the finite number of constant controls satisfying $|v_i(t)| = 1$. Then (8) can be put in the form (3) and each of the $a_i(x)$ is piecewise analytic. It follows that every subsystem of (3) is locally semi-integrable. By direct computation it is easy to show

$$ad^m(a_{i_3})[a_{i_1}, a_{i_2}](x) = ad^m(a_{i_4})[a_{i_1}, a_{i_2}](x) \quad \text{for all } m = 1, 2, 3, \cdots,$$

and for any x^0 there always exists a neighborhood, U, of x^0 and $h > 0$ such that $D^hH_x^+ = DH_x^+$ for all $x \in U^+$ and $D^hH_{x^0}^-$ for all $x \in U^-$, so the result follows from Theorem 3 and Corollary 4. Q.E.D.

Notice that Examples 4 and 5 satisfy the conditions of Theorem 3. As a counterexample, consider this one taken from Filippov [5] as modified by Lobry [14].

Example 6. Let $M = \mathbb{R}^3$, $\dot{x} = ua_1 + (1 - u)a_2$ and $0 \leqq u \leqq 1$, where

$$a_1(x) = \begin{pmatrix} 1 \\ 1 - x_3^2 \\ 1 \end{pmatrix}, \qquad a_2(x) = \begin{pmatrix} 1 \\ 1 - x_3^2 \\ -1 \end{pmatrix},$$

$$[a_1, a_2](x) = \begin{pmatrix} 0 \\ -4x_3 \\ 0 \end{pmatrix}, \qquad [a_1[a_1, a_2]](x) = [a_2[a_2, a_1]](x) = \begin{pmatrix} 0 \\ -4 \\ 0 \end{pmatrix}.$$

Condition (ii) of Theorem 3 fails and the point $(1, 1, 0)$ is accessible from $(0, 0, 0)$ by the singular control $u = 1/2$, but is not bang-bang accessible (see Filippov [5]).

Acknowledgment. The author wishes to acknowledge many fruitful conversations with Professors S. P. Diliberto and H. J. Sussmann, and also the many helpful suggestions of the referee.

REFERENCES

[1] R. W. BROCKETT, *System theory on group manifolds and coset spaces*, this Journal, 10 (1972), pp. 265–285.

[2] W. L. CHOW, *Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann., 117 (1939), pp. 98–105.

[3] S. P. DILIBERTO, *The Pontryagin maximum principle*, Topics in Optimization, Leitmann, ed., Academic Press, New York, 1967.

[4] D. ELLIOT, *A consequence of controllability*, J. Differential Equations, 10 (1971), pp. 364–370.

[5] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 78–84.

[6] R. GABASOV AND F. M. KIRILLOVA, *High order necessary conditions for optimality*, this Journal, 10 (1972), pp. 127–168.

[7] H. HALKIN, *A generalization of LaSalle's bang-bang principle*, this Journal, 2 (1965), pp. 199–202.

[8] G. W. HAYNES AND H. HERMES, *Non-linear controllability via Lie theory*, this Journal, 8 (1970), pp. 450–460.

[9] R. HERMANN, *On the accessibility problem in Control Theory*, International Symposium, Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325–332.

[10] ———, *Differential Geometry and the Calculus of Variations*, Academic Press, New York, 1968.

[11] H. J. KELLEY, R. E. KOPP AND H. G. MOYER, *Singular extremals*, Topics in Optimization, Academic Press, New York, 1965, pp. 63–101.

[12] J. P. LASALLE, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillation, vol. 5, Princeton University Press, Princeton, 1960, pp. 1–24.

[13] N. LEVINSON, *Minimax, Liapunov and "bang-bang"*, J. Differential Equations, 2 (1966), pp. 218–241.

[14] C. LOBRY, *Contrôlabilité des systèmes non linéires*, this Journal, 8 (1970), pp. 573–605.

[15] ———, *Une properiété de l'ensemble des éstats accessibles d'un systèm guidable*, C. R. Acad. Sci. Paris t., 272 (1971).

[16] T. NAGANO, *Linear differential systems with singularities and an application to transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 398–404.

[17] L. S. PONTRYAGIN, U. G. BOLTYANSKII, R. V. GRAMKELIDE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[18] H. J. SUSSMANN, *The bang-bang problem for linear control systems in Lie groups*, this Journal, 10 (1972), pp. 470–476.

[19] H. J. SUSSMANN AND V. J. JURDJEVIC, *Controllability of non-linear systems*, J. Differential Equations, 12 (1962), pp. 95–116.

[20] H. HERMES, *Controllability and the singular problems*, this Journal, 2 (1965), pp. 241–260.

SIAM J. CONTROL
Vol. 12, No. 1, February 1974

SOME RESULTS ON MAX-MIN PURSUIT*

JAMES FLYNN†

Abstract. A pursuer P and an evader E are confined to a subset $\mathcal{S}$ of the Euclidean plane. E whose speed is bounded by $w \geqq 1$ wants to maintain the greatest possible distance between himself and P whose speed is bounded by 1. We will show that if $\mathcal{S}$ is a half-plane or a circle, then E can prevent the distance $|PE|$ from falling below its initial value only if he has a strategy under which $|PE|$ stays constant whenever P moves at maximum speed along a polygonal path. We use this result to characterize d^*, the least upper bound on the values of $|PE|$ that E can maintain for the case of pursuit in the circle.

Introduction. A pursuer P wants to get close to an evader E, who wants to stay away. The players are confined to a subset $\mathcal{S}$ of two-dimensional Euclidean space, P moving with speed bounded by 1 and E moving with speed bounded by $w \geqq 1$. P is required to select his trajectory before play begins, while E is allowed to use a *strategy*; that is, E can continuously observe P and select his trajectory as a function of his observations. However, E is required to tell P his strategy, and P can use this information in selecting his trajectory. We are faced with the following question. Given a specific starting position, what is the largest value of $|PE|$ that can be maintained by E? We attempt to answer that question by considering the following one. Under what conditions can E prevent the distance $|PE|$ from decreasing?

We obtain the following result when $\mathcal{S}$ is a half-plane.

(A) If E can keep $|PE|$ from falling below its initial value whenever P moves at maximum speed along a polygonal path, then E can do so by using an *isometric strategy*; that is, a strategy under which E keeps the distance $|PE|$ constant until an escape position is reached.

(The fact that condition (A) deals only with the case where P is restricted to trajectories in which he travels at maximum speed along a polygonal path involves no essential loss of generality (see Remark 2).)

In a sense, the arguments which we use to establish (A) for the half-plane are unnecessarily long. One can, in fact, use the results of Isaacs [5, § 9.5.2] to develop shorter ones. Unfortunately, the shorter arguments do not work for the circle. Since our main interest is in the circle, we prefer longer arguments which are more general.

To describe our results for the circle, we need the notion of *stable position*. Let O denote the center of the circle. Any position where P lies on the radius through E and satisfies $|OP| = (1/w)|OE|$ is said to be *stable*. The main reason for the importance of such positions is that P can always force the play into a position which is approximately stable (see Remark 7).

We show that (A) holds for stable starting positions when $\mathcal{S}$ is a circle. We use this result to characterize d^*, the least upper bound on the values of $|PE|$ that E

* Received by the editors July 21, 1972, and in revised form December 12, 1972.

† University of Chicago Graduate School of Business, Chicago, Illinois 60637. This research was supported in part by the Office of Naval Research under Contract N00014-67-A-0112-0053 (NR-042-267) at the Statistics Department, Stanford University.

can maintain. One can use our characterization to obtain numerical values for d^*. We use it to obtain bounds (see Theorem 4).

We wish to thank L. Dubins and D. Blackwell for introducing us to these problems. A version of the problem of *pursuit in the half-plane* appears in Isaacs [5, pp. 260–265]. Apparently Isaacs was the first to suggest that the problem be generalized to a circle [5, pp. 265, 270]. Our results, however, owe more to the work of Gerald J. Smith [8] which dealt with the problem of *pursuit in the circle.* Smith attempted to establish result (A) for the circle. He also used (A) to characterize d^*. Some of our arguments are borrowed from him. Other results on *pursuit in the half-plane* and *pursuit in the circle* appear in Flynn [2], [3].

1. Formulation. Let $\mathbf{R}$ represent the real numbers and let $\mathbf{R}^2$ represent the space $\mathbf{R} \times \mathbf{R}$ with the metric determined by the norm $\|\cdot\|$, where $\|(r_1, r_2)\| = (r_1^2 + r_2^2)^{1/2}$. Denote the time axis $[0, \infty)$ by $\mathbf{T}$. Define

$$\begin{aligned} \mathbf{P}(p) &= \{\mathbf{p} | \mathbf{p}: \mathbf{T} \to \mathscr{S}, \mathbf{p}(0) = p \quad \text{and} \quad \|\mathbf{p}(t') - \mathbf{p}(t'')\| \leqq \|t' - t''\| \\ &\qquad \text{for all } t't'' \, E\, \mathbf{T}\} \quad \text{for } p \in \mathscr{S}, \\ \mathbf{E}(e) &= \{\mathbf{e} | \mathbf{e}: \mathbf{T} \to \mathscr{S}, \mathbf{e}(0) = e \quad \text{and} \quad \|\mathbf{e}(t') - \mathbf{e}(t'')\| \leqq w\|t' - t''\| \\ &\qquad \text{for all } t', t'' \in \mathbf{T}\} \quad \text{for } e \in \mathscr{S}, \\ &\qquad \mathbf{P} = \bigcup_{p \in \mathscr{S}} \mathbf{P}(p), \quad \mathbf{E} = \bigcup_{e \in \mathscr{S}} \mathbf{E}(e) \end{aligned} \tag{1}$$

and

$$\mathscr{P}(\mathbf{e}, \mathbf{p}) = \inf_{t \in \mathbf{T}} \|\mathbf{e}(t) - \mathbf{p}(t)\| \quad \text{for } (\mathbf{e}, \mathbf{p}) \in \mathbf{E} \times \mathbf{P}.$$

$\mathbf{P}(p)$ and $\mathbf{E}(e)$ are the sets of pursuit and evasion trajectories, respectively, originating from the positions p and e, while $\mathscr{P}(\mathbf{e}, \mathbf{p})$ is the payoff from P to E when P follows $\mathbf{p}$ and E follows $\mathbf{e}$.

We define strategy as in Ryll-Nardzewski [7, pp. 113–126]. Let $(e, p) \in \mathscr{S} \times \mathscr{S}$. A mapping $\eta: \mathbf{P}(p) \to \mathbf{E}(e)$ is an evasion strategy at (e, p) if it satisfies the *information constraint*: If $\mathbf{p}', \mathbf{p}'' \in \mathbf{P}(p)$ satisfy $\mathbf{p}'(t) = \mathbf{p}''(t)$ for $0 \leqq t \leqq t'$, then $\eta(\mathbf{p}')(t) = \eta(\mathbf{p}'')(t)$ for $0 \leqq t \leqq t'$. Denote by $H(e, p)$ the set of all evasion strategies at (e, p). Define

$$V(e, p) = \sup_{\eta \in H(e,p)} \inf_{\mathbf{p} \in \mathbf{P}(p)} \mathscr{P}(\eta(\mathbf{p}), \mathbf{p}). \tag{2}$$

We call $V(e, p)$ the max–min value at (e, p).

Remark 1. As stated in the Introduction (see paragraph 1), we want to determine when the following situations arise:

(I) $V(e, p) = \|e - p\|$,

(II) $V(e, p) < \|e - p\|$.

(Observe that $V(e, p) \leqq \|e - p\|$ always holds.)

In this paper we shall have occasion to refer to the derivatives of various quantities with respect to time. We will always have in mind the forward or right-hand time derivative. Whenever such a derivative exists we will represent it by the usual "dot" notation. An object's speed will, of course, always refer to the norm of its right-hand velocity vector. Define

(3a) $\mathbf{P}^*(p) = \{\mathbf{p} \in \mathbf{P}(p) | \mathbf{p} \text{ is piecewise linear and } \|\dot{\mathbf{p}}(t)\| = 1, t \in \mathbf{T}\}$ for $p \in \mathscr{S}$

and

(3b) $$\mathbf{P}^* = \bigcup_{p \in \mathscr{S}} \mathbf{P}(p).$$

$\mathbf{P}^*$ is the set of pursuit trajectories under which P travels at maximum speed along a piecewise linear path. Remark 2 tells us that $\mathbf{P}^*$ is almost as good as $\mathbf{P}$.

Let $p \in \mathscr{S}$. One can show that for every positive integer n, there exists an *approximation mapping* α^n from $\mathbf{P}(p)$ to $\mathbf{P}^*(p)$ satisfying

(4) $$\sup_{\mathbf{p} \in \mathbf{P}(p)} \sup_{t \in \mathbf{T}} \|\alpha^n(\mathbf{p})(t) - \mathbf{p}(t)\| \leqq 1/n,$$

such that if $\mathbf{p}', \mathbf{p}'' \in \mathbf{P}(p)$ and $\mathbf{p}'(t) = \mathbf{p}''(t)$ for $0 \leqq t \leqq t'$, then $\alpha^n(\mathbf{p}')(t) = \alpha^n(\mathbf{p}'')(t)$ for $0 \leqq t \leqq t'$.

Remark 2. If the strategy $\eta^* \in H(e, p)$ satisfies

(5) $$\mathscr{P}(\eta^*(\mathbf{p}), \mathbf{p}) = \|e - p\|, \qquad \mathbf{p} \in \mathbf{P}^*(p),$$

then the strategy defined by the composition of η^* and the approximation mapping α^n satisfies

(6) $$\mathscr{P}(\eta^*(\alpha^n(\mathbf{p})), \mathbf{p}) \geqq \|e - p\| - 1/n, \qquad p \in \mathbf{P}(p),$$

for $n = 1, 2, \cdots$. It follows that the existence of such an η^* is sufficient for (I) (see Remark 1).

The above remark justifies the following assumption.

Assumption 1. P is restricted to $\mathbf{p} \in \mathbf{P}^*$.

2. The half-plane. In this section we consider the case where $\mathscr{S}$ is a closed half-plane $\mathscr{H}$ in $\mathbf{R}^2$ with the line $\mathscr{L}_1$ in $\mathbf{R}^2$ as its boundary. After determining conditions under which (I) and (II) hold (see Remark 1), we will establish result (A).

Let $\mathscr{L}_2$ be a line perpendicular to $\mathscr{L}_1$ at some arbitrary point O. Set up a rectangular coordinate system where O is the origin, the coordinate axes are $\mathscr{L}_1$ and $\mathscr{L}_2$ and $\mathscr{H}$ coincides with the first and second quadrants. Positions of P and E are represented by points $p = (p_1, p_2)$ and $e = (e_1, e_2)$, respectively (see Fig. 1).

Let s, $-\pi < s \leqq \pi$, denote the angle that PE makes with $\mathscr{L}_2$ (we give s the same sign as $e_1 - p_1$). Whenever P's and E's right-hand velocity vectors exist, let ϕ and ψ denote the respective angles which these vectors make away from the direction PE, $-\pi < \phi, \psi \leqq \pi$. These angles are positive when they are measured in a counterclockwise arc, e.g., all the angles are positive in Fig. 1. Evidently, whenever P and E travel at maximum speeds, we have

(7) $$\begin{aligned} \dot{\mathbf{p}}_1 &= \sin(s + \phi), & \dot{\mathbf{p}}_2 &= -\cos(s + \phi), \\ \dot{\mathbf{e}}_1 &= w \sin(s + \psi), & \dot{\mathbf{e}}_2 &= -w \cos(s + \psi). \end{aligned}$$

Assumption 2. We will always assume that play starts in a position $(e, p) \in \mathscr{H} \times \mathscr{H}$, where $\|p - e\| > 0$, and $s \geqq 0$. There is, of course, no loss of generality in making such an assumption. We will also assume that $w > 1$. (The case $w = 1$ is left as an exercise for the reader.)

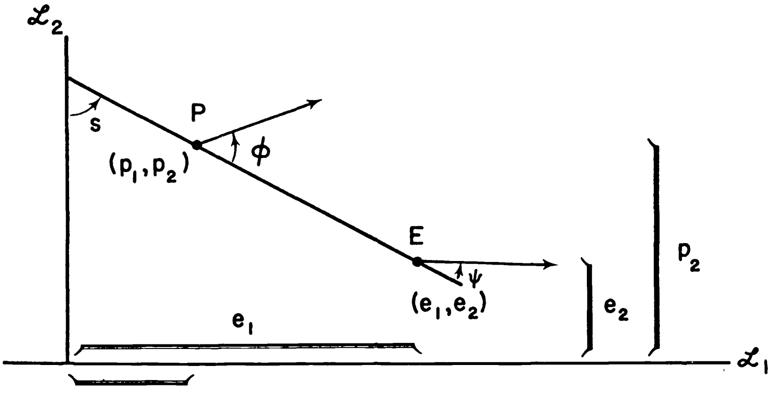

FIG. 1

Define

$$\beta = \text{arc sin}\,(1/w). \tag{8}$$

We say that *P corners E* whenever the game is in a position where $e_2 = 0$ and $|s| < \beta$. The reader should verify that whenever P corners E, P can force a decrease in $|PE|$ by heading directly towards E (see Isaacs [5, pp. 260–264]). Hence (II) holds at any position where P corners E.

We are going to show that (I) is equivalent to the existence of an *isometric strategy* for E, that is, a strategy under which $|PE|$ stays constant until an "escape" position is reached whenever P uses a trajectory in $\mathbf{P}^*$. Formally, an isometric strategy is any evasion strategy under which E follows the *isometric rule* (described below). We define the isometric rule by describing its consequences. For the present we do not worry about whether it is a valid rule, e.g., whether it satisfies the information constraint (see paragraph 2 of § 1).

The isometric rule is that method of play for E under which the following holds when P selects $\mathbf{p} \in \mathbf{P}^*$: E travels at maximum speed along a path with a piecewise continuous derivative, such that if

$$|s(t)| \geqq \beta,$$

we have

$$\dot{\mathbf{e}}_1(t) = w\,\text{sgn}\,(s(t)), \quad \dot{\mathbf{e}}_2(t) = 0,$$

while if

$$\mathbf{e}_2(t) > 0, \quad |s(t)| < \beta, \tag{9}$$

we have

$$\cos\psi(t) = (1/w)\cos\phi(t),$$
$$\psi(t)\begin{cases} > 0 & \text{if } s(t) \geqq 0,\\ < 0 & \text{if } s(t) < 0.\end{cases} \tag{10}$$

The next lemma follows from the arguments of Isaacs [5, pp. 261–264].

LEMMA 1. *If E follows the isometric rule, then as long as* (9) *holds we have*

$$\|\mathbf{p}(t) - \mathbf{e}(t)\| = \|p - e\|,$$
$$\dot{\mathbf{e}}_2(t) = (w^2 - \cos^2\phi(t))^{1/2}\sin s(t) - \cos\phi(t)\cos s(t), \tag{11}$$
$$\dot{s}(t) = [(w^2 - \cos^2\phi(t))^{1/2} - \sin\phi(t)]/\|p - e\| > 0.$$

Also if for some t_0,

$$s(t_0) \geqq \beta,$$

then

$$\|\mathbf{p}(t) - \mathbf{e}(t)\| \geqq \|p - e\|, \qquad t \geqq t_0.$$

(*Observe that* $\|\mathbf{p}(t) - \mathbf{e}(t)\|$ *approaches* ∞ *in this case.*)

Remark 3. At first it appears that the *isometric rule* is not allowable. Certainly we cannot define a strategy directly by rule (10) without violating the information constraint. Fortunately, we can get around this problem by exploiting the fact that (10) is required to hold only when $\mathbf{p} \in \mathbf{P}^*$ (see Assumption 1). The proof of the next lemma is left to the reader.

LEMMA 2. *E can follow the isometric rule.*

Given that E follows the isometric rule, P's objective is to select a $\mathbf{p} \in \mathbf{P}^*$ which corners E. We have the following lemma.

LEMMA 3. *If E follows the isometric rule, then P can corner E only if he can do so with a trajectory* $\mathbf{p}$ *which satisfies*

$$(12) \qquad 0 \leqq \phi(t) \leqq \arccos(w \sin s(t))$$

until cornering takes place.

Proof. By Lemma 1, $\dot{s}(t) > 0$. Hence we want a $\mathbf{p}$ for which $\dot{\mathbf{e}}_2(t) \leqq 0$. This gives us the condition

$$(13) \qquad |\phi(t)| \leqq \arccos(w \sin s(t)) \leqq \pi/2.$$

Consider any $\phi(t) < 0$ which satisfies (13). Lemma 1 implies that $-\phi(t)$ is better since it gives the same value of $\dot{\mathbf{e}}_2(t)$ and a smaller value of $\dot{s}(t)$. The result follows.

Now we need a notion of convexity. We say that a *pursuit trajectory* $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2)$ is *convex* on an interval $\mathbf{T}_0 \subset \mathbf{T}$ if (a) $\mathbf{p}_1(\cdot)$ is a monotone function on $\mathbf{T}_0$, and (b) the set $\bigcup_{t \in \mathbf{T}_0}\{(\mathbf{p}_1(t), p_2) | p_2 \geqq \mathbf{p}_2(t)\}$ is convex. A similar definition holds for evasion trajectories. (We will omit any specific reference to $\mathbf{T}_0$ whenever its identity is clear from the context.)

Lemma 3 implies that P can limit himself to trajectories which satisfy (a) above. That P can restrict himself to convex pursuit trajectories is more difficult to establish. The proof of the next lemma takes up § 3.

LEMMA 4. *If E follows the isometric rule, then P can corner E only if he can do so with a convex* $\mathbf{p}$ *which satisfies* (12) *until cornering takes place.*

We need one more lemma.

LEMMA 5. *If P uses a convex* $\mathbf{p}$ *for which* $\phi(t) \geqq 0$, $t \in \mathbf{T}$, *while E follows the isometric rule, then E's trajectory* $\mathbf{e}$ *is convex on the interval* $[0, t_0]$ *where* t_0 *denotes the first time that condition* (9) *fails.*

Proof. We assume that P's trajectory passes through the polygonal line segment joining P_0 to P_1 and P_1 to P_2 as illustrated in Fig. 2. Let $e^0 = (e_1^0, e_2^0)$, $e^1 = (e_1^1, e_2^1)$ and $e^2 = (e_1^2, e_2^2)$ denote E's positions when P reaches P_0, P_1 and P_2, respectively. Let r_1 denote the angle that the line P_0P_1 makes with the $\mathscr{L}_1$-axis and let r_2 denote the angle that the line P_1P_2 makes with the $\mathscr{L}_1$-axis. As usual, we take these angles as positive when we measure them in a counterclockwise direction.

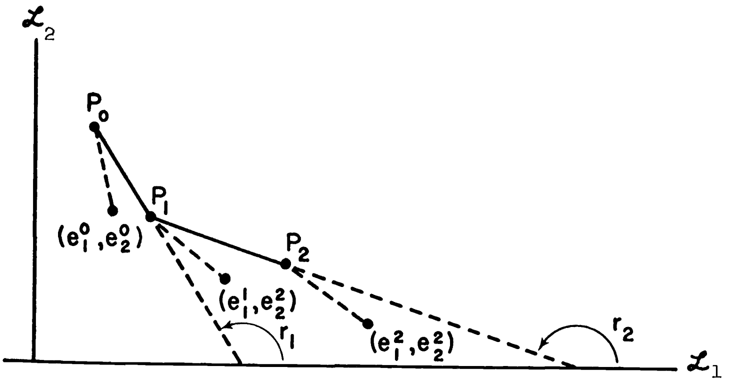

FIG. 2

Now (10), (12), Lemma 1 and the assumption that $s(0) \geqq 0$ imply that $\dot{\mathbf{e}}_1(t)$ is strictly positive. Hence, we can define the real-valued function $\Delta(\cdot)$ on the set $(e_1^0, e_1^1) \cup (e_1^1, e_1^2)$ by

$$\Delta(\mathbf{e}_1(t)) = \dot{\mathbf{e}}_2(t)/\dot{\mathbf{e}}_1(t). \tag{14}$$

We will establish the convexity of $\mathbf{e}$ by showing that $\Delta(\cdot)$ is an increasing function on $(e_1^0, e_1^1) \cup (e_1^1, e_1^2)$.

Using a simple geometric argument one can show that

$$s(t) + \phi(t) = \begin{cases} r_1 - \pi/2 & \text{for } \mathbf{e}_1(t) \in (e_1^0, e_1^1), \\ r_2 - \pi/2 & \text{for } \mathbf{e}_1(t) \in (e_1^1, e_1^2). \end{cases} \tag{15}$$

Consequently, (7) and (15) give us

$$\Delta(\mathbf{e}_1(t)) = \begin{cases} \tan\,[\psi(t) - \phi(t) + r_1] & \text{for } \mathbf{e}_1(t) \in (e_1^0, e_1^1), \\ \tan\,[\psi(t) - \phi(t) + r_2] & \text{for } \mathbf{e}_1(t) \in (e_1^1, e_1^2). \end{cases} \tag{16}$$

Let $\mathbf{e}_1(t) \in (e_1^0, e_1^1)$. Since Lemma 1 implies that $s(t)$ is increasing, (15) implies that $\phi(t)$ is decreasing. Now using (10), one can verify that $\psi - \phi$ is a decreasing function of ϕ. It follows that Δ increases on (e_1^0, e_1^1). A similar argument establishes the same result on (e_1^1, e_1^2). All we have to do is show that

$$\Delta(e_1^1 - 0) \leqq \Delta(e_1^1 + 0). \tag{17}$$

Let t_1 denote the time that P reaches P_1. By (16) we have

$$\begin{aligned} \Delta(e_1^1 - 0) &= \tan\,[\psi(t_1 - 0) - \phi(t_1 - 0) + r_1], \\ \Delta(e_1^1 + 0) &= \tan\,[\psi(t_1 + 0) - \phi(t_1 + 0) + r_2]. \end{aligned} \tag{18}$$

Now (15) implies

$$\phi(t_1 - 0) - r_1 = \phi(t_1 + 0) - r_2. \tag{19}$$

But (12) and the fact that $\mathbf{p} \in \mathbf{P}^*$ imply

$$\pi/2 < r_1 < r_2 < \pi.$$

Thus (19) implies

$$\phi(t_1 + 0) > \phi(t_1 - 0).$$

Hence,

$$\psi(t_1 - 0) < \psi(t_1 + 0) \tag{20}$$

holds. Evidently (17) follows directly from (18), (19), and (20).

Now we have our first main result (see Remark 1).

THEOREM 1. *Let* $(e, p) \in \mathscr{H} \times \mathscr{H}$ *be an initial position with* $\|p - e\| > 0$ *and let* $V(e, p)$ *be defined by* (5). *Then*

$$V(e, p) < \|e - p\| \tag{II}$$

holds if and only if P can corner E with a $\mathbf{p} \in \mathbf{P}^*$ *whenever E uses the isometric rule. Furthermore, whenever* (II) *fails to hold, we have*

$$V(e, p) = \|e - p\|. \tag{I}$$

The following corollary follows immediately from Theorem 1.

COROLLARY 1. (A) *holds for the half-plane.*

Proof of Theorem 1. The second claim and the "only if" part of the first claim follow directly from Remark 2. The only thing left to show is the "if" part of the first claim. Clearly we can assume without loss of generality that $s(0) \geqq 0$. Also we can show that there is no loss of generality in restricting E to strategies for which

$$s(t) \geqq 0, \qquad t \in \mathbf{T}, \tag{21}$$

whenever P uses a trajectory in $\mathbf{P}^*$.

Suppose that P can corner E with a $\mathbf{p} \in \mathbf{P}^*$ whenever E follows the isometric rule. It follows from Lemma 4 that P can also corner E with a convex $\hat{\mathbf{p}} \in \mathbf{P}^*$ for which (12) holds. Lemma 5 implies that whenever P uses $\hat{\mathbf{p}}$, the trajectory $\hat{\mathbf{e}}$, which E follows under the isometric rule, is also convex. The reader should verify that to establish the theorem we need only show that the convexity of $\hat{\mathbf{e}}$ implies that there does not exist an evasion trajectory $\mathbf{e}$ satisfying

$$\|\hat{\mathbf{p}}(t) - \mathbf{e}(t)\| \geqq \|p - e\|, \qquad t \in \mathbf{T}, \tag{22}$$

for which (21) holds.

Observe that (21) is equivalent to

$$\mathbf{e}_1(t) - \hat{\mathbf{p}}_1(t) \geqq 0, \qquad t \in \mathbf{T}. \tag{23}$$

Assume that there exists a trajectory $\mathbf{e}$ satisfying (22) and (23). We will show that this leads to a contradiction. Let t_c denote the time at which cornering takes place when P uses $\hat{\mathbf{p}}$ and E uses $\hat{\mathbf{e}}$ (see Fig. 3). Using (22) and (23) one can show that there exists a $t_0 \in (0, t_c)$ such that the trajectory $\mathbf{e}$ reaches the point $\hat{\mathbf{e}}(t_0)$ strictly before time t_0. Using (22) one can also show that during the time interval $[0, t_0]$, $\mathbf{e}(t)$ must lie below the curve generated by the trajectory $\mathbf{e}$. Now the convexity of $\hat{\mathbf{e}}$ implies that the arc spanned by $\hat{\mathbf{e}}$ between time 0 and time t_0 is the shortest arc joining $\hat{\mathbf{e}}(0)$ and $\hat{\mathbf{e}}(t_0)$ which lies below the curve generated by $\hat{\mathbf{e}}$. By (3) and the definition of isometric rule, E travels at maximum speed when he uses $\hat{\mathbf{e}}$. Hence $\mathbf{e}$ cannot reach the point $\hat{\mathbf{e}}(t_0)$ strictly before time t_0. This contradiction finishes our proof.

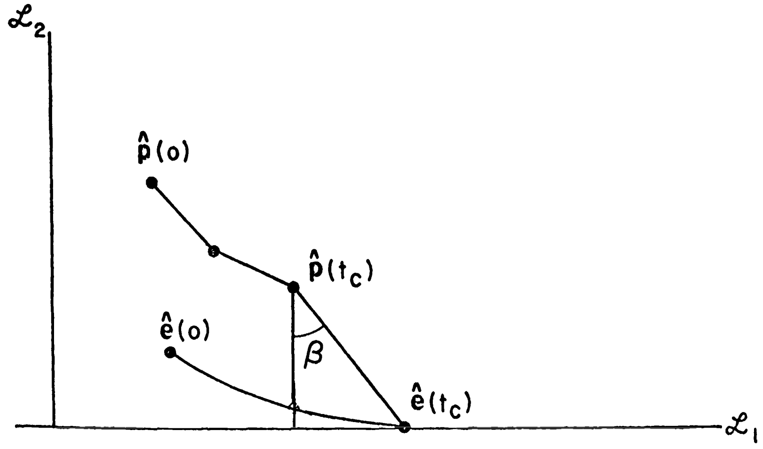

FIG. 3

Remark 4. The above proof borrows heavily from Gerald Smith [8]. In particular the idea that we can assume that E restricts himself to strategies for which (21) holds is his. Smith also uses a similar convexity argument.

3. Proof of Lemma 4. Suppose that P has a $\mathbf{p} \in \mathbf{P}^*(p)$ which corners E. By Lemma 3 we can assume that (12) holds when P uses $\mathbf{p}$. Now (12) implies that $\mathbf{p}$ satisfies part (a) of the definition of convexity. Suppose that $\mathbf{p}$ violates part (b) of that definition. We want to show that P can do better by replacing $\mathbf{p}$ by a convex trajectory. Specifically, we will show that any nonconvex polygonal path segment $P_0P_1P_2$ which satisfies (12) is dominated by a convex polygonal path segment P_0P_2 which also satisfies (12) (see Fig. 4). This result gives us a step-by-step procedure for replacing a nonconvex $\mathbf{p} \in \mathbf{P}^*$ with a better convex trajectory.

Examine Fig. 5. We have a rectangular coordinate system in which P_0 coincides with the origin, P_2 lies on the positive Y-axis, P_1 lies in the second quadrant and E_0 lies in the first. Let d denote the distance $|P_0P_2|$, xd the distance $|P_0P_1|$ and ρ the distance $|P_0E_0|$. Let σ_0 denote the angle that P_0E_0 makes with the X-axis, let u denote the angle that P_0P_1 makes with the Y-axis and let $\gamma(u, x)$ satisfy

$$\gamma(u, x) = \arcsin\left(x \sin u/(1 - 2x \cos u + x^2)^{1/2}\right). \tag{24}$$

Since $s(0) \geqq 0$, we have $\sigma_0 > 0$. Also, condition (12) implies that $s < \pi/2$, $u \geqq 0$, $x \geqq 0$ and $d \geqq 0$.

Suppose that P moves at full speed along P_0P_2. Let $\sigma(t)$ denote the angle that the line through PE makes with the Y-axis at time t. Using Lemma 1, one can show that $\sigma(t)$ satisfies the differential equation

$$\frac{d\sigma}{dt} = \frac{\sqrt{w^2 - \sin^2 \sigma} - \cos \sigma}{\rho}, \tag{25}$$

and using (25), one can show that the value of σ when P reaches P_2 is equal to $I^{-1}(d + I(\sigma_0))$, where I is the elliptic integral

$$I(\sigma) = \frac{\rho}{w^2 - 1} \int_0^{\sigma} \left(\sqrt{w^2 - \sin^2 t} + \cos t\right) dt \tag{26}$$

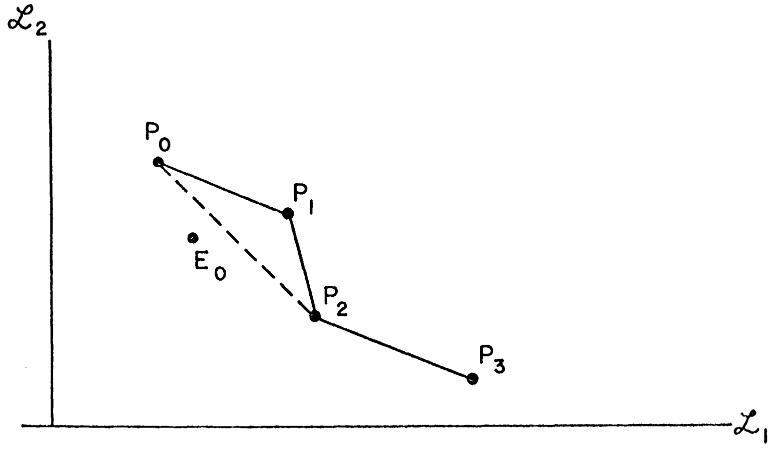

FIG. 4

and I^{-1} denotes the inverse of I.

Another possible trajectory arises when P moves at full speed along the polygonal line $P_0P_1P_2$. Let $\Omega(u, x)$ denote the value of the angle σ when P reaches P_2 under this trajectory. Clearly,

$$\Omega(0, 0) = I^{-1}(d + I(\sigma_0)).$$

More generally, one can show that

$$\Omega(u, x) = C(u, x) - \gamma(u, x), \tag{27}$$

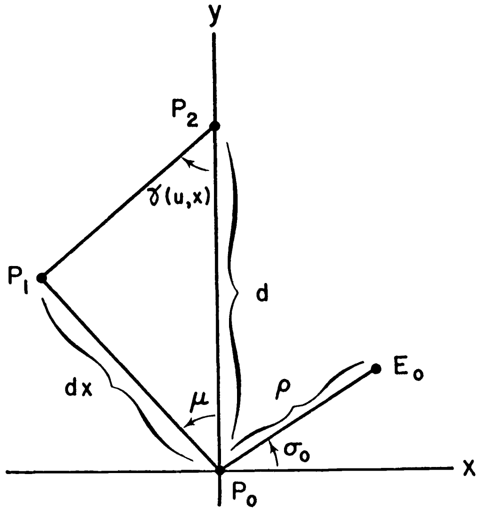

FIG. 5

where

$$\begin{aligned} C(u,x) &= I^{-1}[d(1-2x\cos u + x^2)^{1/2} + I(B(u,x))], \\ B(u,x) &= \gamma(u,x) + u + A(u,x), \\ A(u,x) &= I^{-1}(xd + I(\sigma_0 - u)), \end{aligned} \tag{28}$$

and $\gamma(u,x)$ satisfies (24). The reader should verify that (12) implies

$$0 \leqq \sigma - u \leqq A(u,x) \leqq B(u,x) \leqq C(u,x) \leqq \pi/2. \tag{29}$$

In order to show that the nonconvex polygonal path $P_0P_1P_2$ is dominated by the convex P_0P_2, we must demonstrate that

$$\Omega(u,x) \geqq \Omega(0,0) \tag{30}$$

holds whenever (29) does. To that end, we will establish the following result. If condition

$$0 \leqq d/\rho \leqq (w-1)/(w+1) \tag{31}$$

holds, then (29) implies (30). By repeated application of that result and a uniform convergence argument, one can show that (29) by itself is enough to guarantee (30). Note that condition (31) is only a device introduced to simplify the proof.

Assume that (31) holds. Consider any fixed $x \geqq 0$. We will show that

$$\frac{\partial\Omega(u,x)}{\partial x} \geqq 0, \tag{32}$$

whenever u satisfies (29). (We assume that there is a u satisfying (29), since otherwise the lemma is trivial.) Straightforward calculations give

$$\frac{\partial\Omega(u,x)}{\partial x} = \frac{N(u,x)}{D(u,x)}, \tag{33}$$

where

$$\begin{aligned} D(u,x) &= I'(C(u,x)) > 0, \\ N(u,x) &= N_1(u,x) + N_2(u,x) + N_3(u,x), \\ N_1(u,x) &= \frac{\sin u}{(1-2x\cos u + x^2)}(I'(B(u,x)) - I'(C(u,x))), \\ N_2(u,x) &= \frac{d(x-\cos u)}{(1-2x\cos u + x^2)^{1/2}}, \end{aligned} \tag{34}$$

and

$$N_3(u,x) = dI'(B(u,x))/I'(C(u,x)).$$

(As usual the "prime" notation refers to the first derivative.) Define the function R by

$$R(z) = I'(I^{-1}(z)). \tag{35}$$

The reader should verify that

$$R'(z) = -\sin I^{-1}(z)/(w^2 - \sin^2 I^{-1}(z))^{1/2}. \tag{36}$$

By (28), (29), (34), (35), and the mean value theorem, we have

$$N_1(u,x) \geqq \hat{N}_1(u,x) > 0, \tag{37}$$

where

$$\hat{N}_1(u,x) = \left(\frac{d\sin u}{(1-2x\cos u + x^2)^{1/2}}\right)\cdot\left(\frac{\sin B(u,x)}{(w^2-\sin^2 B(u,x))^{1/2}}\right). \tag{38}$$

Also, using the definition of I, one can show that

$$N_3(u,x) \geqq d(w-1)/(w+1) > 0. \tag{39}$$

Define

$$\hat{N}(u,x) = \hat{N}_1(u,x) + N_2(u,x) + N_3(u,x). \tag{40}$$

Because of (34) and (37),

$$\hat{N}(u,x) \geqq 0 \tag{41}$$

implies (32). Now, by (36), (37) and (39), we have (32) if u satisfies

$$N_2(u,x) \geqq -d(w-1)/(w+1). \tag{42}$$

Hence, we need only show that (41) holds whenever u satisfies (29) and

$$\frac{\cos u - x}{(1-2x\cos u + x^2)^{1/2}} > d(w-1)/(w+1). \tag{43}$$

Observe that the set of all points satisfying (43) is of the form $\{u|0 \leqq u < k(x) < \pi/2\}$, where $k(x)$ is a constant depending on x.

We want to show that the set of points satisfying both (29) and (43) is an interval containing 0. We will do this by showing that

$$\frac{\partial B(u,x)}{\partial u} > 0 \tag{44}$$

whenever u satisfies (43). It follows a fortiori that $\partial C(u,x)/\partial u$ is also positive. Straightforward calculations give us

$$\begin{aligned}\frac{\partial B(u,x)}{\partial u} &= \frac{x(\cos u - x)}{1-2x\cos u + x^2} + 1 - \frac{I'(\sigma_0 - u)}{I'(A(u,x))}\\ &> x(w-1)/(w+1) + \frac{R(dx + I(\sigma_0 - u)) - R(I(\sigma_0 - u))}{I'(A(u,x))}.\end{aligned} \tag{45}$$

But by (31), (36), the mean value theorem, and the definition of I, we have

$$\frac{R(dx + I(\sigma_0 - u)) - R(I(\sigma_0 - u))}{I'(A(u,x))} \geqq \frac{-dx/(w^2-1)^{1/2}}{\rho/(w^2-1)^{1/2}} \geqq -x(w-1)/(w+1). \tag{46}$$

Define

$$\begin{aligned}F(x) &= \{u|N_2(u,x) + N_3(u,x) \geqq 0 \text{ and (29) holds}\},\\ G(x) &= \{u|N_2(u,x) + N_3(u,x) < 0 \text{ and (29) holds}\}.\end{aligned} \tag{47}$$

By (39), $u \in G(x)$ implies (43). And, more important, (37) tells us that $u \in F(x)$ implies (41). We want to show that $u \in G(x)$ also implies (41). To accomplish this, we will establish

$$(48) \qquad \frac{\partial \hat{N}(u,x)}{\partial u} > 0, \qquad u \in G(x).$$

That (41) holds on $F(x) \cup G(x)$ will follow from (48),

$$(49) \qquad \hat{N}(0,x) = 0,$$

and a continuity argument. We leave the details to the reader.

Let $u \in G(x)$. Straightforward calculations give

$$(50) \qquad \begin{aligned} \frac{\partial \hat{N}_1(u,x)}{\partial u} &= N_{1a}(u,x) + N_{1b}(u,x), \\ \frac{\partial N_2(u,x)}{\partial u} &= \frac{d \sin u(1 - x\cos u)}{(1 - 2x\cos u + x^2)^{3/2}}, \\ \frac{\partial N_3}{\partial u}(u,x) &= N_{3a}(u,x) + N_{3b}(u,x), \end{aligned}$$

where

$$(51) \qquad \begin{aligned} N_{1a}(u,x) &= -N_2(u,x)\left(\frac{1 - x\cos u}{1 - 2x\cos u + x^2}\right)\left(\frac{\sin B(u,x)}{(w^2 - \sin^2 B(u,x))^{1/2}}\right), \\ N_{1b}(u,x) &= \left(\frac{dw^2 \cos B(u,x)}{(w^2 - \sin^2 B(u,x))^{1/2}}\right) \cdot \left(\frac{\sin u}{(1 - 2x\cos u + x^2)^{1/2}}\right)\frac{\partial B(u,x)}{\partial u}, \\ N_{3a}(u,x) &= -N_3(u,x)\left(\frac{1 - x\cos u}{1 - 2x\cos u + x^2}\right)\left(\frac{\sin B(u,x)}{(w^2 - \sin^2 B(u,x))^{1/2}}\right), \\ N_{3b}(u,x) &= N_3(u,x)\frac{R(I(\sigma_0 - u))}{R(dx + I(\sigma_0 - u))}\left(\frac{\sin B(u,x)}{(w^2 - \sin^2 B(u,x))^{1/2}}\right. \\ &\qquad \left. - \frac{\sin A(u,x)}{(w^2 - \sin^2 A(u,x))^{1/2}}\right). \end{aligned}$$

Now (43) implies

$$(52) \qquad 1 - x\cos u > 0,$$

while (29), (39), (44), (50), (51) and (52) imply

$$(53) \qquad 0 \leqq N_{1b}(u,x), \frac{\partial N_2(u,x)}{\partial u}, N_{3b}(u,x), \qquad u \in G(x).$$

Also, (50), (51), (52) and the definition of $G(x)$ can be used to establish

$$(54) \qquad 0 \leqq N_{1a}(u,x) + N_{3a}(u,x), \qquad u \in G(x).$$

Evidently (48) follows from (50), (53), and (54). We are done with Lemma 4.

4. The circle. In this section we consider the case where $\mathscr{S}$ is the closed unit disc $\mathscr{D}$ in $\mathbf{R}^2$ with the unit circle $\mathscr{C}$ in $\mathbf{R}^2$ as its boundary. We will show that result

(A) holds whenever play begins at a position which is *stable* (see paragraph 4 of the Introduction for a definition of stable).

We represent positions of P and E by points $p = (p_1, p_2)$ and $e = (e_1, e_2)$ in $\mathscr{D}$. Observe that the center O corresponds to the point $(0, 0)$ in $\mathbf{R}^2$. Let s, $-\pi < s \leqq \pi$, denote the angle that the directed line segment EP makes away from the direction EO (see Fig. 6). Whenever P's and E's right-hand velocity vectors exist, let ϕ and ψ denote the respective angles which these vectors make away from the direction PE, $-\pi < \phi, \psi \leqq \pi$. These angles are positive when they are measured in a counterclockwise arc, e.g., all the angles are positive in Fig. 6.

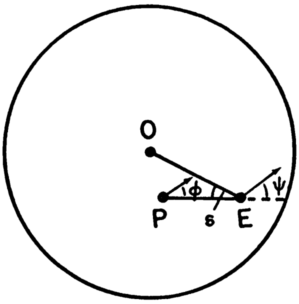

FIG. 6

Assumption 3. We will always assume that play starts at a position $(p, e) \in \mathscr{D} \times \mathscr{D}$ which is *stable* (see the Introduction). We also assume that $w > 1$. (The case $w = 1$ is trivial.)

Whenever $\|e\| = 1$ and $|s| < \beta$ where β satisfies (8), we say that P *corners* E. The reader should verify that whenever P corners E, P can force a decrease in $|PE|$ by heading directly toward E. Basically, we are interested in whether P can corner E at a position where $|PE| \leqq \|p - e\|$ (see paragraph 4 of § 2). We have the following lemma.

LEMMA 6. *If P can corner E at a position where $|PE| \leqq \|p - e\|$, P can do so with a trajectory* $\mathbf{p}$ *which satisfies*

$$(55) \qquad \|\mathbf{p}(t)\| \geqq \|p\|, \qquad t \in \mathbf{T}.$$

Proof. Assumption 3 implies that $s(0) = 0$. One can use the latter to show that restriction (55) does not affect P's ability to corner E.

Because of Lemma 6, we can make the following assumption.

Assumption 4. P is restricted to $\mathbf{p} \in \mathbf{P}^*$ which satisfy (55).

As in § 2, we introduce the notion of an *isometric rule* by describing its consequences. The *isometric rule* is that method of play for E under which the following holds when P selects a $\mathbf{p} \in \mathbf{P}^*(p)$ which satisfies (55). E travels at maximum speed along a path with a piecewise continuous derivative; and as long as

$$(56\text{a}) \qquad \|\mathbf{e}(t)\| < 1, \quad |s(t)| < \beta$$

or

$$(56\text{b}) \qquad \beta \leqq |s(t)| < \pi$$

holds, we have

$$\text{(10a)} \qquad \cos \psi(t) = (1/w) \cos \phi(t)$$

and

$$\text{(10b)} \qquad \psi(t) \begin{cases} >0 & \text{if } s(t) \geqq 0, \\ <0 & \text{if } s(t) < 0. \end{cases}$$

The proof of the next lemma is essentially the same as the proof of Lemma 1. We leave the details to the reader.

LEMMA 7. *If E follows the isometric rule, then as long as* (56) *holds we have*

$$\text{(57a)} \qquad \|\mathbf{p}(t) - \mathbf{e}(t)\| = \|p - e\|,$$

and

$$\text{(57b)} \qquad \frac{d\|\mathbf{e}(t)\|}{dt} = \cos \phi(t) \cos s(t) - (w^2 - \cos \phi(t))^{1/2} \sin |s(t)|.$$

Also, if for some $t_0 \in \mathbf{T}$,

$$|s(t_0)| \geqq \beta,$$

then E can guarantee that

$$\|\mathbf{p}(t) - \mathbf{e}(t)\| \geqq \|p - e\|, \qquad t \geqq t_0.$$

Remark 5. Using the fact that (10) is required to hold only for pursuit trajectories satisfying Assumption 4, one can show that the isometric rule is well-defined and satisfies the information constraint (see Remark 3).

The next lemma implies that given our assumptions, P can restrict himself to trajectories $\mathbf{p}$ for which $\|\mathbf{p}(\cdot)\|$ is an increasing function on $\mathbf{T}$. (This is not true in general. The critical assumption is that the initial value of s is 0.)

LEMMA 8. *If E follows the isometric rule and*

$$\text{(58)} \qquad \|\mathbf{p}(t_2)| \geqq \|\mathbf{p}(t_1)\|$$

holds for some $t_2 > t_1 \geqq 0$, *then either*

$$\text{(59)} \qquad s(t_2) \geqq s(t_1) \geqq 0,$$

or

$$\text{(60)} \qquad \min\{s(t_2), s(t_1)\} \geqq \beta.$$

Proof. Let α, $-\pi < \alpha \leqq \pi$, denote the angle that the directed line segment PO makes away from the direction PE. (Observe that we give α the same sign as s.) We want to show that $\alpha(t)$ is a nonincreasing function. A simple geometric argument establishes that as long as $\alpha > 0$,

$$\frac{d\alpha}{dt} = \frac{\sin[\alpha(t) - \phi(t)]}{\|\mathbf{p}(t)\|} - \frac{[w \sin \psi(t) - \sin \phi(t)]}{\|p - e\|}.$$

Hence (55) and (10) imply that

$$\frac{d\alpha}{dt} \leqq \frac{w - 1 - (w^2 - \cos^2 \phi(t))^{1/2} + \sin \phi(t)}{\|p - e\|} \leqq 0.$$

Suppose (58) holds for t_1 and t_2, where $t_2 > t_1 \geqq 0$. If $s(t_1) < \pi/2$, then the fact $\alpha(t_2) \leqq \alpha(t_1)$ implies (59). Similarly, if $s(t_1) \geqq \pi/2$, then the fact $\alpha(t_2) \leqq \alpha(t_1)$ implies that $s(t_2) \geqq \pi/2$. The observation that $\beta \leqq \pi/2$ finishes the lemma.

LEMMA 9. *If E follows the isometric rule, then P can corner E only if he can do so with a trajectory* **p** *which satisfies*

$$(12) \qquad 0 \leqq \phi(t) \leqq \arccos(w \sin s(t))$$

until cornering takes place.

Proof. Because of Lemma 8, P need only consider trajectories **p** for which $d\|e(t)\|/dt \leqq 0$. By Lemma 7, the latter condition implies that $\phi(t)$ should satisfy

$$(13) \qquad |\phi(t)| \leqq \arccos(w \sin s(t)) \leqq \pi/2.$$

(Assumption 3 allows us to use $s(t)$ instead of $|s(t)|$ in the above formula.) The rest is essentially the same as the proof of Lemma 3.

In order to state the analogues of Lemmas 4 and 5, we need another definition of convexity for trajectories. Suppose that P selects a trajectory **p** which satisfies (55). As the vector $\mathbf{p}(t)$ rotates about O it sweeps out a directed angle. (Following the usual convention, we let positive angles correspond to counterclockwise displacements.) We can represent this angle by a continuous function $\theta(\cdot)$ on $\mathbf{T}$, where $\theta(t)$, $-\infty < \theta(t) < \infty$, is interpreted as the directed angle generated between time 0 and time t. We say that the *pursuit trajectory* **p** is *convex* on an interval $\mathbf{T}_0 \subset \mathbf{T}$ if (a′) $\theta(\cdot)$ is a monotone function on $\mathbf{T}_0$, and (b′) the set

$$\bigcup_{t_0 \leqq t \leqq t_1} \{p \in \mathscr{D} | p = k\mathbf{p}(t) \quad \text{for some } k \in [0, 1]\}$$

is convex whenever $|\theta(t_0) - \theta(t_1)| \leqq \pi$ and $[t_0, t_1] \subset \mathbf{T}_0$. A similar definition holds for evasion trajectories. (As before (see the paragraph preceding Lemma 3), we will omit any specific reference to $\mathbf{T}_0$ whenever its identity is clear from the context.)

LEMMA 10. *If E follows the isometric rule, then P can corner E only if he can do so with a convex trajectory* **p** *which satisfies* (12) *until cornering takes place.*

Proof. See the proof of Lemma 4.

LEMMA 11. *If P uses a convex* **p** *for which* $\phi(t) \geqq 0$, $t \in \mathbf{T}$, *while E follows the isometric rule, then E's trajectory* **e** *is convex on the interval* $[0, t_0]$, *where* t_0 *denotes the first time that condition* (56) *fails.*

Proof. See the proof of Lemma 5.

Now we can state our second main result (see Remark 1).

THEOREM 2. *Let the initial position* $(e, p) \in \mathscr{D} \times \mathscr{D}$ *be stable and let* $V(e, p)$ *be defined by* (5). *Then*

$$(\text{II}) \qquad V(e, p) < \|e - p\|$$

holds if and only if P can corner E with a $\mathbf{p} \in \mathbf{P}^*$ *whenever E uses the isometric rule. Furthermore, whenever* (II) *fails to hold, we have*

$$(\text{I}) \qquad V(e, p) = \|e - p\|.$$

The next corollary follows immediately from Theorem 2.

COROLLARY 2. (A) *holds for stable starting positions in the unit circle.*

Proof of Theorem 2. The theorem follows from Lemmas 6, 10 and 11 via the arguments of Theorem 1.

5. Application. In this section we apply the results of the last section to the problem of *pursuit in the circle.*

Given any d, $0 \leqq d \leqq (\dot{w} - 1)/w$, let $(e^d, p^d) \in \mathscr{D} \times \mathscr{D}$ represent any stable position for which $\|e^d - p^d\| = d$. Define

$$d^* = \sup\{d | d = V(e^d, p^d)\}. \tag{61}$$

The following result is essentially the same as one formulated earlier by Gerald Smith [8].

THEOREM 3. $V(e, p) \leqq d^*$, $(e, p) \in \mathscr{D} \times \mathscr{D}$.

Remark 6. Evidently d^* is the least upper bound of the set of values of $|PE|$ that E can maintain from any starting position. One can—at least in principle—use the characterization of d^* implicit in Theorem 2 to obtain numerical values for d^*. In Theorem 4 we use that characterization to find an upper bound on d^*. The proofs of both of these theorems are deferred until the end of this section.

THEOREM 4. $v^* < d^* < (w^2 - 1)/([w^2 + wE(1/w)]^2 + (w + 1)^2)^{1/2}$, *where*

$$v^* = (1/w)[(w^2 - 1)^{1/2} - \arccos(1/w)] \tag{62a}$$

and

$$E(1/w) = \int_0^{\pi/2} [1 - (\sin u/w)^2]^{1/2} \, du \tag{62b}$$

is an elliptic integral of the second kind.

Some sample values are found in Table 1. The numbers are rounded off to the nearest hundredth.

TABLE 1
Tabled values of bounds on d^*

Value of w	1.00	1.50	2.00	5.00	15.00	∞
Upper bound	.00	.25	.40	.73	.90	1.00
Lower bound	.00	.18	.34	.71	.90	1.00

As mentioned earlier, P can always force play into a position which is approximately stable. Specifically, one can show that given any $\varepsilon > 0$, any initial position $(e, p) \in \mathscr{D} \times \mathscr{D}$ and any evasion strategy $\eta \in H(e, p)$, P has a pursuit trajectory $\mathbf{p} \in \mathbf{P}(p)$ which leads to a position in $\mathscr{D} \times \mathscr{D}$ which is at a distance (in the product norm) less than ε from some stable position. We will use this fact.

Remark 7. Following a suggestion given by L. Dubins, Gerald Smith [8] established that for any $\varepsilon > 0$, P has a *strategy* which brings him to a point Q on OM which satisfies $|OQ| \geqq (1/w)|OM| - \varepsilon$. The idea behind Smith's strategy is to have P go to the center and then move out along the radial line OM. We can apply his results to our situation by having P select a *trajectory* which takes him to the center and then keeps him on the radial line passing through the position which M occupied $\varepsilon/(w + 1)$ time units before. We leave the details to the reader.

We need the following lemma.

LEMMA 12. *If* $V(e^{d_2}, p^{d_2}) = d_1$ *where* $d_1 \neq d_2$, *then* $d_2 > d_1$ *and* $V(e^{d_1}, p^{d_1}) = d_1$. *Furthermore* $V(\cdot\,,\,\cdot)$ *is a continuous function on* $\mathscr{D} \times \mathscr{D}$.

Proof. That $d_2 > d_1$ is immediate. Starting from the position (e^{d_1}, p^{d_1}), let E travel outward at maximum speed along the radius through e^d. Clearly P's best trajectory follows E along this line. Observe that the distance $|PE|$ remains greater than d_1 until a position (e^{d_2}, p^{d_2}) is reached. Since by hypothesis $V(e^{d_2}, p^{d_2}) = d_1$, we have $V(e^{d_1}, p^{d_1}) = d_1$. The proof of the continuity of $V(\cdot\,,\,\cdot)$ is left to the reader.

Now we can prove Theorems 3 and 4.

Proof of Theorem 3. The continuity of $V(\cdot\,,\,\cdot)$ and the remarks preceding Remark 7 give us

$$V(e,p) \leqq \sup_{0 \leqq d \leqq (w-1)/w} V(e^d, p^d), \qquad (e,p) \in \mathscr{D} \times \mathscr{D}. \tag{63}$$

But the first part of Lemma 12 implies that the right-hand side of (63) is equal to d^*. The theorem follows.

Proof of Theorem 4. To get the lower bound we restrict E to the boundary $\mathscr{C}$. That $v^* \leqq d^*$ follows from Theorem 2 of Flynn [2]. (Compare (62) above with (2) of Flynn [2].) One can establish that strict inequality holds by showing that E can do better by traveling along small chords than by staying on the circumference. (This fact is established in [3].)

Now we justify the upper bound. Suppose play starts at (e^d, p^d) and we restrict P to straight-line paths which are perpendicular to the line joining p^d to e^d at p^d. One can show that if d is greater than the right-hand side of Theorem 4, then P can force a decrease in $|PE|$. (See § 10 of Flynn [1] for details.) The theorem follows.

REFERENCES

[1] J. FLYNN, *Lion and man: A bounded pursuit game*, Tech. Rep. 161, Statistics Department, Stanford University, Stanford, Calif., 1970.

[2] ———, *Lion and man: The boundary constraint*, this Journal, 11 (1973), pp. 397–411.

[3] ———, *Lion and man: The general case*, this Journal, to appear.

[4] B. HALPERN, *The robot and the rabbit—a pursuit problem*, Amer. Math. Monthly, 76 (1969), pp. 140–144.

[5] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

[6] J. LITTLEWOOD, *Lion and Man: A Mathematician's Miscellany*, Methuen, London, 1953.

[7] C. RYLL-NARDZEWSKI, *A Theory of Pursuit and Evasion: Advances in Game Theory*, Princeton University Press, Princeton, 1964.

[8] GERALD J. SMITH, *A pursuit evasion game*, unpublished manuscript, 1968.

SIAM J. Control
Vol. 12, No. 1, February 1974

NEUTRAL AUTONOMOUS FUNCTIONAL EQUATIONS WITH QUADRATIC COST*

R. DATKO†

Abstract. In this paper a control problem for neutral functional equations with a quadratic cost function is considered. It is shown that the optimal control is a feedback control. If the problem can be optimized over the positive half-line, then the solution of the problem is obtained by solving a linear homogeneous functional equation which possesses a type of exponential stability.

Introduction. In this paper we extend some of the results in [2] to a control problem involving linear functional equations of neutral type. The problem is to optimize a quadratic functional which has a constraint involving an n-dimensional neutral functional equation. It is shown that the optimal solution is unique and satisfies a linear feedback law. In addition, if the initial functions are continuously differential, then the solution of the problem can be found among the solutions of a $2n$-dimensional linear homogeneous differential-difference equation. In case the problem can be optimized over the positive half-line, we demonstrate that the optimal solution leads to an autonomous linear homogeneous functional equation and the the spectrum of this equation lies in a half-plane $\operatorname{Re} z \leqq -a_0, a_0 > 0$. This last result might lead one to expect that the solutions of the functional equation are exponentially stable. However, we are unable to prove this. The best we can do is to show that for initial conditions which are continuously differentiable, a form of exponential stability holds. This is made precise by Theorem 4. The problem discussed here is the most elementary of its kind, and the results of this paper can be extended to more complex functional equations. The main reason for not considering a more complex problem is that certain computational aspects, particularly in § 2, are involved enough as it is and would soon get out of hand for any significant extension of the problem.

The paper is divided into three parts. Section 1 sets down some conventions and assumptions which will be used throughout the paper. Section 2 discusses the control problem over finite intervals. In this section we compute, in a suitable Hilbert space, the Fréchet derivative of the functional to be optimized, set it equal to the zero vector, and from this obtain the characterization of the optimal control and optimal trajectory. To be more explicit, we show, in Theorem 1, that the point which minimizes the functional satisfies an n-dimensional linear differential-difference equation. Section 3 deals with the problem over the infinite interval, and the basic result, Theorem 3, is obtained as a limiting case of the problem in § 2. General references for the functional equations considered in this paper are [1] or [3] and a reference for the semigroups in § 3 is [4].

1. Preliminary notation.

1. We shall use standard vector and matrix notation. Unless otherwise specified, all matrices and vectors are real. Vectors will be denoted by lower-case letters and matrices by upper-case letters. The complex inner product of two complex

* Received by the editors September 7, 1972.

† Department of Mathematics, Georgetown University, Washington, D.C. 20007.

vectors x and y of the same dimension will be denoted by (x, y). The norm of a vector x in R^n or C^n will be denoted by $|x|$. The conjugate transpose of a matrix A will be denoted by A^*.

2. The symbol $C[-h, 0]$ will stand for the Banach space of all continuous mappings from the closed interval $[-h, 0]$ into R^n or C^n as the cases may be. The norm of a point ϕ in $C[-h, 0]$ will be given by $\|\phi\| = \sup_{-h \leqq s \leqq 0} |\phi(s)|$. If $\phi \in C[-h, 0]$ is continuously differentiable, then its C_1-norm is given by

$$\|\phi\|_1 = \sup_{-h \leqq s \leqq 0} |\phi(s)| + \sup_{-h \leqq s \leqq 0} |\dot{\phi}(s)|.$$

If $t_0 \geqq 0$, then $\phi_{t_0}(\cdot)$ will denote a continuous mapping from $[t_0 - h, t_0]$ into $R^n(C^n)$. Clearly $\phi_{t_0}(\cdot)$ is a point in $C[-h, 0]$. If g is a continuous linear functional on $C[-h, 0]$, its value at a point ϕ will be denoted by $\langle g, \phi \rangle$.

3. L_2 will stand for the equivalence classes of all measurable square integrable mappings from $[0, \infty)$ into $R^m(C^m)$. This space is a Hilbert space with inner product $\langle v, u \rangle = \int_0^\infty (v(t), u(t))\, dt$ and norm $\|u\| = \langle u, u \rangle^{1/2}$. We shall frequently consider measurable mappings over intervals of the form $I = [t_0, T]$. Clearly, if a mapping is square integrable over I, then it is square integrable over $[0, \infty)$ if we extend it by defining it to be the zero vector on the complement of I.

4. If q is a continuous functional on L_2, then q is said to have a Fréchet derivative q' at a point u_0 in L_2 if there exists a point g in L_2 and a functional r on L_2 such that

$$q(u) = q(u_0) + \langle g, u - u_0 \rangle + r(u),$$

$$\frac{r(u)}{\|u - u_0\|} \to 0 \quad \text{as } \|u - u_0\| \to 0.$$

5. A, B and D will denote specific $n \times n$ real constant matrices and E will be a specific real constant $n \times m$ matrix. W will be a real $n \times n$ symmetric positive definite matrix and U will be a real $m \times m$ positive definite symmetric matrix. The conditions on W and U imply that there exist positive constants w_1, w_2, u_1 and u_2 such that

$$w_1|x|^2 \leqq (Wx, x) \leqq w_2|x|^2 \quad \text{and} \quad u_1|y|^2 \leqq (Uy, y) \leqq u_2|y|^2$$

for all n-vectors x and m-vectors y.

2. Statement of the basic problem and some of its properties. Let $0 \leqq t_0 \leqq T < \infty$ and let ϕ_{t_0} be in $C[-h, 0]$. The problem is to minimize the functional on L_2 defined by the equation

$$J(u, \phi_{t_0}, t_0, T) = \int_{t_0}^{T} [(Wx_u(t), x_u(t)) + (Uu(t), u(t))]\, dt, \tag{1}$$

where $x_u(t)$ satisfies the equation

$$(2)\qquad x_u(t) = \begin{cases} \phi(t_0) - D\phi(t_0 - h) + Dx_u(t-h) \\ \quad + \displaystyle\int_{t_0}^{t} [Ax_u(s) + Bx_u(s-h) + Eu(s)]\,ds & \text{if } t \geqq t_0, \\ \phi(t) \quad \text{if } t \in [t_0 - h, t_0]. \end{cases}$$

It is shown in [3] that for any ϕ_{t_0} in $C[-h, 0]$ and $u \in L_2$, there exists a unique solution of (2) in $C[-h, 0]$. Furthermore, if ϕ_{t_0} is continuously differentiable on $[-h, 0]$, then (2) can be replaced by the differential equation

$$(2')\qquad \begin{aligned} \dot{x}_u(t) &= Ax_u(t) + Bx_u(t-h) + D\dot{x}_u(t-h) + Eu(t), \\ x_u(t) &= \phi(t) \quad \text{on } [t_0 - h, t_0]. \end{aligned}$$

Remark 1. We shall sometimes denote the dependence of x_u on ϕ by $x_u(t, \phi)$ or $x_u(\cdot, \phi)$.

DEFINITION 1. Given ϕ_{t_0} in $C[-h, 0]$ we define

$$(3)\qquad j(\phi_{t_0}, t_0, T) = \inf_{u \in L_2} J(u, \phi_{t_0}, t_0, T).$$

If there exists $u \in L_2$ for which the infimum in (3) is attained, this u will be called an optimal control and denoted by $u^m(t, \phi_{t_0}, t_0, T)$ or by $u^m(t)$ if ϕ_{t_0}, t_0 and T are not important to the discussion at hand. The solution of (2) corresponding to an optimal control is called an optimal trajectory and denoted by $x^m(t, \phi_{t_0}, t_0, T)$ or simply by $x^m(t)$.

We shall now construct a variation of parameters formula for (2) or (2′) (see, e.g., [1, pp. 320–323]). The reason for rederiving the formula in [1] is that our notation is somewhat different and we wish to emphasize the role played by the initial function if it is differentiable.

Let $S(t, \alpha)$ denote the unique $n \times n$ matrix which satisfies the conditions

$$(4a)\qquad \frac{\partial S}{\partial \alpha}(t, \alpha) = -S(t, \alpha)A - S(t, \alpha + h)B + \frac{\partial S}{\partial \alpha}(t, \alpha + h)D$$

$$\text{if } t_0 < \alpha < t, \quad \alpha \neq t - nh, \quad n = 0, 1, 2 \cdots,$$

$$(4b)\qquad S(t, t) = I,$$

$$(4c)\qquad S(t, \alpha) = 0 \quad \text{if } \alpha > t$$

and

$$(4d)\qquad S(t, \alpha) - S(t, \alpha + h)D$$

is continuous for all α in $[t_0, t]$.

Remark 2. Hale and Meyer [3, p. 13] have shown that $S(t, \alpha)$ is dependent only on the difference $t - \alpha$. That is, $S(t, \alpha) = S(t - \alpha, 0)$. This is a consequence of the autonomy of (4a).

Let $\phi_{t_0} \in C[-h, 0]$ be continuously differentiable and let $u \in L_2$. Let $x(t)$ be the solution of (2′) for the given pair ϕ_{t_0} and u. Then if $S(t, \alpha)$ satisfies (4), the following identity holds:

$$
\begin{aligned}
&-\int_{t_0}^{t}[S(t,\alpha)A + S(t,\alpha+h)B]x(\alpha)\,d\alpha \\
&\quad = -\int_{t_0}^{t} S(t,\alpha)Ax(\alpha)\,d\alpha - \int_{t_0+h}^{t} S(t,\alpha)Bx(\alpha-h)\,d\alpha \\
&\quad = \int_{t_0}^{t}\frac{\partial}{\partial\alpha}[S(t,\alpha) - S(t,\alpha+h)D]x(\alpha)\,d\alpha \\
(5)\qquad &\quad = x(t) - [S(t,t_0) - S(t,t_0+h)D]\phi(t_0) \\
&\qquad - \int_{t_0}^{t} S(t,\alpha)[Ax(\alpha) + Bx(\alpha-h) + D\dot{x}(\alpha-h) + Eu(\alpha)]\,d\alpha \\
&\qquad + \int_{t_0+h}^{t} S(t,\alpha)D\dot{x}(\alpha+h)\,d\alpha.
\end{aligned}
$$

Comparing the extreme right-hand side of (5) and the terms in the second equality from the left in (5) and making some obvious cancellations and rearrangements, we obtain

$$
\begin{aligned}
(6)\qquad x(t) &= [S(t,t_0) - S(t,t_0+h)D]\phi(t_0) \\
&\quad + \int_{t_0-h}^{t_0} S(t,\alpha+h)[B\phi(\alpha) + D\dot{\phi}(\alpha)]\,d\alpha + \int_{t_0}^{t} S(t,\alpha)Eu(\alpha)\,d\alpha.
\end{aligned}
$$

If $x_0(t)$ is a solution of (2) for ϕ_{t0} and $u = 0$, then on the basis of (6) we can write the solution of (2) for ϕ_{t_0} and u in the form

$$
(7)\qquad x_u(t,\phi_{t_0}) = x_0(t) + \int_{t_0}^{t} S(t,\alpha)Eu(\alpha)\,d\alpha.
$$

Remark 3. If ϕ_{t_0} has a continuous derivative and u in L_2 is continuous on the interval $[t_0, T]$ except possibly at the points $t = nh + t_0$, $n = 0, 1, 2, \cdots$, then $x(t, \phi_{t_0}, u)$ is differentiable on $[t_0, T]$ except possibly at the points $\{nh + t_0\}$. This is an immediate consequence of the representation given for the solution in (6).

Let $\phi_{t_0} \in C[-h, 0]$ and $u_0 \in L_2$ be fixed, and let $h \in L_2$ be arbitrary. Abbreviating $J(u, \phi_{t_0}, t_0, T)$ to $J(u)$ and setting

$$
(8)\qquad y(h,t) = \int_{t_0}^{t} S(t,\alpha)Eh(\alpha)\,d\alpha,
$$

we form the difference

$$
\begin{aligned}
J(u_0 + h) - J(u_0) &= 2\left[\int_{t_0}^{T} (Wx_{u_0}(t), y(h,t))\,dt + \int_{t_0}^{T} (Uu_0(t), h(t))\,dt\right] \\
&\quad + \int_{t_0}^{T} (Uh(t), h(t))\,dt + \int_{t_0}^{T} (Wy(h,t), y(h,t))\,dt.
\end{aligned}
$$

Thus we see that the Fréchet derivative of $J(u_0)$ exists and is given by the expression

$$(9)\qquad \langle J'(u_0), h\rangle = 2\int_{t_0}^{T} (Wx_{u_0}(t), y(h,t))\,dt + 2\int_{t_0}^{T} (Uu_0(t), h(t))\,dt.$$

Making use of (8), we can explicitly compute $J'(u_0)$ for

$$\begin{aligned}\int_{t_0}^{T} (Wx_{u_0}(t), y(h,t))\,dt &= \int_{t_0}^{T}\int_{t_0}^{t} (Wx_{u_0}(t), S(t,\alpha)Eh(\alpha))\,d\alpha\,dt \\ &= \int_{t_0}^{T}\int_{\alpha}^{T} (E^*S^*(t,\alpha)Wx_{u_0}(t), h(\alpha))\,dt\,d\alpha \\ &= \int_{t_0}^{T}\left\{\int_{\alpha}^{T} E^*S^*(t,\alpha)Wx_{u_0}(t)\,dt, h(\alpha)\right\}d\alpha.\end{aligned}$$

Substitution of the above expression into (9) and observing that (9) holds for all h in L_2, we obtain

$$(10)\qquad J'(u_0)(\alpha) = 2\left[Uu_0(\alpha) + E^*\int_{\alpha}^{T} S^*(t,\alpha)Wx_{u_0}(t)\,dt\right].$$

The following two properties of $J(u)$ will be needed in the sequel. Their proofs are omitted since they have been given in [2].

Property 1. For ϕ_{t_0} fixed in $C[-h,0]$, the infimum of $J(u,\phi_{t_0},t_0,T)$ is uniquely attained and satisfies the condition $J'(u^m,\phi_{t_0},t_0,T) = 0$. This is equivalent to the relation

$$(11)\qquad u^m(\alpha) = -U^{-1}E^*\int_{\alpha}^{T} S^*(t,\alpha)Wx^m(t)\,dt.$$

Moreover, for each ϕ_{t_0} in L_2, (11) has a unique solution.

Property 2. Let a and b be real scalars. If u^m and v^m are, respectively, optimal controls for $J(u,\phi_{t_0},t_0,T)$ and $J(u,\psi_{t_0},t_0,T)$, then $au^m + bv^m$ is the optimal control for $J(u, a\phi_{t_0}, +b\psi_{t_0}, t_0, T)$. In other words, the optimal control associated with (1)–(2) induces a linear mapping from $C[-h,0]$ into L_2.

On the basis of Property 2 and equation (11), we introduce the following definition.

DEFINITION 2. Let $\phi_{t_0}\in C[-h,0]$, and let $x^m(t,\phi_{t_0})$ denote the optimal trajectory for the problem (1)–(2). For each $t\in[t_0,T]$, define the *linear mapping from* $C[-h,0]$ *into* R^n by the relation

$$(12a)\qquad L(t,t_0,T)\phi_{t_0} = \int_{t}^{T} S^*(\sigma,t)Wx^m(\sigma,\phi_{t_0})\,d\sigma$$

if $t_0 \leqq t \leqq T$ and if $t \geqq T$ by

$$(12b)\qquad L(t,t_0,T)\phi_{t_0} = 0.$$

Remark 4. The linearity of $L(t,t_0,T)$ is a consequence of Property 2. For by Property 1 and Definition 2,

$$(13)\qquad u^m(t,\phi_{t_0}) = -U^{-1}E^*L(t,t_0,T)\phi_{t_0}.$$

Hence, since equation (2) is linear in u, it follows from Property 2 that $L(t, t_0, T)$ is linear in ϕ_{t_0}.

LEMMA 1. *Let $g:[t_0, T] \to R^n$ be continuous and let*

$$y(t) = \begin{cases} \displaystyle\int_t^T S^*(\sigma, t)g(\sigma)\,d\sigma, & t_0 \leqq t \leqq T, \\ 0 \quad \text{if } t \geqq T. \end{cases}$$

If $t \neq T - nh$, $n = 0, 1, 2, \cdots$, then on $[t_0, T]$, $y(t)$ has a derivative which satisfies the equation

$$\frac{dy}{dt}(t) = -g(t) - A^*y(t) - B^*y(t + h) + D^*\frac{d(y(t + h))}{dt}.$$

Proof. Assume $t = T - nh - \tau$, where $0 < \tau < h$. Then

$$y(t) = \sum_{i=0}^{n-1} \int_{t+ih}^{t+(i+1)h} S^*(\sigma, t)g(\sigma)\,d\sigma + \int_{t+nh}^{T} S^*(\sigma, t)g(\sigma)\,d\sigma.$$

Notice, since $0 < \tau < h$, that for sufficiently small changes in t, n in the above expression does not vary. Moreover, since each term is differentiable, $y(t)$ has a derivative. By Remark 2, $S^*(\sigma, t) = S^*(\sigma - t, 0)$. If we use the notation $S^*(ih^-, 0) = \lim_{t \to ih^-} S^*(t, 0)$ and $S^*(ih^+, 0) = \lim_{t \to ih^+} S^*(t, 0)$, then the derivative of $y(t)$ is given by

$$\begin{aligned} \dot{y}(t) = {} & -g(t) + \sum_{i=1}^{n} [S^*(ih^-, 0) - S^*(ih^+, 0)]g(t + ih) \\ & + \int_t^T \frac{\partial S^*}{\partial t}(\sigma, t)g(\sigma)\,d\sigma. \end{aligned} \tag{14}$$

Since $S(\sigma, t)$ satisfies the relations in (4), we obtain from (14):

$$\begin{aligned} \dot{y}(t) = {} & -g(t) - A^*y(t) - B^*y(t + h) \\ & + \sum_{i=1}^{n} [S^*(ih^-, 0) - S^*(ih^+, 0)]g(t + ih) \\ & - D^* \sum_{i=1}^{n} [S^*((i - 1)h^-, 0) - S^*((i - 1)h^+, 0)]g(t + ih) \\ & + D^* \sum_{i=1}^{n} [S^*((i - 1)h^-, 0) - S^*((i - 1)h^+, 0)]g(t + ih) \\ & + D^* \sum_{i=1}^{n-1} \int_{t+ih}^{t+(i+1)h} \frac{\partial S^*}{\partial t}(\sigma, t + h)g(\sigma)\,d\sigma \\ & + D^* \int_{t+nh}^{T} \frac{\partial S^*}{\partial t}(\sigma, t + h)g(\sigma)\,d\sigma. \end{aligned} \tag{15}$$

Note that by (14), the last three terms in (15) can be replaced by $D^*\dot{y}(t + h)$.

Also by (4d), $S^*(t, 0) - D^*S^*(t - h, 0)$ is continuous for all t. Thus (15) becomes

$$\dot{y}(t) = -g(t) - A^*y(t) - B^*y(t + h) + D^*\dot{y}(t + h),$$

which proves the lemma.

THEOREM 1. *Let ϕ_{t_0} be in $C[-h, 0]$ and let $L(t, t_0, T)\phi_{t_0}$ be given by Definition 2. Then, for t a.e. on $[t_0, T]$, $L(t, t_0, T)\phi_{t_0}$ satisfies the differential equation*

$$\dot{y}(t) = -Wx^m(t, \phi_{t_0}) - A^*y(t) - B^*y(t + h) + D^*\dot{y}(t + h). \tag{16}$$

Proof. $Wx^m(t, \phi_{t_0})$ is continuous on $[t_0, T]$, hence $L(t, t_0, T)\phi_{t_0}$ satisfies the hypotheses of Lemma 1.

COROLLARY. *If ϕ_{t_0} has a continuous derivative on $[t_0 - h, t_0]$, then the solution of the control problem* (1)–(2) *satisfies the* 2*n-th order system of differential-difference equations given by*

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bx(t - h) + D\dot{x}(t - h) - EU^{-1}E^*y(t), \\ \dot{y}(t) &= -Wx(t) - A^*y(t) - B^*y(t + h) + D^*\dot{y}(t + h). \end{aligned} \tag{17}$$

The proof of the following property is straightforward but lengthy. It is therefore omitted. The proof is similar to Lemma 3.1 in [2].

Property 3. If $\{\phi_{t_0}^n\} \to \phi_{t_0}$ in $C[-h, 0]$, then $\{u^m(\cdot, \phi_{t_0}^n)\} \to u^m(\cdot, \phi_{t_0})$ in L_2, and $\{x^m(\cdot, \phi_{t_0}^n)\} \to x(\cdot, \phi_{t_0})$ uniformly on $[t_0, T]$.

DEFINITION 3. Let $\phi_{t_0}^i$, $i = 1, 2$, be in $C[-h, 0]$, and let the pairs (x_1, u_1) and (x_2, u_2) denote the optimal trajectory and optimal control corresponding respectively to $\phi_{t_0}^1$ and $\phi_{t_0}^2$. Let $L(t, t_0, T)$ be abbreviated to $L(t)$. We define the *bilinear form on* $C[-h, 0] \times C[-h, 0]$ given by

$$\begin{aligned} R(t, t_0, T)(\phi_{t_0}^1, \phi_{t_0}^2) &= (L(t)\phi_{t_0}^1, x_2(t) - Dx_2(t - h)) \\ &\quad + \int_{t-h}^{t} (L(\alpha + h)\phi_{t_0}^1, Bx_2(\alpha))\,d\alpha \\ &\quad - \int_{t-h}^{t} \left(\frac{d}{d\alpha}(L(\alpha + h)\phi_{t_0}^1), Dx_2(\alpha)\right) d\alpha. \end{aligned} \tag{18}$$

THEOREM 2. *The bilinear form in Definition* 3 *satisfies the equation*

$$\begin{aligned} R(t_0, t_0, T)(\phi_{t_0}^1, \phi_{t_0}^2) &= R(t_0, t_0, T)(\phi_{t_0}^2, \phi_{t_0}^1) \\ &= \int_{t_0}^{T} [(Wx_1(t), x_2(t)) + (Uu_1(t), u_2(t))]\,dt. \end{aligned} \tag{19}$$

Thus if $\phi_{t_0}^1 = \phi_{t_0}^2$, $R(t_0, t_0, T)(\phi_{t_0}^1, \phi_{t_0}^1) = j(\phi_{t_0}^1, t_0, T)$.

Proof. Let $\phi_{t_0}^1$ and $\phi_{t_0}^2$ be in $C[-h, 0]$, and assume that $\phi_{t_0}^2$ is continuously differentiable on $[t_0 - h, t_0]$. Let $L(t, t_0, T)\phi_{t_0}^1 = L(t)\phi_{t_0}^1 = y_1(t)$. By Remark 3 and equation (11), $x_2(t)$ has a derivative a.e. on $[t_0, T]$. A straightforward computation using (2′) and (16) yields

$$\frac{d}{dt}[R(t, t_0, T)(\phi_{t_0}^1, \phi_{t_0}^2)] = -(Wx_1(t), x_2(t)) + (E^*y_1(t), u_2(t))$$

a.e. on $[t_0, T]$. Since $-U^{-1}E^*y_1(t) = u_1(t)$, it follows that

$$R(t, t_0, T)(\phi_{t_0}^1, \phi_{t_0}^2) = R(t_0, t_0, T)(\phi_{t_0}^1, \phi_{t_0}^2) - \int_{t_0}^{t} [(Wx_1(s), x_2(s)) + (Uu_1(s), u_2(s))]\, ds. \tag{20}$$

However, $R(T, t_0, T)(\phi_{t_0}^1, \phi_{t_0}^2) = 0$ because $L(t, t_0, T) = 0$ for $t \geqq T$. Thus the conclusion of the theorem holds if $\phi_{t_0}^2$ is continuously differentiable. Using Property 3, the fact that the continuously differentiable mappings are dense in $C[-h, 0]$, and that $R(t_0, t_0, T)$ is continuous, we reach the conclusion that the theorem holds for any $\phi_{t_0}^2$ in $C[-h, 0]$.

COROLLARY 1. *For any* $t \in [t_0, T]$ *and* $\phi_{t_0}^1$ *and* $\phi_{t_0}^2$ *in* $C[-h, 0]$,

$$R(t, t_0, T)(\phi_{t_0}^1, \phi_{t_0}^2) = \int_t^T [(Wx_1(t), x_2(t)) + (Uu_1(t), u_2(t))]\, dt. \tag{21}$$

Proof. This corollary is a consequence of Theorem 2 and the observation that if x^m and u^m are optimal for $J(u, \phi_{t_0}, t_0, T)$, then they are also optimal for $J(u, x_t^m, t, T)$.

3. The case $T = \infty$. In this section, we shall consider the optimal control problem posed by equations (1)–(2) when $T = \infty$. To consider this case, it is necessary to make an additional assumption which will be given below. We shall first consider some further properties of the problem in case T is finite. Property 4 is a consequence of Property 1.

Property 4. If $u^m(t, t_0, T, \phi_{t_0})$ is an optimal control for the problem (1)–(2), then it is also optimal for the functional $J(u, x_t^m(\phi_{t_0}), t, T)$.

The next property has been proven in [2, Property 4, § 3].

Property 5. If $t_0 > 0$ and $\phi_{t_0}(t + s) = \psi(s)$, where $s \in [-h, 0]$, then $j(\phi_{t_0}, t_0, T + t_0) = j(\psi, 0, T)$ and $L(t + t_0, t_0, T + t_0)\phi_{t_0} = L(t, 0, T)\psi$.

The next property is the feedback control mentioned in the Introduction. Its proof also can be found in [2, Property 3, § 3].

Property 6. For each ϕ_{t_0} in $C[-h, 0]$, the identity $L(t, t_0, T)\phi_{t_0} = L(t, t, T)x_t^m \cdot (\cdot, \phi_{t_0}, t_0, T)$ holds. Thus $u^m(t, \phi_{t_0}) = -U^{-1}E^*L(t, t, T)x_t^m(\cdot, \phi_{t_0}, t_0, T)$.

Hypothesis H. For all $t_0 \geqq 0$ and each ϕ_{t_0} in $C[-h, 0]$, it will be assumed that $\lim_{T \to \infty} j(\phi_{t_0}, t_0, T) < \infty$.

Remark 5. What Hypothesis H says is that the problem (1)–(2) can be optimized for $T = \infty$ and all ϕ_{t_0}.

The next property is proved in [2, Theorem 3.1 and its corollary].

Property 7. Assume Hypothesis H holds; then there exists a continuous symmetric bilinear form R on $C[-h, 0] \times C[-h, 0]$ and a continuous linear mapping $q: C[-h, 0] \to BV[-h, 0]$ (the space of n-dimensional functions of bounded variation on $[-h, 0]$) such that for ϕ_{t_0} and ψ_{t_0} in $C[-h, 0]$,

$$\lim_{T \to \infty} R(t_0, t_0, T)(\psi_{t_0}, \phi_{t_0}) = R(\psi_{t_0}, \phi_{t_0}) = \langle q\phi_{t_0}, \psi_{t_0} \rangle = \langle q\psi_{t_0}, \phi_{t_0} \rangle. \tag{22}$$

Moreover, for each $\phi_{t_0} \in C[-h, 0]$,

$$\lim_{T \to \infty} L(t_0, t_0, T)\phi_{t_0} = \lim_{s \to 0^-} (q\phi_{t_0}(0) - q\phi_{t_0}(s)).$$

DEFINITION 4. Define the mapping $L: C[-h, 0] \to R^n$ by

$$L\phi_{t_0} = \lim_{s \to 0^-} [(q\phi_{t_0})(0) - (q\phi_{t_0})(s)]. \tag{23}$$

Property 8. The mapping L is linear and continuous.

Proof. The proof is a consequence of the fact that $q: C[-h, 0] \to BV[-h, 0]$ is linear and continuous.

Using Property 8 and Definition 4, we can prove in a manner analogous to Theorem 3.2 in [2] the following theorem.

THEOREM 3. *Assume Hypothesis* H *holds and let L be the linear mapping defined by equation* (23). *Consider the functional equations*

$$x(t) = \begin{cases} \phi(t_0) - D\phi(t_0 - h) + \displaystyle\int_{t_0}^{t} [Ax(s) + Bx(s - h)]\, ds \\ \qquad + Dx(t - h) - \displaystyle\int_{t_0}^{t} EU^{-1}E^*Lx_s\, ds \quad \textit{for } t \geqq t_0, \\ \phi(t) \quad \textit{for } t_0 - h \leqq t \leqq t_0. \end{cases} \tag{24}$$

Denote the solution of (24) by $x_t(\phi_{t_0})$. Then, given any sequence $\{T_n\} \to \infty$, the following holds:

$$\lim_{n \to \infty} x^m(\cdot, \phi_{t_0}, t_0, T_n) = x_t(\phi_{t_0}) \quad \text{for each } t \in [t_0, \infty), \tag{25a}$$

$$\lim_{n \to \infty} j(\phi_{t_0}, t_0, T_n) = R(\phi_{t_0}, \phi_{t_0}), \tag{25b}$$

$$\text{the point } u \in L_2 \text{ given by} \quad u(t) = -U^{-1}E^*Lx_t(\phi_{t_0}) \tag{25c}$$

is the optimal control for the functional $J(u, \phi_{t_0}, t_0, \infty)$ and $x_t(\phi_{t_0})$ is the optimal trajectory.

COROLLARY 2. *If Hypothesis* H *holds, then given any continuously differentiable* ϕ_{t_0} *in* $C[-h, 0]$, *the optimal control and optimal trajectory for the functional* $J(u, \phi_{t_0}, t_0, \infty)$ *are connected by the relations*

$$\dot{x}(t) = Ax(t) + Bx(t - h) + D\dot{x}(t - h) - EU^{-1}E^*q(t), \tag{26}$$

$$\dot{q}(t) = Wx(t) - A^*q(t) - B^*q(t + h) + D^*\dot{q}(t + h), \tag{27}$$

and

$$u(t) = -U^{-1}E^*q(t). \tag{28}$$

Proof. Let ϕ_{t_0} be continuously differentiable. Let $\{T_n\} \to \infty$ and $x^n(t) = x^m(t, \phi_{t_0}, t_0, T_n)$ and set $L(t, t, T_n) = L_n(t)$, $n = 1, 2, \cdots$. By Theorem 3, $\{x^n(t)\} \to x(t, \phi_{t_0})$ uniformly on compact intervals, and by Property 7, $\{L_n(t)\phi\} \to L\phi$ for all ϕ in $C[-h, 0]$. Hence if we set

$$q_n(t) = L_n(t)x_t^n \quad \text{and} \quad q(t) = Lx_t(\phi_{t_0}),$$

it follows that for each $t \in [t_0, \infty)$,

$$\lim_{n\to\infty} q_n(t) = q(t).$$

Notice that $x_t(\phi_{t_0})$ satisfies (26) and that (28) holds by Theorem 3. We have shown in the corollary to Theorem 1 that each $q_n(t)$ is differentiable a.e. and satisfies (16). The function $q(t)$ is also differentiable a.e. by Remark 3 since L is a continuous linear mapping. Hence the sequence of differential equations

$$\frac{d}{dt}[q_n(t) - D^*q_n(t + h)] = -Wx^n(t) - A^*q_n(t) - B^*q_n(t + h)$$

converges a.e. to the function

$$-Wx(t, \phi_{t_0}) - A^*q(t) - B^*q(t + h).$$

Using standard arguments from real variables, it follows that

$$\begin{aligned}\frac{d}{dt}[q(t) - D^*q(t + h)] &= \dot{q}(t) - D^*q(t + h)\\ &= -Wx(t, \phi_{t_0}) - A^*q(t) - B^*q(t + h),\end{aligned}$$

which establishes (27) and proves the corollary.

COROLLARY 3. *If the control problem* (1)–(2) *satisfies Hypothesis* H, *then all solutions of the differential-difference equation*

$$\text{(29a)} \qquad \dot{x}(t) = Ax(t) + Bx(t - h) + D\dot{x}(t - h) - EU^{-1}E^*Lx_t,$$

where

$$\text{(29b)} \qquad x(t) = \phi(t) \quad \textit{is continuously differentiable on } C[-h, 0],$$

have the property that

$$\text{(30)} \qquad \int_0^\infty |x(t)|^2\,dt < \infty.$$

Proof. The inequality (30) is a consequence of the fact that

$$j(\phi, 0, \infty) \geqq w_1 \int_0^\infty |x(t)|^2\,dt, \quad \text{where } w_1 > 0.$$

Remark 6. If A, B, D, E, W and U are real matrices, and $C[-h, 0]$ and L_2 are complex Banach spaces and Hypothesis H holds, then the conclusions of Corollary 3 remain valid.

In [2] it was possible to prove for the case $D = 0$ (i.e., the retarded case) that the system (29) generated an exponentially stable semigroup of operators on $C[-h, 0]$. If $D \neq 0$, the situation is unclear. However, we can obtain certain analogous properties for the system of differential-difference equations given by (29). Thus let us assume that Hypothesis H holds and that $C[-h, 0]$ and L_2 are complex Banach spaces. The functional form of (29) is the system

$$(31)\qquad x(t) = \begin{cases} \phi(t_0) - D\phi(t_0 - h) + \displaystyle\int_{t_0}^{t} [Ax(s) + Bx(s-h)]\,ds \\ \qquad - \displaystyle\int_{t_0}^{t} EU^{-1}E^*Lx_s\,ds + Dx(t-h) \quad \text{if } t \geqq t_0, \\ \phi(t) \quad \text{if } t \in [t-h, t_0]. \end{cases}$$

It is known (see, e.g., [3]) that the solutions of (31) in $C[-h, 0]$ generate a semigroup of operators $T(t)$ which is strongly continuous on $C[-h, 0]$. Let $\mathscr{A}$ denote the infinitesimal generator of $T(t)$. Recall that the domain of $\mathscr{A}$ is dense in $C[-h, 0]$ and consists of those ϕ which are continuously differentiable and satisfy the condition

$$(32)\qquad \frac{d\phi}{ds}(0) = A\phi(0) - B\phi(-h) - EU^{-1}E^*L\phi + D\frac{d\phi}{ds}(-h).$$

Also, if ϕ is in the domain of $\mathscr{A}$, then

$$\frac{d}{dt}(T(t)\phi) = \mathscr{A}T(t)\phi = T(t)\mathscr{A}\phi.$$

Hale and Meyer [3] have proven that the spectrum of $\mathscr{A}$, $\sigma(\mathscr{A})$, consists only of point spectra and that the generalized eigenspace associated with any $\lambda \in \sigma(\mathscr{A})$ is always finite-dimensional. Furthermore, since $T(t)$ is strongly continuous on $[0, \infty)$, there exist constants $M \geqq 1$ and $w > 0$ such that $\|T(t)\phi\| \leqq Me^{wt}\|\phi\|$ for all $\phi \in C$ (see, e.g., [4]).

LEMMA 2. *Let the control problem* (1)–(2) *satisfy Hypothesis* H. *Then there exists* $a_0 > 0$ *such that spectrum* $\sigma(\mathscr{A})$ *of the infinitesimal generator of* $T(t)$ *lies in the left half-plane* $\operatorname{Re} z \leqq -a_0$.

Proof. On $C[-h, 0] \times C[-h, 0]$ we define the bilinear Hermitian form

$$(33)\qquad \beta(\phi, \psi) = \int_0^\infty (Wx(t, \phi), x(t, \psi))\,dt.$$

By the second corollary to Theorem 3 and Remark 5, it follows that β is defined on all $C[-h, 0] \times C[-h, 0]$ and that $\beta(\phi, \phi) \geqq 0$ for all ϕ. Since β is Hermitian and defined everywhere, it is evident that β is continuous. Hence we can find $b > 0$ such that

$$(34)\qquad 0 \leqq \beta(\phi, \phi) \leqq b\|\phi\|^2 \quad \text{for all } \phi \in C[-h, 0].$$

Using the semigroup property of $T(t)$ we have the identity

$$(35)\qquad \begin{aligned} \beta(T(t)\phi, T(t)\phi) &= \int_0^\infty (Wx(s, x_t(\phi)), x(s, x_t(\psi))\,ds \\ &= \int_0^\infty (Wx(s+t, \phi), x(s+t, \psi))\,ds \\ &= \int_t^\infty (Wx(s, \phi), x(s, \psi))\,ds. \end{aligned}$$

Hence if ϕ is in the domain of $\mathscr{A}$,

$$\begin{aligned} \frac{d\beta}{dt}(T(t)\phi, T(t,\phi)) &= \beta(\mathscr{A}T(t)\phi, T(t)\phi) + \beta(T(t)\phi, \mathscr{A}T(t)\phi) \\ &= -(Wx(t,\phi), x(t,\phi)) \\ &= -\mathscr{H}(T(t)\phi, T(t)\phi). \end{aligned} \tag{36}$$

Here $\mathscr{H}$ is the continuous Hermitian form defined by

$$\mathscr{H}(\phi, \psi) = (W\phi(0), \psi(0)). \tag{37}$$

Clearly because of the assumptions on W,

$$0 \leqq \mathscr{H}(\phi, \phi) \leqq w_2\|\phi\|^2. \tag{38}$$

Hale and Meyer [3] have shown that the eigenvectors of $\mathscr{A}$ are of the form $\phi(s) = e^{\lambda s}c$, where c is an n-vector and $s \in [-h, 0]$. Let λ be an eigenvalue of $\mathscr{A}$ and $\phi(s) = e^{\lambda s}c$, $c \neq 0$, a corresponding eigenvector. Then by (36),

$$2\,\mathrm{Re}\,\lambda\beta(\phi, \phi) = -(Wc, c) < 0, \tag{39}$$

which means that $\mathrm{Re}\,\lambda < 0$. Hence the spectrum of $\mathscr{A}$ lies in the left half-plane $\mathrm{Re}\,z < 0$.

Notice that if $\mathrm{Re}\,\lambda < 0$, if $\phi(s) = e^{\lambda s}c$ and $s \in [-h, 0]$, then

$$\|\phi\|^2 = |c|^2\, e^{(-2\mathrm{Re}\lambda)h}. \tag{40}$$

Thus using (36), (40) and the bounds on β and W, we can make the following estimate on any eigenvalue and eigenvector of $\mathscr{A}$:

$$\begin{aligned} -2\,\mathrm{Re}\,\lambda\|\phi\|^2 b \geqq -2\,\mathrm{Re}\,\lambda\beta(\phi,\phi) &= (Wc, c) \\ \geqq w_1|c|^2 &= w_1\, e^{2(\mathrm{Re}\lambda)h}\|\phi\|^2. \end{aligned} \tag{41}$$

Hence

$$\frac{b}{w_1} \geqq \frac{e^{(2\mathrm{Re}\lambda)h}}{-2\,\mathrm{Re}\,\lambda}. \tag{42}$$

The inequality (42) establishes the lemma since there must exist a number $2a_0 > 0$ such that

$$b/w_1 \geqq e^{-xh}/x$$

for all $x \geqq 2a_0$. Thus $-2\,\mathrm{Re}\,\lambda \geqq 2a_0$ or $\mathrm{Re}\,\lambda \leqq -a_0$.

THEOREM 4. *Let the control problem* (1)–(2) *satisfy Hypothesis* H. *Let $\phi \in C[-h, 0]$ be continuously differentiable. Then the solution satisfies the estimate*

$$|x(t, \phi)| \leqq M_1\, e^{-a_1 t}\|\phi\|_1, \tag{43}$$

where $a_1 > 0$, $M \geqq 1$ are independent of the particular choice of ϕ.

Proof. By Lemma 2, the spectrum of the semigroup generated by the solution of (31) lies in the left half-plane $\mathrm{Re}\,z \leqq -a_0$. Hence the estimate (43) with $a_1 = a_0/2$ follows directly from the work of Hale and Meyer [3, p. 38, Theorem 1].

A final word needs to be said concerning Hypothesis H. In [2], necessary and sufficient conditions are given for the satisfaction of this hypothesis when the matrix D is zero (i.e., the retarded case). However, these conditions will not suffice in the case of neutral equations, and it would be interesting to find reasonable sufficient conditions for Hypothesis H to be valid. One obvious condition is that the homogeneous system in (2) be uniformly asymptotically stable.

Thus the value $J(0, \phi_{t_0}, t_0, \infty) < \infty$ for all ϕ_{t_0} in $C[-h, 0]$, and Hypothesis H holds.

REFERENCES

[1] R. Bellman and K. L. Cooke, *Differential–Difference Equations*, Academic Press, New York, 1963.

[2] R. Datko, *Unconstrained control problems with quadratic cost*, this Journal, 11 (1973), pp. 32–52.

[3] J. K. Hale and K. R. Meyer, *A class of functional equations of neutral type*, Amer. Math. Soc., No. 76, 1967.

[4] E. Hille and R. S. Phillips, *Functional Analysis and Semi-groups*, Colloquium Publications, vol. XXXI, American Mathematical Society, Providence, 1957.

SIAM J. CONTROL
Vol. 12, No. 1, February 1974

POSITIVITY CONDITIONS AND INSTABILITY CRITERIA FOR FEEDBACK SYSTEMS*

RONALD A. SKOOG†

Abstract. An instability theorem is obtained for feedback systems which is analogous to the "positive operator" theorem for L_2-stability, and from this theorem instability counterparts to many known stability criteria can be obtained. In particular, a counterpart to the circle criterion is given, and results are given for systems with a time-varying gain with restricted rate of variation in the feedback loop. The most significant feature of the result is that it allows one to use "multiplier" techniques to obtain instability criteria.

1. Introduction. In this paper an instability theorem (Theorem 1) analogous to the "positive operator theorem" [11], [14] for stability is obtained for negative feedback systems of the type shown in Fig. 1. Of course, one cannot hope to prove instability by requiring G and H to be positive operators on $L_2(0, \infty)$, since this leads to stability. Rather, for the simplest case of Theorem 1 it is assumed that G is a linear time invariant operator of the form

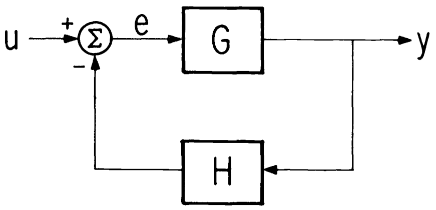

FIG. 1. *A feedback system*

$$(1.1) \qquad (Gx)(t) = g_0x(t) + \int_0^t g(t - \tau)x(\tau)\,d\tau$$

having a Laplace transform $G(s)$ with $\operatorname{Re} G(j\omega) \geqq 0$ for all ω, and in addition $G(s)$ is assumed to have $P \neq 0$ poles in $\operatorname{Re} s > 0$. Then, if H is a positive operator on $L_2(0, \infty)$ it is shown that there exists an input $u \in L_2(0, \infty)$ such that if e satisfies the feedback equation

$$(1.2) \qquad e + HGe = u,$$

then $e \notin L_2(0, \infty)$.

The main idea behind the proof of this result is fairly simple. The proof is by contradiction, and thus one assumes that u and e are elements of $L_2(0, \infty)$, which then implies that $Ge \in L_2(0, \infty)$. The key step is to then show that there is a certain

* Received by the editors June 10, 1971, and in revised form July 29, 1972.

† Bell Laboratories, Holmdel, New Jersey 07733. This research was done while the author was at the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley, and sponsored by the National Science Foundation under Grant GK-5786 and by the Joint Services Electronics Program under Grant AFOSR-68-1488.

bounded noncausal mapping G_{nc} of $L_2(-\infty, \infty)$ into itself with the property that if $Ge \in L_2(0, \infty)$, then $Ge = G_{nc}e$ when e and Ge are viewed as elements of $L_2(-\infty, \infty)$. As a result, it is shown that if $e \in L_2(0, \infty)$ then, when e is viewed as an element of $L_2(-\infty, \infty)$ with $e(t) = 0$ for $t < 0$, it must satisfy the equation

$$e + HG_{nc}e = u. \tag{1.3}$$

The important point here is that (1.3) involves bounded mappings, and is thus easier to deal with than (1.2). In particular, it is shown that there exists a $u \in L_2(-\infty, \infty)$ with $u(t) = 0$ for $t < 0$ such that (1.3) has no solution $e \in L_2(-\infty, \infty)$ with $e(t) = 0$ for $t < 0$. As a result, for this u, (1.2) can have no solution $e \in L_2(0, \infty)$.

The instability theorem described above is applied to the specific case when H is a time varying gain to obtain explicit instability criteria. The first application is to obtain an instability counterpart to the circle criterion. Results of this type were first obtained by Brockett and Lee [3] using Lyapunov methods and by Willems [13] using operator methods. The proof of Willems was simplified and extended somewhat by Bergen and Takeda [1]. The result given here generalizes those of [13] and [1] in that G is allowed to be an unbounded operator. Also, the seemingly superfluous assumption used in [13] and [1] that $(I + HG)$ has a causal inverse on the extended space $L_{2e}(0, \infty)$ is not required in the proofs given here. This advantage has more to do with aesthetics than practicality. Nevertheless, it justifies the remark that stability or instability and causality are two separate issues, and should be treated independently from one another.

The second application of the instability theorem is to obtain an instability counterpart to the stability theorems of Brockett and Forys [2], Gruber and Willems [6], and Freedman [4]. The result given here is in line with that of Freedman in that it does not require the existence of a "multiplier" in the theorem statement. Specifically, the result states that if H is defined by $(Hx)(t) = k(t)x(t)$ with $\alpha < k(t) < \beta$, and if G is a linear time invariant system such that the closed loop system has the same number (not zero) of poles in $\operatorname{Re} s > 0$ for all constant gains between α and β, and either

$$\frac{\dot{k}(t)}{k(t) - \alpha} \leqq \sigma\left(1 - \frac{k(t) - \alpha}{\beta - \alpha}\right) \tag{1.4}$$

or

$$\frac{\dot{k}(t)}{k(t) - \alpha} \geqq \sigma\left(1 - \frac{k(t) - \alpha}{\beta - \alpha}\right), \tag{1.5}$$

where $0 \leqq \sigma \leqq \sigma_*$, σ_* depending on G, then the time-varying feedback system is unstable. Thus, if $\dot{k}$ is sufficiently small, stability or instability can be predicted on the basis of the "frozen-time" systems, so long as the number of poles in $\operatorname{Re} s > 0$ does not change. It has been shown by an example [12] that if the number of poles in $\operatorname{Re} s > 0$ does change, then no matter how slowly varying the system is, the "frozen-time" systems can be unstable and the time-varying system stable.

Although the main instability theorem given here (Theorem 1) can be used to obtain results such as the circle criterion for instability, its main utility lies in proving results as in the second application described above. In these cases instability criteria are obtained by the introduction of certain "multipliers" to create "positive" operators. It is in this direction that the previous methods of proving instability are not applicable.

The reader is assumed to be familiar with L_p-spaces and the notion of extended spaces (in particular the extended $L_2(0, \infty)$-space $L_{2e}(0, \infty)$). The reader is referred to the works of Sandberg [11] and Zames [14] for these details. We recall here two notions. The first is the truncation operator P_T. For $x(\cdot)$ a real-valued function and $T \in R$, $(P_Tx)(t) = x(t)$ for $t \leqq T$ and $(P_Tx)(t) = 0$ for $t > T$. The second notion is causality. A mapping H is called *causal* if and only if $P_THP_T = P_TH$ for all $T \in R$.

2. The main results. The feedback system to be considered is that shown in Fig. 2. The system G is a linear time-invariant system which will be restricted to a certain class $\mathscr{G}$ defined as follows.

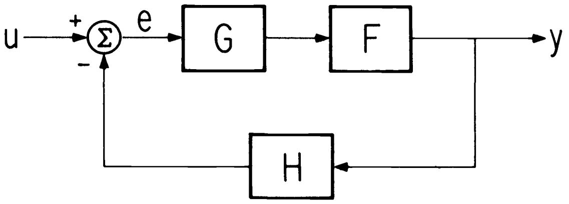

FIG. 2. *Feedback system of Theorem* 1

DEFINITION 1. The *class* $\mathscr{G}$ *of operators mapping* $L_{2e}(0, \infty)$ *into itself* are those which can be represented by

$$(2.1) \qquad (Gf)(t) = g_0 f(t) + \int_0^t g(t - \tau) f(\tau)\, d\tau,$$

where $g(t) = 0$ for $t < 0$ and has the form

$$(2.2) \qquad g(t) = g_1(t) + g_2(t),$$

with $g_1 \in L_1(0, \infty)$, $g_2 e^{-\sigma t} \in L_1(0, \infty)$ for all $\sigma \geqq \sigma_0 > 0$, and where the Laplace transform $G_2(s)$ of g_2 given by

$$(2.3) \qquad G_2(s) = \int_0^\infty g_2(t) e^{-st}\, dt, \qquad \operatorname{Re} s \geqq \sigma_0,$$

is a rational function with a finite number of singularities and no singularities in $\operatorname{Re} s \leqq 0$. (Note that $G_2(s)$ is defined by (2.3) only for $\operatorname{Re} s \geqq \sigma_0$, but since it is a rational function, it is clear that it has a meromorphic continuation to the entire complex plane).

From the conditions placed on $G_2(s)$ it is clear that $g_2(\cdot)$ has the form

$$(2.4) \qquad g_2(t) = \sum_{i=1}^{n} \sum_{j=1}^{l_i} \alpha_{i,j} t^{j-1} e^{\sigma_i t}, \qquad t \geqq 0,$$

where $\alpha_{i,j}$ and σ_i are complex numbers with $\operatorname{Re} \sigma_i > 0$ for all i. Thus it is possible to make a continuation of $g_2(t)$ to negative values of t to obtain $\tilde{g}_2(t)$ defined by

$$\tilde{g}_2(t) = \begin{cases} \sum_{i=1}^{n} \sum_{j=1}^{l_i} \alpha_{i,j} t^{j-1} e^{\sigma_i t}, & t \leqq 0, \\ 0, & t > 0. \end{cases} \tag{2.5}$$

It is observed that $\tilde{g}_2 \in L_1(-\infty, \infty)$.

With this construction of $\tilde{g}_2$ it is now possible to associate with each $G \in \mathscr{G}$ a noncausal mapping G_{nc} of $L_{2e}(-\infty, \infty)$ into itself as follows.

DEFINITION 2. For each $G \in \mathscr{G}$ define G_{nc} *on* $L_{2e}(-\infty, \infty)$ by

$$(G_{nc}f)(t) = g_0 f(t) + \int_{-\infty}^{t} g_1(t-\tau) f(\tau)\, d\tau - \int_{t}^{\infty} \tilde{g}_2(t-\tau) f(\tau)\, d\tau. \tag{2.6}$$

From the fact that $g_1 \in L_1(0, \infty)$ and $\tilde{g}_2 \in L_1(-\infty, 0)$ it follows easily that G_{nc} maps $L_2(-\infty, \infty)$ into itself. The Laplace transform $G_{nc}(s)$ of G_{nc} defined by

$$G_{nc}(s) = g_0 + \int_0^{\infty} g_1(t) e^{-st}\, dt - \int_{-\infty}^{0} \tilde{g}_2(t) e^{-st}\, dt \tag{2.7}$$

exists for $-\sigma_1 < \operatorname{Re} s < \sigma_0$ for some $\sigma_1 \geqq 0$ and is easily seen to be

$$G_{nc}(s) = g_0 + \hat{G}(s), \tag{2.8}$$

where $\hat{G}(s)$ is the meromorphic continuation of

$$\hat{G}(s) = \int_0^{\infty} g(t) e^{-st}\, dt, \qquad \operatorname{Re} s > \sigma_0. \tag{2.9}$$

Thus it is seen that G and G_{nc} have the same Laplace transform, but with different regions of convergence.

Our first theorem is the key result of this paper, and it plays a predominant role in the proofs of the remaining theorems giving explicit instability criteria. This result is analogous to the positive operator theorem for L_2-stability obtained by Sandberg and Zames [11], [14].

THEOREM 1. *Let F, G, and H satisfy the following assumptions*:

(a) $G \in \mathscr{G}$ *and* $\operatorname{Re} G(j\omega) \geqq 0$ *for all* ω.

(b) *F is a causal mapping of $L_2(-\infty, \infty)$ into itself, $F\theta = \theta$, $\langle Fx_1 - Fx_2, x_1 - x_2\rangle \geqq k_1\|x_1 - x_2\|^2$ and $\|Fx_1 - Fx_2\| \leqq k_2\|x_1 - x_2\|$ for all $x_1, x_2 \in L_2(-\infty, \infty)$ and some $k_1, k_2 > 0$.*

(c) *$\langle FG_{nc}x_1 - FG_{nc}x_2, x_1 - x_2\rangle \geqq 0$ for all x_1, $x_2 \in L_2(-\infty, \infty)$, where G_{nc} is as defined in Definition 2.*

(d) *H is a causal mapping of $L_2(-\infty, \infty)$ into itself, $H\theta = \theta$, $\langle Hx_1 - Hx_2, x_1 - x_2\rangle \geqq \alpha\|x_1 - x_2\|^2$, and $\|Hx_1 - Hx_2\| \leqq \beta\|x_1 - x_2\|$ for all x_1, $x_2 \in L_2(-\infty, \infty)$ and some $\alpha, \beta > 0$.*

Then if $G \neq G_{nc}$ (i.e., if $G(s)$ has a singularity in $\operatorname{Re} s > 0$), *there exists some $u \in L_2(0, \infty)$ such that if $e \in L_{2e}(0, \infty)$ and satisfies*

$$(I + HFG)e = u, \tag{2.10}$$

then $e \notin L_2(0, \infty)$.

Note that in the theorem statements, F and H were assumed to be causal mappings of $L_2(-\infty, \infty)$ into itself. For most systems which are of interest, causal mappings on $L_2(0, \infty)$ can be extended to causal mappings on $L_2(-\infty, \infty)$ in a natural way (c.f. [13]). Thus, there will generally be no difficulty in meeting these conditions. Also note that (2.10) is viewed as an equation on $L_{2e}(0, \infty)$. This is well-defined since a causal mapping of $L_{2e}(-\infty, \infty)$ into itself is also a causal mapping of $L_{2e}(0, \infty)$ into itself.

The next result gives an explicit instability criteria for the system of Fig. 3, and is an instability counterpart of the circle criterion. Results of this type were first obtained by Brockett and Lee [3] using Lyapunov methods and by Willems [13] using operator methods. The result given below (Theorem 2) generalizes those in [13] and [1] in that there G was restricted to being a bounded mapping of L_2 into itself, while here G is allowed to be an unbounded operator on $L_2(0, \infty)$.

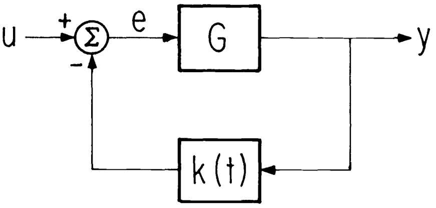

FIG. 3. *Feedback system with time-varying gain*

For the feedback system of Fig. 3, let $k(\cdot)$ be a real-valued function of t, and denote by K the mapping $(Kx)(t) = k(t)x(t)$, where $x(\cdot)$ is any real-valued function of t. Then the feedback equations for Fig. 3 are

$$e + KGe = u. \tag{2.11}$$

THEOREM 2. *For the system of Fig. 3, let $G \in \mathscr{G}$, $G(s)$ have P poles in* Re $s > 0$, *and* $\alpha + \varepsilon \leqq k(t) \leqq \beta - \varepsilon$ *for some* $\varepsilon > 0$ *and all* t.

Case I ($\alpha > 0, \beta > 0$). *If $G(j\omega)$ does not intersect the closed disk* $D[\alpha, \beta]$ (see Fig. 4a) *centered at* $(-(1/2)(1/\alpha + 1/\beta), 0)$ *with radius* $(1/2)(1/\alpha - 1/\beta)$, *and makes $N < \infty$ clockwise encirclements of it as ω goes from $-\infty$ to ∞, and $N \neq -P$, then there exists an input $u \in L_2(0, \infty)$ such that if e is a solution of* (2.11), *then $e \notin L_2(0, \infty)$ and $Ge \notin L_2(0, \infty)$.*

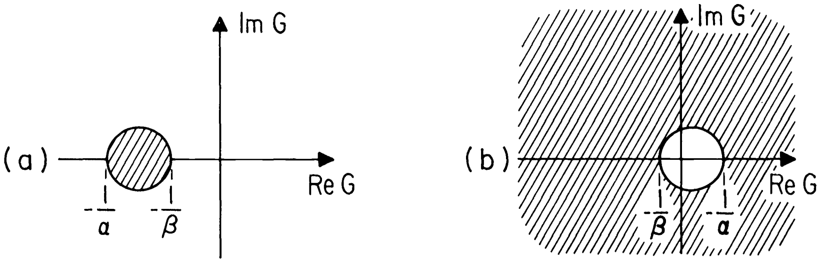

FIG. 4. *The disks $D[\alpha, \beta]$ of Theorem 2*

Case II ($\alpha < 0, \beta > 0$). *If* $G(j\omega)$ *lies inside the open disk* $D(\alpha, \beta)$ (see Fig. 4b) *and* $P \cdot \neq 0$, *then there exists an input* $u \in L_2(0, \infty)$ *such that if* e *is a solution of* (2.11), *then* $Ge \notin L_2(0, \infty)$.

Note that in Case II it is not possible to conclude that $e \notin L_2(0, \infty)$ but only that the output $y = Ge$ is not in $L_2(0, \infty)$. The reason for this lies in the fact that here $k(t)$ is allowed to be equal to zero. As a result, for $k(t) = 0$ it is seen from (2.11) that $e = u$ and is therefore in $L_2(0, \infty)$. Also, it is possible for $k(t)$ to approach zero as $t \to \infty$ at a sufficient rate that $KGe \in L_2(0, \infty)$ even though $Ge \notin L_2(0, \infty)$. For this situation as well, it is seen from (2.11) that $e \in L_2(0, \infty)$.

One further point is worth noting. The result of Case II has not been obtained in any of the previous works on instability [3], [13], [1]. This result cannot be obtained by the methods used by Willems [13] or Bergen and Takeda [1] since their methods of proof depend heavily on the condition that G be bounded on $L_2(0, \infty)$ (i.e., $G(s)$ can have no singularities in $\operatorname{Re} s > 0$). Also, as mentioned previously, the result of Case I generalizes those in [13] and [1] in that G is allowed to be unbounded. Thus, Theorem 2 applies to systems having an unstable open loop, whereas the results in [13] and [1] do not.

The next result (Theorem 3) gives an instability counterpart to stability theorems of the type given by Brockett and Forys [2], Gruber and Willems [6], and Freedman [4] with regard to systems having slowly varying feedback gains. Before giving the result some notation will be needed.

Let $\Phi(\cdot)$ be a differentiable mapping of the real line into the interval $(-\pi, \pi)$ with $\lim_{|\omega| \to \infty} |\Phi(\omega)| = 0$. Then define γ by

$$\gamma \triangleq \pi - \max_{\omega} |\Phi(\omega)|, \tag{2.12}$$

and define Ω by

$$\Omega \triangleq \min \{ w \| \Phi(\omega) | \leqq .(\pi - \gamma)/3 \text{ for all } |\omega| \geqq w \}. \tag{2.13}$$

THEOREM 3. *For the system of Fig.* 3, *let* $G \in \mathscr{G}$, $G(s)$ *have* P *poles in* $\operatorname{Re} s > 0$, $\alpha + \varepsilon \leqq k(t) \leqq \beta - \varepsilon$ *for some* $\varepsilon > 0$ *and all* t, *and let* $k(\cdot)$ *be absolutely continuous.*

Case I ($\alpha > 0, \beta > 0$). *Assume* $G(j\omega)$ *does not intersect the interval* $[-1/\alpha, -1/\beta]$ *on the real axis, and encircles it* N *times in the clockwise direction with* $N \neq -P$.

Case II ($\alpha < 0, \beta > 0$). *Assume* $G(j\omega)$ *does not intersect either the interval* $(-\infty, -1/\beta]$ *or* $[-1/\alpha, \infty)$ *and* $P \neq 0$.

Then, with $\Phi(\omega) = \arg [1 + \beta G(j\omega)][1 + \alpha G(j\omega)]^{-1}$, γ *and* Ω *given by* (2.12) *and* (2.13), *and*

$$\sigma_* = \left(\frac{3\pi\gamma}{16} \right)^2 \Big/ \int_{-\Omega}^{\Omega} \left| \frac{d}{d\omega} \Phi(\omega) \right| d\omega, \tag{2.14}$$

if there exists a $\sigma \in (0, \sigma_*)$ *such that either*

$$\text{(i)} \quad \dot{k}(t)/(k(t) - \alpha) \leqq 2\sigma(1 - (k(t) - \alpha)/(\beta - \alpha))$$

or

$$\text{(ii)} \quad \dot{k}(t)/(k(t) - \alpha) \geqq -2\sigma(1 - (k(t) - \alpha)/(\beta - \alpha))$$

holds for all $t \geqq t_0 > 0$, *there exists a* $u \in L_2(0, \infty)$ *such that if* $e \in L_{2e}(0, \infty)$ *is a* *solution of* (2.11), *then* $e \notin L_2(0, \infty)$ *and* $Ge \notin L_2(0, \infty)$ *for Case* I, *and* $Ge \notin L_2(0, \infty)$ *for Case* II.

It should be noted from the proof of Theorem 3 (in particular (3.17)) that one does not need to require $k(\cdot)$ to be absolutely continuous for Theorem 3 to hold. One can replace conditions (i) and (ii) above by

(i') $(\beta - k(t)/(k(t) - \alpha)\, e^{-\sigma t}$ is monotone nonincreasing,

(ii') $(\beta - k(t)/(k(t) - \alpha)\, e^{\alpha t}$ is monotone nondecreasing,

and the conclusions still·hold. Thus in the case (i'), decreasing discontinuous jumps are allowed, and in case (ii'), increasing discontinuous jumps are allowed.

3. Proofs. As discussed in the Introduction, the essential idea behind the proof of Theorem 1 is the replacement of the unbounded operator G by the bounded noncausal operator G_{nc} defined in Definition 2. The next lemma is the key result concerning G_{nc}.

LEMMA 1. *Let* $G \in \mathscr{G}$ *and* $e \in L_2(0, \infty)$. *If* $Ge \in L_2(0, \infty)$, *then* $Ge = G_{nc}e$, *where* e *and* Ge *are viewed as elements of* $L_2(-\infty, \infty)$.

The proof of Lemma 1 is given in the Appendix. Before a proof of Theorem 1 can be given, one more lemma is needed.

LEMMA 2. *Let* F, G, *and* H *satisfy the assumptions of Theorem* 1. *Then* $(I + HFG_{nc})$ *has a noncausal inverse on* $L_2(-\infty, \infty)$, *and* $P_T(I + HFG_{nc})^{-1} \cdot (I - P_T) \neq 0$ *for all* $T \in R$.

The proof of Lemma 2 is given in the Appendix. It should be noted here that if F and H are linear operators, the noncausality of $(I + HFG_{nc})^{-1}$ implies that $P_T(I + HFG_{nc})^{-1}(I - P_T) \neq 0$. However, for nonlinear operators this is no longer the case.

We are now in a position to give the following proof.

Proof of Theorem 1. If there exists an $e \in L_2(0, \infty)$ satisfying (2.10) for a given $u \in L_2(0, \infty)$, then $Ge \in L_2(0, \infty)$. Indeed, from the causality and positivity conditions on F and H and the Schwarz inequality, it follows that for any $x \in L_{2e}(0, \infty)$,

$$\|P_T F x\| \geqq k_1 \|P_T x\| \tag{3.1}$$

and

$$\|P_T H x\| \geqq \alpha \|P_T x\|. \tag{3.2}$$

Thus if $x \notin L_2(0, \infty)$, $\lim_{T\to\infty} \|P_T F x\|$ and $\lim_{T\to\infty} \|P_T H x\|$ are unbounded and thus $Fx \notin L_2(0, \infty)$ and $Hx \notin L_2(0, \infty)$. Hence if $Ge \notin L_2(0, \infty)$, then $HFGe \notin L_2(0, \infty)$, and $e \in L_2(0, \infty)$ could not satisfy (2.10) when $u \in L_2(0, \infty)$. Therefore, it is seen that $Ge \in L_2(0, \infty)$, and from Lemma 1 it follows that $Ge = G_{nc}e$. As a result, e viewed as an element of $L_2(-\infty, \infty)$ with $e(t) = 0$ for $t < 0$ must be a solution of

$$(I + HFG_{nc})e = u, \tag{3.3}$$

where u is viewed as an element of $L_2(-\infty, \infty)$ with $u(t) = 0$ for $t < 0$. It will now be shown that there exists a $\hat{u} \in L_2(-\infty, \infty)$ with $\hat{u}(t) = 0$ for $t < 0$ such that if $e \in L_2(-\infty, \infty)$ and satisfies (3.3) with $u = \hat{u}$, then $P_0 e \neq 0$. Thus, for $u = \hat{u}$ in (2.10), there can be no solution $e \in L_2(0, \infty)$.

From Lemma 2 it is seen that $(I + HFG_{nc})$ is invertible on $L_2(-\infty, \infty)$ and $P_T(I + HFG_{nc})^{-1}(I - P_T) \neq 0$ for all $T \in R$. Thus, there exists a $\hat{u} \in L_2(-\infty, \infty)$ with $P_0\hat{u} = 0$, (i.e., $\hat{u}(t) = 0$ for $t < 0$) such that $P_0\hat{e} \neq 0$, where $\hat{e} = (I + HFG_{nc})^{-1}\hat{u}$. Furthermore, $e = \hat{e}$ is the only solution of (3.3) in $L_2(-\infty, \infty)$ with $u = \hat{u}$. Therefore, there exists a $u \in L_2(-\infty, \infty)$ with $P_0 u = 0$ for which there is no solution $e \in L_2(-\infty, \infty)$ with $P_0 e = 0$. Hence it follows that for this u there is no solution $e \in L_2(0, \infty)$ of (2.10), and thus if $e \in L_{2e}(0, \infty)$ satisfies (2.10), then $e \notin L_2(0, \infty)$. Q.E.D.

The next result to be proved is Theorem 2. The proof simply involves a change of variables which puts the feedback equations in a form to which Theorem 1 applies.

Proof of Theorem 2. First add and subtract αG from (2.11) to obtain

$$(I + \alpha G)e + (K - \alpha)Ge = u. \tag{3.4}$$

Now, since $G \in \mathscr{G}$ it follows (cf. [8, pp. 141–150]) that for some $\sigma_0 > 0$ there exists a function h where $h\,e^{-\sigma_0 t} \in L_1(0, \infty)$ and a constant h_0 such that $(I + \alpha G)$ has a causal inverse on the weighted L_2-space

$$L_2(0, \infty\,; e^{-\sigma_0 t}) = \{f | e^{-\sigma_0 t} f \in L_2(0, \infty)\},$$

and this inverse is given by

$$[(I + \alpha G)^{-1}x](t) = h_0 x(t) + \int_0^t h(t - \tau)x(\tau)\,d\tau. \tag{3.5}$$

Clearly, this inverse of $(I + \alpha G)$ can be extended to $L_{2e}(0, \infty)$. From the encirclement conditions on $G(j\omega)$ it follows that $\inf_{-\infty<\omega<\infty}|1 + \alpha G(j\omega)| \neq 0$ and that $(I + \alpha G(s))^{-1}$ has $P + N$ poles in $\operatorname{Re} s > 0$ for Case I and P poles in $\operatorname{Re} s > 0$ for Case II. Thus, $(I + \alpha G)^{-1} \in \mathscr{G}$.

Let $\tilde{e} = (I + \alpha G)e$, so that (3.4) becomes

$$\tilde{e} + (K - \alpha)G(I + \alpha G)^{-1}\tilde{e} = u. \tag{3.6}$$

Next, add and subtract $(1/(\beta - \alpha))[K - \alpha]$ from (3.6) to obtain

$$[\beta I - K]\frac{1}{\beta - \alpha}\tilde{e} + [K - \alpha][1 + \beta G][I + \alpha G]^{-1}\frac{1}{\beta - \alpha}\tilde{e} = u. \tag{3.7}$$

From the conditions on K, it is seen that $[\beta I - K]$ has a causal inverse on $L_2(0, \infty)$, and hence also on $L_{2e}(0, \infty)$, given by

$$([\beta I - K]x)(t) = \frac{1}{\beta - k(t)}x(t). \tag{3.8}$$

Thus, applying $[\beta I - K]^{-1}$ to both sides of (3.7) yields

$$e^* + K^*G^*e^* = u^*, \tag{3.9}$$

where

$$e^* = (1/(\beta - \alpha))\tilde{e}, \tag{3.10a}$$

$$u^* = (\beta I - K)^{-1}u, \tag{3.10b}$$

$$K^* = (\beta I - K)^{-1}(K - \alpha), \tag{3.10c}$$

$$G^* = (I + \beta G)(I + \alpha G)^{-1}. \tag{3.10d}$$

Equation (3.9) satisfies the conditions of Theorem 1; for the bilinear mapping $\eta = (1 + \beta z)/(1 + \alpha z)$ maps the disk $D[\alpha, \beta]$ onto the half-plane $\operatorname{Re} \eta < 0$ for $\alpha > 0$ and onto the half-plane $\operatorname{Re} \eta \geqq 0$ for $\alpha < 0$, so $\operatorname{Re} G^*(j\omega) \geqq 0$ for all ω. Also, K^* is given by $(K^*x)(t) = k^*(t)x(t)$, where $k^*(t) = (k(t) - \alpha)/(\beta - k(t))$. Thus, there exists some $\varepsilon > 0$ such that $\varepsilon < k^*(t) < 1/\varepsilon$. Finally, note that $G^* = (1 - \beta/\alpha) \cdot (I + \alpha G)^{-1} + \beta/\alpha$, so $G^*(s)$ has the same number of poles in $\operatorname{Re} s > 0$ as does $(I + \alpha G(s))^{-1}$.

From Theorem 1 it then follows that there exists a $u^* \in L_2(0, \infty)$ such that if e^* is a solution of (3.9), then $e^* \notin L_2(0, \infty)$. This in turn implies that there is a $u \in L_2(0, \infty)$ such that $\tilde{e} \in L_2(0, \infty)$. For Case I it is easily seen from (2.11) and the positivity of K that $e \in L_2(0, \infty)$ if and only if $Ge \in L_2(0, \infty)$. Since $\tilde{e} = (I + \alpha G)e$ and $\tilde{e} \notin L_2(0, \infty)$, it then follows that $e \notin L_2(0, \infty)$ and $Ge \notin L_2(0, \infty)$. For Case II it follows from (2.11) that $e \in L_2(0, \infty)$ if and only if $KGe \in L_2(0, \infty)$. If $e \in L_2(0, \infty)$, then from $\tilde{e} = (I + \alpha G)e$ it follows that $Ge \notin L_2(0, \infty)$. If $e \notin L_2(0, \infty)$, then $KGe \notin L_2(0, \infty)$ and thus from the boundedness of K one has $Ge \notin L_2(0, \infty)$. Q.E.D.

It remains to prove Theorem 3. Before giving the proof of Theorem 3 the following lemma due to Freedman [4] will be needed.

LEMMA 3 [4]. *Let* $h \in L_1(-\infty, \infty)$, $h_0 \in R$, *and* $H(j\omega) = \int_{-\infty}^{\infty} h(t)\, e^{-j\omega t}\, dt$. *Further let* $\Phi(\omega) = \arg[H(j\omega) + h_0]$ *and assume* $|\Phi(\omega)| < \pi$ *for all* ω. *Then with* γ, Ω, *and* σ_* *given by* (2.12) *to* (2.14), *for any* $\sigma \in [0, \sigma_*)$ *there is a* $y(\cdot)$ *in* $L_1(0, \infty)$ *such that with*

$$Z(s) = 1 + \int_0^\infty y(t)\, e^{-\sigma t}\, e^{-st}\, dt, \tag{3.11}$$

the following hold:

(i) $\operatorname{Re}\{Z(j\omega - \sigma)\} \geqq \delta > 0$ *for all* ω,

(ii) $\operatorname{Re}\{Z(j\omega)[H(j\omega) + h_0]\} \geqq \delta > 0$ *for all* ω.

It is remarked that in [4], the above result was stated for $h \in L_1(0, \infty)$; however, from the proof it is clear that it holds as well for $h \in L_1(-\infty, \infty)$.

Proof of Theorem 3. Make the same transformation of variables as in the proof of Theorem 2 to obtain (3.9). Since the bilinear map $\eta = (1 + \beta z)/(1 + \alpha z)$ maps the line segment $\operatorname{Re} z \in [-1/\beta, -1/\alpha]$ onto $\operatorname{Im} \eta = 0$, $\operatorname{Re} \lambda \geqq 0$ when $\alpha < 0$, it follows from the assumptions in Case I and Case II that $|\arg G^*(j\omega)| > \pi$ for all ω. Also, since $G^* \in \mathscr{G}$ there is associated with G^* a G^*_{nc} as defined by Definition 2. Thus, there is an h_0 and $h \in L_1(-\infty, \infty)$ such that $G^*(j\omega) = H(j\omega) + h_0$, and hence by Lemma 3 there exists a causal mapping Z of $L_2(0, \infty)$ into itself given by

$$(Zx)(t) = x(t) + \int_0^t y(t - \tau)\, e^{-\sigma(t-\tau)} x(\tau)\, d\tau, \tag{3.12}$$

where $y \in L_1(0, \infty)$, such that $\operatorname{Re}\{Z(j\omega)G^*(j\omega)\} \geqq \delta > 0$ and $\operatorname{Re}\{Z(j\omega - \sigma)\} \geqq \delta > 0$.

Since $y \in L_1(0, \infty)$, it is seen that $Z(s)$ is analytic for $\operatorname{Re} s \geqq -\sigma$, and thus since $\operatorname{Re} Z(j\omega - \sigma) \geqq \delta > 0$, it follows from the maximum modulus theorem [7] that $\inf_{\operatorname{Re} s \geqq -\sigma} |Z(s)| \neq 0$ and hence that Z has a causal inverse on $L_2(0, \infty)$ of the form

$$(Z^{-1}x)(t) = q_0 x(t) + \int_0^t q(t - \tau)x(\tau)\, d\tau, \tag{3.13}$$

with $q\, e^{\sigma t} \in L_1(0, \infty)$ (cf. [8]). Also, since $Z(j\omega - \sigma)$ is bounded and $\operatorname{Re} Z(j\omega - \sigma) > 0$, it follows that $Z^{-1}(s)$ is analytic for $\operatorname{Re} s \geqq -\sigma$ and $\operatorname{Re}\{Z^{-1}(j\omega - \sigma)\} \geqq \beta > 0$ for all ω.

Now, rewrite (3.9) as follows:

$$e^* + (K^*Z^{-1})(ZG^*)e^* = u^*. \tag{3.14}$$

It is clear that $ZG^* \in \mathscr{G}$, and from the above it is seen that $\operatorname{Re}\{Z(j\omega)G^*(j\omega)\} \geqq 0$ for all ω, and hence ZG^* satisfies the conditions for G of Theorem 1.

It will now be shown that if (i) holds, then K^*Z^{-1} is a positive operator on $L_2(-\infty, \infty)$. First of all, K^* is extended to a mapping of $L_2(-\infty, \infty)$ into itself by defining $k^*(\cdot)$ to be (note that this does not effect the solution of (3.9) for $t \geqq 0$)

$$k^*(t) = \begin{cases} \dfrac{\beta - k(t)}{k(t) - \alpha}, & t \geqq 0, \\[2ex] \dfrac{\beta - k(0)}{k(0) - \alpha} e^{2\sigma t}, & t < 0. \end{cases} \tag{3.15}$$

Since Z^{-1} is defined by a convolution, it has a natural extension to $L_2(-\infty, \infty)$.

Then for any $x \in L_2(-\infty, \infty)$,

$$\begin{aligned} \langle K^*Z^{-1}x, x\rangle &= \int_{-\infty}^{\infty} k^*(t)x(t)(Z^{-1}x)(t)\, dt \\ &= \lim_{T\to\infty} \int_{-\infty}^{T} k^*(t)\, e^{-2\sigma t}x(t)\, e^{2\sigma t}(Z^{-1}x)(t)\, dt. \end{aligned} \tag{3.16}$$

Now condition (i) implies that $k^*(t)\, e^{-2\sigma t}$ is monotone nonincreasing for $t > 0$, and from (3.15) it is seen that $k^*(t)\, e^{-2\sigma t}$ is constant for $t < 0$. Integrating (3.16) by parts gives

$$\begin{aligned} \langle k^*Z^{-1}x, x\rangle = \lim_{T\to\infty} \Bigg\{ & k^*(T)\, e^{-2\sigma T} \int_{-\infty}^{T} x(t)\, e^{2\sigma t}(Z^{-1}x)(t)\, dt \\ & - \int_{-\infty}^{T} \left[\int_{-\infty}^{\tau} x(t)\, e^{2\sigma t}(Z^{-1}x)(t)\, dt \right] d[k^*(\tau)\, e^{-2\sigma\tau}] \Bigg\}. \end{aligned} \tag{3.17}$$

Now, since Z^{-1} is causal and $\operatorname{Re}\{Z^{-1}(j\omega - \sigma)\} \geqq \beta > 0$, it follows from Parseval's theorem that

$$\begin{aligned} \int_{-\infty}^{T} x(t)\, e^{2\sigma t}(Z^{-1}x)(t)\, dt &= \frac{1}{2\pi} \int_{-\infty}^{\infty} Z^{-1}(j\omega - \sigma)|\hat{X}_T(j\omega - \sigma)|^2\, d\omega \\ &\geqq \beta \|P_T x\|^2, \end{aligned} \tag{3.18}$$

where $\hat{X}_T(s)$ denotes the Laplace transform of $P_T x$. Thus using (3.18), the fact that $k^*(t) \geqq \varepsilon > 0$, and $k^*(t)\,e^{-2\sigma t}$ is nonincreasing with (3.17), it is seen that

$$\langle K^*Z^{-1}x, x\rangle \geqq \xi\|x\| \tag{3.19}$$

for some $\xi > 0$.

Finally, taking $F = I$, $G = Z^{-1}G^*$, and $H = K^*Z^{-1}$ in Theorem 1, it follows that there exists a $u^* \in L_2(0, \infty)$ such that if e^* satisfies (3.14), then $e^* \notin L_2(0, \infty)$, and this in turn implies the existence of a $u \in L_2(0, \infty)$ satisfying the conclusion of the theorem.

The proof of the theorem when (ii) holds is similar to the above except that in lieu of (3.14), one applies Z^{-1} to both sides of (3.9) to obtain

$$Z^{-1}e^* + (Z^{-1}K^*)(G^*Z)Z^{-1}e^* = Z^{-1}u^*. \tag{3.20}$$

Condition (ii) implies that $k^*(t)\,e^{2\sigma t}$ is monotone nondecreasing, so $(1/k^*(t))\,e^{-2\sigma t}$ is monotone nonincreasing. Using this fact and the properties of Z^{-1}, it is shown as above that $Z^{-1}K^*$ is positive. The remaining details are omitted. Q.E.D.

4. Concluding remarks. An instability theorem (Theorem 1) analogous to the positive operator theorem for stability has been obtained, and this result has been used to prove the instability counterparts of the circle criterion (Theorem 2) and a criterion with restricted rate of gain variation (Theorem 3). The method employed here for proving instability allows one to deal with a much wider class of systems than did previous methods, in that unbounded operators can be handled, and also multiplier techniques can be employed in obtaining explicit instability criteria.

Although a majority of the stability criteria proved using multiplier techniques can now, through the use of Theorem 1, be given instability counterparts, it is not true in general. In particular, the results of Freedman and Zames [5] do not carry over. The reason for this is that although when G is a positive causal operator and $k(\cdot)$ is a monotone decreasing gain the operator KG is positive (where $(Kx)(t) = k(t)x(t)$), this is no longer the case if G is noncausal. As a result, the factorizations used in [5] will not yield a positive operator when G_{nc} is used in place of G. At this point, it is a perplexing question as to whether this is a result of the method of analysis or if the instability counterpart to the result in [5] is simply not true. Most would probably conjecture that it is the method of analysis which is at fault.

Another deficiency in the methods used here is that the instability circle criterion for nonlinearities in the feedback path cannot be handled satisfactorily. One can easily extend Theorem 2 to cover the case of a time-varying nonlinearity $f(\cdot, t)$ satisfying $\alpha|x - y| < |f(x, t) - f(y, t)| < \beta|x - y|$ but not to the case $\alpha < f(x, t)/x < \beta$, which has been handled successfully in [1] and [3]. Also, counterparts to the results of O'Shea [10] and Zames and Falb [15] cannot be obtained. The reason for this deficiency lies in the fact that the proof of Theorem 1 relies on the inverse of $I + HG_{nc}$ existing on $L_2(-\infty, \infty)$, and therefore incremental bounds on nonlinearities are needed. For finite-dimensional systems, this presents no problem since one can linearize and prove instability locally.

To end on a more positive note, it is pointed out that one need not assume anything (such as having a causal inverse) about the behavior of the operator

$(I + HG)$ on the extended space $L_{2e}(0, \infty)$. In fact one does not even need to deal with $L_{2e}(0, \infty)$, but can make the domain of the operator as large as is desired (e.g., measurable functions). For recalling the proof of Theorem 1, it was shown that if e lies in the domain of $(I + HG)$ and satisfies $(I + HG)e = u$ for a certain $u \in L_2(0, \infty)$, then $e \notin L_2(0, \infty)$.

Appendix.

Proof of Lemma 1. From (2.1), (2.2), and (2.6) it is seen that, with $e(t) = 0$ for $t < 0$, $Ge = G_{nc}e$ if and only if

$$\text{(A.1a)} \qquad \int_0^t g_2(t - \tau)e(\tau)\,d\tau = -\int_t^\infty \tilde{g}_2(t - \tau)e(\tau)\,d\tau, \qquad t \geqq 0,$$

$$\text{(A.1b)} \qquad 0 = \int_0^\infty \tilde{g}_2(t - \tau)e(\tau)\,d\tau, \qquad t < 0.$$

This is equivalent to the condition

$$\text{(A.2)} \qquad \int_0^\infty \hat{g}_2(t - \tau)e(\tau)\,d\tau = 0, \qquad -\infty < t < \infty,$$

where $\hat{g}_2(t) = g_2(t) + \tilde{g}_2(t)$ (note: $g_2(t) = 0$ for $t < 0$ and $\tilde{g}_2(t) = 0$ for $t > 0$).

Let $y = Ge$, so by assumption $y \in L_2(0, \infty)$. The Laplace transform of y is

$$\text{(A.3)} \qquad Y(s) = G(s)E(s),$$

where $E(s)$ is the Laplace transform of e. The region of convergence for $G(s)$ is $\operatorname{Re} s > \sigma_0$, and for $E(s)$ it is $\operatorname{Re} s > 0$. Thus $Y(s)$ is well-defined for $\operatorname{Re} s > \sigma_0$. However, since $y \in L_2(0, \infty)$ it follows that $Y(s)$ is analytic in $\operatorname{Re} s > 0$. Thus if $G(s)$ has an l_ith order pole at $s = \sigma_i$ with $\operatorname{Re} \sigma_i > 0$, then $E(s)$ must have at least and l_ith order zero at $s = \sigma_i$. Thus for $q \leqq l_i - 1$,

$$\text{(A.4)} \qquad \int_0^\infty (-\tau)^q e^{-\sigma_i \tau} e(\tau)\,d\tau = \left.\frac{d^q E(s)}{ds^q}\right|_{s = \sigma_i} = 0.$$

Hence, for $q \leqq l_i - 1$,

$$\text{(A.5)} \qquad \int_0^\infty (t - \tau)^q e^{\sigma_i (t - \tau)} e(\tau)\,d\tau = 0,$$

and since

$$\hat{g}_2(t) = \sum_{i=1}^{n} \sum_{j=1}^{l_i} a_{i,j} t^{j-1} e^{\sigma_i t},$$

it then follows from (2.16) that

$$\int_0^\infty \hat{g}_2(t - \tau)e(\tau)\,d\tau = 0$$

for all t.

Proof of Lemma 2. Two preliminary lemmas are required for the proof of Lemma 2.

LEMMA 4. *Let Q be a mapping of $L_2(-\infty, \infty)$ into itself and be a contraction (i.e., $\|Qx_1 - Qx_2\| \leqq \gamma \|x_1 - x_2\|$ for all $x_1, x_2 \in L_2(-\infty, \infty)$ and some $\gamma < 1$).*

Then $I + Q$ is invertible on $L_2(-\infty, \infty)$, and if $P_TQ(I - P_T) \neq 0$ for some T, then $P_T(I + Q)^{-1}(I - P_T) \neq 0$.

Proof. The fact that $I + Q$ has an inverse if Q is a contraction is well known (cf. [9] and note that $L^2(-\infty, \infty)$ is complete). To prove the remainder of the lemma, note that since $P_TQ(I - P_T) \neq 0$, there exists some $x \in L_2(-\infty, \infty)$ with $P_Tx = 0$ and $P_TQx \neq 0$. Let $y = x + Qx$, so $P_Ty \neq 0$, and let $\hat{x}$ be given by $\hat{x} = (I + Q)^{-1}(I - P_T)y$. It will be shown that $P_T\hat{x} \neq 0$ and hence that $P_T(I + Q)^{-1}(I - P_T) \neq 0$.

Since

$$(I + Q)x = y \tag{A.6}$$

and

$$(I + Q)\hat{x} = (I - P_T)y, \tag{A.7}$$

subtracting (A.2) from (A.1) gives

$$(x - \hat{x}) + Qx - Q\hat{x} = P_Ty. \tag{A.8}$$

Thus, from (A.8),

$$\begin{aligned}\langle x - \hat{x}, P_Ty\rangle &= \langle x - \hat{x}, x - \hat{x} + Qx - Q\hat{x}\rangle \\ &= \|x - \hat{x}\|^2 + \langle x - \hat{x}, Qx - Q\hat{x}\rangle \\ &\geqq \|x - \hat{x}\|^2 - |\langle x - \hat{x}, Qx - Q\hat{x}\rangle| \\ &\geqq \|x - \hat{x}\|^2 - \|x - \hat{x}\|\|Qx - Q\hat{x}\| \\ &\geqq (1 - \gamma)\|x - \hat{x}\|^2.\end{aligned} \tag{A.9}$$

Also, from (A.8) it is seen that $\hat{x} \neq x$ since $P_Ty \neq 0$. Thus, from (A.9), $\langle x - \hat{x}, P_Ty\rangle = -\langle P_T\hat{x}, P_T y\rangle > 0$ and thus $P_T\hat{x} \neq 0$. Q.E.D.

LEMMA 5. *Let G and F satisfy the assumptions in Theorem 1. Then $(I + FG_{nc})$ has a noncausal inverse on $L_2(-\infty, \infty)$, and in fact, $P_T(I + FG_{nc})^{-1}(I - P_T) \neq 0$ for all $T \in R$.*

Proof. Consider the equation

$$(I + FG_{nc})e = u. \tag{A.10}$$

It will be first shown that (A.10) has a unique solution in $L_2(-\infty, \infty)$ for every $u \in L_2(-\infty, \infty)$, and hence that $(I + FG_{nc})^{-1}$ exists on $L_2(-\infty, \infty)$. Add and subtract cG_{nc} from (A.10) with $c > 0$ to obtain

$$(I + cG_{nc})e + (F - cI)G_{nc}e = u. \tag{A.11}$$

From the fact that $\operatorname{Re} G_{nc}(j\omega) \geqq 0$ for all ω, it follows that $[1 + cG_{nc}(j\omega)]^{-1}$ is bounded and thus defines a mapping of $L_2(-\infty, \infty)$ into itself [8], namely, $(I + cG_{nc})^{-1}$. Also, since $G_{nc}(s)$ has poles in $\operatorname{Re} s > 0$ and $\operatorname{Re}[1 + cG_{nc}(j\omega)] > 0$, it follows easily from the principle of the argument [7] that $[1 + cG_{nc}(s)]^{-1}$ has poles in $\operatorname{Re} s > 0$, and hence that $[I + cG_{nc}]^{-1}$ is noncausal.

Having established that $(I + cG_{nc})$ has a noncausal inverse, let $\tilde{e} = (I + G_{nc})e$ so that (A.11) becomes

$$\tilde{e} + (F - cI)G_{nc}(I + cG_{nc})^{-1}\tilde{e} = u. \tag{A.12}$$

Note that since $I - (I + cG_{nc})^{-1} = cG_{nc}(I + cG_{nc})^{-1}$, it follows that $G_{nc}(I + cG_{nc})^{-1}$ is noncausal. It will now be shown that for c sufficiently large, $Q = (F - cI)G_{nc}(I + cG_{nc})^{-1}$ is a contraction. First of all,

$$\text{(A.13)} \qquad \|G_{nc}(I + cG_{nc})^{-1}\| = \sup_{-\infty < \omega < \infty} |G_{nc}(j\omega)(1 + cG_{nc}(j\omega))^{-1}| \leqq 1/c,$$

where $\|\cdot\|$ denotes the induced norm on mappings of $L_2(-\infty, \infty)$ into itself. Secondly, for all $x_1, x_2 \in L_2(-\infty, \infty)$,

$$\begin{aligned} &\|(F - cI)x_1 - (F - cI)x_2\|^2 \\ \text{(A.14)} \qquad &= \|Fx_1 - Fx_2\|^2 - 2c\langle Fx_1 - Fx_2, x_1 - x_2\rangle + c^2\|x_1 - x_2\|^2 \\ &\leqq (k_2^2 - 2ck_1 + c^2)\|x_1 - x_2\|^2. \end{aligned}$$

Thus

$$\begin{aligned} \|Qx_1 - Qx_2\|^2 &\leqq (1/c^2)(k_2^2 - 2ck_1 + c^2)\|x_1 - x_2\|^2 \\ \text{(A.15)} \qquad &= \left(1 - \frac{2k_1}{c} + \frac{k_2^2}{c^2}\right)\|x_1 - x_2\|^2. \end{aligned}$$

For c sufficiently large, $(1 - 2k_1/c + k_2^2/c^2) < 1$, and hence $Q = (F - cI)G_{nc}(I + cG_{nc})^{-1}$ is a contraction.

Since Q is a contraction for c sufficiently large, $I + Q$ is invertible on $L_2(-\infty, \infty)$ for c sufficiently large. Further, since $e = (I + cG_{nc})^{-1}\tilde{e}$, it then follows that $I + FG_{nc}$ is also invertible, and in fact

$$\text{(A.16)} \qquad (I + FG_{nc})^{-1} = (I + cG_{nc})^{-1}(I + Q)^{-1}.$$

It remains to be shown that $P_T(I + FG_{nc})^{-1}(I - P_T) \neq 0$ for all $T \in R$. As shown above, $G_{nc}(I + cG_{nc})^{-1}$ is noncausal. Since it is a linear time-invariant operator, it is easily seen that $P_TG_{nc}(I + cG_{nc})^{-1}(I - P_T) \neq 0$ for all $T \in R$. Then, for $c > k_2$, $P_TQ(I - P_T) \neq 0$. Indeed, since F is causal,

$$\text{(A.17)} \qquad P_TQ(I - P_T) = P_T[F - cI]P_TG_{nc}(I + cG_{nc})^{-1}(I - P_T),$$

and for any $x \in L_2(-\infty, \infty)$,

$$\begin{aligned} \langle P_T(F - cI)P_Tx, P_Tx\rangle &= \langle FP_Tx, P_Tx\rangle - c\|P_Tx\|^2 \\ \text{(A.18)} \qquad &\leqq (k_2 - c)\|P_Tx\|^2, \end{aligned}$$

and thus for $c > k_2$, $P_T(F - cI)P_Tx = 0$ if and only if $P_Tx = 0$. Therefore, since $P_TG_{nc}(I + cG_{nc})^{-1}(I - P_T) \neq 0$, it follows from (A.17) that $P_TQ(I - P_T) \neq 0$. Lemma 4 then establishes that $P_T(I + Q)^{-1}(I - P_T) \neq 0$, and thus for every T there exists a $u \in L_2(-\infty, \infty)$ with $P_Tu = 0$ and $P_T\tilde{e} \neq 0$, where $\tilde{e}$ satisfies (A.12). Finally, it can be concluded that the corresponding solution e of (A.10) satisfies $P_Te \neq 0$. For suppose $P_Te = 0$. Then $P_TFG_{nc}e = 0$, and since F is causal and $\langle P_TFP_TG_{nc}e, P_TG_{nc}e\rangle \geqq k_1\|P_TG_{nc}e\|_2^2$, it then follows that $P_TG_{nc}e = 0$. But $\tilde{e} = e + cG_{nc}e$, so $P_T\tilde{e} = P_Te + cP_TG_{nc}e = 0$, giving a contradiction. Hence if $P_T\tilde{e} \neq 0$, then $P_Te \neq 0$, and therefore $P_T(I + FG_{nc})^{-1}(I - P_T) \neq 0$ for all $T \in R$. Q.E.D.

From Lemmas 4 and 5, the proof of Lemma 2 can now be given. Consider the equation

$$(I + HFG_{nc})e = u, \tag{A.19}$$

and add and subtract $cFG_{nc}e$ to obtain

$$(I + cFG_{nc})e + (H - cI)FG_{nc}e = u. \tag{A.20}$$

Now from Lemma 5 it is known that $(I + cFG_{nc})$ has an inverse on $L_2(-\infty, \infty)$ and satisfies $P_T(I + cFG_{nc})^{-1}(I - P_T) \neq 0$ for all $T \in R$. Thus let $\tilde{e} = (I + cFG_{nc})e$ so that (A.20) becomes

$$\tilde{e} + (H - cI)FG_{nc}(I + cFG_{nc})^{-1}\tilde{e} = u. \tag{A.21}$$

The remainder of the proof proceeds in precisely the same manner as that of Lemma 5. It is only necessary to show that $P_TW(I - P_T) \neq 0$ and $\|Wx_1 - Wx_2\| \leqq (1/c)\|x_1 - x_2\|$ for all $x_1, x_2 \in L_2(-\infty, \infty)$, where $W = FG_{nc}(I + cFG_{nc})^{-1}$. To show $P_TW(I - P_T) \neq 0$, simply observe that $W = -(1/c)[I - (I + cFG_{nc})^{-1}]$, and hence that $P_TW(I - P_T) = -(1/c)P_T(I + cFG_{nc})^{-1}(I - P_T) \neq 0$.

To show $\|Wx_1 - Wx_2\| \leqq (1/c)\|x_1 - x_2\|$, let $x_1, x_2 \in L_2(-\infty, \infty)$ and define y_i and e_i $(i = 1, 2)$ by

$$y_i = FG_{nc}(I + cFG_{nc})^{-1}x_i, \tag{A.22}$$

$$e_i = (I + cFG_{nc})^{-1}x_i. \tag{A.23}$$

Then

$$\langle y_1 - y_2, (x_1 - cy_1) - (x_2 - cy_2)\rangle = \langle FG_{nc}e_1 - FG_{nc}e_2, e_1 - e_2\rangle \geqq 0, \tag{A.24}$$

where condition (c) of Theorem 1 has been used to obtain the inequality. Using the Schwarz inequality it then follows from (A.24) that

$$\|y_1 - y_2\|\,\|x_1 - x_2\| \geqq \langle y_1 - y_2, x_1 - x_2\rangle \geqq c\|y_1 - y_2\|^2, \tag{A.25}$$

and thus $\|Wx_1 - Wx_2\| \leqq (1/c)\|x_1 - x_2\|$. Q.E.D.

REFERENCES

[1] A. R. Bergen and S. Takeda, *On instability of feedback systems with a single nonlinear time-varying gain*, IEEE Trans. Automatic Control, 16 (1971), pp. 462–464.

[2] R. W. Brockett and L. J. Forys, *On the stability of systems containing a time-varying gain*, Proc. 2nd Allerton Conference on Circuit and System Theory, Univ. of Illinois, Urbana, 1964.

[3] R. W. Brockett and H. B. Lee, *Frequency domain instability criteria for time-varying and nonlinear systems*, Proc. IEEE, 55 (1965), pp. 604–619.

[4] M. I. Freedman, *L_2-stability of time-varying systems—Construction of multipliers with prescribed phase characteristics*, this Journal, 6 (1968), pp. 559–578.

[5] M. I. Freedman and G. Zames, *Logarithmic variation criteria for the stability of systems with time-varying gains*, this Journal, 6 (1968), pp. 487–507.

[6] M. Gruber and J. L. Willems, *On a generalization of the circle criterion*, Proc. 4th Allerton Conference on Circuit and System Theory, Univ. of Illinois, Urbana, 1966, pp. 827–848.

[7] E. Hille, *Analytic Function Theory*, vol. I, Ginn, New York, 1959.

[8] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups*, vol. 31, 2nd ed., Colloquium Publications, American Mathematical Society, Providence, R.I., 1957.

[9] A. N. Kolmogorov and S. V. Fomin, *Functional Analysis*, vol. I, Graylock Press, New York, 1957.

[10] R. P. O'Shea, *A combined frequency-time domain stability criterion for autonomous continuous systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 477–484.

[11] I. W. Sandberg, *Some results on the theory of physical systems governed by nonlinear functional equations*, Bell System Tech. J., 43 (1965), pp. 871–898.

[12] R. A. Skoog, *Instability of slowly varying systems*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 86–92.

[13] J. C. Willems, *Stability, instability, invertibility, and causality*, this Journal, 7 (1969), pp. 645–671.

[14] 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 scalar nonlinearities*, IEEE Trans. Automatic Control, 11 (1966), pp. 228–239, 465–476.

[15] G. Zames and P. L. Falb, *Stability conditions for systems with monotone and slope-restricted nonlinearities*, this Journal, 6 (1968), pp. 89–108.

SIAM J. Control
Volume 12, No. 1, February 1974

ON THE EXISTENCE OF MOMENTS OF STATIONARY LINEAR SYSTEMS WITH MULTIPLICATIVE NOISE*

U. G. HAUSSMANN†

Abstract. Conditions are derived under which a control system described by a linear stochastic differential equation with multiplicative noise possesses a stationary probability distribution with finite *p*th moment. Both the Itô and Stratonovich interpretations are considered.

1. Introduction. Consider the control system described by the stochastic differential equation

$$\dot{x} = Ax - Bu - C(u)\dot{w}_1 + D(x)\dot{w}_2 + E\dot{w}_3. \tag{1.1}$$

Here x is the state vector, u the control vector and $\dot{w}_1, \dot{w}_2, \dot{w}_3$ are independent Gaussian white noise disturbances. The matrices $C(u)$ and $D(x)$ are assumed to be linear in their arguments, and hence these terms could model wide band disturbances in the matrices A and B. Some applications of this equation are given in [4].

A problem of some interest is to determine conditions under which the corresponding process has a steady state for which the pth moment, i.e. $\mathscr{E}\{\|x\|^p\}$, is finite. This is a type of stability, equivalent to Lagrange stability of deterministic systems. In the case $p = 2$, such a stability guarantees that there exists a control minimizing the expected value of a quadratic cost criterion in the steady state [1], [7], [8].

Usually (1.1) is interpreted in the sense of Itô; however, it is well known that actual physical processes described by a Langevin equation can best be approximated by equations interpreted in the sense of Stratonovich [5], [6]. In the present article we extend results about the second moments known for the Itô equation [1], [2], [3], [7], [8] to the Stratonovich equation, and then extend these results to higher moments.

In [8] it is shown that if the noise is sufficiently small, then the second moment is finite for (1.1) in the Itô sense. In [3] it is shown for the case $D = 0$ that the same conclusion holds if the system has suitable structure. After some preliminaries in § 2, both these results are shown to hold in § 3, when (1.1) is interpreted in the sense of Stratonovich. In § 4 it is shown that the same result holds for higher moments for either interpretation. Finally, in § 5 conditions on the structure of the system are given to guarantee finite the pth moment for the case $D \neq 0$.

2. Some preliminaries. Consider the stochastic differential equation

$$dx = (Ax - Bu)\,dt - C(u)\,dw_1 + D(x)\,dw_2 + E\,dw_3, \qquad t \geqq 0, \tag{2.1}$$

where x is a vector in R^n, Euclidean n-space with norm $\|x\| = [\sum_{i=1}^n |x_i|^2]^{1/2}$,

* Received by the editors September 25, 1972, and in revised form January 8, 1973.

† Department of Mathematics, University of British Columbia, Vancouver 8, British Columbia, Canada. This research was supported by the National Research Council of Canada under Grant A8051.

$n \geqq 2$, u is in R^m, and w_1, w_2, w_3 are independent Wiener processes of dimensions d_1, d_2, d_3 respectively. $C(u)$ and $D(x)$ are given by

$$C(u) = \sum_{i=1}^{m} C_i u_i,$$

$$D(x) = \sum_{i=1}^{n} D_i x_i,$$

and A, B, C_i, D_i, E are real constant matrices of corresponding dimensions. Hence D, C are linear operators mapping R^n (resp. R^m) into the space of $n \times d_2$ (resp. $n \times d_1$) matrices.

If in (2.1), u has the form $u = Kx$ and the random variable $x(0)$ is independent of the increments of the Wiener processes, then (2.1) determines a diffusion process

$$X_K = \{x(t) : t \geqq 0\}. \tag{2.2}$$

Here we have assumed that the stochastic integrals implicit in (2.1) are taken in the sense of Itô. If they are taken in the sense of Stratonovich [5], then a different diffusion process results,

$$\hat{X}_K = \{x(t) : t \geqq 0\}. \tag{2.3}$$

Although the Stratonovich interpretation of (2.1) is more meaningful physically, since it is the limit of a suitable sequence of Langevin equations [6]; the Itô interpretation is easier to treat mathematically. In view of this we observe that the process $\hat{X}_K$ also satisfies the *Itô equation*

$$dx = (\hat{A}x - \hat{B}(K)u)\,dt - C(u)\,dw_1 + D(x)\,dw_2 + E\,dw_3, \tag{2.4}$$

with $u = Kx$, where $\hat{A} = A + \frac{1}{2}\sum_{j=1}^{d_2} (\bar{D}_j)^2$, $\hat{B}(K) = B - \beta(K)$. Here $(\bar{D}_j)_{ik} = (D_k)_{ij}$, $\beta(K) = \frac{1}{2}\sum_{j=1}^{d_1} \bar{C}_j K \bar{C}_j$, and $(\bar{C}_j)_{ik} = (C_k)_{ij}$ (see [5]).

Let $\mathscr{L}_K$ be the differential operator given by

$$\mathscr{L}_K V(x) = \tfrac{1}{2}x'[\Delta(V_{xx}) + K'\Gamma(V_{xx})K]x + x'(A - BK)'V_x + \tfrac{1}{2}\operatorname{tr}(E'V_{xx}E), \tag{2.5}$$

where tr M is the trace of M, V_x is the vector $\partial V/\partial x$, and V_{xx} is the matrix $\partial^2 V/\partial x^2$. x' denotes the transpose of x. Moreover, $\Delta(P)_{ij} = \operatorname{tr}(D_i'PD_j)$ and $\Gamma(P)_{ij} = \operatorname{tr}(C_i'PC_j)$. Similarly, let $\hat{\mathscr{L}}_K$ be given by

$$\hat{\mathscr{L}}_K V(x) = \tfrac{1}{2}x'[\Delta(V_{xx}) + K'\Gamma(V_{xx})K]x + x'(\hat{A} - \hat{B}(K)K)'V_x + \tfrac{1}{2}\operatorname{tr}(E'V_{xx}E). \tag{2.6}$$

Then $\mathscr{L}_K$ and $\hat{\mathscr{L}}_K$ are the differential generators of X_K and $\hat{X}_K$ respectively.

We shall say that μ, a probability measure on the Borel sets of R^n, is *invariant* provided that $x(t)$ has distribution $\mu, t > 0$, whenever $x(0)$ has distribution μ. Of interest is the pth moment of μ, i.e.,

$$\mathscr{E}_\mu\{\|x\|^p\} \equiv \int_{R^n} \|x\|^p \mu(dx).$$

We wish to investigate conditions guaranteeing that there exist controls $u = Kx$ such that for either X_K or $\hat{X}_K$:

(2.7)
(i) the process admits at least one invariant probability measure;
(ii) all such invariant measures have finite pth moment.

The case of X_K with $p = 2$ and $D \equiv 0$ was discussed in [3].

3. Existence of second moments for $\hat{X}_K$. In [1], [8] it is shown that there exists a control $u = Kx$ such that the process X_K satisfies (2.7) with $p = 2$ provided

$$\inf_{K \in \mathscr{K}} \left\| \int_0^\infty e^{t(A-BK)'}[\Delta(I) + K'\Gamma(I)K\, e^{t(A-BK)}\, dt \right\| < 1, \tag{3.1}$$

where $\mathscr{K} = \{K : A - BK \text{ is stable}\}$. By definition $\mathscr{K} \neq \varnothing$ if (A, B) is stabilizable. The same method applied to (2.4) yields that there exists a control $u = Kx$ such that the process $\hat{X}_K$ satisfies (2.7) with $p = 2$ provided

$$\inf_{K \in \hat{\mathscr{K}}} \left\| \int_0^\infty e^{t(\hat{A} - \hat{B}(K)K)'}[\Delta(I) + K'\Gamma(I)K]\, e^{t(\hat{A} - \hat{B}(K)K)}\, dt \right\| < 1, \tag{3.2}$$

where $\hat{\mathscr{K}} = \{K : \hat{A} - \hat{B}(K) \text{ is stable}\}$.

Moreover, if (A, B) is stabilizable, and if $\|C\|$ and $\|D\|$ are sufficiently small, then $\hat{\mathscr{K}} \neq \varnothing$. Hence assuming (A, B) is stabilizable, we again have that if the noise intensity is sufficiently small, then an invariant measure with finite second moment exists.

However, just as in the case X_K, we can do better *if we assume* $D \equiv 0$ (see [3]). $\alpha(\lambda)$, the minimal polynomial of A, factors into $\alpha(\lambda) = \alpha_+(\lambda)\alpha_-(\lambda)$, where $\alpha_-(\lambda)$ has all zeros in the half-plane $\operatorname{Re}(\lambda) < 0$, and $\alpha_+(\lambda)$ has all zeros in $\operatorname{Re}(\lambda) \geqq 0$. Put

$$\mathscr{S}_\pm(A) = \{x \in R^n : \alpha_\pm(A)x = 0\}. \tag{3.3}$$

Next define

$$\begin{aligned} \mathscr{M}(\mathscr{V}) &= \{Bu : u \in R^m, \bar{C}_j u \in \mathscr{V}, j = 1, 2, \cdots, d_1\} \\ &= \{Bu : u \in R^m, \text{Image } C(u) \subset \mathscr{V}\}. \end{aligned}$$

Let $\mathscr{R}_0 = \mathscr{S}_-(A)$ and

$$\mathscr{R}_{i+1} = \text{span}\{\mathscr{R}_i, x, Ax, \cdots, A^{n-1}x : x \in \mathscr{M}(\mathscr{R}_i)\}, \qquad i \geqq 0.$$

THEOREM 3.1. *If $\mathscr{R}_m = R^n$, then there exists a control $u = Kx$ such that the process $\hat{X}_K$ satisfies* (2.7) *with $p = 2$.*

It should be observed that this theorem is the same as Theorem 3.7 in [3]. Hence if $\mathscr{R}_m = R^n$, then invariant probability measures exist and they have finite second moment irrespective of whether (2.1) is interpreted in the sense of Itô or in the sense of Stratonovich. Of course the associated control problem (cf. [1], [7]) can no longer be solved as easily with the Stratonovich interpretation as it could be with the Itô interpretation.

The proof of the theorem is the same as that of Theorem 3.7 in [3], if we proceed as follows. Let $T_\pm$ be the projection of R^n onto $\mathscr{S}_\pm(A)$ along $\mathscr{S}_\mp(A)$. Then $T_\pm A = AT_\pm \equiv A_\pm$. It follows that A_- restricted to $\mathscr{S}_-(A)$ is stable. Put $x_\pm = T_\pm x, \bar{B}_\pm(K) = T_\pm \hat{B}(K), C_\pm(u) = T_\pm C(u), E_\pm = T_\pm E$. Then (2.4) decomposes into

$$\begin{aligned} dx_- &= [A_- x_- - \bar{B}_-(K)u]\, dt - C_-(u)\, dw_1 + E_-\, dw_3, \\ dx_+ &= [A_+ x_+ - \bar{B}_+(K)u]\, dt - C_+(u)\, dw_1 + E_+\, dw_3. \end{aligned} \tag{3.4}$$

If K is of the form $K = K_+T_+$, i.e., $u = Kx = K_+x_+$, then

$$\bar{B}_+(K) = \hat{B}_+(K_+) \equiv B_+ - \beta_+(K_+),$$

where

$$\beta_+(K_+) = \tfrac{1}{2}\sum_j (\bar{C}_+)_j K_+ (\bar{C}_+)_j.$$

Here $C_+(u) = \sum_i T_+ C_i u_i \equiv \sum_i (C_+)_i u_i$, and $(\bar{C}_+)_j$ is derived from $(C_+)_i$ just as $\bar{C}_j$ is from C_i. Hence with such a control, (3.4) has exactly the same form as (2.4). It is known [3, Lemma 3.1] that a control as required by the theorem exists, provided that there exist real matrices $P \geqq 0$ (i.e., nonnegative definite and symmetric) and K_+ and positive constants λ_+ and ρ_+ such that for all $x_+ \in \mathscr{S}_+(A)$,

$$\hat{\mathscr{L}}^+_{K_+}(x'_+ P x_+) \leqq \lambda_+ - \rho_+ \|x_+\|^2.$$

Here $\hat{\mathscr{L}}^+_{K_+}$ is the operator as defined in (2.6) with + subscripts on K, Γ, A, $\hat{B}$, C_i, E and $D \equiv 0$; hence it is the differential generator of the process described by (3.4) with $u = K_+ x_+$.

As in [3] we try to stabilize A_+ using controls u in $\mathscr{N}$, the kernel of $C_+(\cdot)$. Assume R^{m_2} is isomorphic to $R^m/\mathscr{N}$ and let S be a linear map of R^{m_2} into R^m with the image of S being complementary to $\mathscr{N}$. Then $u = K_+ x_+$ can be written as

$$u = K^1 x_+ + S u^2,$$

where $K^1 x_+ \in \mathscr{N}$ and $u^2 \in R^{m_2}$. Hence

$$\begin{aligned} dx_+ &= (A_+ - B_+ K^1) x_+ \, dt + \beta_+(K_+) K^1 x_+ \, dt \\ &\quad - (B_+ - \beta_+(K_+)) S u^2 \, dt - C_+(S u^2) \, dw_1 + E_+ \, dw_3. \end{aligned} \tag{3.5}$$

As $K^1 x_+ \in \mathscr{N}$, then $\beta_+(K_+) K^1 x_+ = 0$. Put $E^2 = E_+$, $C^2(u^2) = C_+(S u^2)$, $B^2 = B_+ S$, $A^2 = A_+ - B_+ K^1$, $x^2 = x_+$, $u^2 = \bar{K} x^2$. Then

$$\beta_+(K_+) S = \beta_+(K^1 + S\bar{K}) S = \beta_+(S\bar{K}) S = \beta^2(\bar{K}),$$

where β^2 is derived from C^2 exactly as β is from C. Then (3.5) becomes

$$dx^2 = (A^2 x^2 - \hat{B}^2(\bar{K}) u^2) \, dt - C^2(u^2) \, dw_1 + E^2 \, dw_3. \tag{3.6}$$

This has the same form as (2.4) and so we can repeat the process. Now we can follow the method in [3] to complete the proof.

For the process X_K we were able to show that if $\mathscr{R}_m \neq R^n$ and if the noise is sufficiently large, then all second moments must be infinite. In the present case the proof breaks down, although one would certainly expect the result to hold.

4. Existence of higher moments. It is known [3], that if there exist a function $V(x)$ and positive constants k, ρ such that

$$\mathscr{L}_K V(x) \leqq k - \rho \|x\|^p, \tag{4.1}$$

then the process X_K satisfies (2.7). If we set $V(x) = (x'Px)^q$ for $P > 0$ with $q \geqq 1$, then just as in the proof of [3, Thm. 4.3],

$$\begin{aligned} \mathscr{L}_K V(x) &\leqq q(x'Px)^{q-1}\{(2q-1)\operatorname{tr}(E'PE) \\ &\qquad + x'[(2q-1)(\Delta(P) + K'\Gamma(P)K) + (A - BK)'P + P(A - BK)]x\}. \end{aligned} \tag{4.2}$$

But

$$\begin{aligned} &q(x'Px)^{q-1} x'[(2q-1)(\Delta(P) + K'\Gamma(P)K) + (A - BK)'P + P(A - BK)]x \\ &\qquad\qquad \leqq -\rho_0 \|x\|^{2q}, \end{aligned} \tag{4.3}$$

$\rho_0 > 0$, if and only if for some $Q > 0$,

$$(4.4) \quad (2q - 1)(\Delta(P) + K'\Gamma(P)K) + (A - BK)'P + P(A - BK) + Q = 0.$$

In this case (4.2) implies (4.1) with $\rho = \rho_0/2$ and with

$$k = \rho^{1-q}[q(2q - 1)\|P\|^{q-1} \operatorname{tr}(E'PE)]^q.$$

Combining this with Lemma 3.1 from [7] and setting $p = 2q$ we have the following.

LEMMA 4.1. *If*

$$(4.5) \quad \inf_{K \in \mathscr{K}} \left\| \int_0^\infty e^{t(A-BK)'}[\Delta(I) + K'\Gamma(I)K]\, e^{t(A-BK)}\, dt \right\| < (p - 1)^{-1},$$

then there exists a control $u = Kx$ such that the process X_K satisfies (2.7) *with* $p \geqq 2$.

We observe that the same result holds for the process $\hat{X}_K$, if (4.5) is replaced by

$$(4.6) \quad \inf_{K \in \hat{\mathscr{K}}} \left\| \int_0^\infty e^{t(\hat{A}-\hat{B}(K)K)}[\Delta(I) + K'\Gamma(I)K]\, e^{t(\hat{A}-\hat{B}(K)K)}\, dt \right\| < (p - 1)^{-1}.$$

Moreover, the methods of [1], [3] and the previous section yield further information *in the case* $\Delta \equiv 0$.

THEOREM 4.2. *If $\mathscr{R}_m = R^n$, then there exists a control $u = Kx$ such that both X_K and $\hat{X}_K$ satisfy* (2.7) *for any* $p \geqq 2$.

The proof of this theorem proceeds as in [3] for X_K, and as in the previous section for $\hat{X}_K$. As the equivalent of Lemma 3.1 of [3] one needs to show that if there is a mapping P_+ of $\mathscr{S}_+(A)$ onto itself satisfying (4.4) on $\mathscr{S}_+(A)$, then there is a mapping P of R^n onto itself satisfying (4.4) on all of R^n. On $\mathscr{S}_-(A)$, $A = A_-$ is stable, and so (4.4) can be solved for P_- with $Q = I$, $K = 0$ $(\Delta = 0)$. Let $V(x) = x'Px \equiv x'_+P_+x_+ + x'_-P_-x_-$. Mixing notation from (4.18) of [1], we see that the left side of (4.3) is

$$\begin{aligned} q(x'Px)^{q-1}\{&(2q - 1)x'_+K'_+[\Gamma(T'_+P_+T_+) + \beta\Gamma(T'_-P_-T_-)]K_+x_+ \\ &+ x'_+[(A_+ - T_+BK_+)'P_+ + P_+(A_+ - T_+BK_+)]x_+ \\ &+ \beta x'_-[A'_-P_- + P_-A_-]x_- - 2\beta x'_+(T_-BK_+)'P_-x_-\} \\ \leqq {} & -q(x'Px)^{q-1}\rho_1\|x\|^2 \end{aligned}$$

if $\beta > 0$ is sufficiently small as in (4.18) of [1]. Hence P satisfies (4.3) and consequently (4.4).

The remainder of the proof is exactly the same as in [3] for the case $p = 2$.

5. Further results for state-dependent noise. We shall now give some results for the case $D \neq 0$ analogous to Theorems 3.1 and 4.2; that is, we shall prove the existence of controls u such that the corresponding process satisfies (2.7) without requiring the noise to be small. Again we rely on the results of [1, § 4]. *We assume that* $C \equiv 0$. The basic idea is as follows. Let $\mathscr{N}_0$ be the kernel of D. Suppose there exists a matrix K such that $\mathscr{N}_0$ is invariant under $A - BK$, i.e., if x is in $\mathscr{N}_0$ then so is $(A - BK)x$, and such that $\mathscr{N}_0$ is a subset of $\mathscr{S}_-(A - BK)$ (cf. (3.3)). R^n can be split into $\mathscr{N}_0$ and $R^n/\mathscr{N}_0$. Now the proof of Theorem 4.1 of [1] goes through. (There is an extra term of the form $\beta y'Fz$ in (4.18) since $\mathscr{N}_0$ need not decompose

R^n relative to $A - BK$.) The condition (4.8) of [1] is satisfied since it pertains to Δ on $\mathcal{N}_0$, i.e., $\Delta_-(I) = 0$. Finally to satisfy condition (4.9) of [1], we use Lemma 4.4 of [1], i.e., we assume that $R^n/\mathcal{N}_0 \subseteq (\text{Image}(B))/\mathcal{N}_0$.

To formulate the result precisely we proceed as follows:

$$\mathcal{N}_1 \equiv \{u : D(Bu) = 0\} \text{ and } T_1 \text{ is any projection of } R^m \text{ onto } \mathcal{N}_1.$$

$$\mathcal{N}_2 \equiv \ker(T_1). \text{ Hence } R^m = \mathcal{N}_1 \oplus \mathcal{N}_2. \ \mathcal{N}_0 \equiv \{x : D(x) = 0\}.$$

THEOREM 5.1. *Assume*

(i) $R^n = \mathcal{N}_0 \oplus B(\mathcal{N}_2)$,

(ii) $\text{Image}(DAT_0) \subseteq \text{Image}(DB)$,

(iii) $\mathcal{N}_0 \subseteq \text{span}\{R_{BT_1}[T_0A], \mathcal{S}_-(T_0AT_0)\}$,

where T_0 *is the projection of* R^n *onto* $\mathcal{N}_0$ *along* $B(\mathcal{N}_2)$ *and where* $R_{BT_1}[A] =$ image $([BT_1, ABT_1, \cdots, A^{n-1}BT_1])$. *Then there exists a control* $u = Kx$ *such that* X_K *satisfies* (2.7) *for* $p = 2$.

Proof. From (i) and (ii) it follows that there exists a matrix K_0 with kernel $(K_0) \supseteq B(\mathcal{N}_2)$, image $(K_0) \subseteq \mathcal{N}_2$, such that if $u = K_0x + v$ for any v in $\mathcal{N}_1$, then $D(Bu) = D(AT_0x)$. Hence if $K = K_0 + K_1$ where K_1 is any matrix mapping R^n into $\mathcal{N}_1$, then

$$D[(A - BK)x] = 0$$

for x in $\mathcal{N}_0$. Thus $\mathcal{N}_0$ is invariant under $A - BK$.

It follows that

$$A - BK_0 = AT_2 + T_0AT_0, \tag{5.1}$$

where $T_2 = I - T_0$, and so if $\beta(\cdot)$ and $\gamma(\cdot)$ are the minimal polynomials of $A - BK_0$ and T_0AT_0 respectively, then

$$0 = \beta(A - BK_0)T_0 = \beta(T_0AT_0)T_0. \tag{5.2}$$

Hence $\gamma(\lambda)$ divides $\lambda\beta(\lambda)$. As $T_2\beta_-(T_0AT_0) = \beta_-(0)T_2$ and $\beta_-(0) \neq 0$, it follows that

$$\mathcal{S}_-(T_0AT_0) \subseteq \mathcal{N}_0. \tag{5.3}$$

Using (5.2) and (5.3) one sees that $\mathcal{S}_-(T_0AT_0) \subseteq \mathcal{S}_-(A - BK_0)$.

From (iii) it follows that $x \in \mathcal{N}_0$ has the form $x = x_1 + x_2$, where $x_1 \in R_{BT_1}[T_0A]$ and $x_2 \in \mathcal{S}_-(T_0AT_0)$. But (5.1) and $BT_1 = T_0BT_1$ imply $x_1 \in R_{BT_1}[A - BK_0]$. We conclude that

$$\begin{aligned} \mathcal{N}_0 &\subseteq \text{span}\{R_{BT_1}[A - BK_0], \mathcal{S}_-(A - BK_0)\} \\ &= \mathcal{S}_-(A - BK) \end{aligned}$$

for suitable K_1 [3, Thm. 3.2]. The proof is now completed as mentioned above using Theorem 4.1 and Lemma 4.4 from [1], observing that $\Delta > 0$ on $B(\mathcal{N}_2)$.

Some comments are appropriate. First, the work of the previous section shows that the theorem is valid for any $p \geqq 2$. In addition, these results are also true for the process $\hat{X}_K$. This follows readily if we observe that $AT_0 = \hat{A}T_0$, and then apply the theorem to (2.4).

Finally, for the case $C \neq 0$, the theorem is still true if we add the assumption:
(iv) for all u, image $(C(u)) \subseteq \mathcal{N}_0$.

The overall results of the theory are now summarized. If the pair (A, B) is stabilizable and the state- and control-dependent noise are sufficiently small, then (2.1) has a stationary solution whose pth moment is finite no matter whether the Itô interpretation or the Stratonovich interpretation is taken. The greater p is, the smaller the noise must be.

On the other hand, if $D = 0$ or if $C = 0$, and if the system has suitable structure, then just as in the case of no multiplicative noise, there exists a control giving rise to a stationary process with all moments finite. The result is again independent of the interpretation of (2.1). Moreover, if the system does not have the required structure and the control-dependent noise is sufficiently large, then no such control exists provided the Itô interpretation is used and $D = 0$. Finally if both state- and control-dependent noise are present but the control-dependent noise does not produce state-dependent noise, then a similar result holds.

REFERENCES

[1] U. G. HAUSSMANN, *Optimal stationary control with state and control dependent noise*, this Journal, 9 (1971), pp. 184–198.

[2] ———, *Stabilization of linear systems with multiplicative noise*, Proc. Internat. Symp. on the Stability of Stochastical Dynamical Systems, University of Warwick, Coventry, England, 1972; to appear in *Lecture Notes in Mathematics*, Springer-Verlag.

[3] ———, *Stability of linear systems with control dependent noise*, this Journal, 11 (1973), pp. 382–394.

[4] P. J. MCLANE, *Optimal stochastic control of linear systems with state- and control-dependent disturbances*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 793–798.

[5] R. L. STRATONOVICH, *A new representation for stochastic integrals and equations*, this Journal, 4 (1966), pp. 362–371.

[6] E. WONG AND M. ZAKAI, *On the relation between ordinary and stochastic differential equations*, Internat. J. Engrg. Sci., 3 (1965), pp. 213–229.

[7] W. M. WONHAM, *Optimal stationary control of a linear system with state-dependent noise*, this Journal, 5 (1967), pp. 486–500.

[8] ———, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. 2, A. T. Bharucha-Reid, ed., Academic Press, New York, 1969.

SIAM J. Control
Volume 12, No. 1, February 1974

RELAXIVE HILBERT PORTS*

A. H. ZEMANIAN†

Abstract. The completely monotonic behavior of the transient responses of RC n-ports to certain pulses of finite duration is used to characterize a more general kind of system, which we call the relaxive Hilbert port. This generalization encompasses networks having an infinite number of lumped and distributed elements as well as an infinite number of ports.

A primary result of this work is the generalization of Bernstein's theorem on completely monotonic functions to operator-valued functions. This in turn leads to a representation theorem, which states that the frequency-domain system function for a relaxive Hilbert port is the Stieltjes transform of a positive-operator measure. The approximation of the unit-impulse response by a finite sum of damped exponentials is discussed. For n-ports, this provides a means of synthesizing a relaxive Hilbert port as an RC network with perhaps ideal transformers.

Meixner's concept of a relaxation system of the second kind is also extended to operator-valued functions, and relaxive Hilbert ports are shown to be special cases of such systems. In fact, relaxive Hilbert ports are precisely those relaxation systems of the second kind whose unit-impulse responses remain bounded. Actually, the inevitable stray capacitances of any physical electrical system force every relaxation system of the second kind to be relaxive.

An example of a relaxive infinite system is also given.

1. Introduction. The concept of a "relaxive one-port", which was introduced in a prior work [18], extends the idea of a finite lumped RC one port with an initial shunting capacitor to the general context of linear time-invariant passive systems. Thus, infinite and distributed systems are included within this extension. Our present objective is to define and investigate the analogous concept for n-ports and more generally for Hilbert ports. The latter is a system analogous to the n-port but with signals that take their instantaneous values in a complex Hilbert space [14], [15].

To motivate our subsequent definition, consider the finite lumped RC n-port N of Fig. 1, where every port has a shunting capacitor. There are n current generators driving each port and yielding thereby an impressed current vector $u = \{u_1, \cdots, u_n\}$. The responding voltage vector is $v = \{v_1, \cdots, v_n\}$, and the polarities for these quantities are so assigned that $(u(t), v(t))$ is the complex power entering N at the instant t. Here, $(\cdot, \cdot)$ denotes the inner product in n-dimensional complex Euclidean space C^n. Any impulse of current $u = a\delta$, where $a \in C^n$ and δ is Dirac's delta function, deposits charges at the time $t = 0$ throughout a certain capacitive subnetwork N_1 of N. N_1 can be obtained by open-circuiting every resistor in N. N_1 will not in general contain every capacitor in N, but it will include every capacitor that is connected across the terminals of any port. The charge vector inserted through the n ports is the integral of $a\delta$, namely, a. On the other hand, the voltage $v(0)$ generated at $t = 0$ by the inserted charge is $v(0) = Ma$, where M is the inverse of the open-circuit capacitance matrix of N_1.

The charges in N_1 induce dissipating currents in N, which produce in turn a monotonic decay in v. Indeed, it is a fact (see [4, pp. 267–270]) that the open-circuit

* Received by the editors April 18, 1972, and in revised form January 18, 1973.

† Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, New York 11790. This work was supported by the National Science Foundation under Grant GP27958.

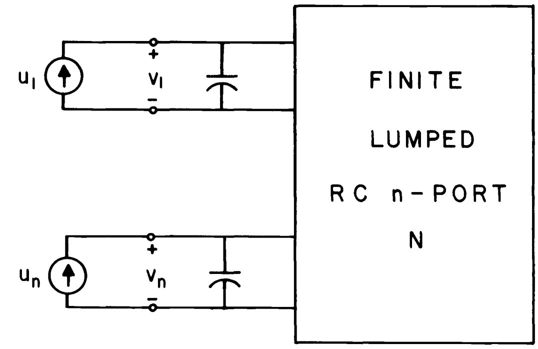

FIG. 1

impedance matrix Z of N has the form

$$Z(\zeta) = \sum_{j=1}^{m} \frac{P_j}{\zeta + \alpha_j}, \tag{1.1}$$

where $\zeta = \sigma + i\omega$ is a complex variable, each P_j is a positive (i.e., nonnegative definite) matrix, $M = \sum_{j=1}^{m} P_j$, and the α_j are real numbers satisfying $0 \leqq \alpha_1 < \alpha_2 < \cdots < \alpha_m$. The response v to the input $u = a\delta$ is therefore

$$v(t) = \sum_{j=1}^{m} P_j a e^{-\alpha_n t} 1_+(t). \tag{1.2}$$

$1_+(t)$ denotes Heaviside's unit step function. Since $(P_j a, a) \geqq 0$, it follows that $(v(t), a)$ is completely monotonic for $t > 0$. In fact, this property of $(v(t), a)$ continues to hold for $t > T$ if u is any current vector of the form $u = a\phi$, where ϕ is any smooth (i.e., infinitely differentiable) function with support contained in the interval $(-\infty, T]$. Moreover, in the latter case,

$$(v(t), a) \leqq (Ma, a) \int_{-\infty}^{\infty} \phi(t)\, dt.$$

It is through these properties that we shall define our general class of relaxive Hilbert ports in § 3. It will be shown in § 4 that a necessary and sufficient condition for a linear time-invariant Hilbert port with a convolution representation to be relaxive is that its response v to the input $u = a\delta$ be of the form

$$v(t) = \int_0^{\infty} dP_\eta a\, e^{-\eta t} 1_+(t), \tag{1.3}$$

where P_η is a PO measure (i.e., a positive-operator measure) on $[0, \infty)$. The proof of this result is based on an extension to operator-valued functions of Bernstein's theorem concerning completely monotonic functions [12, p. 160]. The needed extension is established in the next section. Moreover, the impedance Z of a relaxive Hilbert port is characterized by a Stieltjes integral analogous to (1.1). It follows easily from this that every Hilbert port is passive.

The approximation and synthesis of relaxive n-ports by finite lumped RC n-ports is investigated in § 5. Actually, our results are obtained in the general context of Hilbert ports, but, until synthesis techniques with appropriate building blocks are devised for Hilbert ports, the results of this section have a physical significance only for n-ports.

In § 6 we extend Meixner's class of relaxation systems of the second kind [8, p. 282] to Hilbert ports and then show that our relaxive Hilbert ports comprise a subset of that extended class. We exploit these results in § 7 to develop an example of a relaxive Hilbert port consisting of an infinite lumped RC network with an infinite number of ports.

Our notation is identical to that used in [14]. Thus, given any two topological linear spaces U and V, $[U; V]$ denotes the linear space of all continuous linear mappings of U into V, and $\langle f, \phi \rangle$ or alternatively $f\phi$ is that member of V assigned by $f \in [U; V]$ to $\phi \in U$. $[U; V]$ is supplied with the topology of uniform convergence on the bounded sets in U. $\| \cdot \| = \| \cdot \|_B$ denotes the norm in any Banach space B. H is a complex Hilbert space with the inner product $(\cdot, \cdot)$. R is the real line, R^+ is the closed interval $[0, \infty)$, C is the complex plane, $C_+ = \{\zeta \in C : \operatorname{Re} \zeta > 0\}$, and $C_0 = \{\zeta \in C : \zeta \text{ is not a real nonpositive number}\}$. The kth derivative of any Banach-space-valued function or distribution f is denoted by $f^{(k)} = D^k f$. At times, we use the symbol $D_x{}^k$ to show that the differentiations are with respect to x. We set $f^{[k]} = (-1)^k f^{(k)}$. $\operatorname{supp} f$ is the support of f. We sometimes use $\triangleq$ to denote an equality by definition. Finally, s-lim denotes a limit in the strong operator topology of $[H; H]$.

2. An extension of Bernstein's theorem on completely monotonic functions. Let z be an $[H; H]$-valued function on R^+. We shall say that z is completely monotonic on an open interval $\Omega \subset R$ if, for each nonegative integer k, each $a \in H$, and each $t \in \Omega$, we have that $z^{(k)}(t)$ exists under the strong operator topology and $(z^{[k]}(t)a, a) \geqq 0$. Furthermore, we shall call z completely monotonic on R^+ if $z(t)$ tends to a limit in the strong operator topology as $t \to 0+$ and z is completely monotonic on the interior of R^+. When H is the complex plane, this definition becomes essentially the customary definition of scalar (i.e., numerically-valued) completely monotonic functions [12, p. 145]. We wish to extend the following theorem to $[H; H]$-valued functions.

BERNSTEIN'S THEOREM. *Let f be a scalar function on R^+. A necessary and sufficient condition for f to be completely monotonic on R^+ is that*

$$f(t) = \int_{R^+} d\mu_\eta e^{-\eta t}, \qquad t > 0,$$

where μ_η is a finite positive measure on the Borel subsets of R^+. μ_η is uniquely determined by f.

We shall show that the desired extension can be obtained by replacing μ_η with a PO measure P_η. (For a discussion of PO measures, see [1, Chap. 8], [3], or [17, Chap. 2].) In the following we let $[H; H]_+$ denote the space of positive continuous linear operators on H.

LEMMA 2.1. *Assume that $z(t)$ is an $[H; H]_+$-valued function on the open interval $(0, \infty)$ such that, for all $a \in H$ and as $t \to 0+$, $(z(t)a, a)$ increases monotonically to a finite limit. Then, there exists an $F \in [H; H]_+$ such that, as $t \to 0+$, $z(t) \to F$ in the strong operator topology.*

Proof. By polarization $(z(t)a, b)$ also tends to a finite limit for every $a, b \in H$. Define E by

$$E(a, b) = \lim_{t \to 0+} (z(t)a, b).$$

It follows that $\{a, b\} \mapsto E(a, b)$ is a positive sesquilinear form on $H \times H$. Two applications of the principle of uniform boundedness show that

$$|E(a, b)| = \lim_{t \to 0+} |(z(t)a, b)| \leqq M\|a\| \, \|b\|.$$

Consequently, there exists a unique $F \in [H; H]_+$ such that $(Fa, b) = E(a, b)$. By virtue of [17, Lemma 2.2-1], we can now conclude that $z(t) \to F$ in the strong operator topology.

Our extension of Bernstein's theorem is the following.

THEOREM 2.1. *Let z be an $[H; H]$-valued function on R^+. A necessary and sufficient condition for z to be completely monotonic on R^+ is that*

$$z(t) = \int_{R^+} dP_\eta \, e^{-\eta t}, \qquad t > 0, \tag{2.1}$$

where P_η is a PO measure on the Borel subsets of R^+. P_η is uniquely determined by z.

Note. Equation (2.1) shows that the function $z(t)$ is a Laplace transform and therefore analytic for $t > 0$. Consequently, its derivatives exist in the uniform operator topology as well [17, Thm. 1.7-1].

Proof. Sufficiency. For any scalar continuous bounded function on R^+, we have the estimate

$$\left\| \int dP_\eta g(\eta) \right\|_{[H;H]} \leqq \|P(R^+)\|_{[H;H]} \sup_{\eta \in R^+} |g(\eta)|. \tag{2.2}$$

With the use of this inequality it is straightforward to show that (2.1) can be differentiated under the integral sign any number of times at each $t > 0$. Hence, for every $a \in H$, $t > 0$, and nonnegative integer k,

$$(z^{[k]}(t)a, a) = \int_{R^+} d(P_\eta a, a)\eta^k \, e^{-\eta t} \geqq 0 \tag{2.3}$$

since P_η is a PO measure. Furthermore, (2.3) shows that $(z(t)a, a)$ increases monotonically to the limit $(z(0)a, a)$ as $t \to 0+$, which by Lemma 2.1 implies that $z(t) \to z(0)$ in the strong operator topology.

Necessity. We may write

$$(z^{[k]}(t)a, a) = (-1)^k D_t^k (z(t)a, a), \qquad t > 0. \tag{2.4}$$

By hypothesis, $(z(t)a, a)$ is a completely monotonic scalar function on $0 \leqq t < \infty$. By Bernstein's theorem there exists a unique finite positive measure $\mu_\eta(a)$ depending on a such that

$$(z(t)a, a) = \int_{R^+} d\mu_\eta(a) \, e^{-\eta t}.$$

Now, let $a, b \in H$ and $\beta \in C$. Then,

$$\int_{R^+} d\mu_\eta(\beta a)\, e^{-\eta t} = (z(t)\beta a, \beta a) = |\beta|^2 (z(t)a, a)$$

$$= |\beta|^2 \int_{R^+} d\mu_\eta(a)\, e^{-\eta t}.$$

By the uniqueness of the Laplace transformation [5, Thm. 6.2.3], we have the following equality between measures:

$$\mu_\eta(\beta a) = |\beta|^2 \mu_\eta(a). \tag{2.5}$$

In a similar way, we can show that

$$\mu_\eta(a + b) + \mu_\eta(a - b) = 2\mu_\eta(a) + 2\mu_\eta(b). \tag{2.6}$$

We now define the complex measure $Q_\eta(a, b): E \mapsto [Q_\eta(a, b)](E)$, where E is any Borel subset of R^+, by

$$Q_\eta(a, b) \triangleq \tfrac{1}{4}[\mu_\eta(a + b) - \mu_\eta(a - b) + i\mu_\eta(a + ib) - i\mu_\eta(a - ib)].$$

For fixed E, $a \mapsto [\mu_\eta(a)](E)$ is a functional on H taking only nonnegative values. This fact coupled with (2.5) and (2.6) allows us to conclude (see, for example, Kurepa [6]) that $\{a, b\} \mapsto [Q_\eta(a, b)](E)$ is a positive sesquilinear form on $H \times H$ such that

$$[Q_\eta(a, a)](E) = [\mu_\eta(a)](E) \geqq 0.$$

Furthermore,

$$|[Q_\eta(a, a)](E)| = [\mu_\eta(a)](E) \leqq [\mu_\eta(a)](R) = (z(0)a, a)$$

$$\leqq \|z(0)\| \, \|a\|^2.$$

Therefore, there exists a unique $P_\eta(E) \in [H; H]_+$ such that $(P_\eta(E)a, b) = [Q_\eta(a, b)](E)$ for every $a, b \in H$. Since $(P_\eta(E)a, a) = [\mu_\eta(a)](E)$ and $[\mu_\eta(a)](\cdot)$ is a positive finite measure, we can conclude that P_η is a PO measure on the Borel subsets of R^+. (See [17, Thm. 2.2-1].)

Thus, for all $t > 0$ and all $a \in H$,

$$(z(t)a, a) = \int_{R^+} d(P_\eta a, a)\, e^{-\eta t} = \left(\int_{R^+} dP_\eta\, e^{-\eta t} a, a \right).$$

By means of the polarization equation, we arrive at (2.1). The uniqueness of P_η follows from the fact that $(z(t)a, a)$ uniquely determines $(P_\eta a, a)$ for every $a \in H$.

3. Relaxive Hilbert ports. We shall always assume that the impedance operator $\mathfrak{Z}$ of the Hilbert port at hand is a continuous linear time-invariant causal mapping of $\mathscr{D}(H)$ into $[\mathscr{D}; H]$. As in [14], $\mathscr{D}(H)$ is the linear space of all smooth H-valued functions on R of compact support and is supplied with the customary Schwartz topology. We set $\mathscr{D} = \mathscr{D}(C)$. $[\mathscr{D}; H]$ is the linear space of all H-valued distributions on R. $[\mathscr{D}; H]$ has the topology of uniform convergence on the bounded sets in $\mathscr{D}$. $\mathfrak{Z}$ has the stated properties if and only if $\mathfrak{Z}$ is a convolution operator $\mathfrak{Z} = z *$, where $z \in [\mathscr{D}(H); H]$ and $\operatorname{supp} z \subset R^+$ (see [14, Thm. 6.1]).

By the identification between $[\mathscr{D}(H); H]$ and $[\mathscr{D}; [H; H]]$ given by [14, Thm. 3.1], we can also say that $z \in [\mathscr{D}; [H; H]]$; that is, z is an operator-valued distribution on R. z is called the unit-impulse response, and its Laplace transform $Z = \mathfrak{L}z$ is called the impedance function or simply the impedance.

The domain of $\mathfrak{Z}$ contains every H-valued distribution u whose support is bounded on the left. We set $v = \mathfrak{Z}u = z * u$. Moreover, for any scalar distribution f such that $\operatorname{supp} f$ is bounded on the left, $z * f$ has a sense as an $[H; H]$-valued distribution, and $z * (fa) = (z * f)a$ for every $a \in H$. When $\phi \in \mathscr{D}$, $z * \phi$ is a smooth $[H; H]$-valued function with support bounded on the left, and

$$(z * \phi)(t) = \langle z(x), \phi(t - x)\rangle = \langle z(t - x), \phi(x)\rangle. \tag{3.1}$$

(See [14, Thm. 4.3].)

DEFINITION 3.1. $\mathfrak{Z}$ will be called *relaxive* whenever the following two conditions are satisfied for every nonnegative $\phi \in \mathscr{D}$.

(i) If $T = \sup \operatorname{supp} \phi$, then $z * \phi$ is completely monotonic on the open interval (T, ∞).

(ii) There exists an $M \in [H; H]_+$ not depending on ϕ such that, for all $t \in R$,

$$(z * \phi)(t) \leqq M \int_R \phi(x)\, dx. \tag{3.2}$$

Whenever these conditions hold, we shall also say that the unit-impulse response z, the impedance Z, and the corresponding Hilbert port are relaxive.

Note that condition (i) is equivalent to the assertion that, for $v = \mathfrak{Z}u$ and for every u of the form $u = \phi a$, where $\phi \in \mathscr{D}$, $\phi \geqq 0$, and $a \in H$, we have that $(v(\cdot), a)$ is a completely monotonic scalar function on (T, ∞). Similarly, condition (ii) is equivalent to the requirement that

$$(v(t), a) \leqq (Ma, a) \int_R \phi(x)\, dx$$

for all u of the stated form. These are the conditions we obtained from the physically motivated discussion in the Introduction.

THEOREM 3.1. *$\mathfrak{Z}$ is relaxive if and only if z is an ordinary $[H; H]$-valued function on R and z is completely monotonic on R^+. When this is the case, $z(0+) = s\text{-}\lim_{t \to 0+} z(t)$ is the infimum (in the sense of positive operators) of all M for which (3.2) holds.*

Proof. If. Set $w = z * \phi = \mathfrak{Z}\phi$, where $\phi \in \mathscr{D}$. For every nonnegative integer k, it is permissible to write

$$w^{[k]}(t) = \langle z^{[k]}(x), \phi(t - x)\rangle. \tag{3.3}$$

(See [14, (4.16)].) Let $\phi \geqq 0$ and $T \triangleq \sup \operatorname{supp} \phi$. For $t > T$, the support of the function $\phi(t - x)$ of x is contained in $(0, \infty)$, and therefore

$$w^{[k]}(t) = \int_{\operatorname{supp} \phi} z^{[k]}(x)\phi(t - x)\, dx.$$

Since for each $x > 0$ and $t > T$ the integrand on the right-hand side is a member of $[H; H]_+$, so too is the left-hand side. Condition (i) is hereby established.

Furthermore, for every $t \in R$ and $a \in H$,

$$0 \leqq (w(t)a, a) = \int_0^\infty (z(x)a, a)\phi(t - x)\, dx$$

$$\leqq (z(0+)a, a) \int_0^\infty \phi(x)\, dx.$$

This proves (3.2). That $z(0+)$ is the infimum of all possible M in (3.2) can be seen by choosing ϕ as an arbitrarily sharp pulse.

Only if. Assume that $\mathfrak{Z} = z *$ is relaxive and set

$$w(t) = z * \phi(t) = \langle z(x), \phi(t - x)\rangle, \qquad \phi \in \mathscr{D}.$$

Also, let $\phi \geqq 0$ and $t > T$. Then, for each $k = 0, 1, 2, \cdots$,

$$w^{[k]}(t) = \langle z^{[k]}(x), \phi(t - x)\rangle \in [H; H]_+$$

according to condition (i). So, for all $a \in H$,

$$0 \leqq (w^{[k]}(t)a, a) = (\langle z^{[k]}(x), \phi(t - x)\rangle a, a).$$

Thus, with $\psi \in \mathscr{D}$ and $\operatorname{supp} \psi \subset (0, \infty)$, $\psi \mapsto (\langle z^{[k]}, \psi\rangle a, a)$ is a positive scalar distribution on $(0, \infty)$ and is therefore a positive measure on the Borel subsets of $(0, \infty)$ [10, p. 29]. Since this is so for every k, $\psi \mapsto (\langle z, \psi\rangle a, a)$ is distributionally the same as a smooth nonnegative function f_a on $(0, \infty)$ [10, p. 85]. In fact, f_a is analytic on $(0, \infty)$. Indeed,

$$\begin{aligned} \int f_a^{[k]}\psi\, dt &= \int f_a\psi^{(k)}\, dt = (\langle z, \psi^{(k)}\rangle a, a) \\ &= (\langle z^{[k]}, \psi\rangle a, a) \geqq 0 \end{aligned} \tag{3.4}$$

for all $\psi \geqq 0$. Hence, f_a is completely monotonic on $(0, \infty)$. Whence, f_a is analytic on $(0, \infty)$ according to Bernstein's theorem and the analyticity of a Laplace transform.

Now, for every $a, b \in H$, we define the analytic function $f_{a,b}$ on $(0, \infty)$ by

$$f_{a,b}(\tau) = \tfrac{1}{4}[f_{a+b}(\tau) - f_{a-b}(\tau) + if_{a+ib}(\tau) - if_{a-ib}(\tau)], \qquad 0 < \tau < \infty.$$

It can be shown by expansion that

$$\int f_{a,b}(t)\psi(t)\, dt = (\langle z, \psi\rangle a, b)$$

for every $\psi \in \mathscr{D}$ with $\operatorname{supp} \psi \subset (0, \infty)$. So, for each fixed ψ of this sort,

$$\{a, b\} \mapsto \int f_{a,b}(t)\psi(t)\, dt$$

is a sesquilinear form on $H \times H$, which is positive when $\psi \geqq 0$.

Next, we choose an appropriate sequence $\{\psi_\nu\} \subset \mathscr{D}$ which tends to δ_τ, the delta functional concentrated on the fixed point $\tau \in (0, \infty)$. Then, for any $a, b, c \in H$, we get

$$f_{a+c,b}(\tau) \leftarrow \int f_{a+c,b}\psi_\nu \, dt = (\langle z, \psi_\nu\rangle(a + c), b)$$

$$= \int f_{a,b}\psi_\nu \, dt + \int f_{c,b}\psi_\nu \, dt \rightarrow f_{a,b}(\tau) + f_{c,b}(\tau),$$

and similarly for $f_{a,b+c}$, $f_{\alpha a,b}$, and $f_{a,\alpha b}$, where $\alpha \in C$. This shows that $\{a, b\} \mapsto f_{a,b}(\tau)$ is a sesquilinear form on $H \times H$. It is also positive since $f_{a,a}(\tau) = f_a(\tau) \geqq 0$.

Now, let K be any compact interval contained in $(0, \infty)$ and let $\eta, \tau \in K$ be arbitrary except that $\eta < \tau$. Choose $\psi \in \mathscr{D}$ such that $\psi \geqq 0$, $\operatorname{supp}\psi \subset (0, \infty)$, and $\psi = 1$ on K. Then,

$$\int_\eta^\tau f_a \, dt \leqq \int_0^\infty f_a\psi \, dt = (\langle z, \psi\rangle a, a) \leqq \|\langle z, \psi\rangle\| \, \|a\|^2.$$

Also, by the results of the preceding paragraph, $\{a, b\} \mapsto \int_\eta^\tau f_{a,b} \, dt$ is a positive sesquilinear form on $H \times H$. These facts imply that there exists a mapping Q of the closed intervals $[\eta, \tau] \subset K$ into $[H; H]_+$ such that

$$\int_\eta^\tau f_{a,b} \, dt = (Q([\eta, \tau])a, b).$$

With η fixed, define $g(\tau) \triangleq Q([\eta, \tau])$. Thus,

$$\int_\eta^\tau f_{a,b} \, dt = (g(\tau)a, b). \tag{3.5}$$

Note that the left-hand side is an analytic function of $\tau \in K$. Therefore, g is an $[H; H]_+$-valued analytic function of τ. We may differentiate (3.5) to get

$$f_{a,b}(\tau) = (f(\tau)a, b), \tag{3.6}$$

where $f(\tau) = g^{(1)}(\tau)$. Since η, τ, and K are arbitrary, (3.6) holds for all $\tau \in (0, \infty)$, where again f is analytic on $(0, \infty)$.

Altogether then, for every $\psi \in \mathscr{D}$ with $\operatorname{supp}\psi \subset (0, \infty)$,

$$(\langle z, \psi\rangle a, b) = \int f_{a,b}\psi \, dt = \int (f(t)a, b)\psi(t) \, dt$$

$$= \left(\int f\psi \, dt \, a, b\right),$$

which shows that $z = f$. Hence, z is analytic on $(0, \infty)$. It now follows from (3.4) that $(z^{[k]}(\tau)a, a) \geqq 0$ for all $\tau > 0$ and $a \in H$. That is, z is completely monotonic on $(0, \infty)$.

We next show that $\lim_{t\to 0+}(z(t)a, a)$ exists for every $a \in H$. If this is not so, then, by the complete monotonicity of z, $(z(t)a, a) \to \infty$ for at least one a. Let $\mathscr{P} = \{\phi \in \mathscr{D} : \phi \geqq 0, \int \phi \, dt = 1\}$. We can make $(z(t)a, a) * \phi(t)$ as large as we wish in a neighborhood of $t = 0$ simply by choosing $\phi \in \mathscr{P}$ as a sufficiently sharp pulse. This will violate (3.2) for any fixed M. We conclude that the aforementioned limit must exist. By Lemma 2.1, s-$\lim_{t\to 0+} z(t)$ also exists.

Another thing we have to show is that z does not have any singularities concentrated on the origin. Let w be an $[H;H]$-valued function on R such that $w = 0$ on $(-\infty, 0]$ and $w = z$ on $(0, \infty)$. Then, $h \triangleq z - w$ is an $[H;H]$-valued distribution concentrated on the origin. Therefore, for each $a \in H$,

$$(za, a) = \sum_{j=1}^{n} c_j \delta^{(j)} + (wa, a),$$

where the c_j are complex numbers depending on a. (See [13, p. 98].) Now, with a fixed, $(wa, a) * \mathscr{P}$ is a uniformly bounded set of functions because, for any $\phi \in \mathscr{P}$,

$$\begin{aligned} 0 &\leqq (w * \phi(t)a, a) = [(wa, a) * \phi](t) \\ &= \int (w(x)a, a)\phi(t - x)\, dx \leqq (w(0+)a, a). \end{aligned}$$

On the other hand,

$$\sum c_j \delta^{(j)} * \phi = \sum c_j \phi^{(j)}$$

can be made arbitrarily large at any given t by choosing $\phi \in \mathscr{P}$ appropriately so long as $c_j \neq 0$ for some j and some a. Therefore, $c_j = 0$ for all j and all a if condition (ii) of Definition 3.1 is to be fulfilled.

Consequently, $(za, a) = (wa, a)$. With this fact and the polarization equation, we can show that

$$(\langle z, \psi \rangle a, b) = (\langle w, \psi \rangle a, b)$$

for all $\psi \in \mathscr{D}$ and all $a, b \in H$. Hence, $z = w$ in the sense of equality in $[\mathscr{D}, [H;H]]$. Therefore, z is an ordinary $[H;H]$-valued function on R. This completes the proof.

As was noted immediately after Definition 3.1, $(v(\cdot), a)$ is a completely monotonic function after the termination of the pulse $u = \phi a$, where $\phi \in \mathscr{D}$, $\phi \geqq 0$, and $a \in H$. We can use Theorem 3.1 to show that this property continues to hold even when ϕ is an arbitrary positive distribution of compact support. The proof is quite similar to the argument for the scalar case [18, § IV].

4. Representation theorems. Some representations for relaxive Hilbert ports now follow readily. First, we note that the value at $t = 0$ of any relaxive z can be altered without changing z as a distribution. Henceforth, we shall set $z(0) = s\text{-}\lim_{t \to 0+} z(t)$.

THEOREM 4.1. *A Hilbert port is relaxive if and only if its unit-impulse response z admits the representation*

$$z(t) = \int_{R^+} dP_\eta e^{-\eta t} 1_+(t), \qquad t \in R, \tag{4.1}$$

where P_η is a PO measure on the Borel subsets of R^+. This is the case if and only if the impedance Z has the representation

$$Z(\zeta) = \int_{R^+} dP_\eta \frac{1}{\zeta + \eta}, \qquad \zeta \in C_0. \tag{4.2}$$

(C_0 is the complex plane with the nonpositve real axis deleted.)

Proof. The first assertion follows from the conjunction of Theorems 2.1 and 3.1. To obtain the second assertion, we take the Laplace transform of (4.1) and then reverse the order of integration. The latter can be justified by showing that this can be done weakly by virtue of Fubini's theorem. This yields (4.2) for $\zeta \in C_+$, which is a vector version of the Stieltjes transformation. That the equality in (4.2) holds weakly (and therefore in $[H; H]$ as well) for all $\zeta \in C_0$ follows from the fact that any scalar Stieltjes transform exists for all $\zeta \in C_0$ if it exists for any single $\zeta \in C_0$ [12, p. 326].

In fact, Z is an $[H; H]$-valued analytic function on C_0 since, for every $a, b \in H$, $(Z(\zeta)a, b)$ is a scalar analytic function on C_0 [12, p. 328]. Furthermore, $(Z(\zeta)a, a)$ is a scalar relaxive impedance and the poles of such a function are simple with positive residues [18, Thm. 4]. It follows that the poles of a relaxive $[H; H]$-valued function are also simple and lie on the real nonpositive axis, and their residues are all members of $[H; H]_+$.

The fact that a relaxive $\mathfrak{Z} = z *$ is passive on $\mathscr{D}(H)$ does not appear to be readily obtainable from the definition of relaxivity. However, the representation (4.2) allows us to arrive at this conclusion directly.

THEOREM 4.2. *Every relaxive Z is positive** (see [14, Def. 10.1]) *and the corresponding operator $\mathfrak{Z} = z *$ is passive on $\mathscr{D}(H)$.*

Proof. For every $a \in H$ and $\zeta \in C_+$,

$$\operatorname{Re}(Z(\zeta)a, a) = \int_{R_+} d(P_\eta a, a) \operatorname{Re}\frac{1}{\zeta + \eta} \geqq 0,$$

which shows that Z is positive*. The passivity of $\mathfrak{Z}$ now follows from [14, Thm. 12.2].

Actually, the representation (4.2) shows something more, namely, that $(Z(\sigma)a, a)$ is real for every $\sigma > 0$. This in turn implies that Z is positive*-real [14, Def. 10.2] whenever H is the complexification of a real Hilbert space.

For the rest of this section we discuss the special case where H is n-dimensional complex Euclidean space C^n. In this case the measure P_η takes its values in the space of positive $n \times n$ matrices. If P_η is concentrated on m discrete points of R^+, we get the representation (1.1) of the impedance matrix of a passive RC n-port containing possibly ideal transformers.

For the general case where P_η is a $[C^n; C^n]_+$-valued measure, set $\mu(\eta) \triangleq P([0, \eta))$ for $\eta > 0$ and let $\mu(0)$ be the zero matrix. Then, $\mu(\eta)$ is an $n \times n$ matrix $[\mu_{ik}(\eta)]$; its elements μ_{ik} are functions of η, which are continuous from the left and zero at the origin. Upon choosing $\alpha = \{\alpha_j\}_{j=1}^n \in C^n$ with $\alpha_j = 0$ for $j \neq k$, we see that $(\mu(\eta)\alpha, \alpha) = \mu_{kk}(\eta)|\alpha_k|^2$. Consequently, μ_{kk} is a nondecreasing bounded function on R^+. A similar argument with no more than two of the components of α nonzero shows that

$$f(\alpha_j, \alpha_k, \eta) \triangleq \mu_{kk}(\eta)|\alpha_k|^2 + \mu_{kj}(\eta)\bar{\alpha}_k\alpha_j + \mu_{jk}(\eta)\bar{\alpha}_j\alpha_k + \mu_{jj}(\eta)|\alpha_j|^2$$

is a nondecreasing bounded function of η. Upon first setting $\alpha_k = \alpha_j = 1$, then setting $\alpha_k = 1$, $\alpha_j = i$, and finally combining the results, we see that μ_{jk} and μ_{kj} are functions of bounded variation of R^+. We can therefore conclude with the following.

THEOREM 4.3. *If $Z = [Z_{jk}]$ is the $n \times n$ open-circuit impedance matrix of a relaxive n-port, then each Z_{jk} has the Stieltjes integral representation*

$$Z_{jk}(\zeta) = \int_0^\infty \frac{1}{\zeta + \eta}\, d\mu_{jk}(\eta), \tag{4.3}$$

where every μ_{jj} is a real nondecreasing bounded function on R^+ and every μ_{jk} is a function of bounded variation on R^+. Moreover, $\mu_{jk} = \bar{\mu}_{kj}$.

5. Approximation in the time domain. Our objective in this section is to show that relaxive z can be approximated under a rather strong topology by finite sums of damped exponentials with positive-operator coefficients. For n-ports, the approximating sums will have the significance of the unit-impulse response of a finite lumped RC n-port with perhaps ideal transformers. Thus, the results of this section imply an approximation and synthesis method for relaxive n-ports.

THEOREM 5.1. *Let z be relaxive. Then, there exists a sequence $\{z_q\}_{q=1}^{\infty}$ such that each z_q has the form*

$$\begin{gathered} z_q(t) = \sum_{v=0}^{n_q} c_{q,v}\, e^{-\beta_{q,v}t} 1_+(t), \\ c_{q,v} \in [H;H]_+, \quad \sum_{v=0}^{n_q} c_{q,v} = z(0+), \quad 0 \leqq \beta_{q,0} < \beta_{q,1} < \cdots < \beta_{q,n_q} \end{gathered} \tag{5.1}$$

and such that $\{z_q\}$ converges in the strong operator topology to $z(t)$ uniformly for all $t \in R^+$.

Proof. By virtue of Lemma 2.1 and Theorem 4.1, there exists a $P_1 \in [H;H]_+$ (possibly $P_1 = 0$) such that

$$z(t) = P_1 1_+(t) + \int_{R^+} dQ_\eta\, e^{-\eta t} 1_+(t), \tag{5.2}$$

where Q_η is a PO measure on R^+ such that $(Q_\eta([0,x))a, a) \to 0$ as $x \to 0+$. Now, for each $q = 1, 2, \cdots$, we will choose a partition π_q of R^+ with the endpoints

$$\beta_{q,0} \triangleq 0 < \beta_{q,1} < \beta_{q,2} < \cdots < \beta_{q,n_q} < \infty$$

and will set

$$z_q(t) = P_1 1_+(t) + \int_{R^+} dQ_\eta s(\eta, t) 1_+(t),$$

where

$$s(\eta, t) = \begin{cases} \exp(-\beta_{q,v}t), & \beta_{q,v-1} \leqq \eta < \beta_{q,v}, \quad v = 1, \cdots, n_q, \\ \exp(-\beta_{q,n_q}t), & \beta_{q,n_q} \leqq \eta < \infty. \end{cases}$$

Note that z_q has the form of (5.1) for $c_{q,v} = Q([\beta_{q,v} - \beta_{q,v-1}))$, where $v = 1, \cdots, n_q - 1$, $c_{q,n_q} = Q([\beta_{q,n_q - 1}, \infty))$, and $c_{q,0} = P_1$. Moreover, $\sum_v c_{q,v} = P(R^+) = z(0+)$. Thus,

$$\begin{aligned} z(t) - z_q(t) = &\int_{[0,\beta_{q,1})} dQ_\eta(e^{-\eta t} - e^{-\beta_{q,1}t}) 1_+(t) \\ &+ \int_{[\beta_{q,1},\beta_{q,n_q})} dQ_\eta[e^{-\eta t} - s(\eta, t)] 1_+(t) \end{aligned}$$

(cont.)

$$+\int_{[\beta_{q,n_q},\infty)} dQ_\eta e^{-\eta t}1_+(t) - Q([\beta_{q,n_q},\infty))\exp(-\beta_{q,n_q})1_+(t)$$

$$= A_1(t) + A_2(t) + A_3(t) + A_4(t).$$

Here, the A_i denote the four terms between the two equality signs in the indicated order.

For any $F \in [H; H]_+$ we have the estimate

$$\|Fa\|^2 \leqq (Fa, a)\|F\| \tag{5.3}$$

for all $a \in H$ [17, Lemma 2.2-1]. Let us apply this to $A_3(t)$. By virtue of (2.2),

$$\|A_3(t)\| \leqq \|Q([\beta_{q,n_q}, \infty))\| \leqq \|Q(R^+)\|.$$

Also, $(A_3(t)a, a) \leqq (Q([\beta_{q,n_q}, \infty)a, a) \to 0$ as $\beta_{q,n_q} \to \infty$. So, for every $a \in H$, $\|A_3(t)a\| \to 0$ as $\beta_{q,n_q} \to \infty$ uniformly for all $t \in R^+$. Similarly, $\|A_4(t)a\|$ does the same.

For A_1, we have $\|A_1(t)\| \leqq \|Q([0, \beta_{q,1}))\| \leqq \|Q(R^+)\|$ and $(A_1(t)a, a) \leqq (Q([0, \beta_{q,1}))a, a) \to 0$ as $\beta_{q,1} \to 0+$. Therefore, for every $a \in H$, $\|A_1(t)a\| \to 0$ as $\beta_{q,1} \to 0+$ uniformly for all $t \in R^+$.

Finally,

$$\|A_2(t)\|_{[H;H]} \leqq \|Q(R^+)\| \sup\{|e^{-\eta t} - s(\eta, t)| : \beta_{q,1} \leqq \eta \leqq \beta_{q,n_q}, 0 \leqq t < \infty\}.$$

The right-hand side can be made as small as desired simply by choosing the partition π_q fine enough. This is because $e^{-\beta t}$ is a uniformly continuous function of $\{t, \beta\}$ for $\beta_{q,1} \leqq \eta \leqq \beta_{q,n_q}$ and $0 \leqq t < \infty$.

In fact, we can construct the desired sequence $\{z_q\}$ by choosing a sequence $\{\pi_q\}$ of partitions with the following properties: $\beta_{q+1,1} < \frac{1}{2}\beta_{q,1}$, $\beta_{q+1,n_q} > 2\beta_{q,n_q}$, and the length of the largest interval in the partition of $[\beta_{q+1,1}, \beta_{q+1,n_q}]$ is less than half the length of the smallest interval in the partition of $[\beta_{q,1}, \beta_{q,n_q}]$. This ends the proof.

Actually, Theorem 5.1 can be strengthened by showing that a sequence $\{z_q\}$ can be constructed which converges under a much stronger topology than that implied in the theorem. Indeed, with $f(t)$ denoting an $[H; H]$-valued function on R^+, let

$$\gamma_{0,a}(f) = \sup_{0 \leqq t < \infty} \|f(t)a\|, \qquad a \in H,$$

$$\gamma_{k,a,x}(f) = \sup_{x \leqq t < \infty} \|f^{(k)}(t)a\|, \qquad a \in H, \quad x > 0, \qquad k = 1, 2, \cdots.$$

The set of all f for which all these quantities are finite comprise a locally convex space $\mathscr{V}$ whose topology $\mathscr{T}$ is generated by the seminorms $\gamma_{0,a}$ and all $\gamma_{k,a,x}$.

Now, for any relaxive z, for $t > 0$, and for $k = 1, 2, \cdots$,

$$z^{[k]}(t) = \int_{R^+} dQ_\eta \eta^k e^{-\eta t}.$$

With z_q constructed as in the preceding proof, we have, for t and k as stated,

$$z_q^{[k]}(t) = \int_{R^+} dQ_\eta s_k(\eta, t),$$

where

$$s_k(\eta, t) = \begin{cases} \beta_{q,\nu}^k \exp(-\beta_{q,\nu} t), & \beta_{q,\nu-1} \leqq \eta < \beta_{q,\nu}, \quad \nu = 1, \cdots, n_q, \\ \beta_{q,n_q}^k \exp(-\beta_{q,n_q} t), & \beta_{q,n_q} \leqq \eta < \infty. \end{cases}$$

Choose the partitions π_q as indicated in the last paragraph of the preceding proof. Then, it follows through almost the same argument that, for each $k > 0$ and as $q \to \infty$, $z_q^{[k]}$ tends to $z^{[k]}$ in the strong operator topology uniformly on every interval of the form $[x, \infty)$, where $x > 0$. Thus, $z_q \to z$ under the topology $\mathscr{T}$.

Conversely, let $\{z_q\}$ be a Cauchy sequence under the topology $\mathscr{T}$, where each z_q has the form of (5.1). By virtue of the sequential completeness of $[H; H]$ under the strong operator topology and the uniform convergence of each derivative of $\{z_q\}$, there exists a limit function z to which $\{z_q\}$ converges under $\mathscr{T}$. (In this regard, see [17, Problem 1.6-3].) Moreover, z is relaxive since every z_q is relaxive. Thus, we have proven the following.

THEOREM 5.2. *The set of all relaxive z is the closure in $\mathscr{V}$ of the set of all z_q of the form* (5.1).

6. Relaxation systems of the second kind. We now introduce the vector analogue of Meixner's relaxation systems of the second kind [8, p. 282] and show that relaxive Hilbert ports are special cases of such systems. The physical significance of this will be pointed out at the end of this section.

DEFINITION 6.1. Z is said to be a *relaxation impedance of the second kind* if $(sZ(s^2)a, a)$ is a scalar positive-real function of s for all $a \in H$. The set of all such Z is denoted by $\mathscr{M}_2$. The corresponding Hilbert port is said to be a *relaxation system of the second kind.*

It is understood throughout the following that the range of the argument function is restricted to the branch $(-\pi, \pi]$. We exploit the standard result that F is a positive-real function if and only if F is analytic and $|\arg F(s)| \leqq |\arg s|$ for all $s \in C_+$ [9]. A simple manipulation with $\zeta = s^2$ converts this result into the following criterion for the functions in $\mathscr{M}_2$.

LEMMA 6.1. *$Z \in \mathscr{M}_2$ if and only if Z satisfies the following conditions for every $a \in H$. Z is analytic on C_0 and*

$$(6.1) \qquad -\arg \zeta \leqq \arg (Z(\zeta)a, a) \leqq 0, \qquad \operatorname{Im} \zeta > 0,$$

$$(6.2) \qquad \arg (Z(\sigma)a, a) = 0, \qquad \sigma > 0,$$

$$(6.3) \qquad 0 \leqq \arg (Z(\zeta)a, a) \leqq -\arg \zeta, \qquad \operatorname{Im} \xi < 0.$$

The sum of two impedances in $\mathscr{M}_2$ is also in $\mathscr{M}_2$ in view of Definition 6.1. Therefore, the series connection of two Hilbert ports having such impedances yields another Hilbert port of this kind. A similar assertion holds for parallel connections, but now we must make sure that the parallel combination of impedances has a sense. In the following, $F^{-1}(\zeta) \triangleq [F(\zeta)]^{-1}$.

THEOREM 6.1. *If W and Z are members of $\mathscr{M}_2$ and if $W^{-1}(\zeta)$, $Z^{-1}(\zeta)$, and $[W^{-1}(\zeta) + Z^{-1}(\zeta)]^{-1}$ exist for every $\zeta \in C_0$, then*

$$(6.4) \qquad F \triangleq (W^{-1} + Z^{-1})^{-1} \in \mathscr{M}_2.$$

Proof. F will be analytic wherever W and Z are both analytic. Hence, we need merely investigate $\arg (F(\zeta)a, a)$. By our assumption concerning the existence of the inverses, for any given $a \in H$ there exists a unique $b \in H$ such that $[W^{-1}(\zeta) + Z^{-1}(\zeta)]b = a$. Moreover, there exist unique c, $d \in H$ such that $b = W(\zeta)c = Z(\zeta)d$. Consequently,

$$\begin{aligned}(F(\zeta)a, a) &= (b, [W^{-1}(\zeta) + Z^{-1}(\zeta)]b) \\ &= (W(\zeta)c, c) + (Z(\zeta)d, d).\end{aligned}$$

Now, for any $\alpha, \beta \in C$ with $\operatorname{Im} \alpha < 0$ and $\operatorname{Im} \beta < 0$, we have that

$$\min\{\arg \alpha, \arg \beta\} \leqq \arg(\alpha + \beta) \leqq \max\{\arg \alpha, \arg \beta\}.$$

Consequently, F satisfies (6.1). The same argument establishes (6.2) and (6.3), thereby completing the proof.

To show that every relaxive impedance is a member of $\mathscr{M}_2$, we shall make use of a certain representation for $\mathscr{M}_2$ impedances, which we now establish. We start with Schwindt's representation [14, p. 130] for a positive* function, which we may apply to $sZ(s^2)$ since $sZ(s^2)$ is a positive* function whenever $Z \in \mathscr{M}_2$.

$$sZ(s^2) = sP_1 + P_0 + \int_R dQ_\xi \frac{1 - i\xi s}{s - i\xi}, \qquad s \in C_+. \tag{6.5}$$

Here, $P_1 \in [H; H]_+$, P_0 is a skew-adjoint member of $[H; H]$, and Q_η is a PO measure on the Borel subsets of R. Since, for any $\sigma > 0$ and any $a \in H$, $(\sigma Z(\sigma^2)a, a)$ is a real number and since the real and imaginary parts of the integrand in (6.5) are even and odd functions of ξ, we can conclude that $P_0 = 0$, that $Q_\xi = Q_{-\xi}$ (i.e., $Q_\xi(E) = Q_{-\xi}(E) \triangleq Q_\xi(-E)$ for any Borel subset E of R), and that

$$sZ(s^2) = sP_1 + \int_R dQ_\xi \frac{s(1 + \xi^2)}{s^2 + \xi^2}, \qquad s \in C_+.$$

Upon setting $\zeta = s^2$, we see that there exists a PO measure M_η such that

$$Z(\zeta) = P_1 + \int_{R^+} dM_\eta \frac{1 + \eta}{\zeta + \eta}, \qquad \zeta \in C_0. \tag{6.6}$$

This is the vector analogue to Meixner's representation for a scalar relaxation impedance of the second kind [8, Thm. 3.2].

THEOREM 6.2. *$Z \in \mathscr{M}_2$ if and only if Z admits the representation* (6.6), *where $P_1 \in [H; H]_+$ and M_η is a PO measure on the Borel subsets of R^+.*

We are now ready to relate relaxive impedances to the members of $\mathscr{M}_2$.

THEOREM 6.3. *Z is relaxive if and only if $Z \in \mathscr{M}_2$ and $(\sigma Z(\sigma)a, a)$ tends to a finite limit as $\sigma \to \infty$ for each $a \in H$.*

Proof. Since the representation (4.2) is a special case of (6.6), a relaxive Z is certainly a member of $\mathscr{M}_2$. Moreover, the existence of $\lim_{\sigma \to \infty} (\sigma Z(\sigma)a, a)$ follows easily from (4.2).

Conversely, from (6.6) we get

$$(\sigma Z(\sigma)a, a) = (P_1 a, a)\sigma + \int_{R^+} d(M_\eta a, a) \frac{\sigma(1 + \eta)}{\sigma + \eta}. \tag{6.7}$$

Now,

$$\sigma(1 + \eta)/(\sigma + \eta) \tag{6.8}$$

is a function of η that varies monotonically from the value 1 at $\eta = 0$ toward the value σ as $\eta \to \infty$. Moreover, for each fixed η, (6.8) increases toward $1 + \eta$ as $\sigma \to \infty$. Since $(M_\eta a, a)$ is a positive measure, the integral in (6.7) either increases indefinitely or increases toward a finite limit. Furthermore, $(P_1 a, a)\sigma \to \infty$ as $\sigma \to \infty$ for at least one $a \in H$ if $P_1 \neq 0$. Therefore, under the assumption that $(\sigma Z(\sigma)a, a)$ tends to a finite limit, we must have that $P_1 = 0$ and $\int d(M_\eta a, a)(1 + \eta) < \infty$ for each $a \in H$. We define $P_\eta(E) \triangleq \int_E dM_\eta(1 + \eta)$ for each Borel subset E of R and obtain thereby the PO measure P_η, the representation (4.2), and the conclusion that Z is relaxive.

COROLLARY 6.3a. *Z is relaxive if and only if Z is analytic on C_0 and, for each $a \in H$, Z satisfies conditions* (6.1) *through* (6.3) *and* $\lim_{\sigma \to \infty} (\sigma Z(\sigma)a, a)$ *exists.*

We can interpret Theorem 6.3 physically as follows. Some of the Hilbert ports that are relaxation systems of the second kind will respond to an impulse of current with an infinite initial voltage and will absorb an infinite amount of energy. On the other hand, relaxive Hilbert ports are precisely those relaxation systems of the second kind for which this does not happen. For the n-port discussed in the Introduction, it is the capacitive subnetwork N_1 which prevents infinite initial voltages and infinite energy absorption. Actually, every physical electrical system has stray capacitances which act like the capacitive subnetwork N_1.

7. An example of a relaxive ∞-port. By an ∞-port we mean a Hilbert port for which H is Hilbert's coordinate space l_2 [16]. Actually, we shall let l_2 be the space of all *two-sided* quadratically summable numerical sequences $\{\alpha_k\}_{k=-\infty}^{\infty}$.

The results of the preceding section will be used to show that the ∞-port of Fig. 2 is relaxive. All the series resistances therein are r ohms, the shunt conductances are g mhos, the shunt capacitances are c farads, and we take $rg = 3.43$. It will simplify our formulas a bit if we choose $r = 5.04$. This ∞-port can be considered to be the parallel connection of the two ∞-ports shown in Figs. 3 and 4.

Let us first investigate the ∞-port of Fig. 3. As has been pointed out by H. Flanders [2], care must be taken in choosing the current distribution in the infinite resistive grid. An unreasonable set of currents, which satisfy Kirchhoff's voltage and current laws and for which no input currents are imposed at the ports, is shown in Fig. 5. The only current distribution we will allow when there is no excitation at the ports is the one where all branch currents are zero. When the current u_k is imposed at the kth port and all other ports have no imposed currents, we may apply a suitably modified form of Flanders' analysis [2] to conclude that a unique current exists in each resistance and each conductance. (The modification involves the partitioning of every current distribution into equivalence classes, two current distributions being considered equivalent if they agree on all the branches other than the branches appearing in the ground line, that is, in the

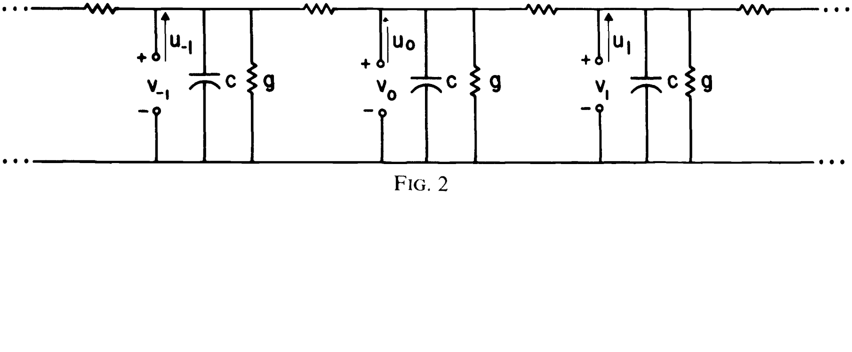

FIG. 2

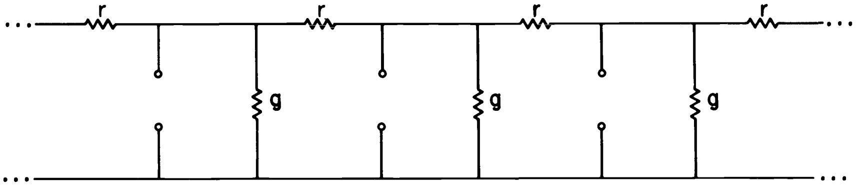

FIG. 3

lower horizontal line of Fig. 3.) The current distribution in the resistances and conductances can be computed by using the standard equations for lumped iterated transmission lines [11, pp. 106, 119]. The current distribution at the instant t under an imposed current vector $\{u_k\}_{k=-\infty}^{\infty} \in l_2$ on the ∞-port is then taken to be the sum of the current distributions due to each u_k applied one at a time.

The aforementioned transmission line equations show that the open-circuit $\infty \times \infty$ impedance matrix for the ∞-port of Fig. 3 is $Z = [Z_{ik}]$, where $Z_{ik}(\zeta)$ $= \rho^{-|i-k|}$ for all $\zeta, i, k = \cdots, -1, 0, 1, \cdots$, and $\rho = 5.23 \cdots$. Thus, $Z(\zeta)$ does not vary with ζ. (Henceforth, Z denotes that fixed range value, and not the function.) We can verify that $Z \in [l_2; l_2]$ as follows. Let $a = \{a_k\}_{k=-\infty}^{\infty} \in l_2$ and set $b = Za$. Then,

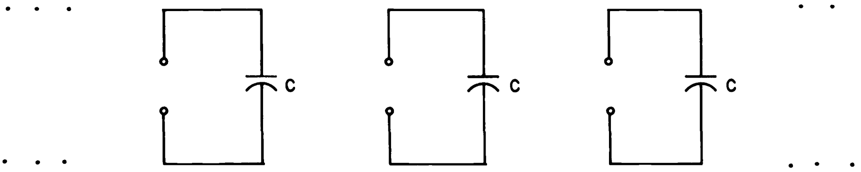

FIG. 4

$$\|b\|^2 = \sum_{i=-\infty}^{\infty}\left[\sum_{k=-\infty}^{\infty} Z_{ik}a_k\right]\left[\sum_{j=-\infty}^{\infty} Z_{ij}\bar{a}_j\right].$$

Upon taking absolute values and then changing the order of summation, we get

$$\begin{aligned}\|b\|^2 &\leqq \sum_k \sum_j |a_k a_j| \sum_i Z_{ik}Z_{ij} \\ &= \sum_k \sum_j |a_k a_j| \rho^{-|k-j|}(2\zeta + |k-j| + 1),\end{aligned}$$

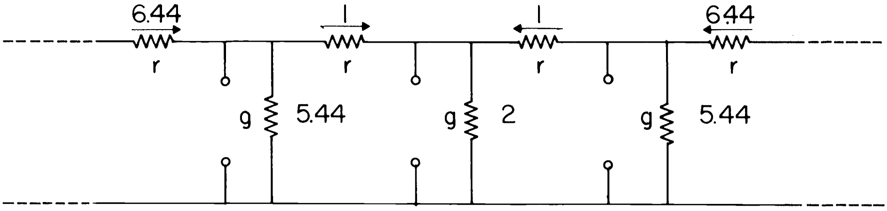

FIG. 5

where

$$\xi = \sum_{\nu=1}^{\infty} \rho^{-2\nu} = .0379 \cdots .$$

Summing diagonally, we obtain

$$\|b\|^2 \leqq \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} |a_n a_{n+m}| \rho^{-|m|}(2\xi + |m| + 1).$$

An application of Schwarz's inequality to the summation on n yields

$$\|b\|^2 \leqq \|a\|^2 \sum_{m=-\infty}^{\infty} \rho^{-|m|}(2\xi + |m| + 1).$$

Since the last summation on m converges, we can conclude that $Z \in [l_2; l_2]$.

Furthermore, Z is invertible. To establish this, it is enough to show that $A \triangleq I - Z$ satisfies $\|A\| < 1$. Here, I is the identity operator on l_2 so that $A = [A_{ik}]$, where $A_{ii} = 0$ and $A_{ik} = -\rho^{-|i-k|}$ for $i \neq k$. This time set $d = Aa$, where $a \in l_2$. The same manipulations as those in the preceding paragraph now yield

$$\begin{aligned}\|d\|^2 &\leqq \|a\|^2\left[2\xi + 4\xi \sum_{m=1}^{\infty} \rho^{-|m|} + 2 \sum_{m=2}^{\infty} \rho^{-|m|}(|m| - 1)\right] \\ &= \|a\|^2(.2227 \cdots).\end{aligned}$$

Thus, $\|A\| \leqq .471 \cdots$, which is what we wished to show.

Still more is true. Z is a positive operator on l_2. Indeed, for any $a \in l_2$,

$$(Za, a) = \sum_{i=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \rho^{-|i-k|} \bar{a}_i a_k.$$

We can conclude that this convergent double series is nonnegative because $\rho^{-|t|}$ is a positive definite function [7, p. 70]. So truly, $Z \in [l_2; l_2]_+$. As a consequence, $Z^{-1} \in [l_2; l_2]_+$ also. In addition, Lemma 6.1 now shows that $Z \in \mathscr{M}_2$.

We turn now to the ∞-port of Fig. 4. Let W denote its open-circuit $\infty \times \infty$ impedance matrix. Clearly, $W = [W_{ik}]$, where $W_{ii}(\zeta) = 1/c\zeta$ for every i and $W_{ik} = 0$ when $i \neq k$. Moreover, $(W(\zeta)a, a) = (c\zeta)^{-1}\|a\|^2$. Thus, by Lemma 6.1, $W \in \mathscr{M}_2$.

We want to investigate the open-circuit $\infty \times \infty$ impedance matrix F of the ∞-port of Fig. 2. This is the parallel combination of W and Z if it exists as an operator. To check the latter condition, consider

$$F(\zeta) = [W^{-1}(\zeta) + Z^{-1}]^{-1} = (c\zeta I + Z^{-1})^{-1}.$$

We have already noted that Z^{-1} is a fixed positive operator. Therefore, its spectrum is contained in the real nonnegative axis. Consequently, $F(\zeta)$ exists for every $\zeta \in C_0$. Theorem 6.1 can now be invoked to conclude that $F \in \mathscr{M}_2$.

To show that F is relaxive, we shall use Theorem 6.3. So, let $\sigma > 0$ and consider

$$(\sigma F(\sigma)a, a) = \frac{1}{c}\left(\left(I + \frac{1}{c\sigma}Z^{-1}\right)^{-1} a, a\right).$$

For all σ sufficiently large, we have $\|(c\sigma Z)^{-1}\| < 1$ and therefore

$$(\sigma F(\sigma)a, a) = c^{-1} \sum_{k=0}^{\infty} ((-c\sigma Z)^{-k}a, a).$$

Moreover, there exists an $x > 0$ for which the last series converges uniformly for $x < \sigma < \infty$. So, we may pass to the limit under the summation sign to conclude that $(\sigma F(\sigma)a, a) \to c^{-1}\|a\|^2$ as $\sigma \to \infty$. By Theorem 6.3, F is relaxive.

We have hereby established that the system of Fig. 2 is an example of a relaxive ∞-port.

REFERENCES

[1] R. Beals, *Topics in Operator Theory*, The University of Chicago Press, Chicago, 1971.

[2] H. Flanders, *Infinite networks*: *I—resistive networks*, IEEE Trans. Circuit Theory, CT-18 (1971), pp. 326–331.

[3] W. Hackenbroch, *Integration vektorwertiger Funktionen nach operatorwertigen Massen*, Math. Z., 105 (1968), pp. 327–344.

[4] D. Hazony, *Elements of Network Synthesis*, Reinhold, New York, 1963.

[5] E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, Colloquium Publications, vol. 31, American Mathematical Society, Providence, R.I., 1957.

[6] S. Kurepa, *Quadratic and sesquilinear functionals*, Glasnik Mat.-Fiz. i Astr., 20 (1965), pp. 79–92.

[7] E. Lukacs, *Characteristic Functions*, Hafner, New York, 1960.

[8] J. Meixner, *On the theory of linear passive systems*, Arch. Rational Mech. Anal., 17 (1964), pp. 278–296.

[9] S. Seshu and N. Balabanian, *Transformations of positive real functions*, IEEE Trans. Circuit Theory, CT-4 (1957), pp. 306–312.

[10] L. Schwartz, *Théorie des distributions*, Hermann, Paris, 1966.

[11] E. Weber, *Linear Transient Analysis*, vol. II, John Wiley, New York, 1956.

[12] D. V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, N.J., 1946.

[13] A. H. Zemanian, *Distribution Theory and Transform Analysis*, McGraw-Hill, New York, 1965.

[14] ———, *The Hilbert port*, SIAM J. Appl. Math., 18 (1970), pp. 98–138.

[15] ———, *A scattering formulism for the Hilbert port*, Ibid., 18 (1970), pp. 467–488.

[16] ———, *Banach systems, Hilbert ports, and ∞-ports*, Network Theory, R. Boite, ed., Gordon and Breach, London, 1972.

[17] ———, *Realizability Theory for Continuous Linear Systems*, Academic Press, New York, 1972.

[18] ———, *Relaxive one ports*, IEEE Trans Circuit Theory, CT-20 (1973), pp. 139–142.

SIAM J. Control
Vol. 12, No. 1, February 1974

NUMERICAL SOLUTION OF ITO INTEGRAL EQUATIONS

N. J. RAO†, J. D. BORWANKAR‡ AND D. RAMKRISHNA§

Abstract. An algorithm is derived for solving a large class of Ito random integral equations. The derivation of the algorithm involves approximate discretization of the given Ito equation. The Ito integrals arising out of discretization are expressed as functions of normal random variables. The algorithm gives a sample pathwise solution and is readily implementable on a digital computer.

1. Introduction. A wide variety of engineering dynamical systems with stochastic inputs and/or parameter disturbances are modeled as Ito random integral (or differential) equations, and consequently the study of these equations is of great interest to the engineer. A stirred tank chemical reactor with stochastic changes in the input concentrations and the lateral, or the longitudinal, dynamics of an aircraft with gust disturbances are good examples of such models. The present paper is concerned with deriving an algorithm to solve a fairly general class of Ito integral equations. The algorithm, which involves discretization of the given equation, is readily implementable on a digital computer and gives a sample pathwise solution. The Ito scalar stochastic integral equation is of the form

$$x(t,\omega) = x_u(\omega) + \int_u^t a(s, x(s,\omega))\,ds + \int_u^t b(s, x(s,\omega))\,dW(s), \tag{1.1}$$

where $x(s,\omega) \in R$, ω is an element of the sample space and $\{W(t)\}$ is a Wiener process. Equation (1.1) is the integral equation formulation of the Ito random differential equation

$$dx(t,\omega) = a(t, x(t,\omega))\,dt + b(t, x(t,\omega))\,dW(t) \tag{1.2}$$

with initial condition $x(u,\omega) = x_u(\omega)$. Under fairly general conditions on $a(.,.)$ and $b(.,.)$ unique and sample pathwise continuous solutions are known to exist both for scalar and vector versions of the Ito equations [4], [6], [8]. The sample pathwise solutions of a small class of Ito equations have been expressed as infinite series [2], [3]. Linear and nonlinear equations have been studied in [1], [5] by solving for the exact or approximate density functions from the Fokker–Planck equations. Either way analytical solutions are difficult to come by and generally are not in a readily usable form. An approximate algorithm will be very useful in the study of Ito equations as well as in solving stochastic control problems.

2. Discretization of the random integral equation. The functions $a(\cdot,\cdot)$ and $b(.,.)$ in the scalar random integral equation (1.1) are explicitly known and assumed to have continuous partial derivatives with respect to t and x at least up to

* Received by the editors December 6, 1971, and in final revised form November 29, 1972.

† Department of Electrical Engineering, Indian Institute of Technology, Kanpur, 208016, U.P., India. Now at School of Automation, Indian Institute of Science, Bangalore, India.

‡ Department of Mathematics, Indian Institute of Technology, Kanpur, 208016, U.P., India.

§ Department of Chemical Engineering, Indian Institute of Technology, Kanpur, 208016, U.P., India.

the third order in the interval of interest. Let T be a time interval $[u, v]$, $0 \leqq u < v < \infty$, and let $\{W(t), t \in T\}$, also written as $\{W_t\}$, be a standard Wiener process, i.e., with variance parameter unity. The integrals involved are assumed to exist in Ito's sense [6]. Divide the interval T, over which the integral equation is defined, into smaller intervals of duration h, i.e., $u = t_1 < t_2 < \cdots < t_N = v$, where $(t_{i+1} - t_i) = h$. Now one may write

$$x(t_{n+1}) = x(t_n) + \int_{t_n}^{t_{n+1}} a(t, x)\, dt + \int_{t_n}^{t_{n+1}} b(t, x)\, dW(t), \qquad n = 1, 2, \cdots, N-1. \tag{2.1}$$

The first strategy for the development of the algorithm consists in replacing the integrals in (2.1) by their Taylor series expansions about (t_n, x_n) where $x_n = x(t_n)$. Terms involving $(x - x_n)$ which arise in the above expansions are successively replaced by the following expression (analogous to (2.1)):

$$x(t) = x(t_n) + \int_{t_n}^{t} a(t', x)\, dt' + \int_{t_n}^{t} b(t', x)\, dW(t'). \tag{2.1a}$$

The number of successive substitutions would be governed by the desired order of error arising from truncation. Thus, for example, suppose one replaces $\int_{t_n}^{t_{n+1}} a(t, x)\, dt$ in (2.1) by $(t_{n+1} - t_n)a(t_n, x_n)$ and $\int_{t_n}^{t_{n+1}} b(t, x)\, dW(t)$ by $b(t_n, x_n)\int_{t_n}^{t_{n+1}} dW(t)$. This would *appear* to be the stochastic analogue of Euler's scheme for the corresponding deterministic case, giving an error of order[1] $o(h)$ and $o_p(h)$.

However, it can be easily shown that the term

$$b(t_n, x_n)\frac{\partial b}{\partial x}(t_n, x_n) \int_{t_n}^{t_{n+1}} \int_{t_n}^{t} dW(s)\, dW(t),$$

arising from the substitution of $(x - x_n)$ in

$$\int_{t_n}^{t_{n+1}} \frac{\partial b}{\partial x}(t_n, x_n)(x - x_n)\, dW(t),$$

is of order $O_p(h)$ and cannot be neglected for an error of $o_p(h)$.

The present paper develops an algorithm for an error of order $o(h^2)$ and $o_p(h^2)$. The development of higher order schemes involves evaluation of stochastic integrals of an increasingly complex nature and the difficulty in evaluating them does not appear to be commensurate with the gain in accuracy.

The following notation is followed in this paper:

$$a_n = a(t_n, x_n), \qquad b_{x_n} = \frac{\partial b}{\partial x}(t_n, x_n),$$

$$a_{t_n} = \frac{\partial a}{\partial t}(t_n, x_n), \qquad b_{t_n} = \frac{\partial b}{\partial t}(t_n, x_n), \quad \text{etc.}$$

[1] $o(h^2)$ means that $\lim_{h \to 0} (1/h^2)o(h^2) = 0$ and x is $o_p(h^2)$ if $\lim_{h \to 0} (1/h^2)P[|x| > \varepsilon] = 0$ for every $\varepsilon > 0$.

We consider terms up to second order in Taylor's expansion of $a(t, x)$ and up to third order terms in Taylor's expansion of $b(t, x)$. We have

$$\begin{aligned} a(t,x) = {} & a_n + (t - t_n)a_{t_n} + (x_t - x_n)a_{x_n} + \tfrac{1}{2}(t - t_n)^2 a_{t_n t_n} \\ & + (t - t_n)(x_t - x_n)a_{t_n x_n} + \tfrac{1}{2}(x_t - x_n)^2 a_{x_n x_n} \\ & + o(|t - t_n|^2) + o(|x_t - x_n|^2), \end{aligned} \tag{2.2}$$

and

$$\begin{aligned} b(t,x) = {} & b_n + (t - t_n)b_{t_n} + (x_t - x_n)b_{x_n} + \tfrac{1}{2}(t - t_n)^2 b_{t_n t_n} \\ & + (t - t_n)(x_t - x_n)b_{t_n x_n} + \tfrac{1}{2}(x_t - x_n)^2 b_{x_n x_n} \\ & + \frac{1}{3!}(t - t_n)^3 b_{t_n t_n t_n} + \tfrac{1}{2}(t - t_n)^2(x_t - x_n)b_{t_n t_n x_n} \\ & + \tfrac{1}{2}(t - t_n)(x_t - x_n)^2 b_{t_n x_n x_n} + \frac{1}{3!}(x_t - x_n)^3 b_{x_n x_n x_n} \\ & + o(|t - t_n|^3) + o(|x_t - x_n|^3). \end{aligned} \tag{2.3}$$

It can be shown that [7]

$$\int_{t_n}^{t_{n+1}} [o(|t - t_n|^2) + o(|x_t - x_n|^2)]\, dt = o(h^2) + o_p(h^2)$$

and

$$\int_{t_n}^{t_{n+1}} [o(|t - t_n|^3) + o(|x_t - x_n|^3)]\, dW(t) = o_p(h^2).$$

The function $b(x, t)$ is expanded up to third order terms unlike $a(x, t)$ because

$$\int_{t_n}^{t_n+h} [o(|t - t_n|^2) + o(|x_t - x_n|)^2]\, dW(t)$$

is only $o_p(h^{3/2})$ and not $o_p(h^2)$. Now equation (2.1) may be written as

$$\begin{aligned} x_{n+1} = x_n + {} & \int_{t_n}^{t_{n+1}} [a_n + (t - t_n)a_{t_n} + (x_t - x_n)a_{x_n} + \tfrac{1}{2}(t - t_n)^2 a_{t_n t_n} \\ & + (t - t_n)(x_t - x_n)a_{t_n x_n} + \tfrac{1}{2}(x_t - x_n)^2 a_{x_n x_n}]\, dt \\ & + \int_{t_n}^{t_{n+1}} \Big[b_n + (t - t_n)b_{t_n} + (x_t - x_n)b_{x_n} + \tfrac{1}{2}(t - t_n)^2 b_{t_n t_n} \\ & + (t - t_n)(x_t - x_n)b_{t_n x_n} + \tfrac{1}{2}(x_t - x_n)^2 b_{x_n x_n} \\ & + \frac{1}{3!}(t - t_n)^3 b_{t_n t_n t_n} + \tfrac{1}{2}(t - t_n)^2(x_t - x_n)b_{t_n t_n x_n} \\ & + \tfrac{1}{2}(t - t_n)(x_t - x_n)^2 b_{t_n x_n x_n} + \frac{1}{3!}(x_t - x_n)^3 b_{x_n x_n x_n} \Big]\, dW(t) \\ & + o(h^2) + o_p(h^2). \end{aligned} \tag{2.4}$$

The lower limit of integration is t_n and the upper limit is t_{n+1} in the rest of the paper unless otherwise specified and as such will be omitted. The terms on the right-hand side of equation (2.4) can be evaluated. We have

$$\int a_n\,dt = a_n h, \tag{2.5}$$

$$\int (t - t_n)a_{t_n}\,dt = \tfrac{1}{2}a_{t_n}h^2, \tag{2.6}$$

$$\int (x_t - x_n)a_{x_n}\,dt = a_{x_n}\int (x_t - x_n)\,dt.$$

$(x_t - x_n)$ is expanded in the form of Taylor's series:

$$\begin{aligned}(x_t - x_n) = &\int^t [a_n + (s - t_n)a_{t_n} + (x_s - x_n)a_{x_n} + o(|s - t_n|) + o(|x_s - x_n|)]\,ds\\ &+ \int^t [b_n + (s - t_n)b_{t_n} + (x_s - x_n)b_{x_n} + o(|s - t_n|)\\ &+ o(|x_s - x_n|)]\,dW(s).\end{aligned}$$

$(x_s - x_n)$ is now expressed as $[a_n(s - t_n) + b_n(W_s - W_n)]$. Now one may write:

$$\begin{aligned}a_{x_n}\int (x_t - x_n)\,dt = a_{x_n}\Big[&a_n \int\int^t ds\,dt + (a_{t_n} + a_{x_n}a_n)\int\int^t (s - t_n)\,ds\,dt\\ &+ a_{x_n}b_n \int\int^t (W_s - W_n)\,ds\,dt + b_n\int\int^t dW(s)\,dt\\ &+ (b_{t_n} + b_{x_n}a_n)\int\int^t (s - t_n)\,dW(s)\,dt\\ &+ b_{x_n}b_n\int\int^t (W_s - W_n)\,dW(s)\,dt\Big] + o(h^2) + o_p(h^2);\end{aligned} \tag{2.7}$$

$$\int \tfrac{1}{2}(t - t_n)^2 a_{t_nt_n}\,dt = \frac{1}{3!}a_{t_nt_n}h^3; \tag{2.8}$$

$$\begin{aligned}&\int (t - t_n)(x_t - x_n)a_{t_nx_n}\,dt\\ &= a_{t_nx_n}\int (t - t_n)[a_n(t - t_n) + b_n(W_t - W_n)]\,dt + o(h^2) + o_p(h^2)\\ &= a_{t_nx_n}\left[a_n \int (t - t_n)^2\,dt + b_n \int (t - t_n)(W_t - W_n)\,dt\right] + o(h^2) + o_p(h^2);\end{aligned} \tag{2.9}$$

$$\begin{aligned}
&\int \tfrac{1}{2}(x_t - x_n)^2 a_{x_n x_n}\,dt \\
&\quad = \tfrac{1}{2}a_{x_n x_n}\int [a_n(t - t_n) + b_n(W_t - W_n)]^2\,dt + o(h^2) + o_p(h^2) \\
&\quad = \tfrac{1}{2}a_{x_n x_n}\left[a_n^2\int (t - t_n)^2\,dt + 2a_n b_n\int (t - t_n)(W_t - W_n)\,dt\right. \\
&\qquad\qquad \left. + b_n^2\int (W_t - W_n)^2\,dt\right] + o(h^2) + o_p(h^2);
\end{aligned} \tag{2.10}$$

$$b_n\int dW(t) = b_n(W_t - W_n); \tag{2.11}$$

$$\int (t - t_n)b_{t_n}\,dW(t) = b_{t_n}\int (t - t_n)\,dW(t); \tag{2.12}$$

$$\begin{aligned}
&\int (x_t - x_n)b_{x_n}\,dW(t) \\
&\quad = b_{x_n}\int\int^t [a_n + (s - t_n)a_{t_n} + (x_s - x_n)a_{x_n} + o(|s - t_n|) + o(|x_s - x_n|)]\,ds\,dW(t) \\
&\qquad + b_{x_n}\int\int^t [b_n + (s - t_n)b_{t_n} + (x_s - x_n)b_{x_n} + \tfrac{1}{2}(s - t_n)^2 b_{t_n t_n} \\
&\qquad + (s - t_n)(x_s - x_n)b_{t_n x_n} + \tfrac{1}{2}(x_s - x_n)^2 b_{x_n x_n} + o(|s - t_n|^2) \\
&\qquad + o(|x_s - x_n|^2)]\,dW(s)\,dW(t) \\
&\quad = b_{x_n}\int\int^t [a_n + (s - t_n)a_{t_n} + a_{x_n}a_n(s - t_n) + a_{x_n}b_n(W_s - W_n)]\,ds\,dW(t) \\
&\qquad + b_{x_n}\int\int^t \Big[b_n + (s - t_n)b_{t_n} + b_{x_n}a_n(s - t_n) + b_{x_n}b_n(W_s - W_n) \\
&\qquad + b_{x_n}b_{t_n}\int^s (u - t_n)\,dW(u) + b_{x_n}b_{x_n}a_n\int^s (u - t_n)\,dW(u) \\
&\qquad + b_{x_n}b_{x_n}b_n\int^s (W_u - W_n)\,dW(u) + \tfrac{1}{2}b_{t_n t_n}(s - t_n)^2 \\
&\qquad + b_{t_n x_n}a_n(s - t_n)^2 + b_{t_n x_n}b_n(s - t_n)(W_s - W_n) \\
&\qquad + \tfrac{1}{2}b_{x_n x_n}a_n^2(s - t_n)^2 + b_{x_n x_n}a_n b_n(s - t_n)(W_s - W_n) \\
&\qquad + \tfrac{1}{2}b_{x_n x_n}b_n^2(W_s - W_n)^2\Big]\,dW(s)\,dW(t) + o_p(h^2) \\
&\quad = b_{x_n}\Big[a_n\int\int^t ds\,dW(t) + (a_{t_n} + a_{x_n}a_n)\int\int^t (s - t_n)\,ds\,dW(t)
\end{aligned} \tag{2.13}$$

(cont.)

$$+ a_{x_n}b_n \int\int^t (W_s - W_n)\,ds\,dW(t) + b_n \int\int^t dW(s)\,dW(t)$$

$$+ (b_{t_n} + b_{x_n}a_n) \int\int^t (s - t_n)\,dW(s)\,dW(t)$$

$$+ b_{x_n}b_n \int\int^t (W_s - W_n)\,dW(s)\,dW(t)$$

$$+ b_{x_n}(b_{t_n} + b_{x_n}a_n) \int\int^t\int^s (u - t_n)\,dW(u)\,dW(s)\,dW(t)$$

$$+ b_{x_n}b_{x_n}b_n \int\int^t\int^s (W_u - W_n)\,dW(u)\,dW(s)\,dW(t)$$

$$+ \tfrac{1}{2}(b_{t_nt_n} + 2b_{t_nx_n}a_n + b_{x_nx_n}a_n^2) \int\int^t (s - t_n)^2\,dW(s)\,dW(t)$$

$$+ (b_{t_nx_n}b_n + b_{x_nx_n}a_nb_n) \int\int^t (s - t_n)(W_s - W_n)\,dW(s)\,dW(t)$$

$$+ \tfrac{1}{2}b_{x_nx_n}b_n^2 \int\int^t (W_s - W_n)^2\,dW(s)\,dW(t)\Big] + o_p(h^2);$$

$$\text{(2.14)} \qquad \int \tfrac{1}{2}(t - t_n)^2 b_{t_nt_n}\,dW(t) = \tfrac{1}{2}b_{t_nt_n} \int (t - t_n)^2\,dW(t);$$

$$\int (t - t_n)(x_t - x_n)b_{t_nx_n}\,dW(t)$$

$$\text{(2.15)} \quad = b_{t_nx_n} \int (t - t_n)[a_n(t - t_n) + b_n(W_t - W_n)]\,dW(t) + o_p(h^2)$$

$$= b_{t_nx_n}\Big[a_n \int (t - t_n)^2\,dW(t) + b_n \int (t - t_n)(W_t - W_n)\,dW(t)\Big] + o_p(h^2);$$

$$\int \tfrac{1}{2}(x_t - x_n)^2 b_{x_nx_n}\,dW(t)$$

$$= \tfrac{1}{2}b_{x_nx_n} \int \Big[a_n(t - t_n) + b_n(W_t - W_n) + b_{t_n} \int^t (s - t_n)\,dW(s)$$

$$+ b_{x_n}a_n \int^t (s - t_n)\,dW(s) + b_{x_n}b_n \int^t (W_s - W_n)\,dW(s)\Big]^2 dW(t) + o_p(h^2)$$

$$= \tfrac{1}{2}b_{x_nx_n}\Big[a_n^2 \int (t - t_n)^2\,dW(t) + b_n^2 \int (W_t - W_n)^2\,dW(t)$$

$$+ 2a_nb_n \int (t - t_n)(W_t - W_n)\,dW(t)$$

$$+ (b_{t_n} + b_{x_n}a_n)^2 \int \Big\{\int^t (s - t_n)\,dW(s)\Big\}^2 dW(t)$$

(cont.)

$$\begin{aligned}
&+ b_{x_n}^2 b_n^2 \int \left\{ \int^t (W_s - W_n)\, dW(s) \right\}^2 dW(t) \\
&+ 2a_n(b_{t_n} + b_{x_n}a_n) \int (t - t_n) \int^t (s - t_n)\, dW(s)\, dW(t) \\
&+ 2b_n(b_{t_n} + b_{x_n}a_n) \int (W_t - W_n) \int^t (s - t_n)\, dW(s)\, dW(t) \\
&+ 2b_{x_n}b_n(b_{t_n} + b_{x_n}a_n) \int \left\{ \int^t (W_s - W_n)\, dW(s) \int^t (s - t_n)\, dW(s) \right\} dW(t) \\
&+ 2a_n b_{x_n} b_n \int (t - t_n) \int^t (W_s - W_n)\, dW(s)\, dW(t) \\
&+ 2b_n b_{x_n} b_n \int (W_t - W_n) \int^t (W_s - W_n)\, dW(s)\, dW(t) \Bigg] + o_p(h^2);
\end{aligned} \tag{2.16}$$

$$\frac{1}{3!} \int (t - t_n)^3 b_{t_n t_n t_n}\, dW(t) = \frac{1}{3!} b_{t_n t_n t_n} \int (t - t_n)^3\, dW(t); \tag{2.17}$$

$$\begin{aligned}
&\frac{1}{2} \int (t - t_n)^2 (x_t - x_n) b_{t_n t_n x_n}\, dW(t) \\
&= \tfrac{1}{2} b_{t_n t_n x_n} \left[a_n \int (t - t_n)^3\, dW(t) + b_n \int (t - t_n)^2 (W_t - W_n)\, dW(t) \right] + o_p(h^2);
\end{aligned} \tag{2.18}$$

$$\begin{aligned}
&\frac{1}{2} \int (t - t_n)(x_t - x_n)^2 b_{t_n x_n x_n}\, dW(t) \\
&= \tfrac{1}{2} b_{t_n x_n x_n} \left[a_n^2 \int (t - t_n)^3\, dW(t) + 2a_n b_n \int (t - t_n)^2 (W_t - W_n)\, dW(t) \right. \\
&\qquad \left. + b_n^2 \int (t - t_n)(W_t - W_n)^2\, dW(t) \right] + o_p(h^2);
\end{aligned} \tag{2.19}$$

$$\begin{aligned}
&\frac{1}{3!} \int (x_t - x_n)^3 b_{x_n x_n x_n}\, dW(t) \\
&= \frac{1}{3!} b_{x_n x_n x_n} \left[a_n^3 \int (t - t_n)^3\, dW(t) + 3a_n^2 b_n \int (t - t_n)^2 (W_t - W_n)\, dW(t) \right. \\
&\qquad \left. + 3a_n b_n^2 \int (t - t_n)(W_t - W_n)^2\, dW(t) + b_n^3 \int (W_t - W_n)^3\, dW(t) \right] + o_p(h^2).
\end{aligned} \tag{2.20}$$

All the integrals that are encountered in equations (2.5)–(2.20) are of Wiener type or of Ito type. So each one of them gives rise to a random variable. They are defined as follows:

$$Z_{1n} = \int dW(t),$$

$$Z_{2n} = \int\int^{t} dW(s)\,dt,$$

$$Z_{3n} = \int\int^{t} (W_s - W_n)\,ds\,dW(t),$$

$$Z_{4n} = \int (W_t - W_n)\,dW(t),$$

$$Z_{5n} = \int (t - t_n)\,dW(t),$$

$$Z_{6n} = \int (W_t - W_n)^2\,dt,$$

$$Z_{7n} = \int (W_t - W_n)^2\,dW(t),$$

$$Z_{8n} = \int (t - t_n)(W_t - W_n)\,dW(t),$$

$$Z_{9n} = \int\int^{t} (s - t_n)\,dW(s)\,dW(t),$$

$$Z_{10n} = \int (W_t - W_n)^3\,dW(t),$$

$$Z_{11n} = \int\int^{t} (W_s - W_n)^2\,dW(s)\,dW(t),$$

$$Z_{12n} = \int\int^{t} (W_s - W_n)\,dW(s)\,dt,$$

$$Z_{13n} = \int\int^{t} (W_s - W_n)\,dW(s)\,dW(t),$$

$$Z_{14n} = \int\int^{t}\int^{s} (W_u - W_n)\,dW(u)\,dW(s)\,dW(t),$$

$$Z_{15n} = \int (W_t - W_n)\int^{t} (W_s - W_n)\,dW(s)\,dW(t),$$

$$Z_{16n} = \int (t - t_n)^2\,dW(t),$$

$$Z_{17n} = \int\int^{t} (s - t_n)\,dW(s)\,dt,$$

$$Z_{18n} = \int\int^t (W_s - W_n)\,ds\,dt,$$

$$Z_{19n} = \int (t - t_n)(W_t - W_n)\,dt,$$

$$Z_{20n} = \int\int^t\int^s (u - t_n)\,dW(u)\,dW(s)\,dW(t),$$

$$Z_{21n} = \int\int^t (s - t_n)^2\,dW(s)\,dW(t),$$

$$Z_{22n} = \int\int^t (s - t_n)(W_s - W_n)\,dW(s)\,dW(t),$$

$$Z_{23n} = \int\int^t (t - t_n)(s - t_n)\,dW(s)\,dW(t),$$

$$Z_{24n} = \int\int^t (t - t_n)(W_s - W_n)\,dW(s)\,dW(t),$$

$$Z_{25n} = \int\left[\int^t (s - t_n)\,dW(s)\right]^2 dW(t),$$

$$Z_{26n} = \int\left[\int^t (W_s - W_n)\,dW(s)\right]^2 dW(t),$$

$$Z_{27n} = \int\left[\int^t (s - t_n)\,dW(s)\right]\left[\int^t (W_u - W_n)\,dW(u)\right] dW(t),$$

$$Z_{28n} = \int\int^t (W_t - W_n)(s - t_n)\,dW(s)\,dW(t),$$

$$Z_{29n} = \int (t - t_n)^3\,dW(t),$$

$$Z_{30n} = \int (t - t_n)^2(W_t - W_n)\,dW(t),$$

$$Z_{31n} = \int (t - t_n)(W_t - W_n)^2\,dW(t).$$

The properties of these random variables and their relationships are studied in detail in the Appendix. One may find from these relationships that $Z_{1n}, Z_{2n}, Z_{16n}, Z_{17n}, Z_{18n}, Z_{19n}$ and Z_{29n} are dependent normal variables and Z_{3n} is uncorrelated to Z_{1n} and Z_{2n} and it may be approximated by a normal variable. The random variables Z_{4n}–Z_{15n} can be expressed in terms of Z_{1n}, Z_{2n} and Z_{3n}. The random variables Z_{16n}–Z_{31n} are $o_p(h^2)$. One may also observe that Z_{ijs} $(i = 1, 2, \cdots, 31)$ are independent for different values of $j, j = 1, 2, \cdots, N - 1$.

3. Algorithm. For a second order algorithm one has to account for deterministic terms which are bigger than $o(h^2)$ and random variables larger than

$o_p(h^2)$. As a result the terms involving $h^3/3!, Z_{16}$—Z_{31n} can be neglected. Regrouping the remaining terms, one gets

$$\begin{aligned} x_{n+1} = {} & x_n + a_n h + b_n Z_{1n} + \frac{h^2}{2}(a_{t_n} + a_{x_n} a_n) + Z_{2n}(a_{x_n} b_n) + Z_{3n}(b_{x_n} a_{x_n} b_n) \\ & + Z_{4n}(b_{x_n} b_n) + Z_{5n}(b_{t_n} + b_{x_n} a_n) + \tfrac{1}{2} Z_{6n}(a_{x_n x_n} b_n^2) + \tfrac{1}{2} Z_{7n}(b_{x_n x_n} b_n^2) \\ & + Z_{8n}(b_{t_n x_n} b_n + b_{x_n x_n} b_n a_n) + Z_{9n}(b_{x_n}^2 a_n + b_{x_n} b_{t_n}) \\ & + \frac{1}{3!} Z_{10n}(b_{x_n x_n x_n} b_n^3) + \tfrac{1}{2} Z_{11n}(b_{x_n} b_{x_n x_n} b_n^2) + Z_{12n}(a_{x_n} b_{x_n} b_n) \\ & + Z_{13n}(b_{x_n}^2 b_n) + Z_{14n}(b_{x_n}^3 b_n) + Z_{15n}(b_{x_n x_n} b_{x_n} b_n^2). \end{aligned} \tag{3.1}$$

By making use of the relationships (A.6)–(A.17) derived in the Appendix, the algorithm may now be written as

$$\begin{aligned} x_{n+1} = {} & x_n + a_n h + b_n Z_{1n} + \tfrac{1}{2} h^2 (a_{t_n} + a_{x_n} a_n - \tfrac{1}{2} a_{x_n} b_{x_n} b_n) \\ & + Z_{2n}(a_{x_n} b_n - b_{t_n} - b_{x_n} a_n + \tfrac{1}{2} b_{x_n}^2 b_n) + \tfrac{1}{2}(Z_{1n}^2 - h) b_{x_n} b_n \\ & + Z_{1n} h (b_{t_n} + b_{x_n} a_n - \tfrac{1}{2} b_{x_n}^2 b_n) + Z_{3n} b_{x_n} a_{x_n} b_n \\ & + \tfrac{1}{2}(Z_{1n} Z_{2n} - Z_{3n})(a_{x_n} b_{x_n} b_n + a_{x_n x_n} b_n^2) \\ & + \tfrac{1}{2}(\tfrac{1}{3} Z_{1n}^3 - Z_{2n})(b_{x_n x_n} b_n^2 + b_{x_n}^2 b_n) \\ & + \tfrac{1}{2}(Z_{1n}^2 h - \tfrac{1}{2} h^2 + Z_{3n} - Z_{1n} Z_{2n})(b_{t_n x_n} b_n + b_{x_n x_n} a_n b_n - \tfrac{1}{2} b_{x_n x_n} b_{x_n} b_n^2) \\ & + \tfrac{1}{2}(Z_{1n}^2 h - \tfrac{1}{2} h^2 - Z_{3n} - Z_{1n} Z_{2n})(b_{x_n}^2 a_n + b_{x_n} b_{t_n} - \tfrac{1}{2} b_{x_n}^3 b_n) \\ & + \tfrac{1}{2}(\tfrac{1}{2} Z_{1n}^4 - 3 Z_{1n} Z_{2n} + 3 Z_{3n}) \left(\frac{1}{3!} b_{x_n x_n x_n} b_n^3 + \tfrac{1}{2} b_{x_n x_n} b_{x_n} b_n^2 \right) \\ & + \tfrac{1}{2}(\tfrac{1}{6} Z_{1n}^4 - Z_{1n} Z_{2n} - Z_{3n})(\tfrac{1}{2} b_{x_n} b_{x_n x_n} b_n^2 + \tfrac{1}{2} b_{x_n}^3 b_n). \end{aligned} \tag{3.2}$$

In this algorithm only three random variables Z_{1n}, Z_{2n} and Z_{3n} appear. Z_{1n} and Z_{2n} are dependent normal variables and Z_{3n} is approximately a normal variable. Now the algorithm given by (3.2) can be very easily implemented on a computer by simulating $Z_{1n}, Z_{2n}, Z_{3n}, n = 1, 2, \cdots, N - 1$.

4. Example. Consider an Ito integral equation given by

$$x(t) = x(t_0) + \int_{t_0}^{t} Cx\, ds + \int_{t_0}^{t} Dx\, dW(s), \tag{4.1}$$

where C and D are constants, $W(t)$ is a standard Wiener process and the interval of definition is $[t_0, t_f]$. It is well known that the process $\{x_t\}$ is Markovian and that the transition density function, $p(x, t/x_0, t_0)$, satisfies the Fokker–Planck equation

$$\frac{\partial p}{\partial t} = -\frac{\partial(Cxp)}{\partial x} + \frac{1}{2}\frac{\partial^2(D^2 x^2 p)}{\partial x^2}. \tag{4.2}$$

The equation can be solved either for the density function or for the moments. If $m_1(t)$ and $m_2(t)$ are the first and second moments respectively, then

$$m_1(t) = E(x_0)\exp(Ct), \tag{4.3}$$

$$m_2(t) = E(x_0^2)\exp(2C + D^2)t. \tag{4.4}$$

The algorithm given by (3.2) gives a sample pathwise solution at discrete points in the interval of definition. The algorithm for this example is given by

$$\begin{aligned} x_{n+1} = x_n[1 &+ Ch + Z_{1n}D + \tfrac{1}{2}h^2(C^2 - \tfrac{1}{2}CD^2) + \tfrac{1}{2}Z_{2n}D^3 + Z_{1n}h(CD - \tfrac{1}{2}D^3) \\ &+ Z_{3n}CD^2 + \tfrac{1}{2}(Z_{1n}^2 - h)D^2 + \tfrac{1}{2}(Z_{1n}Z_{2n} - Z_{3n})CD^2 \\ &+ \tfrac{1}{4}D^3(\tfrac{1}{3}Z_{1n}^3 - Z_{2n}) + \tfrac{1}{2}(CD^2 - \tfrac{1}{2}D^4)(Z_{1n}^2h - \tfrac{1}{2}h^2 - Z_{3n} - Z_{1n}Z_{2n}) \\ &+ \tfrac{1}{4}(\tfrac{1}{6}Z_{1n}^4 - Z_{3n} - Z_{1n}Z_{2n})D^4]. \end{aligned}$$

This equation can be repeatedly solved to produce different sample paths each time by getting different realizations of x_1, Z_{1n}, Z_{2n} and $Z_{3n}, n = 1, 2, \cdots, N - 1$. From these samples paths the moments can be easily computed. The equation given by (4.1) is solved with the following values:

$$[t_0, t_f] = 0.3 \text{ secs.}, \quad C = -2.0; \quad D = 1.0, \quad x_1(w) = 10.0, \quad h = 0.005.$$

Moments are computed with 100 samples. If σ_n^2 is the variance of x_n, then the sample mean should lie between $m_1(t_n) + 3\sigma_n/\sqrt{100}$. The computational results are shown in Figs. 1 and 2.

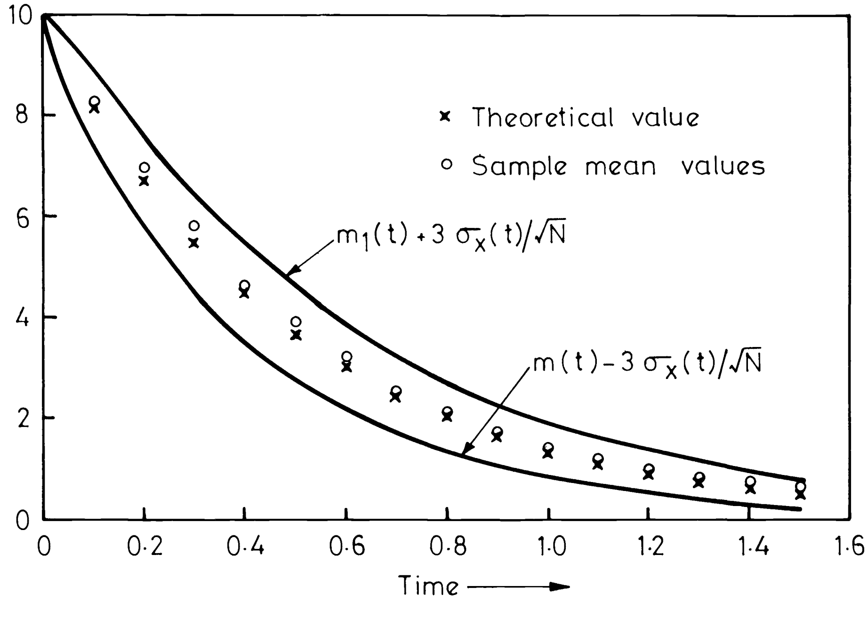

FIG. 1

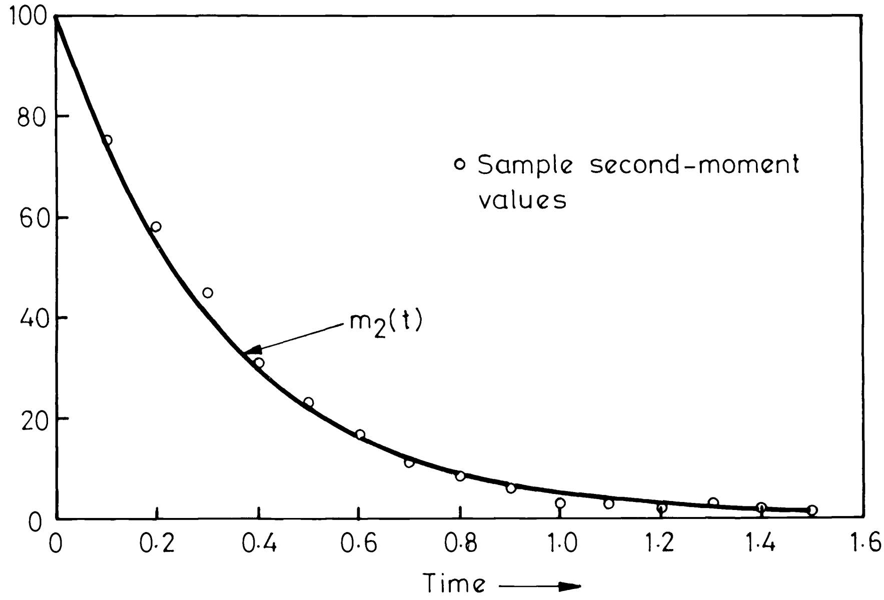

FIG. 2

5. Conclusions. The algorithm derived in this paper is readily implementable on a digital computer. A large class of random integral equations, including the equations where $a(.,.)$ and $b(.,.)$ are nonlinear in t and x, can be studied. The extension of this algorithm to the vector Ito equation, where x_t is a vector but (W_t) is a scalar, is direct. Extension to the vector case, where W_t is also a vector, is not immediate.

Appendix. Let $(\Omega, P, \mathscr{B})$ be the probability space, let $\mathscr{B}^t$, $t \in T$, be a monotone increasing system of Borel subalgebras such that $\mathscr{B}^t$ includes all null sets for each t, and let $W_t \in (\mathscr{B}^t)$ and $(W_{t+h} - W_t)$ be independent of $\mathscr{B}^t$. $f_t \in (\mathscr{B}^t)$ indicates that f_t is $\mathscr{B}^t$-measurable. Let $\mathscr{L}_s$ be the set of all functions, f, such that (i) f is measurable in (t, w), (ii) $f_t \in (\mathscr{B}^t)$ for almost all $t \in T$ and (iii) $\int_T f_t^2\, dt < \infty$ for almost all $w \in \Omega$, where T is the interval over which f_t is defined. The following lemmas are used to derive some of the relationships between the random variables defined earlier.

LEMMA 1. *If $f, g \in \mathscr{L}_s$, then*

$$\left(\int_t^v f_s\, dW_s\right)\left(\int_t^v g_s\, dW_s\right) = \int_t^v f_s G_s\, dW_s + \int_t^v g_s F_s\, dW_s + \int_t^v f_s g_s\, ds,$$

where $F_s = \int_t^s f_\theta\, dW_\theta$ and $G_s = \int_t^s g_\theta\, dW_\theta$.

LEMMA 2. *If $f, g \in \mathscr{L}_s$, then*

$$\left(\int_t^v f_s\, dW_s\right)\left(\int_t^v g_s\, ds\right) = \int_t^v f_s G_s\, dW_s + \int_t^v g_s F_s\, ds,$$

where $F_s = \int_t^s f_\theta\, dW_\theta$ and $g_s = \int_t^s g_\theta\, d\theta$.

Proofs of these lemmas are given in [6]. In what follows the upper limit of integration is t_{n+1} and the lower limit is t_n unless otherwise specified. Consider

$$Z_{1n} = \int dW(t) = W_{t_{n+1}} - W_{t_n}.$$

Z_{1n} is a normal random variable, by definition of the standard Wiener process, with zero mean and variance h:

$$Z_{1n} \overset{d}{=} N(0, h), \tag{A.1}$$

where d means the left-hand side has the distribution specified on the right-hand side. Here $N(0, h)$ represents a normal distribution with mean zero and variance h.

$$Z_{2n} = \int\int^t dW(s)\,dt = \int (W_t - W_{t_n})\,dt.$$

By the definition of the Ito integral Z_{2n} can be expressed as the limit in the mean of linear combinations of random variables which are jointly normal and as such Z_{2n} is normal:

$$Z_{2n} \overset{d}{=} N(0, h^3/3). \tag{A.2}$$

Also one may observe that

$$E[Z_{1n}Z_{2n}] = h/2, \tag{A.3}$$

$$Z_{3n} = \int\int^t \int^s dW(u)\,ds\,dW(t) = \int Z_{2n}(t)\,dW(t).$$

Evidently Z_{3n} is not normal, but its moments can be computed:

$$E[Z_{3n}] = 0; \qquad E[Z_{3n}^2] = \frac{h^4}{12}; \qquad E[Z_{3n}^4] = \frac{61}{1680}h^8;$$

$$E[Z_{3n}^3] = E[Z_{3n}^5] = E[Z_{3n}]^{2k+1} = 0.$$

Its characteristic function may be written as

$$\phi_{Z_{3n}}(t) = 1 - \frac{h^4}{12}\frac{t^2}{2!} + \frac{61}{1680}h^8\frac{t^4}{4!} - \cdots.$$

Consider a normal random variable Y with zero mean and variance $h^4/12$. The characteristic function of Y is given by

$$\phi_Y(t) = 1 - \frac{h^4}{12}\cdot\frac{t^2}{2!} + \frac{3h^8}{144}\cdot\frac{t^4}{4!} - \cdots,$$

$$\phi_{Z_{3n}}(t) - \phi_Y(t) \doteq \frac{h^8}{28}\frac{t^4}{4!} + o(h^8).$$

Therefore for small values of h one can approximate Z_{3n} by a normal variable with zero mean and $h^4/12$ variance:

$$Z_{3n} \overset{d}{=} N(0, h^4/12). \tag{A.4}$$

It may also be noted that

$$E[Z_{1n}Z_{3n}] = 0; \qquad E[Z_{2n}Z_{3n}] = 0. \tag{A.5}$$

The random variables Z_{4n} to Z_{15n} can be evaluated using Lemma 1 and Lemma 2.

1. Let $f(t) = 1$ and $g(t) = 1$ in Lemma 1,

$$\int dW(t) \int dW(t) = \int\int^t dW(s)\,dW(t) + \int\int^t dW(s)\,dW(t) + \int dt,$$

$$Z_{1n} \cdot Z_{1n} = Z_{4n} + Z_{4n} + h,$$

$$Z_{4n} = \tfrac{1}{2}(Z_{1n}^2 - h). \tag{A.6}$$

2. Let $f(t) = 1$ and $g(t) = 1$ in Lemma 2,

$$\int dW(t) \int dt = \int\int^t ds\,dW(t) + \int\int^t dW(s)\,dt,$$

$$Z_{1n} \cdot h = Z_{5n} + Z_{2n},$$

$$Z_{5n} = (Z_{1n}h - Z_{2n}). \tag{A.7}$$

3. Let $f(t) = 1$ and $g(t) = (W_t - W_n)$ in Lemma 2,

$$\int dW(t) \int (W_t - W_n)\,dt = \int\int^t (W_s - W_n)\,ds\,dW(t) + \int (W_t - W_n)^2\,dt,$$

$$Z_{1n}Z_{2n} = Z_{3n} + Z_{6n},$$

$$Z_{6n} = (Z_{1n}Z_{2n} - Z_{3n}). \tag{A.8}$$

4. Let $f(t) = 1$ and $g(t) = (W_t - W_n)$ in Lemma 1,

$$\int dW(t) \int (W_t - W_n)\,dW(t) = \frac{1}{2}\int \{(W_t - W_n)^2 - (t - t_n)\}\,dW(t)$$

$$+ \int (W_t - W_n)^2\,dW(t) + \int (W_t - W_n)\,dt,$$

$$Z_{1n}\tfrac{1}{2}(Z_{1n}^2 - h) = \tfrac{3}{2}Z_{7n} - \tfrac{1}{2}Z_{5n} + Z_{2n},$$

$$Z_{7n} = (\tfrac{1}{3}Z_{1n}^3 - Z_{2n}). \tag{A.9}$$

5. Let $f(t) = (W_t - W_n)$ and $g(t) = 1$ in Lemma 2,

$$\int (W_t - W_n)\,dW(t) \int dt = \int (t - t_n)(W_t - W_n)\,dW(t)$$

$$+ \frac{1}{2}\int [(W_t - W_n)^2 - (t - t_n)]\,dt,$$

$$\tfrac{1}{2}(Z_{1n}^2 - h)h = Z_{8n} + \tfrac{1}{2}Z_{6n} - \tfrac{1}{4}h^2,$$

$$Z_{8n} = \tfrac{1}{2}(Z_{1n}^2h - \tfrac{1}{2}h^2 + Z_{3n} - Z_{1n}Z_{2n}). \tag{A.10}$$

6. Let $f(t) = (t - t_n)$ and $g(t) = 1$ in Lemma 1,

$$\int (t - t_n)\, dW(t) \int dW(t) = \int (t - t_n)(W_t - W_n)\, dW(t) + \int\int^t (s - t_n)\, dW(s)\, dW(t) + \int (t - t_n)\, dt,$$

$$Z_{5n} Z_{1n} = Z_{8n} + Z_{9n} + h^2/2,$$

$$Z_{9n} = \tfrac{1}{2}(Z_{1n}^2 h - \tfrac{1}{2}h^2 - Z_{3n} - Z_{1n}Z_{2n}). \tag{A.11}$$

7. Let $f(t) = (W_t - W_n)$ and $g(t) = (W_t - W_n)$ in Lemma 1,

$$\int (W_t - W_n)\, dW(t) \int (W_t - W_n)\, dW(t) = 2 \int (W_t - W_n)\tfrac{1}{2}\{(W_t - W_n)^2 - (t - t_n)\}\, dW(t) + \int (W_t - W_n)^2\, dt,$$

$$\tfrac{1}{2}(Z_{1n}^2 - h)\tfrac{1}{2}(Z_{1n}^2 - h) = Z_{10n} - Z_{8n} + Z_{6n},$$

$$Z_{10n} = (\tfrac{1}{4}Z_{1n}^4 - \tfrac{3}{2}Z_{1n}Z_{2n} + \tfrac{3}{2}Z_{3n}). \tag{A.12}$$

8. Let $f(t) = (W_t - W_n)^2$ and $g(t) = 1$ in Lemma 1,

$$\int (W_t - W_n)^2\, dW(t) \int dW(t) = \int (W_t - W_n)^3\, dW(t) + \int\int^t (W_s - W_n)^2\, dW(s)\, dW(t) + \int (W_t - W_n)^2\, dt,$$

$$Z_{7n} Z_{1n} = Z_{10n} + Z_{11n} + Z_{6n},$$

$$Z_{11n} = \tfrac{1}{2}(\tfrac{1}{6}Z_{1n}^4 - Z_{1n}Z_{2n} - Z_{3n}), \tag{A.13}$$

$$\begin{aligned} Z_{12n} &= \int\int^t (W_s - W_n)\, dW(s)\, dt \\ &= \frac{1}{2}\int (W_t - W_n)^2\, dt - \frac{1}{2}\int (t - t_n)\, dt \\ &= \tfrac{1}{2}(Z_{1n}Z_{2n} - Z_{3n}) - h^2/4 \quad \text{(by A.8)}, \end{aligned} \tag{A.14}$$

$$\begin{aligned} Z_{13n} &= \int\int^t (W_s - W_n)\, dW(s)\, dW(t) \\ &= \frac{1}{2}\int (W_t - W_n)^2\, dW(t) - \int (t - t_n)\, dW(t) \\ &= \tfrac{1}{2}(\tfrac{1}{3}Z_{1n}^3 - Z_{2n}) - \tfrac{1}{2}(Z_{1n}h - Z_{2n}) \quad \text{(by A.9)}, \end{aligned} \tag{A.15}$$

$$Z_{14n} = \int\int^t \int^s (W_u - W_n)\,dW(u)\,dW(s)\,dW(t)$$

$$= \frac{1}{2}\int\int^t \{(W_s - W_n)^2 - (s - t_n)\}\,dW(s)\,dW(t)$$

$$= \tfrac{1}{2}(Z_{11n} - Z_{9n}), \tag{A.16}$$

$$Z_{15n} = \int (W_t - W_n) \int^t (W_s - W_n)\,dW(s)\,dW(t)$$

$$= \frac{1}{2}\int (W_t - W_n)^3\,dW(t) - \frac{1}{2}\int (t - t_n)(W_t - W_n)\,dW(t)$$

$$= \tfrac{1}{2}(Z_{10n} - Z_{8n}). \tag{A.17}$$

It can be easily seen from the Chebyshev inequality that a random variable Z will be $o_p(h^2)$ if the variance of Z is $o(h^4)$. Using this result it can be easily shown that the random variables Z_{16n} through Z_{31n} are $o_p(h^2)$.

Acknowledgment. The authors are highly indebted to Professor E. J. McShane for his valuable comments and for pointing out an error in an earlier version of the manuscript.

REFERENCES

[1] S. T. ARIARATNAM AND P. W. U. GRAEFE, *Linear systems with stochastic coefficients*, Internat. J. Control, 1 (1965), pp. 239–250; 2 (1965), pp. 161–169; 2 (1965), pp. 205–210.

[2] K. J. ASTROM, *On a first order stochastic differential equation*, Ibid., 1 (1965), pp. 301–326.

[3] F. J. BEUTLER, *Multivariate wide-sense Markov processes and prediction theory*, Ann. Math. Statist., 34 (1963), pp. 424–438.

[4] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[5] A. T. FULLER, *Analysis of nonlinear stochastic systems*, Internat. J. Control, 9 (1969), pp. 603–655.

[6] K. ITO, *Lectures on Stochastic Processes*, Tata Institute of Fundamental Research, Bombay, 1960.

[7] N. J. RAO, *Stochastic optimal control problems: An Algorithmic approach*, Ph.D. thesis, Indian Institute of Technology, Kanpur, India, 1972.

[8] A. V. SKOROKHOD, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Mass., 1965.

SIAM J. CONTROL
Vol. 12, No. 1, February 1974

NECESSARY AND SUFFICIENT OPTIMALITY CONDITIONS FOR THE FRITZ JOHN PROBLEM WITH LINEAR EQUALITY CONSTRAINTS*

KENNETH R. GEHNER†

Abstract. A generalization of Motzkin's theorem of the alternative is developed, and is then used to obtain necessary optimality conditions for the Fritz John optimization problem with linear equality constraints. Under certain constraint qualifications and convexity assumptions these necessary conditions are also sufficient. An application of these results to the problem of Chebyshev approximation with interpolation is given.

1. Introduction. Consider the following general optimization problem:

$$\text{(P)} \qquad \begin{aligned} &\underset{x}{\text{minimize}}\ F(x)\\ &\text{subject to}\\ &\text{(a)}\ G_i(x,t) \leqq 0 \quad \text{for all } t \in T_i, \qquad i = 1, \cdots, l,\\ &\text{(b)}\ H_j(x,s) = 0 \quad \text{for all } s \in S_j, \qquad j = 1, \cdots, m,\\ &\text{(c)}\ x \in X^0, \end{aligned}$$

where

(i) $F(x)$ and each $G_i(x,t)$ are real-valued functions which have continuous partial derivatives with respect to x for each $t \in T_i$, and each $G_i(x,t)$ is continuous in $t \in T_i$ for each $x \in X^0$;

(ii) each $H_j(x,s)$ is a real-valued linear function in x for each $s \in S_j$, and is continuous in $s \in S_j$ for each $x \in X^0$;

(iii) each T_i and S_j is a compact subset of a complete metric space;

(iv) X^0 is an open set in R^n.

Since the Fritz John problem [4] is obtained from (P) by deleting the constraints (b), we shall call (P) the Fritz John problem with linear equality constraints. Although problem (P) can be put into the form of the Fritz John problem by replacing each $H_j(x,s) = 0$ by $H_j(x,s) \leqq 0$ and $-H_j(x,s) \leqq 0$, it is difficult to show that the Fritz John necessary optimality conditions for the resulting problem [4] are also sufficient under reasonable constraint qualifications and convexity assumptions. The specific difficulties are made clear in § 4.

In order to obtain necessary optimality conditions for (P), a generalization of the Motzkin theorem of the alternative [5] is developed in § 2. The necessary optimality conditions in § 3 together with the constraint qualifications in § 4 and appropriate convexity conditions yield the characterization conditions given in § 5. These characterization results are applied in § 6 to obtain characterization

* Received by the editors January 16, 1973.

† Western Electric Engineering Research Center, Princeton, New Jersey 08540. This work was supported in part by the Mathematics Research Center at the University of Wisconsin and by the National Science Foundation under Contract GJ362.

conditions for the solutions of a Chebyshev approximation problem with additional interpolation requirements.

2. Generalized Motzkin theorem. One of the main results needed for proving the necessary optimality conditions is the following theorem.

THEOREM 1. *Let*

U and V be compact sets in R^n,

W be an arbitrary set in R^n.

Then either

$$\begin{aligned} uz &< 0 \quad \text{for all } u \in U, \\ vz &\leqq 0 \quad \text{for all } v \in V, \\ wz &= 0 \quad \text{for all } w \in W, \end{aligned} \tag{2.1}$$

has a solution $z \in R^n$ or for any $u^0 \in U$, there exists $S \leqq n$ with

(2.2)

(i) *S vectors*

$$\begin{aligned} u^i &\in U, \quad i = 1, \cdots, S_1, \\ v^i &\in V, \quad i = S_1 + 1, \cdots, S_2, \\ w^i &\in W, \quad i = S_2 + 1, \cdots, S, \end{aligned}$$

(ii) *$S + 1$ real numbers λ_i, $i = 0, 1, \cdots, S$, such that*

$$\lambda_i \geqq 0 \quad \text{for } i = 0, 1, \cdots, S_2,$$

with either $\lambda_0 > 0$ or $S_1 \geqq 1$ such that

$$\lambda_0 u^0 + \sum_{i=1}^{S_1} \lambda_i u^i + \sum_{i=S_1+1}^{S_2} \lambda_i v^i + \sum_{i=S_2+1}^{S} \lambda_i w^i = 0,$$

but not both.

Proof. The impossibility of both (2.1) and (2.2) follows at once, for if (2.1) holds there would be a $\hat{z} \in R^n$ such that

$$\underbrace{\sum_{i=0}^{S_1} \lambda_i u^i \hat{z}}_{\leftarrow < 0 \rightarrow} + \underbrace{\sum_{i=S_1+1}^{S_2} \lambda_i v^i \hat{z}}_{\leftarrow \leqq 0 \rightarrow} + \underbrace{\sum_{i=S_2+1}^{S} \lambda_i w^i \hat{z}}_{\leftarrow = 0 \rightarrow} < 0,$$

which contradicts (2.2).

Suppose (2.1) does not hold. Define the following subsets of R^n:

$$Z_1 = \left\{ z \,\middle|\, \begin{aligned} uv &\leqq 0 \quad \text{for all } u \in U \\ vz &\leqq 0 \quad \text{for all } v \in V \end{aligned} \right\},$$

$$Z_2 = \{z | wz = 0 \quad \text{for all } w \in W\}.$$

Both Z_1 and Z_2 are convex and $Z_1 \cap Z_2 = \emptyset$.

If $Z_1 \neq \emptyset$, then since $Z_2 \neq \emptyset$ always holds, the separation theorem for disjoint convex sets in R^n [5] yields

$$\begin{aligned} qz &\geqq \alpha \quad \text{for all } z \in Z_1, \\ qz &\leqq \alpha \quad \text{for all } z \in Z_2 \end{aligned} \tag{2.3}$$

for some nontrivial $q \in R^n$ and real number α. Since Z_2 is a subspace, namely $W^{\perp} \triangleq \{z \in R^n | wz = 0 \text{ for all } w \in W\}$, $qz = 0$ for all $z \in Z_2$. Thus, $q \in Z_2^{\perp} = W^{\perp\perp}$ and by a theorem from linear algebra [6] $q \in \operatorname{span} W$, so write $q = \sum_{i=1}^{l} \alpha_i w^i$ with $w^i \in W$ and $l \leqq n$. It is clear from (2.3) that the system

$$\begin{aligned} uz &< 0 \quad \text{for all } u \in U, \\ vz &\leqq 0 \quad \text{for all } v \in V, \\ \left(\sum_{i=1}^{l} \alpha_i w^i\right) z &< 0 \end{aligned} \tag{2.4}$$

has no solution $z \in R^n$.

If $Z_1 = \emptyset$, then the system

$$\begin{aligned} uz &< 0 \quad \text{for all } u \in U, \\ vz &\leqq 0 \quad \text{for all } v \in V \end{aligned} \tag{2.5}$$

has no solution.

Since (2.4) and (2.5) are of the same form, define

$$Q = \begin{cases} U & \text{if } Z_1 = \emptyset, \\ U \cup \left(\sum_{i=1}^{l} \alpha_i w^i\right) & \text{if } Z_1 \neq \emptyset, \end{cases}$$

and then the system

$$\begin{aligned} qz &< 0, \quad q \in Q, \\ vz &\leqq 0, \quad v \in V, \end{aligned} \tag{2.6}$$

has no solution $z \in R^n$. Since both Q and V are compact, for any chosen $u^0 \in U$,

$$Z = \left\{ z \,\middle|\, \begin{array}{l} z = \sum_{i=1}^{I} \beta_i q^i + \sum_{j=1}^{J} \delta_j v^j, \quad q^i = u^0 \text{ for some } i \\ \sum_{i=1}^{I} \beta_i = 1, \beta_i \geqq 0, \delta_j \geqq 0, I, J \text{ arbitrary} \end{array} \right\} \tag{2.7}$$

is closed and convex. If $0 \notin Z$, then by the strict separation theorem for a closed convex set and a point outside that set [5], there is a $d \in R^n$ and a real number $\alpha < 0$ such that

$$dz < \alpha < 0 \quad \text{for all } z \in Z. \tag{2.8}$$

It follows from (2.7) and (2.8) that

$$\begin{aligned} qd &< 0 \quad \text{for all } q \in Q, \\ vd &\leqq 0 \quad \text{for all } v \in V. \end{aligned}$$

Since this contradicts (2.6), $0 \in Z$ must hold, and so for some I_0, J_0, β_i^0 and δ_j^0 we have

$$\begin{aligned} &\sum_{i=1}^{I_0} \beta_i^0 q^i + \sum_{j=1}^{J_0} \delta_j^0 v^j = 0, \\ &q_i = u^0 \quad \text{for some } i = 1, \cdots, I_0, \\ &\sum_{i=1}^{I_0} \beta_i^0 = 1, \quad \beta_i^0 \geqq 0, \quad \delta_j^0 \geqq 0. \end{aligned} \tag{2.9}$$

By replacing the appropriate q^i by $\sum_{i=1}^{l} \alpha_i w^i$ if it occurs in (2.9), the resulting expression is a linear combination in terms of the original vectors in (2.1). By repeatedly using reductions based on the linear dependence of $n + 1$ or more vectors in R^n, such as are used in the proof of the Caratheodory theorem [5], the desired result (2.2) is obtained. Q.E.D.

The classical Motzkin theorem of the alternative [5] is simply Theorem 1 where each of the sets U, V, and W is finite.

3. Necessary optimality conditions. In order to obtain necessary optimality conditions for problem (P), we shall first need a result concerning the linearization of the constraints around any local minimum of (P).

LEMMA 1. *Let $\bar{x}$ be any local minimum of* (P). *Define $\hat{T}_i = \{t \in T_i | G_i(\bar{x}, t) = 0\}$ for $i = 1, \cdots, l$. Then the system*

$$\begin{aligned} &\nabla_x F(\bar{x}) z < 0, \\ &\nabla_x G_i(\bar{x}, t) z < 0 \quad \textit{for all } t \in \hat{T}_i \textit{ for } i = 1, \cdots, l, \\ &\nabla_x H_j(\bar{x}, s) z = 0 \quad \textit{for all } s \in S_j \textit{ for } j = 1, \cdots, m \end{aligned} \tag{3.1}$$

has no solution z in R^n

Since the proof of Lemma 1 is very similar to the proof given in [4], the details are not given here.

THEOREM 2. *Let $\bar{x}$ be a local minimum of problem* (P). *Then there exist integers s_0 and s with $0 \leqq s_0 \leqq s \leqq n$ such that*

(3.2) *there are s_0 indices i_k with $1 \leqq i_k \leqq l$ together with s_0 points $t^k \in \hat{T}_{i_k} = \{t \in T_{i_k} | G_{i_k}(\bar{x}, t) = 0\}$*

for $k = 1, \cdots, s_0$, and

(3.3) *there are $s - s_0$ indices j_k with $1 \leqq j_k \leqq m$ together with $s - s_0$ points $s^k \in S_{j_k}$ for $k = s_0 + 1, \cdots, s$ such that*

(3.4) *there are $s + 1$ real numbers λ_k with $\lambda_0 > 0$ or $s_0 \geqq 1$, and $\lambda_k > 0$ for $k = 1, \cdots, s_0$*

with the property that

$$\lambda_0 \nabla_x F(\bar{x}) + \sum_{k=1}^{s_0} \lambda_k \nabla_x G_{i_k}(\bar{x}, t^k) + \sum_{k=s_0+1}^{s} \lambda_k \nabla_x H_{j_k}(\bar{x}, s^k) = 0. \tag{3.5}$$

Proof. By Lemma 1, there is no solution of the system

$$\begin{aligned} &\nabla_x F(\bar{x})z < 0, \\ (3.6) \qquad &\nabla_x G_i(\bar{x}, t)z < 0 \quad \text{for all } t \in \hat{T}_i \quad \text{for } i = 1, \cdots, l, \\ &\nabla_x H_j(\bar{x}, s)z = 0 \quad \text{for all } s \in S_j \quad \text{for } j = 1, \cdots, m. \end{aligned}$$

Then (3.2)–(3.5) follow from (3.6) and Theorem 1, where we let $u^0 = \nabla_x F(\bar{x})$. Q.E.D.

4. Constraint qualifications for problem (P). Although Theorem 2 gives necessary conditions for an $\bar{x} \in R^n$ to be a local minimum of (P), unless it can be shown that $\lambda_0 > 0$ holds, these conditions are not very meaningful since if $\lambda_0 = 0$ then the conditions say nothing about the objective function $F(x)$ of problem (P). The following constraint qualifications are sufficient to guarantee that $\lambda_0 > 0$.

Constraint qualification 1 (Modified interior point condition). The problem (P) satisfies the *modified interior point condition* if each $G_i(x, t)$ is pseudo-convex in x for all $t \in T_i$ for $i = 1, \cdots, l$ and there exists a point $\tilde{x} \in R^n$ which satisfies

$$\text{(i) } G_i(\tilde{x}, t) < 0 \quad \text{for all } t \in T_i \text{ for } i = 1, \cdots, l,$$

and

$$\text{(ii) } H_j(\tilde{x}, s) = 0 \quad \text{for all } s \in S_j \text{ for } j = 1, \cdots, m.$$

Constraint qualification 2 (Modified strict inequality condition). The problem (P) satisfies the *modified strict inequality condition* at a given point $\bar{x}$, where $\bar{x} \in X = \{x \in X^0 | G_i(x, t) \leqq 0 \text{ for all } t \in T_i \text{ for } i = 1, \cdots, l \text{ and } H_j(x, s) = 0 \text{ for all } s \in S_j \text{ for } j = 1, \cdots, m\}$, if for any choice of integers s_0 and s with $0 \leqq s_0 \leqq s \leqq n$, together with

(i) any choice of s_0 indices i_k with $1 \leqq i_k \leqq l$ and s_0 points $t^k \in \hat{T}_{i_k} = \{t \in T_{i_k} | G_{i_k}(\bar{x}, t) = 0\}$ for $k = 1, \cdots, s_0$

and

(ii) any choice of $s - s_0$ indices j_k with $1 \leqq j_k \leqq m$ and $s - s_0$ points $s^k \in S_{j_k}$ for $k = s_0 + 1, \cdots, s$,

there is a vector $y = (y_1, \cdots, y_n) \in R^n$ such that

$$\text{(iii) } \sum_{q=1}^{n} y_q \nabla_{x_q} G_{i_k}(\bar{x}, t^k) < 0 \quad \text{for } k = 1, \cdots, s_0$$

and

$$\text{(iv) } \sum_{q=1}^{n} y_q \nabla_{x_q} H_{j_k}(\bar{x}, s^k) = 0 \quad \text{for } k = s_0 + 1, \cdots, s.$$

For most problems, it is usually easier to verify constraint qualification 1 rather than constraint qualification 2. Moreover, under the assumption that $G_i(x, t)$ is differentiable in x, constraint qualification 1 implies constraint qualification 2.

THEOREM 3. *Let $\bar{x}$ be a local minimum of problem* (P). *If either constraint qualification* 1 *or* 2 *is satisfied at $\bar{x}$, then $\lambda_0 > 0$ is guaranteed in Theorem* 2.

Proof. Since each $G_i(x, t)$ is differentiable in x, constraint qualification 1 implies constraint qualification 2. So it is sufficient to prove the result under constraint qualification 2. By Theorem 2,

$$\lambda_0 \nabla_x F(\bar{x}) + \sum_{k=1}^{s_0} \lambda_k \nabla_x G_{i_k}(\bar{x}, t^k) + \sum_{k=s_0+1}^{s} \lambda_k \nabla_x H_{j_k}(\bar{x}, s^k) = 0. \tag{4.1}$$

If $\lambda_0 = 0$ holds, then from (4.1) and the $y \in R^n$ from constraint qualification 2 we have

$$\underset{\leftarrow\, < 0\, \rightarrow}{\sum_{k=1}^{s_0} \lambda_k [y \nabla_x G_{i_k}(\bar{x}, t^i)]} + \underset{\leftarrow\, = 0\, \rightarrow}{\sum_{k=s_0+1}^{s} \lambda_k [y \nabla_x H_{j_k}(\bar{x}, s^k)]} = 0$$

which is a contradiction. Thus, $\lambda_0 > 0$ holds. Q.E.D.

Although constraint qualifications are needed to develop meaningful necessary optimality conditions for nonlinear programming problems with a finite number of variables and constraints, this is not necessary for linear problems of the same form [5]. In contrast, for Fritz John type problems some type of constraint qualification is always needed to guarantee $\lambda_0 > 0$, even when the objective function and all the constraints are linear. Consider the following problem which was originally formulated in a different form and used for another purpose [7]:

$$\underset{x_1, x_2, \tau}{\text{minimize}}\ \tau$$

subject to

$$\left.\begin{array}{lll} \text{(i)} & t^4 - x_1 - x_2 t - \tau & \leqq 0 \\ \text{(ii)} & x_1 + x_2 t - t^4 - \tau & \leqq 0 \\ \text{(iii)} & x_1 + x_2 t - t^2 & \leqq 0 \\ \text{(iv)} & -t - x_1 - x_2 t & \leqq 0 \end{array}\right\} \text{ for all } t \in [0, 1]. \tag{4.2}$$

This problem is linear in all the variables x_1, x_2, and τ. By inspection of constraints (iii) and (iv), the only feasible solutions are $x_1 = 0$ and $-1 \leqq x_2 \leqq 0$, so the optimal solution is $\bar{x}_1 = \bar{x}_2 = 0$ and $\bar{\tau} = 1$. By Theorem 2, there exist real numbers λ_i, $i = 0, 1, 2, 3$, with at least one $\lambda_i > 0$ such that

$$\underset{\nabla F(\bar{x}, \bar{\tau})}{\lambda_0 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}} + \underset{\substack{\text{constraint} \\ \text{(i) at } t = 1}}{\lambda_1 \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}} + \underset{\substack{\text{constraint} \\ \text{(iii) at } t = 0}}{\lambda_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}} + \underset{\substack{\text{constraint} \\ \text{(iv) at } t = 0}}{\lambda_3 \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{4.3}$$

Clearly $\lambda_0 > 0$ is impossible for problem (4.2).

Finally, it is appropriate to point out why it is necessary to develop a new characterization theorem for the Fritz John problem with linear equality constraints rather than reduce the equalities to two inequalities and then apply the Fritz John necessary optimality conditions [4] to the resulting problem. If the Fritz John necessary optimality conditions are obtained for the reduced problem with only inequality constraints, it is possible that we can only conclude that $(1/2)(-v) + (1/2)v = 0$, where $vx = 0$ was one of the original equality constraints. There does not appear to be any reasonable way in which to avoid such useless results other than to derive explicit necessary conditions for problems with equality constraints.

5. A characterization theorem for problem (P). Under quite general convexity assumptions on the objective function and constraints of problem (P), the necessary conditions of Theorem 3 are also sufficient. Generalizing from [5], a real-valued function $G(x, t)$, where $x \in R^n$, $t \in T$ and T is an arbitrary set, is said to be *quasi-convex at* $\bar{x}$ if for each x such that $G(x, t) \leqq G(\bar{x}, t)$ for all $t \in T$, then $G((1 - \lambda)\bar{x} + \lambda x, t) \leqq G(\bar{x}, t)$ holds for all $0 \leqq \lambda \leqq 1$ for each $t \in T$. The function $G(x, t)$ is said to be *quasi-convex on a set* $\Gamma \subset R^n$ if it is quasi-convex for each point $x \in \Gamma$. Pseudo-convexity is defined as in [5].

THEOREM 4. *In addition to the assumptions for problem* (P), *let* $F(x)$ *be pseudo-convex on* X^0, *each* $G_i(x, t)$ *be quasi-convex on* X^0, *and assume that either constraint qualification* 1 *or* 2 *holds at* $\bar{x}$. *Then* $\bar{x}$ *solves problem* (P) *if and only if there exist integers* s_0 *and* s *with* $0 \leqq s_0 \leqq s \leqq n$ *such that*

(5.1) *there are* s_0 *indices* i_k *with* $1 \leqq i_k \leqq l$ *together with* s_0 *points* $t^k \in T_{i_k} = \{t \in \hat{T}_{i_k} | G_{i_k}(\bar{x}, t) = 0\}$

for $k = 1, \cdots, s_0$, *and*

(5.2) *there are* $s - s_0$ *indices* j_k *with* $1 \leqq j_k \leqq m$ *together with* $s - s_0$ *points* $s^k \in S_{j_k}$ *for* $k = s_0 + 1, \cdots, s$ *such that*

(5.3) *there are* s *real numbers* λ_k *with* $\lambda_k > 0$ *for* $k = 1, \cdots, s_0$ *with the property that*

$$\nabla_x F(\bar{x}) + \sum_{k=1}^{s_0} \lambda_k \nabla_x G_{i_k}(\bar{x}, t^k) + \sum_{k=s_0+1}^{s} \lambda_k \nabla_x H_{j_k}(\bar{x}, s^k) = 0. \tag{5.4}$$

Proof. The necessity follows at once from Theorem 3.

For the sufficiency, suppose that x were any feasible point, i.e.,

$$x \in \{x \in X^0 | G_i(x, t) \leqq 0 \text{ for all } t \in T_i, i = 1, \cdots, 1;$$
$$H_j(x, s) = 0 \text{ for all } s \in S_j, j = 1, \cdots, m\}.$$

Then for any $i = 1, \cdots, m$, $\nabla_x G_i(\tilde{x}, \tilde{t})(x - \tilde{x}) \leqq 0$ for any $\tilde{x} \in X^0$ and $\tilde{t}$ which satisfy $G_i(\tilde{x}, \tilde{t}) = 0$ by the quasi-convexity and differentiability of each $G_i(x, t)$ from a theorem in [5]. It follows that

$$\sum_{k=1}^{s_0} \lambda_k \nabla_x G_{i_k}(\bar{x}, t^k)(x - \bar{x}) \leqq 0 \tag{5.5}$$

since each $\lambda_k > 0$. By the linearity of $H_j(x, s)$ for $j = 1, \cdots, m$, $\nabla_x H_j(\tilde{x}, \tilde{s})(x - \tilde{x}) = 0$ for any $\tilde{x} \in X^0$ and $\tilde{s}$ which satisfy $H_j(\tilde{x}, \tilde{s}) = 0$, so

$$\sum_{k=s_0+1}^{s} \lambda_k \nabla_x H_{j_k}(\bar{x}, t^k)(x - \bar{x}) = 0. \tag{5.6}$$

Applying (5.5) and (5.6) in (5.4), we conclude

$$\nabla_x F(\bar{x})(x - \bar{x}) \geqq 0. \tag{5.7}$$

The pseudo-convexity of $F(x)$ and (5.7) imply that $F(x) \geqq F(\bar{x})$. Thus $\bar{x}$ solves problem (P). Q.E.D.

6. An application of the characterization theorem. The characterization theorem developed in § 5 can be applied to obtain characterization theorems for numerous important problems in approximation theory; see [2] and [3]. In this paper, for the purposes of illustration, only one application is described, namely the problem of Chebyshev approximation with additional interpolation requirements.

The problem can be described as follows. Let T be a compact subset of a complete metric space, $f(t)$ and $\{\phi_i(t)\}_{i=1}^n$ be real-valued continuous functions defined on T, and $\tilde{T} = \{t^j\}_{j=1}^m$ be a set of m chosen points in T. Then the problem is to find an approximation $\sum_{i=1}^n x_i^* \phi_i(t)$ to $f(t)$ such that

$$\sum_{i=1}^n x_i^* \phi_i(t) = f(t) \quad \text{for } t \in \tilde{T}, \tag{6.1}$$

$$\sup_{t\in T} \left| f(t) - \sum_{i=1}^n x_i^* \phi_i(t) \right| = \inf_{\text{all } x_i\text{'s}} \sup_{t \in T} \left| f(t) - \sum_{i=1}^n x_i \phi_i(t) \right|. \tag{6.2}$$

This can be rewritten in the same form as problem (P):

$$\underset{x,\tau}{\text{minimize}}\ \tau$$

subject to

$$\left.\begin{array}{l} \text{(a)}\ \ -f(t) + \sum_{i=1}^n x_i \phi_i(t) - \tau \leqq 0 \\ \text{(b)}\ \ +f(t) - \sum_{i=1}^n x_i \phi_i(t) - \tau \leqq 0 \\ \text{(c)}\ \ f(t^j) - \sum_{i=1}^n x_i \phi_i(t^j) = 0 \quad \text{for } j = 1, \cdots, m \end{array}\right\} \text{for all } t \in T. \tag{6.3}$$

If there are parameters $\tilde{x}_i$ which satisfy (6.3c), then (6.3) satisfies constraint qualification 1 since by choosing a τ large enough both (6.3a) and (6.3b) can be satisfied as strict inequalities for all $t \in T$. Using Theorem 4 we obtain the following characterization for Chebyshev approximation with interpolation.

THEOREM 5. *Assuming that $f(t)$ is not in the span of $\{\phi_i(t)\}_{i=1}^n$, a vector x^* solves problem* (6.2) *with* (6.1) *if and only if the origin of R^n can be represented as a linear combination of at most $n + 1$ points from the sets*

$$\left\{ e(t) \begin{pmatrix} \phi_1(t) \\ \vdots \\ \phi_n(t) \end{pmatrix} \middle| \, |e(t)| = \|e\|_T \right\}, \tag{6.4}$$

where $e(t) = \sum_{i=1}^n x_i^* \phi_i(t) - f(t)$ is the error function for the approximation problem *and* $\|e\|_T = \sup_T |e(t)|$, *and*

$$(6.5)\qquad \left\{\begin{pmatrix}\phi_1(t)\\ \vdots \\ \phi_n(t)\end{pmatrix}\middle| t\in\tilde{T}\right\} = \left\{\begin{pmatrix}\phi_1(t^j)\\ \vdots \\ \phi_n(t^j)\end{pmatrix}\middle| j = 1,\cdots,m\right\},$$

with at least one point from the set (6.4) *and every point from* (6.4) *having a positive coefficient in the linear combination.*

Proof. All the hypotheses of Theorem 4 are satisfied, so by (5.1)–(5.4), (x^*, τ^*) solves (6.3) if and only if

$$(6.6)\qquad \begin{bmatrix}1\\0\\ \cdot\\ \cdot\\ \cdot\\0\end{bmatrix} + \sum_{k=1}^{s_0} \bar{\lambda}_k \begin{bmatrix}-1\\(-1)^{\varepsilon_k}\phi_1(t^k)\\ \cdot\\ \cdot\\ (-1)^{\varepsilon_k}\phi_n(t^k)\end{bmatrix} + \sum_{k=s_0+1}^{s} \lambda_q \begin{bmatrix}0\\ \phi_1(t^{j_k})\\ \cdot\\ \cdot\\ \phi_n(t^{j_k})\end{bmatrix} = \begin{pmatrix}0\\ \cdot\\ \cdot\\ \cdot\\0\end{pmatrix},$$

where for $k = 1, \cdots, s_0$,

$$\varepsilon_k = \begin{cases}0 & \text{if constraint (6.3a) is active at } t^k,\\ 1 & \text{if constraint (6.3b) is active at } t^k\end{cases}$$

and $\bar{\lambda}_k > 0$. Since $f(t)$ is not in the span of $\{\phi_i(t)\}_{i=1}^n$, $\tau^* > 0$, and so defining $\lambda_k = \bar{\lambda}_k/\tau^*$ for $k = 1, \cdots, s_0$ and observing that

$$e(t^k) = \begin{cases}+\tau^* & \text{if (6.3a) is active at } t^k, \text{ i.e., } \varepsilon_k = 0,\\ -\tau^* & \text{if (6.3b) is active at } t^k, \text{ i.e., } \varepsilon_k = 1,\end{cases}$$

then (6.6) becomes

$$(6.7)\qquad \begin{bmatrix}1\\0\\ \cdot\\ \cdot\\ \cdot\\0\end{bmatrix} + \sum_{k=1}^{s_0} \lambda_k e(t^k) \begin{bmatrix}-1\\ \phi_1(t^k)\\ \cdot\\ \cdot\\ \phi_n(t^k)\end{bmatrix} + \sum_{k=s_0+1}^{s} \lambda_k \begin{bmatrix}0\\ \phi_1(t^{j_k})\\ \cdot\\ \cdot\\ \phi_n(t^{j_k})\end{bmatrix} = \begin{pmatrix}0\\ \cdot\\ \cdot\\ \cdot\\0\end{pmatrix}$$

with $\lambda_k > 0$ for $k = 1, \cdots, s_0$ and $s \leqq n + 1$. Q.E.D.

Theorem 5 generalizes results previously obtained by Deutsch [1] for the case where T is a compact subset of the real line and $\{\phi_i(t)\}_{i=1}^n$ is a Haar set.

Acknowledgment. The author wishes to express his appreciation to Professor O. L. Mangasarian, under whose guidance the major part of this research was conducted as a portion of the author's doctoral dissertation [2] in the Computer Sciences Department at the University of Wisconsin.

REFERENCES

[1] F. Deutsch, *On uniform approximation with interpolatory constraints*, J. Math. Anal. Appl., 24 (1968), pp. 62–79.

[2] K. Gehner, *Optimization problems with an infinite number of constraints and applications to constrained approximation problems*, Doctoral thesis, University of Wisconsin, Madison, 1971.

[3] ———, *Characterization theorems for constrained approximation problems via optimization theory*, submitted to J. Approximation Theory.

[4] F. JOHN, *Extremum problems with inequalities as subsidiary conditions*, Studies and Essays, Courant Anniversary Volume, Interscience, New York, 1948, pp. 187–204.

[5] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[6] E. D. NERING, *Linear Algebra and Matrix Theory*, John Wiley, New York, 1963.

[7] G. D. TAYLOR, *On approximation by polynomials having restricted ranges*, SIAM J. Numer. Anal., 5 (1968), pp. 258–268.

SIAM J. CONTROL
Vol. 12, No. 1, February 1974

A GENERALIZATION TO DUAL BANACH SPACES OF A THEOREM BY BALAKRISHNAN*

RICHARD B. VINTER†

Abstract. A class of optimal control problems is studied in which the controls and outputs are taken as elements in Banach spaces, and the cost functions and constraints are expressible in terms of the norms on these spaces. The paper is principally concerned with generalizing certain results of Balakrishnan relating to optimal control in Hilbert space to this more general setting.

1. Introduction. Let a system be described by a bounded linear map $\mathscr{L}$ from the control space $\mathscr{U}$ into the output space X ($\mathscr{U}$ and X both "dual" Banach spaces), and suppose that the controls u are subject to the constraint $\|u\| \leqq M$, $M > 0$. We set the problem of finding, from among all controls which minimize the distance of the output from the desired output x_d while satisfying the constraint, a control of minimum norm. The norms on $\mathscr{U}$ and X are chosen to give a measure of cost of control and output error respectively. Thus we seek an admissible control whose output approaches most closely to x_d; if there are a number of such controls we seek the most economical. This problem we henceforth call the basic problem.

In connection with the basic problem, Balakrishnan has supplied the following result dealing with the case when $\mathscr{U}$ and X are Hilbert spaces (our notation is adopted).

THEOREM 1.1 [1, p. 120, Thm. 2.3]. *Let $\mathscr{L}$ be a compact linear bounded transformation mapping a Hilbert space $\mathscr{U}$ into another Hilbert space X. Suppose that for a given x_d in X it is required to minimize*

$$\|\mathscr{L}u - x_d\|^2$$

subject to u being in the sphere C in $\mathscr{U}$:

$$\|u\|^2 \leqq M^2.$$

Denote the adjoint of $\mathscr{L}$ by $\mathscr{L}^*$. Then either

$$\sup_{\alpha > 0} \|[\mathscr{L}^*\mathscr{L} + \alpha I]^{-1}\mathscr{L}^*x_d\| \leqq M$$

in which case the sequence

$$u_\alpha = [\mathscr{L}^*\mathscr{L} + \alpha I]^{-1}\mathscr{L}^*x_d$$

is such that u_α converges (strongly) to the optimal element u_0 of minimal norm

$$\lim_{\alpha \downarrow 0} \|\mathscr{L}u_\alpha - x_d\|^2 = \inf_{u \in C} \|\mathscr{L}u - x_d\|^2 = \|\mathscr{L}u_0 - x_d\|^2,$$

or

$$\sup_{\alpha > 0} \|[\mathscr{L}^*\mathscr{L} + \alpha I]^{-1}\mathscr{L}^*x_d\| > M,$$

* Received by the editors October 31, 1972.

† Electronic Systems Laboratory, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. This work was supported by the Commonwealth Fund (Harkness Fellowships).

in which case

$$u_0 = [\mathscr{L}^*\mathscr{L} + \alpha_0 I]^{-1}\mathscr{L}^* x_d,$$

where α_0 is adjusted so that $\|u_0\| = M$ yields the unique solution to the minimization problem.

Below, a generalization of this theorem to dual Banach spaces is provided. The result is thereby made available to a considerably broader range of control problems; in particular, taking $\mathscr{U}$ as L_∞ or NBV, to problems involving "hard" and "fuel" constraints on the control respectively.

Under a fairly mild condition on the input/output map $\mathscr{L}$ we shall establish existence of an optimal control, characterize where possible the solution in terms of a closed hyperplane in the output space and show how this hyperplane may be obtained as the solution to an unconstrained minimization problem over the normed dual of the output space. When such characterization is not possible, we shall obtain the solution instead as the limit of a sequence in the manner of Theorem 1.1 for the Hilbert space problem. Finally § 11 extends certain results of Porter [3, Chap. 4] concerning characterization of "minimum effort" controls from the case when $\mathscr{L}$ is onto to the case when the range of $\mathscr{L}$ is merely dense in X.

Balakrishnan has briefly examined a slight variant on the basic problem above [2, § 3]. In this treatment, however, characterization of the optimal control is limited to the case when the optimal output x_0 is not x_d (this being equivalent to the condition that $k_0 > 0$ in [2, p. 163, line 26]) and no means is provided of determining the appropriate closed hyperplane of support at x_0 (though it is suggested that this may be done through a generalization of the Pontryagin maximum principle).

2. Notations and definitions. By a *dual Banach space* B^* is meant a space isometrically isomorphic with the linear space of all bounded linear functionals f on some Banach space B (the *primal space*), the norm of f being defined in the usual way. We say that B^* is the (normed) *dual* of B and that B is the *pre-dual* of B^*. Only real Banach spaces are here considered.

Suppose that M is a continuous linear map from A into B (A, B Banach spaces), with adjoint $M^*: B^* \to A^*$ (A^*, B^* the duals of A, B respectively). Then we say that M is the *pre-adjoint* of M^*. It is remarked that if B is not (norm) reflexive then not all bounded linear maps from B^* into A^* have pre-adjoints.

The operation of a continuous linear functional b^* on an element b in the primal space will be written $\langle b, b^*\rangle$. The norm on all spaces will be denoted by $\|(\cdot)\|$. θ will denote the null element (in the linear space determined by context).

A Banach space B will be called *strictly normed* if given $x_1, x_2 \in B$,

$$\|x_1 + x_2\| = \|x_1\| + \|x_2\| \quad \text{implies } x_1 = \lambda x_2 \text{ or } x_2 = \theta, \qquad \lambda \geqq 0.$$

We note in particular that L_p, $1 < p < \infty$, and any Hilbert space are strictly normed.

Given any element b in a Banach space B, there exists a nonzero element b^* in the dual of B such that

$$\langle b, b^*\rangle = \|b\| \cdot \|b^*\|$$

([4, p. 186, Thm. 4.3-B]). We say that the two elements b, b^* are *aligned* if and only if this equality holds. Alignment will be indicated by writing $b \parallel b^*$.

In the sequel, solutions to the control problems will be elements in a dual space $\mathcal{U}$. These solutions will be characterized to within alignment with some nonzero element in the pre-dual of $\mathcal{U}$, and to within the magnitude of their norms. For some spaces (e.g., $\mathcal{U} = L_\infty$ or NBV) the characterization may be incomplete unless we place certain restrictions on the input/output map (for example assume system "normality" in the case $\mathcal{U} = L_\infty$). With no such restrictions further analysis will be required to determine the full optimal control history.

In this section we have followed the customary usage of indicating a dual space by *. However in the sequel we shall be working principally in dual spaces and it is convenient to distinguish a Banach space and its dual instead by attaching a superscript $'$ to the primal space.

DEFINITION 2.1. Let X be a topological space. The functional $f: X \to R^+$ is said to be *lower semicontinuous* (l.s.c.) on X if for any $x_0 \in X$ and $\varepsilon > 0$, there exists a neighborhood V of x_0 such that $f(x) > f(x_0) - \varepsilon$ for all $x \in V$.

As is well known [5, p. 219] this definition is equivalent to the requirement that the set $\{x | f(x) \leqq \varepsilon\}$ be closed in X for all real ε.

We note in particular the following.

LEMMA 2.1. *The norm on a dual Banach space B is a weak* l.s.c. functional.*

Proof. For $\varepsilon > 0$, the set $D_\varepsilon = \{x \in B | \|x\| \leqq \varepsilon\}$ is weak* compact by Alaoglu's theorem [4, p. 228, Thm. 4.61-A] and therefore weak* closed [6, p. 424, Cor. 3]. For $\varepsilon \leqq 0$, D_ε is trivially weak* closed. The result follows from the above equivalent definition of l.s.c. 'ty.

3. Problem formulation. Let $\mathcal{U}$, X be dual Banach spaces with pre-duals $\mathcal{U}'$, X' respectively. Let X be reflexive. $\mathcal{U}$ will be called the *control space*, X, the *output space*. We introduce the *input/output map* $\mathcal{L}: \mathcal{U} \to X$. It is assumed:

(i) $\mathcal{L}$ is a bounded, linear map;

(ii) $\mathcal{L}$ has a pre-adjoint $\mathcal{L}': X' \to \mathcal{U}'$.

Let x_d be some nonzero element in X and let $M > 0$. Then we define:

The subproblem.

$$\begin{cases} \text{minimize } \|x_d - \mathcal{L}u\| \\ \text{subject to } \|u\| \leqq M, \quad u \in \mathcal{U}. \end{cases}$$

The subproblem will not be of interest in its own right but as a means to defining the following.

The basic problem. From among all solutions to the subproblem find an element u_0 of minimum norm.

The condition (ii) above on $\mathcal{L}$ is fairly mild: for example it is satisfied when the space $\mathcal{U}$ is reflexive. The condition is introduced to ensure the following continuity property.

LEMMA 3.1. *The map $\mathcal{L}: \mathcal{U} \to X$ is continuous with respect to the weak* topologies on $\mathcal{U}$ and X.*

Proof. Take the basic weak* neighborhood of θ in X:

$$A = \{x \in X | \langle x_i', x\rangle < \varepsilon, i = 1, 2, \cdots, n\}$$

for $x'_i \in X', i = 1, 2, \cdots, n$. Then $\mathscr{L}^{-1}(A) = \{u \in \mathscr{U} | \langle \mathscr{L}' x'_i, u \rangle < \varepsilon, i = 1, 2, \cdots, n\}$; here $\mathscr{L}'$ is the pre-adjoint of $\mathscr{L}$. We see that $\mathscr{L}^{-1}(A)$ is a basic weak* neighborhood of θ in $\mathscr{U}$, whence $\mathscr{L}$ is continuous in the specified sense.

4. Existence and uniqueness of solutions.

PROPOSITION 4.1. *There exists a solution to the basic problem.*

Proof. Let D be the subset of $B = \{u \in \mathscr{U} | \|u\| \leqq M\}$ on which the functional $\|x_d - \mathscr{L}(\cdot)\|$ achieves its minimum. By Alaoglu's theorem, B is weak* compact and in consequence of Lemmas 2.1 and 3.1, $\|x_d - \mathscr{L}(\cdot)\|$ is weak* l.s.c. on $\mathscr{U}$. But a l.s.c. functional on a compact set of a topological space achieves its minimum whence D is nonempty. It readily follows from Definition 2.1 that D is a weak* closed subset of B. Therefore [6, p. 424, Cor. 3] D is weak* compact. Now solutions to the basic problem comprise elements $u \in D$ of minimum norm; we conclude existence of a solution from Lemma 2.1.

PROPOSITION 4.2. *If $\mathscr{U}$ is strictly normed then the solution to the basic problem is unique.*

Proof. This is shown by a simple contradiction argument.

5. The α-problem. Let us define the functional f on $\mathscr{U}$ by

$$f(u) = \|x_d - \mathscr{L}u\|^2 \quad \text{for } u \in \mathscr{U}.$$

The solution to the subproblem is first considered. This problem can be cast as

$$\begin{cases} \text{minimize } f(u^*) \\ \text{subject to } \|u^*\|^2 - M^2 \leqq 0. \end{cases}$$

We find the following.

LEMMA 5.1. *A necessary and sufficient condition for $\bar{u}$ to be a solution to the subproblem is that there exist some $\alpha \geqq 0$ such that*

$$f(\bar{u}) + \alpha[\|\bar{u}\|^2 - M^2] = \min\{f(u) + \alpha[\|u\|^2 - M] | u \in \mathscr{U}\}$$

and

$$\alpha[\|\bar{u}\|^2 - M^2] = 0.$$

Proof. Sufficiency follows from [7, p. 220, Thm. 1] and necessity from [7, p. 217, Thm. 1] on remarking that (i) the required convexity conditions hold, (ii) the infimum is finite and (iii) there exists some $u \in \mathscr{U}$ for which $\|u\| < M$.

The lemma makes it clear that the subproblem can be approached by solution of the following unconstrained problem which we call the α-problem.

The α-problem.

$$\begin{cases} \text{minimize } f(u) + \alpha[\|u\|^2 - M^2], \quad \alpha > 0 \\ \text{subject to } u \in \mathscr{U}. \end{cases}$$

6. Solution of the α-problem. It is a straightforward matter to establish existence of solutions to the α-problem. Further, it is possible to characterize solutions fairly completely and also to compute them in favorable circumstances. This is done in a more general setting in [8].

The following proposition expresses the required control in terms of the solution of an unconstrained minimization problem over the space X'.

PROPOSITION 6.1. *There exists a solution $u(\alpha)$ to the α-problem. Let μ_0 be one of the points on which the functional Γ with domain X' (the pre-dual of X) achieves its minimum where*

$$\Gamma(\mu) = \|\mu\|^2 + \alpha\|\mathscr{L}'\mu\|^2 - 4\langle\mu, x_d\rangle, \qquad \mu \in X'.$$

Then (i) *If $\mathscr{L}'\mu_0 = \theta$ we have the unique trivial solution $u(\alpha) = \theta$.*

(ii) *If $\mathscr{L}'\mu_0 \neq \theta$, $u(\alpha)$ is aligned with $\mathscr{L}'\mu_0$ and $\|u(\alpha)\| = (2\alpha)^{-1}\|\mathscr{L}\mu_0\|$.*

Proof. The proposition is a synthesis of Propositions 6.1, 6.2 and Theorem 3.2 of [8], specialized for the cost functional at hand in the manner indicated in § 10 of [8]. (In [8], it was necessary to assume controllability to guarantee a nontrivial characterization of the optimal control. However with a cost separable in the control and output terms as here, the assumption may be dispensed with.)

7. A particular case. Let us specialize the results of the last section for X, $\mathscr{U}$ both Hilbert spaces. Here, the pre-adjoint map $\mathscr{L}'$, to within isometric isomorphism, is merely the adjoint map $\mathscr{L}^* : X^* \to \mathscr{U}^*$ between the normed duals of X and $\mathscr{U}$.

$\Gamma(\mu)$ is Fréchet differentiable on X. Thus $\partial\Gamma$ takes the value θ at μ_0. A simple computation gives:

$$[\mathscr{L}\mathscr{L}' + \alpha I]\mu_0 = 2\alpha x_d, \tag{7.1}$$

where $I : X \to X$ is the unit map (we loosely regard X and X' as the same space). Now $-\alpha$ lies in the resolvent set of the positive semidefinite self-adjoint linear map $\mathscr{L}\mathscr{L}' : X \to X$ [4, p. 330, Thm. 6.2-B] whence $[\mathscr{L}\mathscr{L}' + \alpha I]^{-1}$ exists and is continuous on its domain of definition.

We can write therefore

$$\mu_0 = 2\alpha[\mathscr{L}\mathscr{L}' + \alpha I]^{-1}x_d.$$

Application of the alignment condition of Proposition 6.1 gives the unique solution to the α-problem as

$$u(\alpha) = \mathscr{L}'[\mathscr{L}\mathscr{L}' + \alpha I]^{-1}x_d \tag{7.2}$$

Remark 7.1. $u(\alpha)$ can be written equivalently as

$$u(\alpha) = [\mathscr{L}'\mathscr{L} + \alpha I]^{-1}\mathscr{L}'x_d \tag{7.3}$$

since from (7.1) we have

$$[\mathscr{L}'\mathscr{L} + \alpha I]\mathscr{L}'\mu_0 = \mathscr{L}'[\mathscr{L}\mathscr{L}' + \alpha I]\mu_0 = 2\alpha\mathscr{L}'x_d.$$

$-\alpha$ lies in the resolvent set of $\mathscr{L}'\mathscr{L} : \mathscr{U} \to \mathscr{U}$, whence $\mathscr{L}\mu_0$ can be expressed

$$\mathscr{L}'\mu_0 = (2\alpha)[\mathscr{L}'\mathscr{L} + \alpha I]^{-1}\mathscr{L}x_d.$$

Equation (7.3) then follows from the alignment condition.

Although (7.3) is more in line with Balakrishnan's results, (7.2) is to be preferred since, in the common case when $\mathscr{L}$ maps a function space into a finite-dimensional space, $\mathscr{L}\mathscr{L}'$ is a finite-dimensional matrix whereas $\mathscr{L}'\mathscr{L}$ is a map between function spaces (albeit with finite-dimensional range).

8. Properties of $u(\alpha)$. In this section some intermediate results are presented. The proofs are routine and therefore confined to Appendix A.

Throughout $u(\alpha)$ will denote a solution to the α-problem. We first show that the functional on $(0, \infty)$ defined by $\|u(\cdot)\|$ is well-defined.

PROPOSITION 8.1. *If two elements u and $\bar{u}$ solve the α-problem, then*

$$\|u\| = \|\bar{u}\|.$$

Additional properties of $\|u(\cdot)\|$ are the following.

PROPOSITION 8.2. $\|u(\alpha)\|$ *is a monotone, nonincreasing function of α on* $(0, \infty)$.

PROPOSITION 8.3. $\|u(\alpha)\|$ *is a continuous function of α on* $(0, \infty)$.

The next result provides a useful characterization of basic problems which still have a solution after removing the constraint $\|u\| \leqq M$.

PROPOSITION 8.4. *There exists a u_0 with $f(u_0) = \min\{f(u) | u \in \mathscr{U}\}$ if and only if* $\sup\{\|u(\alpha)\| \,|\alpha > 0\} < \infty$. *If such a u_0 exists, then*

$$\sup\{\|u(\alpha)\| | \alpha > 0\} \leqq \|u_0\|.$$

Remark 8.1. It is clear from Proposition 8.2 that $\sup\{\|u(\alpha)\| | \alpha > 0\} = \lim_{\alpha \downarrow 0} \|u(\alpha)\|$.

9. Solution to the basic problem, Part 1. Solution of the basic problem will proceed along two different lines depending on which of the two conditions (C1) or (C2) holds:

$$\text{(C1)} \qquad \lim_{\alpha \downarrow 0} \|u(\alpha)\| > M,$$

$$\text{(C2)} \qquad \lim_{\alpha \downarrow 0} \|u(\alpha)\| \leqq M.$$

Case (C1) is the easiest to deal with. In fact we have the following proposition.

PROPOSITION 9.1. *Suppose that condition* (C1) *holds. Let $\alpha_0 > 0$ be such that $\|u(\alpha_0)\| = M$. Then $u(\alpha_0)$ is a solution to the basic problem.*

Proof. Let us first show that α_0 with the stated properties exists. Consider the program

$$\begin{cases} \text{minimize } f(x) \\ \text{subject to } \|u\|^2 - M^2 \leqq 0. \end{cases}$$

We know from Proposition 4.1 that this has a solution $\bar{u}$. Further from Lemma 5.1, there exists some $\alpha \geqq 0$ such that

$$f(\bar{u}) + \alpha\|\bar{u}\|^2 = \min\{f(u) + \alpha\|u\|^2 | u \in \mathscr{U}\}$$

and

$$\alpha[\|\bar{u}\|^2 - M^2] = 0.$$

Suppose that $\alpha = 0$. Then $\bar{u}$ solves the unconstrained problem and from Proposition 8.4 $\|\bar{u}\|^2 - M^2 > 0$ contradicting the fact that $\bar{u}$ solves the program. But if $\alpha > 0$, then $\bar{u} = u(\alpha)$ and $\|u(\alpha)\| = M$. Setting $\alpha_0 = \alpha$ we have achieved our aim.

We have shown that $u(\alpha_0)$ solves the subproblem. It is a simple matter to show that $u(\alpha_0)$ in fact solves the basic problem. For suppose to the contrary that there exists some u with

$$f(u(\alpha_0)) = f(u) \quad \text{and} \quad \|u\| < \|u(\alpha_0)\|.$$

Then $f(u(\alpha_0)) + \alpha_0\|u(\alpha_0)\|^2 > f(u) + \alpha_0\|u\|^2$ since $\alpha_0 > 0$. This strict inequality contradicts the optimality of $u(\alpha_0)$ and concludes the proof.

In view of Remark 8.1, the condition (C1) can be tested by taking some sequence $\{\alpha_n\}$, $\alpha_n \downarrow 0$, and examining whether the monotone nondecreasing sequence $\{\|u(\alpha_n)\|\}$ is bounded above by M or not. The task of matching α_0 to satisfy $\|u(\alpha_0)\| = M$ is aided by knowledge that $\|u(\cdot)\|$ is a continuous, monotone function (Propositions 8.2 and 8.3).

10. Solution to the basic problem, Part 2. We now turn to the more interesting situation when condition (C2) holds. A more direct approach can be followed when the system is controllable (the problem will be examined with this added postulate in the next section). Here we make no controllability hypotheses.

To obtain a Banach space analogue of Balakrishnan's result (Theorem 1.1), we would hope that, for condition (C2), $u(\alpha) \to u_0$ (a solution to the basic problem) as $\alpha \downarrow 0$. Making certain assumptions on $\mathscr{U}$ we will establish the desired convergence property (the nature of convergence, strong or weak*, depending on the severity of the assumptions). In general however we cannot assure convergence (except for certain subsequences) since we have built no conditions into the problem formulation guaranteeing even uniqueness of solutions to the basic problem. However the computational significance of Balakrishnan's result (assuming condition (C2) to hold) is that by taking α a sufficiently small positive number, we can obtain from (7.3) a control whose "suboptimality" is arbitrarily small. It is this aspect which has a parallel for the basic problem in its full generality.

PROPOSITION 10.1. *Suppose condition* (C2) *to hold. Then writing u_0 for a solution to the basic problem, we have*

(i) $\lim_{\alpha\downarrow 0} \alpha^{-1}\{\|x_d - \mathscr{L}u(\alpha)\|^2 - \|x_d - \mathscr{L}u_0\|^2\} = 0$,

(ii) $\lim_{\alpha\downarrow 0} \|u(\alpha)\| = \|u_0\|$,

(iii) *if the basic problem has a unique solution* (*in particular if $\mathscr{U}$ is strictly convex*), *then*

$$u(\alpha) \xrightarrow{\text{weakly*}} u_0 \quad \text{as } \alpha \downarrow 0.$$

If uniqueness fails, we still have that any sequence $\{\alpha_i\}$ of real numbers with $\alpha_i \downarrow 0$ contains a subsequence $\{\alpha_j\}$ such that

$$u(\alpha_j) \xrightarrow{\text{weakly*}} \tilde{u} \quad \text{as } i \to \infty,$$

where $\tilde{u}$ solves the basic problem.

Before embarking on proof of this proposition, we take note of the following lemma.

LEMMA 10.1. *Let K be a map from the real line into the topological space Y. Let α_0 be some real number and suppose that there exists some $y_0 \in Y$ such that, given any sequence $\{\alpha_i\}$ with $\alpha_i \to \alpha_0$, there exists some subsequence $\{\alpha_j\}$ such that $K(\alpha_j) \to y_0$ as $\alpha \to \alpha_0$. Then*

$$K(\alpha) \to y_0 \quad \text{as} \quad \alpha \to \alpha_0.$$

Proof. Suppose that $K(\alpha) \not\to y_0$ as $\alpha \to \alpha_0$. Then there exists some sequence $\{\alpha_i\}$ with $\alpha_i \to \alpha_0$ and a neighborhood $N(y_0)$ of y_0 such that $K(a_i) \notin N(y_0)$ for $i = 1, 2, \cdots$. Clearly for no subsequence $\{\alpha_j\}$ do we have $K(\alpha_j) \to y_0$ as hypothesized. This contradiction establishes that $\lim_{\alpha \to \alpha_0} K(\alpha) = y_0$ as required.

Proof of Proposition 10.1. Define the functional $\Delta f(\cdot)$ by

$$\Delta f(u) = f(u) - f(u_0) \quad \text{for } u \in \mathscr{U}.$$

Part (i). Since by optimality of u_0

$$\liminf_{\alpha \downarrow 0} \{\alpha^{-1} \Delta f(u(\alpha))\} \geqq 0,$$

to establish Part (i), it suffices to show that

$$\limsup_{\alpha \downarrow 0} \{\alpha^{-1} \Delta f(u(\alpha))\} \leqq 0.$$

To this end we first remark that for $0 < \alpha_1 \leqq \alpha_2$,

$$\begin{aligned} \|u_0\|^2 &\geqq \|u(\alpha_1)\|^2 + \alpha_1^{-1} \Delta f(u(\alpha_1)) \\ &\geqq \|u(\alpha_2)\|^2 + \alpha_2^{-1} \Delta f(u(\alpha_2)). \end{aligned} \tag{10.1}$$

This result follows immediately from the inequalities

$$\begin{aligned} \|u_0\|^2 &= \|u_0\|^2 + \alpha_1^{-1} \Delta f(u_0) \geqq \|u(\alpha_1)\|^2 + \alpha_1^{-1} \Delta f(u(\alpha_1)) \\ &\geqq \|u(\alpha_1)\|^2 + \alpha_2^{-1} \Delta f(u(\alpha_1)) \geqq \|u(\alpha_2)\|^2 + \alpha_2^{-1} \Delta f(u(\alpha_2)). \end{aligned}$$

(We have used the optimality of u_0, $u(\alpha_1)$ and $u(\alpha_2)$.)

Inequality (10.1) assures existence of some $Y < \infty$ such that

$$\limsup_{\alpha \downarrow 0} \{\|u(\alpha)\|^2 + \alpha^{-1} \Delta f(u(\alpha))\} = \mathrm{Y}.$$

Clearly,

$$\limsup_{\alpha \downarrow 0} \{\alpha^{-1} \Delta f(u(\alpha))\} \leqq Y.$$

Let us assume contrary to our requirement that

$$\limsup_{\alpha \downarrow 0} \{\alpha^{-1} \Delta f(u(\alpha))\} = 3\varepsilon, \qquad \varepsilon > 0.$$

The positive number α_3 is chosen so that

$$Y < \|u(\alpha_3)\|^2 + \alpha_3^{-1} \Delta f(u(\alpha_3)) + \varepsilon.$$

There exists a positive number α_4 such that $4\alpha_4 < \alpha_3$ and $2\varepsilon < \alpha_4^{-1} \Delta f(u(\alpha_4)) < 4\varepsilon$. But then

$$\begin{aligned} Y &< \|u(\alpha_3)\|^2 + \alpha_3^{-1} \Delta f(u(\alpha_3)) + \varepsilon \leqq \|u(\alpha_4)\|^2 + \alpha_3^{-1} \Delta f(u(\alpha_4)) + \varepsilon \\ &= \|u(\alpha_4)\|^2 + (\alpha_4/\alpha_3)\alpha_4^{-1} \Delta f(u(\alpha_4)) + \varepsilon \leqq \|u(\alpha_4)\|^2 + (1/4)\alpha_4^{-1} \Delta f(u(\alpha_4)) + \varepsilon \\ &< \|u(\alpha_4)\|^2 + 2\varepsilon < \|u(\alpha_4)\|^2 + \alpha_4^{-1} \Delta f(u(\alpha_4)) \leqq Y. \end{aligned}$$

This contradiction concludes the proof of Part (i).

Parts (ii) *and* (iii). Let $\{\alpha_i\}$ be a sequence of real numbers with $\alpha_i \downarrow 0$. Consider $\{u(\alpha_i)\}$. By Proposition 8.4, the sequence is contained in a closed sphere of radius $\|u_0\|$

with center at the origin in $\mathscr{U}$. There therefore exists some subsequence [4, Thms. 2.4-D and 4.61-A] which we write as $\{u(\alpha_n)\}$ such that $u(\alpha_n) \xrightarrow{\text{weakly*}} \bar{u}$, where $\bar{u} \in \mathscr{U}$. Now from (10.1),

$$\limsup_{n\to\infty} \|u(\alpha_n)\| \leqq \|u_0\|. \tag{10.2}$$

Since the set $\{u \in \mathscr{U} | \|u\| \leqq \|u_0\|\}$ is weak* closed [6, p. 424] we have that $\|\bar{u}\| \leqq \|u_0\|$.

Now in consequence of Lemmas 2.1 and 3.1, $\Delta f(\cdot)$ is weak* l.s.c. whence

$$\Delta f(\bar{u}) \leqq \liminf_{n\to\infty} \{\Delta f(u(\alpha_n))\}.$$

Since from Part (i) we have that in particular $\lim_{n\to\infty} \Delta f(u(\alpha_n)) = 0$,

$$\|\bar{u}\| \leqq \|u_0\| \quad \text{and} \quad f(\bar{u}) \leqq f(u_0).$$

The optimality of u_0 is contradicted unless $\|\bar{u}\| = \|u_0\|$ and $f(\bar{u}) = f(u_0)$. It is evident that $\bar{u}$ is a solution to the basic problem. Taking $\tilde{u}$ as $\bar{u}$ we have proved the second part of (iii). The first part of Part (iii) follows from Lemma 10.1.

Again from Lemma 2.1, $\|(\cdot)\|$ is a weak* l.s.c. functional on $\mathscr{U}$. We conclude that

$$\|u_0\| = \|\bar{u}\| \leqq \liminf_{n\to\infty} \|u(\alpha_n)\| \leqq \limsup_{n\to\infty} \|u(\alpha_n)\| \leqq \|u_0\|$$

(we have used (10.2)). In consequence $\lim_{n\to\infty} \|u(\alpha_n)\| = \|u_0\|$. In view of Lemma 10.1 we have proved Part (ii). This concludes the proof of the proposition.

Remark 10.1. Note that Part (i) of Proposition 10.1 is a rather stronger statement than: $\lim_{\alpha\downarrow 0} f(u(\alpha)) = f(u_0)$.

Finally we point out that by imposing the condition that $\mathscr{U}$ be uniformly convex[1] we can ensure that $u(\alpha)$ tends strongly (rather than weakly*) to u_0, the unique solution to the basic problem.

PROPOSITION 10.2. *Let $\mathscr{U}$ be uniformly convex. Then in Proposition* 10.1 *we can replace* (iii) *by*

(iii)′ $u(\alpha) \to u_0$ (*strongly in $\mathscr{U}$*) *as* $\alpha \downarrow 0$, *where u_0 is the unique solution to the basic problem.*

The proof of the proposition is given in Appendix B. The result is of some interest because all Hilbert spaces and also the spaces l_n, L_n, $1 < n < \infty$, are uniformly convex [10].

11. Controllable systems. As promised we now consider problems where condition (C2) holds and the input/output map is controllable. By a *controllable map* (with codomain a topological space) we mean a map whose range is dense. The strongest results are obtained when the map $\mathscr{L}$ is onto.

PROPOSITION 11.1. *Suppose that condition* (C2) *holds and that $\mathscr{L}$ is onto. If μ_0 is any of the points in $X' \sim \{\theta\}$ (X' the pre-dual of X) on which the functional $\Gamma(\mu)$ attains its maximum where*

$$\Gamma(\mu) = \langle \mu, x_d \rangle / \|\mathscr{L}\mu\|,$$

then u is a solution to the basic problem if and only if

$$u \| \mathscr{L}\mu_0 \quad \textit{and} \quad \mathscr{L}u_0 = x_d.$$

[1] A Banach space is said to be *uniformly convex* if for each $\varepsilon \in (0, 2)$, there exists some $\delta > 0$ such that $\|x\| = \|y\| = 1$ and $\|x - y\| > \varepsilon$ imply that $\|(x + y)/2\| < 1 - \delta$.

It will be found that any solution to the basic problem satisfies

$$\|u_0\| = \langle \mu_0, x_d \rangle / \|\mathscr{L}\mu_0\|.$$

Proof. With $\mathscr{L}$ onto, it is clear from the problem formulation that the basic problem therefore reduces to

$$\begin{cases} \text{minimize } \|u\| \\ \text{subject to } x_d = \mathscr{L}u, \quad u \in \mathscr{U}. \end{cases}$$

But this is precisely the minimum effort problem studied in Chapter 4 of [3]. The proposition expresses standard results in the literature.

The assumption that $\mathscr{L}$ be onto is very severe; it effectively limits treatment to systems having finite-dimensional output spaces. It is of interest therefore to examine whether Proposition 11.1 admits generalization to the situation, where the range of $\mathscr{L}$ is merely dense in X. Such a generalization would provide characterization of solutions to the basic problem directly in terms of nonzero elements in X' for the case that $\mathscr{L}$ is controllable and condition (C2) holds. The cumbersome procedure of determining the optimal as a limit of solutions to the α-problem would thereby be circumvented.

Now Proposition 11.1 identifies an optimal control u_0 as being a point in $\mathscr{U}$ whose image under $\mathscr{L}$ lies nearest to the closed hyperplane supporting the set $\mathscr{R}$ at x_d; here

$$\mathscr{R} = \{x \in X | x = \mathscr{L}u, \ \|u\| \leqq \inf\{\|v\| \,|x_d = \mathscr{L}v, v \in \mathscr{U}\}\}.$$

Unfortunately, such identification is not always possible when $\mathscr{L}$ is not onto; indeed, it is easy to construct counterexamples (see, e.g., [9]) where no closed hyperplanes exist supporting $\mathscr{R}$ at certain of its bounding points. We do have, however, the following proposition.

PROPOSITION 11.2. *Let $\mathscr{L}$ be controllable and let $\mathscr{U}$ be reflexive. Suppose that condition* (C2) *holds. Let Γ be the real-valued function with domain $X \sim \{\theta\}$ defined by*

$$\Gamma(\mu) = \langle \mu, x_d \rangle / \|\mathscr{L}'\mu\|, \qquad \mu \in X',$$

and let

$$\Gamma_0 = \sup\{\Gamma(\mu) | \mu \in X' \sim \{\theta\}\}.$$

Then $\Gamma_0 < \infty$ and either (i) *Γ achieves its maximum at some $\mu_0 \in X' \sim \{\theta\}$ in which case $u_0 \| \mathscr{L}\mu_0$ and $\|u_0\| = \Gamma_0$, where u_0 is any solution to the basic problem, or* (ii) *Γ fails to achieve its maximum on $X \sim \{\theta\}$, but there exists some sequence $\{\mu_i\} \subset X'$ such that $\Gamma(\mu_i) \to \Gamma_0$ and $\mathscr{L}'\mu_i \xrightarrow{\text{strongly}} u_0'$ where u_0' is some nonzero element in $\mathscr{U}'$. In this case*

$$u_0 \| u_0' \quad \textit{and} \quad \|u_0\| = \Gamma_0.$$

The proof of this result is supplied in Appendix C.

Remark 11.1. The unsatisfactory (and apparently unavoidable) feature of Proposition 11.2 is that u_0 is aligned not with some element in the range of $\mathscr{L}'$, written $\mathscr{R}(\mathscr{L}')$, but instead with some element in $\overline{\mathscr{R}(\mathscr{L}')}$. If we relinquish the reflexivity of $\mathscr{U}$, the situation is even graver. In this case let us denote the adjoint of $\mathscr{L}$ by $\mathscr{L}^*: X^* \to \mathscr{U}^*$, where $\mathscr{U}^*$, X^* are the normed duals of $\mathscr{U}$, X respectively; in consequence of the assumed existence of the pre-adjoint map $\mathscr{L}': X' \to \mathscr{U}'$ and the assumption that X be reflexive, we have $\mathscr{R}(\mathscr{L}') = \mathscr{R}(\mathscr{L}^*)$ (to within isometric isomorphism). Further analysis reveals that here u_0 is aligned with some element in the orthogonal complement of the null space of $\mathscr{L}$; this we denote $[\mathscr{N}(\mathscr{L})]^{\perp}$. Since for some input/output maps

$$\mathscr{R}(\mathscr{L}') \cong \mathscr{R}(\mathscr{L}^*) \overset{\text{strict}}{\subset} [\mathscr{N}(\mathscr{L})]^{\perp}$$

[4, p. 226, Thm. 4.6-G], we have said somewhat less about the optimal control than in the reflexive case.

In the light of this section, we can distinguish between three different situations when condition (C2) holds:

(i) $\mathscr{R}(\mathscr{L}) = X$, $\mathscr{U}$ not necessarily reflexive. u_0 is characterized through some element $\mu_0 \in X'$ as being aligned with $\mathscr{L}'\mu_0$.

(ii) $\overline{\mathscr{R}(\mathscr{L})} = X$, $\mathscr{U}$ reflexive. u_0 is characterized through some sequence $\{\mu_i\} \subset X'$ as being aligned with $\lim_{i\to\infty} \mathscr{L}'\mu_i$.

(iii) $\overline{\mathscr{R}(\mathscr{L})} = X$, $\mathscr{U}$ not necessarily reflexive. u_0 need not be characterized through any element or sequence in X'.

Notice that the findings of § 10 are not contradicted; although $u(\alpha)$, $\alpha > 0$, is aligned with some $\mathscr{L}\mu$, $\mu \in X$, the weak* limit of $\{u(\alpha_n)\}$ need not be so aligned.

12. Summary of results. We bring together our findings.

THEOREM 12.1. *Let $u(\alpha)$ be a solution to the α-problem. Then either*

$$\lim_{\alpha \downarrow 0} \|u(\alpha)\| > M,$$

in which case a solution to the basic problem is given by $u(\alpha_0)$, where α_0 is so adjusted that

$$\|u(\alpha_0)\| = M,$$

or

$$\lim_{\alpha \downarrow 0} \|u(\alpha)\| \leqq M,$$

in which case $u(\alpha)$, for α sufficiently small, approximates arbitrarily closely to a solution of the basic problem in the sense of Proposition 10.1.

We recall from § 7 that for $\mathscr{U}$, X Hilbert spaces,

$$u(\alpha) = [\mathscr{L}'\mathscr{L} + \alpha I]^{-1}\mathscr{L}'x_d.$$

Noting that when $\mathscr{U}$ is reflexive we have, for the condition (C2), $u(\alpha) \xrightarrow{\text{strongly}} u_0$ (the unique solution to the basic problem) as $\alpha \downarrow 0$ (Proposition 10.2 and succeeding remarks), we recover Balakrishnan's results (Theorem 1.1) in full as a special case of Theorem 12.1. In addition, Part (i) if Proposition 10.1 is a stronger statement than the corresponding statement in Theorem 1.1 and (as observed in [2]) the hypothesis that $\mathscr{L}$ be compact has been shown to be redundant.

Appendix A.

Proof of Proposition 8.1. This is shown by a simple convexity argument. Assume in contraposition that

$$f(u) + \alpha\|u\|^2 = f(\bar{u}) + \alpha\|\bar{u}\|^2 \quad \text{and} \quad \|u\| \neq \|\bar{u}\|.$$

Let $\varepsilon \in (0, 1)$. Then by the convexity of f and the triangle inequality,

$$f[\varepsilon u + (1 - \varepsilon)\bar{u}] + \alpha\|\varepsilon u + (1 - \varepsilon)\bar{u}\|^2$$

$$\leqq \varepsilon[f(u) + \alpha\|u\|^2] + (1 - \varepsilon)[f(\bar{u}) + \alpha\|\bar{u}\|^2] - \varepsilon(1 - \varepsilon)\cdot\alpha[\|u\| - \|\bar{u}\|]^2.$$

Since the last term is positive under the stated hypotheses,

$$f[\varepsilon u + (1 - \varepsilon)\bar{u}] + \alpha\|\varepsilon u + (1 - \varepsilon)\bar{u}\|^2$$

$$< \varepsilon[f(u) + \alpha\|u\|^2] + (1 - \varepsilon)[f(\bar{u}) + \alpha\|\bar{u}\|] = f(u) + \alpha\|u\|^2,$$

in contradiction of the optimality of u.

Proof of Proposition 8.2. Assume in contraposition that there exist α_1, α_2 with $\alpha_1 > \alpha_2 > 0$ for which $\|u(\alpha_1)\| > \|u(\alpha_2)\|$.

By the optimality of $u(\alpha_1)$, $u(\alpha_2)$,

$$f(u(\alpha_2)) - f(u(\alpha_1)) \leqq \alpha_2\{\|u(\alpha_1)\|^2 - \|u(\alpha_2)\|^2\}$$

and

$$\alpha_1\{\|u(\alpha_1)\|^2 - \|u(\alpha_2)\|^2\} \leqq f(u(\alpha_2)) - f(u(\alpha_1)).$$

Since $\alpha_1 > \alpha_2$, $f(u(\alpha_2)) - f(u(\alpha_1)) < f(u(\alpha_2)) - f(u(\alpha_1))$. This proves the proposition.

It is convenient to prove Proposition 8.4 before Proposition 8.3.

Proof of Proposition 8.4. The "only if" part will follow if we can prove the final statement of the proposition. For any $\alpha \in (0, \infty)$, by the optimality of $u(\alpha)$,

$$\|u_0\|^2 = \|u_0\|^2 + \alpha^{-1}[f(u_0) - f(u_0)] \geqq \|u(\alpha)\| + \alpha^{-1}[f(u(\alpha)) - f(u_0)].$$

But $\alpha^{-1}[f(u(\alpha)) - f(u_0)] \geqq 0$ by the optimality of u_0, whence $\|u_0\|^2 \geqq \|u(\alpha)\|^2$ providing the required upper bound.

To prove the "if" part, suppose that

$$\sup\{\|u(\alpha)\| | \alpha > 0\} < k, \qquad k > 0,$$

and consider the program

$$\begin{cases} \text{minimize } f(u) \text{ over } u \in \mathscr{U} \\ \text{subject to } \|u\| - k^2 \leqq 0. \end{cases}$$

By Proposition 4.1 and Lemma 5.1, the program has a solution $u(\alpha)$ and there exists some $\alpha \geqq 0$ such that

$$f(u(\alpha)) + \alpha[\|u(\alpha)\|^2 - k^2] = \min\{f(u) + \alpha[\|u\| - k^2] | u \in \mathscr{U}\}$$

and

$$\alpha[\|u(\alpha)\|^2 - k^2] = 0.$$

Suppose that $\alpha \neq 0$. Then $\|u(\alpha)\| = k$. This contradicts the hypothesis and we conclude that $\alpha = 0$. But in this case f achieves its minimum on $\mathscr{U}$ as required for completion of the proof.

Proof of Proposition 8.3. Given the positive real numbers α', α'' with $\alpha' > \alpha''$ we first show that $\|u(\alpha)\|$ achieves all values between $\|u(\alpha')\|$ and $\|u(\alpha'')\|$ as α ranges over $[\alpha'', \alpha']$: the result follows trivially if $\|u(\alpha')\| = \|u(\alpha'')\|$, so we suppose that $\|u(\alpha'')\| > \|u(\alpha')\|$ (recall that $u(\cdot)$ is monotone nonincreasing). Let M belong to the interval $(\|u(\alpha')\|, \|u(\alpha'')\|)$ and consider the problem

$$\begin{cases} \text{minimize } f(u) \text{ over } u \in \mathscr{U} \\ \text{subject to } \|u\|^2 - M^2 \leqq 0. \end{cases}$$

By Proposition 4.1, this has a solution u_1, and there exists a corresponding $\alpha_1 \geqq 0$ such that

$$f(u_1) + \alpha_1\|u_1\|^2 = \min\{f(u) + \alpha_1\|u\|^2 | u \in \mathscr{U}\} \tag{A.1}$$

and

$$\alpha_1[\|u_1\|^2 - M^2] = 0. \tag{A.2}$$

Suppose that $\alpha_1 = 0$. Then u_1 solves the unconstrained problem, and by Proposition 8.4,

$$\|u_1\| \geqq \|u(\alpha'')\| > M \geqq \|u_1\|.$$

From this contradiction we conclude that $\alpha_1 > 0$. But then from (A.1) we have that $u_1 = u(\alpha_1)$. From (A.2), $\|u(\alpha_1)\| = M$. In view of Propositions 8.1 and 8.2,

$$\alpha_1 \in (\alpha', \alpha''),$$

and the result follows.

Now since $\|u(\cdot)\|$ is monotonic on $(0, \infty)$, it has left and right limits at all points in $(0, \infty)$ [11, p. 78]. Obviously a contradiction arises with the above result unless the left and right limits are the same. We have established that $\|u(\cdot)\|$ is continuous on $(0, \infty)$.

Appendix B.

Proof of Proposition 10.2. With $\mathscr{U}$ uniformly convex, $\mathscr{U}$ is reflexive and strictly normed [12, p. 33]. Take the sequence $\{\alpha_i\}$ with $\alpha_i \downarrow 0$ and write $\{u_i\}$ for the sequence $\{u(\alpha_i)\}$ in $\mathscr{U}$. From Proposition 10.1 (remembering that $\mathscr{U}$ is-in particular reflexive)

$$u_i \xrightarrow{\text{weakly}} u_0,$$

$$\|u_i\| \to \|u_0\|$$

and $\{\|u_i\|\}$ is an increasing sequence. (Here u_0 is the unique solution to the basic problem).

Now suppose that $u_i \xrightarrow{\text{strongly}} u_0$. Then there exists some subsequence of $\{u_i\}$ which we write as $\{u_j\}$ such that

$$\|u_0 - u_j\| > \delta, \qquad j = 1, 2, \cdots,$$

for some $\delta > 0$. Since $\{\|u_j\|\}$ is an increasing sequence and $\mathscr{U}$ is uniformly convex, there exists some $\varepsilon > 0$ such that

$$\|u_0 + u_j\| < (2 - \varepsilon)\|u_0\|.$$

$\mathscr{U}$ is reflexive, so there exists some nonzero u_0' such that $\langle u_0', u_0\rangle = \|u_0'\| \, \|u_0\|$. But then

$$(2 - \varepsilon)\|u_0\| > \|u_0 + u_j\| \geqq \|u_0'\|^{-1}\langle u_0', u_0 + u_j\rangle, \quad j = 1, 2, \cdots.$$

Since $u_j \xrightarrow{\text{weakly}} u_0$, we have

$$\lim \|u_0'\|^{-1}\langle u_0', u_0 + u_j\rangle = 2\|u_0'\|^{-1}\langle u_0', u_0\rangle = 2\|u_0\|.$$

From this contradiction we conclude that $u_j \xrightarrow{\text{strongly}} u_0$. But then by Lemma 10.1, $u(\alpha) \xrightarrow{\text{strongly}} u_0$ as $\alpha \downarrow 0$, and the proposition is proved.

Appendix C. In this section, Proposition 11.2 is proved. The proof leans heavily on duality theory. Rather than refer constantly to the duality literature we prefer to work from the following lemma which condenses the results which we shall require (the lemma can be pieced together from [13] though the more streamlined notation of [14] is employed).

LEMMA A.1. *Let f and g be respectively proper convex l.s.c. and proper concave u.s.c. (extended-valued) functionals (see* [14] *for definition of technical terms) on the Banach space A. Take the primal problem to be*

$$\text{(P)} \qquad \text{minimize } \{f(a) - g(a) | a \in A\}.$$

Suppose that there exists some point a in the effective domain (dom (g)) *of g such that f is (strongly) continuous at a. Define the conjugate functionals f' and g' with domains all of A^* (the normed dual of A) by the formulas*

$$f'(a') = \sup\{\langle a, a'\rangle - f(a) | a \in A\}, \qquad a' \in A^*,$$

$$g'(a') = \inf\{\langle a, a'\rangle - g(a) | a \in A\}, \qquad a' \in A^*.$$

Take the dual problem to be

$$\text{(P}'\text{)} \qquad \text{maximize } \{g'(a') - f'(a') | a' \in A^*\}.$$

Then, if $\inf\{f(a) - g(a) | a \in A\} > -\infty$,

(a) (P′) *has a proper solution* a_0',

(b) $\inf\{f(a) - g(a) | a \in A\} = g'(a_0') - f'(a_0')$,

(c) a_0 *is a solution to the primal problem if and only if the following alignment conditions hold*:

$$\langle a_0, a_0'\rangle - f(a_0) = \max\{\langle a, a_0'\rangle - f(a) | a \in A\},$$

$$\langle a_0, a_0'\rangle - g(a_0) = \min\{\langle a, a_0'\rangle - g(a) | a \in A\}.$$

Proof of Proposition 11.2. In simple consequence of Proposition 8.4, when (C2) holds and $\overline{\mathscr{R}(\mathscr{L})} = X$, there exists some $\bar{u} \in \mathscr{U}$, $\|\bar{u}\| \leqq M$, such that $x_d = \mathscr{L}\bar{u}$. Therefore under the stated conditions the basic problem takes the form:

$$\begin{cases} \text{minimize } \|u\| \\ \text{subject to } x_d = \mathscr{L}u, \quad u \in \mathscr{U}. \end{cases}$$

Let us change variables $u \to \bar{u} - w$ to give

$$\text{(P)} \qquad \begin{cases} \text{minimize } \|\bar{u} - w\| \\ \text{subject to } \mathscr{L}w = \theta, \quad w \in \mathscr{U}. \end{cases}$$

(P) will be identified with the primal problem. To this end define the continuous proper convex functional f with domain $\mathscr{U}$ by

$$f(w) = \|\bar{u} - w\|, \qquad w \in \mathscr{U}.$$

Define also the u.s.c. proper concave extended-valued functional g with domain $\mathscr{U}$ by

$$g(w) = \begin{cases} 0 & \text{if } w \in \mathscr{N}(\mathscr{L}), \\ -\infty & \text{otherwise}. \end{cases}$$

($\mathscr{N}(\mathscr{L})$ denotes the null space of $\mathscr{L}$.)

Then (P) can be cast into the form:

$$\text{(P)} \qquad \text{minimize}\,\{f(w) - g(w) | w \in \mathscr{U}\}.$$

Our next task is construction of the conjugate functionals f', g'. Since by hypothesis $\mathscr{U}$ is reflexive, the domains of f' and g' can be taken as $\mathscr{U}'$. By definition, for any $u' \in \mathscr{U}'$,

$$\begin{aligned} f'(u') &= \sup\,\{\langle u', w\rangle - \|\bar{u} - w\| | w \in \mathscr{U}\} \\ &= \langle u', \bar{u}\rangle + \sup\,\{-\langle u', u\rangle - \|u\| \, | u \in \mathscr{U}\} \\ &= \langle u', \bar{u}\rangle + \sup\,\{\|u\| \cdot [\|u'\| - 1] | \|u\| \in [0, \infty)\}. \end{aligned}$$

Clearly,

$$f'(u') = \begin{cases} \langle u', \bar{u}\rangle & \text{if } \|u'\| \leqq 1, \\ +\infty & \text{otherwise}. \end{cases}$$

We now turn to g'. Again by definition, for any $u' \in \mathscr{U}'$,

$$g'(u') = \inf\,\{\langle u', u\rangle | u \in \mathscr{N}(\mathscr{L})\}.$$

By a standard result [4, p. 226, Thm. 4.6-C] $\mathscr{N}(\mathscr{L}) = \{\overline{\mathscr{R}(\mathscr{L}')}\}^{\perp}$ (since $\mathscr{U}$, X are reflexive $\mathscr{L}'$ is isometrically isomorphic with the adjoint of $\mathscr{L}$). In simple consequence of the definition of the orthogonal complement the infimum is 0 if $u \in {}^{\perp}\{\{\overline{\mathscr{R}(\mathscr{L}')}\}^{\perp}\}$, and $-\infty$ otherwise. But $\overline{\mathscr{R}(\mathscr{L}')} = {}^{\perp}\{\{\overline{\mathscr{R}(\mathscr{L}')}\}^{\perp}\}$ [4, p. 225, Thm. 4.6-A] whence

$$g'(u') = \begin{cases} 0 & \text{if } u \in \overline{\mathscr{R}(\mathscr{L}')}, \\ -\infty & \text{otherwise}. \end{cases}$$

The dual problem

$$\text{(P}'\text{)} \qquad \text{maximize}\,\{g'(u') - f'(u') | u' \in \mathscr{U}'\}$$

therefore becomes

$$\text{(P}')\qquad \begin{cases} \text{maximize } -\langle u', \bar{u}\rangle \text{ over } u' \in \mathscr{U}' \\ \text{subject to } \|u'\| \leqq 1 \text{ and } u' \in \overline{\mathscr{R}(\mathscr{L}')}. \end{cases}$$

Notice that $w = \theta$ is a point in the effective domain of g at which f is continuous. We conclude from Lemma A.1 that there exists some proper solution u'_0 to (P′) and that

$$\|u_0\| = -\langle u'_0, \bar{u}\rangle,$$

where $u_0 = \bar{u} - w_0$ and w_0 is any solution to (P).

Study of the steps in the computation of f' and g' reveals that u_0 satisfies the alignment conditions of Lemma A.1 if and only if

$$u_0 \| - u'_0 \quad \text{and} \quad x_d = \mathscr{L}u_0.$$

Let us show that $u'_0 \neq \theta$, whence the characterization of u_0 is nontrivial. Let μ be a nonzero element in X' such that $-\mu \| x_d$ (remember that X is reflexive). Since $\mathscr{L}$ is controllable, the null space of $\mathscr{L}'$ is simply $\{\theta\}$ (we have used the result $\mathscr{N}(\mathscr{L}') = \{\mathscr{R}(\mathscr{L})\}^{\perp}$); the magnitude of μ can be chosen therefore so that $0 < \|\mathscr{L}'\mu\| \leqq 1$. Then $\mathscr{L}'\mu$ satisfies the constraints on the dual problem. Further

$$-\langle \mathscr{L}'\mu, \bar{u}\rangle = \langle -\mu, \mathscr{L}\bar{u}\rangle = \langle -\mu, x_d\rangle = \|\mu\| \cdot \|x_d\| > 0.$$

Thus the maximum in the dual problem is greater than zero. Clearly $u'_0 \neq \theta$. Indeed since the objective functional of the dual problem is linear in u', we have that $\|u'_0\| = 1$.

We have shown that the problem

$$\text{(P}')\qquad \begin{cases} \text{maximize } \langle -u', \bar{u}\rangle \text{ over } u' \in \mathscr{U}' \\ \text{subject to } \|u'\| \leqq 1 \text{ and } u' \in \overline{\mathscr{R}(\mathscr{L}')} \end{cases}$$

has a proper solution u'_0 with $\|u'_0\| = 1$; that u_0 solves the basic problem if and only if (i) $u \| -u'_0$ and (ii) $x_d = \mathscr{L}u$, and that any solution to the basic problem u_0 satisfies $\|u_0\| = -\langle u_0, \bar{u}\rangle$.

Proposition 11.2 is merely an expansion of this statement, distinguishing between the cases when u_0 lies, and fails to lie, in the range of $\mathscr{L}$ (notice that we have changed variables $u \to -u$ and avoided explicit mention of $\bar{u}$).

REFERENCES

[1] A. V. BALAKRISHNAN, *An operator theoretic formulation of a class of control problems and a steepest descent method of solution*, this Journal, 1 (1963), pp. 109–127.

[2] ———, *Optimal control in Banach spaces*, this Journal, 3 (1965), pp. 152–180.

[3] W. A. PORTER, *Modern Foundations of Systems Engineering*, Macmillan, New York, 1966.

[4] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.

[5] J. HORVATH, *Topological Vector Spaces and Distributions*, Addison–Wesley, Reading, Mass., 1966.

[6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, vol. 1, Interscience, New York, 1958.

[7] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1958.

[8] R. B. Vinter, *Application of duality theory to a class of composite cost control problems*, J. Optimization Theory Appl., to appear.

[9] ———, *Optimum control of linear systems with norm constraints*, Ph.D. thesis, Cambridge University, Cambridge, England, 1972.

[10] J. A. Clarkson, *Uniformly convex spaces*, Trans. Amer. Math. Soc., 40 (1936), pp. 396–414.

[11] T. M. Apostol, *Mathematical Analysis*, Addison-Wesley, Reading, Mass., 1957.

[12] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups*, Colloquium Publications, Amer. Math. Soc., Providence, R.I., 1957.

[13] R. T. Rockafellar, *Extension of Fenchel's duality theory for convex functions*, Duke Math J., 33 (1966), pp. 81–89.

[14] ———, *Duality and stability in extremum problems involving convex functions*, Pacific J. Math, 21 (1967), pp. 167–187.

SIAM J. CONTROL
Volume 12, No. 1, February 1974

DIFFERENTIAL GAMES OF SURVIVAL WITH SPACE-LIKE TERMINAL SET*

RONALD J. STERN†

Abstract. For differential games of survival in which the boundary may be represented as a smooth curve, we give a method for transforming a survival game into a fixed duration game. Comparisons between the upper and lower values of the original game and its transform are derived, and applications of the transformation are given.

1. Introduction. The approach to differential game theory used in this paper is that of A. Friedman [2]–[4]. In this paper we give a procedure for transforming a certain differential game of survival into a differential game of fixed duration, and derive relations between the upper and lower values of these games. We restrict our attention to survival games in which the capture set is space-like; that is, its boundary can be represented as a smooth curve of the form $t = r(x)$.

In § 2 we give the requisite preliminaries. In the third section we give two results concerning δ-games which are needed in what follows. The transformed game and the main results are presented in § 4. Section 5 consists of examples.

2. Preliminaries. Consider a system of m ordinary differential equations

$$\dot{x} = f(t, x, y, z), \qquad t_0 \leqq t \leqq T_0, \tag{2.1}$$

with an initial condition

$$x(t_0) = x_0. \tag{2.2}$$

Let Y and Z be compact subsets of the Euclidean spaces R^n and R^q, respectively. The controls $y(t)$ and $z(t)$ are Lebesgue measurable functions taking values almost everywhere in Y and Z, respectively, defined on $[t_0, T_0]$.

Let F be a closed subset of R^{m+1} such that

$$F \supset [T_1, \infty) \times R^m \quad \text{for some } t_0 < T_1 \leqq T_0. \tag{2.3}$$

Given a trajectory $x = x(t)$, define the *capture time* as

$$\tilde{t} = \tilde{t}(x) = \inf\{t : (t, x(t)) \in F\}. \tag{2.4}$$

Consider a payoff functional

$$P(y, z) = g[x(\tilde{t})] + \int_{t_0}^{\tilde{t}(x)} h(t, x, y, z)\, dt. \tag{2.5}$$

The differential game associated with (2.1)–(2.5) is called a *game of survival*, denoted G.

* Received by the editors December 19, 1972, and in revised form April 2, 1973.

† Department of Applied Mathematics, Technion-Israel Institute of Technology in Haifa, Israel.

Remark 1. Survival games are the most general type of two person, zero-sum differential game, i.e.,

(i) If $g = 0$ and $h \geqq 0$, then G is a game of *generalized pursuit-evasion.*

(ii) If $g = 0$ and $h = 1$, then G is a game of *pursuit-evasion.*

(iii) If $F = [T_0, \infty) \times R^m$, then $\tilde{t} = T_0$. We then call G a *fixed duration* game.

The following assumptions guarantee the existence and uniqueness of a solution of (2.1)–(2.2) for each pair of controls (see [1] and [3]):

(a) $f(t, x, y, z)$ is continuous on $[t_0, T_0] \times R^m \times Y \times Z$.

(b) There exists $k(t) \in L^1(t_0, T_0)$ such that

$$|f(t, x, y, z)| \leqq k(t)(1 + |x|)$$

for all $(t, x, y, z) \in [t_0, T_0] \times R^m \times Y \times Z$.

(c) For each $R > 0$, there exists $k_R(t) \in L^1(t_0, T_0)$ such that

$$|f(t, x, y, z) - f(t, \bar{x}, y, z)| \leqq k_R(t)|x - \bar{x}|$$

for all $t \in [t_0, T_0]$, $y \in Y$, $z \in Z$, $|x| \leqq R$ and $|\bar{x}| \leqq R$.

Regarding the payoff, we state the following assumptions:

(d) $h(t, x, y, z)$ is continuous on $[t_0, T_0] \times R^m \times Y \times Z$.

(e) g is continuous on R^m.

If (a)–(e) hold, then $P(y, z)$ is well-defined on the space of controls. Let n be any positive integer, and $\delta = (T - t_0)/n$. Let

$$I_j = (t_{j-1}, t_j) \quad \text{for } t_j = t_0 + j\delta, \quad 1 \leqq j \leqq n.$$

Define Y_j and Z_j to be the classes of measurable functions on I_j which almost everywhere take values in Y and Z, respectively.

Let $\Gamma^{\delta,j}$ be any map of $Z_1 \times Y_1 \times Z_2 \times Y_2 \times \cdots \times Y_{j-1} \times Z_j$ into Y_j. We then call the n-tuple

$$\Gamma^\delta = (\Gamma^{\delta,1}, \cdots, \Gamma^{\delta,n})$$

an *upper δ-strategy* for y. Similarly, we define an upper δ-strategy for z, Δ^δ whose components $\Delta^{\delta,j}$ are maps from $Y_1 \times Z_1 \times Y_2 \times Z_2 \times \cdots \times Z_{j-1} \times Y_j$ into Z_j.

For $2 \leqq j \leqq n$, let $\Gamma_{\delta,j}$ be any map from $Y_1 \times Z_1 \times Y_2 \times Z_2 \times \cdots \times Y_{j-1} \times Z_{j-1}$ into Y_j, and let $\Gamma_{\delta,1}$ be any element of Y_1. We then call the n-tuple

$$\Gamma_\delta = (\Gamma_{\delta,1}, \cdots, \Gamma_{\delta,n})$$

a *lower-δ-strategy* for y. Analogously, one defines a lower δ-strategy Δ_δ for z.

Given a pair $(\Delta_\delta, \Gamma^\delta)$, we uniquely obtain control functions (z_δ, y^δ), and a trajectory x^δ. Then (z_δ, y^δ) is called the *outcome* of $(\Delta_\delta, \Gamma^\delta)$. Analogously, a pair $(\Gamma_\delta, \Delta^\delta)$ yields an outcome (y_δ, z^δ) and a trajectory x_δ.

The *upper δ-value* of G is the number

$$V^\delta(G) = \inf_{\Delta_{\delta,1}} \sup_{\Gamma^{\delta,1}} \inf_{\Delta_{\delta,2}} \sup_{\Gamma^{\delta,2}} \cdots \inf_{\Delta_{\delta,n}} \sup_{\Gamma^{\delta,n}} P[\Delta_\delta, \Gamma^\delta],$$

and the *lower δ-value* of the game is defined by

$$V_\delta(G) = \sup_{\Gamma_{\delta,1}} \inf_{\Delta^{\delta,1}} \sup_{\Gamma_{\delta,2}} \inf_{\Delta^{\delta,2}} \cdots \sup_{\Gamma_{\delta,n}} \inf_{\Delta^{\delta,n}} P[\Gamma_\delta, \Delta^\delta].$$

We say the differential game has *value* $V(G)$ if the limits $\lim_{\delta\to 0} V_\delta(G) = V^-(G)$ and $\lim_{\delta\to 0} V^\delta(G) = V^+(G)$ exist and are equal. $V^-(G)$ and $V^+(G)$ are the *lower* and *upper values* of G, respectively.

Proof of the following theorem may be found in [3]:

THEOREM 2.1. *Let assumptions* (a)–(e) *hold. Then*

$$V^\delta(G) = \inf_{\Delta_\delta}\sup_{\Gamma^\delta} P[\Delta_\delta, \Gamma^\delta] = \sup_{\Gamma^\delta}\inf_{\Delta_\delta} P[\Delta_\delta, \Gamma^\delta], \tag{2.6}$$

and

$$V_\delta(G) = \sup_{\Gamma_\delta}\inf_{\Delta^\delta} P[\Gamma_\delta, \Delta^\delta] = \inf_{\Delta^\delta}\sup_{\Gamma_\delta} P[\Gamma_\delta, \Delta^\delta]. \tag{2.7}$$

Some further assumptions now will be stated.

(f) F has C^2 boundary ∂F and

$$\begin{aligned} &v_0 + \min_Z \max_Y \sum_{j=1}^m v_j f_j(t, x, y, z) < 0, \\ &v_0 + \min_Y \max_Z \sum_{j=1}^m v_j f_j(t, x, y, z) < 0 \end{aligned} \tag{2.8}$$

for all $(t, x) \in \partial F$, where $v = (v_0, v_1, \cdots, v_m)$ is the outward normal to ∂F at (t, x), and the f_j are the component functions of f.

$$\begin{aligned} &\max_Y \min_Z \{f(t, x, y, z)\cdot p + h(t, x, y, z)\} \\ &= \min_Z \max_Y \{f(t, x, y, z)\cdot p + h(t, x, y, z)\} \end{aligned} \tag{g}$$

for all $p \in R^m$.

THEOREM 2.2. *Let assumptions* (a)–(f) *hold. Then*

(i) *$V^+(G)$ and $V^-(G)$ exist.*

(ii) *If also* (g) *holds, then G has value.*

The proof (i) follows from arguments in [3], in particular §§ 3.3, 5.2, and Problem 3.4.3. Part (ii) is proven in [2] and [4] in case G is of fixed duration. Extensions of the results of [4] found in [5] yield (ii) for the general survival case.

3. Two lemmas on δ-games.

DEFINITION 1. B^y_δ is the class of Γ_δ such that for any Δ^δ, the y-outcome, y_δ, of $(\Gamma_\delta, \Delta^\delta)$ is a *Borel measurable* function. B^z_δ is defined similarly.

LEMMA 3.1. *Let G be a game of survival satisfying* (a)–(e). *We then have*

$$V^\delta(G) = \inf_{\Delta_\delta \in B^z_\delta} \sup_y P[\Delta_\delta, y] \tag{3.1}$$

and

$$V_\delta(G) = \sup_{\Gamma_\delta \in B^y_\delta} \inf_z P[\Gamma_\delta, z]. \tag{3.2}$$

(Here "$\sup_y$" means the sup taken over all admissible controls for the player y; similarly for "$\inf_z$".

Proof. From Theorem 2.1, it is easily seen that

$$V^\delta(G) = \inf_{\Delta_\delta}\sup_y P[\Delta_\delta, y], \tag{3.3}$$

and

$$V_\delta(G) = \sup_{\Gamma_\delta}\inf_z P[\Gamma_\delta, z]. \tag{3.4}$$

We will use (3.3) to prove (3.1). Statement (3.2) will follow similarly from (3.4).

Let $\varepsilon > 0$ be given. Then (3.3) implies the existence of $\tilde{\Delta}_\delta$ such that

$$V^\delta(G) \geqq P[\tilde{\Delta}_\delta, y] - \varepsilon \tag{3.5}$$

for all controls $y = y(t)$.

Given any Lebesgue measurable function r valued in R^q on an interval I, there exists a Borel measurable function, say $b[r]$, which agrees with r almost everywhere on I. We will use the above fact in order to modify $\tilde{\Delta}_\delta$ into $\tilde{\tilde{\Delta}}_\delta$, which is a member of B_δ^z.

Let $y = (y_1, y_2, \cdots, y_n)$ be any control for player y. Define

(i) $\tilde{\tilde{\Delta}}_{\delta,1} = b(\tilde{z}_1]$, where $\tilde{z}_1 = \tilde{\Delta}_{\delta,1}$,

(ii) $\tilde{\tilde{\Delta}}_{\delta,2}(\tilde{\tilde{z}}_1, y_1) = b[\tilde{z}_2]$, where $\tilde{z}_2 = \tilde{\Delta}_{\delta,2}(\tilde{z}_1, y_1)$,

(iii) $\tilde{\tilde{\Delta}}_{\delta,j}(\tilde{\tilde{z}}_1, y_1, \tilde{\tilde{z}}_2, y_2, \cdots, \tilde{\tilde{z}}_{j-1}, y_{j-1}) = b[\tilde{z}_j]$, where $\tilde{z}_j = \tilde{\Delta}_{\delta,j}(z_1, y_1, z_2, y_2, \cdots, z_{j-1}, y_{j-1})$, for $2 \leqq j \leqq n$.

Define $\tilde{\tilde{\Delta}}_\delta = (\tilde{\tilde{\Delta}}_{\delta,1}, \tilde{\tilde{\Delta}}_{\delta,2}, \cdots, \tilde{\tilde{\Delta}}_{\delta,n})$. It is not difficult to see that $P[\tilde{\tilde{\Delta}}_\delta, y] = P[\tilde{\Delta}_\delta, y]$ for all controls y. This fact and the fact that ε in (3.5) was arbitrary yield (3.3). Statement (3.4) follows similarly.

DEFINITION 2. S_δ^y is the class of Γ_δ such that for any Δ^δ, the y-outcome, y_δ, of $(\Gamma_\delta, \Delta^\delta)$ is a *step function*. We define S_δ^z similarly.

LEMMA 3.2. *Let G be a game of fixed duration satisfying* (a)–(e). *We then have*

$$V^\delta(G) = \inf_{\Delta\delta \in S_\delta^z} \sup_y P[\Delta_\delta, y], \tag{3.6}$$

and

$$V_\delta(G) = \sup_{\Gamma_\delta \in S_\delta^y} \inf_z P[\Gamma_\delta, z]. \tag{3.7}$$

Proof. Let $\gamma > 0$ and $\alpha > 0$ be given. Given any Lebesgue measurable function r valued in R on an interval I, there exists a step function $s[r]$ and a measurable subset $J \subset I$ such that

$$|s[r] - r| \leqq \gamma \quad \text{for all } t \in I - J$$

and meas $J \leqq \alpha$.

We will prove (3.6), the case for (3.7) being similar. Let $\tilde{\Delta}_\delta$ be as in the proof of Lemma 3.1, where $\varepsilon > 0$ was given. Proceeding as we did there, we can define a lower δ-strategy $\tilde{\tilde{\Delta}}_\delta$ with the following property: given any control y, there exists a measurable set $J \subset [t_0, T_0]$ such that

$$|\tilde{z}_\delta - \tilde{\tilde{z}}_\delta| \leqq \gamma \quad \text{for all } t \in I - J, \tag{3.8}$$

and

$$\text{meas } J \leqq n\alpha, \tag{3.9}$$

where $\tilde{z}_\delta$ is the z-outcome of $(\tilde{\Delta}_\delta, y)$ and $\tilde{\tilde{z}}_\delta$ is the z-outcome of $(\tilde{\tilde{\Delta}}_\delta, y)$. We claim that

$$|P(y, \tilde{z}_\delta) - P(y, \tilde{\tilde{z}}_\delta)| \leqq O(\gamma, n\alpha) \tag{3.10}$$

for a positive function $O(\cdot, \cdot)$ which does not depend on the control y and tends to zero with its arguments.

Statement (3.10), (3.5) and the arbitrariness of ε, γ and α will yield (3.6). Thus it remains to establish (3.10) from (3.8) and (3.9). We will show that for a positive constant C,

$$\sup_{t \in [t_0, T_0]} |\tilde{x}_\delta - \tilde{\tilde{x}}_\delta| \leqq C(\gamma + n\alpha), \tag{3.11}$$

where $\tilde{x}_\delta$ is the trajectory associated with $(\tilde{\Delta}_\delta, y)$ and $\tilde{\tilde{x}}_\delta$ is the trajectory associated with $(\tilde{\tilde{\Delta}}_\delta, y)$. Verification of (3.10) is then quite routine upon making use of the continuity of h. To verify (3.11) we write

$$\begin{aligned} |\dot{\tilde{x}}_\delta - \dot{\tilde{\tilde{x}}}_\delta| &= |f(t, \tilde{x}_\delta, y, \tilde{z}_\delta) - f(t, \tilde{\tilde{x}}_\delta, y, \tilde{\tilde{z}}_\delta)| \\ &\leqq |f(t, \tilde{x}_\delta, y, \tilde{z}_\delta) - f(t, \tilde{x}_\delta, y, \tilde{\tilde{z}}_\delta)| + |f(t, \tilde{x}_\delta, y, \tilde{\tilde{z}}_\delta) - f(t, \tilde{\tilde{x}}_\delta, y, \tilde{\tilde{z}}_\delta)|. \end{aligned} \tag{3.12}$$

From (3.8), (3.9), the continuity of f, the fact that the family of trajectories of G is uniformly bounded, and condition (c) of § 2, we have the existence of a positive constant B and a positive integrable function $\bar{k}(t)$ such that

$$|\dot{\tilde{x}}_\delta - \dot{\tilde{\tilde{x}}}_\delta| \leqq B(\gamma + n\alpha) + \bar{k}(t)|\tilde{x}_\delta - \tilde{\tilde{x}}_\delta|. \tag{3.13}$$

From a fundamental inequality of differential equations (see [1, p. 37]), we then obtain

$$\sup_{t \in [t_0, T_0]} |\tilde{x}_\delta - \tilde{\tilde{x}}_\delta| \leqq B(\gamma + n\alpha) \exp\left(\int_{t_0}^{T_0} \bar{k}(t)\, dt\right), \tag{3.14}$$

completing the proof.

4. Transformed survival games. Let G denote the game of survival introduced in § 2. Let R_0 denote a uniform bound on the family of trajectories for G. Let

$$c = \max_{[t_0, T_0] \times \{|x| \leqq R_0\} \times Y \times Z} |f(t, x, y, z)|.$$

We will require the following condition in much of what follows:

(F_r) ∂F is represented by an equation $t = r(x)$ where r is continuously differentiable on $\{|x| \leqq R_0\}$ and is piecewise twice continuously differentiable on $\{|x| \leqq R_0\}$. Furthermore, the gradient of r satisfies

$$\max_{\{|x| \leqq R_0\}} |r_x(x)| < 1/c.$$

Remark 2. An elementary argument shows that (F_r) implies (f) of § 2.

We now define a certain *coordinate transformation* $(t, x) \to (t', x')$ which transforms ∂F into the surface $t' = r(x_0)$, via the relations

$$t' = t + r(x_0) - r(x) \tag{4.1}$$

and

$$x' = x. \tag{4.2}$$

Let $y(t)$ and $z(t)$ be a given pair of control functions, and let $x(t)$ be the corresponding trajectory. Using the above transformation, let us write

$$(t, x(t)) \to (t', x'(t')). \tag{4.3}$$

Define

$$t'(t) = t + r(x_0) - r(x(t)). \tag{4.4}$$

Notice that

$$x(t) = x'(t'(t)).$$

LEMMA 4.1. *Let G satisfy assumptions* (a)–(e) *and* (F_r). *Then the following all hold*:

(i) *$t'(t)$ is strictly monotone increasing on $[t_0, T_0]$.*

(ii) *$x'(t')$ is uniformly Lipschitz continuous on $[t_0, r(x_0)]$ and thus is differentiable almost everywhere on $[t_0, r(x_0)]$.*

(iii) *At almost every $t' \in [t_0, r(x_0)]$, we have*

$$\frac{dx'(t')}{dt'} = \frac{f[t' - r(x_0) + r(x'(t')), x'(t'), y(t' - r(x_0) + r(x'(t'))), z(t' - r(x_0) + r(x'(t')))]}{\begin{aligned}1 - r_x[x'(t')]\cdot f[t' - r(x_0) + r(x'(t')), x'(t'),&\\ y(t' - r(x_0) + r(x'(t'))), z(t' - r(x_0) + r(x'(t')))]&\end{aligned}}$$

(iv) $P(y, z) = g(x'(r(x_0)))$

$$+ \int_{t_0}^{r(x_0)} \frac{h[t' - r(x_0) + r(x'(t'))\, x'(t'), y(t' - r(x_0) + r(x'(t'))), z(t' - r(x_0) + r(x'(t')))]}{\begin{aligned}1 - r_x[x'(t')]\cdot f[t' - r(x_0) + r(x'(t')), x'(t'),&\\ y(t' - r(x_0) + r(x'(t'))), z(t' - r(x_0) + r(x'(t')))]&\end{aligned}}\, dt'.$$

Proof. Let $t_2 > t_1$, where t_2 and t_1 are in the interval $[t_0, T_0]$. We have

$$|r(x(t_2)) - r(x(t_1))| \leqq c \max_{\{|x| \leqq R_0\}} |r_x(x)|(t_2 - t_1) < t_2 - t_1, \tag{4.5}$$

by (F_r). This proves (i).

Let t'_1 and t'_2 be in the interval $[t_0, r(x_0)]$. We have

$$\begin{aligned}|x'(t'_2) - x'(t'_1)| &= |x(t'_2 - r(x_0) + r(x'(t'_2))) - x(t'_1 - r(x_0) + r(x'(t'_1)))| \\ &\leqq c|t'_2 - t'_1| + c \max_{\{|x| \leqq R_0\}} |r_x(x)| \cdot |x'(t'_2) - x'(t'_1)|,\end{aligned} \tag{4.6}$$

and therefore

$$|x'(t'_2) - x'(t'_1)| \leqq \frac{c}{1 - c \max\limits_{\{|x| \leqq R_0\}} |r_x(x)|} |t'_2 - t'_1|. \tag{4.7}$$

This proves (ii).

We have, by the chain rule,

$$\begin{aligned}\frac{dx(t)}{dt} &= \frac{d}{dt'} x'(t'(t)) \cdot \frac{d}{dt} t'(t) \\ &= \frac{d}{dt'} x'(t'(t)) \left[1 - r_x(x(t)) \cdot \frac{dx(t)}{dt}\right].\end{aligned} \tag{4.8}$$

Straightforward substitutions now yield (iii) and (iv), completing the proof.

Define now the following functions:

$$f'(t', x', y', z') = \frac{f(t' - r(x_0) + r(x'), x', y', z')}{1 - r_x(x') \cdot f(t' - r(x_0) + r(x'), x', y', z')},$$

$$h'(t', x', y', z') = \frac{h(t' - r(x_0) + r(x'), x', y', z')}{1 - r_x(x') \cdot f(t' - r(x_0) + r(x'), x', y', z')}.$$

We now will define a certain game of *fixed duration*, G'. The dynamics of G' are

$$dx'/dt' = f'(t', x', y', z'), \tag{4.9}$$

with initial condition

$$x'(t_0) = x_0. \tag{4.10}$$

The payoff is given by

$$P'(y', z') = g(x(r(x_0))) + \int_{t_0}^{r(x_0)} h'(t', x', y', z')\, dt'. \tag{4.11}$$

Lemma 4.1 implies that

$$P(y, z) = P'(y', z'), \tag{4.12}$$

where

$$y(t) = y'(t'(t)), \quad z(t) = z'(t'(t)). \tag{4.13}$$

THEOREM 4.1. *Let G satisfy conditions* (a)–(e), (2.3), (F_r), *and let G' satisfy conditions* (a)–(e). *Then*

$$V^-(G') \leqq V^+(G) \tag{4.14}$$

and

$$V^-(G) \leqq V^+(G'). \tag{4.15}$$

Proof. We will give the details for the proof of (4.14). The proof of (4.15) is similar.

Let $\eta > 0$ be such that

$$\max_{\{|x| \leqq R_0\}} |r_x(x)| \leqq 1/c + \eta. \tag{4.16}$$

Let $\delta = (T_0 - t_0)/n$ and $\delta' = (r(x_0) - t_0)/n'$, where n' and n are taken to be such that

$$(3 + c\eta)\delta \leqq \delta'. \tag{4.17}$$

Since the upper and lower values of G and G' are guaranteed to exist by Theorem 2.2, then (4.14) will hold if it can be proven that

$$\sup_{\Gamma_{\delta'} \in S_{\delta'}^y} \inf_{z'} P'[\Gamma_{\delta'}, z'] \leqq \inf_{\Delta_\delta \in B_\delta^z} \sup_y P[\Delta_\delta, y]. \tag{4.18}$$

Let $\Gamma_{\delta'} \in S_{\delta'}^y$ and $\Delta_\delta \in B_\delta^z$ be arbitrary. Then (4.18) will hold if there exist controls z' and y such that

$$P'[\Gamma_{\delta'}, z'] = P[\Delta_\delta, y]. \tag{4.19}$$

Let y' denote the y-outcome of $(\Gamma_{\delta'}, z')$ and let x' denote the corresponding path. Let z denote the z-outcome of (Δ_δ, y), and let x denote the corresponding path. Suppose we can construct z' and y such that

$$z'(t'(t)) = z(t), \qquad t \in [t_0, T_0], \tag{4.20}$$

and

$$y'(t'(t)) = y(t), \qquad t \in [t_0, T_0], \tag{4.21}$$

where $t'(t) = t + r(x_0) - r(x(t))$.

By Lemma 4.1, we then would have (4.19), which is all we require.

Denote the intervals of game G by I_j, $j = 1, \cdots, n$, and denote the intervals of G' by $I'_{j'}$, $j' = 1, \cdots, n'$. Here $I_j = (t_{j-1}, t_j)$ and $I'_{j'} = (t'_{j'-1}, t'_{j'})$.

The component $\Gamma_{\delta',1}$ of $\Gamma_{\delta'}$ determines $y'(t')$ on I'_1, and this is a step function given by

$$y'(t') = \sum_{i=1}^{k} c_{i,1}\psi_{C_{i,1}}.$$

Here $\psi_{C_{i,1}}$ is the characteristic function of $C_{i,1}$,$\cdot i = 1, \cdots, k$. The $C_{i,1}$ are disjoint intervals $C_{i,1} = (t'_{i-1}, t'_{i,1})$ such that

$$\bigcup_{i=1}^{k} C_{i,1} = I'_1.$$

Here $t'_{0,1} = t_0$ and $t'_{k,1} = t'_1$. The $c_{i,1}$ are constants.

The component $\Delta_{\delta,1}$ of Δ_δ determines $z(t)$ on I_1. We will now construct $y(t)$ on I_1, *making use of the trajectory* $x(t)$ *as it evolves.*

By Lemma 4.1(i), we can define the following:

$$t(t'_{i,1}) = \text{the unique } t \text{ such that } t'_{i,1} = t + r(x_0) - r(x(t)).$$

Now define the following function on I_1:

$$y(t) = \begin{cases} c_{1,1} & \text{for } t_0 \leqq t \leqq t(t'_{1,1}), \\ c_{2,1} & \text{for } t(t'_{1,1}) \leqq t \leqq t(t'_{2,1}), \\ \quad\vdots & \\ c_{p,1} & \text{for } t(t'_{p-1,1}) \leqq t \leqq t_1, \end{cases} \tag{4.22}$$

where $p \leqq k$ due to

$$|t'_{i,1} - t'_{i-1,1}| \leqq (2 + c\eta)|t(t'_{i,1}) - t(t'_{i-1,1})| \tag{4.23}$$

and (4.17).

The inequality (4.23) follows from

$$t(t'_{i,1}) - t(t'_{i-1,1}) = t'_{i,1} - t'_{i-1,1} + r(x(t(t'_{i,1}))) - r(x(t(t'_{i-1,1}))) \tag{4.24}$$

and assumption (F_r).

Thus $y(t)$ has been constructed on I_1 by using $y'(t')$ on only the first part of the interval I'_1.

The map $\Delta_{\delta,2}$ gives $z(t)$ on I_2. We continue in the above manner until $y(t)$ has been constructed up to time $t(t'_{k,1})$, which occurs in some interval $I_m, m > 1$.

In the construction we made use of the maps $\Delta_{\delta,j}, j = 1, \cdots, m$, and the trajectory x up to time $t(t'_{k,1})$.

We now use $z(t), t \in [t_0, t(t'_{k,1})]$ to define $z'(t')$ on $[t_0, t'_{k-1}]$, as follows:

$$(4.25) \qquad z'(t') = z(t' - r(x_0) + r(x(t(t')))), \qquad t' \in I'_1.$$

The Borel measurability of z implies that z' thus defined is a measurable function.

Thus $y(t)$ is constructed on $I_1, I_2, \cdots, I_m$, and $z'(t')$ is constructed on I'_1. We continue in this fashion until $t = \tilde{t}$, which implies $t'(t) = r(x_0)$.

The controls which have been constructed can be seen to satisfy relations (4.20) and (4.21), completing the proof of (4.14). The proof of (4.15) is analogous to the arguments above.

Remark 3. By Theorem 4.1, it is clear that if both G and G' have value, then their values are equal.

5. Applications. In the first two examples which follow, Theorem 4.1 is applied to certain classes of differential games, and insights are readily obtained. In the third example, a specific type of "attrition-attack" model is analyzed by employing Theorem 4.1, and a synthesized saddle point is obtained. (See [3] for terminology.)

Example 1. Let G be a game of pursuit-evasion. It follows that in this case, we have

$$P'(y', z') = \int_{t_0}^{r(x_0)} \frac{dt'}{1 - r_x(x') \cdot f(t' - r(x_0) + r(x'), x', y', z')}.$$

Suppose G has value. We immediately obtain the following estimate for $V(G)$:

$$\max_Y \min_Z \min_{[t_0, r(x_0)] \times \{|x| \leqq R_0\}} h'(t', x', y', z') \leqq \frac{V(G)}{r(x_0) - t_0}$$
$$\leqq \min_Z \max_Y \max_{[t_0, r(x_0)] \times \{|x| \leqq R_0\}} h'(t', x', y', z') ,$$

where h' is the integrand of $P'(y', z')$.

Example 2. Let G be survival game with linear terminal set, let w denote the gradient $r_x(x)$ of the terminal surface, where this is now a constant vector. Furthermore, assume that the game has integral payoff and that x does not appear in f or h. In other words, the trajectory is manifest in the payoff only via $t = \tilde{t}(x)$. It then follows that the payoff of G' is given by

$$P'(y', z') = \int_{t_0}^{r(x_0)} \frac{h(y', z')}{1 - w \cdot f(y', z')} dt' = \int_{t_0}^{r(x_0)} h'(y', z') \, dt'.$$

Thus the problem G' is dynamics free, and the value, if it exists, is given by

$$\inf_{z'} \sup_{y'} P'(y', z') = \sup_{y'} \inf_{z'} P'(y', z').$$

In the next example, we consider a game which is not of the type mentioned in the preceeding example. Nevertheless, a dynamics free transformed game results. This enables us, as will be seen, to give a synthesized saddle point for G (see [3] for terminology.)

Example 3. Consider the following game. The dynamics are given by

$$\left.\begin{aligned}\dot{x}_1 &= f_1(y) - p_1(z) + m_1\\ \dot{x}_2 &= f_2(z) - p_2(y) + m_2\end{aligned}\right\} = d(y, z),$$

$$x_1(0) = x_{10},$$

$$x_2(0) = x_{20},$$

and the payoff is given by

$$P(y, z) = \int_0^{\bar{t}(x)} h(x_1(t) - x_2(t))\, dt,$$

where $\bar{t}$ is the first t such that $x_1(t) = x_2(t)$. This is a type of attrition-attack game. Other attrition-attack models are analyzed in [3] and [6]. The quantities $x(t)$ may be thought of as weapons on hand at time t. The quantity f is the part of a player's effort devoted to increasing his weapons level, while p denotes his effort at destroying the weapons of the enemy. We will assume that $x_{10} > x_{20}$, and that for any controls, the path eventually satisfies $x_1 = x_2$; at this time y will "surrender". The payoff may be thought of as the overall damage y inflicts upon z. Notice that we allow h to take negative values. Let $w = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

The capture time of this game is the first instant such that $w \cdot x(t) = 0$. Next we will put the above game into the form of the survival games discussed in the previous sections. Write

$$\hat{w} = \begin{pmatrix} a^2 w \\ ka \end{pmatrix}, \quad \hat{d} = \begin{pmatrix} d \\ 1/ka \end{pmatrix} \quad \text{and} \quad \hat{x}(t) = \begin{pmatrix} x(t) \\ t/ka \end{pmatrix},$$

where a and k are positive numbers. We have $\hat{w} \cdot \hat{x} = a^2 w \cdot x + t$.

Let $\hat{G}$ be the game with dynamics

$$\dot{\hat{x}} = \hat{d}(y, z) \qquad \hat{x}(0) = \begin{pmatrix} x(0) \\ 0 \end{pmatrix},$$

with payoff given by

$$\hat{P}(y, z) = \int_0^{\bar{t}(\hat{x})} h((t - \hat{w} \cdot \hat{x}(t))/a^2)\, dt,$$

where $\bar{t}$ is the first time the trajectory $\hat{x}(t)$ enters F. The boundary of F is given by $t = \hat{w} \cdot \hat{x}$. As easily verified, the numbers a and k may be chosen to guarantee that condition (F_r) holds. We now apply Theorem 4.1 to the game $\hat{G}$. The transformed game, which is of fixed duration, has a payoff given by

$$P'(y', z') = \int_0^{w \cdot x(0)} \frac{h((t' + w \cdot x(0)/a^2)}{1 - f(y') - p(y') + f(z') + p(z')}\, dt'$$

Thus G', the transformed game, is dynamics free. This game obviously has value and an open-loop saddle point which is determined as follows: at times t' when the numerator of P' is positive, y chooses a vector in Y which maximizes $f_1 + p_2$;

when the numerator is negative, he minimizes $f_1 + p_2$, and when the numerator is zero, he plays any control. Player z's optimal control in G' is similarly computed. Denote by $\{y'(t'), z'(t')\}$ the saddle point for G' thus constructed. By the proof of Theorem 4.1, we have that a synthesized saddle point for the original game in this example is given by $\{y'(t + w \cdot x(0) - w \cdot x(t)),\ z'(t + w \cdot x(0) - w \cdot x(t))\}$. In the attrition-attack setting, the interpretation of this saddle point is as follows: when the argument of h is positive, player y maximizes the sum of his attrition and attack effort; when the argument of h is negative, he minimizes this sum, thereby hastening the termination of the game, or the entry of the trajectory into a region where h again becomes positive. Player z's behavior under his rule may be described similarly.

Acknowledgment. The author thanks Avner Friedman and Richard Scalzo for their stimulating comments.

REFERENCES

[1] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[2] R. Elliott and N. Kalton, *The existence of value in differential games*, Mem. Amer. Math. Soc., 126 (1972).

[3] A. Friedman, *Differential Games*, Wiley-Interscience, New York, 1971.

[4] ———, *Upper and lower values for differential games*, J. Differential Equations, 12 (1972).

[5] ———, *Remarks on differential games of survival*, to appear.

[6] R. Isaacs, *Differential Games*, Wiley-SIAM, New York, 1967.

SIAM J. CONTROL
Vol. 12, No. 2, May 1974

LUCIEN W. NEUSTADT—IN MEMORIAM

Lucien W. Neustadt, who was managing editor of this journal from 1964 until his death in 1972, was one of the principal contributors to the vigorous growth and development of optimal control theory in the last twenty years. This issue of the *SIAM Journal on Control* is dedicated to the honor of his memory and contributions.

Lucien Neustadt was born in 1928 in Berlin, Germany, where his father practiced medicine. After the Nazis came to power, the family moved to France and then to the United States. He received his high school education in New York and his B.A. degree in mathematics from New York University in 1948. He received his M.S. degree from the University of Wisconsin in 1950 and his Ph.D. in mathematics from New York University in 1960. From 1948 to 1964 his studies and research overlapped his work in industry as a mathematician at Aberdeen Proving Grounds, Bell Aircraft Corporation, Reeves Instrument Corporation, T.R.W. Space Technology Laboratories and Aerospace Corporation. During 1964 he was a professor of electrical engineering at the University of Southern California. He spent the 1971–72 academic year on sabbatical leave at the Center for Dynamical Systems, Brown University. Lucien Neustadt died in Los Angeles, California on October 9, 1972 at the age of 44.

As a person, Lucien Neustadt was kind, warm, good-natured and considerate of others. He was also very courageous. Although he suffered from leukemia for five years or so before his death, he carried on a full and active scientific life and maintained his good humor. In fact, some of his most important work was done in the five years preceding his death. His death came as a shock, even to those of us who were aware of his condition. Those of us who knew him personally sorely miss a friend and colleague.

Lucien Neustadt made two types of important contributions to the growth and development of optimal control theory. First, there were his own important scientific contributions. These will be discussed later. Second, there were his contributions connected with what could be summarized as "communication" in the sense of disseminating scientific knowledge and communicating with other workers in the field. Probably his best-known contribution in this area was his work as managing editor of this journal. He was a member of the original editorial board and became managing editor with Volume 2 in 1964. He served the journal with extreme dedication. He insisted on high standards for all papers and for fair and rapid service to contributors. Nevertheless, he always dealt tactfully with authors and members of the editorial board. The present editorial board feels that the present status of this journal is due, in large measure, to Lucien Neustadt's stewardship.

Lucien Neustadt also helped organize several international symposia and conferences on optimal control theory and other areas related to optimization. He edited or assisted with the editing of the proceedings of many of these conferences.

His circle of friends in the scientific community was large, and his knowledge of their work was extensive. He shared this knowledge freely and warmly with all who were interested. He was personally responsible for establishing and developing important and lasting contacts and avenues of communication between American and Soviet workers in control theory. He developed strong personal friendships with some of the leading Soviet contributors to the field. These led to ever-widening contacts and interactions between American and Eastern European workers in control theory. He was also a vigorous advocate of freedom of movement for scientists and others.

In the early sixties Neustadt edited the translation from the Russian into English of the important book, *The Mathematical Theory of Optimal Processes*, by Boltyanskii, Gamkrelidze, Mishchenko and Pontryagin. In this book, these authors summarized their important work to date. In editing the translation, Neustadt brought the material up to date, corrected some errors, and pointed out the relationship of this work to other works. The translation of this book into English was an important stimulus to the growth of interest in optimal control theory in this country and in Western Europe.

He also edited the translation into English of the book, *Necessary Conditions for an Extremum*, by B. Pshenichnyi. This book is concerned with necessary conditions for general optimization problems, and is related to Neustadt's own interests from the mid-1960's onwards.

Lucien Neustadt was also a member of the SIAM Council and the SIAM Program Committee.

His enthusiasm for his subject and for communicating ideas extended into the classroom. In 1967 he received a University of Southern California Associates Award for excellence in teaching. The award was given by a vote of his students, who paid him high tribute.

We now turn to a brief survey of Lucien Neustadt's scientific contributions. A complete list of his scientific publications is given at the end of this article. The numbers in square brackets refer to this list.

His early papers were closely related to applications and were motivated by his industrial experience. One of the most important and influential of these was [3], in which a computational procedure for solving a minimum time control problem is presented. It was the first general method for computing "bang-bang" controls and was based on a geometric idea which he later exploited in solving other optimal control problems [7]. In the early to mid-sixties he was particularly interested in minimum effort control problems, which originated in aerospace applications. See [6], [7], [16] and [20]. In [16] and [20] a rigorous mathematical basis was established for fuel optimal impulsive controls.

Another important paper in this period is [14]. Here Neustadt showed that for systems of the form

$$dx/dt = A(t)x + h(t, u(t)), \qquad x(0) = x_0, \tag{1}$$

subject to constraints $u(t) \in \Omega$, where Ω is compact, it is true that the set of attainability is compact and convex. Note that no assumption is made as to the convexity of the sets $h(t, \Omega(t))$. From this result one gets that the problem of minimizing

$$\int_0^T [\langle a(t), x(t)\rangle + h^0(t, u(t))]\, dt$$

subject to (1) has a solution.

From the mid-1960's until his death, Neustadt was concerned with developing a general theory of necessary conditions for optimization problems. This ranks among his most important work. It appeared in various papers ([21]–[26], [29], [30], [31], [34], [39]) and is summarized and refined in the posthumous book [40].

By the mid-1960's several writers had noted similarities in the techniques and results concerned with necessary conditions in various areas of optimization. These similarities were not always on the surface, especially in the techniques used to obtain the results. The program that Neustadt and others—notably Halkin, Hestenes, Gamkrelidze, and Dubovitskii and Milyutin—had embarked on was the following:

First, formulate a very general optimization problem that would include as special cases all optimization problems of interest, such as ordinary control problems, control problems with bounded states, control problems with lags, control problems with distributed parameters, nonlinear programming problems, etc., etc.

Second, develop a meaningful set of necessary conditions for the general problem under reasonable hypotheses. The necessary conditions must be such that one obtains useful necessary conditions for the special problems when appropriate specializations and identifications are made. One of the difficulties in the formulation of the general problem is that the hypotheses must be specific enough to yield necessary conditions with some structure, yet they should be general enough to include all the special problems of interest.

Neustadt's contributions to this unified theory have been numerous. He has introduced or sharpened several of its key concepts, and in many technically difficult papers he has given the detailed application of this theory to specific hard problems. On the conceptual side, one of Neustadt's most important contributions was his introduction of "convex differentials". If an optimal control problem with bounded states is expressed as a mathematical programming problem in infinite-dimensional space, then the corresponding inequality constraints are not differentiable, i.e., they cannot be approximated by affine functions. Neustadt observed that those inequality contraints could, however, be approximated by convex functions. The exploitation of this observation led to a better and more profound theory of necessary conditions for optimal control problems with bounded states. Neustadt also applied the unified theory to control problems governed by a very general class of Volterra-type operator equations which includes certain differential difference equations, functional differential equations, and Volterra integral equations as special cases. All this material can be found in [40]. In addition to his own work, Neustadt encouraged and influenced others who were applying the theory to problems involving hereditary systems.

THE EDITORIAL BOARD

PUBLICATIONS OF L. W. NEUSTADT

[1] *A method of computing eigenvectors and eigenvalues on an analog computer*, Mathematical Tables and Other Aids to Computation, vol. XIII, no. 67, 1959, pp. 194–201.

[2] *Analog computer techniques for plotting bode and Nyqist diagrams*, with G. A. BEKEY, Proceedings of the Western Joint Computer Conference, May 1960, pp. 165–172.

[3] *Synthesizing time optimal control systems*, J. Math. Anal. Appl., 1 (1960), pp. 484–493.

[4] *The moment problem and weak convergence in L^2*, Pacific J. Math., 11 (1961), pp. 715–721.

[5] *Applications of linear and non-linear programming techniques*, Third International Analogue Computation Meetings, Opatija, Yugoslavia, 1961; published in Proceedings, Gordon and Breach, New York, pp. 197–200.

[6] *Time optimal control systems with position and integral limits*, J. Math. Anal. Appl., 3 (1961), pp. 406–427.

[7] *Minimum effort control systems*, this Journal, 1 (1962), pp. 16–31.

[8] *Simulating money allocation problems on an analog computer*, Annales de l'Association Internationale pour le Calcul Analogique, 1 (1962), pp. 37–39.

[9] *A synthesis method for optimal controls*, Optimum System Synthesis Conf., Wright–Patterson AFB, Ohio, September 1962; published in Proceedings.

[10] *The Mathematical Theory of Optimal Processes*, L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze and E. F. Mishchenko. English translation edited by L. W. Neustadt. Interscience (division of John Wiley), New York, 1962.

[11] *Discrete time optimal control systems*, International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, J. P. LaSalle and S. Lefschetz, ed., Academic Press, New York, 1963, pp. 267–283.

[12] *On synthesizing optimal controls*, Second Congress of the International Federation of Automatic Control (IFAC), Basel, Switzerland, September 1963; published in Proceedings, Butterworth, London, 1964.

[13] *An application of optimal control to midcourse guidance*, with J. S. Meditch, Second Congress of the International Federation of Automatic Control (IFAC), Basel, Switzerland, September 1963; published in Proceedings, Butterworth, London, 1964.

[14] *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

[15] *A simple reentry guidance system*, with J. A. Love, AIAA Guidance and Control Conf., M.I.T., August 1963; published in Progress in Aeronautics and Astronautics, 13 (1964).

[16] *Optimization, a moment problem and nonlinear programming*, this Journal, 2 (1964), pp. 33–53.

[17] *The status of optimal computation*, Joint Automatic Control Conference, 1964; invited paper.

[18] *Computing Methods in Optimization Problems*, edited with A. V. Balakrishnan, Academic Press, New York, 1964.

[19] *Minimum-fuel impulses for space trajectories*, IBM Scientific Computing Sympos. on Control Theory and Application, Yorktown Heights, N.Y., October 1964; invited paper. Published in Proceedings, pp. 201–208.

[20] *A general theory of minimum-fuel space trajectories*, this Journal, 3 (1965), pp. 317–356.

[21] *Optimal control problems as extremal problems in a Banach space*, Polytechnic Institute of Brooklyn Symposium on System Theory, April 1965; invited paper. Published in Proceedings, pp. 215–224.

[22] *A general theory of variational problems with applications to optimal control problems*, Dokl. Akad. Nauk. SSSR, 171 (1966), pp. 48–50.

[23] *A general theory of optimal processes with applications to optimal control problems with restricted phase coordinates and minimax problems*, Internat. Sympos. on Differential Equations and Dynamical Systems, Mayaguez, Puerto Rico, December 27–30, 1965.

[24] *An abstract variational theory with applications to a broad class of optimization problems. I: General theory*, this Journal, 4 (1966), pp. 505–527. Russian transl., Kibernetika (Kiev), 1 (1967), pp. 77–91.

[25] *An abstract variational theory with applications to a broad class of optimization problems. II: Applications*, this Journal, 5 (1967), pp. 90–137.

[26] *General necessary conditions for optimization problems*, with Hubert Halkin. Proc. Nat. Acad. Sci. U.S.A., 56 (1966), pp. 1066–1071.

[27] *Mathematical Theory of Control*, edited with A. V. Balakrishnan, Academic Press, New York, 1967.

[28] *A survey of certain aspects of control theory*, Mathematics of the Decision Sciences, part 2, G. B. Dantzig and A. F. Veinott, eds., American Mathematical Society, Providence, R.I., 1968, pp. 3–16.

[29] *Control as programming in general normed linear spaces*, with HUBERT HALKIN, Mathematical Systems Theory and Economics, H. W. Kuhn and G. P. Szego, eds., Springer-Verlag, Berlin, 1969, pp. 23–40.

[30] *Optimal control problems as mathematical programmimg in an unorthodox function space*, Control Theory and the Calculus of Variations, A. V. Balakrishnan, ed., Academic Press, New York, 1969, pp. 175–207.

[31] *A general theory of extremals*, J. Comput. System Sci. 3 (1969), pp. 57–92.

[32] *Computing Methods in Optimization Problems 2*, edited with L. A. ZADEH AND A. V. BALAKRISHNAN, Academic Press, New York, 1969.

[33] *On the solutions of certain integral-like operator equations, existence, uniqueness, and dependence theorems*, Arch. Rational Mech. Anal., 38 (1970), pp. 131–160.

[34] *Optimal control problems with operator equation restrictions*, Symposium on Optimization, A. V. Balakrishnan, et al., eds., Springer-Verlag, Berlin, 1970, p. 292–306.

[35] *Existence and continuous dependence theorems for certain operator equations*, Proc. Fifth Internat. Conf. on the Theory of Nonlinear Oscillations, Kiev, USSR, 1970.

[36] *Sufficiency conditions and a duality theory for mathematical programming problems in arbitrary linear spaces*, Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York, 1970, pp. 323–348.

[37] *Optimal control: A theory of necessary conditions*, Proc. 1970 Internat. Congress of Mathematicians, vol. 3, pp. 183–185.

[38] *Necessary Conditions for an Extremum*, by B. N. PSHENICHNYI, English translation edited by L. W. Neustadt. M. Dekker, New York, 1971.

[39] *Optimal control problems with mixed control-phase variable equality and inequality constraints*, with KAROL MAKOWSKI, this Journal, 12 (1974), pp. 184–228.

[40] *Optimization: A Theory of Necessary Conditions*, Princeton Univ. Press, Princeton, N.J., 1974.

SIAM J. CONTROL
Vol. 12, No. 2, May 1974

OPTIMAL CONTROL PROBLEMS WITH MIXED CONTROL-PHASE VARIABLE EQUALITY AND INEQUALITY CONSTRAINTS*

KAROL MAKOWSKI† AND LUCIEN W. NEUSTADT‡

Abstract. In this paper,[1] necessary conditions are obtained for optimal control problems containing equality constraints defined in terms of functions of the control and phase variables. The control system is assumed to be characterized by an ordinary differential equation, and more conventional constraints, including phase inequality constraints, are also assumed to be present. Because the first-mentioned equality constraint must be satisfied for all t (the independent variable of the differential equation) belonging to an arbitrary (prescribed) measurable set, this problem gives rise to infinite-dimensional equality constraints. To obtain the necessary conditions, which are in the form of a maximum principle, an implicit-function-type theorem in Banach spaces is derived.

1. Introduction.

This paper is devoted to a study of necessary conditions for optimal control problems with mixed control-phase variable equality constraints. Specifically, we shall consider optimal control systems whose evolution is described by an ordinary differential equation of the form

$$\dot{x}(t) = f(x(t), u(t), t), \qquad t_1 \leqq t \leqq t_2, \tag{1.1}$$

where x is an n-dimensional "phase" variable and u is an m-dimensional "control" variable. We shall suppose that x and u are further constrained by equalities of the form

$$p^i(x(t), u(t), t) = 0 \quad \text{for almost all } t \in I_1, \quad i = 1, \cdots, l, \tag{1.2}$$

I_1 being some given measurable subset of $[t_1, t_2]$. Additional, more conventional constraints of the form

$$\begin{aligned} &\chi^i(x(\tau_1), \cdots, x(\tau_\sigma)) = 0 \quad \text{for } i = 1, \cdots, k, \\ &\chi^i(x(\tau_1), \cdots, x(\tau_\sigma)) \leqq 0 \quad \text{for } i = k+1, \cdots, k+k_1, \end{aligned} \tag{1.3}$$

$$q^i(u(t), t) \leqq 0 \quad \text{for } i = 1, \cdots, r \text{ and almost all } t \in [t_1, t_2], \tag{1.4}$$

* Received by the editors May 8, 1972. This work was supported in part by the National Institutes of Health under Grant 16197-03, in part by the United States Air Force Office of Scientific Research under Grant AF-AFOSR-1029-67, in part by the National Science Foundation under Grant GK-15787, and in part by the National Aeronautics and Space Administration under Grant NGL 40-002-015.

† Department of Electrical Engineering, University of Southern California, Los Angeles, California 90007. Now at 3761 Jasmine Ave., Apt. 10, Los Angeles, California 90034.

‡ Department of Electrical Engineering, University of Southern California, Los Angeles, California 90007, and Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. Deceased October 9, 1972.

[1] The idea, most of its development, and the final written version of this paper belong to Professor Lucien Neustadt. The connection can be clearly seen with his earlier papers, of which even some notation was preserved. I took part mainly in the development of Sections 4, 6, 8, 10, 11, and 12— K. Makowski.

as well as phase inequality constraints of the form

$$\tilde{\chi}^i(x(t), t) \leqq 0 \quad \text{for all } t \in I'_i, \quad i = 1, \cdots, l_1,$$

will also be allowed in our problem ($\tau_1, \cdots, \tau_\sigma$ are fixed points in $[t_1, t_2]$, and $I'_1, \cdots, I'_l$ are given closed subsets of $[t_1, t_2]$).

The constraint (1.2) is unorthodox in that it represents an "infinite-dimensional" *equality* constraint on the problem. We shall investigate the problem by viewing it as a mathematical programming problem over a Banach space with generalized inequality and equality constraints. These constraints are defined in terms of functions which take on their values in appropriate infinite-dimensional Banach spaces.

The fact that the equality constraint takes place in an infinite-dimensional space causes very serious complications in the analysis. In order to obtain necessary conditions for mathematical programming problems with nonlinear equality constraints, it is generally necessary to invoke implicit function theorems (or fixed-point theorems) in the space which contains the range of the function defining the equality constraint. When the equality constraint is finite-dimensional, then this space is Euclidean k-space (for some positive integer k), and the Brouwer fixed-point theorem has turned out to be the appropriate fixed-point theorem to use for optimal control problems (see, e.g., [1], [2]). Here, with equality constraints of the form (1.2), (1.3), it is necessary to use not only the Brouwer theorem, but also an implicit-function-type theorem in $L^l_\infty(I_1)$ (the space of all essentially bounded functions from I_1 into Euclidean l-space). The latter implicit function theorem is based on the fixed-point theorem for contraction mappings in a Banach space.

As is to be expected, we shall have to make certain "regularity" assumptions on the constraints (1.2). Roughly speaking, we shall have to suppose that the l equality constraints (1.2) are independent with respect to the control variable (at least to first order, near the optimum control and trajectory) as well as compatible with the control inequality constraints (1.4). This means that the functions p^i must all depend explicitly on u, which makes it appear that we are excluding from consideration pure phase equality constraints of the form $p^i(x(t), t) = 0$. However, it turns out that pure phase equality constraints can often be transformed to equality constraints of the form we require if we simply differentiate the phase equality constraint with respect to t.

In the language of the classical calculus of variations, in our proofs, it will be necessary for us to make both "strong" and "weak" variations in the controls, whereas weak variations generally are redundant in problems with finite-dimensional equality constraints.

Our necessary conditions will be in the form of a maximum principle, much like the Pontryagin maximum principle. Results similar to ours have previously been obtained, primarily by Hestenes [3], [7] and Virsan [4], [5], but under more restrictive assumptions and in a somewhat weaker form. Virsan's approach to the problem had much in common with ours, but Hestenes used a finite-dimensional implicit function theorem to reduce the problem to one with finite-dimensional equality constraints. A detailed discussion comparing our results with earlier ones is given in § 15.

Problems in which the equality constraints (1.2) are supplemented by similar inequality type constraints of the form

$$p^i(x(t), u(t), t) \leqq 0 \quad \text{for almost all } t \in I_1$$

are touched upon in § 14. In fact, we shall show that such inequality constraints can be formulated in such a way that they become equality constraints, but at the expense of creating additional constraints of the form of (1.4).

2. Problem statement. Let there be given positive integers n, m, k, k_1, l, l_1, r, and σ, where $l \leqq m$. Let R denote the real line, and let R^j (for each integer $j > 1$) denote Euclidean j-space. Further, let there be given a compact interval $I = [t_1, t_2]$, an open interval $\tilde{I} \supset I$, a Lebesgue measurable subset I_1 of I, closed subsets $I'_1, \cdots, I'_{l_1}$ of I, a finite subset $\{\tau_1, \cdots, \tau_\sigma\}$ of I such that $t_1 = \tau_1 < \tau_2 < \cdots < \tau_\sigma = t_2$, and open sets G and U_0 in R^n and R^m, respectively. Finally, let there be given functions $f: G \times U_0 \times I \to R^n, p = (p^1, \cdots, p^l): G \times U_0 \times I \to R^l$, $\chi = (\chi^0, \chi^1, \cdots, \chi^{k+k_1}): (G)^\sigma \to R^{k+k_1+1}$ (where $(G)^\sigma$ denotes the direct product of G with itself σ times), $q = (q^1, \cdots, q^r): U_0 \times I \to R^r$, and $\tilde{\chi} = (\tilde{\chi}^1, \cdots, \tilde{\chi}^{l_1})$: $G \times \tilde{I} \to R^{l_1}$.

Let $\mathscr{U}_0$ denote the set of all measurable functions $u: I \to R^m$ such that (possibly neglecting a subset of I of measure zero) the closure of the set $\{u(t): t \in I\}$ is a *compact* subset of U_0. Let $\mathscr{U}$ denote the set of all functions $u \in \mathscr{U}_0$ such that

(2.1) $$q^j(u(t), t) \leqq 0 \quad \text{for } j = 1, \cdots, r \text{ and almost all } t \in I.$$

The elements of $\mathscr{U}$ will be called *admissible controls.*

Let $\mathscr{G}$ denote the set of all continuous functions $x: I \to G$.

We shall be concerned with ordinary differential equations of the form

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

where $u \in \mathscr{U}_0$. By a solution of (2.2), we mean an absolutely continuous function $x \in \mathscr{G}$ such that (2.2) holds for almost all $t \in I$. The absolutely continuous functions $x \in \mathscr{G}$ which satisfy (2.2) for some *admissible* control u, i.e., some $u \in \mathscr{U}$, will be called *admissible trajectories.*

Our problem consists in finding a pair $(x, u) \in \mathscr{G} \times \mathscr{U}$ such that

(2.3) $$\begin{gathered} x \text{ is a solution of (2.2)}, \\ \chi^i(x(\tau_1), \cdots, x(\tau_\sigma)) = 0 \quad \text{for } i = 1, \cdots, k, \end{gathered}$$

(2.4) $$\chi^i(x(\tau_1), \cdots, x(\tau_\sigma)) \leqq 0 \quad \text{for } i = k + 1, \cdots, k + k_1,$$

(2.5) $$p^i(x(t), u(t), t) = 0 \quad \text{for almost all } t \in I_1 \text{ and each } i = 1, \cdots, l,$$

(2.6) $$\tilde{\chi}^i(x(t), t) \leqq 0 \quad \text{for all } t \in I'_i \text{ and each } i = 1, \cdots, l_1,$$

and which, in so doing, achieves a minimum for $\chi^0(x(\tau_1), \cdots, x(\tau_\sigma))$.

In § 14, we shall also investigate problems where I_1 is replaced by I_i in (2.5)—$I_1, \cdots, I_l$ being given Lebesgue measurable subsets of I—as well as problems where there are additional constraints of the form

$$p^i(x(t), u(t), t) \leqq 0 \quad \text{for almost all } t \in I_1 \text{ and each } i.$$

We shall show that such problems can be reduced to the problem just described.

In order to obtain meaningful results for this problem, we shall have to make certain smoothness and continuity assumptions on the functions $f, p, \chi, \tilde{\chi}$, and q. These are as follows:

A1. The function χ is continuously differentiable.

A2. The function $\tilde{\chi}$ is twice continuously differentiable.

A3. For each $t \in I$, the functions $(x, u) \to f(x, u, t): G \times U_0 \to R^n$, $(x, u) \to p(x, u, t): G \times U_0 \to R^l$, and $u \to q(u, t): U_0 \to R^r$ are continuously differentiable.

A4. The continuity of the function $(x, u) \to f(x, u, t): G \times U_0 \to R^n$ is uniform with respect to $t \in I$, and similarly for f replaced by p, f_x, f_u, p_x, and p_u (where f_x, etc., denote the evident matrices of partial derivatives). Also, the continuity of the functions $u \to q(u, t)$ and $u \to q_u(u, t)$ is uniform with respect to $t \in I$.

A5. For each $(x, u) \in G \times U_0$, the functions $t \to f(x, u, t): I \to R^n$, $t \to p(x, u, t): I \to R^l$, and $t \to q(u, t): I \to R^r$ are measurable.

A6. For each compact subset C of $G \times U_0$, there is a number $\zeta > 0$ such that, for all $(x, u) \in C$ and almost all $t \in I$,

(a) $|f(x, u, t)| + |f_x(x, u, t)| + |f_u(x, u, t)| \leqq \zeta$,

(b) $|p(x, u, t)| + |p_x(x, u, t)| + |p_u(x, u, t)| \leqq \zeta$,

(c) $|q(u, t)| + |q_u(u, t)| \leqq \zeta$,

where the vertical bars are used to denote Euclidean norms for vectors and matrix norms for matrices.

Assumptions A3–A6 ensure that, for any $u \in \mathscr{U}_0$, equation (2.2) has a unique local solution for any given initial value $x(t_1) \in G$.

Note that if A3 holds, and if, in addition, the functions $f, p, q, f_x, f_u, p_x, p_u$, and q_u are continuous in *all* of their arguments, then A4–A6 automatically hold.

For each positive integer j, let $\mathscr{C}^j$ (resp., L^j_∞, L^j_1) denote the linear vector space of all continuous (resp., essentially bounded, integrable) functions from I into R^j. We shall define the following norms on $\mathscr{C}^j, L^j_\infty, L^j_1$, respectively:

$$\|x\| = \max_{t \in I} |x(t)| \quad \text{for } x \in \mathscr{C}^j, \qquad \|y\| = \operatorname*{ess\,sup}_{t \in I} |y(t)| \quad \text{for } y \in L^j_\infty,$$

$$\|z\| = \int_{t_1}^{t_2} |z(t)|\, dt \quad \text{for } z \in L^j_1 .$$

With these norms, $\mathscr{C}^j$, L^j_∞, and L^j_1 become Banach spaces. The spaces $\mathscr{C}^1, L^1_\infty, L^1_1$, will simply be denoted by $\mathscr{C}, L_\infty, L_1$. If I is replaced by a Lebesgue measurable subset I' of I in the preceding definitions, then the corresponding spaces will be denoted by $\mathscr{C}^j(I'), L^j_\infty(I')$, etc.

Note that $\mathscr{G}$ is an open set in $\mathscr{C}^n$ and that $\mathscr{U}_0$ is an open set in L^m_∞. In the sequel, unless the contrary is specified, $\mathscr{G}$ and $\mathscr{U}_0$ are to be considered as sets in $\mathscr{C}^n$ and L^m_∞, respectively.

For ease of notation in the subsequent developments, we shall define the functions $P^i: \mathscr{G} \times \mathscr{U}_0 \to L_\infty(I_1)$ for $i = 1, \cdots, l$, $X_0: \mathscr{G} \to R$, $X_1: \mathscr{G} \to R^k$, $X_2: \mathscr{G} \to R^{k_1}$, $Q^i: \mathscr{U}_0 \to L_\infty$ for $i = 1, \cdots, r$, and $\tilde{X}^i: \mathscr{G} \to \mathscr{C}$ for $i = 1, \cdots, l_1$ as follows:

$$P^i(x(\cdot), u(\cdot)) \text{ is the function } t \to p^i(x(t), u(t), t): I_1 \to R, \tag{2.7}$$

$$X_0(x(\cdot)) = \chi_0(x(\tau_1), \cdots, x(\tau_\sigma)), \quad \text{where } \chi_0 = \chi^0, \tag{2.8}$$

$$X_1(x(\cdot)) = \chi_1(x(\tau_1), \cdots, x(\tau_\sigma)), \quad \text{where } \chi_1 = (\chi^1, \cdots, \chi^k), \tag{2.9}$$

$$X_2(x(\cdot)) = \chi_2(x(\tau_1), \cdots, x(\tau_\sigma)), \quad \text{where } \chi_2 = (\chi^{k+1}, \cdots, \chi^{k+k_1}), \tag{2.10}$$

$$Q^i(u(\cdot)) \quad \text{is the function} \quad t \to q^i(u(t), t) : I \to R, \tag{2.11}$$

$$\tilde{X}^i(x(\cdot)) \quad \text{is the function} \quad t \to \chi^i(x(t), t) : I \to R. \tag{2.12}$$

Also, let us denote $(P^1, \cdots, P^l)$ by P, $(Q^1, \cdots, Q^r)$ by Q, $(\tilde{X}^1, \cdots, \tilde{X}^{l_1})$ by $\tilde{X}$, and (X_0, X_1, X_2) by X.

Further, let us define the following sets (each of which is easily seen to be convex):

$$R_-^{k_1} = \{\xi = (\xi^1, \cdots, \xi^{k_1}) : \xi^i < 0 \text{ for each } i = 1, \cdots, k_1\}, \tag{2.13}$$

$$Y = \{y = (y^1, \cdots, y^{l_1}) : y \in \mathscr{C}^{l_1}, y^i(t) \leqq 0 \text{ for all } t \in I_i' \text{ and each } i = 1, \cdots, l_1\}, \tag{2.14}$$

$$Y_0 = \{y = (y^1, \cdots, y^{l_1}) : y \in \mathscr{C}^{l_1}, y^i(t) < 0 \text{ for all } t \in I_i' \text{ and each } i = 1, \cdots, l_1\}, \tag{2.15}$$

$$W = \{w = (w^1, \cdots, w^r) : w \in L_\infty^r, \operatorname*{ess\,sup}_{t \in I} w^i(t) \leqq 0 \text{ for each } i = 1, \cdots, r\}, \tag{2.16}$$

$$W_0 = \{w = (w^1, \cdots, w^r) : w \in L_\infty^r, \operatorname*{ess\,sup}_{t \in I} w^i(t) < 0 \text{ for each } i = 1, \cdots, r\}, \tag{2.17}$$

$$W_1 = \{w = (w^1, \cdots, w^r) : w \in L_\infty^r, \operatorname*{ess\,sup}_{t \in I_1} w^i(t) < 0 \text{ for each } i = 1, \cdots, r\}, \tag{2.18}$$

$$W_2 = \{w = (w^1, \cdots, w^r) : w \in L_\infty^r, \operatorname*{ess\,sup}_{t \in I_1} w^i(t) \leqq 0 \text{ for each } i = 1, \cdots, r\}. \tag{2.19}$$

Note that Y, W, and W_2 are all closed, and that $Y_0 = \text{int } Y$, $W_0 = \text{int } W$, and $W_1 = \text{int } W_2$, where int denotes interior. Further, $R_-^{k_1}$ is open and

$$\text{cl } R_-^{k_1} = \{\xi = (\xi^1, \cdots, \xi^{k_1}) : \xi^i \leqq 0 \text{ for each } i = 1, \cdots, k_1\}, \tag{2.20}$$

where cl denotes closure. Also, $\mathscr{U} = \{u : u \in \mathscr{U}_0, Q(u) \in W\}$.

Our basic problem may now be restated as follows: Find a pair $(x, u) \in \mathscr{G} \times \mathscr{U}_0$ such that (i) x is a solution of (2.2), (ii) $Q(u) \in W$, (iii) $X_1(x) = 0$, (iv) $X_2(x) \in \text{cl } R_-^{k_1}$, (v) $P(x, u) = 0$, and (vi) $\tilde{X}(x) \in Y$, and which, in so doing, achieves a minimum for $X_0(x)$.

It easily follows from Assumptions A3 and A4 that the map P is Fréchet differentiable at each $(x, u) \in \mathscr{G} \times \mathscr{U}_0$, with partial differentials at any $(x, u) \in \mathscr{G} \times \mathscr{U}_0$—which we shall denote by $D_iP(x, u; \cdot)$, $i = 1, 2$—given by the following formulas:

$$D_1 P(x, u; \delta x) \quad \text{is the function} \quad t \to p_x(x(t), u(t), t)\delta x(t) : I_1 \to R^l \quad \text{for all } \delta x \in \mathscr{C}^n, \tag{2.21}$$

$$D_2P(x, u; \delta u) \quad \text{is the function} \quad t \to p_u(x(t), u(t), t)\delta u(t) : I_1 \to R^l \quad \text{for all } \delta u \in L_\infty^m. \tag{2.22}$$

Further, A3 and A4 imply that Q is Fréchet differentiable at each $u \in \mathscr{U}_0$ with Fréchet differential (at any $u \in \mathscr{U}_0$) given by

$$DQ(u; \delta u) \quad \text{is the function} \quad t \to q_u(u(t), t)\delta u(t) : I \to R^r \quad \text{for all } \delta u \in L_\infty^m. \tag{2.23}$$

Finally, A1 and A2 imply that X_0, X_1, X_2, and $\tilde{X}$ are Fréchet differentiable at each $x \in \mathscr{G}$, with Fréchet differentials (at any $x \in \mathscr{G}$) given by

$$DX_j(x; \delta x) = \sum_{i=1}^{\sigma} \chi_{jx_i}(x(\tau_1), \cdots, x(\tau_\sigma))\delta x(\tau_i) \quad \text{for all } \delta x \in \mathscr{C}^n \text{ and } j = 0, 1, 2, \tag{2.24}$$

$$D\tilde{X}(x; \delta x) \quad \text{is the function} \quad t \to \tilde{\chi}_x(x(t), t)\delta x(t) : I \to R^{l_1} \quad \text{for all } \delta x \in \mathscr{C}^n. \tag{2.25}$$

3. Admissible control variations. Let $(\bar{x}, \bar{u})$ be a solution of our problem. Our aim is to find necessary conditions which $(\bar{x}, \bar{u})$ must satisfy.

Let us denote by $\bar{f}$ the function $t \to f(\bar{x}(t), \bar{u}(t), t) : I \to R^n$ (so that $\bar{f} \in L_\infty^n$), and similarly define the functions $\bar{p}, \bar{q}, \bar{f}_x, \bar{f}_u, \bar{p}_x, \bar{p}_u$, and $\bar{q}_u$. Further, we shall denote $D_iP(\bar{x},\bar{u}; \cdot)$, for $i = 1, 2$, by $D_i\bar{P}(\cdot)$;$DX_j(\bar{x}; \cdot)$, for $j = 0, 1, 2$, by $D\bar{X}_j(\cdot)$;$DX(\bar{x}; \cdot)$ by $D\bar{X}(\cdot)$; and $DQ(\bar{u}; \cdot)$ by $D\bar{Q}(\cdot)$. Also, let $\chi_{x_i}(\bar{x}(\tau_1), \cdots, \bar{x}(\tau_\sigma))$ be denoted by $\bar{\chi}_{x_i}$ for each $i = 1, \cdots, \sigma$.

Let $\tilde{\mathscr{U}}$ denote the set of all $u \in \mathscr{U}$ such that $P(\bar{x}, u) = 0$; i.e., $\tilde{\mathscr{U}}$ consists of all admissible controls which, together with the optimum phase trajectory $\bar{x}$, satisfy the problem constraint $P = 0$. Note that $\bar{u} \in \tilde{\mathscr{U}}$.

We shall largely confine our attention in what follows to controls $u \in \tilde{\mathscr{U}}$ which satisfy the following condition:

Condition C1. For almost all $t \in I_1$, the $l \times l$ matrix $p_u(\bar{x}(t), u(t), t) \cdot (p_u(\bar{x}(t), u(t), t))^T$—where T denotes transposition—is nonsingular, and the function $t \to [p_u(\bar{x}(t), u(t), t)(p_u(\bar{x}(t), u(t), t))^T]^{-1} : I_1 \to R^{l^2}$ is in $L_\infty^{l^2}(I_1)$.

In order to obtain our necessary conditions, we shall have to suppose that $\bar{u}$ satisfies Condition C1 (as well as Condition C2, which is described later in this section). As we shall see, the strength of these necessary conditions is directly related to the number of controls in $\tilde{\mathscr{U}}$ which satisfy C1 and C2.

The first part of C1 is equivalent to the assertion that the matrix $p_u(\bar{x}(t), u(t), t)$ has, for almost all $t \in I_1$, maximum rank, i.e., rank l. Consequently, C1 cannot hold for any $u \in \mathscr{U}$, if, for example, some component of the function p is independent of u. Thus, it appears that if we require $\bar{u}$, as well as a reasonable number of other controls in $\tilde{\mathscr{U}}$, to satisfy C1, then we shall have to exclude from consideration many interesting types of equality constraints. However, it turns out that, by a suitable reformulation, we can often transform our problem from one in which C1 is violated to one in which C1 holds. For example, suppose that I_1 is a subinterval of I and that p is independent of u, so that we can write (2.5) in the form

$p(x(t), t) = 0$. Then, differentiating this last equation with respect to t and using (2.2), we see that it is equivalent to the equations

$$p_1(x(t), u(t), t) = 0 \quad \text{for almost all } t \in I_1, \tag{3.1}$$

$$p(x(\tilde{t}), \tilde{t}) = 0, \tag{3.2}$$

where p_1 is the function defined by

$$p_1(x, u, t) = p_x(x, t)f(x, u, t) + p_t(x, t), \qquad x \in G, \quad u \in U_0, \quad t \in I,$$

and $\tilde{t}$ is an arbitrary point of I_1. Thus, we may replace constraints (2.5) by the equivalent pair of constraints (3.1) and (3.2). Evidently, (3.1) is of the same form as (2.5), and (3.2) may be adjoined to (2.3). Further, if p is twice continuously differentiable, and A1–A6 hold, then these same assumptions hold with p replaced by p_1. Hopefully, when p is replaced by p_1 in C1, then this condition will be satisfied by a broad class of functions in $\mathscr{U}$, including $\bar{u}$. If p_1 also turns out to be independent of u, then the just indicated procedure may be repeated.

If only some components of p are independent of u, then the procedure which we have described must be applied to only those components.

Let us return to C1. If $u \in \tilde{\mathscr{U}}$ satisfies C1, then we may consider (for almost every $t \in I_1$) the $m \times l$ matrix

$$(p_u(\bar{x}(t), u(t), t))^T[p_u(\bar{x}(t), u(t), t)(p_u(\bar{x}(t), u(t), t))^T]^{-1}. \tag{3.3}$$

This matrix is commonly referred to as the *pseudo-inverse* of $p_u(\bar{x}(t), u(t), t)$ because, if the matrix (3.3) is premultiplied by $p_u(\bar{x}(t), u(t), t)$, the result is the identity matrix.

For each $u \in \tilde{\mathscr{U}}$ satisfying C1, let us define the continuous linear operator $\widetilde{D_2P}(\bar{x}, u; \cdot): L^l_\infty(I_1) \to L^m_\infty$ as follows: For any $z \in L^l_\infty(I_1)$, let $\widetilde{D_2P}(\bar{x}, u; z)$ denote the function δu in L^m_∞ defined by

$$\delta u(t) = \begin{cases} 0 & \text{for } t \in I \setminus I_1, \\ (p_u(\bar{x}(t), u(t), t))^T[p_u(\bar{x}(t), u(t), t)(p_u(\bar{x}(t), u(t), t))^T]^{-1}z(t) & \text{for } t \in I_1. \end{cases}$$

Clearly, $\widetilde{D_2P}(\bar{x}, u)$ behaves like a pseudo-inverse of $D_2P(\bar{x}, u)$ in the sense that (see (2.22)) $D_2P(\bar{x}, u) \circ \widetilde{D_2P}(\bar{x}, u)$ is the identity operator on $L^l_\infty(I_1)$.

For the special case where $u = \bar{u}$, we shall refer to the function (3.3) as $\tilde{p}(t)$; i.e.,

$$\tilde{p}(t) = (\bar{p}_u(t))^T[\bar{p}_u(t)(\bar{p}_u(t))^T]^{-1} \quad \text{for } t \in I_1. \tag{3.4}$$

For convenience, let us extend the function $\tilde{p}$ to I by setting

$$\tilde{p}(t) = 0 \quad \text{for } t \in I \setminus I_1. \tag{3.5}$$

Note that

$$\bar{p}_u(t)\tilde{p}(t) = \text{the } l \times l \text{ identity matrix for almost all } t \in I_1, \tag{3.6}$$

and that

$$\widetilde{D_2P}(\bar{x}, \bar{u}; z)(t) = \tilde{p}(t)z(t) \quad \text{for all } t \in I \text{ and } z \in L^l_\infty(I_1).$$

Not only shall we confine ourselves to admissible controls u such that $P(\bar{x}, u) = 0$ and which satisfy C1, but we shall impose yet another requirement on the controls that we shall consider. To describe this requirement, we must introduce some new sets. For each $u \in \mathscr{U}$, let us define the set $\tilde{\mathscr{Q}}(u)$ as follows:

$$\tilde{\mathscr{Q}}(u) = \{\delta u : \delta u \in L_\infty^m, [Q(u) + DQ(u; \delta u)] \in W_1\}. \tag{3.7}$$

It is easily seen that, for each $u \in \mathscr{U}$, $\tilde{\mathscr{Q}}(u)$ is an open convex set in L_∞^m and $\alpha\tilde{\mathscr{Q}}(u) \subset \tilde{\mathscr{Q}}(u)$ for all α, $0 < \alpha \leqq 1$. Let $\mathscr{Q}(u)$ denote the cone spanned by $\tilde{\mathscr{Q}}(u)$, i.e.,

$$\mathscr{Q}(u) = \{\alpha\delta u : \alpha > 0, \delta u \in \tilde{\mathscr{Q}}(u)\}. \tag{3.8}$$

Since $\tilde{\mathscr{Q}}(u)$ is open and convex, $\mathscr{Q}(u)$ is an open convex cone, which implies (as is easily seen) that

$$\mathscr{Q}(u) + \operatorname{cl} \mathscr{Q}(u) = \mathscr{Q}(u). \tag{3.9}$$

Finally, let

$$\Delta_0 = \{\delta u : \delta u \in L_\infty^m, \delta u(t) = 0 \text{ for almost all } t \in I \diagdown I_1\} \tag{3.10}$$

and

$$\mathscr{Q}_0(u) = \mathscr{Q}(u) \cap \Delta_0.$$

It is easily verified that $\mathscr{Q}_0(u)$ and Δ_0 are also convex cones.

We now state our second requirement on the admissible controls u which we shall consider.

Condition C2. $\{D_2P(\bar{x}, u; \delta u) : \delta u \in \mathscr{Q}_0(u)\} = L_\infty^l(I_1)$.

We shall denote by $\hat{\mathscr{U}}$ the set of all $u \in \mathscr{U}$ which satisfy Conditions C1 and C2, and shall confine our attention to controls in $\hat{\mathscr{U}}$. Indeed, when constructing "strong variations" of $\bar{u}$ in the derivation of our necessary conditions, we shall only allow variations to functions in $\hat{\mathscr{U}}$, so that the maximum principle—which will be the form that our necessary conditions will take— which we shall obtain will be valid only for such u. (We shall also allow "weak variations" of $\bar{u}$ from $\mathscr{Q}(\bar{u})$.)

Remark 3.1. Note that if u_1 and u_2 belong to $\hat{\mathscr{U}}$ and if $u_3 \in \mathscr{U}$ is such that, for some measurable subset I' of I, $u_3(t) = u_1(t)$ for all $t \in I'$ and $u_3(t) = u_2(t)$ for all $t \in I \diagdown I'$, then also $u_3 \in \hat{\mathscr{U}}$.

We shall suppose that $\bar{u}$ satisfies C2 *as well as* C1, so that $\bar{u} \in \hat{\mathscr{U}}$.

Conditions C1 and C2 should be interpreted as regularity conditions or as compatibility conditions—to "first order" in u— of the constraints $p = 0$ and $q^j \leqq 0$ for $j = 1, \cdots, r$. They replace conditions found, for example, in [6] and [7] on the linear independence of the vectors $p_u^j(\bar{x}(t), u(t), t)$ for $j = 1, \cdots, l$, and $q_u^j(u(t), t)$ for those $j = 1, \cdots, r$ for which $q^j(u(t), t) = 0$.

We close this section with two lemmas regarding the sets $\mathscr{Q}(u)$ which we shall need in the sequel.

LEMMA 3.1. *For every $\hat{u} \in \hat{\mathscr{U}}$ and any $\delta\hat{u} \in \mathscr{Q}(\hat{u})$, there is an $\hat{\varepsilon} > 0$ such that $(\hat{u} + \varepsilon\delta u) \in \mathscr{U}$ whenever $0 \leqq \varepsilon < \hat{\varepsilon}$, $\delta u \in \Delta_0$, and $\|\delta u - \delta\hat{u}\| < \hat{\varepsilon}$.*

Proof. If $\hat{u} \in \hat{\mathscr{U}}$ and $\delta\hat{u} \in \mathscr{Q}(\hat{u})$, then $\delta\hat{u} = \tilde{\alpha}\delta\tilde{u}$ for some $\tilde{\alpha} > 0$ and some $\delta\tilde{u} \in \tilde{\mathscr{Q}}(\hat{u})$. Since $\mathscr{U}_0$ is open, there is an $\varepsilon_0 > 0$ such that $(\hat{u} + \varepsilon\delta u) \in \mathscr{U}_0$ for all $\varepsilon, 0 < \varepsilon < \varepsilon_0$, and all $\delta u \in L_\infty^m$ such that $\|\delta u - \delta\hat{u}\| < \varepsilon_0$. By definition of $\tilde{\mathscr{Q}}(\hat{u})$ (see (3.7) and (2.18)), there is an $\eta > 0$ such that

$$\operatorname*{ess\,sup}_{t \in I_1} [Q^j(\hat{u})(t) + DQ^j(\hat{u}; \delta\tilde{u})(t)] < -\eta \quad \text{for } j = 1, \cdots, r. \tag{3.11}$$

Because $DQ(\hat{u};\cdot)$ is the Fréchet differential of Q at $\hat{u}$, it easily follows that, for every $\delta u \in L_\infty^m$ and $j = 1, \cdots, r$,

$$\frac{Q^j(\hat{u} + \varepsilon\delta u) - Q^j(\hat{u})}{\varepsilon} \xrightarrow[\varepsilon\to 0]{} DQ^j(\hat{u};\delta u),$$

with the convergence uniform with respect to δu in any bounded subset of L_∞^m. From this, we can easily deduce that there is an $\varepsilon_1 \in (0, 1)$ such that $\varepsilon_1 < \varepsilon_0$ and

$$\left\| \frac{Q^j(\hat{u} + \varepsilon\delta u) - Q^j(\hat{u})}{\varepsilon} - DQ^j(\hat{u};\delta\tilde{u}) \right\| < \frac{\eta}{2} \quad \text{for } j = 1, \cdots, r \tag{3.12}$$

whenever $0 < \varepsilon < \varepsilon_1$ and $\|\delta u - \delta\tilde{u}\| < \varepsilon_1$.

If we write

$$Q^j(\hat{u} + \varepsilon\delta u) = \varepsilon\left[\frac{Q^j(\hat{u} + \varepsilon\delta u) - Q^j(\hat{u})}{\varepsilon} - DQ^j(\hat{u};\delta\tilde{u})\right]$$
$$+ \varepsilon[Q^j(\hat{u}) + DQ^j(\hat{u};\delta\tilde{u})] + (1 - \varepsilon)Q^j(\hat{u}),$$

and take into account (3.11), (3.12), and the fact that $\hat{u} \in \mathscr{U}$ (so that $Q^j(\hat{u})(t) \leqq 0$ for almost all $t \in I$ and each j), we quickly conclude that, if $0 < \varepsilon < \varepsilon_1$, $\delta u \in \Delta_0$, and $\|\delta u - \delta\tilde{u}\| < \varepsilon_1$, then $(\hat{u} + \varepsilon\delta u) \in \mathscr{U}_0$ and

$$Q^j(\hat{u} + \varepsilon\delta u)(t) \leqq 0 \quad \text{for almost all } t \in I_1 \text{ and each } j,$$

$$Q^j(\hat{u} + \varepsilon\delta u)(t) = Q^j(\hat{u})(t) \leqq 0 \quad \text{for almost all } t \in I \diagdown I_1 \text{ and each } j,$$

i.e., $(\hat{u} + \varepsilon\delta u) \in \mathscr{U}$ for all ε and δu as just indicated. Setting $\hat{\varepsilon} = \min\{\varepsilon_1\tilde{\alpha}, \varepsilon_1/\tilde{\alpha}\}$, we arrive at our desired conclusion.

Using a standard compactness argument, we can strengthen Lemma 3.1 as follows.

LEMMA 3.2. *For any $\hat{u} \in \hat{\mathscr{U}}$ and any compact subset $\mathscr{Q}_c$ of $\mathscr{Q}(\hat{u})$, there is an $\varepsilon_0 > 0$ such that $(\hat{u} + \varepsilon\delta u) \in \mathscr{U}$ whenever $0 \leqq \varepsilon < \varepsilon_0$, $\delta u \in \Delta_0$, and $\|\delta u - \delta\hat{u}\| < \varepsilon_0$ for some $\delta\hat{u} \in \mathscr{Q}_c$.*

4. Linearized equations. For each finite subset $H = \{(\beta^1, u_1), \cdots, (\beta^s, u_s)\}$ of $\bar{R}_+ \times \hat{\mathscr{U}}$—where $\bar{R}_+$ denotes the set of nonnegative numbers—let us define the function $\Delta f_H \in L_\infty^n$ as follows:

$$\Delta f_H(t) = \sum_{j=1}^{s} \beta^j[f(\bar{x}(t), u_j(t), t) - f(\bar{x}(t), \bar{u}(t), t)], \qquad t \in I.$$

The set of all such functions Δf_H, as H ranges over all finite subsets of $\bar{R}_+ \times \hat{\mathscr{U}}$, will be denoted by M. Evidently, M is a convex cone in L_∞^n, and $0 \in M$.

We shall consider linear inhomogeneous differential equations of the form

$$\delta\dot{x}(t) = \bar{f}_x(t)\delta x(t) + \bar{f}_u(t)\delta u(t) + \Delta f(t), \qquad t \in I, \tag{4.1}$$

where $\delta u \in L_\infty^m$ and $\Delta f \in M$. Equation (4.1) may be viewed as a "linearization" of the basic equation (2.2) about $(\bar{x}, \bar{u})$. Let us denote by Z the set of all pairs $(\delta x, \delta u) \in \mathscr{C}^n \times L_\infty^m$ such that δx is absolutely continuous and δx and δu satisfy equation (4.1) for some $\Delta f \in M$. Since M is a convex cone, it follows at once that Z also is.

For each $z \in L^l_\infty(I_1)$, let $\delta x_z \in \mathscr{C}^n$ denote the function which satisfies the linear inhomogeneous ordinary differential equation

$$(4.2) \qquad \delta\dot{x}_z(t) = \bar{f}_x(t)\delta x_z(t) + \bar{f}_u(t)\tilde{p}(t)[z(t) - \bar{p}_x(t)\delta x_z(t)] \quad \text{a.e. on } I$$

with initial value

$$(4.3) \qquad \delta x_z(t_1) = 0.$$

(In (4.2), $z(t)$ is undefined for $t \in I \diagdown I_1$, but this does not matter because $\tilde{p}(t) = 0$ for $t \in I \diagdown I_1$, by definition—see (3.5).) Equation (4.2) may be looked upon as a differential equation with a "feedback law" expressed by the term in brackets. Indeed, if we denote this term by δu_z, i.e., $\delta u_z \in L^m_\infty$ is defined by

$$(4.4) \qquad \delta u_z(t) = \tilde{p}(t)[z(t) - \bar{p}_x(t)\delta x_z(t)], \qquad t \in I,$$

then we have (see (3.6)) that

$$(4.5) \qquad \bar{p}_x(t)\delta x_z(t) + \bar{p}_u(t)\delta u_z(t) = z(t) \quad \text{for almost all } t \in I_1,$$

or, equivalently, that

$$(4.6) \qquad D_1\bar{P}(\delta x_z) + D_2\bar{P}(\delta u_z) = z \quad \text{for all } z \in L^l_\infty(I_1).$$

In fact, the feedback law was specifically designed in order that (4.5) (or (4.6)) hold.

Note that δx_z satisfies the linear inhomogeneous differential equation

$$(4.7) \qquad \delta\dot{x}_z(t) = \bar{f}_x(t)\delta x_z(t) + \bar{f}_u(t)\delta u_z(t), \quad t \in I; \qquad \delta x_z(t_1) = 0,$$

where δu_z is defined by (4.2)–(4.4), so that $(\delta x_z, \delta u_z) \in Z$. Also note that (see (4.4), (3.5), and (3.10)),

$$(4.8) \qquad \delta u_z \in \Delta_0 \quad \text{for all } z \in L^l_\infty(I_1).$$

LEMMA 4.1. *The map $z \to (\delta x_z, \delta u_z): L^l_\infty(I_1) \to Z$ is linear and continuous.*

Proof. Since δx_z is defined by the linear inhomogeneous differential equation (4.2) with zero initial value (see (4.3)), the map $z \to \delta x_z : L^l_\infty(I_1) \to \mathscr{C}^n$ is evidently linear. Using the variations of parameters formula for the solution of (4.2), (4.3), we at once see that this map is also continuous. Because the map $(z, \delta x_z) \to \delta u_z : L^l_\infty(I_1) \times \mathscr{C}^n \to L^m_\infty$ defined by (4.4) is evidently linear and continuous, our lemma follows at once.

COROLLARY 4.1. *If z_0 denotes the function in $L^l_\infty(I_1)$ which vanishes identically, then $\delta x_{z_0} = 0$ and $\delta u_{z_0} = 0$.*

We observe that equations (4.2) and (4.3) have a solution not only for every $z \in L^l_\infty(I_1)$, but also for every $z \in L^l_1(I_1)$, and the resulting function δu_z given by (4.4) is then in L^m_1. Further, the map $z \to (\delta x_z, \delta u_z): L^l_1(I_1) \to \mathscr{C}^n \times L^m_1$ is continuous.

5. The basic theorem. We can now state our main theorem, on the basis of which we shall obtain our necessary conditions.

Let

$$Z_1 = \{(\delta x, \delta u) : \delta x \in \mathscr{C}^n, \delta u \in \mathscr{Q}_0(\bar{u}), X_2(\bar{x}) + D\bar{X}_2(\delta x) \in R^{k_1}_{-}, \tilde{X}(\bar{x}) + D\tilde{X}(\bar{x}; \delta x) \in Y_0, D\bar{X}_0(\delta x) < 0\}, \tag{5.1}$$

$$Z_2 = \{(\delta x, \delta u) : \delta x \in \mathscr{C}^n, \delta u \in L^m_\infty, D_1\bar{P}(\delta x) + D_2\bar{P}(\delta u) = 0\}, \tag{5.2}$$

$$Z_3 = Z_1 \cap Z_2 \cap Z. \tag{5.3}$$

Roughly speaking, Z_3 consists of those perturbations of $(\bar{x}, \bar{u})$ which, to first order, satisfy all of the problem constraints except (2.3), and also yield a lower value for the cost functional X_0. Note that Z_i is convex for $i = 1, 2, 3$ because $\mathscr{Q}_0(\bar{u})$, $R^{k_1}_{-}$, Y_0, and Z are.

Our main theorem is the following.

THEOREM 5.1. *The origin is not an interior point* (*in* R^k) *of the set*

$$K = \{D\bar{X}_1(\delta x) : (\delta x, \delta u) \in Z_3 \text{ for some } \delta u \in L^m_\infty\}. \tag{5.4}$$

In § 11, we shall employ Theorem 5.1 to appeal to a series of separation theorems on the basis of which we shall obtain our necessary conditions. Note that K is convex because Z_3 is.

Theorem 5.1 essentially asserts that, to "first order," there is no perturbation of $(\bar{x}, \bar{u})$ which satisfies all of the problem constraints while at the same time yielding a lower value for X_0.

We shall argue by contradiction. Indeed, we shall prove that if $0 \in \operatorname{int} K$, i.e., if to "first order" one can improve X_0 while satisfying the problem constraints, then one can improve X_0, even including higher order terms, while satisfying the problem constraints (again including "higher order" terms), which violates that $(\bar{x}, \bar{u})$ is a solution of our problem.

6. An auxiliary lemma.

LEMMA 6.1. *If* $0 \in \operatorname{int} K$, *then there are a simplex* $S \subset R^k$ *with vertices* $v_1, \cdots, v_{k+1}$ *and with* $0 \in \operatorname{int} S$, *functions* $\Delta_1 f, \cdots, \Delta_{k+1} f$ *in* M, *elements* $(\delta x_i, \delta u_i)$, *for* $i = 1, \cdots, k+1$, *in* Z *with* $\delta u_i \in \mathscr{Q}_0(\bar{u})$, *and a number* $\varepsilon_0 > 0$ *with the following property*:

For each $\gamma = (\gamma^1, \cdots, \gamma^{k+1}) \in S^k = \{\gamma : \gamma \in R^{k+1}, \gamma^i \geqq 0 \text{ for each } i, \sum_{i=1}^{k+1} \gamma^i = 1\}$ *and each* $z \in L^l_\infty(I_1)$ *satisfying* $\|z\| < \varepsilon_0$, *there are pairs* $(\delta x_{z,\gamma}, \delta u_{z,\gamma}) \in Z$ *such that* (*for all such* z *and* γ)

$$\delta \dot{x}_{z,\gamma} = \bar{f}_x \delta x_{z,\gamma} + \bar{f}_u \delta u_{z,\gamma} + \sum_{i=1}^{k+1} \gamma^i \Delta_i f, \quad \text{a.e. on } I, \tag{6.1}$$

$$\delta x_{z,\gamma}(t_1) = \sum_{i=1}^{k+1} \gamma^i \delta x_i(t_1), \tag{6.2}$$

$$D\bar{X}_1(\delta x_{z_0,\gamma}) = \sum_{i=1}^{k+1} \gamma^i v_i, \quad \text{where } z_0 \text{ is the origin in } L^l_\infty(I_1), \tag{6.3}$$

$$D_1\bar{P}(\delta x_{z,\gamma}) + D_2\bar{P}(\delta u_{z,\gamma}) = z, \tag{6.4}$$

$$X_2(\bar{x}) + D\bar{X}_2(\delta x_{z_0,\gamma}) \in R^{k_1}_{-} \quad \text{and} \quad \tilde{X}(\bar{x}) + D\tilde{X}(\bar{x}; \delta x_{z_0,\gamma}) \in Y_0, \tag{6.5}$$

$$D\bar{X}_0(\delta x_{z_0,\gamma}) < 0, \tag{6.6}$$

$$(6.7)\qquad \delta x_{z,\gamma} = \delta x_z + \sum_{i=1}^{k+1} \gamma^i \delta x_i,$$

$$(6.8)\qquad \delta u_{z,\gamma} = \delta u_z + \sum_{i=1}^{k+1} \gamma^i \delta u_i,$$

$$(6.9)\qquad (\bar{u} + \varepsilon\delta u_{z,\gamma}) \in \mathscr{U} \quad \textit{whenever } 0 \leqq \varepsilon < \varepsilon_0,$$

$$(6.10)\qquad \delta u_{z,\gamma} \in \Delta_0.$$

Proof. If $0 \in \operatorname{int} K$, then there is a simplex $S \subset K$ such that $0 \in \operatorname{int} S$. Let the vertices of S be $v_1, \cdots, v_{k+1}$. Since each $v_i \in K$ (see (5.1)–(5.4)), there are functions $\Delta_i f \in M$ and $(\delta x_i, \delta u_i) \in Z$ with $\delta u_i \in \mathscr{Q}_0(\bar{u})$, for $i = 1, \cdots, k+1$, such that, for each i,

$$D\bar{X}_1(\delta x_i) = v_i,$$

$$\delta \dot{x}_i = \bar{f}_x \delta x_i + \bar{f}_u \delta u_i + \Delta_i f \text{ a.e. on } I,$$

$$X_2(\bar{x}) + D\bar{X}_2(\delta x_i) \in R^{k_1}_{-}, \qquad \tilde{X}(\bar{x}) + D\tilde{X}(\bar{x}; \delta x_i) \in Y_0, \qquad D\bar{X}_0(\delta x_i) < 0,$$

$$D_1\bar{P}(\delta x_i) + D_2\bar{P}(\delta u_i) = 0.$$

For each $z \in L^l_\infty(I_1)$ and $\gamma \in S^k$, let $\delta x_{z,\gamma}$ and $\delta u_{z,\gamma}$ be defined by (6.7) and (6.8). Taking into account the properties of δx_z and δu_z discussed in § 4 (in particular, see Corollary 4.1 and relations (4.3) and (4.6)–(4.8)), we easily deduce that (6.1)–(6.6) and (6.10) hold. (Recall that Y_0 and $R^{k_1}_{-}$ are convex.) Thus, it only remains to show that, for a suitable $\varepsilon_0 > 0$, (6.9) holds for all $\gamma \in S^k$, so long as $\|z\| < \varepsilon_0$. But, since $\{\sum_{i=1}^{k+1} \gamma^i \delta u_i : \gamma = (\gamma^1, \cdots, \gamma^{k+1}) \in S^k\}$ is a compact subset of $\mathscr{Q}(\bar{u})$, the existence of such an ε_0 follows directly from (6.8) and Lemmas 3.2 and 4.1.

Lemma 6.1 essentially asserts that if $0 \in \operatorname{int} K$, then there is a simplex $S \subset R^k$, having 0 in its interior, with the following property: For any $\xi \in S$ (with barycentric coordinates $\gamma^1, \cdots, \gamma^{k+1}$), there is a "perturbation" $(\delta x, \delta u)$ of $(\bar{x}, \bar{u})$ (δx should be chosen as $\delta x_{z_0,\gamma}$ and δu as $\delta u_{z_0,\gamma}$, where $\gamma = (\gamma^1, \cdots, \gamma^{k+1})$) such that, to "first order," at the perturbed x and u, (i) the basic differential equation is satisfied, (ii) X_1 takes on the value ξ, (iii) the problem inequality constraints are "strictly" satisfied, (iv) the equality constraint $P = 0$ is satisfied, (v) the perturbed control is admissible, and (vi) X_0 takes on a lower value than it does at $\bar{x}$. Further, there is a neighborhood of 0 in $L^l_\infty(I_1)$ such that, for any z in this neighborhood, we can make an additional small perturbation $(\delta x_z, \delta u_z)$ of $(\bar{x}, \bar{u})$ which, "to first order," will achieve $P = z$ while only slightly perturbing the values of X_0, X_1, X_2, and $\tilde{X}$, will still satisfy the basic differential equation, and will result in an admissible control.

In what follows, we shall show that, for each $\xi \in \varepsilon S$ (where ε is some sufficiently small positive number), $(\bar{x}, \bar{u})$ can be perturbed in such a way that (i) the perturbed u is admissible and, together with the perturbed x, satisfies the basic differential equation (2.2) as well as the equality constraint $P = 0$, (ii) the perturbed x satisfies the problem inequality constraints and assigns to X_0 a value less than $X_0(\bar{x})$, and (iii) the perturbed x assigns to X_1 (with a "small" error) the value ξ. All these assertions are true exactly, not just to first order.

Armed with the preceding intermediate result, we shall, in § 10, use the Brower fixed-point theorem to show that we can find a perturbation of $(\bar{x}, \bar{u})$ which has properties (i) and (ii) described in the preceding paragraph, and, in addition, assigns 0 to X_1. But then this perturbation satisfies all of the problem constraints while giving a lower value to X_0 than $\bar{x}$ does, which contradicts the fact that $(\bar{x}, \bar{u})$ is a solution of our problem. As a result, we shall be able to conclude that $0 \notin \operatorname{int} K$.

The next two sections are devoted to the construction of some suitable "perturbations" of our optimal control $\bar{u}$.

7. Chattering controls. Continuing with our contradiction argument, let $\Delta_1 f, \cdots, \Delta_{k+1} f$ be the functions in M whose existence is asserted in Lemma 6.1. By definition of M, this means that there are functions $u_1, \cdots, u_s \in \hat{\mathscr{U}}$ and nonnegative numbers β_i^j $(i = 1, \cdots, k + 1; j = 1, \cdots, s)$ such that

$$\Delta_i f(t) = \sum_{j=1}^{s} \beta_i^j [f(\bar{x}(t), u_j(t), t) - f(\bar{x}(t), \bar{u}(t), t)], \qquad t \in I, \quad i = 1, \cdots, k + 1.$$

Thus, equation (6.1) for $\delta x_{z,\gamma}$ can be rewritten as follows:

$$\begin{aligned}\delta \dot{x}_{z,\gamma}(t) &= \bar{f}_x(t)\delta x_{z,\gamma}(t) + \bar{f}_u(t)\delta u_{z,\gamma}(t)\\ &\quad + \sum_{j=1}^{s} \beta^j(\gamma)[f(\bar{x}(t), u_j(t), t) - f(\bar{x}(t), \bar{u}(t), t)],\end{aligned}$$

where, for each $\gamma = (\gamma^1, \cdots, \gamma^{k+1}) \in S$,

$$\beta^j(\gamma) = \sum_{i=1}^{k+1} \gamma^i \beta_i^j, \qquad j = 1, \cdots, s,$$

so that

$$0 \leqq \beta^j(\gamma) \leqq \sum_{i=1}^{k+1} \beta_i^j \quad \text{for all } \gamma \in S^k \text{ and } j = 1, \cdots, s.$$

Now, for any $\varepsilon \geqq 0$ and $\gamma \in S^k$, consider the differential equation

$$\begin{aligned}\dot{x}(t) &= f(x(t), \bar{u}(t), t)\\ &\quad + \varepsilon \sum_{j=1}^{s} \beta^j(\gamma)[f(x(t), u_j(t), t) - f(x(t), \bar{u}(t), t)] \quad \text{a.e. on } I\end{aligned} \tag{7.1}$$

with initial condition

$$x(t_1) = \bar{x}(t_1) + \varepsilon \sum_{i=1}^{k+1} \gamma^i \delta x_i(t_1), \tag{7.2}$$

where the δx_i are as indicated in Lemma 6.1. For $\varepsilon = 0$, $\bar{x}$ is the solution of (7.1) and (7.2). It follows directly from conventional theorems on the differentiability of solutions of ordinary differential equations with respect to parameters that there is an $\hat{\varepsilon} > 0$ such that (7.1) and (7.2) have a unique solution defined on the entire interval I for all $\varepsilon \in [0, \hat{\varepsilon})$ and $\gamma \in S^k$, and that, if we denote the corresponding solutions by $\tilde{x}_{\gamma,\varepsilon}$, then

$$\left\| \frac{\tilde{x}_{\gamma,\varepsilon} - \bar{x}}{\varepsilon} - \delta x_\gamma \right\| \xrightarrow[\varepsilon \to 0^+]{} 0 \quad \text{uniformly with respect to } \gamma \in S^k, \tag{7.3}$$

where δx_γ, for each $\gamma \in S^k$, is the solution of the linear inhomogeneous differential equation

$$\begin{aligned} \delta \dot{x}_\gamma(t) &= \bar{f}_x(t)\delta x_\gamma(t) + \sum_{i=1}^{k+1} \gamma^i \Delta_i f(t) \quad \text{a.e. on } I, \\ \delta x_\gamma(t_1) &= \sum_{i=1}^{k+1} \gamma^i \delta x_i(t_1). \end{aligned} \tag{7.4}$$

Now, in general, $\tilde{x}_{\gamma,\varepsilon}$ will not be an admissible trajectory because there is no reason to expect the existence of a function $u \in \mathscr{U}$ such that

$$\begin{aligned} f(x, \bar{u}(t), t) + \varepsilon \sum_{j=1}^{s} \beta^j(\gamma)[f(x, u_j(t), t) \\ - f(x, \bar{u}(t), t)] = f(x, u(t), t) \quad \text{for all } x \in G \text{ and } t \in I. \end{aligned}$$

However, $\tilde{x}_{\gamma,\varepsilon}$ may be approximated by an admissible trajectory in the following way. (The idea of this approximation is due to Gamkrelidze in [2] who coined the term quasiconvexity in connection therewith.)

Without loss of generality, we shall suppose that $\hat{\varepsilon}$ is sufficiently small that (i) $\|\tilde{x}_{\gamma,\varepsilon} - \bar{x}\| < \rho_0$ for all $\gamma \in S^k$ and $\varepsilon \in [0, \hat{\varepsilon})$, where $\rho_0 > 0$ is such that $x \in G$ whenever $x \in R^n$ and $|x - \bar{x}(t)| \leqq 2\rho_0$ for some $t \in I$, and (ii) $\hat{\varepsilon}\beta^j(\gamma) \leqq 1/s$ for all j and $\gamma \in S^k$. For each $\varepsilon \in [0, \hat{\varepsilon})$, we perform the following construction. We first partition (in a manner to be specified later) I into a finite number of subintervals $J_1^\varepsilon, \cdots, J_{\nu(\varepsilon)}^\varepsilon$. For each $\gamma \in S^k$, we then partition each J_i^ε into $s + 1$ subintervals $J_{i,0}^{\varepsilon,\gamma}, J_{i,1}^{\varepsilon,\gamma}, \cdots, J_{i,s}^{\varepsilon,\gamma}$—these subintervals are indexed in such a way that $J_{i,j}^{\varepsilon,\gamma}$ is immediately to the left of $J_{i,j+1}^{\varepsilon,\gamma}$ for all i, j—such that, for each i,

$$|J_{i,0}^{\varepsilon,\gamma}| = \left(1 - \varepsilon \sum_{j=1}^{s} \beta^j(\gamma)\right)|J_i^\varepsilon| \quad \text{and} \quad |J_{i,j}^{\varepsilon,\gamma}| = \varepsilon\beta^j(\gamma)|J_i^\varepsilon|$$

$$\text{for each } j = 1, \cdots, s,$$

where vertical bars here denote lengths of intervals. Now set

$$I_{j,\varepsilon,\gamma} = \bigcup_{i=1}^{\nu(\varepsilon)} J_{i,j}^{\varepsilon,\gamma} \quad \text{for every } j = 0, \cdots, s, \quad \varepsilon \in [0, \hat{\varepsilon}), \quad \text{and } \gamma \in S^k, \tag{7.5}$$

and define the function $u_{\gamma,\varepsilon} \in \mathscr{U}$ as follows:

$$u_{\gamma,\varepsilon}(t) = \sum_{j=0}^{s} e_{j,\varepsilon,\gamma}(t)u_j(t), \quad \text{where } u_0 = \bar{u}, \tag{7.6}$$

and where, for each j, ε, γ, $e_{j,\varepsilon,\gamma}$ is the characteristic function of $I_{j,\varepsilon,\gamma}$, i.e.,

$$e_{j,\varepsilon,\gamma}(t) = \begin{cases} 0 & \text{for } t \in I \setminus I_{j,\varepsilon,\gamma}, \\ 1 & \text{for } t \in I_{j,\varepsilon,\gamma}. \end{cases} \tag{7.7}$$

Since $u_{\gamma,\varepsilon}(t) = u_j(t)$ for all $t \in J_{i,j}^{\varepsilon,\gamma}$ for each i, $u_{\gamma,\varepsilon}$ is often referred to as a "chattering combination" of $\bar{u}, u_1, \cdots, u_s$, or as a "chattering" control. Note that $u_{\gamma,0} \equiv \bar{u}$ for all $\gamma \in S^k$. Also, by Remark 3.1, $u_{\gamma,\varepsilon} \in \hat{\mathscr{U}} \subset \tilde{\mathscr{U}} \subset \mathscr{U}$ for all γ, ε, so that (by definition of $\tilde{\mathscr{U}}$)

$$P(\bar{x}, u_{\gamma,\varepsilon}) = 0 \quad \text{for all } \gamma, \varepsilon. \tag{7.8}$$

We show in the Appendix that if, for each ε, the partition of I into subintervals is sufficiently fine, i.e., if $\max_i |J_i^\varepsilon|$ is sufficiently small, then, for each γ, the equation

$$\dot{x}(t) = f(x(t), u_{\gamma,\varepsilon}(t), t) \tag{7.9}$$

with initial condition $x(t_1) = \tilde{x}_{\gamma,\varepsilon}(t_1)$ has a solution on the entire interval I such that, if we denote this solution by $x_{\gamma,\varepsilon}$, then

$$\|x_{\gamma,\varepsilon} - \tilde{x}_{\gamma,\varepsilon}\| < \varepsilon^2 \quad \text{for all } \gamma \in S^k. \tag{7.10}$$

We shall suppose that such a partition of I has been made for each $\varepsilon \in [0, \hat{\varepsilon})$. By definition,

$$\begin{aligned} \dot{x}_{\gamma,\varepsilon}(t) &= f(x_{\gamma,\varepsilon}(t), u_{\gamma,\varepsilon}(t), t) \quad \text{a.e. on } I, \\ x_{\gamma,\varepsilon}(t_1) &= \tilde{x}_{\gamma,\varepsilon}(t_1) = \bar{x}(t_1) + \varepsilon \sum_{i=1}^{k+1} \gamma^i \delta x_i(t_1). \end{aligned} \tag{7.11}$$

Thus, $x_{\gamma,\varepsilon}$ is an admissible trajectory for all $\gamma \in S^k$ and $\varepsilon \in [0, \hat{\varepsilon})$. Also, note that, for all $\varepsilon \in [0, \hat{\varepsilon})$ and $\gamma \in S^k$,

$$\begin{aligned} |I_{j,\varepsilon,\gamma}| &= \varepsilon \beta^j(\gamma)|I| \quad \text{for each } j = 1, \cdots, s; \\ |I_{0,\varepsilon,\gamma}| &= \left(1 - \varepsilon \sum_{j=1}^{s} \beta^j(\gamma)\right)|I|, \end{aligned} \tag{7.12}$$

so that (see (7.6) and (7.7))

$$|\{t : t \in I, u_{\gamma,\varepsilon}(t) \neq \bar{u}(t)\}| \xrightarrow[\varepsilon \to 0^+]{} 0 \quad \text{uniformly with respect to } \gamma \in S^k. \tag{7.13}$$

Combining (7.3) and (7.8), we conclude that

$$\left\| \frac{x_{\gamma,\varepsilon} - \bar{x}}{\varepsilon} - \delta x_\gamma \right\| \xrightarrow[\varepsilon \to 0^+]{} 0 \qquad \text{uniformly with respect to } \gamma \in S^k. \tag{7.14}$$

8. Perturbed admissible controls.

In the preceding section, we described one way to "perturb" (or make a variation in) the optimal control $\bar{u}$—by constructing a chattering control. Note that the chattering control coincides with $\bar{u} = u_0$ except for t in a set whose measure is proportional to ε (see (7.6), (7.7), and (7.12)). However, at those t where the chattering control differs from $\bar{u}$, this difference is in no sense "small of the order of ε." Thus, chattering controls correspond, in the language of the calculus of variations, to strong variations of $\bar{u}$.

We shall also admit another kind of perturbation of $\bar{u}$—the addition of a term of the form $\varepsilon \delta u$, where $\delta u \in \Delta_0$. This, of course, is a "weak variation." Further, we shall consider combinations of the two kinds of perturbations, i.e., we shall first add a term of the form $\varepsilon \delta u$ to $\bar{u}$, and then shall replace $\bar{u} + \varepsilon \delta u$ by a chattering control, with a term of the form $\varepsilon \delta u_j$ added to u_j (where δu and $\delta u_j \in \Delta_0$). The resultant control functions, which, for sufficiently small $\varepsilon > 0$, will be in $\mathscr{U}$, will be called perturbed admissible controls.

Note that, in our chattering controls, the functions u_j are not only in $\mathscr{U}$, but also in $\hat{\mathscr{U}}$. The reasons for this will now become apparent.

Perturbed controls will be constructed for each $\gamma \varepsilon S^k$, $\varepsilon \in [0, \tilde{\varepsilon})$, and $z \in L^l_\infty(I_1)$ with $\|z\| < \varepsilon_0$, where $\varepsilon_0 > 0$ is specified in Lemma 6.1 and $\tilde{\varepsilon}$ is some sufficiently small positive number. Let

$$B_0 = \{z : z \in L^l_\infty(I_1),\ \|z\| < \varepsilon_0\}.$$

We begin by constructing functions $\delta u_{z,\gamma,j} \in L^m_\infty$ (for each $\gamma \in S^k$, $z \in B_0$, and $j = 1, \cdots, s$) such that (see (6.4))

$$D_1P(\bar{x}, u_j; \delta x_{z,\gamma}) + D_2P(\bar{x}, u_j; \delta u_{z,\gamma,j}) = D_1\bar{P}(\delta x_{z,\gamma}) + D_2\bar{P}(\delta u_{z,\gamma}) = z, \tag{8.1}$$

$$(u_j + \varepsilon\delta u_{z,\gamma,j}) \in \mathscr{U} \quad \text{for all } \varepsilon > 0 \text{ sufficiently small,} \tag{8.2}$$

where $\delta x_{z,\gamma}$ and $\delta u_{z,\gamma}$ are the functions specified in Lemma 6.1.

To do this, we define the continuous linear operators $\tilde{P}_j : \mathscr{C}^n \times L^m_\infty \to L^m_\infty$ for $j = 0, 1, \cdots, s$ by the relations

$$\begin{aligned} &\tilde{P}_j(\delta x, \delta u) = \widetilde{D_2P}(\bar{x}, u_j; D_1\bar{P}(\delta x) - D_1P(\bar{x}, u_j; \delta x) + D_2\bar{P}(\delta u)), \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = 1, \cdots, s, \\ &\tilde{P}_0(\delta x, \delta u) = \delta u, \end{aligned} \tag{8.3}$$

and set

$$\delta\tilde{u}_{z,\gamma,j} = \tilde{P}_j(\delta x_{z,\gamma}, \delta u_{z,\gamma}) \quad \text{for all } z \in B_0, \quad \gamma \in S^k, \quad \text{and } j = 0, \cdots, s. \tag{8.4}$$

Note that, by definition of $\widetilde{D_2P}$ and (6.10), for every z, γ, and j, $\delta\tilde{u}_{z,\gamma,j} \in \Delta_0$. Since $D_2P(\bar{x}, u_j) \circ \widetilde{D_2P}(\bar{x}, u_j)$ is the identity operator on $L^l_\infty(I_1)$ for each j, we have

$$\begin{aligned} &D_2P(\bar{x}, u_j; \tilde{P}_j(\delta x, \delta u)) + D_1P(\bar{x}, u_j; \delta x) = D_1\bar{P}(\delta x) + D_2\bar{P}(\delta u) \\ &\qquad\qquad \text{for all } \delta x \in \mathscr{C}^n, \quad \delta u \in L^m_\infty, \quad j = 1, \cdots, s, \end{aligned} \tag{8.5}$$

so that, in particular, (8.1) holds with $\delta u_{z,\gamma,j}$ replaced by $\delta\tilde{u}_{z,\gamma,j}$. Inasmuch as there is no reason to expect that (8.2) also holds (with this substitution), we must modify $\delta\tilde{u}_{z,\gamma,j}$.

Since (for each j) $u_j \in \hat{\mathscr{U}}$, so that u_j satisfies C2, we can find a function $\delta u_{\theta,j} \in \mathscr{Q}_0(u_j)$ such that $D_2P(\bar{x}, u_j; \delta u_{\theta,j}) = 0$. By Lemma 3.1, there is an $\hat{\varepsilon}_j > 0$ such that $(u_j + \varepsilon\delta u) \in \mathscr{U}$ whenever $0 \leqq \varepsilon < \hat{\varepsilon}_j$, $\delta u \in \Delta_0$, and $\|\delta u - \delta u_{\theta,j}\| < \hat{\varepsilon}_j$.

Now it follows from the conclusions of Lemmas 4.1 and 6.1 and from (8.3) and (8.4) that the set

$$\{\delta\tilde{u}_{z,\gamma,j} : z \in B_0, \gamma \in S^k, j = 1, \cdots, s\}$$

is bounded; say $\|\delta\tilde{u}_{z,\gamma,j}\| < \mu$ for all z, γ, j. If we set (for all z, γ)

$$\begin{aligned} &\delta u_{z,\gamma,j} = \delta\tilde{u}_{z,\gamma,j} + \frac{\mu}{\hat{\varepsilon}_j}\delta u_{\theta,j} \quad \text{for } j = 1, \cdots, s, \\ &\delta u_{z,\gamma,0} = \delta u_{z,\gamma}, \end{aligned} \tag{8.6}$$

then it is easily seen that (for all z, γ, j) (8.1) holds and $\delta u_{z,\gamma,j} \in \Delta_0$. Hence, if we set $\tilde{\varepsilon} = \min [\{\hat{\varepsilon}_j^2/\mu : j = 1, \cdots, s\} \cup \{\varepsilon_0, \hat{\varepsilon}\}]$, we easily conclude, recalling (6.9), that (setting $u_0 = \bar{u}$)

$$(u_j + \varepsilon\delta u_{z,\gamma,j}) \in \mathscr{U} \quad \text{whenever } 0 \leqq \varepsilon < \tilde{\varepsilon}, \quad \text{for all } \gamma \in S^k, \quad z \in B_0, \quad \text{and } j = 0, \cdots, s. \tag{8.7}$$

Note that (see (6.7) and (6.8)) the maps $\gamma \to \delta u_{z,\gamma,j} : S^k \to L_\infty^m$ are continuous for each $z \in B_0$ and $j = 0, \cdots, s$, and that this continuity is uniform with respect to z.

We can now define the perturbed admissible controls. Namely, for each $\gamma \in S^k, \varepsilon \in [0, \tilde{\varepsilon})$, and $z \in B_0$, let the perturbed admissible control $u_{z,\gamma,\varepsilon} \in L_\infty^m$ be defined by

$$u_{z,\gamma,\varepsilon} = u_{\gamma,\varepsilon} + \varepsilon\delta u_{z,\gamma,\varepsilon}, \tag{8.8}$$

where (see (7.7))

$$\delta u_{z,\gamma,\varepsilon}(t) = \sum_{j=0}^{s} e_{j,\varepsilon,\gamma}(t)\delta u_{z,\gamma,j}(t), \qquad t \in I. \tag{8.9}$$

Observe that (see (7.12) and 8.6))

$$|\{t : t \in I, \delta u_{z,\gamma,\varepsilon}(t) \neq \delta u_{z,\gamma}(t)\}| \xrightarrow[\varepsilon \to 0^+]{} 0 \quad \text{uniformly with respect to } z \in B_0 \text{ and } \gamma \in S^k. \tag{8.10}$$

Further, it is evident that (see (7.6))

$$u_{z,\gamma,\varepsilon}(t) = \sum_{j=0}^{s} e_{j,\varepsilon,\gamma}(t)[u_j(t) + \varepsilon\delta u_{z,\gamma,j}(t)] \quad (\text{where } \bar{u} = u_0), \tag{8.11}$$

so that, by (8.7), $u_{z,\gamma,\varepsilon}$ is a chattering combination of functions in $\mathscr{U}$, and is therefore itself in $\mathscr{U}$, for all $z \in B_0, \gamma \in S^k$, and $\varepsilon \in [0, \tilde{\varepsilon})$.

Note that $u_{z,\gamma,0} \equiv \bar{u}$ for all $z \in B_0$ and $\gamma \in S^k$, and that

$$\begin{aligned} &\sup \{\|u_{z,\gamma,\varepsilon}\| : z \in B_0, \gamma \in S^k, \varepsilon \in [0\ \tilde{\varepsilon})\} < \infty, \\ &\sup \{\|\delta u_{z,\gamma,\varepsilon}\| : z \in B_0, \gamma \in S^k, \varepsilon \in [0, \tilde{\varepsilon})\} < \infty. \end{aligned} \tag{8.12}$$

It is also worth noting that, by virtue of (8.9), (7.6), (7.7), (8.1), (2.21), and (2.22), we can conclude that

$$D_1P(\bar{x}, u_{\gamma,\varepsilon}; \delta x_{z,\gamma}) + D_2P(\bar{x}, u_{\gamma,\varepsilon}; \delta u_{z,\gamma,\varepsilon}) = D_1\bar{P}(\delta x_{z,\gamma}) + D_2\bar{P}(\delta u_{z,\gamma}). \tag{8.13}$$

It follows directly from the continuous dependence Theorem A.1 in the Appendix, by virtue of (7.13) and (8.12), that the differential equation

$$\dot{x}(t) = f(x(t), u_{z,\gamma,\varepsilon}(t), t) = f(x(t), u_{\gamma,\varepsilon}(t) + \varepsilon\delta u_{z,\gamma,\varepsilon}(t), t) \quad \text{a.e. on } I$$

with initial value

$$x(t_1) = \bar{x}(t_1) + \varepsilon \sum_{i=1}^{k+1} \gamma^i \delta x_i(t_1)$$

has a solution in $\mathscr{G}$ (for all $\gamma \in S^k$ and $z \in B_0$) so long as $\varepsilon \geqq 0$ is sufficiently small (for convenience, and without loss of generality, we shall say for all $\varepsilon, 0 \leqq \varepsilon < \tilde{\varepsilon}$), and that, if we denote this solution by $x_{z,\gamma,\varepsilon}$, then

(8.14) $$\|x_{z,\gamma,\varepsilon} - \bar{x}\| \xrightarrow[\varepsilon \to 0^+]{} 0 \quad \text{uniformly with respect to } z \text{ and } \gamma.$$

In the same way (see Remark A.1 in the Appendix) we can also conclude—recalling the construction of the sets $I_{j,\varepsilon,\gamma}$—that

(8.15) $$\|x_{z,\gamma,\varepsilon} - x_{z,\gamma_1,\varepsilon}\| \xrightarrow[\substack{\gamma \to \gamma_1 \\ \gamma \in S^k}]{} 0$$
$$\text{uniformly with respect to } z \in B_0 \text{ and } \gamma_1 \in S^k, \quad \text{for each } \varepsilon \in [0, \tilde{\varepsilon}).$$

Note that $x_{z,\gamma,0} = \bar{x}$ for all z and γ.

It follows from Theorem A.2 in the Appendix that, if we denote by $\delta\tilde{x}_{z,\gamma,\varepsilon}$ the solution of the equation

$$\delta\dot{x}(t) = f_x(x_{\gamma,\varepsilon}(t), u_{\gamma,\varepsilon}(t), t)\delta x(t) + f_u(x_{\gamma,\varepsilon}(t), u_{\gamma,\varepsilon}(t), t)\delta u_{z,\gamma,\varepsilon}(t), \qquad \delta x(t_1) = 0,$$

then

(8.16) $$\left\| \frac{x_{z,\gamma,\varepsilon} - x_{\gamma,\varepsilon}}{\varepsilon} - \delta\tilde{x}_{z,\gamma,\varepsilon} \right\| \xrightarrow[\varepsilon \to 0^+]{} 0 \quad \text{for all } \gamma \in S^k \quad \text{and} \quad z \in B_0.$$

It is a consequence of the mean value theorem of differential calculus in Banach spaces (see [8, Thm. (8.5.4), p. 155]), as well as Theorems A.1 and A.2 in the Appendix, that the convergence in (8.16) is uniform with respect to z and γ. On the other hand, it follows from the continuous dependence Theorem A.1 in the Appendix (see (7.13), (8.10), and (8.12), and note that, because of (7.14), $\|x_{\gamma,\varepsilon} - \bar{x}\| \to 0$ as $\varepsilon \to 0$ uniformly with respect to $\gamma \in S^k$) that, if we denote by $\delta\tilde{\tilde{x}}_{z,\gamma}$ the solution of the equation

(8.17) $$\delta\dot{x}(t) = \bar{f}_x(t)\delta x(t) + \bar{f}_u(t)\delta u_{z,\gamma}(t) \quad \text{a.e. on } I, \qquad \delta x(t_1) = 0,$$

then

(8.18) $$\|\delta\tilde{x}_{z,\gamma,\varepsilon} - \delta\tilde{\tilde{x}}_{z,\gamma}\| \xrightarrow[\varepsilon \to 0+]{} 0 \quad \text{uniformly with respect to } z, \gamma.$$

Note that, by (7.4), (6.1), (6.2), and (8.17),

(8.19) $$\delta x_{z,\gamma} = \delta x_\gamma + \delta\tilde{\tilde{x}}_{z,\gamma}.$$

Combining (8.16), (8.18), (8.19), and (7.14), we conclude that

(8.20) $$\left\| \frac{x_{z,\gamma,\varepsilon} - \bar{x}}{\varepsilon} - \delta x_{z,\gamma} \right\| \xrightarrow[\varepsilon \to 0^+]{} 0 \quad \text{uniformly with respect to } z, \gamma.$$

9. An implicit function type theorem. For each $\gamma \in S^k$ and $\varepsilon \in [0, \tilde{\varepsilon})$, we shall consider the mapping

$$z \to P(x_{z,\gamma,\varepsilon}, u_{z,\gamma,\varepsilon}): B_0 \to L_\infty^l(I_1).$$

Let us denote this map by $P_{\gamma,\varepsilon}$, so that

$$P_{\gamma,\varepsilon}(z) = P(x_{z,\gamma,\varepsilon}, u_{z,\gamma,\varepsilon}), \qquad z \in B_0. \tag{9.1}$$

Note that

$$P_{\gamma,0}(z) = P(x_{z,\gamma,0}, u_{z,\gamma,0}) = P(\bar{x}, \bar{u}) = 0 \quad \text{for all } z \text{ and } \gamma.$$

Continuing our contradiction argument, we shall prove the following result, which is a special kind of implicit function theorem.

THEOREM 9.1. *If* $0 \in \operatorname{int} K$, *then, for every* $\rho \in (0, \varepsilon_0)$, *there exist an* $\varepsilon_\rho \in (0, \tilde{\varepsilon})$ *and functions* $z_{\rho,\gamma} \in L^l_\infty(I_1)$ *defined for each* $\gamma \in S^k$ *such that*

(i) $\|z_{\rho,\gamma}\| \leqq \rho$ *for all* $\gamma \in S^k$,

(ii) $P_{\gamma,\varepsilon_\rho}(z_{\rho,\gamma}) = 0$ *for all* $\gamma \in S^k$,

(iii) *the mapping* $\gamma \to x_{z_{\rho,\gamma},\gamma,\varepsilon_\rho} : S^k \to \mathscr{C}^n$ *is continuous,*

(iv) $0 < \varepsilon_\rho < \rho$.

Note that the pairs $(x_{z_{\rho,\gamma},\gamma,\varepsilon_\rho}, u_{z_{\rho,\gamma},\gamma,\varepsilon_\rho})$ of Theorem 9.1 are admissible and satisfy equation (2.2) as well as the equality constraint $P = 0$.

Our proof of Theorem 9.1 will be carried out by showing that—for an arbitrary $\rho \in (0, \varepsilon_0)$—for every $\varepsilon > 0$ sufficiently small, the mapping (for any $\gamma \in S^k$) $\hat{P}_{\gamma,\varepsilon} : B_0 \to L^l_\infty(I_1)$ defined by

$$\hat{P}_{\gamma,\varepsilon}(z) = z - \frac{1}{\varepsilon} P_{\gamma,\varepsilon}(z), \qquad z \in B_0, \tag{9.2}$$

has a fixed point in $\{z : z \in L^l_\infty(I_1), \|z\| \leqq \rho\}$. To do this, we shall use the fixed-point theorem for contraction mappings.

(Actually, to first order in ε, the mapping (9.2) vanishes identically. In fact, the functions $x_{z,\gamma,\varepsilon}$ and $u_{z,\gamma,\varepsilon}$ were specifically constructed to achieve this. Indeed (see (9.1), (7.8), (8.20), (8.8), (8.13), and (6.4)),

$$\begin{aligned} P_{\gamma,\varepsilon}(z) &= P(x_{z,\gamma,\varepsilon}, u_{z,\gamma,\varepsilon}) - P(\bar{x}, u_{\gamma,\varepsilon}) \\ &= P(\bar{x} + \varepsilon \delta x_{z,\gamma} + \cdots, u_{\gamma,\varepsilon} + \varepsilon \delta u_{z,\gamma,\varepsilon}) - P(\bar{x}, u_{\gamma,\varepsilon}) \\ &= \varepsilon(D_1 \bar{P}(\delta x_{z,\gamma}) + D_2 \bar{P}(\delta u_{z,\gamma})) + \cdots = \varepsilon z + \cdots \quad \text{for all } \gamma \in S^k, \end{aligned}$$

where three dots denote terms of higher than first order in ε.)

Proof of Theorem 9.1. For each $\gamma \in S^k$, let us define the map $V_\gamma : B_0 \to \mathscr{C}^n \times L^m_\infty$ as follows:

$$V_\gamma(z) = (\delta x_{z,\gamma}, \delta u_{z,\gamma}). \tag{9.3}$$

Further, for each $\gamma \in S^k$ and $\varepsilon \in [0, \tilde{\varepsilon})$, let $H_{\gamma,\varepsilon}$ be the function from $\mathscr{C}^n \times L^m_\infty$ into L^m_∞ defined by (see (8.3))

$$H_{\gamma,\varepsilon}(\delta x, \delta u) = \sum_{j=1}^{s} E_{j,\varepsilon,\gamma}\left(u_j + \varepsilon \tilde{P}_j(\delta x, \delta u) + \varepsilon \frac{\mu}{\hat{\varepsilon}_j} \delta u_{\theta,j}\right) + E_{0,\varepsilon,\gamma}(\bar{u} + \varepsilon \delta u), \tag{9.4}$$

where, for each $j = 0, \cdots, s$, $\varepsilon \in [0, \tilde{\varepsilon})$, and $\gamma \in S^k$, $E_{j,\varepsilon,\gamma}$ is the continuous linear function from L^m_∞ into L^m_∞ defined by (see (7.7))

$$(E_{j,\varepsilon,\gamma} u)(t) = e_{j,\varepsilon,\gamma}(t) u(t), \qquad t \in I, \quad u \in L^m_\infty. \tag{9.5}$$

Also, let $F_{\gamma,\varepsilon}$ be the map which assigns to each $u \in \mathscr{U}_0$ such that equation (2.2), with initial condition $x(t_1) = \bar{x}(t_1) + \varepsilon \sum_{i=1}^{k+1} \gamma^i \delta x_i(t_1)$, has a (necessarily unique) solution defined on all of I, this particular solution. We point out that (see (8.4), (8.6), and (8.11)) $H_{\gamma,\varepsilon} \circ V_\gamma(z) = u_{z,\gamma,\varepsilon}$ and that $F_{\gamma,\varepsilon}(u_{z,\gamma,\varepsilon}) = x_{z,\gamma,\varepsilon}$ for all $\varepsilon \in [0, \tilde{\varepsilon})$, $\gamma \in S^k$, and $z \in B_0$, from which it immediately follows that (see (9.1))

$$(9.6) \qquad P_{\gamma,\varepsilon}(z) = P(F_{\gamma,\varepsilon} \circ H_{\gamma,\varepsilon} \circ V_\gamma(z), H_{\gamma,\varepsilon} \circ V_\gamma(z)) \quad \text{for all } z, \gamma, \varepsilon.$$

Note that the maps V_γ and $H_{\gamma,\varepsilon}$ are continuous and affine (see (6.7), (6.8), and Lemma 4.1), and therefore Fréchet differentiable with constant differentials. Further, it follows from Theorem A.2 in the Appendix that $F_{\gamma,\varepsilon}$ is Fréchet differentiable, and we have seen in § 2 that P also is. Hence, $P_{\gamma,\varepsilon}$ is Fréchet differentiable, and by the chain rule of differentiation, its differential at any $z \in B_0$ is given by

$$(9.7) \qquad \begin{aligned} DP_{\gamma,\varepsilon}(z) = {} & D_1P(x_{z,\gamma,\varepsilon}, u_{z,\gamma,\varepsilon}) \circ DF_{\gamma,\varepsilon}(u_{z,\gamma,\varepsilon}) \circ DH_{\gamma,\varepsilon} \circ DV_\gamma \\ & + D_2P(x_{z,\gamma,\varepsilon}, u_{z,\gamma,\varepsilon}) \circ DH_{\gamma,\varepsilon} \circ DV_\gamma \quad \text{for all } z, \gamma, \varepsilon. \end{aligned}$$

Let us evaluate each of the differentials in the right-hand side of (9.7), in order to obtain estimates for $\|D\hat{P}_{\gamma,\varepsilon}(z)\|$. In fact, we shall show that both $\|D\hat{P}_{\gamma,\varepsilon}(z)\|$ and $\|\hat{P}_{\gamma,\varepsilon}(0)\|$ tend to 0 as $\varepsilon \to 0^+$, for all $\gamma \in S^k$ and $z \in B_0$, which implies that, for $\varepsilon > 0$ sufficiently small, $\hat{P}_{\gamma,\varepsilon}$ is a contraction mapping on B_0. We shall thus be able to conclude, by the fixed-point theorem for contraction mappings, that $\hat{P}_{\gamma,\varepsilon}$—for all $\gamma \in S^k$ and $\varepsilon > 0$ sufficiently small—has a fixed point in B_0, i.e. (see (9.2)), that $P_{\gamma,\varepsilon}$ has a zero in B_0.

Recall that D_1P and D_2P are given by (2.21) and (2.22), and (see (9.3), (6.7), (6.8), and Lemma 4.1) DV_γ is the map $\delta z \to (\delta x_{\delta z}, \delta u_{\delta z})$, independent of γ. Also, since the maps $\tilde{P}_j$ and $E_{j,\varepsilon,\gamma}$ are linear (see (8.3) and (9.5)), $DH_{\gamma,\varepsilon}$ is given by

$$(9.8) \qquad DH_{\gamma,\varepsilon} = \varepsilon \sum_{j=0}^{s} E_{j,\varepsilon,\gamma} \tilde{P}_j.$$

Finally, $DF_{\gamma,\varepsilon}(u_{z,\gamma,\varepsilon})$ is the map which assigns to each $\delta u \in L_\infty^m$ the function $\delta x \in \mathscr{C}^n$ which is a solution of (see Theorem A.2 in the Appendix)

$$(9.9) \qquad \begin{aligned} \delta\dot{x}(t) &= f_x(x_{z,\gamma,\varepsilon}(t), u_{z,\gamma,\varepsilon}(t), t)\delta x(t) + f_u(x_{z,\gamma,\varepsilon}(t), u_{z,\gamma,\varepsilon}(t), t)\delta u(t), \qquad t \in I, \\ \delta x(t_1) &= 0. \end{aligned}$$

It follows from our continuous dependence Theorem A.1 in the Appendix, by virtue of (8.14), (8.8), (7.13), and (8.12), that the solutions of (9.9) tend, as $\varepsilon \to 0^+$, to the solution of

$$(9.10) \qquad \delta\dot{x}(t) = f_x(t)\delta x(t) + \bar{f}_u(t)\delta u(t), \qquad \delta x(t_1) = 0,$$

uniformly with respect to $t \in I$, $\gamma \in S^k$, $z \in B_0$, and $\delta u \in L_\infty^m$ with $\|\delta u\| \leqq 1$. This means that, if we denote by F_0 the linear continuous map from L_∞^m into $\mathscr{C}^n$ which assigns to each $\delta u \in L_\infty^m$ the δx defined by (9.10), then

$$(9.11) \qquad \|DF_{\gamma,\varepsilon}(u_{z,\gamma,\varepsilon}) - F_0\| \xrightarrow[\varepsilon \to 0^+]{} 0 \quad \text{uniformly with respect to } \gamma \in S^k \text{ and } z \in B_0.$$

Further, since $DH_{\gamma,\varepsilon}$ is given by (9.8), it is a consequence of (9.5), (7.7), (7.12), and (8.3) that

$$|\{t : t \in I,\ DH_{\gamma,\varepsilon}(\delta x, \delta u)(t) \neq \varepsilon \delta u(t)\}| \xrightarrow[\varepsilon \to 0^+]{} 0$$

uniformly with respect to $\gamma \in S^k$,

from which we conclude that

$$\left\| \frac{1}{\varepsilon} F_0 \circ DH_{\gamma,\varepsilon}(\delta x, \delta u) - F_0(\delta u) \right\| \xrightarrow[\varepsilon \to 0^+]{} 0 \tag{9.12}$$

uniformly with respect to $\gamma \in S^k$, $\delta u \in L_\infty^m$, and $\delta x \in \mathscr{C}^n$, $\|\delta x\| \leqq 1$, $\|\delta u\| \leqq 1$.

Combining (9.11) and (9.12), we obtain that

$$\left\| \frac{1}{\varepsilon} DF_{\gamma,\varepsilon}(u_{z,\gamma,\varepsilon}) \circ DH_{\gamma,\varepsilon} \circ DV_\gamma - F_0 \circ D_2 V_\gamma \right\| \xrightarrow[\varepsilon \to 0^+]{} 0 \tag{9.13}$$

uniformly with respect to $\gamma \in S^k$ and $z \in B_0$,

where (for any γ) $D_2 V_\gamma$ denotes the map $\delta z \to \delta u_{\delta z} : L_\infty^l(I_1) \to L_\infty^m$.

Also, by virtue of (2.21) and (2.22), Hypothesis A4 on the function p, and (8.14), and because $\|u_{z,\gamma,\varepsilon} - u_{\gamma,\varepsilon}\| \to 0$ as $\varepsilon \to 0^+$ uniformly with respect to $\gamma \in S^k$ and $z \in B_0$ (see (8.8) and (8.12)), we have

$$\|D_i P(x_{z,\gamma,\varepsilon}, u_{z,\gamma,\varepsilon}) - D_i P(\bar{x}, u_{\gamma,\varepsilon})\| \xrightarrow[\varepsilon \to 0^+]{} 0 \tag{9.14}$$

uniformly with respect to γ, z for $i = 1, 2$.

It follows from (9.7), (9.13), and (9.14) that

$$\left\| \frac{1}{\varepsilon} DP_{\gamma,\varepsilon}(z) - D_1 P(\bar{x}, u_{\gamma,\varepsilon}) \circ F_0 \circ D_2 V_\gamma - \frac{1}{\varepsilon} D_2 P(\bar{x}, u_{\gamma,\varepsilon}) \circ DH_{\gamma,\varepsilon} \circ DV_\gamma \right\| \xrightarrow[\varepsilon \to 0^+]{} 0 \tag{9.15}$$

uniformly with respect to $\gamma \in S^k$ and $z \in B_0$.

But, for any $\delta z \in L_\infty^l(I_1)$ and $\gamma \in S^k$ (see (9.10) and (4.7)),

$$F_0 \circ D_2 V_\gamma(\delta z) = F_0(\delta u_{\delta z}) = \delta x_{\delta z}, \tag{9.16}$$

and, for any $\varepsilon \in [0, \tilde{\varepsilon})$ and $\delta z \in L_\infty^l(I_1)$, $(1/\varepsilon) D_2 P(\bar{x}, u_{\gamma,\varepsilon}) \circ DH_{\gamma,\varepsilon} \circ DV_\gamma(\delta z)$ is the function (see (2.21), (2.22), (9.8), (9.5), (7.6), (7.7), and (8.5))

$$\begin{aligned} t \to p_u(\bar{x}(t), u_{\gamma,\varepsilon}(t), t) &\sum_{j=0}^{s} e_{j,\varepsilon,\gamma}(t)[(\tilde{P}_j(\delta x_{\delta z}, \delta u_{\delta z}))(t)] \\ &= \sum_{j=0}^{s} e_{j,\varepsilon,\gamma}(t) p_u(\bar{x}(t), u_j(t), t)[(\tilde{P}_j(\delta x_{\delta z}, \delta u_{\delta z}))(t)] \\ &= \sum_{j=0}^{s} e_{j,\varepsilon,\gamma}(t)[(D_2 P(\bar{x}, u_j; \tilde{P}_j(\delta x_{\delta z}, \delta u_{\delta z})))(t)] \end{aligned}$$

(cont.)

$$= \sum_{j=0}^{s} e_{j,\varepsilon,\gamma}(t)[(D_1\bar{P}(\delta x_{\delta z}) + D_2\bar{P}(\delta u_{\delta z}) - D_1P(\bar{x}, u_j; \delta x_{\delta z}))(t)]$$

$$= (D_1\bar{P}(\delta x_{\delta z}) + D_2\bar{P}(\delta u_{\delta z}) - D_1P(\bar{x}, u_{\gamma,\varepsilon}; \delta x_{\delta z}))(t): I_1 \to R^l,$$

i.e.,

$$\frac{1}{\varepsilon}D_2P(\bar{x}, u_{\gamma,\varepsilon}) \circ DH_{\gamma,\varepsilon} \circ DV_\gamma(\delta z) = D_1\bar{P}(\delta x_{\delta z}) + D_2\bar{P}(\delta u_{\delta z}) - D_1P(\bar{x}, u_{\gamma,\varepsilon}; \delta x_{\delta z})$$

$$\text{for all } \gamma \in S^k, \quad \varepsilon \in [0, \tilde{\varepsilon}), \quad \text{and} \quad \delta z \in L^l_\infty(I_1). \tag{9.17}$$

Combining (9.15)–(9.17), and taking into account (4.6), we see that

$$\left\|\frac{1}{\varepsilon}DP_{\gamma,\varepsilon}(z) - E\right\| \xrightarrow[\varepsilon\to 0^+]{} 0 \quad \text{uniformly with respect to } \gamma \in S^k \quad \text{and} \quad z \in B_0,$$

where E denotes the identity operator on $L^l_\infty(I_1)$, or (see (9.2))

$$\|D\hat{P}_{\gamma,\varepsilon}(z)\| \xrightarrow[\varepsilon\to 0^+]{} 0 \quad \text{uniformly with respect to } \gamma \in S^k \quad \text{and} \quad z \in B_0.$$

Hence, there is an $\tilde{\varepsilon}_1 \in (0, \tilde{\varepsilon})$ such that $\|D\hat{P}_{\gamma,\varepsilon}(z)\| < \frac{1}{2}$ for all $z \in B_0$ and $\gamma \in S^k$, as long as $0 < \varepsilon < \tilde{\varepsilon}_1$. Applying the mean value theorem of the differential calculus (see [8, Thm. (8.5.4), p. 155]), we conclude that

$$\|\hat{P}_{\gamma,\varepsilon}(z_1) - \hat{P}_{\gamma,\varepsilon}(z_2)\| < \tfrac{1}{2}\|z_1 - z_2\| \quad \text{for all } z_1, z_2 \in B_0, \gamma \in S^k, \quad \text{and} \quad \varepsilon \in (0, \tilde{\varepsilon}_1).$$

Let us arbitrarily fix a number $\rho \in (0, \varepsilon_0)$ for the remainder of the argument, and let us show that, for all $\varepsilon > 0$ sufficiently small,

$$\|\hat{P}_{\gamma,\varepsilon}(0)\| < \rho/2 \quad \text{for all } \gamma \in S^k. \tag{9.18}$$

Now (see (9.1) and (9.2))

$$\hat{P}_{\gamma,\varepsilon}(0) = -\frac{1}{\varepsilon}P(x_{z_0,\gamma,\varepsilon}, u_{z_0,\gamma,\varepsilon}) \quad \text{for all } z \in B_0, \ \gamma \in S^k, \quad \text{and} \quad \varepsilon \in [0, \tilde{\varepsilon}),$$
$$\text{where } z_0 \equiv 0. \tag{9.19}$$

But (see (8.8), (7.8), (8.13), and (6.4))

$$P(x_{z_0,\gamma,\varepsilon}, u_{z_0,\gamma,\varepsilon})$$
$$= P\left(\bar{x} + \varepsilon\delta x_{z_0,\gamma} + \varepsilon\left(\frac{x_{z_0,\gamma,\varepsilon} - \bar{x}}{\varepsilon} - \delta x_{z_0,\gamma}\right), u_{\gamma,\varepsilon} + \varepsilon\delta u_{z_0,\gamma,\varepsilon}\right) \tag{9.20}$$
$$- P(\bar{x}, u_{\gamma,\varepsilon}) - \varepsilon[D_1P(\bar{x}, u_{\gamma,\varepsilon}; \delta x_{z_0,\gamma}) + D_2P(\bar{x}, u_{\gamma,\varepsilon}; \delta u_{z_0,\gamma,\varepsilon})].$$

It follows from Assumptions A3 and A4 (also see (2.21) and (2.22)) that the map P is continuously differentiable and that, because $\operatorname{cl}\{u_{\gamma,\varepsilon}(t) : t \in I, \gamma \in S^k, \varepsilon \in [0, \tilde{\varepsilon})\}$ is a compact subset of U_0 (by definition of $\mathscr{U}$), the map $(x, u) \to D_iP(x, u)$ (for $i = 1$ or 2) is uniformly continuous on the set $\{(\bar{x}, u_{\gamma,\varepsilon}) : \gamma \in S^k, \varepsilon \in (0, \tilde{\varepsilon})\}$. Appealing to an evident corollary of the mean value theorem of the differential calculus (see [8, Thm. (8.6.2), p. 156]) and recalling (8.12) and (8.20), we see that (9.19) and (9.20) imply that

$$\|\hat{P}_{\gamma,\varepsilon}(0)\| \xrightarrow[\varepsilon\to 0^+]{} 0 \quad \text{uniformly with respect to } \gamma \in S^k,$$

so that we may choose an $\varepsilon_\rho \in (0, \tilde{\varepsilon}_1)$ such that $\varepsilon_\rho < \rho$ and such that (9.18) holds whenever $0 < \varepsilon \leqq \varepsilon_\rho$.

We may now apply the fixed-point theorem for contraction mappings—see [8, Thm. (10.1.2), p. 261]—to the maps $\hat{P}_{\gamma,\varepsilon_\rho}$ (for $\gamma \in S_k$), and conclude that, for each $\gamma \in S^k$, there is a unique element $z_\gamma \in L^l_\infty(I_1)$ such that $\|z_\gamma\| < \rho$ and such that $\hat{P}_{\gamma,\varepsilon_\rho}(z_\gamma) = z_\gamma$, i.e. (see (9.2)), such that $P_{\gamma,\varepsilon_\rho}(z_\gamma) = 0$. Since the elements z_γ depend on ρ, we shall write $z_{\rho,\gamma}$ in place of z_γ.

It only remains to prove that the map $\gamma \to x_{z_{\rho,\gamma},\gamma,\varepsilon_\rho}: S^k \to \mathscr{C}^n$ is continuous. Let us examine the map

$$(z, \gamma) \to \hat{P}_{\gamma,\varepsilon_\rho}(z): B_0 \times S^k \to L^l_\infty(I_1).$$

Unfortunately, this map is *not* continuous. However, if we consider the $L^l_1(I_1)$-topology on B_0 (rather than the $L^l_\infty(I_1)$-topology), and similarly for the elements $\hat{P}_{\gamma,\varepsilon_\rho}(z)$, then this map is continuous; i.e., the map

$$(9.21) \qquad (z, \gamma) \to \hat{P}_{\gamma,\varepsilon_\rho}(z): B_0 \times S^k \to L^l_1(I_1)$$

is continuous when B_0 is considered to be a subset of $L^l_1(I_1)$. Indeed (see (9.2), (9.6), and (9.3)), it is sufficient to show that the maps

$$(9.22) \qquad (z, \gamma) \to (\delta x_{z,\gamma}, \delta u_{z,\gamma}): B_0 \times S^k \to \mathscr{C}^n \times L^m_1,$$

$$(9.23) \qquad (\delta x, \delta u, \gamma) \to H_{\gamma,\varepsilon_\rho}(\delta x, \delta u): \{(\delta x_{z,\gamma}, \delta u_{z,\gamma}): z \in B_0, \gamma \in S^k\} \times S^k \to L^m_1,$$

$$(9.24) \qquad (u, \gamma) \to F_{\gamma,\varepsilon_\rho}(u): \{u_{z,\gamma,\varepsilon_\rho}: z \in B_0, \gamma \in S^k\} \times S^k \to \mathscr{C}^n,$$

$$(9.25) \qquad (x, u) \to P(x, u): \mathscr{G} \times \{u_{z,\gamma,\varepsilon_\rho}: z \in B_0, \gamma \in S^k\} \to L^l_1(I_1)$$

are continuous when B_0 and $\{u_{z,\gamma,\varepsilon_\rho}: z \in B_0, \gamma \in S^k\}$ are considered to be subsets of $L^l_1(I_1)$ and L^m_1, respectively, and $\{(\delta x_{z,\gamma}, \delta u_{z,\gamma}): z \in B_0, \gamma \in S^k\}$ is to be considered a subset of $\mathscr{C}^n \times L^m_1$.

But the map (9.22) is continuous by the remarks at the end of § 4 and by (6.7) and (6.8). Further, by (8.3), (2.21), (2.22), and the definition of $\widetilde{D_2 P}(\bar{x}, u_j)$, the maps $\tilde{P}_j$, when considered from $\mathscr{C}^n \times L^m_1$ into L^m_1, are continuous, so that, by (9.4), (9.5), (7.7), the construction of the sets $I_{j,\varepsilon,\gamma}$, and the boundedness in the L^m_∞-norm of the sets

$$\left\{\left(u_j + \varepsilon_\rho \tilde{P}_j(\delta x_{z,\gamma}, \delta u_{z,\gamma}) + \varepsilon_\rho \frac{\mu}{\hat{\varepsilon}_j} \delta u_{\theta,j}\right): z \in B_0, \gamma \in S^k, j = 1, \cdots, s\right\},$$

$$\{\bar{u} + \varepsilon_\rho \delta u_{z,\gamma}: z \in B_0, \gamma \in S^k\},$$

the map (9.23) is continuous as well. Finally, we observe that, in any subset of L^m_∞ which is bounded (in the L^m_∞-norm), convergence in measure is equivalent to convergence in the L^1_∞-norm, and similarly with L^m_∞ and L^1_∞ replaced by $L^l_\infty(I_1)$ and $L^1_\infty(I_1)$, respectively. This is easily seen to imply, by virtue of (8.12) and Assumptions A4 and A6 in § 2, that the map (9.25) is continuous and, by virtue of the continuous dependence Theorem A.1 in the Appendix, that the map (9.24) is continuous as well. Thus, (9.21) is continuous when B_0 is considered to be a subset of $L_1^l(I_1)$.

Now the elements $z_{\rho,\gamma}$, which are fixed points of the contraction mappings $\hat{P}_{\gamma,\varepsilon_\rho}$, may be constructed through a Picard iteration process as follows (see [8, pp. 260–261]):

$$z_{\rho,\gamma}^{(0)} = 0,$$
$$z_{\rho,\gamma}^{(j+1)} = \hat{P}_{\gamma,\varepsilon_\rho}(z_{\rho,\gamma}^{(j)}), \qquad j = 0, 1, \cdots,$$

and $z_{\rho,\gamma} = \lim_{j\to\infty} z_{\rho,\gamma}^{(j)}$, where the limit is to be understood in the $L_\infty^l(I_1)$-norm. Arguing as in [8, pp. 260–261], we can show that the continuity of the map (9.21) implies that the map $\gamma \to z_{\rho,\gamma}: S^k \to B_0$ is continuous, when B_0 is considered a subset of $L_1^l(I_1)$. Since the map $(z,\gamma) \to x_{z,\gamma,\varepsilon_\rho}: B_0 \times S^k \to \mathscr{C}^n$ (with B_0 considered a subset of $L_1^l(I_1)$), which is the composite of the maps (9.22)–(9.24), has already been shown to be continuous, this means that the map $\gamma \to x_{z_{\rho,\gamma},\gamma,\varepsilon_\rho}: S^k \to \mathscr{C}^n$ is continuous, as was to be shown.

10. The completion of the contradiction argument. We shall now complete our contradiction argument for the proof of Theorem 5.1. Indeed, we shall show that, for some $\tilde{\rho} \in (0, \varepsilon_0)$ and some $\tilde{\gamma} \in S^k$, when we denote $x_{z_{\tilde{\rho},\tilde{\gamma}},\tilde{\gamma},\varepsilon_{\tilde{\rho}}}$ by $\tilde{x}$ and $u_{z_{\tilde{\rho},\tilde{\gamma}},\tilde{\gamma},\varepsilon_{\tilde{\rho}}}$ by $\tilde{u}$ (so that $\tilde{x} \in \mathscr{G}$ and $\tilde{u} \in \mathscr{U}$, and $\tilde{x}$ and $\tilde{u}$ satisfy equation (2.2)), then $X_1(\tilde{x}) = 0$, $P(\tilde{x}, \tilde{u}) = 0$, $X_2(\tilde{x}) \in R_-^{k_1}$, $\tilde{X}(\tilde{x}) \in Y$, and $X_0(\tilde{x}) < X_0(\bar{x})$, contradicting the fact that $(\bar{x}, \bar{u})$ is a solution of our problem.

Since $X_1(\bar{x}) = 0$, we have that (for all $z \in B_0$, $\gamma \in S^k$, and $\varepsilon \in (0, \bar{\varepsilon})$)

$$\begin{aligned}\frac{1}{\varepsilon} X_1(x_{z,\gamma,\varepsilon}) &= D\bar{X}_1(\delta x_{z,\gamma}) \\ &+ \frac{\|x_{z,\gamma,\varepsilon} - \bar{x}\|}{\varepsilon} \, \frac{X_1(x_{z,\gamma,\varepsilon}) - X_1(\bar{x}) - D\bar{X}_1(x_{z,\gamma,\varepsilon} - \bar{x})}{\|x_{z,\gamma,\varepsilon} - \bar{x}\|} \\ &+ D\bar{X}_1\left(\frac{x_{z,\gamma,\varepsilon} - \bar{x}}{\varepsilon} - \delta x_{z,\gamma}\right),\end{aligned}$$

so that, by (8.20) and the definition of $D\bar{X}_1$, and because the set $\{\delta x_{z,\gamma}: z \in B_0, \gamma \in S^k\}$ is bounded (see Lemma 4.1 and (6.7)),

$$\begin{gathered}\frac{1}{\varepsilon} X_1(x_{z,\gamma,\varepsilon}) \xrightarrow[\varepsilon\to 0^+]{} D\bar{X}_1(\delta x_{z,\gamma}) \\ \text{uniformly with respect to } z \in B_0 \quad \text{and} \quad \gamma \in S^k.\end{gathered} \tag{10.1}$$

In the same way, we can prove that

$$\begin{gathered}\frac{X_i(x_{z,\gamma,\varepsilon}) - X_i(\bar{x})}{\varepsilon} \xrightarrow[\varepsilon\to 0^+]{} D\bar{X}_i(\delta x_{z,\gamma}) \quad \text{for } i = 0, 2 \\ \text{uniformly with respect to } z, \gamma,\end{gathered} \tag{10.2}$$

$$\frac{\tilde{X}(x_{z,\gamma,\varepsilon}) - \tilde{X}(\bar{x})}{\varepsilon} \xrightarrow[\varepsilon\to 0^+]{} D\tilde{X}(\bar{x}; \delta x_{z,\gamma}) \quad \text{uniformly with respect to } z, \gamma. \tag{10.3}$$

Since (see (6.7)) the set $\{\delta x_{z_0,\gamma} : \gamma \in S^k\}$ is compact, it follows from (6.5) and (6.6), because $R^{k_1}_-$ and Y_0 are open, that there is a number $\zeta_0 > 0$ such that

$$D\bar{X}_0(\delta x_{z_0,\gamma}) < -\zeta_0 \quad \text{for all } \gamma \in S^k, \tag{10.4}$$

$$\xi \in R^{k_1}_- \quad \text{whenever } \xi \in R^{k_1} \quad \text{and} \quad |\xi - X_2(\bar{x}) - D\bar{X}_2(\delta x_{z_0,\gamma})| < \zeta_0 \quad \text{for some } \gamma \in S^k, \tag{10.5}$$

$$y \in Y_0 \quad \text{whenever } y \in \mathscr{C}^{l_1} \quad \text{and} \quad \|y - \tilde{X}(\bar{x}) - D\tilde{X}(\bar{x}; \partial x_{z_0,\gamma})\| < \zeta_0 \quad \text{for some } \gamma \in S^k. \tag{10.6}$$

Let S be the simplex in R^k with vertices $v_1, \cdots, v_{k+1}$ whose existence was asserted in Lemma 6.1. Since $0 \in \operatorname{int} S$, there is a $\theta_0 > 0$ such that the θ_0-neighborhood of 0 in R^k is contained in S. Let $\tilde{\rho} \in (0, \varepsilon_0)$ be such that (see (10.1)–(10.3))

$$\left|\frac{1}{\varepsilon} X_1(x_{z,\gamma,\varepsilon}) - D\bar{X}_1(\delta x_{z,\gamma})\right| < \frac{\theta_0}{2}, \tag{10.7}$$

$$\left|\frac{X_i(x_{z,\gamma,\varepsilon}) - X_i(\bar{x})}{\varepsilon} - D\bar{X}_i(\delta x_{z,\gamma})\right| < \frac{\zeta_0}{2} \quad \text{for } i = 0, 2, \tag{10.8}$$

$$\left\|\frac{\tilde{X}(x_{z,\gamma,\varepsilon}) - \tilde{X}(\bar{x})}{\varepsilon} - D\tilde{X}(\bar{x}; \delta x_{z,\gamma})\right\| < \frac{\zeta_0}{2} \tag{10.9}$$

for all $z \in B_0$ and $\gamma \in S^k$ whenever $0 < \varepsilon < \tilde{\rho}$, and, in addition, such that (see Lemma 4.1)

$$|D\bar{X}_1(\delta x_z)| < \theta_0/2, \tag{10.10}$$

$$|D\bar{X}_i(\delta x_z)| < \zeta_0/2 \quad \text{for } i = 0, 2, \quad \text{and} \quad \|D\tilde{X}(\bar{x}; \delta x_z)\| < \zeta_0/2 \tag{10.11}$$

whenever $\|z\| \leqq \tilde{\rho}$. Without loss of generality, suppose that $\tilde{\rho} \leqq 1$.

Note that (see (6.7) and Corollary 4.1),

$$\delta x_{z,\gamma} = \delta x_z + \delta x_{z_0,\gamma} \quad \text{for all } z \in B_0 \quad \text{and} \quad \gamma \in S^k. \tag{10.12}$$

It now follows from (10.7), (10.10), (10.12), and (6.3) that

$$\left|\frac{1}{\varepsilon} X_1(x_{z,\gamma,\varepsilon}) - \sum_{i=1}^{k+1} \gamma^i v_i\right| < \theta_0 \quad \text{whenever } \|z\| \leqq \tilde{\rho}, \quad \gamma = (\gamma^1, \cdots, \gamma^{k+1}) \in S^k, \quad \text{and} \quad 0 < \varepsilon < \tilde{\rho}. \tag{10.13}$$

Further, it follows from (10.8), (10.9), (10.11), and (10.12) that

$$\left|\frac{X_i(x_{z,\gamma,\varepsilon}) - X_i(\bar{x})}{\varepsilon} - D\bar{X}_i(\delta x_{z_0,\gamma})\right| < \zeta_0 \quad \text{for } i = 0, 2, \tag{10.14}$$

$$\left\|\frac{\tilde{X}(x_{z,\gamma,\varepsilon}) - \tilde{X}(\bar{x})}{\varepsilon} - D\tilde{X}(\bar{x}; \delta x_{z_0,\gamma})\right\| < \zeta_0 \tag{10.15}$$

whenever $\|z\| \leqq \tilde{\rho}$, $\gamma \in S^k$, and $0 < \varepsilon < \tilde{\rho}$. But (10.14), (10.4), and (10.5) imply that

$$X_0(x_{z,\gamma,\varepsilon}) < X_0(\bar{x}), \tag{10.16}$$

$$\frac{X_2(x_{z,\gamma,\varepsilon}) - X_2(\bar{x})}{\varepsilon} + X_2(\bar{x}) \in R^{k_1}_- . \tag{10.17}$$

Since $X_2(\bar{x}) \in \operatorname{cl} R^{k_1}_-$ and $(1 - \varepsilon)[\operatorname{cl} R^{k_1}_-] + \varepsilon R^{k_1}_- \subset R^{k_1}_-$ whenever $0 < \varepsilon \leqq 1$, (10.17) implies that

$$X_2(x_{z,\gamma,\varepsilon}) \in R^{k_1}_- . \tag{10.18}$$

Similarly, we can show that, if $0 < \varepsilon < 1$, then (see (2.14) and (2.15))

$$\tilde{X}(x_{z,\gamma,\varepsilon}) \in Y. \tag{10.19}$$

Hence, (10.13), (10.16), (10.18), and (10.19) hold whenever $\|z\| \leqq \tilde{\rho}$, $\gamma \in S^k$, $0 < \varepsilon < \tilde{\rho}$.

Appealing to Theorem 9.1, we can conclude that there are a number $\varepsilon_{\tilde{\rho}} \in (0, \tilde{\varepsilon})$ and functions $z_{\tilde{\rho},\gamma} \in L^l_\infty(I_1)$, defined for all $\gamma = (\gamma^1, \cdots, \gamma^{k+1}) \in S^k$, such that, if we denote $x_{z_{\tilde{\rho},\gamma},\gamma,\varepsilon_{\tilde{\rho}}}$ by x_γ, and similarly define u_γ (for each γ), then

$$\left| \frac{1}{\varepsilon_{\tilde{\rho}}} X_1(x_\gamma) - \sum_{i=1}^{k+1} \gamma^i v_i \right| < \theta_0, \tag{10.20}$$

$$X_0(x_\gamma) < X_0(\bar{x}), \tag{10.21}$$

$$X_2(x_\gamma) \in R^{k_1}_- \quad \text{and} \quad \tilde{X}(x_\gamma) \in Y, \tag{10.22}$$

$$P(x_\gamma, u_\gamma) = 0 \tag{10.23}$$

for all $\gamma \in S^k$, and such that the map

$$\gamma \to X_1(x_\gamma) : S^k \to R^k \tag{10.24}$$

is continuous.

Now let us consider the map Γ which assigns to each point $\eta \in S$, with barycentric coordinates $\gamma^1, \cdots, \gamma^{k+1}$, the point $[\eta - (1/\varepsilon_{\tilde{\rho}})X_1(x_\gamma)]$ in R^k, where $\gamma = (\gamma^1, \cdots, \gamma^{k+1})$. Since the map (10.24) is continuous, Γ also is. By (10.20), Γ maps S into itself, so that, by the Brouwer fixed-point theorem, $\Gamma(\tilde{\eta}) = \tilde{\eta}$ for some $\tilde{\eta} \in S$, i.e.,

$$X_1(x_{\tilde{\gamma}}) = 0 \quad \text{for some } \tilde{\gamma} \in S^k. \tag{10.25}$$

But, by (10.22), (10.23), and (10.25), $x_{\tilde{\gamma}}$ and $u_{\tilde{\gamma}}$ satisfy all of the problem equality and inequality constraints. Since also $x_{\tilde{\gamma}} \in \mathscr{G}$, $u_{\tilde{\gamma}} \in \mathscr{U}$, and $x_{\tilde{\gamma}}$ and $u_{\tilde{\gamma}}$ satisfy equation (2.2), (10.21) contradicts that $(\bar{x}, \bar{u})$ is a solution of our problem, which completes the proof of Theorem 5.1.

11. A multiplier rule. In this section, we shall show that Theorem 5.1 leads to a generalized multiplier rule. In § 12, we shall expand this multiplier rule to obtain a maximum principle.

THEOREM 11.1. *Let* $(\bar{x}, \bar{u})$ *be a solution of our optimal control problem subject to Assumptions* A1–A6 *on the problem data, and suppose that* $\bar{u}$ *satisfies Conditions* C1 *and* C2. *Then there exist a vector* $\beta = (\beta^0, \beta^1, \cdots, \beta^{k+k_1}) \in R^{k+k_1+1}$, *and continuous linear functionals* $l' \in (L^r_\infty)^*$, $l'' \in (L^l_\infty(I_1))^*$, *and* $l''' \in (\mathscr{C}^{l_1})^*$ (*where the* *asterisk denotes dual space), not all zero, such that*

$$\beta \cdot D\bar{X}(\delta x) + (l' \circ D\bar{Q} + l'' \circ D_2\bar{P})(\delta u) + l'' \circ D_1\bar{P}(\delta x) + l''' \circ D\tilde{X}(\bar{x};\delta x) \leqq 0$$

(11.1) $$\textit{for all } (\delta x, \delta u) \in Z \quad \textit{with} \quad \delta u \in \Delta_0,$$

(11.2) $$\beta^j \leqq 0 \quad \textit{for } j = 0, k+1, \cdots, k+k_1,$$

(11.3) $$(\beta^{k+1}, \cdots, \beta^{k+k_1}) \cdot X_2(\bar{x}) = 0,$$

(11.4) $$l'(w) \geqq 0 \quad \textit{for all } w \in W_2,$$

(11.5) $$l'''(y) \geqq 0 \quad \textit{for all } y \in Y,$$

(11.6) $$l' \circ Q(\bar{u}) = l''' \circ \tilde{X}(\bar{x}) = 0.$$

Proof. By Theorem 5.1, $0 \notin \operatorname{int} K$. If $0 \notin K$ or if $\operatorname{int} K \neq \emptyset$, then by the separation theorem in finite-dimensional spaces (see [9, p. 162, Lemma 2] and recall that K is convex), there is a nonzero vector $\alpha = (\alpha^1, \cdots, \alpha^k) \in R^k$ such that

(11.7) $$\alpha \cdot D\bar{X}_1(\delta x) \leqq 0 \quad \text{whenever } (\delta x, \delta u) \in Z_3 \quad \text{for some } \delta u \in L_\infty^m.$$

If $0 \in K$ and $\operatorname{int} K = \emptyset$, then K lies in a linear manifold in R^k of dimension less than k, and there is consequently also a nonzero vector $\alpha \in R^k$ such that (11.7) holds—even with equality.

Now consider the set

$$\tilde{K} = \{(D_1\bar{P}(\delta x) + D_2\bar{P}(\delta u), \alpha \cdot D\bar{X}_1(\delta x) - \gamma) : (\delta x, \delta u) \in Z_1 \cap Z, \gamma > 0\}.$$

Since Z_1 and Z are convex, $\tilde{K}$ is a convex set in $L_\infty^l(I_1) \times R$. By (11.7), (5.2), and (5.3),

$$\tilde{K} \cap [\{0\} \times R_+] = \emptyset,$$

where R_+ denotes the set of positive numbers. Note that because $\mathscr{Q}(\bar{u})$, $R_-^{k_1}$, and Y_0 are open and $D\bar{X}_j(\cdot)$ $(j = 0, 2)$ and $D\tilde{X}(\bar{x};\cdot)$ are continuous, Z_1 (see (5.1)) is open in the relative topology of $\mathscr{C}^n \times \Delta_0$ (see (3.10)). Recalling that Z and Δ_0 are convex cones, and taking into account (4.6), (4.8), and Lemma 4.1, we at once conclude that $\tilde{K}$ is open.

Hence, by the principal separation theorem for convex sets in linear topological spaces (see [10, Thm. V.2.8, p. 417]), there is a nonzero continuous linear functional $\tilde{l} \in (L_\infty^l(I_1) \times R)^*$ such that $\tilde{l}(z, \zeta) \leqq \tilde{l}(0, \gamma)$ for all $(z, \zeta) \in \tilde{K}$ and $\gamma > 0$, i.e., there are a continuous linear functional $\tilde{l}'' \in (L_\infty^l(I_1))^*$ and a number α^0, not both zero, such that

$$\tilde{l}'' \circ D_1\bar{P}(\delta x) + \tilde{l}'' \circ D_2\bar{P}(\delta u) + \alpha^0\alpha \cdot D\bar{X}_1(\delta x) \leqq 0 \quad \text{for all } (\delta x, \delta u) \in Z_1 \cap Z.$$

By definition of Z_1 (see (5.1), (3.7), (3.8), and (3.10)), this means that the convex sets

(11.8) $$\begin{gathered}\{(Q(\bar{u}) + D\bar{Q}(\delta u), X_2(\bar{x}) + D\bar{X}_2(\delta x), \tilde{X}(\bar{x}) + D\tilde{X}(\bar{x};\delta x), D\bar{X}_0(\delta x), \\ \tilde{l}'' \circ D_1\bar{P}(\delta x) + \tilde{l}'' \circ D_2\bar{P}(\delta u) + \alpha^0\alpha \cdot D\bar{X}_1(\delta x)) : (\delta x, \delta u) \in Z, \delta u \in \Delta_0\}\end{gathered}$$

and

(11.8′) $$W_1 \times R_-^{k_1} \times Y_0 \times R_- \times R_+$$

(where R_- denotes the set of all negative numbers) in $L_\infty^r \times R^{k_1} \times \mathscr{C}^{l_1} \times R \times R$ have an empty intersection. Also, observe that because $0 \in Z$ and $(\bar{x}, \bar{u})$ satisfies the problem constraints, $(Q(\bar{u}), X_2(\bar{x}), \tilde{X}(\bar{x}), 0, 0)$ belongs to the set (11.8) as well as to the closure of the set (11.8′). Since the set (11.8′) is open, we may again appeal to the principal separation theorem for convex sets in linear topological spaces, and conclude that there are continuous linear functionals $l' \in (L_\infty^r)^*$ and $l''' \in (\mathscr{C}^{l_1})^*$, a vector $\tilde{\beta} = (\tilde{\beta}^1, \cdots, \tilde{\beta}^{k_1}) \in R^{k_1}$, and numbers β^0 and $\hat{\beta}^0$, not all of them zero, such that, if we set $\beta^i = \hat{\beta}^0\alpha^0\alpha^i$ for $i = 1, \cdots, k$, $\beta^{k+1} = \tilde{\beta}^i$ for $i = 1, \cdots, k_1$, $\beta = (\beta^0, \beta^1, \cdots, \beta^{k+k_1})$, and $l'' = \hat{\beta}^0\tilde{l}''$ (so that $l'' \in (L_\infty^l(I_1))^*$ and $\beta \in R^{k+k_1+1}$), then (11.1) holds, l', l'', l''', and β are not all zero, and

$$l'(w) \geqq l' \circ Q(\bar{u}) \quad \text{for all } w \in W_2, \tag{11.9}$$

$$\tilde{\beta} \cdot \xi \geqq \tilde{\beta} \cdot X_2(\bar{x}) \quad \text{for all } \xi \in \operatorname{cl} R_-^{k_1}, \tag{11.10}$$

$$l'''(y) \geqq l''' \circ \tilde{X}(\bar{x}) \quad \text{for all } y \in Y, \tag{11.11}$$

$$\beta^0\gamma^0 \geqq 0 \quad \text{for all } \gamma^0 < 0. \tag{11.12}$$

Since $\frac{1}{2}Q(\bar{u})$ and $2Q(\bar{u})$ both belong to W_2, (11.9) implies that $l' \circ Q(\bar{u}) = 0$ and that (11.4) holds. Similarly, (11.10) and (11.12) imply that (11.2) and (11.3) hold, and it is a consequence of (11.11) that (11.5) holds and that $l''' \circ \tilde{X}(\bar{x}) = 0$, so that (11.6) holds.

THEOREM 11.2. *The linear functionals l' and l'' in Theorem* 11.1 *may be extended to L_1^r and to $L_1^l(I_1)$, respectively, in such a way that the extensions belong to $(L_1^r)^*$ and to $(L_1^l(I_1))^*$, respectively.*

Proof. For each $j = 1, \cdots, l$, let z_j^* denote the function in $L_\infty^l(I_1)$ whose jth component is identically equal to one on I_1 and whose other components identically vanish. Since $\bar{u}$ satisfies C2, there are functions $\delta u_j'$ and $\delta u_j''$ $(j = 1, \cdots, l)$ in $\mathscr{Q}(\bar{u}) \cap \Delta_0$ such that $D_2\bar{P}(\delta u_j') = z_j^*$ and $D_2\bar{P}(\delta u_j'') = -z_j^*$ for each $j = 1, \cdots, l$.

If $z = (z^1, \cdots, z^l) \in L_\infty^l(I_1)$, then clearly

$$z = \sum_{j=1}^{l} (z_+^j z_j^* - z_-^j z_j^*),$$

where (for each j) $z_+^j(t) = \max\{0, z^j(t)\}$ and $z_-^j(t) = \max\{0, -z^j(t)\}$ for all $t \in I_1$. Let us extend the functions z_+^j and z_-^j to I by setting them equal to 0 on $I \setminus I_1$. Then (see (2.22))

$$D_2\bar{P}\left(\sum_{j=1}^{l} (z_+^j \delta u_j' + z_-^j \delta u_j'')\right) = z. \tag{11.13}$$

For each $z = (z^1, \cdots, z^l) \in L_\infty^l(I_1)$, let

$$\delta\tilde{u}_z = \sum_{j=1}^{l} (z_+^j \delta u_j' + z_-^j \delta u_j''). \tag{11.14}$$

It is not difficult to verify that there is an $\alpha > 0$ such that $Q(\bar{u}) + D\bar{Q}(\alpha\delta\tilde{u}_z) \in W_2$ (see (3.7), (3.8), (2.18), and (2.19)), so that, by (11.4) and (11.6),

$$l' \circ D\bar{Q}(\delta\tilde{u}_z) \geqq 0. \tag{11.15}$$

Now let $\delta\tilde{x}_z$ be the solution of

$$\delta\dot{x} = \bar{f}_x\delta x + \bar{f}_u\delta\tilde{u}_z, \qquad \delta x(t_1) = 0, \tag{11.16}$$

so that $(\delta\tilde{x}_z, \delta\tilde{u}_z) \in Z$. Also, $\delta\tilde{u}_z \in \Delta_0$. Hence, (11.1) and (11.13)–(11.15) imply that

$$\begin{gathered} l''(z) \leqq -\beta \cdot D\bar{X}(\delta\tilde{x}_z) - l'' \circ D_1\bar{P}(\delta\tilde{x}_z) - l''' \circ D\tilde{X}(\bar{x}; \delta\tilde{x}_z) \\ \text{for all } z \in L^l_\infty(I_1). \end{gathered} \tag{11.17}$$

By the variations of parameters formula for the solution of (11.16) and by (11.14), the map $z \to \delta\tilde{x}_z : L^l_\infty(I_1) \to \mathscr{C}^n$ is continuous, even with the $L^l_1(I_1)$-topology on $L^l_\infty(I_1)$. Also, $\delta\tilde{x}_{z_0} = 0$, where $z_0 \equiv 0$. Hence, (11.17) implies that $l''(z) \to 0$ as $z \to 0$, $z \in L^l_\infty(I_1)$, where $z \to 0$ is to be understood in the $L^l_1(I_1)$-topology. Thus, l'' is continuous on $L^l_\infty(I_1)$, when $L^l_\infty(I_1)$ is viewed as a linear manifold in $L^l_1(I_1)$. Hence, by the Hahn–Banach theorem, l'' can be extended to $L^l_1(I_1)$ in such a way that the extension belongs to $(L^l_1(I_1))^*$.

We now turn to l'. Let δu_0 be an arbitrary fixed function in $\tilde{\mathscr{Q}}(\bar{u}) \cap \Delta_0$, so that there is a constant $\alpha_0 > 0$ such that $Q(\bar{u}) + D\bar{Q}(\delta u_0) + \alpha_0 w_1 \in W_2$, where w_1 is the function in L^r_∞ all of whose components are identically equal to 1. For every $w \in L^r_\infty$, let $|w|$ denote the function $t \to |w(t)| : I \to R$. Note that, for every $w \in L^r_\infty$, $|w|Q(\bar{u}) \in W_2$ (since $Q(\bar{u}) \in W_2$), so that, by (11.4), $l'(|w|Q(\bar{u})) \geqq 0$. On the other hand, if $w \in L^r_\infty$ and $w \neq 0$, then evidently

$$-\frac{1}{\|w\|}|w|Q(\bar{u}) + Q(\bar{u}) \in W_2,$$

from which it follows, by virtue of (11.4) and (11.6), that $l'(|w|Q(\bar{u})) \leqq 0$, so that

$$l'(|w|Q(\bar{u})) = 0 \quad \text{for all } w \in L^r_\infty. \tag{11.18}$$

Further, for each $w \in L^r_\infty$, let $\delta\hat{x}_w$ be the solution of

$$\delta\dot{x}(t) = \bar{f}_x(t)\delta x(t) + \bar{f}_u(t)|w(t)|\delta u_0(t), \quad \delta x(t_1) = 0, \tag{11.19}$$

so that $(\delta\hat{x}_w, |w|\delta u_0) \in Z$ and $|w|\delta u_0 \in \Delta_0$ for all $w \in L^r_\infty$.

Now, for every $w \in L^r_\infty$,

$$\begin{gathered} |w|[Q(\bar{u}) + D\bar{Q}(\delta u_0) + \alpha_0 w_1] = |w|Q(\bar{u}) + D\bar{Q}(|w|\delta u_0) + \alpha_0|w|w_1 \in W_2, \\ -\alpha_0|w|w_1 + \alpha_0 w \in W_2. \end{gathered}$$

Hence, by (11.4) and (11.18),

$$-\alpha_0 l'(w) \leqq l' \circ D\bar{Q}(|w|\delta u_0).$$

Appealing to (11.1), with $\delta u = |w|\delta u_0$ and $\delta x = \delta\hat{x}_w$, we obtain that, for all $w \in L^r_\infty$,

$$\begin{aligned} -\alpha_0 l'(w) \leqq -\beta \cdot D\bar{X}(\delta\hat{x}_w) - l''(D_2\bar{P}(|w|\delta u_0) + D_1\bar{P}(\delta\hat{x}_w)) \\ - l''' \circ D\tilde{X}(\bar{x}; \delta\hat{x}_w). \end{aligned} \tag{11.20}$$

As before, examining equation (11.19), we conclude that the map $w \to \delta\hat{x}_w : L^r_\infty \to \mathscr{C}^n$ is continuous, even with the L^r_1-topology on L^r_∞. On the basis of (11.20) and the previously derived continuity property of l'' with respect to the $L^l_1(I_1)$-topology on $L^l_\infty(I_1)$, we then obtain that $l'(w) \to 0$ as $w \to 0$, $w \in L^r_\infty$, where $w \to 0$ is to be understood in the L^r_1-topology. Appealing to the Hahn–Banach theorem as before, we obtain our desired conclusion.

12. The maximum principle. In this section, we shall expand the necessary conditions of Theorem 11.1 by taking into account the special character of the problem data X, P, etc.

Indeed, using the standard representation theorems for L^*_1, $(L_1(I_1))^*$, and $\mathscr{C}^*$, we conclude, on the basis of Theorems 11.1 and 11.2, that there exist functions $\lambda = (\lambda^1, \cdots, \lambda^{l_1}): I \to R^{l_1}$, $\mu = (\mu^1, \cdots, \mu^r) \in L^r_\infty$, and $v = (v^1, \cdots, v^l) \in L^l_\infty$ with the properties that (i) λ is of bounded variation and is continuous from the right in (t_1, t_2), (ii) $\lambda(t_2) = 0$, and (iii) $v(t) = 0$ for almost all $t \in I \setminus I_1$, together with a vector $\beta = (\beta^0, \beta^1, \cdots, \beta^{k+k_1}) \in R^{k+k_1+1}$ such that (see (2.21)–(2.25) and (2.8)–(2.12))

$$\begin{aligned}&\sum_{i=1}^{\sigma} \beta \cdot \bar{\chi}_{x_i}\delta x(\tau_i) + \int_{t_1}^{t_2} \mu(t)\cdot \bar{q}_u(t)\delta u(t)\,dt + \int_{t_1}^{t_2} v(t)\cdot[\bar{p}_u(t)\delta u(t) + \bar{p}_x(t)\delta x(t)]\,dt \\ &\qquad + \sum_{j=1}^{l_1}\int_{t_1}^{t_2} \tilde{\chi}^j_x(\bar{x}(t), t)\delta x(t)\,d\lambda^j(t) \leqq 0 \quad \text{for all } (\delta x, \delta u) \in Z \quad \text{with } \delta u \in \Delta_0,\end{aligned} \tag{12.1}$$

$$\int_{t_1}^{t_2} \mu(t)\cdot w(t)\,dt \geqq 0 \quad \text{for all } w \in W_2 \quad \text{and} \quad \int_{t_1}^{t_2} \mu(t)\cdot \bar{q}(t)\,dt = 0, \tag{12.2}$$

$$\begin{aligned}&\sum_{j=1}^{l_1}\int_{t_1}^{t_2} y^j(t)\,d\lambda^j(t) \geqq 0 \quad \text{for all } y = (y^1, \cdots, y^{l_1}) \in Y \\ &\qquad \text{and} \quad \sum_{j=1}^{l_1}\int_{t_1}^{t_2} \tilde{\chi}^j(\bar{x}(t), t)\,d\lambda^j(t) = 0,\end{aligned} \tag{12.3}$$

$$\sum_{j=k+1}^{k+k_1} \beta^j \chi^j(\bar{x}(\tau_1), \cdots, \bar{x}(\tau_\sigma)) = 0, \tag{12.4}$$

$$\beta^j \leqq 0 \quad \text{for } j = 0, k+1, \cdots, k+k_1, \tag{12.5}$$

and such that β, λ, μ, and v do not all vanish.

It easily follows from (12.3) and (2.14), because $\tilde{\chi}^j(\bar{x}(t), t) \leqq 0$ for all $t \in I_j$ and each j, that (for each j) λ^j is nonincreasing on I and is constant on any subinterval of I which does not meet the set

$$\{t : t \in I'_j, \tilde{\chi}^j(\bar{x}(t), t) = 0\}. \tag{12.6}$$

Further, (12.2) and (2.19) are easily seen to imply that, for each $j = 1, \cdots, r$, $\mu^j(t) \leqq 0$ a.e. on I and $\mu^j(t) = 0$ a.e. on $I \setminus I_1$, and, because $\bar{q}^j(t) \leqq 0$ a.e. on I, that $\bar{q}^j(t)\mu^j(t) = 0$ a.e. on I.

Recall that $Z = \{(\delta x, \delta u): \delta x \in \mathscr{C}^n, \delta u \in L^m_\infty, \delta x$ and δu satisfy equation (4.1) for some $\Delta f \in M\}$. By the variations of parameters formula, the solutions of (4.1) can be written in the form

$$\delta x(t) = \Phi(t)\delta x(t_1) + \Phi(t)\int_{t_1}^{t} \Phi^{-1}(s)[\bar{f}_u(s)\delta u(s) + \Delta f(s)]\,ds, \qquad t \in I, \tag{12.7}$$

where Φ is the absolutely continuous $(n \times n)$-matrix-valued function defined on I which satisfies

$$(12.8) \qquad \dot{\Phi}(t) = \bar{f}_x(t)\Phi(t) \quad \text{a.e. on } I, \quad \Phi(t_1) = \text{the identity matrix}.$$

Using (12.7) in (12.1), we obtain that

$$\begin{aligned}
&\sum_{i=1}^{\sigma} \beta \cdot \bar{\chi}_{x_i}\Phi(\tau_i)\xi + \sum_{i=1}^{\sigma} \beta \cdot \bar{\chi}_{x_i}\Phi(\tau_i)\int_{t_1}^{\tau_i} \Phi^{-1}(s)[\bar{f}_u(s)\delta u(s) + \Delta f(s)]\,ds \\
&\quad + \int_{t_1}^{t_2} [\mu(t)\cdot \bar{q}_u(t) + \nu(t)\cdot \bar{p}_u(t)]\delta u(t)\,dt + \int_{t_1}^{t_2} \nu(t)\cdot \bar{p}_x(t)\Phi(t)\,dt\xi \\
(12.9) \qquad &\quad + \int_{t_1}^{t_2} \nu(t)\cdot \bar{p}_x(t)\Phi(t)\int_{t_1}^{t} \Phi^{-1}(s)[\bar{f}_u(s)\delta u(s) + \Delta f(s)]\,ds\,dt \\
&\quad + \sum_{j=1}^{l_1}\int_{t_1}^{t_2} \bar{\chi}_x^j(\bar{x}(t), t)\Phi(t)\,d\lambda^j(t)\xi \\
&\quad + \sum_{j=1}^{l_1}\int_{t_1}^{t_2} \bar{\chi}_x^j(\bar{x}(t), t)\Phi(t)\int_{t_1}^{t} \Phi^{-1}(s)[\bar{f}_u(s)\delta u(s) + \Delta f(s)]\,ds\,d\lambda^j(t) \leqq 0 \\
&\qquad\qquad \text{for all } \xi \in R^n, \quad \delta u \in \Delta_0, \quad \text{and} \quad \Delta f \in M.
\end{aligned}$$

Let us define the n-vector-valued functions $\psi_i, i = 1, 2, 3$, and $\hat{\psi}$ on I as follows:

$$(12.10) \qquad \psi_1(s) = \sum_{j=i+1}^{\sigma} \beta \cdot \bar{\chi}_{x_i}\Phi(\tau_i)\Phi^{-1}(s)$$

$$\text{for } \tau_i \leqq s < \tau_{i+1} \quad \text{and} \quad i = 1, \cdots, \sigma - 1,$$

$$(12.11) \qquad \psi_1(t_2) = \beta \cdot \bar{\chi}_{x_0},$$

$$(12.12) \qquad \psi_2(s) = \int_{s}^{t_2} \nu(t)\cdot \bar{p}_x(t)\Phi(t)\,dt\Phi^{-1}(s), \qquad s \in I,$$

$$(12.13) \qquad \psi_3(s) = \sum_{j=1}^{l_1}\int_{s}^{t_2} \bar{\chi}_x^j(\bar{x}(t), t)\Phi(t)\,d\lambda^j(t)\Phi^{-1}(s), \quad s \in I,$$

$$(12.14) \qquad \hat{\psi} = \sum_{i=1}^{3} \psi_i.$$

If we interchange the order of integration in the double integrals of (12.9), we easily conclude that this inequality holds if and only if (recall that $\tau_1 = t_1$)

$$(12.15) \qquad \hat{\psi}(t_1) = -\beta \cdot \bar{\chi}_{x_1},$$

$$(12.16) \qquad \int_{t_1}^{t_2} [\hat{\psi}(s)\cdot \bar{f}_u(s) + \mu(s)\cdot \bar{q}_u(s) + \nu(s)\cdot \bar{p}_u(s)]\delta u(s)\,ds \leqq 0 \quad \text{for all } \delta u \in \Delta_0,$$

$$(12.17) \qquad \int_{t_1}^{t_2} \hat{\psi}(s) f(\bar{x}(s), u(s), s)\,ds \leqq \int_{t_1}^{t_2} \hat{\psi}(s) f(\bar{x}(s), \bar{u}(s), s)\,ds \quad \text{for all } u \in \hat{\mathscr{U}}.$$

Integrating by parts in (12.13), and taking into account that $\lambda(t_2) = 0$, we obtain that

$$\psi_3(s) + \lambda(s) \cdot \tilde{\chi}_x(\bar{x}(s), s) = \psi_4(s), \tag{12.18}$$

where

$$\psi_4(s) = -\int_s^{t_2} \lambda(t) \cdot \frac{d}{dt}[\tilde{\chi}_x(\bar{x}(t), t)\Phi(t)]\, dt \Phi^{-1}(s). \tag{12.19}$$

Further, if we differentiate the identity $\Phi(t)\Phi^{-1}(t) =$ the identity matrix, and take into account (12.8), we obtain

$$\begin{gathered} \frac{d}{dt}(\Phi^{-1}(t)) = -\Phi^{-1}(t)\bar{f}_x(t) \quad \text{a.e. on } I, \\ \Phi^{-1}(t_1) = \text{the identity matrix.} \end{gathered} \tag{12.20}$$

If we set $\psi = \psi_1 + \psi_2 + \psi_4$, and take into account (12.8), (12.10)–(12.12), (12.19), (12.20), and the fact that $\bar{x}$ and $\bar{u}$ satisfy equation (2.2) a.e. on I, we quickly conclude that, on each of the intervals $[t_1, \tau_2), [\tau_2, \tau_3), \cdots, [\tau_{\sigma-2}, \tau_{\sigma-1}), [\tau_{\sigma-1}, t_2]$, ψ is absolutely continuous and satisfies the linear inhomogeneous ordinary differential equation

$$\begin{aligned} \dot{\psi}(t) = {} & -\psi(t)f_x(\bar{x}(t), \bar{u}(t), t) + \lambda(t)\tilde{p}_x(\bar{x}(t), \bar{u}(t), t) \\ & - v(t)p_x(\bar{x}(t), \bar{u}(t), t) \quad \text{almost everywhere,} \end{aligned} \tag{12.21}$$

where $\tilde{p}$ is the function from $G \times U_0 \times I \to R^{l_1}$ defined by

$$\tilde{p}(x, u, t) = \tilde{\chi}_x(x, t)f(x, u, t) + \tilde{\chi}_t(x, t), \qquad x \in G, \quad u \in U_0, \quad t \in I \tag{12.22}$$

(in (12.21), ψ, λ, and v are to be regarded as row vectors), and that, in addition (see (12.14), (12.15), and (12.18)),

$$\psi(t_1) = -\beta \cdot \bar{\chi}_{x_1} + \lambda(t_1)\tilde{\chi}_x(\bar{x}(t_1), t_1), \tag{12.23}$$

$$\psi(t_2) = \beta \cdot \bar{\chi}_{x_\sigma}, \tag{12.24}$$

$$\psi(\tau_i) - \psi(\tau_i^-) = -\beta \cdot \bar{\chi}_{x_i} \quad \text{for } i = 2, \cdots, \sigma - 1. \tag{12.25}$$

Further, (12.17) and (12.16) take the forms

$$\begin{aligned} & \int_{t_1}^{t_2} [\psi(t) - \lambda(t)\tilde{\chi}_x(\bar{x}(t), t)]f(\bar{x}(t), u(t), t)\, dt \\ & \qquad \leqq \int_{t_1}^{t_2} [\psi(t) - \lambda(t)\tilde{\chi}_x(\bar{x}(t), t)]f(\bar{x}(t), \bar{u}(t), t)\, dt \quad \text{for all } u \in \hat{\mathscr{U}}, \end{aligned} \tag{12.26}$$

$$\begin{aligned} & \int_{I_1} \{[\psi(t) - \lambda(t)\tilde{\chi}_x(\bar{x}(t), t)]\bar{f}_u(t) + \mu(t)\bar{q}_u(t) + v(t)\bar{p}_u(t)\}\delta u(t)\, dt \\ & \qquad \leqq 0 \quad \text{for all } \delta u \in \Delta_0. \end{aligned} \tag{12.27}$$

But (12.27) can hold only if

$$\begin{aligned} & [\psi(t) - \lambda(t)\bar{\chi}_x(\bar{x}(t), t)]f_u(\bar{x}(t), \bar{u}(t), t) + \mu(t)q_u(\bar{u}(t), t) \\ & \qquad + v(t)p_u(\bar{x}(t), \bar{u}(t), t) = 0 \quad \text{a.e. on } I_1. \end{aligned} \tag{12.28}$$

Let us show that β and λ cannot both vanish. Indeed, suppose that λ and β are both zero. Then (see (12.10)–(12.13) and (12.18)) $\psi_1 \equiv \psi_3 \equiv \psi_4 \equiv 0$, and

$$\psi(t) - \lambda(t)\tilde{\chi}_x(\bar{x}(t), t) = \psi(t) = \psi_2(t) = \int_t^{t_2} v(s)\bar{p}_x(s)\Phi(s)\,ds\Phi^{-1}(t), \quad t \in I. \tag{12.29}$$

Since $\bar{u}$ satisfies C2, there is a function $\delta u_1 \in \mathcal{Q}_0(\bar{u})$ such that

$$\bar{p}_u(t)\delta u_1(t) = \begin{cases} v(t)/|v(t)| & \text{if } v(t) \neq 0, \quad t \in I_1, \\ 0 & \text{if } v(t) = 0, \quad t \in I_1. \end{cases} \tag{12.30}$$

Hence, (12.28)–(12.30) imply that

$$\begin{aligned} |v(t)| = &- \int_t^{t_2} v(s)\bar{p}_x(s)\Phi(s)\,ds\Phi^{-1}(t)\bar{f}_u(t)\delta u_1(t) \\ &- \mu(t)\bar{q}_u(t)\delta u_1(t) \quad \text{a.e. on } I. \end{aligned} \tag{12.31}$$

Since the functions $\bar{p}_x$, Φ, Φ^{-1}, $\bar{f}_u$, and δu_1 are all essentially bounded on I, it follows from (12.31), the properties of μ and the definition of $\mathcal{Q}_0(\bar{u})$ that, for some $\gamma > 0$,

$$|v(t)| \leqq \int_t^{t_2} \gamma |v(s)|\,ds \quad \text{a.e. on } I,$$

which is possible only if $v = 0$ a.e. on I. But then (12.31) implies that $\mu(t) = 0$ a.e. on I_1, i.e., β, λ, μ, and v all vanish, which is a contradiction.

Thus, we have shown the following theorem.

THEOREM 12.1. *Let $(\bar{x}, \bar{u})$ be a solution of our optimal control problem, subject to Assumptions* A1–A6 *on the problem data, and suppose that $\bar{u}$ satisfies Conditions* C1 *and* C2. *Then there exist a vector $\beta = (\beta^0, \beta^1, \cdots, \beta^{k+k_1}) \in R^{k+k_1+1}$ and functions $\lambda = (\lambda^1, \cdots, \lambda^{l_1}) : I \to R^{l_1}$, $\mu = (\mu^1 \cdots, \mu^r) \in L^r_\infty$, $v = (v^1, \cdots, v^l) \in L^l_\infty$, and $\psi : I \to R^n$, all of them to be considered row-vector-valued, such that*

(i) $|\beta| + |\lambda(t_1)| > 0$;

(ii) *on each of the intervals $[t_1, \tau_2), \cdots, [\tau_{\sigma-2}, \tau_{\sigma-1}), [\tau_{\sigma-1}, t_2]$, ψ is absolutely continuous and satisfies* (12.21), *where $\tilde{p}$ is defined by* (12.22);

(iii) *ψ satisfies the boundary (transversality) conditions* (12.23) *and* (12.24) *as well as the "jump" conditions* (12.25), *where $\bar{\chi}_{x_i} = \chi_{x_i}(\bar{x}(\tau_1), \cdots, \bar{x}(\tau_\sigma))$ for each i*;

(iv) *$\bar{u}$ satisfies the maximum condition* (12.26), *where $\hat{\mathcal{U}}$ consists of all functions $u \in \mathcal{U}$ such that $p(\bar{x}(t), u(t), t) = 0$ a.e. on I_1 and which satisfy Conditions* C1 *and* C2;

(v) *relation* (12.28) *holds*;

(vi) *for each $j = 1, \cdots, l_1$, λ^j is nonincreasing on I and continuous from the right on (t_1, t_2), satisfies $\lambda^j(t_2) = 0$, and is constant on any subinterval of I which does not meet the set* (12.6);

(vii) $|\mu(t)| = |v(t)| = 0$ *for almost all $t \in I \setminus I_1$*;

(viii) *for each $j = 1, \cdots, r$ and almost all $t \in I_1$, $\mu^j(t) \leqq 0$ and $\mu^j(t)q^j(\bar{u}(t), t) = 0$*;

(ix) *β satisfies* (12.4) *and* (12.5).

13. The pointwise maximum principle. We shall now show that if we strengthen Assumptions A3–A5 by requiring that the functions f, p, q, and p_u be continuous, then the maximum principle (12.26) in Theorem 12.1 can also be written in a pointwise, rather than in an integral, form.

For each $x \in G$ and $t \in I_1$, let $\omega(x, t)$ denote the set of all $v \in U_0$ such that

(i) $q^j(v, t) \leqq 0$ for $j = 1, \cdots, r$,

(ii) $p(x, v, t) = 0$,

(iii) $p_u(x, v, t)[p_u(x, v, t)]^T$ is nonsingular,

(iv) $\{p_u(x, v, t)\delta v : \delta v \in R^m,\ q_u^j(v, t)\delta v < 0$ for each $j = 1, \cdots, r$ such that $q^j(v, t) = 0\} = R^l$.

Note that $\bar{u}(t) \in \omega(\bar{x}(t), t)$ for almost all $t \in I_1$ because $\bar{u}$ satisfies C1 and C2.

If e is any function from I into R, then $t_0 \in I$ will be called a *regular* (or *Lebesgue*) point for e if, for every neighborhood O of $e(t_0)$,

$$\lim_{\text{meas } J \to 0} \frac{\text{meas}\,(J \cap e^{-1}(O))}{\text{meas } J} = 1,$$

where J is an arbitrary subinterval of I such that $t_0 \in J$. (A regular (or Lebesgue) point for a function $u : I \to R^m$ is similarly defined.) If $e : I \to R$ (resp., $u : I \to R^m$) is measurable, then almost every point of I is a regular point for e (resp., u) (see [11, *pp.* 255–256]).

Now, for each $t_0 \in I$, let $U(t_0)$ denote the set of all $v \in U_0$ for which there exists a function $u \in \mathscr{U}$ such that (a) $u(t_0) = v$ and (b) t_0 is a regular point for u.

Note that, if $u \in \mathscr{U}$, then $u(t) \in U(t)$ for almost all $t \in I$.

If the function q is independent of $t \in I$, then evidently

$$U(t) = \{v : v \in U_0, q^j(v) \leqq 0 \text{ for } j = 1, \cdots, r\} \quad \text{for every } t \in I.$$

More generally, it is not difficult to verify that, if q is continuously differentiable, then, for every $t \in I$,

$$\begin{aligned} U(t) \supset \{v : v \in U_0, q^j(v, t) \leqq 0 \text{ for } j = 1, \cdots, r, \text{ there exists a} \\ \delta v \in R^m \text{ such that } q_u^j(v, t)\delta v < 0 \text{ for all} \\ j = 1, \cdots, r \text{ for which } q^j(v, t) = 0\} \\ \supset \{v : v \in U_0, q^j(v, t) \leqq 0 \text{ for } j = 1, \cdots, r, \text{ the vectors } q_u^j(v, t), \\ \text{for those } j = 1, \cdots, r \text{ for which } q^j(v, t) = 0, \text{ are linearly independent}\}. \end{aligned}$$

THEOREM 13.1. *Suppose that the hypotheses of Theorem* 12.1 *hold, and that, in addition, the functions f, p, q, and p_u are continuous. Then the conclusion in Theorem* 12.1 *that* (12.26) *holds can be replaced by the conclusions that*

$$\hat{\psi}(t)f(\bar{x}(t), \bar{u}(t), t) = \max_{v \in \omega(\bar{x}(t), t)} \hat{\psi}(t)f(\bar{x}(t), v, t) \quad \textit{for almost all } t \in I_1, \tag{13.1}$$

$$\hat{\psi}(t)f(\bar{x}(t), \bar{u}(t), t) = \max_{v \in U(t)} \hat{\psi}(t)f(\bar{x}(t), v, t) \quad \textit{for almost all } t \in I \setminus I_1, \tag{13.2}$$

where

$$\hat{\psi}(t) = \psi(t) - \lambda(t)\chi_x(\bar{x}(t), t), \qquad t \in I.$$

Proof. We first point out that, since each component of λ is monotone, λ has at most a countable number of points of discontinuity. Since ψ is continuous at all $t \in I$ except possibly $\tau_2, \cdots, \tau_{\sigma-1}$, and $\tilde{\chi}_x$ and $\bar{x}$ are continuous, the set of points of discontinuity of $\hat{\psi}$ form a set of measure zero.

Let us show that (13.1) holds. We argue by contradiction and thus suppose that, for some $t_0 \in I_1$ which is a point of continuity of $\hat{\psi}$ as well as a regular point for $\bar{u}$ and for e_{I_1} (the characteristic function of I_1) and for some $v_0 \in \omega(\bar{x}(t_0), t_0)$,

$$\hat{\psi}(t_0)f(\bar{x}(t_0), v_0, t_0) > \hat{\psi}(t_0)f(\bar{x}(t_0), \bar{u}(t_0), t_0). \tag{13.3}$$

Consequently, $v_0 \in U_0$ and

$$\begin{aligned} &q^j(v_0, t_0) \leqq 0 \quad \text{for } j = 1, \cdots, r, \\ &p(\bar{x}(t_0), v_0, t_0) = 0, \end{aligned} \tag{13.4}$$

$$p_u(\bar{x}(t_0), v_0, t_0) \cdot [p_u(\bar{x}(t_0), v_0, t_0)]^T \quad \text{is nonsingular,} \tag{13.5}$$

and there is a vector $\delta v_0 \in R^m$ such that

$$p_u(\bar{x}(t_0), v_0, t_0)\delta v_0 = 0, \tag{13.6}$$

$$q_u^j(v_0, t_0)\delta v_0 < 0 \quad \text{for each } j = 1, \cdots, r \quad \text{such that } q^j(v_0, t_0) = 0. \tag{13.7}$$

First let us suppose that $q^j(v_0, t_0) = 0$ for some $j = 1, \cdots, r$. Then we must have that $m > l$ and that $\delta v_0 \neq 0$. If $m = l + 1$, let us consider the system of $l + 1 = m$ equations (for $v \in U_0$, with θ a real parameter)

$$\begin{aligned} &p^i(\bar{x}(t_0), v, t_0) = 0, \qquad i = 1, \cdots, l, \\ &(v - v_0) \cdot \delta v_0 - \theta = 0, \end{aligned} \tag{13.8}$$

which, for $\theta = 0$, have the solution $v = v_0$. Because the vectors $p_u^i(\bar{x}(t_0), v_0, t_0)$, $i = 1, \cdots, l$, and δv_0 are linearly independent in R^m (recall that $p_u(\bar{x}(t_0), v_0, t_0)$ has rank l, and see (13.6)), we may appeal to the conventional implicit functions theorem (see, e.g., [12, Thms. (9.3) and (9.8), pp. 122–125]), and conclude that there is a continuously differentiable function $\hat{v}$ from some neighborhood N of $0 \in R$ into U_0 such that $\hat{v}(0) = v_0$, and such that $0 = p^i(\bar{x}(t_0), \hat{v}(\theta), t_0) = (\hat{v}(\theta) - v_0) \cdot \delta v_0 - \theta = 0$ for all $\theta \in N$ and each $i = 1, \cdots, l$. Differentiating the last set of equalities with respect to θ, and setting $\theta = 0$, we conclude that $p_u(\bar{x}(t_0), v_0, t_0)\hat{v}'(0) = 0$ and $\hat{v}'(0) \cdot \delta v_0 = 1$, which means—since $p_u^i(\bar{x}(t_0), v_0, t_0)$, for $i = 1, \cdots, l$, and δv_0 span R^m, and (13.6) holds—that $\hat{v}'(0) = \alpha_0 \delta v_0$ for some $\alpha_0 > 0$.

If $m > l + 1$, then we choose any $m - l - 1$ vectors $\delta v_1, \cdots, \delta v_{m-l-1}$ which span the orthogonal complement in R^m of the $(l + 1)$-dimensional linear manifold spanned by δv_0 and the vectors $p_u^i(\bar{x}(t_0), v_0, t_0)$, $i = 1, \cdots, l$, and adjoin to (13.8) the $m - l - 1$ equations

$$(v - v_0) \cdot \delta v_i = 0 \quad \text{for } i = 1, \cdots, m - l - 1.$$

In the same way, we can then conclude that there is a continuously differentiable function $\hat{v}$ from some neighborhood N of $0 \in R$ into U_0 satisfying the relations $\hat{v}(0) = v_0$, $p(\bar{x}(t_0), \hat{v}(\theta), t_0) = 0$ for all $\theta \in N$, and $\hat{v}'(0) = \alpha_0 \delta v_0$ for some $\alpha_0 > 0$.

Consequently, by virtue of (13.3)–(13.5) and (13.7) and the continuity of f and p_u, we can conclude that there is a $\theta_1 > 0$ such that $v_1 = \hat{v}(\theta_1) \in U_0$ has the following properties:

$$q^j(v_1, t_0) < 0 \quad \text{for each } j = 1, \cdots, r, \tag{13.9}$$

$$p(\bar{x}(t_0), v_1, t_0) = 0, \tag{13.10}$$

$$\hat{\psi}(t_0) f(\bar{x}(t_0), v_1, t_0) > \hat{\psi}(t_0) f(\bar{x}(t_0), \bar{u}(t_0), t_0), \tag{13.11}$$

$$p_u(\bar{x}(t_0), v_1, t_0) \cdot [p_u(\bar{x}(t_0), v_1, t_0)]^T \quad \text{is nonsingular.} \tag{13.12}$$

Note that (13.12) implies that $p_u(\bar{x}(t_0), v_1, t_0)$ has maximum rank, i.e., rank l.

If $q^j(v_0, t_0) < 0$ for every $j = 1, \cdots, r$, we simply set $v_1 = v_0$, in which case (13.9)–(13.12) are immediately seen to hold.

Because the functions p, q, p_u, and $\bar{x}$ are continuous in all of their arguments, we may again appeal to the implicit functions theorem, and conclude that there are a negative number $\hat{\theta}$, a closed interval I_0 containing t_0, and a continuous function $\hat{u}$ from $I_0 \cap I$ into a compact subset of U_0 such that (i) $\hat{u}(t_0) = \text{v}_1$, (ii)

$$p(\bar{x}(t), \hat{u}(t), t) = 0 \quad \text{for all } t \in I_0 \cap I, \tag{13.13}$$

and (see (13.9)) (iii)

$$q^j(\hat{u}(t), t) < \hat{\theta} < 0 \quad \text{for all } t \in I_0 \cap I \quad \text{and each } j = 1, \cdots, r. \tag{13.14}$$

Further, because of the continuity of p_u, $\bar{x}$, and $\hat{u}$, and because the matrix

$$p_u(\bar{x}(t), \hat{u}(t), t) \cdot [p_u(\bar{x}(t), \hat{u}(t), t)]^T \tag{13.15}$$

is nonsingular for $t = t_0$ (see (13.12)), we may suppose that this matrix is nonsingular for all $t \in I_0 \cap I$, and that the elements of its inverse remain in some bounded set as t ranges over $I_0 \cap I$.

For any function $z \in L^l_\infty(I_1)$, let us define the function $\delta v_z \in L^m_\infty(I_0 \cap I_1)$ as follows:

$$\delta v_z(t) = [p_u(\bar{x}(t), \hat{u}(t), t)]^T \{p_u(\bar{x}(t), \hat{u}(t), t) \cdot [p_u(\bar{x}(t), \hat{u}(t), t)]^T\}^{-1} z(t), \quad t \in I_0 \cap I_1.$$

Evidently, then

$$p_u(\bar{x}(t), \hat{u}(t), t) \delta v_z(t) = z(t) \quad \text{for all } t \in I_0 \cap I_1.$$

Finally, by (13.11) and the continuity of the function f and of the functions $\hat{u}$, $\bar{x}$, and $\hat{\psi}$ at t_0, and because t_0 is a regular point for both $\bar{u}$ and e_{I_1}, we can conclude that, for some subset J_0 of $I_0 \cap I_1$ of positive measure,

$$\hat{\psi}(t) f(\bar{x}(t), \hat{u}(t), t) > \hat{\psi}(t) f(\bar{x}(t), \bar{u}(t), t) \quad \text{for all } t \in J_0. \tag{13.16}$$

If we now define the function $\tilde{u} \in L^m_\infty$ by the relation

$$\tilde{u}(t) = \begin{cases} \hat{u}(t) & \text{for } t \in J_0, \\ \bar{u}(t) & \text{for } t \in I \diagdown J_0, \end{cases} \tag{13.17}$$

then it follows directly from what we have shown (see (13.13) and (13.14)) that $\tilde{u} \in \tilde{\mathscr{U}}$, and that $\tilde{u}$ satisfies Conditions C1 and C2, i.e., that $\tilde{u} \in \hat{\mathscr{U}}$. But (13.16) and (13.17) imply that

$$\int_{t_1}^{t_2} \hat{\psi}(t) f(\bar{x}(t), \tilde{u}(t), t)\, dt > \int_{t_1}^{t_2} \hat{\psi}(t) f(\bar{x}(t), \bar{u}(t), t)\, dt, \tag{13.18}$$

which then contradicts (12.26). Consequently, (13.1) holds.

To prove that (13.2) holds, we again argue by contradiction, and thus suppose that there are a point $t_0 \in I$ and a function $\hat{u} \in \mathscr{U}$ such that (i) t_0 is a regular point for $\hat{u}, \bar{u}$ and $e_{I \setminus I_1}$ (the characteristic function of $I \setminus I_1$) as well as a point of continuity of $\hat{\psi}$, and (ii)

$$\hat{\psi}(t_0)f(\bar{x}(t_0), \hat{u}(t_0), t_0) > \hat{\psi}(t_0)f(\bar{x}(t_0), \bar{u}(t_0), t_0).$$

Hence, there is a subset J_0 of $I \setminus I_1$ of positive measure such that (13.16) holds. Defining $\tilde{u}$ by (13.17), we immediately obtain that $\tilde{u} \in \hat{\mathscr{U}}$ and that (13.18) holds, again contradicting (12.26). Thus (13.2) holds. The proof is complete.

Recall that $\bar{u}(t) \in \omega(\bar{x}(t), t)$ for almost all $t \in I_1$, and note that (for each $(x, t) \in G \times I_1$) the set of all $v \in U_0$ that satisfy conditions (iii) and (iv) in the definition of $\omega(x, t)$ is open. Hence, (13.1) may be interpreted as follows: For almost every $t \in I_1$, there is a neighborhood O of $\bar{u}(t)$ such that $\bar{u}(t)$ solves the problem of maximizing the function $v \to \hat{\psi}(t)f(\bar{x}(t), v, t): O \to R$ subject to the constraints $q^j(v, t) \leqq 0$ for $j = 1, \cdots, r$ and $p(\bar{x}(t), v, t) = 0$. Thus, conclusions (v) and (viii) of Theorem 12.1 show that, for almost all $t \in I_1$, $\mu^1(t), \cdots, \mu^r(t), v^1(t), \cdots, v^l(t)$ may be interpreted as Lagrange multipliers for this finite-dimensional constrained maximization problem. If, for some $t \in I_1$ for which (12.28) holds, the vectors $\bar{q}^i_u(t), i = 1, \cdots, r$, and $\bar{p}^i_u(t), i = 1, \cdots, l$, are linearly independent, then (12.28) uniquely determines $\mu(t)$ and $v(t)$. If, in addition, $\hat{\psi}$ and $\bar{u}$ are continuous at t, f_u, p_u, and q_u are all continuous functions, and $t \in \operatorname{int} I_1$, then μ and v are also continuous at t.

The requirement in Theorem 13.1 that the functions f, p, q, and p_u are continuous can be relaxed. For example, the theorem remains in force if there is a finite subset I_f of I such that f, p, q, and p_u are continuous at all $(x, u, t) \in G \times U_0 \times (I \setminus I_f)$. (Of course, A3–A6 must still hold.)

14. Some generalizations. The problem wherein the equality constraints (2.5) are replaced by the more general constraints

$$p^i(x(t), u(t), t) = 0 \quad \text{for almost all } t \in I_i, \quad i = 1, \cdots, l, \tag{14.1}$$

where $I_1, \cdots, I_l$ are given Lebesgue measurable subsets of I, is also of interest, and can, by a straightforward reformulation, be included in the problem studied in the earlier sections of this paper.

Let us indicate how to do this. For simplicity, let us suppose that $l = 2$. Let $\mathbf{I}_1 = I_1 \cup I_2$, and let us adjoin to $u = (u^1, \cdots, u^m)$ two new scalar "control" variables u^{m+1}, u^{m+2}. Let $\mathbf{u} = (u, u^{m+1}, u^{m+2})$, let $\mathbf{U}_0 = U_0 \times R^2$, and define the functions $\mathbf{f}: G \times \mathbf{U}_0 \times I \to R^n$, $\mathbf{p}: G \times \mathbf{U}_0 \times I \to R^2$, and $\mathbf{q}: U_0 \times I \to R^r$ as follows:

$$\begin{aligned} &\mathbf{f}(x, \mathbf{u}, t) = f(x, u, t) \quad \text{and} \quad \mathbf{q}(\mathbf{u}, t) = q(u, t) \quad \text{for all } (x, \mathbf{u}, t), \\ &\mathbf{p}^j(x, \mathbf{u}, t) = \begin{cases} p^j(x, u, t) & \text{for all } (x, \mathbf{u}) \text{ and all } t \in I_j \cup (I \setminus \mathbf{I}_1), \\ u^{m+j} & \text{for all } (x, \mathbf{u}) \text{ and all } t \in \mathbf{I}_1 \setminus I_j \text{ for } j = 1, 2. \end{cases} \end{aligned} \tag{14.2}$$

It is easily seen that the generalized problem reduces to our original problem with I_1, u, U_0, f, p, and q replaced by $\mathbf{I}_1$, etc., and that if f, p, and q satisfy A3–A6, then $\mathbf{f}$, $\mathbf{p}$, and $\mathbf{q}$ also do. Further, if $u \in \mathscr{U}$ satisfies Conditions C1 and C2, modified in an evident manner, then $(u, 0, 0)$ also satisfies these conditions with p replaced by $\mathbf{p}$.

Appealing to Theorem 12.1, as applied to the "expanded" problem, we can easily convince ourselves that the necessary conditions of this theorem carry over to the problem where (2.5) is replaced by (14.1)—provided that we replace I_1 by $\bigcup_{j=1}^{l} I_j$, redefine $\hat{\mathscr{U}}$ in an evident manner in conclusion (iv), and add the conclusion that $v^j(t) = 0$ for almost all $t \in I \diagdown I_j$, for each $j = 1, \cdots, l$.

As another generalization, suppose that the equality constraints (2.5) are replaced by equality *and inequality* constraints of the form

$$\text{(iva)} \quad p^i(x(t), u(t), t) \leqq 0 \quad \text{for almost all } t \in I_1 \text{ and each } i = 1, \cdots, l', \tag{14.3}$$

$$\begin{gathered}\text{(ivb)} \quad p^i(x(t), u(t), t) = 0 \quad \text{for almost all } t \in I_1 \text{ and} \\ \text{each } i = l' + 1, \cdots, l,\end{gathered} \tag{14.4}$$

where l' is some positive integer with $l' \leqq l$ (and we replace the restriction that $l \leqq m$ in the problem statement by the inequality $l - l' \leqq m$).

We shall show that the problem in which (2.5) is replaced by (14.3) and (14.4) can, by a simple reformulation, also be included in the problem which we investigated in the preceding sections.

We first introduce some notation. Let $p_1 = (p^1, \cdots, p^{l'})$ and $p_2 = (p^{l'+1}, \cdots, p^l)$, and define $P_1 = (P^1, \cdots, P^{l'})$, $P_2 = (P^{l'+1}, \cdots, P^l)$, where the P^i are defined by (2.7). If $(\bar{x}, \bar{u})$ is the solution of the problem for which we wish to obtain necessary conditions, let

$$\begin{aligned}\tilde{\tilde{\mathscr{U}}} = \{u : u \in \mathscr{U},\ & p^i(\bar{x}(t), u(t), t) \leqq 0 \text{ for almost all } t \in I_1 \\ & \text{and each } i = 1, \cdots, l',\ p^i(\bar{x}(t), u(t), t) = 0 \text{ for almost all } t \in I_1 \\ & \qquad\qquad \text{and each } i = l' + 1, \cdots, l\},\end{aligned} \tag{14.5}$$

and, for each $u \in \tilde{\tilde{\mathscr{U}}}$, let

$$\begin{aligned}\tilde{\mathscr{P}}(u) = \{\delta u : \delta u \in L_\infty^m, & \operatorname*{ess\,sup}_{t \in I_1} [p^i(\bar{x}(t), u(t), t) \\ & + p_u^i(\bar{x}(t), u(t), t)\delta u(t)] < 0 \text{ for } i = 1, \cdots, l'\},\end{aligned} \tag{14.6}$$

$$\mathscr{P}(u) = \{\alpha\delta u : \alpha > 0, \delta u \in \tilde{\mathscr{P}}(u)\}. \tag{14.7}$$

Corresponding to Conditions C1 and C2 we introduce the following conditions:

Condition C1′. For almost all $t \in I_1$, the $(l - l') \times (l - l')$ matrix $p_{2_u}(\bar{x}(t), u(t), t) \cdot [p_{2_u}(\bar{x}(t), u(t), t)]^T$ is nonsingular, and the function $t \to [p_{2_u}(\bar{x}(t), u(t), t) \cdot (p_{2_u} \cdot (\bar{x}(t), u(t), t))^T]^{-1} : I_1 \to R^{(l-l')^2}$ is in $L_\infty^{(l-l')^2}(I_1)$.

Condition C2′. $\{D_2P_2(\bar{x}, u; \delta u) : \delta u \in (\mathscr{Q}_0(u) \cap \mathscr{P}(u))\} \twoheadrightarrow L_\infty^{l-l'}(I_1)$, where $\mathscr{Q}_0(u)$ is given by (3.7), (3.8), and (3.10).

Conditions C1′ and C2′ should (as were Conditions C1 and C2) be interpreted as regularity conditions, or as compatibility conditions—to "first order" in u—for the constraints (14.3), (14.4), and (2.1).

Now let us reformulate our problem. We adjoin to $u = (u^1, \cdots, u^m)$, l' new scalar "control" variables $u^{m+1}, \cdots, u^{m+l'}$. Let $\mathbf{u} = (u, u^{m+1}, \cdots, u^{m+l'})$, let $\mathbf{U}_0 = U_0 \times R^{l'}$, and let the functions $\mathbf{f} : G \times \mathbf{U}_0 \times I \to R^n$, $\mathbf{p} : G \times \mathbf{U}_0 \times I \to R^l$, and $\mathbf{q}:\mathbf{U}_0 \times I \to R^{r+l'}$ be defined as follows:

$$\begin{aligned}
\mathbf{f}(x, \mathbf{u}, t) &= f(x, u, t), && \\
\mathbf{p}^i(x, \mathbf{u}, t) &= p^i(x, u, t) + u^{m+i} && \text{for } i = 1, \cdots, l', \\
\mathbf{p}^i(x, \mathbf{u}, t) &= p^i(x, u, t) && \text{for } i = l' + 1, \cdots, l, \\
\mathbf{q}^i(\mathbf{u}, t) &= q^i(u, t) && \text{for } i = 1, \cdots, r, \\
\mathbf{q}^{r+i}(\mathbf{u}, t) &= -u^{m+i} && \text{for } i = 1, \cdots, l',
\end{aligned}$$

for all $(x, \mathbf{u}, t) \in G \times \mathbf{U}_0 \times I$. It is then easy to see that the problem in which (2.5) is replaced by (14.3) and (14.4) is equivalent to the original problem, provided that we therein replace m by $(m + l')$, r by $(r + l')$, and u, U_0, f, p, and q by $\mathbf{u}$, $\mathbf{U}_0$, etc. Further, if f, p, q, and U_0 satisfy A3–A6, then $\mathbf{f}$, $\mathbf{p}$, $\mathbf{q}$, and $\mathbf{U}_0$ also do.

Let us consider the set $\boldsymbol{\hat{\mathscr{U}}}$ defined in terms of $\mathbf{p}$, $\mathbf{q}$, etc. in the same way that $\hat{\mathscr{U}}$ was defined in terms of p, q, etc. Then one may convince oneself without too much difficulty that $\mathbf{u} = (u, u^{m+1}, \cdots, u^{m+l'}) \in \boldsymbol{\hat{\mathscr{U}}}$ if and only if (i) $u \in \tilde{\tilde{\mathscr{U}}}$, (ii) for each $i = 1, \cdots, l'$, $u^{m+i} \in L_\infty$, $u^{m+i}(t) \geqq 0$ for almost all $t \in I \setminus I_1$ and $u^{m+i}(t) + p^i \cdot (\bar{x}(t), u(t), t) = 0$ a.e. in I_1, and (iii) u satisfies Conditions C1′ and C2′. If $l = l'$, so that the constraints (14.4) are absent, then (iii) is to be replaced by: (iii′) $\tilde{\mathscr{Q}}(u) \cap \tilde{\mathscr{P}}(u) \neq \varnothing$.

Appealing to Theorem 12.1 for the reformulated problem, we can quickly obtain the following theorem.

THEOREM 14.1. *Let* $(\bar{x}, \bar{u})$ *be a solution of our optimal control problem with* (2.5) *replaced by* (14.3) *and* (14.4), *where* $1 \leqq l' \leqq l$, *and suppose that Assumptions* A1–A6 *hold. Further, suppose that* $\bar{u}$ *satisfies Conditions* C1′ *and* C2′ *if* $l' < l$, *and that* $\tilde{\mathscr{Q}}(u) \cap \tilde{\mathscr{P}}(u) \neq \varnothing$ *if* $l' = l$. *Then there exist a vector* $\beta = (\beta^0, \beta^1, \cdots, \beta^{k+k_1}) \in R^{k+k_1+1}$ *and functions* $\lambda = (\lambda^1, \cdots, \lambda^{l_1}): I \to R^{l_1}$, $\mu = (\mu^1, \cdots, \mu^r) \in L^r_\infty$, $v = (v^1, \cdots, v^l) \in L^l_\infty$, *and* $\psi: I \to R^n$, *all of them to be considered row vectors, such that conclusions* (i)–(ix) *of Theorem* 12.1 *hold, except that, in* (iv) (*i.e., in* (12.26)) $\hat{\mathscr{U}}$ *is to be the set of all* $u \in \tilde{\tilde{\mathscr{U}}}$ *which satisfy Conditions* C1′ *and* C2′ *if* $l < l'$ *and the set of all* $u \in \tilde{\tilde{\mathscr{U}}}$ *such that* $\tilde{\mathscr{P}}(u) \cap \tilde{\mathscr{Q}}(u) \neq \varnothing$ *if* $l = l'$ (*see* (14.5) *and* (14.6)), *and, in addition,* (x) *for each* $j = 1, \cdots, l'$ *and almost all* $t \in I_1$, $v^j(t) \leqq 0$ *and* $v^j(t)p^j(\bar{x}(t), \bar{u}(t), t) = 0$.

In order to obtain here a pointwise maximum principle which is analogous to Theorem 13.1, we introduce the set $\omega_1(x, t)$. Indeed, for each $x \in G$ and $t \in I_1$, let $\omega_1(x, t)$ denote the set of all $v \in U_0$ such that

(i) $q^j(v, t) \leqq 0$ for $j = 1, \cdots, r$;

(14.8) (ii) $p^j(x, v, t) \leqq 0$ for $j = 1, \cdots, l'$;

(14.9) (iii) $p^j(x, v, t) = 0$ for $j = l' + 1, \cdots, l$;

(iv) $p_{2_u}(x, v, t) \cdot (p_{2_u}(x, v, t))^T$ is nonsingular;

(va) $\{p_{2_u}(x, v, t)\delta v : \delta v \in R^m$, $q^j_u(v, t)\delta v < 0$ for each $j = 1, \cdots, r$
such that $q^j(v, t) = 0$, $p^j_u(x, v, t)\delta v < 0$ for each $j = 1, \cdots, l'$
such that $p^j(x, v, t) = 0\} = R^{l-l'}$ if $l' < l$,

(vb) $\{\delta v : \delta v \in R^m$, $q^j_u(v, t)\delta v < 0$ for each $j = 1, \cdots, r$
such that $q^j(v, t) = 0$, and $p^j_u(x, v, t)\delta v < 0$ for each
$j = 1, \cdots, l'$ such that $p^j(x, v, t) = 0\} \neq \varnothing$ if $l' = l$.

If we define $\boldsymbol{\omega}(x, t)$ in terms of $\mathbf{p}$, $\mathbf{q}$, and $\mathbf{U}_0$ in the same way that $\omega(x, t)$ was defined in terms of p, q, and U_0, then it is easy to see that $\mathbf{v} = (v, v^{m+1}, \cdots, v^{m+l'}) \in \boldsymbol{\omega}(x, t)$ if and only if $v \in \omega_1(x, t)$ and $v^{m+j} = -p^j(x, v, t)$ for $j = 1, \cdots, l'$.

Appealing to Theorem 13.1 for the reformulated problem, we arrive at the following pointwise maximum principle.

THEOREM 14.2. *Suppose that the hypotheses of Theorem* 14.1 *hold and that, in addition, the functions f, p, q, and p_u are continuous. Then the conclusion that* (12.26) *holds can be replaced by the conclusion that* (13.1) *and* (13.2) *hold, but with ω replaced by ω_1 in* (13.1).

As before, relation (13.1)—with ω replaced by ω_1—together with conclusions (v), (viii), and (x) of Theorem 14.1 allow us to interpret (for almost every $t \in I_1$) $\mu^1(t), \cdots, \mu^r(t), \nu^1(t), \cdots, \nu^l(t)$ as Lagrange multipliers for the problem of maximizing the function $v \to \hat{\psi}(t) f(\bar{x}(t), v, t) : O \to R$ (where O is some neighborhood of $\bar{u}(t)$) subject to the constraints of $q^j(v, t) \leqq 0$ for $j = 1, \cdots, r$; $p^j(\bar{x}(t), v, t) \leqq 0$ for $j = 1, \cdots, l'$; and $p^j(\bar{x}(t), v, t) = 0$ for $j = l' + 1, \cdots, l$. The statements made in § 13 regarding the uniqueness and continuity of the functions μ and ν carry over here essentially unchanged.

The two types of generalizations of our basic problem which we described in this section may of course be combined, but we shall leave the details of the analysis to the interested reader.

15. Comparison with earlier results. Optimal control problems of the type examined in this paper were first studied by Gamkrelidze (see [13] and [6]). Gamkrelidze primarily investigated the problem with constraints of the form (2.1)–(2.3) and (2.5), and confined himself to the case where I_1 is a closed interval, where f, p, and q are independent of t (in which case Assumptions A4–A6 automatically hold), and where χ has a particularly simple form. He also assumed that the solution $(\bar{x}, \bar{u})$ of the problem was such that $\bar{u}$ was piecewise-continuous and piecewise-smooth, and that, for each $t \in I_1$, $\bar{u}(t) \in \tilde{\omega}(\bar{x}(t), t)$, where (for each (x, t)) $\tilde{\omega}(x, t)$ differs from our $\omega(x, t)$ in that conditions (iii) and (iv) (in our definition of ω) are replaced by the much stronger requirement that the vectors

$$p_u^j(x, v), \quad j = 1, \cdots, l, \quad \text{and} \quad q_u^j(v) \quad \text{for all } j = 1, \cdots, r \text{ such that } q^j(v) = 0$$

are linearly independent. All this implies, as is not hard to see, that the function $\bar{u}$ satisfies our Conditions C1 and C2. The necessary conditions obtained in [6] and [13] under these stronger assumptions are essentially the same as those in our Theorem 13.1, except that the maximum condition (13.1) was shown to hold only for all $v \in \tilde{\omega}(\bar{x}(t), t)$ (rather than for all v in the larger set $\omega(\bar{x}(t), t)$). Gamkrelidze also investigated the problem with constraints of the form (14.3) by using essentially the same device as we used, but under the previously indicated stronger assumptions on the problem data. Finally, he also considered the problem with constraints of the form of (2.6), but only for the case where $\tilde{\chi}^i(\bar{x}(t), t) \equiv 0$ on some subinterval of I_i', in which case he could reduce this constraint to one of the form of (2.5), by using the procedure we described at the beginning of § 3.

Hestenes, in [3, pp. 256–260] (also see [7]), considered problems with constraints of the form of (2.2)–(2.5), where f and p are continuously differentiable. He also confined himself to piecewise-continuous control functions u. (An extension to the case of more general f and p—much as indicated in A3–A6—and to measurable u was made by Guinn in [14].) Finally, he supposed that $I_1 = I$ and that the function χ had a particularly simple form. Under the additional assumption that, for all $t \in I$, $\omega(\bar{x}(t), t) = \{v : v \in U_0, p(\bar{x}(t), v, t) = 0\}$ (which implies that $\bar{u}$ satisfies our Conditions C1 and C2), he obtained necessary conditions essentially the same as those in our Theorem 13.1. In [3, pp. 260–262], the problem with constraints of the form (2.2)–(2.4) and (14.3), (14.4) was also considered, and necessary conditions which are the same as those in Theorem 14.2 were obtained, under the very strong assumption that, for all $t \in I$ and all $v \in U_0$ that satisfy (14.8) and (14.9) with $x = \bar{x}(t)$, the vectors $p_u^j(\bar{x}(t), v, t)$, for those $j = 1, \cdots, l$ such that $p^j(\bar{x}(t), v, t) = 0$, are linearly independent. This assumption implies that $\omega_1(\bar{x}(t), t) = \{v : v \in U_0, p^j(\bar{x}(t), v, t) \leqq 0$ for $j = 1, \cdots, l'$ and $p^j(\bar{x}(t), v, t) = 0$ for $j = l' + 1, \cdots, l\}$ and that $\bar{u}$ satisfies Conditions C1′ and C2′, but the converse is far from true.

Virsan in [4] considered problems with constraints of the form (2.2), (2.3), and (2.5), under the assumption that f and p are continuous and continuously differentiable in x and u and that I_1 is a subinterval of I. He also confined himself to piecewise-continuous controls. Under the additional assumption that $\bar{u}(t) \in \omega(\bar{x}(t), t)$ for all $t \in I_1$ (which implies that $\bar{u}$ satisfies Conditions C1 and C2), he obtained necessary conditions essentially the same as those in our Theorem 13.1. Virsan's approach to the problem had much in common with ours. Namely, he viewed the constraints (2.5) as an equality constraint in the function space $L_1^l(I_1)$, and brought to bear the tools of functional analysis much as we did in § 11.

In [5], Virsan extended his results to measurable controls under the additional assumptions that $I = I_1$, that the functions f and p are independent of t, that, for each $t \in I$, $\omega(\bar{x}(t), t) = \{v : v \in U_0, p(\bar{x}(t), v, t) = 0\}$, and that the sets $\omega(\bar{x}(t), t)$ are uniformly bounded. His approach in this paper was even closer to ours than it was in [4].

Dubovitskii and Milyutin, in [15], considered a problem containing both equality and inequality constraints of the form (14.3) and (14.4), under assumptions much like our Conditions C1′ and C2′ in § 14. However, their necessary conditions did not include the maximum principle (12.26) (or (13.1)) but only the "local" maximum condition (12.28). Earlier, in [16], they had obtained similar necessary conditions for problems without the equality constraints (14.4).

Appendix. We first prove that, in the notation of § 7, if $\max_i |J_i^\varepsilon|$ is sufficiently small (for each $\varepsilon \in [0, \hat{\varepsilon})$) then $\|x_{\gamma,\varepsilon} - \tilde{x}_{\gamma,\varepsilon}\| < \varepsilon^2$ for all $\gamma \in S^k$. Let us denote the compact set $\{x : x \in R^n, |x - \bar{x}(t)| \leqq 2\rho_0$ for some $t \in I\}$ by G_c. By definition of ρ_0, $G_c \subset G$, and by choice of $\hat{\varepsilon}$, $\|\tilde{x}_{\gamma,\varepsilon} - \bar{x}\| < \rho_0$ for all $\gamma \in S^k$ and $\varepsilon \in [0, \hat{\varepsilon})$. Let us arbitrarily fix $\varepsilon \in [0, \hat{\varepsilon})$ for the remainder of the argument.

For each $\gamma \in S^k$, let us define the function $\delta f(\cdot, \cdot\,; \gamma) : G \times I \to R^n$ as follows:

$$\delta f(x, t; \gamma) = f(x, u_{\gamma,\varepsilon}(t), t) - f(x, \bar{u}(t), t)$$
$$- \varepsilon \sum_{j=1}^{s} \beta^j(\gamma)[f(x, u_j(t), t) - f(x, \bar{u}(t), t)].$$

It follows from Assumption A6 of § 2 that there is a number $\zeta > 0$ such that (see (7.1))

$$|f_x(x, u_{\gamma,\varepsilon}(t), t)| \leqq \zeta, \quad |\delta f_x(x, t; \gamma)| \leqq \zeta, \quad |\dot{\tilde{x}}_{\gamma,\varepsilon}(t)| \leqq \zeta$$

for all $x \in G_c$ and $\gamma \in S^k$ and almost all $t \in I$, independent of the way in which the J_i^ε are chosen.

For each γ, equation (7.9), with initial condition $x(t_1) = \tilde{x}_{\gamma,\varepsilon}(t_1) \in G_c$, has a solution which takes on its values in G_c—at least for t in some neighborhood of t_1. Let us denote this solution by $x_{\gamma,\varepsilon}$. Then, for all t in this neighborhood of t_1, we have (see (7.9) and (7.1))

$$\text{(A.1)} \qquad \begin{aligned} |x_{\gamma,\varepsilon}(t) - \tilde{x}_{\gamma,\varepsilon}(t)| \leqq & \int_{t_1}^{t} |f(x_{\gamma,\varepsilon}(s), u_{\gamma,\varepsilon}(s), s) - f(\tilde{x}_{\gamma,\varepsilon}(s), u_{\gamma,\varepsilon}(s), s)|\, ds \\ & + \left| \int_{t_1}^{t} \delta f(\tilde{x}_{\gamma,\varepsilon}(s), s; \gamma)\, ds \right|. \end{aligned}$$

But

$$\text{(A.2)} \qquad \begin{aligned} & \int_{t_1}^{t} |f(x_{\gamma,\varepsilon}(s), u_{\gamma,\varepsilon}(s), s) - f(\tilde{x}_{\gamma,\varepsilon}(s), u_{\gamma,\varepsilon}(s), s)|\, ds \\ & \quad \leqq \int_{t_1}^{t} (\max_{x \in G_c} |f_x(x, u_{\gamma,\varepsilon}(s), s)|) |x_{\gamma,\varepsilon}(s) - \tilde{x}_{\gamma,\varepsilon}(s)|\, ds \\ & \quad \leqq \int_{t_1}^{t} \zeta |x_{\gamma,\varepsilon}(s) - \tilde{x}_{\gamma,\varepsilon}(s)|\, ds. \end{aligned}$$

Also, if we denote by $\tilde{\tilde{x}}_{\gamma,\varepsilon}$ some step-function approximation to $\tilde{x}_{\gamma,\varepsilon}$, we similarly obtain

$$\text{(A.3)} \qquad \left| \int_{t_1}^{t} \delta f(\tilde{x}_{\gamma,\varepsilon}(s), s; \gamma)\, ds \right| \leqq \left| \int_{t_1}^{t} \delta f(\tilde{\tilde{x}}_{\gamma,\varepsilon}(s), s; \gamma)\, ds \right| + \zeta(t_2 - t_1) \|\tilde{\tilde{x}}_{\gamma,\varepsilon} - \tilde{x}_{\gamma,\varepsilon}\|.$$

Combining inequalities (A.1)–(A.3), we arrive at the inequality

$$\text{(A.4)} \qquad \begin{aligned} |x_{\gamma,\varepsilon}(t) - \tilde{x}_{\gamma,\varepsilon}(t)| \leqq & \int_{t_1}^{t} \zeta |x_{\gamma,\varepsilon}(s) - \tilde{x}_{\gamma,\varepsilon}(s)|\, ds \\ & + \zeta(t_2 - t_1) \|\tilde{\tilde{x}}_{\gamma,\varepsilon} - \tilde{x}_{\gamma,\varepsilon}\| + \left| \int_{t_1}^{t} \delta f(\tilde{\tilde{x}}_{\gamma,\varepsilon}(s), s; \gamma)\, ds \right|. \end{aligned}$$

Note that inequality (A.4) holds for every $\gamma \in S^k$, for every step function $\tilde{\tilde{x}}_{\gamma,\varepsilon}$ taking on its values in G_c, for any choice of the J_i^ε, and for every $t \in I$ such that $x_{\gamma,\varepsilon}(s) \in G_c$ for all $s \in [t_1, t]$. Applying the Gronwall inequality to (A.4), we can conclude that, for every $\gamma \in S^k$, $x_{\gamma,\varepsilon}$ is defined and takes on its values in G_c for all $t \in I$ and satisfies the inequality $\|x_{\gamma,\varepsilon} - \tilde{x}_{\gamma,\varepsilon}\| < \varepsilon^2$, provided that we can show that, by a proper choice of the step functions $\tilde{\tilde{x}}_{\gamma,\varepsilon}$ (depending on γ) and of the J_i^ε (independent of γ), we can make the last two terms in the right-hand side of (A.4) arbitrarily small in magnitude, for all $t \in I$ (and $\gamma \in S^k$).

But $\|\tilde{\tilde{x}}_{\gamma,\varepsilon} - \tilde{x}_{\gamma,\varepsilon}\|$ can evidently be made arbitrarily small by a suitable choice of $\tilde{\tilde{x}}_{\gamma,\varepsilon}$. We may evidently suppose that $\tilde{\tilde{x}}_{\gamma,\varepsilon}$ takes on its values in G_c. Further (since $|\dot{\tilde{x}}_{\gamma,\varepsilon}(t)| \leqq \zeta$ for all γ and almost all t) for a given upper bound on $\|\tilde{\tilde{x}}_{\gamma,\varepsilon} - \tilde{x}_{\gamma,\varepsilon}\|$, we may even choose the points of discontinuity of the $\tilde{\tilde{x}}_{\gamma,\varepsilon}$ to be the same for every $\gamma \in S^k$.

According to a lemma of Gamkrelidze (see [2, Lemma 4.1]), for any $\zeta_0 > 0$, there is a choice of the J_i^ε such that

$$\left| \int_{t'}^{t''} \delta f(x, s; \gamma)\, ds \right| < \zeta_0 \quad \text{for all } t', t'' \in I, \gamma \in S^k, \text{ and } x \in G_c,$$

which means that the last term in (A.4) can be made arbitrarily small by a suitable choice of the J_i^ε, which completes our proof.

Let us denote by $\mathscr{F}$ the set of all functions $F: G \times I \to R^n$ (where G is an open set in R^n and I is the compact interval $[t_1, t_2]$) such that (i) for almost every $t \in I$, the function $x \to F(x,t): G \to R^n$ is continuously differentiable, (ii) for each $x \in R^n$, the function $t \to F(x,t): I \to R^n$ is measurable, and (iii) for every compact subset G_c of G, there is a number $\zeta > 0$ such that, for almost every $t \in I$, $|F(x,t)| + |F_x(x,t)| \leqq \zeta$ for all $x \in G_c$. Thus we have the following continuous dependence theorem for ordinary differential equations.

THEOREM A.1. *Let $\tilde{F} \in \mathscr{F}$ and $\tilde{x}(\cdot) \in \mathscr{C}^n$ be such that $\tilde{x}(\cdot)$ is absolutely continuous and satisfies the differential equation $\dot{\tilde{x}}(t) = \tilde{F}(\tilde{x}(t), t)$ a.e. on I. Then there is a compact subset G_c of G with the following property: For every $\varepsilon > 0$, there is a $\delta > 0$ such that, whenever $F \in \mathscr{F}$ and $\xi_0 \in R^n$ satisfy the inequalities*

$$\int_{t_1}^{t_2} (\sup_{x \in G_c} |F(x,s) - \tilde{F}(x,s)|)\, ds < \delta, \qquad |\xi_0 - \tilde{x}(t_1)| < \delta,$$

the differential equation $\dot{x}(t) = F(x(t), t)$, with initial condition $x(t_1) = \xi_0$, has a solution defined on all of I satisfying $|x(t) - \tilde{x}(t)| < \varepsilon$ for all $t \in I$.

Proof. Choose $\delta_1 > 0$ such that the compact set $G_c = \{x : x \in R^n, |x - \tilde{x}(t)| \leqq \delta_1 \text{ for some } t \in I\}$ is contained in G, so that, for some $\zeta > 0$, $|\tilde{F}_x(x,t)| \leqq \zeta$ for all $x \in G_c$, for almost all $t \in I$. If $F \in \mathscr{F}$ and $|\xi_0 - \tilde{x}(t_1)| \leqq \delta_1/2$, then the equation $\dot{x}(t) = F(x(t), t)$, with initial value $x(t_1) = \xi_0$, has, at least in some neighborhood of t_1, a solution satisfying $|x(t) - \tilde{x}(t)| \leqq \delta_1$. For each $t \in I$ such that $|x(s) - \tilde{x}(s)| \leqq \delta_1$ for all $s \in [t_1, t]$, we then have

$$\begin{aligned}
|x(t) - \tilde{x}(t)| &= \left| \xi_0 - \tilde{x}(t_1) + \int_{t_1}^{t} [F(x(s), s) - \tilde{F}(\tilde{x}(s), s)]\, ds \right| \\
&\leqq |\xi_0 - \tilde{x}(t_1)| + \int_{t_1}^{t} |F(x(s), s) - \tilde{F}(x(s), s)|\, ds \\
&\qquad + \int_{t_1}^{t} |\tilde{F}(x(s), s) - \tilde{F}(\tilde{x}(s), s)|\, ds \\
&\leqq |\xi_0 - \tilde{x}(t_1)| + \int_{t_1}^{t_2} (\sup_{x \in G_c} |F(x,s) - \tilde{F}(x,s)|)\, ds \\
&\qquad + \int_{t_1}^{t} \zeta |x(s) - \tilde{x}(s)|\, ds,
\end{aligned}$$

and our desired result follows at once from the Gronwall inequality.

Remark A.1. In the preceding theorem, δ depends on $\tilde{F}$ only in that δ is a function of ζ.

We also have the following differentiable dependence theorem.

THEOREM A.2. *Let I, G, U_0, f, $\mathscr{G}$, $\mathscr{U}_0$, and $\mathscr{U}$ be as described in* § 2, *and suppose that f satisfies Assumptions* A3–A6 *of* § 2. *Suppose that $\tilde{x}(\cdot) \in \mathscr{G}$ and $\tilde{u}(\cdot) \in \mathscr{U}$ satisfy the differential equation $\dot{\tilde{x}}(t) = f(\tilde{x}(t), \tilde{u}(t), t)$ a.e. on I, with $\tilde{x}$ absolutely continuous. Then there is a neighborhood N of $\tilde{u}$ in L_∞^m such that, for all $u \in N$, the equation $\dot{x}(t) = f(x(t), u(t), t)$, with initial value $x(t_1) = \tilde{x}(t_1)$, has a solution which is defined on all of I, and* (*denoting this solution by x_u to emphasize its dependence on u*)

$$x_u = \tilde{x} + \delta x_u + o(u - \tilde{u}),$$

where (*for each $u \in \mathscr{U}$*) *δx_u denotes the solution of*

$$\delta\dot{x}(t) = f_x(\tilde{x}(t), \tilde{u}(t), t)\delta x(t) + f_u(\tilde{x}(t), \tilde{u}(t), t)(u(t) - \tilde{u}(t)), \qquad \delta x(t_1) = 0,$$

and

$$\frac{o(u)}{\|u\|} \xrightarrow[u \to 0]{} 0 \quad \textit{in the } \mathscr{C}^n\textit{-norm}.$$

Proof. Let us denote by T the mapping from $\mathscr{G} \times \mathscr{U}_0$ into $\mathscr{C}^n$ defined by

$$(T(x, u))(t) = \tilde{x}(t_1) + \int_{t_1}^{t} f(x(s), u(s), s)\, ds, \qquad t \in I, \quad x \in \mathscr{G}, \quad u \in \mathscr{U}_0.$$

Then the differential equation $\dot{x}(t) = f(x(t), u(t), t)$ with initial value $x(t_1) = \tilde{x}(t_1)$ is evidently equivalent to the equation $x = T(x, u)$, so that $\tilde{x} = T(\tilde{x}, \tilde{u})$. It easily follows from Assumptions A.3–A.6 that T is continuously (Fréchet) differentiable.

Our result now follows directly from the conventional implicit function theorem in Banach spaces (see [8, Thm. 10.2.1, p. 265]).

Acknowledgment. The authors wish to acknowledge some fruitful conversations with Leonard M. Silverman. Indeed, the construction of the functions δx_z and δu_z in § 4 are based on suggestions made by him.

REFERENCES

[1] L. W. NEUSTADT, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.

[2] R. V. GAMKRELIDZE, *On some extremal problems in the theory of differential equations with applications to the theory of optimal control*, this Journal, 3 (1965), pp. 106–128.

[3] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966, Chap. 6.

[4] C. VIRSAN, *Necessary conditions for optimization problems with operatorial constraints*, this Journal, 8 (1970), pp. 527–558.

[5] ———, *Necessary conditions for optimal control problems with equality-type mixed constraint*, Rev. Roumaine Math. Pures Appl., 16 (1971), pp. 127–147.

[6] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962, Chap. VI.

[7] M. R. HESTENES, *On variational theory and optimal control theory*, this Journal, 3 (1965), pp. 23–48.

[8] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[9] C. BERGE, *Topological Spaces*, Macmillan, New York, 1963.

[10] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I: General Theory*, Interscience, New York, 1958.

[11] I. P. NATANSON, *The Theory of Functions of a Real Variable*, Frederick Ungar, New York, 1955.

[12] E. J. MCSHANE AND T. A. BOTTS, *Real Analysis*, Van Nostrand, Princeton, N.J., 1959.

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

[14] T. GUINN, *Weakened hypotheses for the variational problem considered by Hestenes*, this Journal, 3 (1965), pp. 418–423.

[15] A. Y. DUBOVITSKII AND A. A. MILYUTIN, *Maximum principle in a class of variations of small absolute value for optimal control problems with mixed constraints of equality and inequality type*, Soviet Math., Dokl., 10 (1969), pp. 1567–1571.

[16] ———, *Necessary conditions for a weak extremum in optimal control problems with mixed inequality-type constraints*, Ž. Vyčisl. Mat. i Fiz., 8 (1968), pp. 725–779.

SIAM J. CONTROL
Vol. 12, No. 2, May 1974

IMPLICIT FUNCTIONS AND OPTIMIZATION PROBLEMS WITHOUT CONTINUOUS DIFFERENTIABILITY OF THE DATA*

HUBERT HALKIN†

Abstract. Let ϕ be a function from a normed linear space X into a finite-dimensional Euclidean space Y and let A be a continuous linear mapping from X *onto* Y. We assume that ϕ is continuous in a neighborhood of some point $\hat{x} \in X$ and that ϕ admits A as its differential at the point $\hat{x}$. In this paper we prove that under those conditions there exist a neighborhood U of $\hat{x}$ and a mapping ζ from U into X such that $\phi(x + \zeta(x)) = \phi(\hat{x}) + A(x - \hat{x})$ for all $x \in U$ and such that $\lim_{\eta \to 0+} \sup_{|x-\hat{x}| \leqq \eta} |\zeta(x)|/\eta = 0$. Besides the above "correction function theorem" this paper contains related implicit and inverse function theorems. The correction function theorem is applied to the proof of the multiplier rule for mathematical programming problems with equality and inequality constraints without assuming the continuous differentiability of the data in a neighborhood of the optimal solution.

1. Introduction.

If ϕ is a function from a normed linear space X into a normed linear space Y and if a continuous linear mapping A from X into Y is the differential of ϕ at a point $\hat{x} \in X$, then by definition we have

$$\phi(x) = \phi(\hat{x}) + A(x - \hat{x}) + g(x), \tag{1.1}$$

where $\lim_{\eta \to 0+} \sup_{|x-\hat{x}| \leqq \eta} |g(x)|/\eta = 0$. In other words, the relation

$$\phi(x) = \phi(\hat{x}) + A(x - \hat{x}) \tag{1.2}$$

is "almost" correct and to obtain the "really" correct relation (1.1) we have added to the right side of relation (1.2) a "little" term $g(x)$. Instead of making the above classical correction we might try a correction of another type by shifting the argument x in the left side of relation (1.2) by a "little" amount $\zeta(x)$ in order to obtain a relation of the type

$$\phi(x + \zeta(x)) = \phi(\hat{x}) + A(x - \hat{x}). \tag{1.3}$$

The fact that such a correction is possible whenever ϕ is continuous in a neighborhood of $\hat{x}$, Y is a finite-dimensional Euclidean space and A maps X *onto* Y constitutes the "correction" function theorem stated below[1] as Theorem F and proved in § 2 of this paper.

THEOREM F. *Let X be a normed linear space and let Y be a finite-dimensional Euclidean space. Let ϕ be a mapping from X into Y, let $\hat{x} \in X$ and let A be a continuous linear mapping of X onto Y such that*

(i) *ϕ is continuous in a neighborhood of $\hat{x}$,*

(ii) *A is the differential of ϕ at the point $\hat{x}$.*

* Received by the editors April 10, 1972.

† Department of Mathematics, University of California, La Jolla, California 92037. The first version of this article was written at the Center for Operations Research and Econometrics (CORE), Louvain, where the author was on sabbatical leave from the University of California supported in part by a Guggenheim fellowship. Preliminary research for this paper was supported by a grant from the U.S. Air Force Office of Scientific Research. The final version of this paper was written while the author was a visiting member of the Centre de Recherches Mathématiques, Université de Montréal.

[1] The results of this paper are given in Theorems A through G. The order A through G corresponds to the logical order of construction of the proofs given in § 2. In the Introduction we state Theorems F, C, E and G in that order for "motivational" reasons. Theorems A, B and D are stated only in § 2.

Then there exist a neighborhood U of $\hat{x}$ and a mapping ζ of U into X such that

(α) $\phi(x + \zeta(x)) = \phi(\hat{x}) + A(x - \hat{x})$ *for all* $x \in U$ *and*

(β) $\lim_{\eta \to 0+} \sup_{|x-\hat{x}| \leqq \eta} |\zeta(x)|/\eta = 0.$

In § 2 we first prove a string of theorems including the two implicit function theorems stated below as Theorems C and E, and then we use Theorem E to prove Theorem F.

THEOREM C. *Let X be a metric space, let Y be a normed linear space, let Z be a finite-dimensional Euclidean space, let ϕ be a mapping from $X \times Y$ into Z, let $(\hat{x}, \hat{y}) \in X \times Y$ and let B be a continuous linear mapping of Y onto Z such that*

(i) *ϕ is continuous in a neighborhood of $(\hat{x}, \hat{y})$,*

(ii) *ϕ admits B as its differential with respect to the second variable at the point $(\hat{x}, \hat{y})$, i.e.,*

$$\lim_{\eta \to 0+} \sup_{|y-\hat{y}| \leqq \eta} |\phi(\hat{x}, \hat{y}) + B(y - \hat{y}) - \phi(\hat{x}, y)|/\eta = 0.$$

Then for any neighborhood V of $\hat{y}$ there exist a neighborhood U of $\hat{x}$ and a function ψ from U into V such that

(α) $\psi(\hat{x}) = \hat{y}$,

(β) $\phi(x, \psi(x)) = \phi(\hat{x}, \hat{y})$ *for all* $x \in U$, *and*

(γ) *ψ is continuous at the point $\hat{x}$.*

THEOREM E. *Let X and Y be normed linear spaces, let Z be a finite-dimensional Euclidean space, let ϕ be a mapping from $X \times Y$ into Z, let $(\hat{x}, \hat{y}) \in X \times Y$, let A be a continuous linear mapping from X into Z and let B be a continuous linear mapping from Y onto Z such that*

(i) *ϕ is continuous in a neighborhood of $(\hat{x}, \hat{y})$,*

(ii) *(A, B) is the differential of ϕ at the point $(\hat{x}, \hat{y})$, i.e.,*

$$\lim_{\eta \to 0+} \sup_{|x-\hat{x}|+|y-\hat{y}| \leqq \eta} |\phi(\hat{x}, \hat{y}) + A(x - \hat{x}) + B(y - \hat{y}) - \phi(x, y)|/\eta = 0.$$

Then there exist a neighborhood U of $\hat{x}$, a mapping ψ of U into Y and a continuous linear mapping C from X into Y such that

(α) $\psi(\hat{x}) = \hat{y}$,

(β) $\phi(x, \psi(x)) = \phi(\hat{x}, \hat{y})$ *for all* $x \in U$,

(γ) *C is the differential of ψ at the point $\hat{x}$, and*

(δ) $A + BC = 0$.

In § 2 we also apply Theorem E to the proof of the following inverse function theorem.

THEOREM G. *Let X be a finite-dimensional Euclidean space, let ϕ be a mapping from X into X, let $\hat{x} \in X$ and let A be a linear mapping of X onto X such that*

(i) *ϕ is continuous in a neighborhood of $\hat{x}$ and*

(ii) *A is the differential of ϕ at the point $\hat{x}$.*

Then there exist a neighborhood U of $\phi(\hat{x})$ and a mapping ζ from U into X such that

(α) $\zeta(\phi(\hat{x})) = \hat{x}$,

(β) $\phi(\zeta(y)) = y$ *for all* $y \in U$, *and*

(γ) *ζ admits A^{-1} as its differential at the point $\phi(\hat{x})$.*

The implicit and inverse function theorems stated above, Theorems E, C and G, are closely related to their "classical" counterparts (see, for instance, Dieudonné [1, p. 265]) in which the given functions are assumed to be continuously differentiable and in which the resulting functions are themselves guaranteed to be continuously differentiable. There does not appear to be any "classical" correction function theorem stated explicitly in the literature. We give below the statement of such a theorem which can be proved as an easy corollary of the "classical" implicit function theorem.

CLASSICAL CORRECTION FUNCTION THEOREM. *Let X and Y be Banach spaces, and let T be a closed linear subspace of X. Let $\phi : X \to Y$, let $\hat{x} \in X$ and let A be a continuous linear mapping of X into Y such that*

(i) *ϕ is continuously differentiable in a neighborhood of $\hat{x}$,*

(ii) *A is the differential of ϕ at the point $\hat{x}$ and*

(iii) *A maps T one-to-one onto Y.*

Then there exist a neighborhood U of $\hat{x}$ and a continuously differentiable mapping ζ of U into X such that

(α) *$\phi(x + \zeta(x)) = \phi(\hat{x}) + A(x - \hat{x})$ for all $x \in U$, and*

(β) $\lim_{\eta \to 0+} \sup_{|x-\hat{x}| \leqq \eta} |\zeta(x)|/\eta = 0$.

In § 3 we apply Theorem F, the correction function theorem, to the proof of the multiplier rule for a mathematical programming problem with equality and inequality constraints without assuming the continuous differentiability of the data in a neighborhood of the optimal solution. If the data is assumed to be continuously differentiable in a neighborhood of the optimal solution, then this "classical" multiplier rule (which combines the results of Carathéodory [2, p. 177] and John [3]) can be found in Mangasarian [4]. In Mangasarian [4] the classical multiplier rule is derived from the classical implicit function theorem. The proof of the classical multiplier rule would be greatly simplified by patterning it on § 3 of the present paper, i.e., by using the classical correction function theorem instead of the classical implicit function theorem.

The proofs of § 2 are based on the Brouwer fixed-point theorem.[2] The result of § 3 is a particular case of some general necessary conditions in mathematical programming (Halkin–Neustadt [5], Halkin [6]). Those necessary conditions are all derived by a specific application of the Brouwer fixed-point theorem. This type of application of the Brouwer fixed-point theorem to optimization problems was introduced in Halkin [7].

2. Proofs of the implicit, correction, and inverse function theorems.

THEOREM A. *Let X be a metric space, let Y be a finite-dimensional Euclidean space, let ϕ be a mapping from $X \times Y$ into Y, let $(\hat{x}, \hat{y}) \in X \times Y$ and let B be a linear mapping of Y onto Y such that*

(i) *ϕ is continuous in a neighborhood of $(\hat{x}, \hat{y})$,*

(ii) *ϕ admits B as its differential with respect to the second variable at the point $(\hat{x}, \hat{y})$, i.e.,*

$$\lim_{\eta \to 0+} \sup_{|y-\hat{y}| \leqq \eta} |\phi(\hat{x}, \hat{y}) + B(y - \hat{y}) - \phi(\hat{x}, y)|/\eta = 0.$$

Then for each neighborhood V of $\hat{y}$ there exist a neighborhood U of $\hat{x}$ and a function ψ from U into V such that $\phi(x, \psi(x)) = \phi(\hat{x}, \hat{y})$ for all $x \in U$.

[2] The key use of the Brouwer fixed-point theorem is in the proof of Theorem A. The assumption of finite-dimensionality of space X in Theorem G, of space Y in Theorems A, B, D and F and of space Z in Theorems C and E is conditioned by that use of the Brouwer fixed-point theorem.

Proof of Theorem A. We may assume without loss of generality that $\hat{y} = \phi(\hat{x}, \hat{y}) = 0$. For any $\varepsilon > 0$ let

$$N_\varepsilon(\hat{x}) = \{x : x \in X, d(x, \hat{x}) \leqq \varepsilon\},$$

where $d(\cdot, \cdot)$ is the metric on X, and let

$$B_\varepsilon = \{y : y \in Y, |y| \leqq \varepsilon\}.$$

Let $\sigma > 0$ be such that $B_\sigma \subset V$ and such that ϕ is continuous over $N_\sigma(\hat{x}) \times B_\sigma$. Let $\alpha \in (0, \sigma]$ be such that

$$|\phi(\hat{x}, y) - B(y)| \leqq \alpha/(2|B^{-1}|)$$

whenever $y \in B_\alpha$, and let[3] $\beta \in (0, \sigma]$ be such that

$$|\phi(x, y) - \phi(\hat{x}, y)| \leqq \alpha/(2|B^{-1}|)$$

whenever $x \in N_\beta(\hat{x})$ and $y \in B_\alpha$. For all $x \in N_\beta(\hat{x})$, let h_x be the mapping from B_α into Y defined by

$$h_x(y) = y - B^{-1}\phi(x, y).$$

The mapping h_x is continuous and maps B_α into itself since for all $y \in B_\alpha$, we have

$$|h_x(y)| \leqq |y - B^{-1}\phi(\hat{x}, y)| + |B^{-1}\phi(\hat{x}, y) - B^{-1}\phi(x, y)| \leqq \alpha/2 + \alpha/2 = \alpha.$$

Let $\psi(x)$ be the fixed point of h_x given by the Brouwer fixed-point theorem. For all $x \in N_\beta(\hat{x})$ we have then $\psi(x) \in B_\alpha \subset V$ and $\psi(x) - B^{-1}\phi(x, \psi(x)) = \psi(x)$, i.e.,

$$\phi(x, \psi(x)) = 0.$$

This concludes the proof of Theorem A.

THEOREM B. *Under the assumptions of Theorem A then for any neighborhood V of $\hat{y}$ there exist a neighborhood U of $\hat{x}$ and a function ψ from U into V such that*

(α) $\psi(\hat{x}) = \hat{y}$,

(β) $\phi(x, \psi(x)) = \phi(\hat{x}, \hat{y})$ *for all* $x \in U$, *and*

(γ) ψ *is continuous at the point* $\hat{x}$.

Proof of Theorem B. Let V_i, $i = 1, 2, \cdots$, be a sequence of neighborhoods of $\hat{y}$ such that $V_i = V$, $V_{i+1} \subset V_i$ for all $i = 1, 2, \cdots$ and such that for all $\varepsilon > 0$ there is an $n < +\infty$ with $V_n \subset B_\varepsilon$. For every $i = 1, 2, \cdots$ we know by Theorem A that there exist a neighborhood U_i of $\hat{x}$ and a function ψ_i from U_i into V_i such that $\phi(x, \psi_i(x)) = \phi(\hat{x}, \hat{y})$ for all $x \in U_i$. We may assume without loss of generality that $U_{i+1} \subset U_i$ for all $i = 1, 2, \cdots$ and that for all $\varepsilon > 0$ there is an $n < +\infty$ with $U_n \subset N_\varepsilon(\hat{x})$. Let $U = U_1$, let $\psi(\hat{x}) = \hat{y}$ and let $\psi(x) = \psi_i(x)$ whenever $x \in U_i \sim U_{i+1}$ for some $i = 1, 2, \cdots$. The function ψ satisfies conditions (α), (β) and (γ). This concludes the proof of Theorem B.

THEOREM C. *Let X be a metric space, let Y be a normed linear space, let Z be a finite-dimensional Euclidean space, let ϕ be a mapping from $X \times Y$ into Z, let $(\hat{x}, \hat{y}) \in X \times Y$ and let B be a continuous linear mapping of Y onto Z such that conditions* (i) *and* (ii) *of Theorem* A *hold. Then the conclusions of Theorems* A *and* B *are still valid.*

Proof of Theorem C. Let T be a finite-dimensional linear subspace of Y and let B^{-1} be a linear mapping from Z *onto* T such that $BB^{-1}z = z$ for all $z \in Z$. This is possible since B maps Y onto Z and Z is finite-dimensional. The spaces

[3] This is possible since ϕ is continuous and B_α is compact.

T and Z are thus finite-dimensional Euclidean spaces of equal dimension and may be considered as the same space. We conclude the proof of Theorem C by applying Theorems A and B to the restriction of the function ϕ to the domain $X \times T$.

THEOREM D. *Let X be a normed linear space, let Y be a finite-dimensional Euclidean space, let $\phi : X \times Y \to Y$, let $(\hat{x}, \hat{y}) \in X \times Y$, let A be a continuous linear mapping from X into Y and let B be a linear mapping from Y onto Y such that*

(i) *ϕ is continuous in a neighborhood of $(\hat{x}, \hat{y})$,*

(ii) *(A, B) is the differential of ϕ at the point $(\hat{x}, \hat{y})$, i.e.,*

$$\lim_{\eta \to 0+} \sup_{|x-\hat{x}|+|y-\hat{y}| \leqq \eta} |\phi(\hat{x}, \hat{y}) + A(x - \hat{x}) + B(y - \hat{y}) - \phi(x, y)|/\eta = 0.$$

Then there exist a neighborhood U of $\hat{x}$, a mapping ψ of U into Y such that

(α) *$\psi(\hat{x}) = \hat{y}$,*

(β) *$\phi(x, \psi(x)) = \phi(\hat{x}, \hat{y})$ for all $x \in U$, and*

(γ) *ψ admits $-B^{-1}A$ as its differential at $\hat{x}$.*

Proof of Theorem D. We may assume without loss of generality that $\hat{x} = 0$, $\hat{y} = 0$ and $\phi(0, 0) = 0$. By Theorem B we know that there exist a neighborhood U of $\hat{x} = 0$, a mapping ψ of U into Y such that conditions (α) and (β) hold, and such that ψ is continuous at $\hat{x} = 0$. It remains to prove that this function ψ satisfies condition (γ). For all $x \in U$ we have $Ax + B\psi(x) + g(x, \psi(x)) = 0$, where $g(x, y)$ is a function such that

$$\lim_{\eta \to 0+} \sup_{|x|+|y| \leqq \eta} |g(x, y)|/\eta = 0.$$

We have then

$$\psi(x) = -B^{-1}Ax - B^{-1}g(x, \psi(x)) \quad \text{whenever } x \in U.$$

It remains to prove that

$$\lim_{\eta \to 0+} \sup_{|x| \leqq \eta} |B^{-1}g(x, \psi(x))|/\eta = 0.$$

This will be done by first proving that ψ is Lipschitz continuous at $\hat{x} = 0$. Let $\sigma > 0$ be such that $\{x : |x| \leqq \sigma\} \subset U$ and $|g(x, y)| \leqq (|x| + |y|)/(2|B^{-1}|)$ whenever $|x|$ and $|y| \leqq \sigma$. Let $\alpha \in (0, \sigma]$ be such that $|\psi(x)| \leqq \sigma$ whenever $|x| \leqq \alpha$. We then have

$$|\psi(x)| \leqq |B^{-1}A|\,|x| + |B^{-1}|\,|g(x, \psi(x))| \leqq (|B^{-1}A| + \tfrac{1}{2})|x| + \tfrac{1}{2}|\psi(x)|,$$

i.e.,

$$|\psi(x)| \leqq (1 + 2|B^{-1}A|)|x| \quad \text{whenever } |x| \leqq \alpha.$$

We then have

$$\lim_{\eta \to 0+} \sup_{|x| \leqq \eta} |B^{-1}g(x, \psi(x))|/\eta \leqq \lim_{\eta \to 0+} \sup_{|x|+|y| \leqq 2(1+|B^{-1}A|)\eta} |B^{-1}g(x, y)|/\eta = 0.$$

This concludes the proof of Theorem D.

THEOREM E. See the statement given in the Introduction.

Proof of Theorem E. From Theorem D one obtains immediately Theorem E in the same manner as Theorem C was obtained from Theorems A and B.

It is possible to prove Theorem E directly, i.e., without first proving Theorems A, B and D. The direct proof of Theorem E given below combines, but also obscures, several aspects of the proofs given above.

Direct proof of Theorem E. Let T be a finite-dimensional linear subspace of Y and let B^{-1} be a linear mapping from Z onto T such that $BB^{-1}z = z$ for all $z \in Z$. This is possible since B maps Y *onto* Z and Z is *finite-dimensional.* Without loss of generality we may assume that $\hat{x} = 0$, $\hat{y} = 0$ and $\phi(\hat{x}, \hat{y}) = \phi(0, 0) = 0$. For $t > 0$ let

$$g(t) = \sup_{|x| + |y| \leqq t} |Ax + By - \phi(x, y)|.$$

Let $\varepsilon > 0$ be such that ϕ is continuous on $\{(x, y) : |x|, |y| \leqq \varepsilon\}$. Let $\eta \in (0, \varepsilon)$ be such that for all $x \in X$ with $|x| \leqq \eta$ we have

$$|B^{-1}|E_1(|x|) \leqq \varepsilon \quad \text{and} \quad E_2(|x|) \leqq E(|x|),$$

where

$$E(|x|) = 2g(2(|A| + 1)|x|),$$

$$E_1(|x|) = E(|x|) + |A|\,|x| + g(|x|),$$

$$E_2(|x|) = g(|x|) + g(|x| + E_1(|x|)).$$

This is possible since g is nonnegative, nondecreasing and such that $\lim_{t\to 0+} g(t)/t = 0$. Let $x \in X$ with $|x| \leqq \eta$, let

$$S_x = \{z : z \in Z, |z| \leqq E(|x|)\}$$

and let h_x be a mapping from S_x into Z defined by

$$h_x(z) = z - \phi(x, B^{-1}(z - \phi(x, 0))).$$

For $|x| \leqq \eta$ and $z \in S_x$ we have

$$|z| + |\phi(x, 0)| \leqq E(|x|) + |A|\,|x| + g(|x|) \leqq E_1(|x|)$$

and hence (i) the mapping h_x is continuous since

$$|x| \leqq \varepsilon \quad \text{and} \quad |B^{-1}(z - \phi(x, 0))| \leqq |B^{-1}|E_1(|x|) \leqq \varepsilon,$$

and (ii) the mapping h_x maps S_x into S_x since

$$\begin{aligned} |h_x(z)| &\leqq |z - Ax - (z - \phi(x, 0))| + g(|x| + |z| + |\phi(x, 0)|) \\ &\leqq g(|x|) + g(|x| + E_1(|x|)) \leqq E_2(|x|) \leqq E(|x|). \end{aligned}$$

From the Brouwer fixed-point theorem we know that there exists a point $z_x \in S_x$ such that $h_x(z_x) = z_x$, i.e., such that $\phi(x, B^{-1}(z_x - \phi(x, 0))) = 0$. Let $\psi(x) = B^{-1}(z_x - \phi(x, 0))$. We have

$$\begin{aligned} |\psi(x) + B^{-1}Ax| &\leqq |B^{-1}|\,|z_x - \phi(x, 0) + Ax| \\ &\leqq |B^{-1}|(|z_x| + g(|x|)) \leqq |B^{-1}|(2g(2(|A| + 1)|x|) + g(|x|)) \end{aligned}$$

which implies that ψ is differentiable at the point 0 and admits $-B^{-1}A$ as its differential. This concludes the direct proof of Theorem E.

THEOREM F (Correction function theorem).[4]

THEOREM G (Inverse function theorem).[4]

[4] See the statement given in the Introduction.

Proof of Theorem F. Let T be a finite-dimensional subspace of X such that A maps T one-to-one onto Y. Let ϕ_1 be a mapping from $X \times T$ into Y defined by $\phi_1(x, y) = \phi(x + y) - A(x - \hat{x})$. Let (A_1, B_1) be the differential of ϕ_1 at the point $(\hat{x}, \hat{y} = 0)$. We have $A_1 = 0$ and $B_1 = A$. We apply Theorem D at the point $(\hat{x}, \hat{y} = 0)$ for the function ϕ_1. There thus exist a neighborhood U of $\hat{x}$, a mapping ψ of U into T and a continuous linear mapping C_1 from X into T such that (α) $\psi(\hat{x}) = 0$, (β) $\phi_1(x, \psi(x)) = \phi_1(\hat{x}, \hat{y}) = \phi(\hat{x})$ for all $x \in U$, (γ) C_1 is the differential of ψ at the point $\hat{x}$ and (δ) $A_1 + B_1C_1 = 0$. By letting $\zeta = \psi$ we obtain $\phi(x + \zeta(x)) = \phi(\hat{x}) + A(x - \hat{x})$ for all $x \in U$. Moreover we have $C_1 = 0$ (since $A_1 = 0$ and since $B_1 = A$ maps T one-to-one onto Y) and hence $\lim_{\eta \to 0+} \sup_{|x - \hat{x}| \leqq \eta} |\zeta(x)|/\eta = 0$. This concludes the proof of Theorem F.

Proof of Theorem G. We are given a finite-dimensional Euclidean space X, a mapping ϕ from X into X, an element $\hat{x} \in X$ and a linear mapping A from X onto X such that

(i) ϕ is continuous in a neighborhood of $\hat{x}$ and

(ii) A is the differential of ϕ at the point $\hat{x}$.

Let ϕ_1 be a mapping from $X \times X$ into X defined by $\phi_1(u, v) = u - \phi(v)$. Let (A_1, B_1) be the differential of ϕ_1 at the point $\hat{u} = \phi(\hat{x})$, $\hat{v} = \hat{x}$. We have $A_1 = I$, the identity mapping, and $B_1 = -A$. We apply Theorem E at the point $(\hat{u}, \hat{v})$ for the function ϕ_1. There thus exist a neighborhood U of $\hat{u} = \phi(\hat{x})$, a mapping ψ of U into X and a linear mapping C_1 from X into X such that

(α) $\psi(\hat{u}) = \hat{v}$, i.e., $\psi(\phi(\hat{x})) = \hat{x}$,

(β) $\phi_1(u, \psi(u)) = \phi_1(\hat{u}, \hat{v})$ for all $u \in U$, i.e., $\phi(\psi(y)) = y$ for all $y \in U$,

(γ) C_1 is the differential of ψ at the point $\hat{u}$ and

(δ) $A_1 + B_1C_1 = 0$.

We have $C_1 = A^{-1}$ since $A_1 = I$ and $B_1 = -A$ is onto. By letting $\zeta = \psi$ we conclude the proof of Theorem G.

3. Multiplier rule for optimization problems with equality and inequality constraints which are not assumed to be continuously differentiable. We are given a normed linear space X and functions $\phi_{-\mu}, \cdots, \phi_{-1}, \phi_0, \phi_1, \cdots, \phi_m$ from X into R^1. The problem is to find an $\hat{x} \in A \equiv \{x : x \in X, \phi_i(x) \leqq 0 \text{ for } i = -\mu, \cdots, -1$ and $\phi_i(x) = 0$ for $i = 1, \cdots, m\}$ such that $\phi_0(\hat{x}) \leqq \phi_0(x)$ for all $x \in A$. An element $\hat{x} \in A$ satisfying this condition will be called *optimal*. We shall prove the following result.

MULTIPLIER RULE. *If $\hat{x}$ is an optimal solution, if $\phi_1, \cdots, \phi_m$ are continuous in a neighborhood of $\hat{x}$, and if $\phi_{-\mu}, \cdots, \phi_m$ admit differentials $a_{-\mu}, \cdots, a_m$ at the point $\hat{x}$, then there exists a nonzero vector $\lambda = (\lambda_{-\mu}, \cdots, \lambda_m) \in R^{\mu+m+1}$ such that*

(i) $\sum_{i=-\mu,\cdots,m} \lambda_i a_i = 0$,

(ii) $\lambda_i \leqq 0$ *for* $i = -\mu, \cdots, 0$,

(iii) $\lambda_i \phi_i(\hat{x}) = 0$ *for* $i = -\mu, \cdots, -1$.

Proof of the multiplier rule. We may assume without loss of generality that $\hat{x} = 0$. We may also assume that the continuous linear functionals $a_1, \cdots, a_m$ are linearly independent. Indeed, if $\sum_{i=1,\cdots,m} \mu_i a_i = 0$ for some $(\mu_1, \cdots, \mu_m) \neq 0$, then by letting $\lambda_i = 0$ for $i = -\mu, \cdots, 0$ and $\lambda_i = \mu_i$ for $i = 1, \cdots, m$ we obtain the required multiplier rule. If the continuous linear functionals $a_1, \cdots, a_m$ are linearly independent, then by Theorem F there exist a neighborhood U of $\hat{x}$ and a mapping ζ from U into X such that

$$\lim_{\eta \to 0+} \sup_{|x| \leqq \eta} |\zeta(x)|/\eta = 0,$$

and such that $\phi_i(x + \zeta(x)) = a_i(x)$ for all $x \in U$ and for all $i = 1, \cdots, m$. Let $S = \{x : x \in X, a_i(x) < -\phi_i(\hat{x})$ for $i = -\mu, \cdots, -1$; $a_0(x) < 0$ and $a_i(x) = 0$ for $i = 1, \cdots, m\}$. We shall first prove that S is empty, or equivalently that $S \cap U$ is empty. If $x^* \in S \cap U$, let $g(t) = tx^* + \zeta(tx^*)$ for all $t \in [0, 1]$. Let $f_i(t) = \phi_i(g_i(t))$ for all $i = -\mu, \cdots, m$ and all $t \in [0, 1]$. The functions f_i have one-sided derivatives at $t = 0$ and we have $f_i'(0+) = a_i(x^*)$ for $i = -\mu, \cdots, m$. Since $\phi_i(\hat{x}) \leqq 0$ for $i = -\mu, \cdots, -1$ there thus exists a $\tau \in (0, 1]$ such that $\phi_i(g(\tau)) < 0$ for $i = -\mu, \cdots, -1$; $\phi_0(g(\tau)) < \phi_0(\hat{x})$ and $\phi_i(g(\tau)) = 0$ for $i = 1, \cdots, m$. The existence of $g(\tau)$ contradicts the optimality of $\hat{x}$. We have thus proved that S is empty. Let $K_1 = \{z = (z_{-\mu}, \cdots, z_m)$: for some $x \in X$ we have $z_i = a_i(x)$ for $i = -\mu, \cdots, m\}$ and let $K_2 = \{z = (z_{-\mu}, \cdots, z_m) : z_i < -\phi_i(\hat{x})$ for $i = -\mu, \cdots, -1$; $z_0 < 0$ and $z_i = 0$ for $i = 1, \cdots, m\}$. Since S is empty the sets K_1 and K_2 are convex disjoint subsets of $R^{\mu+m+1}$. Moreover we have $0 \in K_1 \cap \bar{K}_2$, hence there exists a nonzero vector $\lambda = (\lambda_{-\mu}, \cdots, \lambda_m)$ such that

(a) $\lambda \cdot z \leqq 0$ for all $z \in K_1$,

(b) $\lambda \cdot z \geqq 0$ for all $z \in K_2$.

From (a) we know that for all $x \in X$ we have

$$\sum_{i = -\mu, \cdots, m} \lambda_i a_i(x) = 0,$$

which implies (i). From (b) we have immediately (ii) and (iii). This concludes the proof of the multiplier rule.

Acknowledgment. I thank the referee, Gérard Debreu and Freddy Delbaen for valuable comments on this paper.

REFERENCES

[1] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[2] C. Carathéodory, *Calculus of Variations and Partial Differential Equations of the First Order*, Holden-Day, San Francisco, 1967.

[3] F. John, *Extremum problems with inequalities as subsidiary conditions*, Studies and Essays: Courant Anniversary Volume, K. O. Friedrichs, O. E. Neugebauer and J. J. Stoker, eds., Interscience, New York, 1948, pp. 187–204.

[4] O. L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[5] H. Halkin and L. W. Neustadt, *General necessary conditions for optimization problems*, Proc. Nat. Acad. Sci. U.S.A., 56 (1966), pp. 1066–1071.

[6] H. Halkin, *Nonlinear nonconvex programming in an infinite dimensional space*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1968, pp. 10–25.

[7] ———, *On the necessary condition for optimal control of nonlinear systems*, J. Analyse Math., 12 (1964), pp. 1–82.

SIAM J. CONTROL
Vol. 12, No. 2, May 1974

ON THE APPROXIMATION OF ITO INTEGRALS USING BAND-LIMITED PROCESSES*

A. V. BALAKRISHNAN†

Abstract. Ito integrals involving observed data arise in many applications where we idealize the observation noise to be white. Since the Wiener process is not realizable, and any observed process must be smooth, the need arises for approximation in terms of a smooth process such as a band-limited process which has no frequency components outside a finite band. We show that under a sufficient condition, such an approximation is possible provided we also add suitable "correction" terms.

1. Introduction. Let $W(t;\omega)$, $-\infty < t < +\infty$, denote a standard Wiener process, say one-dimensional for simplicity, and let $\beta(t)$ denote the associated "growing" sigma algebra of sets (that is the sigma algebra generated by $W(s;\omega), s \leqq t$). Let $f(t;\omega)$ be a function jointly measurable in t and ω, measurable $\beta(t)$ for each t, and further with say (for simplicity),

$$\int_0^1 E|f(t;\omega)|^2\,dt < \infty.$$

In many problems of filtering and control, we need to evaluate the Ito integral

$$\int_0^1 f(t;\omega)\,dW(t;\omega). \tag{1.1}$$

The standard approximation is to use partial sums of the form

$$\sum f(t_i;\omega)(W(t_{i+1},\omega) - W(t_i;\omega)). \tag{1.2}$$

However, in practice, there is a serious difficulty with this procedure because, in dealing with Ito integrals with respect to observed data, one does not have a true Wiener process to work with (see the application in § 2.1 for more details on this). What one has rather is a "smooth" approximation to the Wiener process. A most convenient approximation is the "band-limited" version, by which is meant, precisely, the process

$$y(t;\omega) = \int_{-\infty}^{\infty} M(t-s)\,dW(s;\omega),$$

where

$$\int_{-\infty}^{\infty} e^{2\pi i f t} M(t)\,dt = \psi(f)$$

vanishes outside a finite interval, and is thus "limited" to the finite band

$$\psi(f) = 0, \qquad |f| > m > 0, \tag{1.3}$$

* Received by the editors April 9, 1973.

† Systems Science Department, School of Engineering and Applied Sciences, University of California at Los Angeles, Los Angeles, California 90024. This research was supported in part by the United States Air Force, Applied Mathematics Division, under Grant 73-2492.

and[1]

$$\psi(f) = 1, \qquad |f| \leqq m.$$

For any function $g(\cdot)$ in $L_2(0, 1)$ we note that $\int_0^1 g(t)y(t;\omega)\,dt$ converges as m goes to infinity with probability one and in the L_2-mean to

$$\int_0^1 g(t)\,dW(t;\omega). \tag{1.4}$$

However, the situation is quite different in the case in which $g(\cdot)$ is not deterministic. To take a classic example, consider the Ito integral

$$\int_0^1 (W(t;\omega) - W(0,\omega))\,dW(t;\omega). \tag{1.5}$$

If we approximate

$$\int_0^1 \int_0^t y(s;\omega)\,ds \quad y(t;\omega)\,dt \tag{1.6}$$

by the band-limited processes above, the limit as m goes to infinity is

$$(W(1;\omega) - W(0;\omega))^2/2, \tag{1.7}$$

while the Ito integral itself is equal to

$$(W(1;\omega) - W(0;\omega))^2/2 - 1/2. \tag{1.8}$$

In other words, (1.6) does not converge to (1.5), but it does if we add the correction term $(-\frac{1}{2})$. Wong and Zakai [1] showed that if we consider the Ito integral arising from a stochastic differential equation in the Ito sense, for example,

$$dx(t;\omega) = m(x(t;\omega))\,dt + \sigma(x(t;\omega))\,dW(t;\omega),$$

then if we take the sequence of solutions obtained by replacing $W(t;\omega)$ by a sequence of smooth processes (such as the band-limited process), we do not have convergence to the solution of the Ito equation but rather to another equation obtained by adding a correction term

$$\tfrac{1}{2}\sigma(x(t;\omega))\sigma'(x(t;\omega))\,dt,$$

analagous to (1.8). They do not, however, deal with integrals of the form (1.1) directly. McShane in his recent work[2] [2] has examined many approximations to (1.1), but they are time-domain approximations (extension of the form (1.2)), and in fact he cites the need for examining band-limited approximations in view of the negative results of Wong–Zakai. In this paper we study the problem of approximating integrals of the form (1.1) by functionals on the band-limited process, one area of application being in the calculation of likelihood functionals. We show that under a sufficiency condition on the function $f(t;\omega)$, it is possible to approximate in the desired manner, and indicate what the precise "correction"

[1] We can generalize this condition, of course, so long as (1.4) holds.

[2] See also his *Stochastic Equations and Stochastic Models*, Holt, Rhinehart and Winston, New York, to appear 1974.

terms must be. We begin with a special case first along with a direct application of it. Our general result is given in § 3.

2. The linear case. We begin first with the notationally simple case, in which $f(t;\omega)$ is a linear functional but still exhibits the essential features of the more general situation. Let $W(t;\omega)$ denote an n-by-1 standard Wiener process on $(-\infty,\infty)$. See [3] for an explicit construction. As therein, we may set $W(0;\omega)$ to be zero. Let

$$f(t;\omega) = \int_0^t L(t;s)\,dW(s;\omega), \qquad 0 \leqq t \leqq 1, \tag{2.1}$$

where $L(t;s)$ is an $n \times n$ matrix function, Lebesgue measurable in s, t, and such that[3]

$$\int_0^1 \int_0^t \|L(t;s)\|^2\,ds\,dt < \infty. \tag{2.2}$$

In particular, of course, condition (2.2) implies that

$$\int_0^1 E(\|f(t;\omega)\|^2)\,dt = \int_0^1 \int_0^t \|L(t;s)\|^2\,ds\,dt < \infty.$$

Our first result is the following theorem.

THEOREM 2.1. *Let H denote the real Hilbert space of $n \times 1$ square integrable functions $L_2((0,1),E_n)$. With $L(t;s)$ as in* (2.2), *define the linear transformation L by*

$$Lf = g; \quad g(t) = \int_0^t L(t;s)f(s)\,ds, \qquad 0 < t < 1,$$

mapping H into itself. Suppose $(L + L^*)$ is trace-class (or, "nuclear", as it is referred to in the more recent literature). *Then*

$$\begin{aligned}\eta &= \int_0^1 \left[\int_0^t L(t;s)\,dW(s;\omega), dW(t;\omega)\right] \\ &= \lim_m \int_0^1 \left[\int_0^t L(t;s)y_m(s;\omega)\,ds, y_m(t;\omega)\right] dt - \tfrac{1}{2}\operatorname{tr}(L + L^*),\end{aligned} \tag{2.3}$$

where

$$y_m(t;\omega) = \int_{-\infty}^{\infty} M(t-s)\,dW(s;\omega),$$

$$M(s) = I_n \int_{-m}^{m} e^{2\pi i f s}\,df = I_n(\sin 2\pi m s)/\pi s,$$

$$I_n = n \times n \text{ identity matrix},$$

and the limit may be taken in the L_2-mean.

Proof. We begin with a lemma (cf. [4]).

[3] $\|B\|^2 = \operatorname{tr}(BB^*)$; $[A, B] = \operatorname{tr}(AB^*)$.

LEMMA. *Let $\phi_i(\cdot)$ denote any orthonormal basis of functions in H. Then*

$$\eta = \sum_i \sum_j [L\phi_i, \phi_j]\zeta_i\zeta_j - \tfrac{1}{2}\operatorname{tr}(L + L^*), \tag{2.4}$$

where

$$\zeta_i = \int_0^1 [\phi_i(t), dW(t;\omega)],$$

and the convergence is in the L_2-mean.

Proof. See [4]. Here we shall indicate the main steps. Let $a_{ij} = [L\phi_i, \phi_j]$. Then since we are in a real Hilbert space, $\frac{1}{2}\operatorname{tr}(L + L^*) = \sum a_{ii}$, and we note that

$$\sum |a_{ii}| < \infty. \tag{2.5}$$

For any finite subdivision of [0, 1] with subdivision points t_i,

$$\{t_i\}, \quad t_0 = 0, \quad t_p = 1,$$

we note that

$$\begin{aligned}\eta_p &= \sum_{i=0}^{p-1} [f(t_i;\omega), W(t_{i+1};\omega) - W(t_i;\omega)] \\ &= \sum_i \sum_j [L_p\phi_i, \phi_j]\zeta_i\zeta_j,\end{aligned}$$

where L_p is the operator on H defined by

$$L_p f = g; \quad g(t) = \int_0^{t_i} L(t_i; s) f(s)\, ds, \qquad t_i < t < t_{i+1},$$

and we observe that L_p is trace-class with trace zero. Next we can calculate that, if we assume for the moment that $L(t;s)$ is continuous in $0 \leqq s \leqq t \leqq 1$,

$$\begin{aligned}&E\left(\left(\sum_i \sum_j ([L_p\phi_i, \phi_j] - a_{ij})\zeta_i\zeta_j + \sum_i a_{ii}\right)^2\right) \\ &\qquad\qquad \leqq 2\|L_p - L\|^2_{\text{H-S}} \to 0 \text{ as } p \to \infty,\end{aligned}$$

where H–S stands for the Hilbert–Schmidt norm. Since η_p converges in the L_2-mean to η, we obtain (2.4). Next given an arbitrary (that is, noncontinuous) kernel $L(t;s)$, we can approximate it by continuous kernels L_n such that

$$\begin{aligned}&\|L_n - L\|^2_{\text{H-S}} \to 0, \\ &\operatorname{tr}(L_n + L_n^*) \to \operatorname{tr}(L + L^*).\end{aligned}$$

Since (2.4) holds for each L_n, we can proceed to take limits on both sides to obtain the desired result for L.

Remark. In the case that $L(t;s)$ is continuous in $s \leqq t$, we know that

$$\tfrac{1}{2}\operatorname{tr}(L + L^*) = \tfrac{1}{2}\int_0^1 \operatorname{tr} L(t;t)\, dt.$$

On the other hand, even if $\int_0^1 |\operatorname{tr} L(t;t)|\, dt < \infty$, it is not necessary that $(L + L^*)$ be trace-class, and (2.4) need not hold. For example, we know that we can, by a classical construction due to Carleman, find a continuous function $h(t)$, $0 \leqq t \leqq 1$, such that $h(t)$ has the Fourier series expansion

$$h(t) = \sum_0^\infty c_k \cos 2\pi k t,$$

where $\sum_0^\infty |c_k| = +\infty$. If we now take $L(t;s) = h(t-s)$, we note that L is not trace-class, and (2.4) does not hold if we take the orthonormal basis of trigonometric functions.

Let us now return to the proof of the theorem. Let us first assume that $L(t;s)$ is continuous in $s \leqq t$. Then because $(L + L^*)$ is trace-class, we know that

$$\operatorname{tr}(L + L^*) = \int_0^1 \operatorname{tr} L(t;t)\,dt.$$

First we note that by letting (for fixed m)

$$R(t;s) = E(y_m(t;\omega)y(s;\omega)^*),$$

we have

$$R(t;s) = I_n(\sin 2\pi m(t-s))/\pi(t-s) = M(t-s). \tag{2.6}$$

Define the transformation, mapping H into itself, by

$$Rf = g; \quad g(t) = \int_0^1 R(t;s)f(s)\,ds, \qquad 0 \leqq t \leqq 1.$$

Let ϕ_i denote the orthonormalized eigenfunctions of R, and it is an easy matter to see that R is trace-class and that

$$\int_0^1 \left[\int_0^t L(t;s)y_m(s;\omega)\,ds; y_m(t;\omega)\right] dt = \sum_i \sum_i a_{ij}\zeta_i^m \zeta_j^m,$$

where

$$\zeta_i^m = \int_0^1 [y_m(t;\omega), \phi_i(t)]\,dt,$$

$$a_{ij} = [L\phi_i, \phi_j].$$

Let ζ_i be defined as before:

$$\zeta_i = \int_0^1 [\phi_i(t), dW(t;\omega)].$$

The $\phi_i(\cdot)$ being an orthonormal basis, we have, of course,

$$\sum [L\phi_k, \phi_k] = \tfrac{1}{2}\operatorname{tr}(L + L^*).$$

Next let us note that, because of the circumstance (2.6),

$$E(\zeta_i \zeta_j^m) = [R\phi_i, \phi_j],$$

as a ready calculation shows. Next let

$$\eta_m = \sum_i \sum_j a_{ij} \zeta_i^m \zeta_j^m,$$

$$\eta_0 = \sum_i \sum_j a_{ij} \zeta_i \zeta_j.$$

Denoting by λ_i the eigenvalues corresponding to ϕ_i, we can readily calculate that

$$E((\eta_m - \eta_0)^2) = (\sum a_{ii}(1 - \lambda_i))^2 + 2 \sum_i \sum_j (b_{ij})^2 (1 - \lambda_i \lambda_j),$$

where $b_{ij} = (a_{ij} + a_{ji})/2$. If we denote by $k(t;s)$ the kernel corresponding to $(L + L^*)$, we note that

$$\sum a_{ii}(1 - \lambda_i) = \tfrac{1}{2} \operatorname{tr}(L + L^*) - \int_0^1 \int_0^1 \operatorname{tr} k(t;s) R(s;t)\, ds\, dt$$

$$= \int_0^1 \operatorname{tr} L(t;t)\, dt - \int_0^1 \int_0^1 \operatorname{tr} L(t;s) \frac{\sin 2\pi m(t-s)}{\pi(t-s)}\, ds\, dt.$$

Again,

$$(b_{ij})^2 \lambda_i \lambda_j = \sum_i \sum_j [\tfrac{1}{2}(L + L^*) R \phi_i, \phi_j] \cdot [R \tfrac{1}{2}(L + L^*) \phi_i, \phi_j]$$

$$= [\tfrac{1}{2}(L + L^*)R, R\tfrac{1}{2}(L + L^*)]_{\text{H-S}}.$$

(H–S denoting the inner product in the space of H–S operators). But now it is standard analysis to show that

$$\lim_m \int_0^1 \int_0^1 \operatorname{tr} L(t;s) \frac{\sin 2\pi m(t-s)}{(t-s)}\, ds\, dt \to \int_0^1 \operatorname{tr} L(t;t)\, dt,$$

$$\lim_m [\tfrac{1}{2}(L + L^*)R, R\tfrac{1}{2}(L + L^*)]_{\text{H-S}} \to [\tfrac{1}{2}(L + L^*), \tfrac{1}{2}(L + L^*)]_{\text{H-S}} = \sum_i \sum_j b_{ij}^2.$$

Hence (2.3) has been proven for the case where the kernel $L(t;s)$ is continuous. If $L(t;s)$ is not continuous, then we can use the approximation (as in [5], for example)

$$L_h(t;s) = \frac{1}{4h^2} \int_{t-h}^{t+h} \int_{s-h}^{s+h} L(u;v)\, du\, dv,$$

apply the theorem for each sufficiently small h, then as h goes to zero exploit the fact that

$$\operatorname{tr}(L + L^*) = \lim_h \operatorname{tr}(L_h + L_h^*),$$

$$\|L_h - L\|_{\text{H-S}}^2 \to 0.$$

2.1. An application. We shall now indicate one application of Theorem 2.1, which was in fact the motivation for the present work. Consider the linear stochastic differential system

$$x(t;\omega) = \int_0^t Ax(s;\omega)\, ds + BW(t;\omega), \qquad 0 \leqq t \leqq 1, \tag{2.7}$$

$$Y(t;\omega) = \int_0^t Cx(s;\omega)\,ds + DW(t;\omega), \qquad -\infty < t < \infty, \tag{2.8}$$

$$x(t;\omega) = 0, \quad t \leqq 0, \quad DD^* = \text{identity matrix}, \quad BD^* = 0.$$

$W(t;\omega)$ is a standard Wiener process as before, with $W(0;\omega)$ zero. We know that the process $Y(t;\omega)$, $0 \leqq t \leqq 1$, induces a probability measure on the Banach space $C(0, 1)$ which is absolutely continuous with respect to the Wiener measure induced by $W(t;\omega)$, $0 \leqq t \leqq 1$. Moreover the R–N derivative is given by

$$\exp -\frac{1}{2}\left\{\int_0^1 [Cx(s;\omega), Cx(s;\omega)]\,ds - 2\int_0^1 [Cx(s;\omega), dY(s;\omega)]\right\}, \tag{2.9}$$

where

$$x(t;\omega) = \int_0^t \phi(t)\phi(s)^{-1}P(s)C^*\,dY(s;\omega),$$

where $\phi(t)$ is a fundamental matrix solution of

$$\dot{\phi}(t) = (A - P(t)C^*C)\phi(t),$$

and $P(t)$ is the nonnegative definite solution of

$$\dot{P}(t) = AP(t) + P(t)A^* + BB^* - P(t)C^*CP(t), \qquad P(0) = 0.$$

Unfortunately, what is observed in practice is not (2.8), but a band-limited version, albeit of large enough bandwidth to allow the use (in theory) of (2.8). The main question, then, is the approximation of the Ito integral in (2.9). Here we can use Theorem 2.1 to state the following.

THEOREM 2.2. *Let $M(\cdot)$ be as in Theorem 2.1 and define*

$$y(t;\omega) = \int_{-\infty}^{\infty} M(t-s)\,dY(s;\omega), \qquad 0 \leqq t \leqq 1.$$

Then the Ito integral in (2.9) can be approximated

$$\begin{aligned}\int_0^1 [Cx(s;\omega), dY(s;\omega)] &= \lim \int_0^1 \left[\int_0^t CL(t;s)y(s;\omega)\,ds, y(t;\omega)\right] dt \\ &\quad - \int_0^1 \operatorname{tr} CP(t)C^*\,dt,\end{aligned} \tag{2.10}$$

where

$$L(t;s) = \phi(t)\phi(s)^{-1}P(s)C^*.$$

Proof. We note, first of all, that we can write

$$y(t;\omega) = Cx_m(t;\omega) + z(t;\omega),$$

$$Cx_m(t;\omega) = \int_{-\infty}^{\infty} M(t-s)Cx(s;\omega)\,ds = \int_0^1 M(t-s)Cx(s;\omega)\,ds,$$

$$z(t;\omega) = \int_{-\infty}^{\infty} M(t-s)D\,dW(s;\omega).$$

We note that using Theorem 2.1 and the fact the operator

$$Lf = g; \quad g(t) = \int_0^t CL(t; s)f(s)\,ds, \qquad 0 \leqq t \leqq 1,$$

is such that $(L + L^*)$ is trace-class (see [4] for a proof if necessary), we have

$$\int_0^1 \left[\int_0^t CL(t; s)z(s; \omega)\,ds, z(t; \omega)\right] dt \to \int_0^1 \left[\int_0^t CL(t; s)D\,dW(s; \omega), D\,dW(t; \omega)\right] - \int_0^1 \operatorname{tr} CP(t)C^*\,dt.$$

The theorem is thus proved if we can show that

$$\int_0^1 \left[\int_0^t CL(t; s)z(s; \omega)\,ds, z(t; \omega)\right] dt \to \int_0^1 \left[\int_0^t CL(t; s)Cx(s; \omega), D\,dW(t; \omega)\right] \tag{2.11}$$

and that

$$\int_0^1 \left[\int_0^t CL(t; s)Cx_m(s; \omega), Cx_m(t; \omega)\right] dt \to \int_0^1 \left[\int_0^t CL(t; s)Cx(s; \omega)\,ds, Cx(t; \omega)\right] dt. \tag{2.12}$$

Because random variables are involved, we shall proceed to prove this in some detail. Let

$$\psi(f; \omega) = \int_0^1 e^{2\pi i f s} x(s; \omega)\,ds.$$

Then

$$Cx_m(t; \omega) = \int_{-m}^m e^{2\pi i f t} \psi(f; \omega)\,df.$$

Since $x(t; \omega)$ is continuous in t, omitting a set of measure zero, we note that

$$\int_0^1 \|Cx_m(t; \omega) - Cx(t; \omega)\|^2\,dt \leqq \int_{|f| > m} \|\psi(f; \omega)\|^2\,df,$$

and since

$$\int_0^1 \|Cx_m(t; \omega) - Cx(t; \omega)\|^2\,dt \leqq \int_{|f| > m} \|\psi(t; \omega)\|^2\,df,$$

it follows that

$$\int_0^1 E(\|Cx_m(t; \omega) - Cx(t; \omega)\|^2)\,dt \to 0 \quad \text{as } m \to \infty.$$

This is clearly enough to establish (2.12). To handle (2.11), it is convenient first to integrate by parts. Thus,

$$\int_0^1 \left[\int_0^t CL(t;s)Cx(s;\omega), dW(t;\omega)\right]$$

$$= \int_0^1 \left[\int_0^t \phi(s)^{-1}P(s)C^*Cx(s;\omega)\,ds, \phi(t)^*C^*D\,dW(t;\omega)\right]$$

$$= \left[\int_0^1 \phi(s)^{-1}P(s)C^*Cx(s;\omega)\,ds, \int_0^1 \phi(s)^*C^*D\,dW(s;\omega)\right]$$

$$- \int_0^1 \left[\phi(t)^{-1}P(t)C^*Cx(t;\omega), \int_0^t \phi(s)^*C^*D\,dW(s;\omega)\right].$$

We perform a similar integration by parts on the left-hand member of (2.11), and establish the necessary convergence term by term. For $0 \leqq t \leqq 1$, let

$$\eta(t;\omega) = \int_0^t \phi(d)^*C^*z(s;\omega)\,ds - \int_0^t \phi(s)^*C^*D\,dW(s;\omega)$$

$$= \int_{-\infty}^{\infty} (h_m(s) - h(s))D\,dW(s;\omega),$$

where

$$h(s) = \begin{cases} \phi(s)^*C^*, & 0 \leqq s \leqq t, \\ 0 & \text{otherwise}, \end{cases}$$

$$h_m(s) = \int_0^t M(s - \sigma)h(\sigma)\,d\sigma.$$

Then

$$E(\|\eta(t;\omega)\|^2) = \int_{-\infty}^{\infty} \|h_m(s) - h(s)\|^2\,ds,$$

and the integral on the right goes to zero as m goes to infinity. With this additional estimate, we can see that (2.11) follows.

3. Generalization. Let us now go on to consider the general case. In order to avoid notational complication, we shall restrict ourselves to the case, in which $W(t;\omega)$, the standard Wiener process, is one-dimensional. The extension to the multidimensional case can be made using either polynomials as in [6] or tensor-product Hilbert spaces as in [7].

Suppose, then, we are given an Ito integral of the form $\int_0^1 f(t;\omega)\,dW(t;\omega)$, where $\int_0^1 E(|f(t;\omega)|^2)\,dt < \infty$. Note that $\eta(\omega) = \int_0^1 f(t;\omega)\,dW(t;\omega)$ defines a measurable, square integrable functional on the Wiener process $W(t;\omega), 0 < t < 1$, and as Ito has shown in [8], it can be approximated by sums of the form

$$\sum_1^n \int_0^1 \cdots \int_0^1 K_p(t_1, \cdots, t_p)\,dW(t_1;\omega) \cdots dW(t_p;\omega),$$

where each term is an Ito multiple integral. For example, the "linear" case of § 2,

$$\int_0^1 \left(\int_0^t L(t; s)\, dW(s; \omega) \right) dW(t; \omega),$$

can be expressed using the convention that $L(t; s) = 0$ for $s > t$:

$$\int_0^1 \int_0^1 \frac{L(t; s) + L(s; t)}{2}\, dW(s; \omega)\, dW(t; \omega).$$

Hence our main result can be stated as follows.

THEOREM 3.1. *Let $K(t_1, \cdots, t_p)$ be a continuous symmetric (real-valued) function on $0 \leqq t_i \leqq 1$, $i = 1, \cdots, p$. For each integer v, $2v \leqq p$, and each fixed function $h(t_1, \cdots, t_{p-2v})$ continuous in $0 \leqq t_i \leqq 1$, $i = 1, \cdots, p - 2v$, define the operator L_v by*

$$L_v f = g;$$

$$g(t_1, \cdots, t_v) = \int_0^1 \cdots \int_0^1 K(t_1, \cdots, t_v, s_1, \cdots, s_v, \tau_{2v+1}, \cdots, \tau_p)$$
$$h(\tau_{2v+1}, \cdots, \tau_p)\, d\tau_{2v+1} \cdots d\tau_p f(s_1, \cdots, s_v)\, ds_1 \cdots ds_v$$

mapping $L_2((0, 1)^v)$ into itself. Suppose L_v is trace-class for each v and each arbitrary chosen $h(\cdots)$. Then the Ito integral

$$\int_0^1 \cdots \int_0^1 K(t_1, \cdots, t_p)\, dW(t_1; \omega) \cdots dW(t_p; \omega)$$

$$= \lim \sum_{v=0}^{[p/2]} \frac{p!(-1)^v}{(p - 2v)!2^v v!} \int_0^1 \cdots \int_0^1 K(\sigma_1, \sigma_1, \sigma_2, \cdots, \sigma_v, \sigma_v, t_{2v+1}, \cdots, t_p) \tag{3.1}$$
$$\cdot d\sigma_1 \cdots d\sigma_v y(t_{2v+1}; \omega) \cdots y(t_p; \omega)\, dt_{2v+1} \cdots dt_p,$$

where

$$y(t; \omega) = \int_{-\infty}^{\infty} \frac{(\sin 2\pi(t - s))}{\pi(t - s)} \cdot dW(s; \omega),$$

the limit being taken in the L_2-norm, and $[c]$ denotes the largest integer $\leqq c$.

Proof. To clarify the notation in (3.1), let us look at (3.1) for the case $p = 2$. We have

$$\int_0^1 \int_0^1 K(t_1, t_2)\, dW(t_1; \omega)\, dW(t_2; \omega)$$
$$= \lim \int_0^1 \int_0^1 K(t_1, t_2) y(t_1; \omega) y(t_2; \omega)\, dt_t\, dt_2 - \int_0^1 K(s, s)\, ds.$$

But this is a special case of Theorem 2.1. For, the Ito double integral is given by the sum of the integrated integrals,

$$\int_0^1 \int_0^t K(t; s)\, dW(s; \omega)\, dW(t; \omega) + \int_0^1 \int_0^t K(s; t)\, dW(s; \omega)\, dW(t; \omega),$$

which, since the kernel is symmetric, is equal to

$$2\int_0^1 \int_0^t K(t; s)\, dW(s; \omega)\, dW(t; \omega),$$

and this in turn, by Theorem 2.1,

$$\begin{aligned} &= \lim 2 \int_0^1 \int_0^t K(t; s) y(s; \omega)\, ds\, y(t; \omega)\, dt - \int_0^1 K(t; t)\, dt \\ &= \lim \int_0^1 \int_0^1 K(t; s) y(s; \omega) y(t; \omega)\, ds\, dt - \int_0^1 K(t; t)\, dt. \end{aligned}$$

Illustrating the case when p is odd, let us calculate (3.1) for $p = 3$:

$$\begin{aligned} &\int_0^1 \int_0^1 \int_0^1 K(t_1, t_2, t_3)\, dW(t_1, \omega)\, dW(t_2; \omega)\, dW(t_3; \omega) \\ &\quad = \lim \int_0^1 \int_0^1 \int_0^1 K(t_1, t_2, t_3) y(t_1; \omega) y(t_2; \omega) y(t_3; \omega)\, dt_1\, dt_2\, dt_3 \\ &\quad\quad - 3 \int_0^1 \int_0^1 K(s; s; t)\, ds\, y(t; \omega)\, dt. \end{aligned}$$

The main tool we shall use in the proof is the decomposition formula for multiple Ito integrals. Let $\phi_i(\cdot)$ denote an orthonormal basis in $L_2(0, 1)$. Let us use the notation $I_n(K(t_1, \cdots, t_n))$ for the associated Ito multiple integral. Then we have, using the Ito decomposition formula (cf. [8]),

$$\begin{aligned} I_p(\phi_{i_1}(t_1)\phi_{i_2}(t_2) \cdots \phi_{i_p}(t_p)) &= I_{p-1}(\phi_{i_1}(t_1) \cdots \phi_{i_{p-1}}(t_{p-1}))\zeta_{i_p}(\omega) \\ &\quad - \sum_{k=1}^{p-1} I_{p-2}(\phi_{i_1}(t_1) \cdots \phi_{i_{k-1}}(t_{k-1})\phi_{i_{k+1}} \cdots \phi_{i_{p-1}}(t_{p-1}))[\phi_{i_k}, \phi_{i_p}], \end{aligned} \tag{3.2}$$

where $\zeta_j(\omega) = \int_0^1 \phi_j(t)\, dW(t; \omega)$. Next let us note that $\phi_{i_1}(t_1) \cdots \phi_{i_v}(t_v)$ is an orthonormal basis for $L_2(0, 1)^v)$, $v \leqq p$, and in particular,

$$K(t_1, \cdots, t_p) = \sum_{i_1} \cdots \sum_{i_p} a_{i_1 i_2 \ldots i_p} \phi_{i_1}(t_1) \cdots \phi_{i_p}(t_p), \tag{3.3}$$

the series converging in $L_2((0, 1)^p)$, where the Fourier coefficients are also symmetric in the variables. Because of the trace-class condition on the operators L_v, it is readily seen that

$$\sum_{i_1} \cdots \sum_{i_v} |a_{i_1 i_1 i_2 i_2 \ldots i_v i_v i_{2v+1} \ldots i_p}| < \infty \tag{3.4}$$

for each fixed set of indices $i_{2v+1}, \cdots, i_p$, and every v, $2v \leqq p$. Because of (3.3), we have that the Ito integral in (3.1) is the limit in the L_2-norm of the series

$$\sum_{i_1} \cdots \sum_{i_p} a_{i_1 i_2 i_3 \ldots i_p} I_p(\phi_{i_1}(t_1) \cdots \phi_{i_p}(t_p)). \tag{3.5}$$

This series we shall now show is expressible as

$$\sum_{\nu=0}^{[p/2]} \frac{p!(-1)^\nu}{(p-2\nu)!2^\nu \nu!} \sum_{i_1} \cdots \sum_{i_\nu} \sum_{i_{2\nu+1}} \cdots \sum_{i_p} a_{i_1 i_1 i_2 i_2 \ldots i_{2\nu+1} \ldots i_p} \cdot \zeta_{i_{2\nu+1}}(\omega) \cdots \zeta_{i_p}(\omega) \cdots . \tag{3.6}$$

We can prove this by induction using (3.2). (The coefficients are obtained by the same combinatorial argument as in Wiener [9]). Thus, substituting in (3.5) from (3.2), and assuming the result true for integers less than p, we have

$$\sum_{i_p} \cdots \sum_{i_p} a_{i_1 \ldots i_p} I_p(\phi_{i_1}(t_1) \cdots \phi_{i_p}(t_p))$$
$$= \sum_{0}^{[(p-1)/2]} \frac{(p-1)!(-1)^\nu}{(p-1-2\nu)!2^\nu \nu!} \sum_{i_1} \cdots \sum_{i_\nu} \cdots \sum_{i_p} a_{i_1 i_1 \ldots i_\nu i_\nu i_{2\nu+1} \ldots i_p} \cdot \zeta_{2\nu+1}(\omega) \cdots \zeta_p(\omega)$$
$$- (p-1) \sum_{0}^{[(p-2)/2]} \frac{(p-2)!(-1)^\nu}{(p-2-2\nu)!2^\nu \nu!} \sum_{i_1} \cdots \sum_{j_1} \cdots \sum_{j_{p-2}} a_{i_1 i_1 j_1 j_1 \ldots j_\nu j_\nu j_{2\nu+1} \ldots j_{p-2}} \zeta_{2\nu+1}(\omega) \cdots \zeta_{p-2}(\omega).$$

And combining the two sums, taking the ν sum in the first term with the $(\nu - 1)$ sum in the second term for $\nu \geqq 1$, and noting that

$$\frac{(p-1)!}{(p-1-2\nu)!2^\nu \nu!} + \frac{(p-1)(p-2)!}{(p-2-2(\nu-1))!2^{\nu-1}(\nu-1)!} = \frac{p!}{(p-2\nu)!2^\nu \nu!}$$

we have (3.6). Since we have already proved the result for $p = 2$, the induction is complete. Next for each m, let ϕ_i be the orthonormalized eigenfunctions of the operator R defined by

$$Rf = g; \quad g(t) = \int_0^1 (\sin 2\pi(t-s))/\pi(t-s)\, ds$$

mapping $L_2(0, 1)$ into itself, and letting

$$\zeta_i^m(\omega) = \int_0^1 \phi_i(t) y(t; \omega)\, dt,$$

we note that

$$\int_0^1 \cdots \int_0^1 K(t_1, t_1, \cdots, t_\nu, t_\nu, t_{2\nu+1}, \cdots, t_p) \cdot \phi_{i_{2\nu+1}}(t_{2\nu+1}) \cdots \phi_{i_p}(t_p)\, dt_1 \cdots dt_\nu\, dt_{2\nu+1} \cdots dt_p$$
$$= \sum_{i_1} \cdots \sum_{i_\nu} a_{i_1 i_1 \ldots i_\nu i_\nu i_{2\nu+1} \ldots i_p},$$

by virtue of the trace-class assumption of L_ν. Hence we can readily see that for fixed m, the right-hand side of (3.1) is given by:

$$\sum_{0}^{[p/2]} \frac{p!(-1)^\nu}{(p-2\nu)!2^\nu \nu!} \sum_{i_1} \cdots \sum_{i_\nu} \sum_{i_{2\nu+1}} \cdots \sum_{i_p} a_{i_1 i_1 \ldots i_\nu i_\nu i_{2\nu+1} \ldots i_p} \zeta_{i_{2\nu+1}}^m(\omega) \cdots \zeta_{i_p}^m(\omega) \cdots , \tag{3.7}$$

which is the same as (3.6) except for replacing $\zeta_i(\omega)$ by $\zeta_i^m(\omega)$. It only remains to show that as m goes to infinity, (3.7) converges to (3.6), to conclude the proof of (3.1). Here we can again exploit the fact that $E(\zeta_i^m(\omega)\zeta_j(\omega) = [R\phi_i, \phi_j]$. The proof is similar to the one we used in Theorem 2.1, only more tedious. Let us first note that the difference of the $v = 0$ sums in (3.6) and (3.7) is

$$\sum_{i_1} \cdots \sum_{i_p} a_{i_1 i_2 \ldots i_p}(\zeta_{i_1}\zeta_{i_2} \cdots \zeta_{i_p} - \zeta_{i_1}^m \zeta_{i_2}^m \cdots \zeta_{i_p}^m),$$

and we can readily calculate that the expected value of the square of this is given by

$$\begin{aligned} & p! \sum_{i_1} \cdots \sum_{i_p} (a_{i_1 \ldots i_p})^2 (1 - \lambda_{i_1}\lambda_{i_2} \cdots \lambda_{i_p}) \\ (3.8) \qquad & + \sum_{1}^{[p/2]} \frac{p!}{2^v v!} \sum_{i_{2v+1}} \cdots \sum_{i_p} \left(\sum_{i_1} \cdots \sum_{i_v} a_{i_1 i_1 \ldots i_v i_v i_{2v+1} \ldots i_p} \lambda_{i_1}\lambda_{i_2} \cdots \lambda_{i_v} \right)^2 \lambda_{i_{2+1}} \cdots \lambda_{i_p} \\ & - \sum_{1}^{[p/2]} \frac{p!}{2^v v!} \sum_{i_{2v+1}} \cdots \sum_{i_p} \left(\sum_{i_1} \cdots \sum_{i_v} a_{i_1 i_1 \ldots i_v i_v i_{2v+1} \ldots i_p} \right)^2 . \end{aligned}$$

We can now proceed by induction. If we assume the result to be true for integers less than p, then to prove it for p we only need to show that (3.8) goes to zero. But this is readily done in a manner which is analogous to the case $p = 2$. Thus the first term in (3.8) can be expressed

$$(3.9) \qquad p!([K_p, K_p] - [R_p K_p, K_p]),$$

where we denote by R_v the operator

$$R_v f = g; \quad g(t_1, \cdots, t_v) = \int_0^1 \cdots \int_0^1 \frac{\sin 2\pi m(t_1 - s_1)}{\pi(t_1 - s_1)}$$
$$\cdots \frac{\sin 2\pi m(t_v - s_v)}{\pi(t_v - s_v)} f(s_1 \cdots s_v)\, ds_1 \cdots ds_v$$

mapping $L_2((0, 1)^v)$ into itself. By K_p we mean the function $K(t_1, \cdots, t_p)$ as an element of $L_2((0, 1)^p)$. Clearly (3.9) goes to zero as m goes to infinity. Next let us look at the second term. Note that we can write

$$\sum_{i_1} \cdots \sum_{i_v} a_{i_1 i_1 \ldots i_v i_v i_{2v+1} \ldots i_p} \lambda_{i_1} \cdots \lambda_{i_v} = \operatorname{tr} R_v L_v,$$

where L_v is defined by

$$L_v f = g; \quad g(t_1 \cdots t_v)$$
$$= \int_0^1 \cdots \int_0^1 \int_0^1 \cdots \int_0^1 K(t_1, \cdots, t_v, s_1, \cdots, s_v, \sigma_{2v+1}, \cdots, \sigma_p) \phi_{i_{2v+1}}(\sigma_{2v+1})$$
$$\cdots \phi_{i_p}(\sigma_p)\, d\sigma_{2v+1} \cdots d\sigma_p f(s_1, \cdots, s_v)\, ds_1 \cdots ds_v$$

mapping $L_2((0, 1)^\nu)$ into itself. And hence

$$\sum_{i_{2\nu+1}} \cdots \sum_{i_p} \left(\sum_{i_1} \cdots \sum_{i_\nu} a_{i_1 i_1 \dots i_\nu i_\nu i_{2\nu+1} \dots i_p} \lambda_{i_1} \cdots \lambda_{i_\nu} \right)^2 \lambda_{i_{2\nu+1}} \cdots \lambda_{i_p}$$

$$= \int_0^1 \cdots \int_0^1 \left(\int_0^1 \cdots \int_0^1 \frac{\sin 2\pi m(t_1 - s_1)}{\pi(t_1 - s_1)} \cdots \frac{\sin 2\pi m(t_\nu - s_\nu)}{\pi(t_\nu - s_\nu)} \right.$$

$$\left. \cdot K(s_1, \cdots, s_\nu, t_1, \cdots, t_\nu, \sigma_{2\nu+1}, \cdots, \sigma_p) \, ds_1 \cdots ds_\nu \cdots dt_1 \cdots dt_\nu \right)^2$$

$$\cdot d\sigma_{2\nu+1} \cdots d\sigma_p,$$

which, as m goes to infinity, clearly goes to

$$\int_0^1 \cdots \int_0^1 \left(\int_0^1 \cdots \int_0^1 K(t_1, \cdots, t_\nu, t_1, \cdots, t_\nu, \sigma_{2\nu+1}, \cdots, \sigma_p) \, dt_1 \cdots dt_\nu \right)^2$$

$$\cdot d\sigma_{2\nu+1} \cdots d\sigma_p$$

$$= \sum_{i_{2\nu+1}} \cdots \sum_{i_p} \left(\sum_{i_1} \cdots \sum_{i_\nu} a_{i_1 i_1 \dots i_\nu i_\nu i_{2\nu+1} \dots i_p} \right)^2 .$$

Hence (3.8) goes to zero, thus concluding the proof of the theorem.

Finally let us remove the condition of continuity on the kernel.

COROLLARY. *Suppose $K(t_1, \cdots, t_p)$ is symmetric and $\in L_2((0, 1)^p)$. Suppose further that the operator L_ν defined by*

$$L_\nu f = g; \quad g(t_1, \cdots, t_\nu) = \int_0^1 \cdots \int_0^1 K(t_1, \cdots, t_\nu, s_1, \cdots, s_\nu, \sigma_{2\nu+1}, \cdots, \sigma_p)$$

$$\cdot d\sigma_{2\nu+1} \cdots d\sigma_p \cdot f(s_1, \cdots, s_\nu) \, ds_1 \cdots ds_\nu$$

mapping $L_2[(0, 1)^\nu]$ into itself is trace-class a.e., in the variables σ_i, $2\nu + 1 \leqq i \leqq p$, and

$$\int_0^1 \cdots \int_0^1 (\operatorname{tr} L_\nu(\sigma_{2\nu+1}, \cdots, \sigma_p))^2 \, d\sigma_{2\nu+1} \cdots d\sigma_p < \infty$$

for each ν, $2\nu \leqq p$. Then the Ito integral

$$\int_0^1 \cdots \int_0^1 K(t_1, \cdots, t_p) \, dW(t_1, \omega) \cdots dW(t_p, \omega)$$

$$(3.10) \qquad = \lim \sum_{\nu=0}^{[p/2]} \frac{p!(-1)^\nu}{(p - 2\nu)! 2^\nu \nu!} \int_0^1 \cdots \int_0^1 \operatorname{tr} L_\nu(t_{2\nu+1}, \cdots, t_p) y(t_{2\nu+1}, \omega)$$

$$\cdots y(t_p, \omega) \, dt_{2\nu+1} \cdots dt_p .$$

Proof. For each h, the kernel

$$K_h(t_1, \cdots, t_p) = (1/(2h)^p) \int_{t_1 - h}^{t_1 + h} \cdots \int_{t_p - h}^{t_p + h} K(s_1, \cdots, s_p) \, ds_1, \cdots, ds_p$$

satisfies the conditions of the theorem, and hence we can obtain (3.1) for $K_h(\cdot\cdot\cdot)$. Because of the trace-class conditions imposed on $K(\cdot\cdot\cdot)$, we may proceed to take limits on both sides and obtain (3.10) as required.

REFERENCES

[1] E. WONG AND M. ZAKAI, *On the relation between ordinary integrals and stochastic integrals*, Internat. J. Engrg. Sci., 3 (1965), pp. 213–229.

[2] E. MCSHANE, *Stochastic functional equations: Continuity properties and relation to ordinary equations*, Calculus of Variations and Control Theory, Academic Press, New York, 1969.

[3] E. P. MCKEAN, *Stochastic Inegrals*, Academic Press, New York, 1969.

[4] A. V. BALAKRISHNAN, *Stochastic Differential Systems*, Springer-Verlag, Berlin, 1973.

[5] I. C. GOHBERG AND M. KREIN, *Introduction to the Theory of Non-Selfadjoint Operators*, vol. 18, American Mathematical Society Translations, Providence, R.I., 1969.

[6] A. V. BALAKRISHNAN, *Stochastic control—a function space approach*, this Journal, 10 (1972), pp. 285–297.

[7] G. KALLIANPUR, *The role of reproducing kernel Hilbert spaces in the study of Gaussian processes*, Advances in Probability, Marcel-Dekker, New York, 1972.

[8] K. ITO, *Multiple Wiener integral*, Journal Math. Soc. Japan, 3 (1951), no. 1, pp. 157–169.

[9] N. WIENER, *Nonlinear Problems in Random Theory*, Technology Press, Mass. Inst. of Tech., Cambridge, Mass., 1958.

SIAM J. CONTROL
Vol. 12, No. 2, May 1974

ON LOCAL AND GLOBAL CONTROLLABILITY*

H. HERMES†

Abstract. Computable sufficient conditions to determine local controllability along a reference trajectory are developed both by considering controllability of the linear variational equation and by direct use of differential geometric techniques with special types of control perturbations. The equivalence of the results obtained by the two methods is shown.

A collection, $\mathscr{S}$, of smooth vector fields on a manifold M is said to be *controllable on* M if every pair of points of M can be connected by a solution of $\mathscr{S}$. If the set of points attainable by solutions of $\mathscr{S}$ from every point $x \in M$ has nonempty interior, $\mathscr{S}$ is said to have the *accessibility property*. Jurdjevic has posed the problem of whether every family $\mathscr{S}$ of analytic vector fields on a connected analytic manifold M which has the accessibility property is controllable on M. We give a counterexample on the two-torus. It is next shown that every commuting two-field on the two-torus is controllable. We also show that any n-manifold admits a collection $\mathscr{S}$ of $2n$ smooth (C^∞) vector fields with the accessibility property, but such that $\mathscr{S}$ is not controllable on M.

1. Introduction. Let M be a smooth (C^∞) n-dimensional manifold with tangent space at x denoted TM_x. For each α in an index set A, let X^α denote a smooth tangent vector field on M; let $\mathscr{S} = \{X^\alpha : \alpha \in A\}$ and $\mathscr{S}_x = \{X^\alpha(x) \in TM_x : \alpha \in A\}$. A *solution* of the collection of vector fields $\mathscr{S}$ is an absolutely continuous map $\varphi : [0, \beta] \to M, \beta > 0$, such that $d\varphi(t)/dt \in \mathscr{S}_{\varphi(t)}$ almost everywhere. For simplicity of exposition, we shall assume that solutions exist for all $t \geqq 0$. Note that we do *not* allow $t < 0$.

We begin with the problem of local controllability along a reference trajectory. Specifically, let $\varphi(\cdot\,; p^0)$ denote a solution of $\mathscr{S}$ with $\varphi(0; p^0) = p \in M$. The system $\mathscr{S}$ is *locally controllable along* φ *at time* $t_2 \geqq 0$ if all points in some n-dimensional neighborhood of $\varphi(t_2; p^0)$ can be attained at time t_2 by solutions of $\mathscr{S}$ initiating from p^0. The computable sufficient conditions which we obtain for $\mathscr{S}$ to be locally controllable along a reference trajectory are local in nature, hence there is no loss of generality in assuming $M = \mathbb{R}^n$. Here we shall consider the collection of differential equations associated with $\mathscr{S}$ to have one of the following control representations (X, Y, f are C^∞ functions with ranges in $\mathbb{R}^n$; $\dot{x}$ denotes dx/dt):

$$\dot{x} = X(x) + Y(x)u, \qquad -1 \leqq u \leqq 1, \tag{1.1}$$

$$\dot{x} = X(x) + \sum_{j=1}^{s} Y^j(x)u_j, \qquad -1 \leqq u_j \leqq 1 \tag{1.2}$$

$$\dot{x} = f(x, u), \qquad u \in U \subset \mathbb{R}^s, \quad \text{interior } U \neq \varnothing. \tag{1.3}$$

We shall give sufficient conditions for the system considered to be locally controllable at some time t_2 along a reference trajectory φ generated by a control u which is piecewise constant and takes values in the interior of its set of admissible values. Many sufficient conditions can be found throughout the literature of the

* Received by the editors March 22, 1973, and in revised form June 30, 1973. This research was supported by the National Science Foundation under Grant GP27957.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80302.

last several years. One of the goals of this section will be to unify and show the relations among the various approaches taken on this problem.

For the question of global controllability, we follow Sussmann and Jurdjevic [1], and define: $\mathscr{S}$ *is controllable on* M if every pair of points of M can be joined by a solution of $\mathscr{S}$; $\mathscr{S}$ has the *accessibility property* if the set of points attainable from each point $x \in M$, by solutions of $\mathscr{S}$, has nonempty interior. In [1], computable criteria are given to determine when a collection of analytic vector fields on an analytic manifold M has the accessibility property. Jurdjevic [2] shows that if $\mathscr{S}$ is a collection of right-invariant, analytic vector fields, with the accessibility property, on a compact, connected Lie group M, then $\mathscr{S}$ is controllable. The problem is posed, in [2], as to whether the accessibility property always implies $\mathscr{S}$ is controllable on a compact, connected manifold M. We give a counterexample with M the two-torus, T^2. On the other hand, we show that every pair of linearly independent, *commuting* (under the Lie product) vector fields on T^2 is controllable. We next show that any n-manifold M carries a collection, $\mathscr{S}$, of $2n$ tangent vector fields $X^1, \cdots, X^{2n}$ such that $\mathscr{S}$ has the accessibility property (indeed we even have dim span $\{X^1(x), \cdots, X^{2n}(x)\} = n$ for all $x \in M$), yet $\mathscr{S}$ is not controllable on M.

2. Local controllability along a reference trajectory. Let $\varphi(t; p^0)$ denote a reference trajectory for one of the systems (1.1)–(1.3). If $t_2 > t_1$, the map $\varphi(t_2, \cdot)$, considered as a map of initial data, carries a neighborhood of $\varphi(t_1; p^0)$ homeomorphically onto a neighborhood of $\varphi(t_2; p^0)$. Thus if a system is locally controllable along φ at time t_1, this is also true for any $t_2 \geqq t_1$.

2.1. Controllability of the variational equation. Let $u \equiv 0$ generate the reference trajectory φ of equation (1.1), and let $X_x(x)$ denote the $n \times n$ Jacobian matrix of partial derivatives of the vector function X. The linear variational equation along φ is

$$\dot{z}(t) = X_x(\varphi(t))z(t) + Y(\varphi(t))u(t). \tag{2.1}$$

If this equation is controllable at some time $t_1 > 0$ (see [3, § 19]), it easily follows from the implicit function theorem that the system (1.1) is locally controllable along φ at time t_1. For notational simplicity, let $A(t) = X_x(\varphi(t))$ and $B(t) = Y(\varphi(t))$. Then A and B are smooth matrix-valued functions and [3, Thm. 19.3] can be applied as follows. Let $\Gamma = d/dt - A(t)$. If there exists a positive integer k and a $t_1 \geqq 0$ such that

$$\operatorname{rank}\,[B(t_1), (\Gamma B)(t_1), \cdots, (\Gamma^k B)(t_1)] = n,$$

then the linear system (2.1) is controllable at t_1.

For the special form of $A(t)$, $B(t)$ as above, we compute

$$(\Gamma B)(t_1) = \left[\frac{d}{dt} Y(\varphi(t)) - X_x(\varphi(t))Y(\varphi(t))\right]_{t=t_1} = -[X, Y](\varphi(t_1)), \tag{2.2}$$

where $[X, Y]$ denotes the Lie bracket product. Define the operator ad as follows: $(\mathrm{ad}^0\, X, Y)(p^1) = Y(p^1)$, $(\mathrm{ad}\, X, Y)(p^1) = [X, Y](p^1)$ and inductively $(\mathrm{ad}^k\, X, Y)(p^1) = [X, (\mathrm{ad}^{k-1}\, X, Y)](p^1)$. Then, inductively from (2.2), $\Gamma(\Gamma B)(t_1) = (\mathrm{ad}^2\, X, Y)(\varphi(t_1))$, etc., and Theorem 19.3 of [3] applied to equation (2.1) yields the following.

PROPOSITION 1. *A sufficient condition that the variational equation* (2.1) *be controllable at time* $t_2 > 0$ *is that there exist a* $t_1 \in (0, t_2]$ *and integer* k *such that*

$$\text{dim span}\,\{Y(\varphi(t_1)), (\text{ad}\, X, Y)(\varphi(t_1)), \cdots, (\text{ad}^k X, Y)(\varphi(t_1))\} = n. \tag{2.3}$$

It follows that (2.3) provides a sufficient condition for system (1.1) to be locally controllable along the solution φ generated by $u \equiv 0$, at time $t_2 > 0$.

Remark 1. Theorems 19.2 and 19.4 of [3] give necessary and sufficient conditions for controllability of the variational equation (2.1). These do not, however, yield necessary conditions for controllability of the system (1.1) along φ. Indeed the system (1.1) may be locally controllable along φ at t_1, yet the variational equation (2.1) is not controllable at t_1. For an example of this, see [4, Ex. 2.2].

2.2. Differential geometric methods of Hermann and Krener. For notational convenience, we shall use either $\psi(t; p)$ or $S^t(p)$ to denote the solution, at time t, of $\dot{x} = X(x) + Y(x)$, $x(0) = p$, and either $\varphi(t; p)$ or $T^t(p)$ to denote the solution of $\dot{x} = X(x), x(0) = p$. Then ψ corresponds to $u \equiv 1$ in (1.1) and φ to $u \equiv 0$. Again, φ will be the reference trajectory.

For the moment, assume that we may proceed either forwards or backwards in time along a trajectory. This will be corrected later. Let $t_1 > 0$ and $\varphi(t_1; p^0) = p^1$. For real σ, r in a neighborhood of zero, define

$$q(r; \sigma) = T^\sigma S^r T^{t_1 - \sigma - r}(p^0).$$

Then $q(0; \sigma) \equiv p^1$ for all σ, and

$$\frac{\partial q}{\partial r}(0; \sigma) = D\varphi(\sigma; \varphi(-\sigma, p^1))Y(\varphi(-\sigma; p^1)), \tag{2.4}$$

where D denotes the differential. Geometrically, this equation tells how the variational equation along φ, which has fundamental solution $D\varphi$, carries the vector field Y forward to time t_1 which is $\sigma = 0$. Indeed, consider the situation on a manifold M with tangent space at q denoted TM_q rather than on $\mathbb{R}^n$. Then

$$D\varphi(\sigma; q): TM_q \to TM_{\varphi(\sigma; q)},$$

thus

$$D\varphi(\sigma, \varphi(-\sigma, p^1)): TM_{\varphi(-\sigma; p^1)} \to TM_{p^1}.$$

Now

$$[X, Y](p^1) = \lim_{t \to 0} \frac{1}{t}[D\varphi(-t, \varphi(t; p^1))Y(\varphi(t; p^1)) - Y(p^1)],$$

hence

$$\frac{d}{d\sigma}[D\varphi(\sigma; \varphi(-\sigma; p^1))Y(\varphi(-\sigma; p^1))]_{\sigma = 0} = -[X, Y](p^1).$$

In the terminology of differential geometry, $[X, Y](p^1)$ is called *the Lie derivative of* Y *with respect to* X *at* p^1.

Another interpretation of the Lie derivative is as follows. Pick a basis for TM_{p^1}. Then $D\varphi(-\sigma; p^1)$ carries this basis into a basis for $TM_{\varphi(-\sigma; p^1)}$.

Express $Y(\varphi(-\sigma; p^1))$ in terms of this latter basis, and the Lie derivative of Y with respect to X at $\varphi(-\sigma; p^1)$ is just the derivative (with respect to σ) of the components of Y. But for any basis in TM_{p^1}, $D\varphi(\sigma; \varphi(-\sigma; p^1))Y(\varphi(-\sigma; p^1))$ merely gives the components of Y at $\varphi(-\sigma; p^1)$ in terms of the induced basis for $TM_{\varphi(-\sigma; p^1)}$. Thus, again $(d/d\sigma)D\varphi(\sigma; \varphi(-\sigma; p^1))Y(\varphi(-\sigma; p^1))$ is the Lie derivative of Y with respect to X. Also, inductively,

$$\frac{d^2}{d\sigma^2}D\varphi(\sigma; \varphi(-\sigma; p^1))Y(\varphi(-\sigma; p^1))|_{\sigma=0} = [X, [X, Y]](p^1).$$

In general, we obtain the Taylor series expansion

$$\begin{aligned} &D\varphi(\sigma; \varphi(-\sigma; p^1))Y(\varphi(-\sigma; p^1)) \\ (2.5)\qquad &= Y(p^1) - \sigma[X, Y](p^1) + (\sigma^2/2)(\mathrm{ad}^2\, X, Y)(p^1) - \cdots \\ &= (\exp(-\sigma\, \mathrm{ad}\, X))Y(p^1). \end{aligned}$$

In forming the derivative of q with respect to r in (2.4), we assumed $r < 0$ was admissible. Since r represents time along a trajectory, this presents a difficulty. Following ideas of Krener [5], we use the fact that both $X \pm Y$ are admissible vector fields. Let $Q^t(p)$ denote a solution, at time t, of $\dot{x} = X(x) - Y(x)$, $x(0) = p$ and define $q^-(r; \sigma) = T^\sigma Q^r T^{t_1-\sigma-r}(p^0)$. A simple calculation shows

$$\lim_{\substack{r\to 0\\ r>0}} \frac{\partial q}{\partial r}(r; \sigma) = -\lim_{\substack{r\to 0\\ r>0}} \frac{\partial q^-}{\partial r}(r; \sigma),$$

hence we may consider q, with $r \geqq 0$, as having a derivative with respect to r at $r = 0$ as given by (2.4).

Suppose, for some positive integer k,

$$(2.6)\qquad \mathrm{rank}\,\{Y(p^1), (\mathrm{ad}\, X, Y)(p^1), \cdots, (\mathrm{ad}^k\, X, Y)(p^1)\} = n.$$

For any set $\sigma_1, \cdots, \sigma_k$ of real numbers in a neighborhood of zero, define

$$q(r_1, \cdots, r_k; \sigma_1, \cdots, \sigma_k) = T^{\sigma_k}S^{r_k} \cdots T^{\sigma_1}S^{r_1}T^{t_1-\Sigma(\sigma_i+r_i)}(p^0).$$

From (2.4) and (2.5) it easily follows that

$$(2.7)\qquad \frac{\partial q}{\partial r_j}(0, \cdots, 0; \sigma_1, \cdots, \sigma_k) = Y(p^1) - \sigma_j(\mathrm{ad}\, X, Y)(p^1) + \cdots.$$

Now consider $\sigma_1, \cdots, \sigma_k$ as fixed parameters and $q: \mathbb{R}^k \to \mathbb{R}^n$. We have $q(0, \cdots, 0; \sigma_1, \cdots, \sigma_k) = p^1$, while from (2.6) and (2.7) one can easily conclude that in any neighborhood of zero, there are values $\sigma_1, \cdots, \sigma_k$ such that $Dq(0, \cdots, 0; \sigma_1, \cdots, \sigma_k)$ has rank n. For such a choice of the values σ_i, q maps a neighborhood of zero in $\mathbb{R}^k$ onto a neighborhood of p^1 in $\mathbb{R}^n$.

Remark 2. The conditions (2.3) and (2.6) are the same. Both provide only a sufficient condition for controllability of the linear variational equation.

We may now use either the methods of § 2.1, or of this section, to immediately obtain the following local controllability results.

PROPOSITION 2. *Let φ be a solution of* (1.1) *corresponding to control $u \equiv 0$. A sufficient condition that the system* (1.1) *is locally controllable along φ at time*

$t_2 > 0$ is that there exist a $t_1 \in (0, t_2]$ and integer k such that

$$\operatorname{rank}\{Y(\varphi(t_1)), (\operatorname{ad} X, Y)(\varphi(t_1)), \cdots, (\operatorname{ad}^k X, Y)(\varphi(t_1))\} = n.$$

PROPOSITION 3. *Let φ be a solution of system (1.2) corresponding to $u_j = 0$ for all j. A sufficient condition that system (1.2) be controllable along φ at time $t_2 > 0$ is that there exist integers $k_1, \cdots, k_s$ and a $t_1 \in (0, t_2]$ such that*

$$\operatorname{rank}\{Y^1(p^1), \cdots, Y^s(p^1), (\operatorname{ad} X, Y^1)(p^1), \cdots, (\operatorname{ad} X, Y^s)(p^1),$$
$$\cdots, (\operatorname{ad}^{k_1} X, Y^1)(p^1), \cdots, (\operatorname{ad}^{k_s} X, Y^s)(p^1)\} = n,$$

where $p^1 = \varphi(t_1)$.

PROPOSITION 4. *Let u^* be an admissible control for system (1.3) which generates a reference trajectory φ. A sufficient condition that system (1.3) is controllable along φ at time $t_2 > 0$ is that there exist a $t_1 \in (0, t_2]$ such that $u^*(t_1) \in$ interior U, u^* is constant in a neighborhood of t_1, and if $p^1 = \varphi(t_1)$,*

$$\operatorname{rank}\{Y^1(p^1), \cdots, Y^s(p^1), (\operatorname{ad} X, Y^1)(p^1), \cdots, (\operatorname{ad}^{k_s} X, Y^s)(p^1)\} = n$$

for some integers $k_1, \cdots, k_s$, where $X(x) = f(x, u^*(t_1))$ and $Y^j(x) = (\partial/\partial u_j)f(x, u^*(t_1))$.

3. Global controllability. Let M be a smooth, n-dimensional manifold. For X, Y any two C^∞ vector fields on M, we let $[X, Y]$ denote their Lie bracket product. Let $V(M)$ denote the set of all C^∞ vector fields on M, which we consider as a Lie algebra over the reals with product the Lie product. If $\mathscr{S}$ is a collection of vector fields on M, the fact that we required a solution of $\mathscr{S}$ to be defined only for $t \geqq 0$ again creates difficulties. These are eliminated by assuming $\mathscr{S}$ is *symmetric*, i.e., $X \in \mathscr{S}$ implies $-X \in \mathscr{S}$, for then we can effectively reverse time by reversing the vector field.

For the moment, consider $\mathscr{S}$ to be C^∞ and symmetric. Let $\mathscr{T}(\mathscr{S})$ denote the smallest subalgebra of the Lie algebra $V(M)$ containing $\mathscr{S}$; i.e., $\mathscr{T}(\mathscr{S})$ consists of the linear span of all elements of $\mathscr{S}$ together with all products of elements of $\mathscr{S}$, and products of products, etc. We can now state a case of Chow's theorem, as follows.

THEOREM (Chow [6]). *Let $\mathscr{S}$ be a symmetric set of C^∞ vector fields on M. If for each $x \in M$, $\dim \mathscr{T}(\mathscr{S})_x = n$, then $\mathscr{S}$ has the accessibility property on M.*

For symmetric $\mathscr{S}$, this theorem has as an immediate consequence a relation between $\mathscr{S}$ having the accessibility property and $\mathscr{S}$ being controllable.

COROLLARY. *Suppose M is connected, while $\mathscr{S}$ is a symmetric set of C^∞ vector fields with $\dim \mathscr{T}(\mathscr{S})_x = n$ for all $x \in M$. Then $\mathscr{S}$ is controllable on M.*

The verification of this is easy. Indeed, join any two points $p^1, p^2 \in M$ by an arc in M. For each point p on this arc, there is a neighborhood of points accessible from p, and this neighborhood contains p as an interior point. The arc is compact, hence we can take a finite subcover from the above neighborhoods. Call these $N(x^1), \cdots, N(x^k)$, and assume the labeling taken so $p^1 \in N(x^1)$, $N(x^i) \cap N(x^{i+1}) \neq \varnothing$ and $p^2 \in N(x^k)$. Then there is a trajectory of $\mathscr{S}$ which joins p^1 to a point $q^1 \in N(x^1) \cap N(x^2)$, a trajectory joining q^1 to x^2 in $N(x^2)$, a trajectory joining x^2 to $q^2 \in N(x^2) \cap N(x^3)$, etc., until we get to p^2. For details, see [7, pp. 664–665], or [2, Prop. 5] for a different argument.

If $\mathscr{S}$ is not symmetric, matters are not so easy. Sussmann and Jurdjevic obtain the following result in [1].

THEOREM (Sussmann–Jurdjevic). *Let $\mathscr{S}$ be a set of analytic vector fields on the analytic manifold M. If* $\dim \mathscr{T}(\mathscr{S})_x = n$ *for all $x \in M$, then $\mathscr{S}$ has the accessibility property.*

Now, however, it is in general not true that x is an interior point of the neighborhood of points attainable from x by solutions of $\mathscr{S}$. Thus the relation between $\mathscr{S}$ being controllable and having the accessibility property is no longer as easy as in the corollary following Chow's theorem. Some known results are as follows. Let $M = \mathbb{R}^n$ and $\mathscr{S}_x = \{Ax + b^iu, i = 1, 2, \cdots, k, -\infty < u < \infty\}$ where A is a given real $n \times n$ matrix and $b^1, \cdots, b^k \in \mathbb{R}^n$ are given. Then if $\mathscr{S}$ has the accessibility property, $\mathscr{S}$ is controllable. Here one can easily verify that $\dim \mathscr{T}(\mathscr{S})_x = \dim \mathscr{T}(\mathscr{S})_0 = \dim \operatorname{span} \{b^1, \cdots, b^k, Ab^1, \cdots, Ab^k, \cdots, A^{n-1}b^1, \cdots, A^{n-1}b^k\}$. The result now follows from standard controllability theory of linear control systems. Jurdjevic [2] shows that if $\mathscr{S}$ is a set of right-invariant analytic vector fields with the accessibility property on a compact, connected Lie group M, then $\mathscr{S}$ is controllable on M. This motivates the question, asked by Jurdjevic, as to whether the accessibility property for an analytic set of vector fields, $\mathscr{S}$, on a compact, connected, analytic manifold M implies $\mathscr{S}$ is controllable on M. We shall next give an example to show this is not true.

DEFINITION. A *tangent k-field on a manifold M* is a set of k tangent vector fields $X^1, \cdots, X^k$ which are linearly independent at each point of M.

Remark 3. Clearly, a k-field on a k-dimensional manifold has the accessibility property. For results on when an n-manifold admits a k-field, see [8].

EXAMPLE 3.1. *An analytic two-field on the torus, T^2, which is not controllable.*

Let θ, φ designate coordinates for $\mathbb{R}^2$, and define

$$X^1(\theta, \varphi) = (1/2 + (1/2)\cos \theta)\partial/\partial\theta + (1/2 - (1/2)\cos \theta)\partial/\partial\varphi$$

$$X^2(\theta, \varphi) = (-1/2 + (1/2)\cos \theta)\partial/\partial\theta + (1/2 + (1/2)\cos \theta)\partial/\partial\varphi.$$

These functions are 2π-periodic in θ and φ and analytic. We consider T^2 as the plane with points whose coordinates differ by integer multiples of 2π identified. Letting Z denote the integer multiples of 2π, the covering space map from $\mathbb{R}^2 \to T^2 = \mathbb{R}^2/Z \times Z$ induces a Riemannian metric on T^2. We now consider X^1 and X^2 as (induced) vector fields on T^2. Relative to the inner product derived from this metric, we see X^1 and X^2 are always orthogonal and neither is zero. Thus $\mathscr{S} = \{X^1, X^2\}$ is a two-field on T^2. (Note for later use, that $[X^1, X^2](\theta, \varphi) = ((1/2)\sin \theta, (1/2)\sin \theta \cos \theta)$, hence $\mathscr{S}$ is *not* a commuting two-field.) We see that for $\theta = \pi$, X^1 has a periodic solution $\theta(t) = \pi$, $\varphi(t) = t$. Denote this solution by γ_1. For $\theta = 0$, X^2 has a periodic solution γ_2. Also any solution which begins in the "half-torus" (cylinder) $0 \leqq \theta \leqq \pi$, $0 \leqq \varphi \leqq 2\pi$, remains there, i.e., this is an invariant set for the flow of $\mathscr{S}$, hence $\mathscr{S}$ is not controllable on T^2. (See Fig. 1.)

Remark. The referee has informed me that C. Lobry has also given an example of two analytic vector fields on the two-sphere which satisfy the accessibility property but do not yield a controllable system. His paper, *Controllability of nonlinear systems on compact manifolds*, appears in this Journal, 12 (1974), pp. 1–4.

DEFINITION. A k-field $\{X^1, \cdots, X^\kappa\}$ on M is called a *commuting* k-field if $[X^i, X^j] = 0$ for all $1 \leqq i, j \leqq k$. We say a manifold M has *rank k* if k is the largest

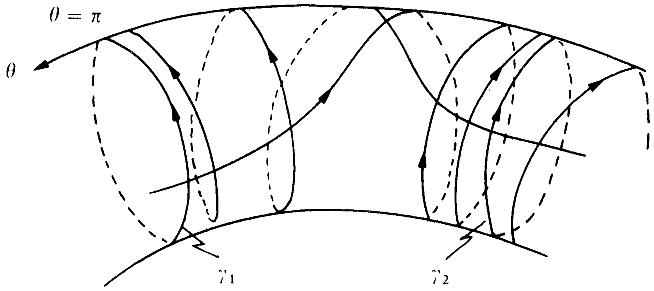

FIG. 1

integer such that there exists a commuting k-field on M.

For example (see [8]), rank $T^2 = 2$, while Lima [9] has shown that the rank of the three-sphere, S^3, is one.

THEOREM 3.1. *Let $\mathcal{S} = \{X^1, X^2\}$ be a commuting two-field on the two-torus T^2. Then $\mathcal{S}$ is controllable on T^2.*

Proof. (We utilize many ideas from the proof of [9, Thm. 1].) Since X^1, X^2 are pointwise linearly independent and commute, they determine an action φ of the Lie group $\mathbb{R}^2$ on T^2. Specifically, let $\xi, \eta : \mathbb{R}^1 \times T^2 \to T^2$ denote, respectively, the flows generated by X^1 and X^2. Then for any $r = (t, s) \in \mathbb{R}^2$ and $x \in T^2$, $\varphi_r(x) = \xi_t\eta_s(x) = \eta_s\xi_t(x)$. Since for controllability we restrict the time parameter to values $t \geqq 0$, what we wish to show is that the image $\{\varphi_r(x) : r = (t, s), t, s \geqq 0\}$ is all of T^2 for any $x \in T^2$. We shall call this the *positive orbit of φ through x.*

Consider, first, the positive semiorbit $\xi_t(x)$, $t \geqq 0$. If this is dense in T^2, the linear independent of X^1 and X^2 immediately yields the desired result, thus we assume the positive semiorbits of X^1 and X^2 are not dense in T^2. By the theorem of A. Schwartz for smooth (C^2 at least) vector fields on two-manifolds, since X^1 does not vanish, $\xi_t(x)$ must have a minimal set homeomorphic to S^1 for some $x_0 \in T^2$, i.e., there exists an $x_0 \in T^2$ such that the ξ-orbit through x_0 is periodic with least period $t_0 > 0$.

Now let K be the closure of the positive orbit of φ through x_0. Then K is compact, nonempty, φ positively invariant, hence contains a positive minimal set M for φ. The positive φ-orbit of $x_0 \in M$ is dense in K, since K is its closure, hence is dense in M. Thus if x is any point in M, there exists a sequence $r_n = (t_n, s_n)$, $t_n \to \infty$, $s_n \to \infty$, such that $\varphi_{r_n}(x_0) \to x$. Then $\xi_{t_0}(x) = \lim \xi_{t_0} \circ \varphi_{r_n}(x_0) = \lim \varphi_{r_n} \circ \xi_{t_0}(x_0) = \lim \varphi_{r_n}(x_0) = x$, where the second equality uses the commutivity of ξ, η. This shows that every ξ-orbit through a point of M is periodic of period t_0.

Pick any $x_1 \in M$, and let $\gamma_1 = \{\xi(t, x_1) : 0 \leqq t \leqq t_0\}$ be the closed orbit of X^1 through x_1. Since $X^1 \neq 0$, γ_1 does not bound a disc, hence $\xi(\cdot, x_1)$ represents a nontrivial element of $\pi_1(T^2)$. Since X^1 and X^2 are linearly independent, the full orbit of φ through x_1 is either a cylinder (the case in which the isotropy subgroup is Z, the integers), or the full orbit is T^2 (the case in which the isotropy subgroup is $Z \times Z$). In the latter case, it clearly will also be true that the positive orbit of φ through x_1 will be T^2. To complete the proof, it suffices to rule out the case that the full orbit of φ through x_1 is a cylinder. We shall show that the η-orbit through some point of M is homeomorphic to S^1 but that this orbit is not in the homotopy class of $\xi(\cdot, x_1)$. Then by the above argument, every η-orbit through a point of M is periodic (with the same period) and the isotropy subgroup of φ is $Z \times Z$, as desired.

Let $x \in \gamma_1$ and $\eta_{+\infty}(x)$ be the ω-limit set of the η-flow through x. Let $L = \xi_{\mathbb{R}}(\eta_{+\infty}(x))$ be the union of all ξ-orbits through points of $\eta_{+\infty}(x)$. Since each such ξ-orbit has period t_0, $L = \xi_{[0,t_0]}(\eta_{+\infty}(x))$. Now $\eta_{+\infty}(x)$ is compact, so L is compact. Since L is also nonempty and positively φ-invariant, $L = M$. In particular, $x \in L$, thus $x = \xi_t(y)$ for some $t \in [0, t_0]$ and $y \in \eta_{+\infty}(x)$, so $y = \xi_{-t}(x)$, i.e., $y \in \gamma_1 \cap \eta_{+\infty}(x)$. Since we assumed that neither X^1 or X^2 generated a dense flow, and neither is zero, Schwartz's theorem again gives that $\eta_{+\infty}(x)$ is a periodic orbit of X^2, which we denote by $\gamma_2 = \{\eta(s, y) : 0 \leqq s \leqq s_0\}$, s_0 being the minimum period. We know that $y \in (\gamma_1 \cap \gamma_2) \neq \emptyset$. By linear independence of X^1 and X^2, γ_2 is not identical with γ_1, nor does it bound a disc. If $\eta(\cdot, y)$ is homotopic to $\xi(\cdot, x_1)$, the orbits γ_1 and γ_2 must have at least one more intersection, say at a point y_1 where the orientations of the two-field $\{X^1(y), X^2(y)\}$ and $\{X^1(y_1), X^2(y_1)\}$ must be reversed. This would contradict the linear independence of the vector fields on T^2, hence $\eta(\cdot, y), \xi(\cdot, x_1)$ must not be homotopic. They then represent generators of $Z \times Z$, and the full orbit of φ is diffeomorphic to $\mathbb{R}^2/Z \times Z$ or T^2. The same must be true of the positive orbit of φ, and the proof is complete.

In general, a manifold M cannot be expected to admit a k-field. Indeed, for $k = 1$, a 1-field is a nonsingular vector field, and a necessary (and sufficient) condition for M to admit such is that the Euler characteristic $\chi(M) = 0$. It is a well-known result of Whitney [10] that any n-dimensional manifold can be imbedded in $\mathbb{R}^{2n}$. Let $f: M \to \mathbb{R}^{2n}$ be such an imbedding. Then relative to a choice of coordinates for $\mathbb{R}^{2n}$, f has the form $(f^1, \cdots, f^{2n})$, with each f^i real-valued. Let X^i be the gradient field of f^i. Rank $Df(x) = n$ for all $x \in M$ since f is an imbedding, hence dim span $\{X^1(x), \cdots, X^{2n}(x)\} = n$ for all $x \in M$. Thus we have the following.

PROPOSITION 5. *Any smooth n-manifold M admits a collection $\mathcal{S} = \{X^1, \cdots, X^{2n}\}$ of $2n$ smooth vector fields such that* dim span $\{X^1(x), \cdots, X^{2n}(x)\} = n$ *for all $x \in M$.*

Remark 4. All that was required for the above construction was that rank $Df(x) = n$ for all $x \in M$; thus it would have sufficed to have f an immersion. Whitney [11] has shown that every smooth n-manifold can be immersed in $\mathbb{R}^{2n-1}$, for $n > 1$. One may easily sharpen Proposition 5 by replacing $2n$ with $(2n - 1)$. (The case $n = 1$ provides no difficulty.) An interesting question which arises is: can one replace $2n$ by $(n + 1)$ in Proposition 5?

A collection of vector fields, as in Proposition 5, will always possess the accessibility property. We shall next show (geometrically) how to locally modify any such collection $\mathcal{S}$ so that

(a) the modified collection $\mathcal{S}'$ still satisfies the condition dim span $\mathcal{S}'_x = n$ for all $x \in M$, but $\mathcal{S}'$ is not controllable on M.

One may conclude, from this, that if $\mathcal{S}$ has the accessibility property, in order to show $\mathcal{S}$ is controllable on M, one must impose additional conditions on the vector fields of $\mathcal{S}$ (such as right-invariant with M a connected Lie group as done by Jurdjevic, or commuting as in Theorem 3.1 above). Controllability of $\mathcal{S}$ on M cannot be accomplished by any conditions on the topology of M alone.

We next outline, geometrically, the construction to obtain the modified collection $\mathscr{S}'$ as in (a) above. To do this analytically, with all details, would not serve much purpose, in this author's opinion.

Let $\mathscr{S} = \{X^1, \cdots, X^{2n}\}$ be as in Proposition 5. Pick $x^0 \in M$. Our modification will be local, hence we assume we work in $\mathbb{R}^n$ rather than M. By a slight perturbation of the vector fields of $\mathscr{S}$, if necessary, we may assume that

(i) every collection $\{X^{i_1}, \cdots, X^{i_n}\}$ of n distinct elements of $\mathscr{S}$ satisfies $\dim \operatorname{span}\{X^{i_1}(x^0), \cdots, X^{i_n}(x^0)\} = n$. This also insures $X^j(x^0) \neq 0$ for $j = 1, \cdots, 2n$, hence we further assume that

(ii) each $X^j(x^0)$ has unit length, and

(iii) for some $\varepsilon > 0$, all of the vector fields of $\mathscr{S}$ are constant in a 4ε-neighborhood N of x^0.

This allows us to picture each X^i as producing a parallel flow in N. We shall modify each X^i to a vector field Y^i in N so that X^i and Y^i match on the boundary, but each Y^i has exactly two critical points in N. Let $p^i = x^0 + \varepsilon X^i(x^0)$, $q^i = x^0 + 3\varepsilon X^i(x^0)$, $i = 1, \cdots, 2n$, and choose Y^i to have an attractive node at p^i, a hyperbolic critical point with $(n-1)$-dimensional stable manifold at q^i, and no other critical points in N. Figure 2 shows the flow of the modification Y^i of X^i.

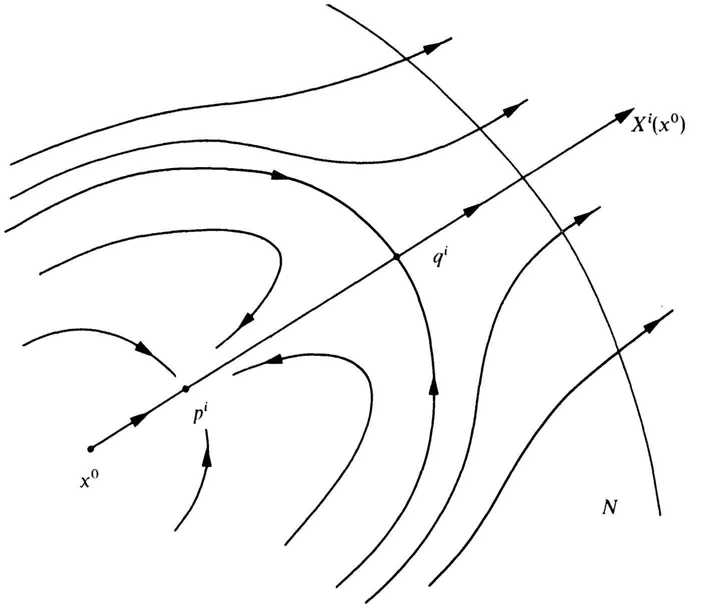

FIG. 2

By assumptions (i), (ii), the points p^i, q^i are distinct for different values of i, hence the vector fields Y^i do not have common zeros. Thus one can do such a construction so that $\dim \operatorname{span}\{Y^1(x), \cdots, Y^{2n}(x)\} = n$ for all $x \in N$. Also, we may have $Y^i = X^i$ on the boundary of N, while a 2ε-neighborhood of x^0 becomes an invariant set for the flow of $\mathscr{S}' = \{Y^1, \cdots, Y^{2n}\}$ in N. Letting Y^i be X^i in the complement of N, we obtain a collection $\mathscr{S}'$ with $\dim \operatorname{span} \mathscr{S}'_x = n$ for all $x \in M$, so $\mathscr{S}$ has the accessibility property, yet the 2ε-neighborhood of x^0 is invariant under the flow of $\mathscr{S}'$; hence $\mathscr{S}'$ is not controllable on M.

REFERENCES

[1] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

[2] V. JURDJEVIC, *Certain controllability properties of analytic control systems*, this Journal, 10 (1972), pp. 354–360.

[3] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[4] H. HERMES, *Controllability and the singular problem*, this Journal, 2 (1965), pp. 241–260.

[5] A. J. KRENER, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems*, this Journal, 12 (1974),

[6] W. L. CHOW, *Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann., 117 (1939), pp. 98–105.

[7] G. W. HAYNES, *On the optimality of a totally singular vector control: An extension of the Green's theorem approach to higher dimensions*, this Journal, 4 (1966), pp. 662–677.

[8] E. THOMAS, *Vector fields on manifolds*, Bull. Amer. Math. Soc., 75 (1969), pp. 643–683.

[9] E. LIMA, *Commuting vector fields on S^3*; Ann. of Math., 81 (1965), pp. 70–81.

[10] H. WHITNEY, *The self-intersections of a smooth n-manifold in 2n space*, Ibid., 45 (1949), pp. 220–246.

[11] ———, *The singularities of a smooth n-manifold in* $(2n - 1)$ *space*, Ibid., 45 (1949), pp. 247–293.

SIAM J. Control
Vol. 12, No. 2, May 1974

LINEAR DIFFERENTIAL GAMES*

L. S. PONTRYAGIN†

Abstract. A pursuit problem and an evasion problem are formulated and results are obtained for the case in which the dynamics are governed by linear differential equations and the terminal set is a linear manifold in the state space. Conditions are given ensuring the existence of an open set in the phase space such that if the initial state belongs to this set, termination of the pursuit game can be achieved. Conditions are also given ensuring that for any initial state not in the terminal manifold the evasion game can be prolonged indefinitely.

This paper is devoted to the study of the pursuit process of one controlled object by another controlled object. The general nonlinear problem will be formulated, but the results will only concern a linear case. Even in this simple case they are nontrivial and interesting. Stronger results than those which will be given here have already been published [1], [2], [3]. Those given here, however, will be in a form that is simple and easy to remember.

The problem considered is technological in origin, and the success of the mathematical investigation depends very much on the idealization which is made. I consider here only my idealization which leads to a rather simple mathematical consideration. There are other idealizations of the problem, but I do not intend to mention them here.

To have a technological example let us imagine that one airplane pursues another. The objective of the first airplane is to intercept the second one. The objective of the second airplane is to prevent interception. Each pilot controls his plane, having in mind his objective and using the information about the situation. The information consists of two parts. The first part is the complete knowledge of the performance capabilities of both planes. The second part of the information concerns the present and the past behavior of the airplanes, but nothing is known about their future behavior. We must give a mathematical idealization that retains the essential features of the technological problem.

We denote the phase vector of a controlled object by x and assume that the motion of this object is described by the ordinary differential equation

$$\dot{x} = f(x, u), \tag{1}$$

where the dot denotes derivative with respect to time t and u is the control. The variable u may take its values from the given set P; thus $u \in P$. This equation describes the performance capabilities of the object. Indeed, it gives all the motions of which the object is capable. To obtain a concrete motion we have to specify the initial values t_0, x_0 and we have to prescribe the values of the control u as a function of time t; $u = u(t)$. It is supposed that u is a measurable function of t. The possibility of the choice of different functions $u(t)$ just means that the object is a controlled one. As x is the phase vector, it consists of two parts $x = (x_1, x_2)$, where x_1 is the geometrical position, and x_2 is the velocity of the object. In the

* Received by the editors April 27, 1973.

† Steklov Mathematical Institute, Academy of Sciences of the USSR, Moscow B-333, USSR.

case of an airplane, equation (1) gives its performance capabilities and the values of the control u correspond to the various settings of the plane.

In the pursuit process there are two controlled objects, the pursuer x and the evader y. The capabilities of the evader are described by the equation

$$\dot{y} = g(y, v). \tag{2}$$

Here $v \in Q$ is the control, and y consists of two parts, $y = (y_1, y_2)$, where y_1 is the geometrical position and y_2 is the velocity of the object.

Interception occurs when the objects coincide geometrically, that is,

$$x_1 = y_1. \tag{3}$$

If interception occurs, we shall also say that the pursuit is completed.

The first part of the information which I mentioned above is contained in equations (1) and (2), and it is always assumed to be known. In considering the second part, different mathematical idealizations of the problem arise. I am going to describe one of them.

The pursuit process itself can be considered from two different viewpoints.

In the first point of view we identify ourselves with the pursuer and assume that the evader moves in an arbitrary manner in accordance with his capabilities. The control u is in our hands and our aim is to achieve interception, or to complete the pursuit. Having our aim in mind, we have to calculate the value $u(t)$ at each instant of time t, using the knowledge of the functions

$$x(s), \quad y(s), \quad v(s)$$

which are defined on the interval $t - \theta \leqq s \leqq t$, where θ is a suitable positive number. In symbols,

$$u(t) = U_t(x(s), y(s), v(s), t - \theta \leqq s \leqq t),$$

where U_t is a functional which we call the *pursuit rule*.

In the second point of view we identify ourselves with the evader and assume that the pursuer moves in an arbitrary manner in accordance with his capabilities. Then the control v is in our hands and our aim is to prevent interception, or completion of the pursuit. So having our aim in mind, we have to calculate the value $v(t)$ at each instant of time t, using the knowledge of the functions

$$x(s), \quad y(s), \quad u(s)$$

which are defined on the interval $t - \theta \leqq s \leqq t$, where θ is a suitable positive number. In symbols,

$$v(t) = V_t(x(s), y(s), u(s), t - \theta \leqq s \leqq t),$$

where V_t is a functional which we call the *evasion rule*.

This is the idealization which I will consider. Here we have two problems: the pursuit problem and the evasion problem, which are quite different.

To simplify the notations we transform the pursuit process into a differential game by coupling the phase vectors x and y of the objects into a single vector $z = (x, y)$. The vector z is the phase vector of the game. It belongs to the phase vector space R of the game, where R is the direct sum of the phase vector spaces of the two objects. We can now rewrite our two differential equations (1) and (2) into the single differential equation

$$\dot{z} = F(z, u, v). \tag{4}$$

Condition (3) defines a certain subset M in the space R. Now we can consider the game independently of the pursuit process. The game is given if its phase vector space R, differential equation (4), and the subset M on which the game is completed are given. Here we have $z \in R$ and two controls: u—the pursuit control, and v—the evasion control.

Here we also have two problems: the pursuit game and the evasion game. The pursuit game is as follows: At each instant of time t find the value $u(t)$ in order to complete the game. The value $u(t)$ is a functional of the two functions $z(s)$ and $v(s)$ which are defined for $t - \theta \leqq s \leqq t$, where θ is a suitable positive number. In symbols,

$$u(t) = U_t(z(s), v(s), t - \theta \leqq s \leqq t),$$

where U_t is a functional which we call a pursuit rule. In other words, the problem is to select a function U_t, or pursuit rule, that will result in completion of the game. The evasion game is as follows: At each instant of time t find the value $v(t)$ in order to prevent completion of the game. The value $v(t)$ is a functional of the two functions $z(s)$ and $u(s)$ which are defined for $t - \theta \leqq s \leqq t$, where θ is a suitable positive number. In symbols,

$$v(t) = V_t(z(s), u(s), t - \theta \leqq s \leqq t),$$

where V_t is a functional which we call an evasion rule. In other words, the problem is to select a functional V_t, or evasion rule, that will prevent completion of the game.

To obtain some concrete results I will restrict myself to linear differential games. The differential equation of the game is written in the following form:

$$\dot{z} = Cz - u + v + a, \tag{5}$$

where $z \in R$, the phase vector space R is a Euclidean space, C is a given linear mapping of the phase vector space R of the game into itself, a is a constant vector in the space R, and the controls u and v are also vectors in the space R. The controls are not arbitrary vectors, but satisfy the conditions

$$u \in P, \quad v \in Q, \tag{6}$$

where P and Q are given compact convex subsets of the space R. These sets are sets of arbitrary dimension. The set M on which the game is completed is a vector subspace of the space R. We denote the orthogonal complement of M in R by L, and its dimension by ν. Thus $R = M \oplus L$ and

$$\dim L = \nu, \qquad \nu \geqq 2. \tag{7}$$

To obtain a solution of the pursuit game, i.e., to achieve completion, we have to have superiority of the pursuit control over the evasion control. Similarly, we have to have superiority of the evasion control over the pursuit control in order to obtain a solution of the evasion game, i.e., to prevent completion. The capabilities of the pursuit control u are defined by the set P, and the capabilities of the evasion control v are defined by the set Q. Therefore, to express these superiorities we have to compare the sets and express the fact that one of them is in some sense richer than the other. This comparison of the sets P and Q must be done by taking into consideration the linear differential equation of the game and the final set M, and therefore it must depend on the linear mapping C of the space R into itself and on the linear subspace M.

Let us denote by π the operation of orthogonal projection of the space R onto its subspace L. Thus if z is an arbitrary vector of the space R, πz is its orthogonal projection onto the space L. Since C is a linear mapping of the space R into itself, $e^{\tau C}$, where τ is a real number, is a linear mapping of the space R into itself. This mapping depends analytically on the variable τ. Now we can consider the linear mapping $\pi e^{\tau C}$ of the space R onto the space L. This mapping also depends on the variable τ. Let us apply this linear transformation to the sets P and Q. We obtain two compact convex subsets of the space L:

$$P_\tau = \pi e^{\tau C} P, \qquad Q_\tau = \pi e^{\tau C} Q. \tag{8}$$

Now the superiorities which I mentioned above may be expressed essentially as follows:

$$\begin{aligned} &\mu Q_\tau \subset P_\tau \quad \text{for the pursuit game,} \\ &\mu P_\tau \subset Q_\tau \quad \text{for the evasion game,} \end{aligned} \tag{9}$$

where μ is a constant >1.

Now I shall formulate the results: the first for the pursuit game and the second for the evasion game.

The pursuit control u has superiority over the evasion control v if the following two conditions hold:

(A) The dimension of the set P_τ is equal to ν for all sufficiently small positive values of τ, i.e.,

$$\dim P_\tau = \dim L = \nu. \tag{10}$$

(B) For all sufficiently small positive values of τ the following inclusion holds:

$$\mu Q_\tau \subset P_\tau, \tag{11}$$

where μ is a constant, $\mu > 1$.

The result for the pursuit game is the following: If conditions (A) and (B) hold, then there exists a nonempty open set Ω in the phase space R of the game such that if the initial value z_0 of the game belongs to Ω, then the game with this initial value z_0 can be completed.

The evasion control v has superiority over the pursuit control u if the following two conditions (C) and (D) hold. (The statement of these conditions can be obtained by interchanging the sets P_τ and Q_τ in conditions (A) and (B).)

(C) The dimension of the set Q_τ is equal to ν for all sufficiently small positive values of τ, i.e.,

$$\dim Q_\tau = \dim L = \nu. \tag{12}$$

(D) For all sufficiently small positive values of τ the following inclusion takes place:

$$\mu P_\tau \subset Q_\tau, \tag{13}$$

where μ is a constant, $\mu > 1$.

The result for the evasion game is the following: If conditions (C) and (D) hold, then in the evasion game completion can be prevented for arbitrary initial values z_0 which do not belong to M. This means that for arbitrary $z_0 \notin M$ it is possible to choose an evasion rule such that the phase vector $z(t)$ of the game never reaches the set M; i.e., for all $t < \infty$, $z(t) \notin M$. Moreover, the distance $\xi(t)$ of the point $z(t)$ from the set M cannot become too small, and this distance can be estimated in the following way.

Let us denote by $\eta(t)$ the distance between $z(t)$ and L. Then there exist positive constants c and ε and a positive integer k, all depending only on the data defining the game, such that if $\xi(0) \leqq \varepsilon$, then

$$\xi(t) > \frac{c\xi^k(0)}{[1 + \eta(t)]^k}. \tag{14}$$

If $\xi(0) > \varepsilon$, then we wait until $\xi(t)$ becomes equal to ε, say at time t_0. The estimate (14) will then hold for $t \geqq t_0$, while for $t < t_0$ we have the estimate $\xi(t) > \varepsilon$.

To illustrate the results let us consider two examples.

Let E be a geometrical Euclidean vector space of dimension $v \geqq 2$. Let x and y be two points in the space E, x—the pursuer and y—the evader. Here x and y are not the phase vectors of the objects, but their geometrical positions. Thus the pursuit process is completed when $x = y$.

Example 1. The differential equations which describe the motions of x and y are the following:

$$\ddot{x} = u, \qquad \dot{y} = v. \tag{15}$$

Here the controls u and v are vectors in the space E subject to the following constraints:

$$|u| \leqq 1, \qquad |v| \leqq 1. \tag{16}$$

It turns out that this pursuit process, considered as a game, satisfies conditions (C) and (D). The evasion can therefore be continued indefinitely provided only that $x(0) \neq y(0)$. This result is intuitively clear as the evader y has the maneuvering superiority over the pursuer x. This is so because the evader's velocity is controlled directly while the velocity of the pursuer x is controlled indirectly by controlling the acceleration. Thus the pursuer's velocity can only be changed slowly.

Example 2. The differential equations which describe the motions of x and y are the following:

$$\ddot{x} + \alpha\dot{x} = u, \qquad \ddot{y} + \beta\dot{y} = v. \tag{17}$$

Here the controls u and v are vectors in the space E subject to the following constraints:

$$|u| \leqq \rho, \qquad |v| \leqq \sigma. \tag{18}$$

Here α, β, ρ, σ are positive numbers. The motions of the points x and y, described by equations (17), are motions under the action of forces u and v with linear friction.

It turns out that, if

$$\sigma > \rho, \tag{19}$$

then the differential game which corresponds to the pursuit process satisfies conditions (C) and (D). The evasion process can therefore be continued indefinitely provided only that $x(0) \neq y(0)$. If, however, the inequality

$$\rho > \sigma \tag{20}$$

holds, then conditions (A) and (B) are satisfied. Therefore a nonempty open set Ω exists in the phase space such that if the initial state $(x(0), \dot{x}(0), y(0), \dot{y}(0))$ of the game belongs to Ω, then the pursuit process will be completed in a finite time. If besides the inequality (20) the following inequality

$$\rho/\alpha > \sigma/\beta \tag{21}$$

holds, then the set Ω is the whole phase space R of the game. The pursuit process can then be completed for any arbitrary initial state $(x(0), \dot{x}(0), y(0), \dot{y}(0))$ and the time of the pursuit can be estimated in terms of the initial state of the game.

REFERENCES

[1] L. S. Pontryagin, *Linear differential games, I*, Dokl. Akad. Nauk SSSR, 174 (1967), pp. 1278–1280 = Soviet Math. Dokl., 8 (1967), pp. 769–771.

[2] L. S. Pontryagin and E. F. Misčenko, *A problem on the escape of one controlled object from another*, Dokl. Acad. Nauk SSSR, 189 (1969), pp. 721–723 = Soviet Math. Dokl., 10 (1969), pp. 1488–1490.

[3] L. S. Pontryagin, *The linear differential game of evasion*, Trudy Mat. Inst. Steklov., 112 (1971), pp. 30–63.

SIAM J. CONTROL
Vol. 12, No. 2, May 1974

AUGMENTED LAGRANGE MULTIPLIER FUNCTIONS AND DUALITY IN NONCONVEX PROGRAMMING*

R. TYRRELL ROCKAFELLAR†

Abstract. If a nonlinear programming problem is analyzed in terms of its ordinary Lagrangian function, there is usually a duality gap, unless the objective and constraint functions are convex. It is shown here that the gap can be removed by passing to an augmented Lagrangian which involves quadratic penalty-like terms. The modified dual problem then consists of maximizing a concave function of the Lagrange multipliers and an additional variable, which is a penalty parameter. In contrast to the classical case, the multipliers corresponding to inequality constraints in the primal are not constrained a priori to be nonnegative in the dual. If the maximum in the dual problem is attained (and conditions implying this are given), optimal solutions to the primal can be represented in terms of global saddle points of the augmented Lagrangian. This suggests possible improvements of existing penalty methods for computing solutions.

1. Introduction. Let $f_0, f_1, \cdots f_m$ be real-valued functions defined on a set $S \subset R^n$. We shall be concerned with the nonlinear programming problem:

$$\text{(P)} \qquad \begin{aligned} &\text{minimize} \quad f_0(x) \quad \text{over all } x \in S \text{ satisfying} \\ &f_i(x) \leqq 0 \quad \text{for } i = 1, \cdots, m. \end{aligned}$$

The ordinary Lagrangian function associated with problem (P) is

$$(1.1) \qquad L_0(x, y) = f_0(x) + y_1 f_1(x) + \cdots + y_m f_m(x) \quad \text{for } (x, y) \in S \times R^m_+,$$

and this corresponds to the dual problem:

$$(\text{D}_0) \qquad \begin{aligned} &\text{maximize} \quad g_0(y) \quad \text{over all } y \in R^m_+, \quad \text{where} \\ &g_0(y) = \inf_{x \in S} L_0(x, y). \end{aligned}$$

It is well known that the optimal values in these two problems satisfy

$$(1.2) \qquad \inf(\text{P}) \geqq \sup(\text{D}_0),$$

but equality cannot be expected to hold, aside from freakish cases, unless S and the functions f_i are convex. The discrepancy in (1.2) is termed a "duality gap".

In recent years a number of authors have addressed the question of whether this duality gap in nonconvex programming could be eliminated by changing the Lagrangian function. Such a change might also be of benefit computationally in some situations, even in convex programming, where the plurality of useful Lagrangians and dual problems has been known for some time. Computational considerations in nonconvex problems with equality constraints have led in particular to algorithms based on an augmented Lagrangian in which "penalty" terms of the form $rf_i(x)^2$, $i = 1, \cdots, m$, are added to $L_0(x, y)$; cf. Arrow and Solow [2], Bertsekas [3], Buys [4], Fletcher [6], [7], [8], Haarhoff and Buys [9], Hestenes [10], Kort and Bertsekas [11], Lill [12], Miele et al. [14], [15], [16], [17], Poljak [30],

* Received by the editors March 5, 1973, and in revised form August 30, 1973.

† Department of Mathematics, University of Washington, Seattle, Washington 98195. This work was supported in part by the Air Force Office of Scientific Research under Grant AFOSR-72-2269.

Tripathi and Narendra [25], and Wierzbicki [27], [28], [29]. For the inequality-constrained problem (P), the simple terms $rf_i(x)^2$ are not suitable, and the analogous augmented Lagrangian (suggested in [21] and investigated by Buys in his thesis [4]) turns out to be

$$
\begin{aligned}
L(x,y,r) &= f_0(x) + \sum_{i=1}^{m} [y_i \max\{f_i(x), -y_i/2r\} + r\max{}^2\{f_i(x), -y_i/2r\}] \\
&= f_0(x) + r\sum_{i=1}^{m} \psi(f_i(x), y_i/r) \quad \text{for } x \in S, \quad (y,r)\in T,
\end{aligned} \tag{1.3}
$$

where $T = R^m \times (0, +\infty)$ and

$$
\begin{aligned}
\psi(\alpha,\beta) &= [\max{}^2\{0, 2\alpha+\beta\} - \beta^2]/4 \\
&= \begin{cases} \alpha\beta + \alpha^2 & \text{if } \alpha \geqq -\beta/2, \\ -\beta^2/4 & \text{if } \alpha \leqq -\beta/2. \end{cases}
\end{aligned} \tag{1.4}
$$

We have demonstrated in [13] that in the convex case this augmented Lagrangian is not only a natural choice but has a number of strong properties not possessed by the ordinary Lagrangian L_0. In [14], we have derived some consequences of these properties for the multiplier method of Hestenes and Powell. It is the purpose of the present paper to develop general properties of L in the nonconvex case, especially with regard to duality.

Arrow, Gould and Howe [1, Thm. 2] have already shown that if $\bar{x}$ is an isolated local solution to (P) satisfying the standard second order sufficiency conditions for optimality with strict complementarity, the Lagrange multiplier vector being $\bar{y}$, and if $\bar{r}$ is sufficiently large, then there is a neighborhood N of $\bar{x}$ in S such that

$$
\min_{x\in N} L(x,\bar{y},\bar{r}) = L(\bar{x},\bar{y},\bar{r}) = \max_{y\in R^m_+} L(\bar{x},y,\bar{r}), \tag{1.5}
$$

with the minimum in (1.5) attained uniquely at $\bar{x}$. This saddle-point theorem is strengthened below (Corollary 6.1) in three ways: by extending the maximum in (1.5) to the maximum of $L(\bar{x}, y, r)$ over all $(y, r) \in T$ (thus in particular removing the constraint $y \geqq 0$), by deleting the strict complementarity assumption, and (under the hypothesis that $\bar{x}$ is the unique globally optimal solution to (P) "in the strong sense") by extending the minimum in (1.5) to the minimum over all $x \in S$. Introducing the ordinary perturbations associated with (P), we also give necessary and sufficient conditions in terms of stability for the existence of a global saddle point $(\bar{x}, \bar{y}, \bar{r})$ of L with respect to $S \times T$ and more generally characterize the case where at least the global "inf sup" and "sup inf" of L are equal.

These results correspond to a detailed study of the following dual problem in place of (D_0):

$$
\begin{aligned}
&\text{maximize} \quad g(y,r) \quad \text{over all } (y,r)\in T, \quad \text{where} \\
&g(y,r) = \inf_{x\in S} L(x,y,r) < +\infty.
\end{aligned} \tag{D}
$$

Of course, the optimal value in (D) is by definition

$$
\sup(\mathrm{D}) = \sup_{(y,r)\in T}\ \inf_{x\in S} L(x,y,r). \tag{1.6}
$$

On the other hand, the optimal value in (P) satisfies

$$\inf(\mathrm{P}) = \inf_{x\in S}\sup_{(y,r)\in T} L(x,y,r), \tag{1.7}$$

inasmuch as

$$\sup_{(y,r)\in T} L(x,y,r) = \begin{cases} f_0(x) & \text{if } x \text{ is feasible,} \\ +\infty & \text{if } x \text{ is not feasible.}\end{cases} \tag{1.8}$$

The latter is immediate from the fact that

$$\sup_{\beta\in R}\psi(\alpha,\beta) = \begin{cases} 0 & \text{if } \alpha \leqq 0, \\ +\infty & \text{if } \alpha > 0.\end{cases} \tag{1.9}$$

Thus the relation

$$\inf(\mathrm{P}) \geqq \sup(\mathrm{D}) \tag{1.10}$$

holds, and minimax theorems for L are equivalent to duality theorems asserting the equality and attainment of the optimal values in (1.10). (For related work on duality since this paper was submitted for publication, see Mangasarian [13], Pollatschek [18] and Rockafellar [24].)

For notational simplicity, only inequality constraints are treated in this paper. However, the same results apply with only the obvious changes if explicit equality constraints are also allowed (the corresponding terms $r\psi(f_i(x), y_i/r)$ in (1.3) being replaced by $y_i f_i(x) + rf_i(x)^2$). The routine alterations in the proofs are left to the reader.

Except for Theorem 6, which requires second order differentiability of the functions f_i, the results remain valid if S is a subset of an arbitrary topological real vector space.

2. The nature of the dual problem. Let $p: R^m \to [-\infty, +\infty]$ be the ordinary perturbation function (min-value function) associated with (P), that is,

$$p(u) = \inf_{x\in S} F(x,u), \tag{2.1}$$

where for each $(x,u)\in S \times R^m$:

$$F(x,u) = \begin{cases} f_0(x) & \text{if } f_i(x) \leqq u_i \text{ for } i = 1,\cdots,m, \\ +\infty & \text{otherwise.}\end{cases} \tag{2.2}$$

Then

$$\inf_{u\in R^m}\{F(x,u) + y\cdot u\} = \begin{cases} L_0(x,y) & \text{if } y\in R^m_+, \\ -\infty & \text{if } y \notin R^m_+,\end{cases} \tag{2.3}$$

$$\inf_{u\in R^m}\{p(u) + y\cdot u\} = \begin{cases} g_0(y) & \text{if } y \in R^m_+, \\ -\infty & \text{if } y \notin R^m_+.\end{cases} \tag{2.4}$$

More generally, it is elementary to calculate that

$$L(x,y,r) = \inf_{u\in R^m} \{F(x,u) + y\cdot u + r|u|^2\} \quad \text{for all } (y,r)\in T, \tag{2.5}$$

$$g(y,r) = \inf_{u\in R^m} \{p(u) + y\cdot u + r|u|^2\} \quad \text{for all } (y,r)\in T. \tag{2.6}$$

In order that (2.3) and (2.4) can be regarded as instances of (2.5) and (2.6), we adopt the convention that

$$L(x,y,0) = \begin{cases} L_0(x,y) & \text{if } y\in R^m_+, \\ -\infty & \text{if } y\notin R^m_+, \end{cases} \tag{2.7}$$

$$g(y,0) = \begin{cases} g_0(y) & \text{if } y\in R^m_+, \\ -\infty & \text{if } y\notin R^m_+. \end{cases} \tag{2.8}$$

This extends the definition of $L(x,\cdot,\cdot)$ and g to cl T.

THEOREM 1. *The functions $L(x,y,r)$ and $g(y,r)$ are concave and upper semicontinuous in $(y,r)\in \operatorname{cl} T = R^m \times R^1_+$ and nondecreasing in $r\in R^1_+$, nowhere $+\infty$. Furthermore, whenever $r > s \geqq 0$ one has*

$$g(y,r) \geqq \max_{z\in R^m} \{g(z,s) - |y-z|^2/4(r-s)\}. \tag{2.9}$$

Proof. The first assertion is implied by (2.5) and (2.6), since the pointwise infimum of a collection of affine functions of (y,r) which are nondecreasing in r is an upper semicontinuous, concave function which is nondecreasing in r. For any (y,r) and (z,s) satisfying $r > s \geqq 0$, we have from (2.6) that

$$\begin{aligned} g(y,r) &= \inf_{u\in R^m} \{p(u) + z\cdot u + s|u|^2 + (y-z)\cdot u + (r-s)|u|^2\} \\ &\geqq \inf_{u\in R^m} \{p(u) + z\cdot u + s|u|^2\} + \inf_{u\in R^m} \{(y-z)\cdot u + (r-s)|u|^2\} \\ &= g(z,s) - |y-z|^2/4(r-s), \end{aligned}$$

and this yields (2.9). The maximum (instead of supremum) in (2.9) is valid because $g(\cdot,s)$ is an upper semicontinuous concave function nowhere having the value $+\infty$ and hence in particular is majorized by at least one affine function. (Thus the function of z being maximized is upper semicontinuous; its level sets are bounded because it is majorized by a negative definite quadratic function of z.)

Remark. In the convex case (i.e., where S and the functions f_i are all convex), $L(x,y,r)$ is convex in x and relation (2.9) holds as an equation [13]. Then for every $r > 0$ the function $g(\cdot,r)$ has the same maximum and even the same maximizing set as $g(\cdot,0)$, since in the formula

$$g(y,r) = \max_{z\in R^m} \{g(z,0) - |y-z|^2/4r\}$$

the bracketed expression is maximized jointly in y and z if and only if y maximizes $g(\cdot,0)$ and $z = y$. In other words, in the convex case a pair $(\bar{y},\bar{r})$ with $\bar{r} > 0$ is an optimal solution to the dual problem (D) if and only if $\bar{y}$ is an optimal solution to the ordinary dual (D_0). In the nonconvex case this is no longer true, although the monotonicity of $g(y,r)$ in r still implies that if $(\bar{y},\bar{r})$ is an optimal solution to (D) and $r > \bar{r}$, then $(\bar{y},r)$ is also an optimal solution to (D).

COROLLARY 1.1. *There is an r_0, $0 \leqq r_0 \leqq +\infty$, such that $g(y,r)$ is finite for all $y \in R^m$ if $r_0 < r < +\infty$, whereas $g(y,r) = -\infty$ for all $y \in R^m$ if $0 \leqq r < r_0$.*

Proof. This is obvious from (2.9), according to which $g(y,r) > -\infty$ if there exists some $s \in [0,r)$ and $z \in R^m$ such that $g(z,s) > -\infty$.

In view of the fact that $g(y,r)$ is nondecreasing in r, Corollary 1.1 says there are no real constraints at all in (D), even implicit ones. This is in contrast to the situation for (D_0), where the feasibility condition $g_0(y) > -\infty$ requires the satisfaction of $y \geqq 0$, as well as other possible constraints. (It is not always possible a priori to specify for (D) an r such that $r > r_0$, although, for example, one has $r_0 = 0$ if f_0 is bounded below on S. In this connection, see the remarks preceding Theorem 2 in the next section.)

COROLLARY 1.2. *For every $y \in R^m$, one has*

$$\lim_{r \to +\infty} g(y,r) = \sup_T g = \sup(\mathrm{D}). \tag{2.10}$$

Proof. Given any $(z,s) \in T$ and $\varepsilon > 0$, one has $g(y,r) \geqq g(z,s) - \varepsilon$ for all r sufficiently large by (2.9).

The last result brings out the close relationship between the dual (D) and penalty methods for solving (P). By definition, we have

$$L(x,0,r) = f_0(x) + r \sum_{i=1}^{m} \max{}^2 \{0, f_i(x)\}, \tag{2.11}$$

and consequently

$$g(0,r) = \inf_{x \in S} \left\{ f_0(x) + r \sum_{i=1}^{m} \max{}^2 \{0, f_i(x)\} \right\}. \tag{2.12}$$

The limit of the infimum (2.12) as $r \to +\infty$ is the optimal value sup (D), according to Corollary 1.2. Thus the relationship between sup (D) and inf (P) is of fundamental importance for the penalty method in which (2.12) is calculated for a sequence of r values tending to $+\infty$. Note that if we fix any $y \in R^m$ and minimize $L(\cdot, y, r)$, instead of $L(\cdot, 0, r)$, for a sequence of r values tending to $+\infty$, the limit of the infima is still sup (D) by Corollary 1.2. This procedure can be regarded as a modified penalty method. Still more broadly, one can try to solve (P) by minimizing $L(\cdot, y, r)$ for a sequence of vectors $(y,r) \in T$ such that $g(y,r) \to \sup(\mathrm{D})$. If the sequence can be generated in such a manner that the r values remain *bounded*, there is the advantage that the numerical instabilities associated with minimizing (2.11) for ever-larger values of r could be avoided. The results below demarcate the region of validity and potential effectiveness of such algorithms, from a theoretical point of view. Theorem 6 indicates that indeed, penalty methods can be constructed which are capable of solving "most" problems *without* $r \to +\infty$.

3. Solving (P) in the asymptotic sense. We say that (P) satisfies the *quadratic growth condition* if there is an $r \geqq 0$ such that the expression (2.11) is bounded below as a function of $x \in S$. This certainly holds if f_0 is bounded below on S, and in particular if S is compact and f_0 lower semicontinuous. In general, since by (2.6) and the definition of g we have

$$\inf_{x\in S} L(x,0,r) = g(0,r) = \inf_{u\in R^m} \{p(u) + r|u|^2\}, \tag{3.1}$$

the quadratic growth condition holds if and only if there exist real numbers $r \geqq 0$ and q such that

$$p(u) \geqq q - r|u|^2 \quad \text{for all } u \in R^m. \tag{3.2}$$

The condition is therefore equivalent also to the relation

$$\liminf_{|u|\to+\infty} p(u)/|u|^2 > -\infty.$$

Observe that the r_0 in Corollary 1.1 is the infimum of all the numbers $r \geqq 0$ for which the quadratic growth condition holds, since it is the infimum of all the numbers $r \geqq$ such that $g(0, r) > -\infty$. Thus (P) satisfies the condition if and only if g is not identically $-\infty$ on T, or, in other words, if and only if (D) has "feasible solutions". This also shows that the quadratic growth condition is equivalent to the seemingly more general condition that for some $y \in R^m$ (not necessarily $y = 0$) and some $r \geqq 0$, the infimum of $L(x, y, r)$ over all $x \in S$ is not $-\infty$.

THEOREM 2. *If* (P) *satisfies the quadratic growth condition, one has*

$$\begin{aligned} -\infty < \sup(\mathrm{D}) &= \liminf_{u\to 0} p(u) \\ &\leqq p(0) = \inf(\mathrm{P}). \end{aligned} \tag{3.3}$$

If (P) *does not satisfy the quadratic growth condition, one has* $\sup(\mathrm{D}) = -\infty$.

Proof. The preceding remark makes clear that $\sup(\mathrm{D}) = -\infty$ if and only if the quadratic growth condition fails to be satisfied. Assume henceforth that the condition is satisfied; thus (3.2) holds for a certain $\bar{q}$ and $\bar{r}$. From (3.1) we see that

$$g(0,r) \leqq \liminf_{u\to 0} p(u) \quad \text{for all } r \geqq 0.$$

Taking the limit as $r \to +\infty$ and invoking Corollary 1.2, we obtain

$$\sup(\mathrm{D}) \leqq \liminf_{u\to 0} p(u).$$

To establish the opposite inequality, and thereby complete the proof of the theorem, consider now an arbitrary real number q such that

$$q < \liminf_{u\to 0} p(u). \tag{3.4}$$

Choose ε sufficiently small that $p(u) \geqq q$ whenever $|u| < \varepsilon$. For r sufficiently large, we have

$$q - r|u|^2 \leqq \bar{q} - \bar{r}|u|^2 \quad \text{if } |u| \geqq \varepsilon$$

(with $\bar{q}$ and $\bar{r}$ as above), and therefore

$$q - r|u|^2 \leqq p(u) \quad \text{for all } u.$$

But then

$$q \leqq \inf_{u\in R^m} \{p(u) + r|u|^2\} = g(0,r) \leqq \sup(\mathrm{D}).$$

Since q was any real number satisfying (3.4), this shows that

$$\sup(\mathrm{D}) \geqq \liminf_{u \to 0} p(u),$$

and we are done.

The quantity

$$\liminf_{u \to 0} p(u) \tag{3.5}$$

in Theorem 2 is the *asymptotic optimal value* in (P). It can also be described as the minimum of

$$\limsup_{k \to \infty} f_0(x^k) \tag{3.6}$$

over all asymptotically feasible sequences $(x^k)_{k=1}^{\infty}$ for (P): that is, sequences in S satisfying

$$\limsup_{k \to \infty} f_i(x^k) \leqq 0 \quad \text{for } i = 1, \cdots, m. \tag{3.7}$$

Indeed, according to the definition of p, (3.5) is the lowest possible limit achievable by any sequence $(\alpha_k)_{k=1}^{\infty}$ such that there exist $u^k \in R^m$ and $x^k \in S$ with $u^k \to 0$, $f_i(x^k) \leqq u_i^k$ for $i = 1, \cdots, m$, and $f_0(x^k) \leqq \alpha_k$.

Let us call a sequence $(x^k)_{k=1}^{\infty}$ *asymptotically minimizing* for (P) if it is asymptotically feasible and yields the minimum possible value for (3.6). We can then obtain from Theorem 2 a result which shows how any procedure for solving (D) can be used to solve (P) in the sense of constructing an asymptotically minimizing sequence. (A similar result involving more detailed estimates in the convex case has been demonstrated in [22].)

THEOREM 3. *Let $(y^k, r_k)_{k=1}^{\infty}$ be a sequence such that for some $\delta > 0$ one has $(y^k, r_k - \delta) \in T$ and*

$$\lim_{k \to \infty} g(y^k, r_k - \delta) = \sup(\mathrm{D}) < +\infty. \tag{3.8}$$

Let $x^k \in S$ satisfy

$$L(x^k, y^k, r_k) \leqq \inf_{x \in S} L(x, y^k, r_k) + \alpha_k, \tag{3.9}$$

where $\alpha_k \to 0$. Then $(x^k)_{k=1}^{\infty}$ is asymptotically feasible and

$$\liminf_{k \to \infty} y_i^k / r_k \geqq 0 \quad \textit{for } i = 1, \cdots, m. \tag{3.10}$$

If in addition $(y^k)_{k=1}^{\infty}$ is bounded, then $(x^k)_{k=1}^{\infty}$ is an asymptotically minimizing sequence for (P).

Proof. From (3.9) and (3.8) we have

$$L(x^k, y^k, r_k) \leqq g(y^k, r_k) + \alpha_k \leqq \sup(\mathrm{D}) + \alpha_k < +\infty. \tag{3.11}$$

In particular, sup (D) is finite. On the other hand, (2.5) and (2.2) imply

$$L(x^k, y^k, r_k) = f_0(x^k) + y^k \cdot u^k + r_k |u^k|^2, \tag{3.12}$$

where

$$u_i^k = \max\{f_i(x^k), -y_i^k / 2r_k\} \quad \text{for } i = 1, \cdots, m. \tag{3.13}$$

Therefore, using (2.1) and (2.6),

$$\begin{aligned} L(x^k, y^k, r_k) &\geqq p(u^k) + y^k \cdot u^k + (r_k - \delta)|u^k|^2 + \delta|u^k|^2 \\ &\geqq g(y^k, r_k - \delta) + \delta|u^k|^2 . \end{aligned} \tag{3.14}$$

We combine (3.14) with (3.11) to obtain

$$\delta|u^k|^2 \leqq \sup(\mathrm{D}) - g(y^k, r_k - \delta) + \alpha_k \to 0. \tag{3.15}$$

Thus $u^k \to 0$, and this establishes in view of (3.13) that (3.7) and (3.10) hold. Next we argue from (3.11) and (3.14) that

$$\lim_{k \to \infty} L(x^k, y^k, r_k) = \sup(\mathrm{D}). \tag{3.16}$$

If the y^k sequence is bounded, then (3.12) and the fact that $u^k \to 0$ give us

$$\lim_{k \to \infty} f_0(x^k) = \sup(\mathrm{D}).$$

But sup (D), since it is finite, is the asymptotic optimal value in (P) by Theorem 2. This completes the proof.

The need for the boundedness of $(y^k)_{k=1}^{\infty}$ in Theorem 3, even in the convex case, is illustrated by the following counterexample.

Example 1. Define f_0, f_1, f_2 for $x = (x_1, x_2, x_3) \in R^3$ by $f_0(x) = x_3$, $f_1(x) = x_1$, $f_2(x) = x_2$. Let

$$S = \{x \in R^3 | x_1 y + x_2 y_2 - x_3 \leqq 0 \text{ for all } (y_1, y_2) \in C\},$$

where

$$C = \{y \in R^2 | y_1 \leqq 0, y_1^2 + 2y_2 \leqq 0\}.$$

Note that S is a closed convex cone which can also be expressed as

$$S = \{x \in R^3 | x_3 \geqq \phi(x_1, x_2)\},$$

where ϕ is the *support function* of C:

$$\begin{aligned} \phi(x_1, x_2) &= \sup\{x_1 y_1 + x_2 y_2 | (y_1, y_2) \in C\} \\ &= \begin{cases} x_1^2/2x_2 & \text{if } x_1 \leqq 0 \text{ and } x_2 > 0, \\ 0 & \text{if } x_1 \geqq 0, x_2 \geqq 0, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

The function ϕ is nonincreasing in x_1 and x_2, so obviously

$$p(u_1, u_2) = \phi(u_1, u_2) \quad \text{for all } u_1, u_2.$$

It can be shown, incidentally, from this fact and formula (2.6) by means of elementary results about conjugate functions, that

$$g(y, r) = -(1/4r)\,\mathrm{dist}^2(-y, C).$$

All we really need to know at the moment, however, is that $g(y, r) \leqq 0$ everywhere and

$$g(y, r) = (1/r)g(y, 1) \quad \text{for } r > 0. \tag{3.17}$$

These relations follow from (2.6) because $p(0) = 0$ and $p(u/r) = p(u)/r$. Let

$$u^k = (-k^{-1}, k^{-3}), \qquad x^k = (-k^{-1}, k^{-3}, k),$$
$$r_k \equiv 1, \qquad y^k = -\nabla p(u^k) - 2u^k = (k^2 + 2k^{-1}, (1/2)k^4 - 2k^{-3}).$$

Then if y^k and r_k are substituted into (2.6), the minimum is attained uniquely at u^k, indicating that

$$g(y^k, r_k) = -|u^k|^2 \to 0 = \sup(\mathrm{D}).$$

Hence also $g(y^k, r_k - \delta) \to \sup(\mathrm{D})$ by (3.17), if $0 < \delta < 1$. On the other hand, the minimum in (2.1) for $u = u^k$ is attained uniquely at x^k. Thus x^k uniquely minimizes $L(\cdot, y^k, r_k)$ over S (cf. (2.5)), and all the assumptions in Theorem 3 are satisfied except for the boundedness of $(y^k)_{k=1}^{\infty}$. But $f_0(x^k) = k \to +\infty$, so that $(x^k)_{k=1}^{\infty}$ is certainly not an asymptotically minimizing sequence for (P).

Two corollaries of Theorem 3 may now be stated.

COROLLARY 3.1. *Assume the asymptotic optimal value in* (P) *is not* $+\infty$. *Fix any* $y \in R^m$. *Let* x^k *satisfy*

$$L(x^k, y, r_k) \leqq \inf_{x \in S} L(x, y, r_k) + \alpha_k, \tag{3.18}$$

where $r_k \to +\infty$ *and* $\alpha_k \to 0$. *Then* $(x^k)_{k=1}^{\infty}$ *is an asymptotically minimizing sequence for* (P).

Proof. With $y^k \equiv y$, we have (3.8) by Corollary 1.2 so that the conclusions of Theorem 3 are justified.

COROLLARY 3.2. *Let* $(\bar{y}, \bar{r})$ *be such that for some* $\delta > 0$ *one has* $(\bar{y}, \bar{r} - \delta) \in T$ *and*

$$-\infty < g(\bar{y}, \bar{r} - \delta) = \sup(\mathrm{D}). \tag{3.19}$$

Let $(x^k)_{k=1}^{\infty}$ *be a minimizing sequence in* S *for the function* $L(\cdot, \bar{y}, \bar{r})$. *Then* $\bar{y} \geqq 0$, *and* $(x^k)_{k=1}^{\infty}$ *is an asymptotically minimizing sequence for* (P). *Moreover, if* $\bar{x}$ *is a point at which the minimum of* $L(\cdot, \bar{y}, \bar{r})$ *over* S *is attained, then* $\bar{x}$ *is actually an optimal solution to* (P).

Proof. Take $(y^k, r^k) \equiv (\bar{y}, \bar{r})$ in Theorem 3. For the final assertion of the corollary, take $x^k \equiv \bar{x}$.

Theorem 3 makes clear the computational relevance of the questions of when sup (D) equals inf (P) and when sup (D) is attained. These questions are answered in the next section in terms of the stability of (P).

4. Duality theorems and stability. Problem (P) will be called (*lower*) *stable of degree* k (where k is a nonnegative integer) if there is an open neighborhood U of the origin in R^m and a function $\pi : U \to R$ of class C^k such that

$$p(u) \geqq \pi(u) \quad \text{for all } u \in U, \quad \text{with } p(0) = \pi(0). \tag{4.1}$$

This implies of course that inf (P) is finite.

Stability of degree 0 is equivalent to the property that

$$p(0) = \liminf_{u \to 0} p(u) \quad \text{(finite)}. \tag{4.2}$$

The necessity of (4.2) is evident. On the other hand, if (4.2) holds, then the nonincreasing function

$$\theta(s) = \inf_{|u| \leqq s} p(u), \qquad s \geqq 0,$$

satisfies $\theta(s) \to \theta(0)$ as $s \to 0$. Choose $\varepsilon > 0$ small enough that $\theta(\varepsilon) > -\infty$, and define the function θ_0 on $[0, \varepsilon/2]$ as follows: $\theta_0(0) = \theta(0)$, $\theta_0(\varepsilon/(j+1)) = \theta(\varepsilon/j)$ for positive integers j, θ_0 interpolated linearly over the intervals $[\varepsilon/(j+1), \varepsilon/j]$. Then θ_0 is continuous and $\theta_0 \leqq \theta$. The definition of stability of degree 0 is therefore satisfied by $\pi(u) = \theta_0(|u|)$.

Theorem 2 therefore gives us the following.

THEOREM 4. *Suppose that* (P) *satisfies the quadratic growth condition. In order that the duality relation*

$$\inf(\mathrm{P}) = \sup(\mathrm{D}) \tag{4.3}$$

hold, or equivalently

$$\inf_S \sup_T L(x, y, r) = \sup_T \inf_S L(x, y, r), \tag{4.4}$$

it is necessary and sufficient that (P) *be stable of degree* 0.

Various conditions are known which guarantee stability of degree 0, i.e., (4.2). The most basic perhaps is the following: S is closed, the functions f_i are all lower semicontinuous, and for some $u \in \operatorname{int} R^m_+$ and $\alpha > \inf(\mathrm{P})$ the set

$$\{x \in S | f_0(x) \leqq \alpha, f_1(x) \leqq u_1, \cdots, f_m(x) \leqq u_m\} \tag{4.5}$$

is compact. (This is evident from the characterization of (3.5) in terms of asymptotically minimizing sequences.) In the convex case, the Slater condition and its variants suffice [20], [24].

Stability of degree 1 is a generalization of the stability condition in convex programming that p be subdifferentiable at $u = 0$. As a matter of fact, in the convex case stability of degree 1 implies stability of all higher orders. In the absence of convexity, however, stability of degree 2 plays an essential role.

THEOREM 5. *Suppose that* (P) *satisfies the quadratic growth condition. In order that the duality relation*

$$\inf(\mathrm{P}) = \max(\mathrm{D}) \tag{4.6}$$

hold, or equivalently,

$$\inf_S \sup_T L(x, y, r) = \max_T \inf_S L(x, y, r), \tag{4.7}$$

it is necessary and sufficient that (P) *be stable of degree* 2. *Indeed,* $(\bar{y}, \bar{r})$ *is an optimal solution to* (D) *for some* $\bar{r} > 0$ *if and only if* $\bar{y} = -\nabla\pi(0)$ *for some function* π *as in the definition of stability of degree* 2.

Proof. Clearly (4.6) is equivalent to the existence of $(\bar{y}, \bar{r}) \in T$ such that

$$\inf(\mathrm{P}) \leqq g(\bar{y}, \bar{r}) > -\infty, \tag{4.8}$$

since $\inf(\mathrm{P}) \geqq \sup(\mathrm{D})$ in general, while $\sup(\mathrm{D}) > -\infty$ by Theorem 2. Using (2.6), we can write (4.8) in the form

$$-\infty < p(0) \leqq p(u) + \bar{y} \cdot u + \bar{r}|u|^2 \quad \text{for all } u \in R^m. \tag{4.9}$$

If this is fulfilled, then $p(0)$ is finite and the condition for (P) to be stable of degree 2 is satisfied with

$$\pi(u) = p(0) - \bar{y} \cdot u - \bar{r}|u|^2, \qquad U = R^m.$$

Here $\bar{y} = -\nabla\pi(0)$.

Assume now conversely that the stability condition is satisfied for a certain π and U. Then $\pi(0) = p(0)$ (finite). Define $\bar{y} = -\nabla\pi(0)$, and choose $\varepsilon > 0$ small enough that $|u| \leqq \varepsilon$ implies $u \in U$. Since π is of class C^2, there is an $r_1 > 0$ such that

$$z \cdot \nabla^2\pi(u)z \geqq -2r_1|z|^2 \quad \text{for all } z \in R^m \text{ if } |u| \leqq \varepsilon. \tag{4.10}$$

Then

$$\pi(u) \geqq p(0) - \bar{y} \cdot u - r_1|u|^2 \quad \text{if } |u| \leqq \varepsilon. \tag{4.11}$$

This follows from the fact that for $h(t) = \pi(tu)$, $0 \leqq t \leqq 1$, one has

$$h(1) = h(0) + \int_0^1 \left[h'(0) + \int_0^t h''(\tau)\,dt \right] dt,$$

where

$$h''(\tau) = u \cdot \nabla^2\pi(tu)u.$$

Since (P) satisfies the quadratic growth condition, there exist numbers q and r such that (3.2) holds. We can choose $r_2 > 0$ so that

$$q - r|u|^2 \geqq p(0) - \bar{y} \cdot u - r_2|u|^2 \quad \text{if } |u| \geqq \varepsilon. \tag{4.12}$$

Then (4.12) and (3.2) imply

$$p(u) \geqq p(0) - \bar{y} \cdot u - r_2|u|^2 \quad \text{if } |u| \geqq \varepsilon,$$

while (4.11) and (4.1) imply

$$p(u) \geqq p(0) - \bar{y} \cdot u - r_1|u|^2 \quad \text{if } |u| \leqq \varepsilon.$$

Taking $\bar{r} = \max\{r_1, r_2\}$, we have (4.9), and hence equivalently (4.6) as already noted.

COROLLARY 5.1. *Suppose* (P) *satisfies the quadratic growth condition and is stable of degree* 0. *Then* (D) *has an optimal solution if and only if* (P) *is stable of degree* 2.

Proof. This is obtained by combining Theorem 5 with Theorem 2.

COROLLARY 5.2. *Suppose* (P) *satisfies the quadratic growth condition and is stable of degree* 2. *In order that $\bar{x} \in S$ be an optimal solution to* (P), *it is necessary and sufficient that there exist $(\bar{y}, \bar{r}) \in T$ such that*

$$L(x, \bar{y}, \bar{r}) \geqq L(\bar{x}, \bar{y}, \bar{r}) \geqq L(\bar{x}, y, r) \quad \textit{for all } x \in S, \quad (y, r) \in T. \tag{4.13}$$

Moreover, this condition is satisfied by $(\bar{y}, \bar{r})$ if and only if $(\bar{y}, \bar{r})$ is an optimal solution to (D).

Proof. The saddle-point condition (4.13) is equivalent by virtue of (1.8) and (1.10) to $\bar{x}$ being a feasible solution to (P) such that

$$f_0(\bar{x}) = \min(\mathrm{P}) = \max(\mathrm{D}) = g(\bar{y}, \bar{r}), \tag{4.14}$$

in which case the common value in (4.14) is $L(\bar{x}, \bar{y}, \bar{r})$.

Remark. If there exist $\bar{x} \in S$ and $(\bar{y}, \bar{r}) \in T$ satisfying (4.13), and therefore (4.14), then (P) must satisfy the quadratic growth condition (cf. remark preceding Theorem 2) and hence be stable of degree 2 (Theorem 5). Compare also with Corollary 3.2.

Corollary 5.2 may be regarded as a generalization of the Kuhn–Tucker theorem in convex programming. Qualitatively, we may expect that most problems encountered in practice will be stable of degree 2, so that the result will be applicable. But, as in the case of "constraint qualifications" and other familiar conditions in the theory of nonlinear programming, it is hard to give verifiable criteria directly in terms of the constraint functions (rather than an unknown optimal solution) which imply such stability. Of course, convexity plus some form of the Slater condition is sufficient. In the next section we investigate the nonconvex case further in terms of the local conditions which are usually satisfied by optimal solutions to (P).

It should be emphasized that the saddle-point relation (4.13) does yield the usual differential Kuhn–Tucker conditions if $\bar{x} \in \operatorname{int} S$ and the functions f_i are differentiable at $\bar{x}$. Indeed, (4.13) implies

$$0 = \frac{\partial L}{\partial y_i}(\bar{x}, \bar{y}, \bar{r}) = \max\{f_i(\bar{x}), -\bar{y}_i/2\bar{r}\} \quad \text{for } i = 1, \cdots, m, \tag{4.15}$$

$$\begin{aligned} 0 = \nabla_x L(\bar{x}, \bar{y}, \bar{r}) &= \nabla f_0(\bar{x}) + \sum_{i=1}^{m} \max\{0, \bar{y}_i + 2\bar{r} f_i(\bar{x})\} \nabla f_i(\bar{x}) \\ &= \nabla f_0(\bar{x}) + \sum_{i=1}^{m} [\bar{y}_i + 2\bar{r} \max\{f_i(\bar{x}), -\bar{y}_i/2\bar{r}\}] \nabla f_i(\bar{x}), \end{aligned} \tag{4.16}$$

or in other words,

$$f_i(\bar{x}) \leqq 0, \quad \bar{y}_i \geqq 0, \quad \bar{y}_i f_i(\bar{x}) = 0 \quad \text{for } i = 1, \cdots, m, \tag{4.17}$$

$$\nabla f_0(\bar{x}) + \bar{y}_1 \nabla f_1(\bar{x}) + \cdots + \bar{y}_m \nabla f_m(\bar{x}) = 0. \tag{4.18}$$

At all events, the vectors $\bar{y}$ involved in Theorem 5 and its corollaries can be interpreted in terms of "equilibrium prices" for perturbations of (P). As seen at the beginning of the proof of Theorem 5, a pair $(\bar{y}, \bar{r}) \in T$ satisfies

$$\inf(\mathrm{P}) = \sup(\mathrm{D}) = g(\bar{y}, \bar{r})$$

if and only if

$$p(u) + \bar{y} \cdot u + \bar{r}|u|^2 \tag{4.19}$$

is minimized in u when $u = 0$. Let us imagine an "economic" situation where we are allowed to perturb (P) by replacing the constraint functions f_i by $f_i - u_i$, so as to obtain perhaps a lower minimum "cost" value $p(u)$, but the cost associated with the perturbation vector $u = (u_1, \cdots, u_m)$ is $\bar{y} \cdot u + \bar{r}|u|^2$. The expression in (4.19) gives the resulting total cost associated with the perturbed problem. Thus (4.19) describes the "equilibrium" where the costs are such that no advantage is to be gained from perturbation, and we are "content with (P) as it is." In particular, we would have (assuming $p(0)$ finite):

$$-\bar{y}\cdot u \leqq \liminf_{\lambda\downarrow 0}\frac{p(\lambda u)-p(0)}{\lambda} \quad \text{for all } u. \tag{4.20}$$

As is well known, such a global "equilibrium" cannot be achieved with $\bar{r} = 0$ unless, at the very least, p coincides at 0 with its convexification, a property which is very unlikely in nonconvex programming.

5. Local criterion for stability of degree 2. We consider now an $\bar{x}$ which is an optimal solution to (P) and show that, if certain conditions slightly stronger then those usually necessary for optimality are satisfied at $\bar{x}$, (P) must be stable of degree 2. In doing this, we extend a result of Arrow, Gould and Howe [1], as described in the Introduction.

The point $\bar{x}$ is said to be the *unique optimal solution to* (P) *in the strong sense* if every asymptotically minimizing sequence for (P) converges to $\bar{x}$. This condition is milder than it might seem. For example, assuming the functions f_i are lower semicontinuous, it is satisfied if S is replaced by any compact subset in which $\bar{x}$ is the only locally optimal solution to (P).

The following conditions are well known to be sufficient (and "almost necessary") for $\bar{x}$ to be an isolated locally optimal solution to (P) (cf. [5, p. 30]):

(a) S contains an open neighborhood N_0 of $\bar{x}$ on which the functions f_i are all of class C^2;

(b) there is a vector $\bar{y}\in R^m$ such that the Kuhn–Tucker conditions (4.17) and (4.18) hold;

(c) for the Hessian matrix

$$H = \nabla^2 f_0(\bar{x}) + \bar{y}_1\nabla^2 f_1(\bar{x}) + \cdots + \bar{y}_m\nabla^2 f_m(\bar{x}) = \nabla_x^2 L_0(\bar{x},\bar{y}) \tag{5.1}$$

and the index sets

$$I_0 = \{i \neq 0 | f_i(\bar{x}) = 0, \bar{y}_i > 0\}, \qquad I_1 = \{i \neq 0 | f_i(\bar{x}) = 0, \bar{y}_i = 0\}, \tag{5.2}$$

one has $z\cdot Hz > 0$ for every nonzero $z\in R^m$ such that

$$z\cdot\nabla f_i(\bar{x}) = 0 \quad \text{for all } i\in I_0 \quad \text{and} \quad z\cdot\nabla f_i(\bar{x}) \leqq 0 \quad \text{for all } i \in I_1. \tag{5.3}$$

These will be referred to as the *standard* (*second order*) *sufficiency conditions.*

THEOREM 6. *Suppose* (P) *satisfies the quadratic growth condition. Let $\bar{x}$ be the unique optimal solution to* (P) *in the strong sense, and assume that $\bar{x}$ satisfies the standard sufficiency conditions with $\bar{y}$ as the vector of multipliers. Then* (P) *is stable of degree* 2, *and for all $\bar{r}$ sufficiently large the pair $(\bar{y}, \bar{r})$ is an optimal solution to* (D).

Proof. Let $N \subset S$ denote a neighborhood of $\bar{x}$, the nature of which will be specified later, and define

$$p^0(u) = \inf\{f_0(x) | x\in N \text{ and } f_i(x) \leqq u_i, i = 1, \cdots, m\}. \tag{5.4}$$

Since $\bar{x}$ is the unique optimal solution to (P) in the strong sense, there exists $\varepsilon > 0$ such that $x\in N$ whenever $x\in S, f_i(x) \leqq \varepsilon$ for $i = 1,\cdots, m$, and $f_0(x) \leqq f_0(\bar{x}) + \varepsilon$. Then

$$p(u) = p^0(u) \quad \text{for all } u\in U_0, \tag{5.5}$$

where

$$U_0 = \{u \in R^m | u_i \leqq \varepsilon \text{ for } i = 1, \cdots, m \text{ and } p(u) < p(0) + \varepsilon\}. \tag{5.6}$$

Suppose we can construct a function π of class C^2 on an open neighborhood U_1 of the origin in R^m such that

$$p^0(u) \geqq \pi(u) \quad \text{for all } u \in U_1, \quad \text{with } p^0(0) = \pi(0). \tag{5.7}$$

We will then have

$$p(u) \geqq \pi(u) \quad \text{for all } u \in U_0 \cap U_1 \quad \text{with } p(0) = \pi(0), \tag{5.8}$$

so that the definition of stability of degree 2 will be satisfied with

$$U = \{u \in U_1 | \pi(u) < \pi(0) + \varepsilon \text{ and } u_i < \varepsilon \text{ for } i = 1, \cdots, m\}. \tag{5.9}$$

(If $u \in U$ but $u \notin U_0$, we have $p(u) \geqq p(0) + \varepsilon = \pi(0) + \varepsilon$ and hence $p(u) > \pi(u)$.) If also $\nabla\pi(0) = -\bar{y}$, then $(\bar{y}, \bar{r})$ is an optimal solution to (D) for all $\bar{r}$ sufficiently large by Theorem 5 and the monotonicity of $g(y, r)$ in r. Thus the proof of the theorem is reduced to the construction of N, U_1 and π satisfying (5.7), such that π is of class C^2 on U_1 and $\nabla\pi(0) = -\bar{y}$.

It will be enough actually to show the existence of N such that, for some $\bar{r} > 0$,

$$L(x, \bar{y}, \bar{r}) \geqq L(\bar{x}, \bar{y}, \bar{r}) = f_0(\bar{x}) \quad \text{for all } x \in N. \tag{5.10}$$

Indeed, this will imply from (2.5) that

$$\begin{aligned} f_0(\bar{x}) &= \inf_{x \in N} L(x, \bar{y}, \bar{r}) = \inf_{x \in N} \inf_{u \in R^m} \{F(x, u) + \bar{y} \cdot u + \bar{r}|u|^2\} \\ &= \inf_{u \in R^m} \inf_{x \in N} \{F(x, u) + \bar{y} \cdot u + \bar{r}|u|^2\} \\ &= \inf_{u \in R^m} \{p^0(u) + \bar{y} \cdot u + \bar{r}|u|^2\}. \end{aligned} \tag{5.11}$$

Since $p^0(0) = f_0(\bar{x})$, we will then have

$$p^0(u) \geqq p^0(0) - \bar{y} \cdot u - \bar{r}|u|^2 \quad \text{for all } u \in R^m. \tag{5.12}$$

In other words, the desired properties will hold for $\pi(u) = p^0(0) - \bar{y} \cdot u - \bar{r}|u|^2$ and $U_1 = R^m$.

Let I_0 and I_1 be the index sets in (5.2), and let

$$I_2 = \{i \neq 0 | f_i(\bar{x}) < 0\}. \tag{5.13}$$

Let N_0 be the neighborhood of $\bar{x}$ in the standard sufficiency conditions. For all $r > 0$, define

$$N_1(r) = N_0 \bigcap_{i \in I_0} \{x | f_i(x) > -\bar{y}_i/2r\} \bigcap_{i \in I_2} \{x | f_i(x) < 0\}. \tag{5.14}$$

Then $N_1(r)$ is an open neighborhood of $\bar{x}$, and for all $x \in N_1(r)$ we have

$$L(x, \bar{y}, r) = f_0(x) + \sum_{i \in I_0} [\bar{y}_i f_i(x) + r f_i(x)^2] + r \sum_{i \in I_1} \theta(f_i(x))^2, \tag{5.15}$$

where

$$\theta(\alpha) = \max\{\alpha, 0\}. \tag{5.16}$$

Observe that $L(\bar{x}, \bar{y}, r) = f_0(\bar{x})$, and by the Kuhn–Tucker conditions

$$\text{(5.17)} \qquad \nabla_x L(\bar{x}, \bar{y}, r) = \nabla f_0(\bar{x}) + \sum_{i \in I_0} \bar{y}_i \nabla f_i(\bar{x}) = 0.$$

We shall show next that in fact

$$\text{(5.18)} \qquad L(x, \bar{y}, r) = f_0(\bar{x}) + h(x - \bar{x}) + rk(x - \bar{x}) + o(|x - \bar{x}|^2),$$

where $h(z) = z \cdot Hz$ and

$$\text{(5.19)} \qquad k(z) = \sum_{i \in I_0} (z \cdot \nabla f_i(\bar{x}))^2 + \sum_{i \in I_1} \theta(z \cdot \nabla f_i(\bar{x}))^2 \geqq 0.$$

Since by (5.17)

$$\text{(5.20)} \qquad f_0(x) + \sum_{i \in I_0} \bar{y}_i f_i(x) = f_0(\bar{x}) + h(x - \bar{x}) + o(|x - \bar{x}|^2),$$

and since (for $f_i(\bar{x}) = 0$)

$$\text{(5.21)} \qquad f_i(x)^2 = ((x - \bar{x}) \cdot \nabla f_i(\bar{x}))^2 + o(|x - \bar{x}|^2),$$

we need only prove that the expansion

$$\text{(5.22)} \qquad \theta(f_i(x))^2 = \theta((x - \bar{x}) \cdot \nabla f_i(\bar{x}))^2 + o(|x - \bar{x}|^2)$$

is valid when $f_i(\bar{x}) = 0$. This amounts to establishing that

$$\text{(5.23)} \qquad 0 = \lim_{t \downarrow 0} \frac{\theta(f_i(\bar{x} + tz))^2 - \theta(tz \cdot \nabla f_i(\bar{x}))^2}{t^2}$$

uniformly in $z \in B$, where

$$B = \{z \in R^n | |z| = 1\}.$$

But the latter is obvious from the continuity of θ and the fact that the difference quotient in (5.23) can be rewritten as

$$\theta(f_i(\bar{x} + tz)/t)^2 - \theta(z \cdot \nabla f_i(\bar{x}))^2 = \theta(z \cdot \nabla f_i(\bar{x}) + w(tz))^2 - \theta(z \cdot \nabla f_i(\bar{x}))^2,$$

where $w(tz) \to 0$ uniformly in $z \in B$ as $t \downarrow 0$.

We now demonstrate the existence of $\bar{r} > 0$ and $\delta > 0$ with

$$\text{(5.24)} \qquad h(z) + \bar{r}k(z) \geqq 2\delta \quad \text{for all } z \in B.$$

Let $B_0 = \{z \in B | h(z) \leqq 0\}$. According to part (c) of the sufficiency conditions, if $k(z) = 0$, i.e., (5.3) holds, we have $h(z) > 0$. Thus $k(z) > 0$ for all $z \in B_0$, implying that the quotient $-h(z)/k(z)$ is well-defined and bounded above as a function of $z \in B_0$. Choose any $\bar{r} > 0$ such that

$$\bar{r} > -h(z)/k(z) \quad \text{for all } z \in B_0.$$

Then $h(z) + \bar{r}k(z) > 0$ for all $z \in B_0$; the same inequality also holds trivially for $z \in B \diagdown B_0$, because there $h(z) > 0$ and $k(z) \geqq 0$. Thus $h + \bar{r}k$ is a positive, continuous function on the compact set B, and (5.24) is valid for some $\delta > 0$ as claimed. Of course (5.24) implies

$$\text{(5.25)} \qquad h(z) + \bar{r}k(z) \geqq 2\delta |z|^2 \quad \text{for all } z \in R^m,$$

because h and k are both positively homogeneous of degree 2.

It remains only to combine (5.25) with (5.18). There exists by (5.18) a neighborhood N of $\bar{x}$, $N \subset N_1(\bar{r})$, such that

$$L(x,\bar{y},\bar{r}) \geqq f_0(\bar{x}) + h(x - \bar{x}) + \bar{r}k(x - \bar{x}) - \delta|x - \bar{x}|^2 \tag{5.26}$$

for all $x \in N$. Using (5.25), we obtain

$$L(x,\bar{y},\bar{r}) \geqq f_0(\bar{x}) + \delta(x - \bar{x})^2 \quad \text{for all } x \in N. \tag{5.27}$$

Thus (5.10) holds for N and $\bar{r}$, and the proof of Theorem 6 is complete.

COROLLARY 6.1. *Under the assumptions in Theorem* 6, *the global saddle-point condition* (4.13) *holds for all* $\bar{r}$ *sufficiently large.*

Proof. The proof is immediate from Corollary 5.2.

We conclude this section with a counterexample demonstrating the need for the second order condition in the hypothesis of Theorem 6.

Example 2. Here all the assumptions in Theorem 6 are satisfied, except for a slight weakening of part (c) of the sufficiency conditions, and (P) is stable of degree 1. But (P) is not stable of degree 2. The problem consists of minimizing

$$f_0(x_1, x_2) = 4x_1(x_2 - 1) + x_2^4$$

over $S = \{x = (x_1, x_2) \in R^2 | -1 \leqq x_1 \leqq 1\}$ subject to

$$0 \geqq f_1(x_1, x_2) = x_1.$$

The minimum of $f_0(x_1, x_2)$ in x_2 for fixed x_1 is $-4x_1 - 3x_1^{4/3}$, attained only at $x_2 = -x_1^{1/3}$, and this minimum is a strictly decreasing function of x_1 as long as $x_1 \geqq -1$. Thus $\bar{x} = (0,0)$ is the unique optimal solution to (P) in the strong sense. The quadratic growth condition is satisfied, because f_0 is bounded below on S. Furthermore, the Kuhn–Tucker conditions hold at $\bar{x}$ with $\bar{y}_1 = 4$ and with the gradients $\nabla f_0(\bar{x})$ and $\nabla f_1(\bar{x})$ nonzero (thus one has "strict complementarity" in (4.17), and moreover "the gradients of the active constraints at $\bar{x}$ form a linearly independent set"). Although the Hessian matrix H of the function $l(x) = f_0(x) + \bar{y}_1 f_1(x)$ at $\bar{x}$ does not have the positive definiteness property required in (c) of the sufficiency conditions, it is true at least that $l(\bar{x} + z) > l(\bar{x})$ for every nonzero z such that (5.3) holds (i.e., $z \cdot \nabla f_1(\bar{x}) = 0$). However,

$$p(u_1) = -4u_1 - 3u_1^{4/3} \quad \text{for } u_1 \in [-1, 1].$$

The function p is continuously differentiable around $u_1 = 0$, but it does not majorize near 0 any function π of class C^2 such that $\pi(0) = p(0) = 0$. Thus (P) is stable of degree 1 but not of degree 2.

Remark. We have already noted towards the end of § 4 that, if $(\bar{x}, \bar{y}, \bar{r})$ is a saddle point of L and the functions f_i are differentiable at $\bar{x}$ (and $\bar{x} \in \operatorname{int} S$), then $\bar{x}$ and $\bar{y}$ satisfy the Kuhn–Tucker conditions. In fact, if every f_i is twice-differentiable at $\bar{x}$, then the standard *second order* necessary conditions [5, p. 25] are satisfied, i.e., besides the Kuhn–Tucker conditions one has condition (c) at the beginning of this section, but with the inequality $z \cdot Hz > 0$ weakened to $z \cdot Hz \geqq 0$. This is true because (5.15) holds (with $\bar{r}$ in place of r) for all x in some neighborhood of $\bar{x}$, so that the right side of (5.15) must have a local minimum at $x = \bar{x}$. From (5.21) and (5.22), it is clear that the latter implies

$$0 \leqq \frac{d^2}{dt^2}[f_0(\bar{x} + tz) + \sum_{i \in I_0} \bar{y}_i f_i(\bar{x} + tz)]|_{t=0} \tag{5.28}$$

for all z satisfying (5.3), and this derivative equals $z \cdot Hz$.

Thus for twice-differentiable functions f_i and S open, the situation can be summarized as follows. *If* $(\bar{x}, \bar{y}, \bar{r})$ *is a saddle point of* L *for some* $\bar{r} \geqq 0$, *then* $\bar{x}$ *and* $\bar{y}$ *satisfy the standard second order necessary conditions for optimality, and* $\bar{x}$ *is (globally) optimal. On the other hand, if* $\bar{x}$ *and* $\bar{y}$ *satisfy the standard second order sufficient conditions and* $\bar{x}$ *is the unique (globally) optimal solution in the strong sense, and the quadratic growth condition is satisfied, then* $(\bar{x}, \bar{y}, \bar{r})$ *is a saddle point of* L *for some* $\bar{r} \geqq 0$.

REFERENCES

[1] K. J. ARROW, P. J. GOULD AND S. M. HOWE, *A general saddle point result for constrained optimization*, Institute of Statistics Mimeo Series No. 774, Univ. of North Carolina, Chapel Hill, 1971. (Revised version: Nov. 1972.)

[2] K. J. ARROW AND R. M. SOLOW, *Gradient methods for constrained maxima, with weakened assumptions*, Studies in Linear and Nonlinear Programming, K. Arrow, L. Hurwicz and H. Uzawa, eds., Stanford Univ. Press, Stanford, Calif., 1958.

[3] D. P. BERTSEKAS, *Combined primal-dual and penalty methods for constrained minimization*, this Journal, 13 (1975), to appear.

[4] J. D. BUYS, *Dual algorithms for constrained optimization*, Thesis, Leiden, 1972.

[5] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Optimization Techniques*, John Wiley, New York, 1968.

[6] R. FLETCHER, *A class of methods for nonlinear programming with termination and convergence properties*, Integer and Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1970.

[7] ———, *A class of methods for non-linear programming. III: Rates of convergence*, Numerical Methods for Non-Linear Optimization, F. A. Lootsma, ed., Academic Press, New York, 1973.

[8] R. FLETCHER AND S. LILL, *A class of methods for nonlinear programming, II: Computational experience*, Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York, 1970.

[9] P. C. HAARHOFF AND J. D. BUYS, *A new method for the optimization of a nonlinear function subject to nonlinear constraints*, Comput. J., 13 (1970), pp. 178–184.

[10] M. R. HESTENES, *Multiplier and gradient methods*, J. Optimization Theory Appl., 4 (1969), pp. 303–320.

[11] B. W. KORT AND D. P. BERTSEKAS, *A new penalty function method for constrained minimization*, Proc. IEEE Decision and Control Conference, New Orleans, 1972.

[12] S. A. LILL, *Generalization of an exact method for solving equality constrained problems to deal with inequality constraints*, Numerical Methods for Nonlinear Optimization, F. A. Lootsma, ed., Academic Press, New York, 1973.

[13] O. L. MANGASARIAN, *Unconstrained Lagrangians in nonlinear programming*, Computer Sciences Tech. Rep. 174, Univ. of Wisconsin, Madison, 1973.

[14] A. MIELE, E. E. CRAGG, R. R. IVER AND A. V. LEVY, *Use of the augmented penalty function in mathematical programming, part I*, J. Optimization Theory Appl., 8 (1971), pp. 115–130.

[15] A. MIELE, E. E. CRAGG AND A. V. LEVY, *Use of the augmented penalty function in mathematical programming problems, part II*, Ibid., 8 (1971), pp. 131–153.

[16] A. MIELE, P. E. MOSELEY AND E. E. CRAGG, *A modification of the method of multipliers for mathematical programming problems*, Techniques of Optimization, A. V. Balakrishnan, ed., Academic Press, New York, 1972.

[17] A. MIELE, P. E. MOSELEY, A. V. LEVY AND G. M. COGGINS, *On the method of multipliers for mathematical programming problems*, J. Optimization Theory Appl., 10 (1972), pp. 1–33.

[18] M. A. POLLATSCHEK, *Generalized duality theory in nonlinear programming*, Operations Research Mimeograph Series No. 122, Technion, Haifa, Israel, 1973.

[19] M. J. D. POWELL, *A method for nonlinear optimization in minimization problems*, Optimization, R. Fletcher, ed., Academic Press, New York, 1969.

[20] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, N.J., 1970.

[21] ———, *New applications of duality in convex programming*, Proc. 4th Conference on Probability, Brasov, Romania, 1971. (This is the written version of a talk given at several conferences, including the 7th International Symposium on Mathematical Programming, the Hague, 1970.)

[22] ———, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Prog., to appear.

[23] ———, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optimization Theory Appl., 12 (1973).

[24] ———, *Conjugate Duality and Optimization*, SIAM/CBMS lecture note series, Philadelphia, 1974.

[25] ———, *Penalty methods and augmented Lagrangians in nonlinear programming*, Proc. 5th IFIP Conference on Optimization Techniques, Rome, Springer-Verlag, Berlin, 1973.

[26] S. S. TRIPATHI AND K. S. NARENDRA, *Constrained optimization problems using multiplier methods*, J. Optimization Theory Appl., 9 (1972), pp. 59–70.

[27] A. P. WIERZBICKI, *Convergence properties of a penalty shifting algorithm for nonlinear programming problems with inequality constraints*, Arch. Automat. i Telemech. (1970).

[28] ———, *A penalty function shifting method in constrained static optimization and its convergence properties*, Ibid., 16 (1971), pp. 395–416.

[29] A. P. WIERZBICKI AND A. HATKO, *Computational methods in Hilbert space for optimal control problems with delays*, Proc. 5th IFIP Conference on Optimization Techniques, Rome, Springer-Verlag, Berlin, 1973.

[30] B. T. POLJAK, *Iteration methods using Lagrange multipliers for the solution of extremal problems with constraints of the equality type*, USSR Comput. Math. and Math. Phys., 10 (1970), no. 5, pp. 1098–1106.

SIAM J. CONTROL
Vol. 12, No. 2, May 1974

OPTIMAL CONTROLS WITH PSEUDODELAYS*

J. WARGA†

Abstract. This paper investigates questions of existence and necessary conditions for ordinary, relaxed, and approximate solutions of optimal control problems defined by functional-integral equations of the form

$$y(t) = \int \kappa(d\tau) \int f(t, \tau, s, \xi(y)(\tau), u(s), b)\nu(\tau)(ds)$$

and subject to unilateral restrictions. Here $t, \tau \in T$, $s \in S$, T and S are compact metric spaces, $u : S \to R$ and b are the control function and parameter, κ and $\nu(\tau)$ are given positive Radon measures, and ξ is a given functional transformation. These equations generalize differential and integral equations with delays that are characterized by the additive coupling of the delays of the control functions. The paper also includes a brief discussion of the ways of extending these results to conflicting control problems defined by similar equations containing the additional arguments $u_P(s), b_P$ that refer to adverse controls. The methods, concepts and arguments used here strongly rely on the author's previous work on similar problems in which the controls were not subject to pseudodelays.

1. Introduction. There is a fairly large and growing literature concerning optimal control problems involving delays or, more generally, functional-integral equations (see, e.g., [1]–[2], [4]–[6], [8]–[12]). Two types of such problems have attracted particular attention. The first type involves equations such as

$$dy(t)/dt = f(t, y(h_1(t)), \cdots, y(h_k(t)), u(t)) \quad \text{a.e. in } [t_0, t_1], \tag{1.1}$$

or, more generally,

$$y(t) = \int f(t, \tau, y, u(\tau))\mu(d\tau) \qquad (t \in T), \tag{1.2}$$

where, for each τ, $f(t, \tau, y, u(\tau))$ may depend on the function y and not only on $y(\tau)$, and thus the state function y may be subject to delays or more general transformations. The second, more general, type involves equations such as

$$dy(t)/dt = f(t, y(h_1(t)), \cdots, y(h_k(t)), u(\tilde{h}_1(t)), \cdots, u(\tilde{h}_l(t))) \quad \text{a.e. in } [t_0, t_1], \tag{1.3}$$

or, more generally,

$$y(t) = \int f(t, \tau, y, u)\mu(d\tau) \qquad (t \in T), \tag{1.4}$$

where both the state function y and the control function u may undergo transformations.

For a large class of problems of the first type there exists a reasonably complete theory, including existence theorems for various kinds of optimal solutions (relaxed, ordinary, and approximate), statements about the interrelationship of these solutions, and necessary conditions for both relaxed and ordinary minima;

* Received by the editors April 24, 1973.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115. This research was supported in part by the National Science Foundation under Grant GP-34189.

and this theory encompasses functional restrictions such as, for example, the "unilateral" restrictions of the form $y(t) \in A(t)$ $(t \in T)$, or, more generally, $g(y)(p) \in A(p)$ $(p \in P)$, where P may, but need not, coincide with T. This theory is worked out in detail in [14, Chaps. VII, VIII] for equations of the form

$$(1.5) \qquad y(t) = \int f(t, \tau, \xi(y)(\tau), u(\tau), b)\mu(d\tau) \qquad (t \in T),$$

where T is a compact metric space, μ a nonatomic positive Radon measure in T, u is a control function subject to a restriction of the form $u(t) \in R^{\#}(t)$ μ-a.e., b is a control parameter, the state function y is an element of either $C(T, \mathbb{R}^n)$ or $L^p(\mu, \mathbb{R}^n)$, and ξ is a differentiable transformation between appropriate Banach spaces. Moreover, as we shall see below, the same methods can be applied with relatively small modifications to the general problem defined by equation (1.2).

The situation appears to be quite different for nonlinear problems of the second type. With one exception, all of the published results in this area are concerned with either necessary or sufficient conditions for ordinary minima. The one exception concerns problems defined by equation (1.3) for which it is assumed that each h_i or $\tilde{h}_j$ is either the identity, or an iterate of some function h, or of its inverse h^{-1}. Subject to this special assumption, there exist results [13, pp. 119–120], [14, pp. 402–406] similar to those previously mentioned for problems of the first type, as well as results of Sabbagh [12] applicable to certain special variational problems.

Our present purpose is to derive an analogous theory for a class of problems of the second type, in which the delays or their analogues are additively coupled. Specifically, we consider functional-integral equations of the form

$$(1.6) \qquad y(t) = \int_T \kappa(d\tau) \int_S f(t, \tau, s, \xi(y)(\tau), u(s), b)\nu(\tau)(ds) \qquad (t \in T),$$

where T and S are compact metric spaces and κ (resp. $\nu(\tau)$ $(\tau \in T)$) is a given positive Radon measure in T (resp. S). In its general form, equation (1.6) provides an example of an optimal control problem with state functions y and control functions u defined on different domains, namely T and S. In the special case where $S = T = [t_0, t_1] \subset \mathbb{R}$, δ_β denotes the Dirac measure at β, κ is the Borel measure in T,

$$\xi(y) = (y \circ h_1, \cdots, y \circ h_k), \qquad \nu(\tau) = \sum_{j=1}^{l} \delta_{\tilde{h}_j(\tau)}$$

and

$$f(t, \tau, \cdot, \cdot, \cdot, \cdot) = \begin{cases} \tilde{f}(\tau, \cdot, \cdot, \cdot, \cdot), & \tau \leqq t, \\ 0, & \tau > t, \end{cases}$$

equation (1.6) reduces to the functional-differential equation

$$(1.7) \qquad \frac{dy(\tau)}{d\tau} = \sum_{j=1}^{l} \tilde{f}(\tau, \tilde{h}_j(\tau), y(h_1(\tau)), \cdots, y(h_k(\tau)), u(\tilde{h}_j(\tau)), b) \qquad \text{a.e. in } T.$$

In particular, if $\tilde{h}_j(\tau) \leqq \tau$ $(j = 1, 2, \cdots, l, \tau \in T)$, then it is proper to say that the control function in (1.7) is subjected to delays. By analogy, we refer to the control function as it appears in (1.6) as a *control with pseudodelays*. Thus a pseudodelay is a transformation which, in special cases, may involve discrete or continuous delays and advances but which can also be defined in situations where there is no past or future.

We shall limit our present investigations to the case where the state functions y are continuous. However, entirely parallel arguments can be applied, on the pattern of [14, Chap. VIII], in the case where y is chosen from L^p or even Orlicz spaces.

Our approach to equation (1.6) is conceptually quite simple. A relatively simple assumption about κ and v (Assumption 2.1(i)) ensures that there exist a nonatomic Radon measure μ in S and Radon measures $\lambda(s)$ $(s \in S)$ in T such that

$$\int \kappa(d\tau) \int h(\tau, s)v(\tau)(ds) = \int \mu(ds) \int h(\tau, s)\lambda(s)(d\tau)$$

for appropriate functions h, and thus equation (1.6) can be rewritten in the form

$$y(t) = \int \phi(t, s, y, u(s), b)\mu(ds) \qquad (t \in T), \tag{1.8}$$

where

$$\phi(t, s, y, u(s), b) = \int f(t, \tau, s, \xi(y)(\tau), u(s), b)\lambda(s)(d\tau).$$

We then study equation (1.8) in much the same way as we studied the equation

$$y(t) = \int f(t, \tau, \xi(y)(\tau), u(\tau), b)\mu(d\tau) \qquad (t \in T),$$

in [14, Chap. VII].

Existence theorems for optimal solutions (ordinary, relaxed, and approximate) are presented in § 2. In § 3 we discuss necessary conditions for ordinary and relaxed optimal solutions, as well as necessary conditions for the existence of a strict $\mathscr{U}$-solution, that is, an optimal ordinary solution that is not optimal among relaxed solutions. In § 4 we discuss, without going into detail, ways of extending these results to conflicting control problems with "simultaneous" pseudodelays. Finally, the proofs appear in § 5.

2. Definitions and existence theorems. We shall use, with slight modifications, the notation of [14], and we briefly summarize the pertinent portions of it here. If X is a compact metric space, we denote by $C(X, \mathbb{R}^n)$ the Banach space of continuous functions on X to $\mathbb{R}^n$ with the sup norm $|\cdot|_{\sup}$, by $\Sigma_{\text{Borel}}(X)$ the σ-field of Borel subsets of X, and by frm (X) the set of all Radon measures in X identified with $C(X)^*$ (the topological dual of $C(X) = C(X, \mathbb{R})$) and endowed with the weak star topology; then frm$^+$ (X) (resp. rpm (X)) represents the subsets of frm (X) whose elements are positive (resp. probability) measures. We denote by $C(X, W)$ the set of all $h \in C(X, \mathbb{R}^n)$ such that $h(X) \subset W \subset \mathbb{R}^n$. For given $\lambda \in \text{frm}^+(X)$, $p \in [1, \infty]$ and $m \in \{1, 2, \cdots\}$, we represent by $L^p(\lambda, \mathbb{R}^m)$ the usual space $L^p(X, \Sigma_{\text{Borel}}(X), \lambda, \mathbb{R}^m)$, by $L^p(\lambda, V)$ the set of all $h \in L^p(\lambda, \mathbb{R}^m)$ such that $h(x) \in V$ λ-a.e., and we write $L^p(\lambda)$ for $L^p(\lambda, \mathbb{R})$. We denote by A° the interior and by $\bar{A}$ the closure of a set A, by $\mathscr{P}'(Z)$ (resp. $\mathscr{K}(Z)$) the space of all nonempty (resp. compact and nonempty) subsets of a metric space Z (with the topology of the Hausdorff semimetric (resp. metric)), and by $d[x, A]$ the distance from a point x to a set A. We also write $\triangleq$ for "equal by definition" or "is defined as".

We assume given compact metric spaces P, R, S and T, a set B of "control parameters" which is a convex subset of a real vector space with a sequentially compact topology in B, closed sets $W \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^m$, a nonzero measure $\kappa \in \operatorname{frm}^+(T)$, and functions

$$\xi : C(T, W) \to L^\infty(\kappa, V), \quad \nu : T \to \operatorname{frm}^+(S), \quad A : P \to \mathscr{K}(\mathbb{R}^{m_2}),$$

$$f : T \times T \times S \times V \times R \times B \to \mathbb{R}^n, \quad g = (g_0, g_1, g_2) : C(T, V) \times B \to \mathbb{R} \times \mathbb{R}^m \times C(P, \mathbb{R}^{m_2}).$$

We now state our first set of assumptions.

ASSUMPTION 2.1.

(i) $\nu : T \to \operatorname{frm}^+(S)$ is κ-measurable [or, equivalently, $\tau \to \int c(s)\nu(\tau)(ds)$ is κ-measurable for each $c \in C(S)$], $0 < \kappa\text{-ess sup}\, \nu(\tau)(S) < \infty$, $\int \nu(\tau)(\{s\})\kappa(d\tau) = 0$ for each $s \in S$;

(ii) $f(\cdot, \cdot, \cdot, v, r, b)$ is Borel measurable on $T \times T \times S$ for each $(v, r, b) \in V \times R \times B$; $f(t, \tau, s, \cdot, \cdot, \cdot)$ is continuous on $V \times R \times B$ for each $(t, \tau, s) \in T \times T \times S$;

$$\sup_{t \in T} \int \kappa(d\tau) \int |f(t, \tau, s, \cdot, \cdot, \cdot)|_{\sup} \nu(\tau)(ds) < \infty;$$

$$\lim_{t' \to t} \int \kappa(d\tau) \int |f(t, \tau, s, \cdot, \cdot, \cdot) - f(t', \tau, s, \cdot, \cdot, \cdot)|_{\sup} \nu(\tau)(ds) = 0 \qquad (t \in T);$$

(iii) g and ξ are continuous.

We can deduce from Assumption 2.1(i) that the relation

$$\mu(E) \triangleq \int \nu(\tau)(E)\kappa(d\tau) \qquad [E \in \Sigma_{\text{Borel}}(S)],$$

defines a nonzero nonatomic measure $\mu \in \operatorname{frm}^+(S)$. (The proof of this assertion follows from [14, X.1.1, p. 482] and the observation that a measure in $\operatorname{frm}^+(S)$ is atomic only if it has point atoms.) We assume given a (set-valued) mapping $R^\# : S \to \mathscr{P}'(R)$ and an "abundant set" $\mathscr{U}$ of "original control functions" as defined in [14, IV.3, pp. 279–281] (with T replaced by S), but suggest that the reader who is unfamiliar with these definitions restrict himself to the important special case[1] where $R^\#$ is a μ-measurable mapping on S to $\mathscr{K}(R)$ and $\mathscr{U}$ is the set of all μ-measurable selections of $R^\#$ (that is, functions $u : S \to R$ such that $u(s) \in R^\#(s)$ μ-a.e.). We define $\mathscr{S}^\#$ as the set of all μ-measurable $\sigma : S \to \operatorname{rpm}(R)$ such that $\sigma(\overline{R^\#(s)}) = 1$ μ-a.e. The original control functions are imbedded in $\mathscr{S}^\#$ by identifying each $u : S \to R$ with the function $s \to \delta_{u(s)}$ (where, as before, δ_r is the

[1] In other special cases, an "abundant set" $\mathscr{U}$ of original control functions may be restricted to include only μ-simple or μ-piecewise continuous selections of $R^\#$, and under certain conditions [14, Remark, p. 287] we may even choose $\mathscr{U}$ as the set of all continuous $u : T \to R$.

Dirac measure at r). The set $\mathscr{S}^{\#}$ is, in turn, embedded in $L^1(\mu, C(R))^*$ and endowed with its relative weak star topology [14, Chap. IV]; the embedding is accomplished by identifying each $\sigma \in \mathscr{S}^{\#}$ with the linear functional $h \to \int \mu(dt) \int h(t, r)\sigma(t)(dr)$.

The optimal control problem that we consider is specified, in its "original" form, by setting

$$F(y, u, b)(t) \triangleq \int \kappa(d\tau) \int f(t, \tau, s, \xi(y)(\tau), u(s), b)\nu(\tau)(ds)$$

$$(y \in C(T, W), u \in \mathscr{U}, b \in B, t \in T),$$

if this relation defines $F(y, u, b)$ as an element of $C(T, W)$ and, otherwise, setting

$$F(y, u, b)(t) \triangleq y(t) + (1, 0, \cdots, 0) \in \mathbb{R}^n;$$

and by letting

$$\mathscr{H}(\mathscr{U}) \triangleq \{(y, u, b) \in C(T, W) \times \mathscr{U} \times B | y = F(y, u, b)\}$$

and

$$\mathscr{A}(\mathscr{U}) \triangleq \{(y, u, b) \in \mathscr{H}(\mathscr{U}) | g_1(y, b) = 0, g_2(y, b)(p) \in A(p)(p \in P)\}.$$

A triplet $(\bar{y}, \bar{u}, \bar{b}) \in \mathscr{A}(\mathscr{U})$ is a *minimizing $\mathscr{U}$-solution* if

$$g_0(\bar{y}, \bar{b}) = \inf \{g_0(y, b) | (y, u, b) \in \mathscr{A}(\mathscr{U})\}.$$

A sequence $((y_j, u_j, b_j))$ in $\mathscr{H}(\mathscr{U})$ is an *approximate $\mathscr{U}$-solution* if

$$\lim_j (|g_1(y_j, b_j)| + \sup_{p \in P} d[g_2(y_j, b_j)(p), A(p)]) = 0.$$

An approximate $\mathscr{U}$-solution $((\bar{y}_j, \bar{u}_j, \bar{b}_j))$ is a *minimizing approximate $\mathscr{U}$-solution* if

$$\lim_j g_0(\bar{y}_j, \bar{b}_j) \leqq \liminf_j g_0(y_j, b_j)$$

for every approximate $\mathscr{U}$-solution $((y_j, u_j, b_j))$.

We introduce the relaxed optimal control problem by setting

$$f(t, \tau, s, v, \sigma(s), b) \triangleq \int f(t, \tau, s, v, r, b)\sigma(s)(dr)$$

$$((t, \tau, s, v, b) \in T \times T \times S \times V \times B, \sigma \in \mathscr{S}^{\#}),$$

and defining $F(y, \sigma, b)$, $\mathscr{H}(\mathscr{S}^{\#})$, $\mathscr{A}(\mathscr{S}^{\#})$ and a minimizing $\mathscr{S}^{\#}$-solution just as before but with $\mathscr{U}, u$ replaced by $\mathscr{S}^{\#}, \sigma$, respectively. We refer to a minimizing $\mathscr{S}^{\#}$-solution as a *minimizing relaxed solution.*

We can now state our first existence theorem.

THEOREM 2.2. *Let Assumption* 2.1 *be satisfied and $\mathscr{A}(\mathscr{S}^{\#})$ be nonempty. Then there exists a minimizing relaxed solution $(\bar{y}, \bar{\sigma}, \bar{b})$.*

If, furthermore, $\bar{y}$ is the unique solution in $C(T, W)$ of the equation $y = F(y, \bar{\sigma}, \bar{b})$ and there exists a neighborhood G of $\bar{\sigma}$ in $\mathscr{S}^{\#}$ such that the equation $y = F(y, u, \bar{b})$ has a (*not necessarily unique*) *solution y for all $u \in \mathscr{U} \cap G$, then there exists a minimizing approximate $\mathscr{U}$-solution $((y_j, u_j, \bar{b}))$ such that* $\lim_j (y_j, u_j) = (\bar{y}, \bar{\sigma})$ *in $C(T, W) \times \mathscr{S}^{\#}$.*

The applicability of the second part of Theorem 2.2 is contingent on our ability to determine whether the equation $y = F(y, \sigma, b)$ has a solution y for fixed σ and b, and whether such a solution is unique. The uniqueness question also comes up in the necessary conditions that we discuss in § 3. We refer the reader to Hale's book [7] for results pertaining to functional-differential equations, and to [14, II.5.A-C, pp. 200–210], where questions of existence are discussed for fairly general functional-integral equations and uniqueness criteria are established for "hereditary" functional-integral equations over an arbitrary compact metric space T. (These equations are generalizations of Volterra-type integral equations over an interval.)

The next theorem describes certain convexity conditions that ensure the existence of a minimizing $\mathscr{U}$-solution, and it generalizes the well-known existence theorem of Filippov [3, Thm. 1, p. 76], as well as [14, Thm. VII.1.4, p. 410] that applies to problems defined by equation (1.5). Before stating this theorem, we observe that it follows from Assumption 2.1(i) (see, e.g., [14, X.1.1–X.1.3, pp. 482–487]) that there exist $\gamma \in \mathrm{frm}^+ (T \times S)$ and a μ-measurable $\lambda : S \to \mathrm{rpm}(T)$ such that

$$\begin{aligned}\int h(t,s)\gamma(d(t,s)) &= \int \kappa(dt) \int h(t,s)\nu(t)(ds) \\ &= \int \mu(ds) \int h(t,s)\lambda(s)(dt) \qquad (h \in L^1(\gamma)).\end{aligned} \tag{2.3}$$

THEOREM 2.4. *Let Assumption* 2.1 *be verified,* $R^{\#} : S \to \mathscr{K}(R)$ *be a* μ*-measurable (set-valued) mapping,* $\mathscr{U}$ *the set of all* μ*-measurable selections of* $R^{\#}$*, and* $\mathscr{S}^{\#}$ *the set of all* μ*-measurable* $\sigma : S \to \mathrm{rpm}(R)$ *with* $\sigma(s)(R^{\#}(s)) = 1$ μ*-a.e. Let*

$$\phi(t,s,y,r,b) \triangleq \int f(t,\tau,s,\xi(y)(\tau),r,b)\lambda(s)(d\tau)$$
$$[(t,s,r,b) \in T \times S \times R \times B,\ y \in C(T,W)],$$

and assume that $\mathscr{A}(\mathscr{S}^{\#})$ *is nonempty and there exists* $\beta \in \mathrm{frm}^+ (T)$ *such that* $\beta(E) > 0$ *for every open* $E \subset T$ *and, for* μ*-a.a.* $s \in S$ *and all* $b \in B$ *and* $y \in C(T, W)$*, the set*

$$\{\phi(\cdot, s, y, r, b) \mid r \in R^{\#}(s)\}$$

is a convex subset of the vector space of all (equivalence classes of) β*-measurable functions on* T *to* $\mathbb{R}^n$.

Then for every choice of a minimizing relaxed solution $(\bar{y}, \bar{\sigma}, \bar{b})$ *(of which at least one must exist by Theorem* 2.2*), there exists* $\bar{u} \in \mathscr{U}$ *such that* $(\bar{y}, \bar{u}, \bar{b})$ *is a minimizing relaxed solution (and, a fortiori, a minimizing* $\mathscr{U}$*-solution).*

Remark. It is easily verified that the assumption about the existence of β can be replaced by the assumption that there exists a dense subset T' of T such that, for μ-a.a. $s \in S$ and all $b \in B$ and $y \in C(T, W)$, the set of functions

$$\{\phi(\cdot, s, y, r, b)|_{T'} | r \in R^{\#}(s)\}$$

is convex, where $\phi(\cdot, s, y, r, b)|_{T'}$ is the restriction to T'. Indeed, given such a set T', we may choose a dense denumerable subset $\{\tau_1, \tau_2, \cdots\}$ and let

$$\beta(E) \triangleq \sum_{\tau_j \in E} 2^{-j} \qquad [E \in \Sigma_{\mathrm{Borel}}(T)].$$

3. Necessary conditions for relaxed and original minimum. In order to discuss necessary conditions for a minimum, we shall require some additional definitions and assumptions. We shall apply, in particular, the concept of a (Fréchet) derivative relative to a set [14, p. 167]. Specifically, if $\mathscr{X}$ and $\mathscr{Y}$ are Banach spaces, A a convex subset of $\mathscr{X}$ with a nonempty interior, $a \in A$, and $h: A \to \mathscr{Y}$, then we say that $h'(a)$ is the derivative of h at a whenever $h'(a) \in B(\mathscr{X}, \mathscr{Y})$ (the Banach space of continuous linear operators on $\mathscr{X}$ to $\mathscr{Y}$) and

$$\lim |\Delta x|^{-1}|h(a + \Delta x) - h(a) - h'(a)\,\Delta x| = 0 \quad \text{as} \quad \Delta x \to 0, \Delta x \in A - a, \Delta x \neq 0.$$

We define partial derivatives accordingly, and write $h_a(a, b)$ or $\mathscr{D}_1 h(a, b)$ and $h_b(a, b)$ or $\mathscr{D}_2 h(a, b)$ for the derivatives of $h(\cdot, b)$ at a and of $h(a, \cdot)$ at b. If A is a convex subset of a vector space, $\mathscr{Y}$ a Banach space, $\bar{a}, a \in A$, and $h: A \to \mathscr{Y}$, then we denote by

$$Dh(\bar{a}; a - \bar{a}) \triangleq \lim_{\alpha \to +0} \alpha^{-1}[h(\bar{a} + \alpha(a - \bar{a})) - h(\bar{a})]$$

the directional derivative of h at $\bar{a}$ in the direction of a. We write $D_1 h(\bar{a}, \bar{b}; a - \bar{a})$ and $D_2 h(\bar{a}, \bar{b}; b - \bar{b})$ for the partial directional derivatives of a function $h(\cdot, \cdot)$. Finally, we denote by $\mathscr{T}_{m+1}$ the simplex

$$\left\{(\theta^0, \cdots, \theta^m) \in \mathbb{R}^{m+1} \,|\, \theta^i \geqslant 0, \sum_{j=0}^{m} \theta^j \leqq 1\right\},$$

and define $s \to \lambda(s): S \to \operatorname{rpm}(T)$ as in (2.3).

THEOREM 3.1. *Let* $(\bar{y}, \bar{\sigma}, \bar{b})$ *be a minimizing relaxed solution, and assume that*

(i) ξ *has a continuous derivative*;

(ii) v *satisfies Assumption* 2.1 (*i*);

(iii) $A(p)$ *is a convex body* (*i.e., a closed convex set with a nonempty interior*) *for each* $p \in P$ *and the set*

$$G(A^\circ) \triangleq \{(p, v) \in P \times \mathbb{R}^{m_2} \,|\, v \in A(p)^\circ\}$$

is an open subset of $P \times \mathbb{R}^{m_2}$;

(iv) *for each choice of* $L \triangleq (b_0, \cdots, b_m) \in B^{m+1}$ *there exist a closed convex neighborhood* $V^L \subset V$ *and a closed convex neighborhood* $\mathscr{T}^L$ *of* 0 *in* $\mathscr{T}_{m+1}$ *such that* $\xi(\bar{y})$ *is in the interior of* $L^\infty(\kappa, V^L)$ *and the functions*

$$(t, \tau, s, v, r, \theta) \to f^L(t, \tau, s, v, r, \theta) \triangleq f\left(t, \tau, s, v, r, \bar{b} + \sum_{j=0}^{m} \theta^j(b_j - \bar{b})\right):$$

$$T \times T \times S \times V^L \times R \times \mathscr{T}^L \to \mathbb{R}^n$$

and

$$(y, \theta) \to g^L(y, \theta) \triangleq g\left(y, \bar{b} + \sum_{j=0}^{m} \theta^j(b_j - \bar{b})\right) : \xi^{-1}(L^\infty(\kappa, V^L)) \times \mathscr{T}^L$$

$$\to \mathbb{R} \times \mathbb{R}^m \times C(P, \mathbb{R}^{m_2})$$

have the following properties:

(a) $f^L(t, \tau, s, \cdot, r, \cdot)$ *has a derivative* $f^L_{(v,\theta)}(t, \tau, s, v, r, 0)$ *for all* $(t, \tau, s, v, r, \theta) \in T \times T \times S \times V^L \times R \times \mathscr{T}^L$, *and both* f^L *and* $f^L_{(v,\theta)}$ *satisfy the same conditions* *as f in Assumption* 2.1(ii) (*with b replaced by θ*); *and*

(b) *$g^L(\cdot,\cdot)$ is continuous and has a derivative at* $(\bar{y}, 0)$.

We assume, furthermore, that

(c) *the equation $\Delta y = \mathscr{F}\Delta y$ has $\Delta y = 0$ as its only solution in $C(T, \mathbb{R}^n)$, where $\mathscr{F}$ is defined by*

$$(\mathscr{F}\Delta y)(t) \triangleq \int \kappa(d\tau) \int f_v(t, \tau, s, \xi(\bar{y})(\tau), \bar{\sigma}(s), \bar{b})(\xi'(\bar{y})\,\Delta y)(\tau) v(\tau)(ds)$$

$$(\Delta y \in C(T, \mathbb{R}^n),\ t \in T).$$

Then $I - \mathscr{F}$ has a continuous inverse, $(I - \mathscr{F})^{-1} - I$ is a compact operator in $C(T, \mathbb{R}^n)$, and there exist $l_0 \geqq 0$, $l_1 \in \mathbb{R}^m$, $\zeta \in \mathrm{frm}^+ (T)$, $\tilde{\zeta} \in L^1(\zeta, \mathbb{R}^n)$, $\omega \in \mathrm{frm}^+ (P)$ and $\tilde{\omega} \in L^1(\omega, \mathbb{R}^{m_2})$ such that

$$|\tilde{\zeta}(t)| = 1 \quad (t \in T), \quad |\tilde{\omega}(p)| = 1 \quad (p \in P), \quad l_0 + |l_1| + \omega(T) > 0,$$

$$(1) \qquad \int \tilde{\zeta}(t) \cdot x(t)\zeta(dt) = \sum_{i=0}^{1} l_i \mathscr{D}_1 g_i(\bar{y}, \bar{b})(I - \mathscr{F})^{-1} x + \int \tilde{\omega}(p) \cdot [\mathscr{D}_1 g_2(\bar{y}, \bar{b})(I - \mathscr{F})^{-1} x](p)\omega(dp)$$

$$(x \in C(T, \mathbb{R}^n);$$

$$(2) \qquad \int \zeta(dt) \int \tilde{\zeta}(t) \cdot f(t, \tau, s, \xi(\bar{y})(\tau), \bar{\sigma}(s), \bar{b})\lambda(s)(d\tau) = \min_{r \in R^{\#}(s)} \int \zeta(dt) \int \tilde{\zeta}(t) \cdot f(t, \tau, s, \xi(\bar{y})(\tau), r, \bar{b})\lambda(s)(d\tau)$$

for μ-a.a. $s \in S$;

$$(3) \qquad \int \zeta(dt) \int \mu(ds) \int \tilde{\zeta}(t) \cdot D_6 f(t, \tau, s, \xi(\bar{y})(\tau), \bar{\sigma}(s), \bar{b}; b - \bar{b})\lambda(s)(d\tau) + \sum_{i=0}^{1} l_i \cdot D_2 g_i(\bar{y}, \bar{b}; b - \bar{b}) + \int \tilde{\omega}(p) \cdot D_2 g_2(\bar{y}, \bar{b}; b - \bar{b})(p)\omega(dp) \geqq 0$$

$$(b \in B);$$

$$(4) \qquad \tilde{\omega}(p) \cdot g_2(\bar{y}, \bar{b})(p) = \max_{a \in A(p)} \tilde{\omega}(p) \cdot a \quad \textit{for } \omega\textit{-a.a. } p \in P.$$

THEOREM 3.2. *Let $\mathscr{U}$ be an abundant set* (*or consider the special case where $\mathscr{U}$ is the set of all μ-measurable selections of a μ-measurable $R^{\#} : S \to \mathscr{K}(R)$*), *and let all the assumptions of Theorem* 3.1 *be satisfied except that $(\bar{y}, \bar{\sigma}, \bar{b})$ is assumed to be a minimizing $\mathscr{U}$-solution and not necessarily a minimizing relaxed solution. Assume, furthermore, that for each choice of $L \triangleq (b_0, \cdots, b_m) \in B^{m+1}$, $\sigma \in \mathscr{S}^{\#}$ and $\theta \in \mathscr{T}^L$, the equation*

$$y = F\left(y, \sigma, \bar{b} + \sum_{j=0}^{m} \theta^j (b_j - \bar{b})\right)$$

has a unique solution $y = \tilde{y}^L(\sigma, \theta)$ *in* $\xi^{-1}(L^\infty(T, V^L))$. *Then the conclusions of Theorem* 3.1 *remain valid.*

The next theorem refers to *strict* $\mathcal{U}$*-solutions*, that is, minimizing $\mathcal{U}$-solutions that are not at the same time minimizing relaxed solutions. We shall refer to a point $(\bar{y}, \bar{\sigma}, \bar{b}) \in C(T, W) \times \mathcal{S}^\# \times B$ as *admissible* if it belongs to $\mathcal{A}(\mathcal{S}^\#)$; we shall call it *extremal* if there exist corresponding $l_0, l_1, \omega, \zeta, \tilde{\omega}$ and $\tilde{\zeta}$, as described in the statement of Theorem 3.1, and that satisfy together with $(\bar{y}, \bar{\sigma}, \bar{b})$ the relations (1)–(4) of that theorem. Thus Theorems 3.1 and 3.2 assert that, under appropriate conditions, every minimizing relaxed or $\mathcal{U}$-solution is admissible and extremal (but, as it is well known, the converse is generally not true). We refer to an extremal point as *abnormal* if there exists an appropriate choice of corresponding $l_0, l_1, l_2, \omega, \zeta, \tilde{\omega}, \tilde{\zeta}$ with $l_0 = 0$.

THEOREM 3.3. *Let* $\mathcal{U}$ *be an abundant set,* $(\tilde{y}, \tilde{u}, \tilde{b}) \in \mathcal{A}(\mathcal{U})$,

$$\mathcal{M} \triangleq \{(y, \sigma, b) \in \mathcal{A}(\mathcal{S}^\#) | g_0(y, b) \leqq g_0(\tilde{y}, \tilde{b})\},$$

and assume that every point $(y, \sigma, b) \in \mathcal{M}$ *satisfies all the assumptions made in Theorems* 3.1 *and* 3.2 *about* $(\bar{y}, \bar{\sigma}, \bar{b})$ *except that it need not be a minimizing* ($\mathcal{U}$*- or relaxed*) *solution. Then* $(\tilde{y}, \tilde{u}, \tilde{b})$ *is a strict* $\mathcal{U}$*-solution only if it is extremal, the set*

$$\mathcal{M}^< \triangleq \{(y, \sigma, b) \in \mathcal{A}(\mathcal{S}^\#) | g_0(y, b) < g_0(\tilde{y}, \tilde{b})\}$$

is nonempty, and every element of $\mathcal{M}^<$ *is extremal and abnormal.* (*It follows therefore, under these conditions, that the problem can have a strict* $\mathcal{U}$*-solution only if every minimizing relaxed solution is abnormal.*)

4. Conflicting controls with pseudodelays. We observe that the results of [14, Chaps. IX and X] are applicable in the case of conflicting control problems with both controls subject to "simultaneous" pseudodelays. Let R_P and B_P be compact metric spaces, $R_P^\# : S \to \mathcal{P}'(R_P)$ defined similarly to $R^\#$ but with R_P replacing R, and let $\mathcal{U}_P$ be an abundant set of "adverse" original control functions $u_P : S \to R_P$. Let the "original" conflicting control problem be defined by the equations

$$(1) \qquad y(t) = \int \kappa(d\tau) \int f(t, \tau, s, \xi(y)(\tau), u(s), b) v(\tau)(ds) \qquad (t \in T),$$

$$(2) \qquad \hat{y}(t) = \int \hat{\kappa}(d\tau) \int \hat{f}(t, \tau, s, \hat{\xi}(\hat{y})(\tau), u(s), b, u_P(s), b_P) \hat{v}(\tau)(ds) \qquad (t \in T),$$

We assume that these equations have unique solutions $y(u, b)$ and $\hat{y}(u, b, u_P, b_P)$, and that the problem consists in choosing $(u, b) \in \mathcal{U} \times B$ to minimize $x_0(u, b) \triangleq h_0(y(u, b))$ subject to the restrictions

$$x_1(u, b) \triangleq h_1(y(u, b)) = 0 \in \mathbb{R}^m,$$

$$x_2(u, b)(u_P, b_P) \triangleq h_2(\hat{y}(u, b, u_P, b_P)) \in A \subset \mathbb{R}^{m_2} \quad \text{for all } (u_P, b_P) \in \mathcal{U}_P \times B_P.$$

We may assume that $\kappa = \hat{\kappa}$ and $v = \hat{v}$, otherwise determining the Radon–Nikodym derivatives of κ and $\hat{\kappa}$ with respect to $\kappa + \hat{\kappa}$ and of $v(\tau)$ and $\hat{v}(\tau)$ with respect to $v(\tau) + \hat{v}(\tau)$, and modifying f and $\hat{f}$ accordingly. With this accomplished, we determine the corresponding $\mu \in \text{frm}^+ (S)$ and "relax" the problem, replacing

u by a μ-measurable $\sigma : S \to \operatorname{rpm}(R)$ and u_P by the corresponding hyperrelaxed control function [14, Chap. X]. We can verify (by arguments analogous to those of Theorem 2.2) that the conditions of the existence theorem [14, X.1.8, p. 493] for optimal relaxed and approximate solutions are satisfied when f satisfies Assumption 2.1(ii) and $\hat{f}$ its analogue. It appears that the necessary conditions of [14, X.2.4, p. 499] can be applied by methods analogous to those used for autonomous ordinary differential equations in [14, X.3.5]; but we have not carried out the details of these derivations or determined the exact form of conditions that must be imposed on $f, \hat{f}, \xi$ and h_i.

5. Proofs.

5.1. Proof of Theorem 2.2. Let γ and λ be defined as in (2.3) and ϕ as in the statement of Theorem 2.4. By Assumption 2.1, (2.3), and Fubini's theorem, for each $(y, \sigma, b) \in \mathscr{A}(\mathscr{S}^{\#})$ and $t \in T$ the function $(\tau, s) \to f(t, \tau, s, \xi(y)(\tau), \sigma(s), b)$ is γ-integrable and

$$
\begin{aligned}
y(t) &= \int \kappa(d\tau) \int f(t, \tau, s, \xi(y)(\tau), \sigma(s), b) \nu(\tau)(ds) \\
&= \int \mu(ds) \int \lambda(s)(d\tau) \int f(t, \tau, s, \xi(y)(\tau), r, b) \sigma(s)(dr) \\
&= \int \mu(ds) \int \sigma(s)(dr) \int f(t, \tau, s, \xi(y)(\tau), r, b) \lambda(s)(d\tau) \\
&= \int \mu(ds) \int \phi(t, s, y, r, b) \sigma(s)(dr) \\
&= \int \phi(t, s, y, \sigma(s), b) \mu(ds),
\end{aligned} \tag{1}
$$

where, as we did for f, we write

$$\phi(t, s, y, \sigma(s), b) \quad \text{for} \int \phi(t, s, y, r, b) \sigma(s)(dr).$$

Furthermore, we have

$$
\begin{aligned}
|y(t)| &\leqq \int \kappa(d\tau) \int |f(t, \tau, s, \xi(y)(\tau), \sigma(s), b)| \nu(\tau)(ds) \\
&\leqq \sup_{t' \in T} \int \kappa(d\tau) \int |f(t', \tau, s, \cdot, \cdot, \cdot)|_{\sup} \nu(\tau)(ds) < \infty
\end{aligned}
$$

and

$$|y(t) - y(t')| \leqq \int \kappa(d\tau) \int |f(t, \tau, s, \cdot, \cdot, \cdot) - f(t', \tau, s, \cdot, \cdot, \cdot)|_{\sup} \nu(\tau)(ds) \underset{t' \to t}{\to} 0.$$

Thus the set $Y \triangleq \{y | (y, \sigma, b) \in \mathscr{A}(\mathscr{S}^{\#})\}$ is bounded and equicontinuous and therefore conditionally compact in $C(T, \mathbb{R}^n)$. Since $\mathscr{S}^{\#}$ is sequentially compact [14, IV.3.11, p. 287] and so is, by assumption, the set B, it follows that every sequence in $\mathscr{A}(\mathscr{S}^{\#})$ has a convergent subsequence.

Now let $(y_j, \sigma_j, b_j) \in \mathscr{A}(\mathscr{S}^\#)$ $(j = 1, 2, \cdots)$ and $\lim_j (y_j, \sigma_j, b_j) = (\bar{y}, \bar{\sigma}, \bar{b})$ in $C(T, \mathbb{R}^n) \times \mathscr{S}^\# \times B$. Since V and W are closed, it follows that $\bar{y}(T) \subset W$ and $\xi(\bar{y})(\tau) \in V$ κ-a.e. Thus, for all $(t, s, r) \in T \times S \times R$ and κ-a.a. $\tau \in T$,

$$\lim_j f(t, \tau, s, \xi(y_j)(\tau), r, b_j) = f(t, \tau, s, \xi(\bar{y})(\tau), r, \bar{b}),$$

and, since $f(t, \tau, s, \cdot, \cdot, \cdot)$ is continuous and R is compact, the convergence is uniform for all $r \in R$. It follows, by a variant of the dominated convergence theorem [14, I.4.36, p. 87], that

$$\begin{aligned} &\lim_j [\phi(t, s, y_j, r, b_j) - \phi(t, s, \bar{y}, r, \bar{b})] \\ &\qquad = \lim_j \int [f(t, \tau, s, \xi(y_j)(\tau), r, b_j) - f(t, \tau, s, \xi(\bar{y})(\tau), r, \bar{b})]\lambda(s)(d\tau) \\ &\qquad = 0 \quad \text{for all } (t, s) \in T \times S, \quad \text{uniformly for } r \in R; \end{aligned}$$

hence

$$\lim_j [\phi(t, s, y_j, \sigma_j(s), b_j) - \phi(t, s, \bar{y}, \sigma_j(s), \bar{b})] = 0 \qquad [(t, s) \in T \times S],$$

and, again by the dominated convergence theorem,

$$(2) \qquad \lim_j \int [\phi(t, s, y_j, \sigma_j(s), b_j) - \phi(t, s, \bar{y}, \sigma_j(s), \bar{b})]\mu(ds) = 0 \qquad (t \in T).$$

Next we observe that, for each $(t, y, b) \in T \times C(T, W) \times B$, the function $(s, r) \to \phi(t, s, y, r, b)$ belongs to $L^1(\mu, C(R))$. It follows [14, IV.1.11, p. 272] that

$$\lim_j \int \phi(t, s, \bar{y}, \sigma_j(s), \bar{b})\mu(ds) = \int \phi(t, s, \bar{y}, \bar{\sigma}(s), \bar{b})\mu(ds) \qquad (t \in T).$$

We combine this relation with (1) and (2) and conclude that

$$\begin{aligned} \bar{y}(t) = \lim_j y_j(t) &= \lim_j \int \phi(t, s, y_j, \sigma_j(s), b_j)\mu(ds) \\ &= \int \phi(t, s, \bar{y}, \bar{\sigma}(s), \bar{b})\mu(ds) \\ &= F(\bar{y}, \bar{\sigma}, \bar{b})(t) \qquad (t \in T), \end{aligned}$$

and

$$(g_1(\bar{y}, \bar{b}), g_2(\bar{y}, \bar{b})(p)) = \lim_j (g_1(y_j, b_j), g_2(y_j, b_j)(p)) \in \{0\} \times A(p) \qquad (p \in P).$$

Thus $(\bar{y}, \bar{\sigma}, \bar{b}) \in \mathscr{A}(\mathscr{S}^\#)$, showing that $\mathscr{A}(\mathscr{S}^\#)$ is sequentially compact and $F|\mathscr{A}(\mathscr{S}^\#)$ sequentially continuous. The existence of a minimizing relaxed solution now follows from [14, V.1.1, p. 296].

To prove the second part of the theorem, we observe that the same argument as above shows that $\mathscr{H}(\mathscr{S}^\#)$ is sequentially compact and $F|\mathscr{H}(\mathscr{S}^\#)$ sequentially continuous. Then our conclusion follows from [14, V.1.2, p. 297]. Q.E.D.

5.2. Proof of Theorem 2.3. Let $(\bar{y}, \bar{\sigma}, \bar{b})$ be a minimizing relaxed solution, and

$$h(t, s, r) \triangleq \phi(t, s, \bar{y}, r, \bar{b}).$$

Then the arguments of Theorems IV.3.14 and VII.1.4 of [14, p. 291 and p. 410] apply directly, with (t, τ), $T \times T$, $v \times \mu$ replaced by (t, s), $T \times S$, $\beta \times \mu$, respectively; our conclusion follows from [14, Thm. VII.1.4], as modified. Q.E.D.

5.3. Proof of Theorem 3.1. Let $K \triangleq ((\sigma_0, b_0), \cdots, (\sigma_m, b_m)) \in (\mathscr{S}^{\#} \times B)^{m+1}$, $L \triangleq (b_0, \cdots, b_m)$, $Y_K \triangleq \xi^{-1}(L^\infty(\kappa, V^L))$,

$$\sigma^K(\theta) \triangleq \bar{\sigma} + \sum_{j=0}^{m} \theta^j(\sigma_j - \bar{\sigma}), \qquad b^K(\theta) \triangleq \bar{b} + \sum_{j=0}^{m} \theta^j(b_j - \bar{b}),$$

$$\tilde{f}^K(t, \tau, v, \theta) \triangleq \int f(t, \tau, s, v, \sigma^K(\theta)(s), b^K(\theta))v(\tau)(ds)$$

$$\int f^L(t, \tau, s, v, \sigma^K(\theta)(s), \theta)v(\tau)(ds)$$

$$(t, \tau \in T, v \in V^L, \theta = (\theta^0, \cdots, \theta^m) \in \mathscr{T}^L),$$

and

$$\tilde{F}^K(y, \theta) \triangleq F(y, \sigma^K(\theta), b^K(\theta)) \qquad (y \in Y_K, \theta \in \mathscr{T}^L).$$

Then we verify (essentially as in [14, p. 415, Step 2]) that

$$\tilde{F}^K(y, \theta)(t) = \int \tilde{f}^K(t, \tau, \xi(y)(\tau), \theta)\kappa(d\tau) \qquad (y \in Y_K, \theta \in \mathscr{T}^L, t \in T),$$

and $\tilde{F}^K$ has a continuous derivative such that

$$\mathscr{D}\tilde{F}^K(y, \theta)(\Delta y, \Delta\theta)(t) = \int \tilde{f}^K_{(v.\theta)}(t, \tau, \xi(y)(\tau), \theta)((\xi'(y)\,\Delta y)(\tau), \Delta\theta)\kappa(d\tau)$$

$$(\Delta y \in C(T, \mathbb{R}^n), \Delta\theta \in \mathbb{R}^{m+1}).$$

It follows that $F_y(\bar{y}, \bar{\sigma}, \bar{b}) = \mathscr{F}$,

$$\begin{aligned} D_2F(\bar{y}, (\bar{\sigma}, \bar{b}); (\sigma, b) - (\bar{\sigma}, \bar{b}))(t) = \int \kappa(d\tau) \int [&f(t, \tau, s, \xi(\bar{y})(\tau), \sigma(s) - \bar{\sigma}(s), \bar{b}) \\ &+ D_6f(t, \tau, s, \xi(\bar{y})(\tau), \bar{\sigma}(s), \bar{b}; \\ &b - \bar{b})]v(\tau)(ds) \end{aligned} \tag{5}$$

$$((\sigma, b) \in \mathscr{S}^{\#} \times B, t \in T),$$

and, by [14, II.5.5, p. 207], $\mathscr{F}$ is a compact operator, $I - \mathscr{F}$ has a continuous inverse, and $(I - \mathscr{F})^{-1} - I$ is compact. We may therefore apply [14, V.2.3, p. 303] to conclude that there exists $l = (l_0, l_1, l_2) \in [0, \infty) \times \mathbb{R}^m \times C(P, \mathbb{R}^{m_2})^*$ such that $l \neq 0$,

$$l[\mathscr{D}_1g(\bar{y}, \bar{b}) \circ (I - \mathscr{F})^{-1}D_2F(\bar{y}, (\bar{\sigma}, \bar{b}); (\sigma, b) - (\bar{\sigma}, \bar{b})) + D_2g(\bar{y}, \bar{b}; \bar{b} - \bar{b})] \geqq 0 \tag{6}$$

$$((\sigma, b) \in \mathscr{S}^{\#} \times B),$$

and

$$(7) \qquad l_2(g_2(\bar y, \bar b)) \geqq l_2(c) \qquad (c \in C(P, \mathbb{R}^{m_2}), c(p) \in A(p), p \in P).$$

We can represent l_2 (e.g., [14, I.5.9(2), p. 117]) by $\omega \in \mathrm{frm}^+(P)$ and $\tilde\omega \in L^1(\omega, \mathbb{R}^{m_2})$ such that

$$|\tilde\omega(p)| = 1 \quad (p \in P) \quad \text{and} \quad l_2 x = \int \tilde\omega(p) \cdot x(p)\omega(dp) \qquad (x \in C(P, \mathbb{R}^{m_2})).$$

We may similarly represent $l \circ \mathscr{D}_1 g(\bar y, \bar b) \circ (I - \mathscr{F})^{-1}$ by $\zeta \in \mathrm{frm}^+(T)$ and $\tilde\zeta \in L^1(\zeta, \mathbb{R}^n)$, and these two combined representations and $l \neq 0$ yield statement (1) of the theorem. Statement (4) of the theorem follows from (7) and [14, V.2.5, p. 307]. When we apply these representations and (5) to relation (6), we obtain

$$(8) \qquad \begin{aligned} &\int \zeta(dt) \int \kappa(d\tau) \int \tilde\zeta(t) \cdot [f(t, \tau, s, \xi(\bar y)(\tau), \sigma(s) - \bar\sigma(s), \bar b) \\ &\qquad + D_6 f(t, \tau, s, \xi(\bar y)(\tau), \bar\sigma(s), \bar b; b - \bar b)]\nu(\tau)(ds) \\ &\qquad + \sum_{i=0}^{1} l_i D_2 g_i(\bar y, \bar b; b - \bar b) + \int \tilde\omega(p) \cdot D_2 g_2(\bar y, \bar b; b - \bar b)(p)\omega(dp) \\ &\geqq 0 \qquad ((\sigma, b) \in \mathscr{S}^{\#} \times B). \end{aligned}$$

For $\sigma = \bar\sigma$, this relation and (2.3) yield statement (3) of the theorem.

It remains, therefore, to prove statement (2) of the theorem. We set $b = \bar b$ in (8), and apply (2.3) and Fubini's theorem to obtain

$$(9) \quad \int \mu(ds) \int \zeta(dt) \int \tilde\zeta(t) \cdot f(t, \tau, s, \xi(\bar y)(\tau), \sigma(s) - \bar\sigma(s), \bar b)\lambda(s)(d\tau) \geqq 0 \qquad (\sigma \in \mathscr{S}^{\#}).$$

Because of our choice of $R^{\#}$ (see [14, IV.3.1, IV.3.2, pp. 280–281]; in particular, by a theorem of Castaing [14, I.7.8, p. 152] if $R^{\#}(s)$ is closed μ-a.e.), there exists an at most denumerable set $\{u_1, u_2, \cdots\}$ of μ-measurable selections of $R^{\#}$ such that $\{u_1(s), u_2(s), \cdots\}$ is dense in $R^{\#}(s)$ for μ-a.a. $s \in S$. For each $j = 1, 2, \cdots$ and each μ-measurable set E, we set

$$\sigma(s) = \begin{cases} u_j(s) & (s \in E), \\ \bar\sigma(s) & (s \notin E). \end{cases}$$

Then (9) yields

$$(10) \qquad \begin{aligned} &\int \zeta(dt) \int \tilde\zeta(t) \cdot f(t, \tau, s, \xi(\bar y)(\tau), u_j(s), \bar b)\lambda(s)(d\tau) \\ &\qquad \geqq \int \zeta(dt) \int \tilde\zeta(t) \cdot f(t, \tau, s, \xi(\bar y)(\tau), \bar\sigma(s), \bar b)\lambda(s)(d\tau) \\ &\qquad\qquad\qquad (j = 1, 2, \cdots, \quad \mu\text{-a.a. } s \in S). \end{aligned}$$

Since the function $r \to \int\zeta(dt)\int\tilde\zeta(t) \cdot f(t, \tau, s, \xi(\bar y)(\tau), r, \bar b)\lambda(s)(d\tau)$ is continuous for μ-a.a. $s \in S$ (as a consequence of the Lebesgue dominated convergence theorem) and $\{u_1(s), u_2(s), \cdots\}$ is dense in $R^{\#}(s)$ for μ-a.a. $s \in S$, statement (2) of the theorem follows from (10). Q.E.D.

5.4. Proof of Theorem 3.2. We shall continue to use the notation introduced in the proof of Theorem 3.1. We shall also write

$$F^K(y, \sigma, \theta) \triangleq F(y, \sigma, b^K(\theta))$$

and

$$\hat{f}^K(t, \tau, v, \sigma, \theta) \triangleq \int f(t, \tau, s, v, \sigma(s), b^K(\theta))v(\tau)(ds).$$

If we replace B, f by $\mathscr{T}_K, \hat{f}^K$ in the proof of Theorem 2.2, then this argument effectively shows that $F^K|Y_K \times \mathscr{S}^\# \times \mathscr{T}_K$ is continuous. (This argument remains unchanged if we choose the convergent sequence $((y_j, \sigma_j, \theta_j))$ in $Y_K \times \mathscr{S}^\# \times \mathscr{T}_K$, its limit remaining in the same set because $Y_K \times \mathscr{S}^\# \times \mathscr{T}_K$ is closed.) Furthermore, the same argument shows that the set $\tilde{y}^L(\mathscr{S}^\# \times \mathscr{T}_K)$ is conditionally compact in $C(T, \mathbb{R}^n)$. These observations and the properties of $\tilde{F}^K$ derived in the proof of Theorem 3.1 show that [14, V.3.2, p. 310] is applicable and there exists, therefore, a nonzero $l = (l_0, l_1, l_2) \in [0, \infty) \times \mathbb{R}^m \times C(P, \mathbb{R}^{m_2})^*$ satisfying relations (6) and (7) in the proof of Theorem 3.1. From this point on, our arguments proceed as in that proof except that the set $\{u_1, u_2, \cdots\}$ (with $\{u_1(s), u_2(s), \cdots\}$ dense in $R^\#(s)$ μ-a.e.) is chosen out of $\mathscr{U}$. Q.E.D.

5.5. Proof of Theorem 3.3. By [14, V.3.4, p. 314], $\mathscr{M}^<$ is nonempty and every element $(\bar{y}, \bar{\sigma}, \bar{b})$ of it is extremal and abnormal in the sense of [14, Def. V.2.0, p. 298]; that is, there exists a corresponding nonzero $l = (l_0, l_1, l_2) \in [0, \infty) \times \mathbb{R}^m \times C(P, \mathbb{R}^{m_2})^*$ such that relations (6) and (7) in the proof of Theorem 3.1 are satisfied and $l_0 = 0$. It follows then, as in the proof of Theorem 3.1, that $(\bar{y}, \bar{\sigma}, \bar{b})$ is extremal and abnormal in the sense defined here. Q.E.D.

REFERENCES

[1] H. T. BANKS, *Necessary conditions for control problems with variable time lags*, this Journal, 6 (1968), pp. 9–47.

[2] ———, *Variational problems involving functional differential equations*, this Journal, 7 (1969), pp. 1–17.

[3] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.

[4] R. GABASOV AND S. V. CHURAKOVA, *Necessary optimality conditions in time-lag systems*, Automat. Remote Control, 1968, pp. 37–54.

[5] ———, *Sufficient conditions for optimality in systems with a delay*, Ibid., 1968, pp. 193–209.

[6] A. HALANAY, *Optimal controls for systems with time-lag*, this Journal, 6 (1968), pp. 215–234.

[7] J. K. HALE, *Functional differential equations*, Applied Mathematical Sciences, vol. 3, Springer-Verlag, New York, 1971.

[8] D. K. HUGHES, *Variational and optimal control problems with delayed argument*, J. Optimization Theory Appl., 2 (1968), pp. 1–14.

[9] G. A. KAMENSKII AND E. A. KHVILON, *Necessary conditions for the optimal control of systems with deviating argument of neutral type*, Automat. Remote Control, 1969, pp. 327–329.

[10] G. A. KENT, *A maximum principle for optimal control problems with neutral functional differential systems*, Bull. Amer. Math. Soc., 77 (1971), pp. 565–570.

[11] V. P. MILYUTKIN, *Maximum principle for problems with delays, fixed durations, and free trajectory right ends*, Automat. Remote Control, 1968, pp. 882–889.

[12] L. D. SABBAGH, *Variational problems with lags*, J. Optimization Theory Appl., 3 (1969), pp. 34–51.

[13] J. WARGA, *The reduction of certain control problems to an "ordinary differential" type*, SIAM Rev., 10 (1968), pp. 219–222.

[14] ———, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. CONTROL
Vol. 12, No. 2, May 1974

ON THE PROBLEM OF EVADING THE ENCOUNTER IN DIFFERENTIAL GAMES*

E. F. MISHCHENKO

Abstract. In this paper is given a sufficient condition for evading the encounter in a differential game with nonlinear controls.

Introduction. At present, several works are well known which advanced the solution of the problem on evading the encounter in differential games quite far (see, e.g., [1], [2] and [3]). However, the result of [1] remains simplified and, at the same time, made sufficiently broad for applications. In this paper, a generalization of this result is given to the case of games with nonlinear controls. This generalization has been obtained by me together with N. Satimov.

1. Statement of the problem and formulation of the result. We shall consider the differential game given by the equation

$$\dot{z} = Cz + f(u, v), \tag{1}$$

where z is a vector of an n-dimensional Euclidean space R^n, C is a constant square matrix, u and v are the control parameters with $u \in P$ and $v \in Q$, P and Q are given nonempty compact subsets of p- and q-dimensional Euclidean spaces R^p and R^q, respectively, and $f(u, v)$ is an arbitrary function, continuous in both of its arguments, defined on $R^p \times R^q$. We shall assume that the set of game terminations M is a linear subspace of the space R^n.

We shall say that it is possible to *avoid the encounter* in the game (1) (or that it is possible to *flee*) if the following is true: for any initial value $z_0 \in R^n$ with $z_0 \notin M$ of the vector z, and for any measurable function $u(t)$ according to which the parameter u changes, there exists a measurable change $v(t)$ of the parameter v such that the point $z(t)$, which is the solution of the equation

$$\dot{z} = Cz + f(u(t), v(t)), \qquad z_0 = z(0), \tag{2}$$

does not enter M for any value of the time t, $0 < t < \infty$. In this connection, in order to find the value $v(t)$ of the parameter v at every instant of time $t \geqq 0$, one is allowed to make use only of the values $u(s)$ and $z(s)$ of the parameter u and vector z for $s \leqq t$, and one is not allowed to utilize these values for $s > t$.

We now pass to the formulation of the result. We denote by L the orthogonal complement of the subspace M in R^n, and let $\dim L = \nu$. Further, let W be a linear, so far arbitrary, subspace of the space L. We denote by π the operation of the orthogonal projection from R^n onto W.

We shall say that the evasion condition is satisfied in the game (1) if there exist a two-dimensional subspace W of the space L and a positive integer k such that:

(a) each of the sets

$$\pi f(P, Q), \pi C f(P, Q), \pi C^2 f(P, Q), \cdots, \pi C^{k-2} f(P, Q)$$

* Received by the editors May 29, 1973. This translation into English was prepared by K. Makowski.

† Steklov Mathematical Institute, Academy of Sciences of the USSR, Moscow B-333, USSR.

consists of a single point, and

(b) the set

$$R = \bigcap_{u \in P} \pi C^{k-1} f(u, Q)$$

contains an interior (with respect to W) point.

Henceforth, we shall denote by ξ the distance from the point z to M, and by η the distance from the point z to W. We shall denote by M_1 the orthogonal complement of W in R^n, and by ξ_1 the distance from the point z to M_1.

THEOREM (The theorem on evading the encounter). *If the conditions* (a) *and* (b) *are satisfied in the game* (1), *then the evasion is possible. Moreover, by an appropriate choice of the evasion control* $v = v(t)$, *one can guarantee the following estimate of the distance* $\xi(t)$ *from the point* $z(t)$ *to* M:

$$(3)\qquad \begin{aligned} &\xi(t) \geqq \xi_1^k(0)\gamma[\eta(t)] \quad \textit{for all } t \geqq 0 \quad \textit{and} \quad 0 < \xi_1(0) < \varepsilon, \\ &\xi(t) \geqq \xi_1^k(\tau_1)\gamma[\eta(t)] \quad \textit{for all } t \geqq \tau_1 \quad \textit{and} \quad \xi_1(0) = 0, \end{aligned}$$

where $\gamma[\eta]$ *is a monotonically nonincreasing function of its argument,* ε *is a positive number which depends only on the game* (1) *and does not depend either on the initial point of the game or on its progress, and* τ_1 *is a positive number which depends only on the initial point of the game but not on its progress.*

2. The proof of the theorem on evading the encounter. In this section, ε, c, c_0, c_1, c_2, c_3 and c_4 denote positive constants which depend only on the game (1) but neither on its initial point nor on the controls $u(t)$ and $v(t)$. We shall denote by ζ the orthogonal projection of the point z onto W: $\pi z = \zeta$. We shall assume that an orthogonal coordinate system has been chosen in W and denote the coordinates of the point ζ in this system by ζ^1 and ζ^2.

ASSERTION A. *Since the set* R *contains an interior* (*with respect to* W) *point, there exist a vector* $l \in R$ *and a constant* c_1 *such that if a vector* $w_1 \in W$ *satisfies the condition*

$$(4)\qquad |w_1| < c,$$

and $u(s) \in P$ *is a measurable control, the equation*

$$(5)\qquad \pi C^{k-1} f(u(s), v) = l + w_1,$$

given on the interval $0 \leqq s \leqq \tau$, *has a measurable solution* $v(s) \in Q$, *defined on the interval* $0 \leqq s \leqq \tau$.

Proof. Let l be an arbitrary interior point of the set R. If the constant c_1 is sufficiently small, then, obviously, all the points $l + w_1$ belong to R for $|w_1| < c_1$. Therefore, for an arbitrary $u \in P$, we have

$$(6)\qquad l + w_1 \in \pi C^{k-1} f(u, Q).$$

In particular, $l + w_1 \in \pi C^{k-1} f(u(s), Q)$.

We consider the equation

$$(7)\qquad \pi C^{k-1} f(u(s), v) = l + w_1, \qquad 0 \leqq s \leqq \tau.$$

By virtue of (6), there exist one or many solutions of (7). We shall show that a measurable one can be chosen among them. For this purpose, we shall make use of the procedure which was employed in [4]. Namely, of all the values $v(s)$ which satisfy (7), we shall always choose $v(s) = (v^1(s), \cdots, v^q(s))$ whose component $v^1(s)$ is smallest; from these, we shall always choose the ones whose component $v^2(s)$ is smallest, and so on. (Smallest values do exist since, due to the continuity of the function f, the set of $v(s)$ which satisfy (7) is closed.) We denote the solution thus obtained by $v_0(s)$.

We shall prove by induction that the function $v_0(s), 0 \leqq s \leqq \tau$, is measurable. We assume that the components $v_0^1(s), v_0^2(s), \cdots, v_0^{r-1}(s)$ of the vector-valued function $v_0(s)$ are measurable on the interval $0 \leqq s \leqq \tau$ (there is nothing to assume for $r = 1$) and prove that the component $v_0^r(s)$ is also measurable on this interval. According to the well-known theorem of N. N. Luzin, there exists, for an arbitrarily small number $\delta > 0$, a closed subset F of the interval $[0, \tau]$ such that meas $F > \tau - \delta$ and the functions $u(s), v_0^1(s), \cdots, v_0^{r-1}(s)$ are continuous on F. We shall show that for any number a, the set of $s \in F$ for which $v_0^r(s) \leqq a$ is closed. If this is not true, then there exists a sequence $s_n \in F$, $n = 1, 2, \cdots$, such that

$$s_n \to \hat{s}, \qquad v_0^r(s_n) \leqq v_0^r(\hat{s}) - \delta_1, \quad \delta_1 > 0. \tag{8}$$

Since $|v_0^i(s)| \leqq$ const. for all $i = 1, 2, \cdots, q$ and s, a subsequence can be chosen from the sequence s_n on which the values of all the functions $v_0^i(s)$ converge to some limit values $\tilde{v}_0^i$, $i = 1, 2, \cdots, q$. Obviously, $(\tilde{v}_0^1, \cdots, \tilde{v}_0^q) \in Q$. It follows from the continuity of the functions $v_0^i(s)$, $i = 1, 2, \cdots, r - 1$, on the set F and from formula (8) that

$$\begin{aligned} &\tilde{v}_0^i = v_0^i(\hat{s}), \qquad i = 1, 2, \cdots, r - 1, \\ &\tilde{v}_0^r \leqq v_0^r(\hat{s}) - \delta_1. \end{aligned} \tag{9}$$

Passing to the limit in the identity

$$\pi C^{k-1} f(u(s), v_0^1(s), \cdots, v_0^q(s)) = l + w_1$$

along the subsequence chosen, and making use of the continuity of the function f, we obtain

$$\pi C^{k-1} f(u(\hat{s}), v_0^1(\hat{s}), \cdots, v_0^{r-1}(\hat{s}), \tilde{v}_0^r, \cdots, \tilde{v}_0^q) = l + w_1. \tag{10}$$

By virtue of (9) and (10), $v_0^r(\hat{s})$ is not a smallest value of v_0^r which satisfies the equation

$$\pi C^{k-1} f(u(\hat{s}), v_0^1(\hat{s}), \cdots, v_0^{r-1}(\hat{s}), v_0^r, \cdots, v_0^q) = l + w_1,$$

which contradicts the definition of the function $v_0^r(s)$. Thus, the function $v_0^r(s)$ is measurable on F and, since meas $F > \tau - \delta$, where δ is arbitrarily small, $v_0^r(s)$ is measurable for $0 \leqq s \leqq \tau$.

Substituting the control $v_0(s)$ chosen into (7), multiplying the identity thus obtained by $(t - s)^{k-1}/(k - 1)!$, $t \leqq \tau$, and carrying out the integration, we obtain

$$\frac{1}{(k-1)!} \int_0^t (t - s)^{k-1} \pi C^{k-1} f(u(s), v_0(s))\, ds = w t^k + l(t), \tag{11}$$

where

$$w = w^1/k!, \qquad l(t) = t^k l/k!.$$

Thus, in the game (1), there exists a constant c such that if any vector $w \in L$ satisfies the constraint $|w| < c$, and $u(s)$ is any measurable control given on the interval $0 \leqq s \leqq \tau$, then there always exists a special control $v_0(s)$ for which the relation (11) holds. In this connection, the vector-valued function $l(t)$ does not depend on $u(s)$ and $v_0(s)$.

ASSERTION B. *Let the game* (1) *start at* $t = 0$ *from the point* $z_0, |\pi z_0| = \xi_{1,0} < 1$, *and let it develop, for a given vector* w *which satisfies the constraint* $|w| < c$, *under the influence of the control* $v_0(s)$ *chosen in accordance with Assertion* A. *Then the projection of the game trajectory onto* W *can be written down for* $t \leqq \tau$ *in the form*

$$\zeta = \zeta(t) = \zeta_0 + a_1 t + \cdots + a_k t^k + w t^k + h(t) t^k, \tag{12}$$

where $a_1, a_2, \cdots, a_k$ *are constant vectors which depend on the point* z_0 *but not on the controls* $u(s)$ *and* $v_0(s)$, *and* $h(t)$ *is a vector-valued function which, although it depends on the choice of the controls, satisfies the constraint*

$$|h(t)| < c_2(1 + \eta(0))t, \tag{13}$$

where c_2 *is a constant. Moreover, the constant* c_2 *can be chosen such that the following inequalities hold*:

$$\begin{aligned} &|a_i| < c_2(1 + \eta(0)), \qquad i = 1, 2, \cdots, k, \\ &|\eta(t) - \eta(0)| < c_2(1 + \eta(0))t. \end{aligned} \tag{14}$$

Proof. According to the Cauchy formula, we have

$$\zeta(t) = \pi\, e^{tC} z_0 + \pi \int_0^t e^{(t-s)C} f(u(s), v_0(s))\, ds \tag{15}$$

Expanding the first term into series in powers of t, the integrand into series in powers of $t - s$, and taking into account the condition (a) of § 1 and relation (11), we obtain formula (12). The estimates (13) and (14) are obvious.

Assertions A and B reduce the problem of evading the encounter on the time interval $0 \leqq t \leqq \tau$ to the problem of choosing a vector $w = w(z_0) = w_0$ constant on this interval and such that inequality (3) holds for the curve (12).

The following three assertions C, D and E show that such a choice of a vector w is possible if the length of the interval $[0, \tau]$ is reasonably limited. In order to prove this, we choose in the space W a coordinate system such that the point ζ_0 has the coordinates $(\zeta_0, 0)$, and let, in this coordinate system, $a_i = (a_i^1, a_i^2)$, $w = (w^1, w^2)$ and $h(t) = (h^1(t), h^2(t))$. Then the equation of the curve (12) can be rewritten in the form

$$\begin{aligned} \zeta^1 &= \zeta^1(t) = \xi_0 + a_1^1 t + \cdots + a_{k-1}^1 t^{k-1} + (a_k^1 + w^1 + h^1(t))t^k, \\ \zeta^2 &= \zeta^2(t) = a_1^2 t + \cdots + a_{k-1}^2 t^{k-1} + (a_k^2 + w^2 + h^2(t))t^k. \end{aligned} \tag{16}$$

ASSERTION C. *There exist constants* $\varepsilon \leqq 1$, c_0 *and* w_0^1 *such that, for*

$$\tau_0 = c_0[\sqrt[k]{\varepsilon}/(1 + \eta(0))] \tag{17}$$

and $w^1 = w_0^1$, *the following inequality holds for the first component of the curve* (16):

$$|\zeta^1(\tau_0)| \geqq \varepsilon/[1 + \eta(\tau_0)]^k. \tag{18}$$

Proof. We choose

$$c_0 = \sqrt[k]{6/c}, \qquad \varepsilon \leqq 1/(c_0\sqrt[k]{c_0c_2}). \tag{19}$$

Further, we set

$$w_0^1 = \begin{cases} \frac{1}{2}c & \text{if } (\xi_0 + a_1^1\tau_0 + \cdots + a_k^1\tau_0^k) \geqq 0, \\ -\frac{1}{2}c & \text{if } (\xi_0 + a_1^1\tau_0 + \cdots + a_k^1\tau_0^k) < 0. \end{cases} \tag{20}$$

It can be verified directly that the following inequality holds for $w^1 = w_0^1$:

$$|\zeta^1(\tau_0)| \geqq 2\varepsilon/[1 + \eta(0)]^k. \tag{21}$$

Therefore in order to prove Assertion C, it is sufficient to prove that decreasing, if necessary, the number ε, we obtain the inequality

$$2\varepsilon/(1 + \eta(0))^k \geqq \varepsilon/(1 + \eta(\tau_0))^k, \tag{22}$$

or, equivalently, the inequality

$$\begin{aligned} &\frac{\sqrt[k]{2}}{1 + \eta(0)} - \frac{1}{1 + \eta(\tau_0)} \\ &\quad = \frac{1}{1 + \eta(0)} \cdot \left[\sqrt[k]{2} - \frac{1}{1 + \dfrac{\eta(\tau_0) - \eta(0)}{1 + \eta(0)}} \right] \geqq 0. \end{aligned} \tag{23}$$

By virtue of (14), we have

$$\frac{\eta(\tau_0) - \eta(0)}{1 + \eta(0)} \leqq c_2c_0\sqrt[k]{\varepsilon}. \tag{24}$$

Therefore (23) holds for a sufficiently small ε, and Assertion C has been proved.

We first fix the value of ε chosen and then, by the end of the proof of Assertion D, decrease it again if needed.

Remark. We choose w_0^1 in an entirely determined way: $w_0^1 = \frac{1}{2}c$ or $w_0^1 = -\frac{1}{2}c$ (see (20)). However, it is not difficult to see that any w_0^1 chosen, e.g., on the interval $[\frac{1}{2}c, \frac{3}{4}c]$ or, correspondingly, on $[-\frac{3}{4}c, -\frac{1}{2}c]$, is suitable.

The following assertion is most essential in the proof of the theorem on evading the encounter.

ASSERTION D. *Let the initial point* ζ_0 *of the curve* (16) *satisfy the condition*

$$0 \neq \xi_{1,0} = |\zeta_0| = |\xi_0| \leqq \varepsilon.$$

Then there exist a vector $w = w_0 = (w_0^1, w_0^2)$ *and a monotonically nonincreasing function* γ *such that the following inequality holds on the interval* $0 \leqq t \leqq \tau_0$:

$$|\zeta(t)| \geqq \xi_{1,0}^k\gamma[\eta(t)]. \tag{25}$$

In this connection, the first component w_0^1 *of the vector* w_0 *can be chosen in such a way that inequality* (18) *holds simultaneously.*

Proof. We denote the polar coordinates of the point $\zeta(t)$ of the curve (16) by $\rho(t)$, $\varphi(t)$. Further, we set

$$\begin{aligned} \alpha &= a_k^1 + w^1 + h^1(t), \\ \beta &= a_k^2 + w^2 + h^2(t). \end{aligned} \tag{26}$$

Then we obtain

$$\begin{aligned} \rho(t)\cos\varphi(t) &= \xi_0 + a_1^1 t + \cdots + a_{k-1}^1 t^{k-1} + \alpha t^k, \\ \rho(t)\sin\varphi(t) &= a_1^2 t + \cdots + a_{k-1}^2 t^{k-1} + \beta t^k. \end{aligned} \tag{27}$$

Multiplying the relations (27) successively by $1, t, t^2, \cdots, t^{k-1}$, we obtain

$$\begin{aligned} \rho(t)\cos\varphi(t) &= \xi_0 + a_1^1 t + \cdots + a_{k-1}^1 t^{k-1} + \alpha t^k \cdots, \\ t\rho(t)\cos\varphi(t) &= 0 + \xi_0 t + \cdots + a_{k-2}^1 t^{k-1} + a_{k-1}^1 t^k + \alpha t^{k+1} \cdots, \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \rho(t)\sin\varphi(t) &= 0 + a_1^2 t + \cdots + a_{k-1}^2 t^{k-1} + \beta t^k \cdots, \\ t\rho(t)\sin\varphi(t) &= 0 + a_1^2 t^2 + \cdots + a_{k-1}^2 t^k + \beta t^{k+1} \cdots, \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \end{aligned} \tag{28}$$

We shall consider these relations as a system of $2k$ linear algebraic equations in the unknowns $1, t, t^2, \cdots, t^{2k-1}$. Solving it formally with respect to the unknown 1, we find

$$1 = D_1/D, \tag{29}$$

where D is the determinant of the system (28), and D_1 is the determinant obtained from D by replacing the first column by the column of the free terms of system (28). We take out the common factor $\rho(t)$ from the first column and set

$$D_1 = \rho(t)\tilde{D}. \tag{30}$$

Then we obtain

$$\rho(t)\tilde{D} = D. \tag{31}$$

Therefore, in order to obtain an estimate for $\rho(t)$, one must estimate the determinants $\tilde{D}$ and D.

The determinant D is a function of the parameters α and β. We denote this dependence by the subscripts α, β: $D = D_{\alpha,\beta}$. Moreover, this determinant depends on $\xi_0, a_1^1, \cdots, a_{k-1}^1, a_1^2, \cdots, a_{k-1}^2$, which in turn depend on z_0. We shall not explicitly denote this dependence. We shall prove that given any fixed rectangle Π determined by the inequalities

$$\alpha_1 \leqq \alpha \leqq \alpha_2, \qquad \beta_1 \leqq \beta \leqq \beta_2 \tag{32}$$

there always exist a point $(\alpha_0, \beta_0) \in \Pi$ and positive numbers r and δ, which depend only on the size of the rectangle Π and on the number k but not on $\xi_0, a_1^1, \cdots, a_{k-1}^1, \cdots, a_{k-1}^2$, such that

$$|D_{\alpha_0+\Delta\alpha,\beta_0+\Delta\beta}| \geqq r\xi_{10}^{k}, \tag{33}$$

if only

$$|\Delta\alpha| < \delta, \qquad |\Delta\beta| < \delta. \tag{34}$$

Indeed, one can see directly that $D_{\alpha,\beta}$ is a polynomial in parameters α and β of the following form:

$$D_{\alpha,\beta} = \xi_0^k \beta^k + \sum_{i,j=0}^{k-1} p_{ij}\alpha^i\beta^j. \tag{35}$$

Let p be the coefficient of the polynomial (35) with the largest absolute value. Then

$$D_{\alpha,\beta} = p\, d_{\alpha\beta}, \qquad |p| \geqq \xi_{10}^k, \tag{36}$$

where $d_{\alpha\beta}$ belongs to the set $\{d\}$ of polynomials of bounded degree, in each of which there is at least one coefficient equal to one and all the remaining coefficients are not larger than one. This set is compact. Hence it follows easily that there exists a positive constant which bounds from below the maximum values of the modules of all polynomials of the set. Since the derivatives with respect to α and β of a polynomial belonging to the set $\{d\}$ are uniformly bounded from above, the existence of numbers r and δ, which provide inequality (33), has been proved.

We shall now choose an entirely defined rectangle Π. Namely, we shall choose numbers α_1 and α_2 such that, if the inequality

$$\alpha_1 \leqq \alpha_k^1 + w^1 \leqq \alpha_2 \tag{37}$$

holds, then the number w^1 guarantees that (18) holds (see Assumption C). Then we choose numbers β_1 and β_2 such that, if the following inequality holds:

$$\beta_1 \leqq a_k^2 + w^2 \leqq \beta_2, \tag{38}$$

then the inequality

$$(w^1)^2 + (w^2)^2 \leqq c \tag{39}$$

holds. Let (α_0, β_0) be a maximum point of polynomial (35) on this rectangle.

We set

$$w_0^1 = \alpha_0 - a_k^1, \qquad w_0^2 = \beta_0 - a_k^2. \tag{40}$$

Then, obviously, we have

$$D = D_{\alpha,\beta} = D_{\alpha_0 + h^1(t), \beta_0 + h^2(t)}. \tag{41}$$

Finally, we choose ε so small that the following inequalities hold on the entire interval $0 \leqq t \leqq \tau_0$ (see Assumption C):

$$|h^1|(t)| < \delta, \qquad |h^2(t)| < \delta. \tag{42}$$

Then the following estimate from below for the determinant D follows at once from (33):

$$|D| \geqq r\xi_{1,0}^k. \tag{43}$$

We shall now estimate from above the module of the determinant $\tilde{D}$. It follows from inequality (14) that there exists a constant c_3 such that the following inequality holds for $0 \leqq t \leqq \tau_0$:

$$|\tilde{D}| \leqq c_3(1 + \eta(0))^{2k-1}. \tag{44}$$

Comparing the inequalities (43) and (44) and setting

$$\gamma_1[\eta_0] = \frac{r}{c_3}\cdot\frac{1}{[1+\eta(0)]^{2k-1}}, \tag{45}$$

we obtain from (31)

$$\rho(t) \geqq \gamma_1[\eta(0)]\xi_{10}^k \tag{46}$$

for any t in the interval $[0,\tau_0]$. Obviously, $\gamma_1[\eta]$ is a monotonically decreasing function of its argument, which depends only on the game.

We now define the function $\gamma_0[\eta]$ by the following formula:

$$\gamma_0[\eta] = \gamma_1[\eta + c_0c_2]. \tag{47}$$

Obviously, we have then for $0 \leqq t \leqq \tau_0$,

$$\gamma_0[\eta(t)] = \gamma_1[\eta(t) + c_2c_0] = \gamma_1[\eta(0) + \eta(t) - \eta(0) + c_2c_0] \leqq \gamma_1[\eta(0)], \tag{48}$$

since $|\eta(t) - \eta(0)| \leqq c_2c_0$. Hence (25) follows from this and from (46). Assertion D has been proved.

ASSERTION E. *Let $z_0 \in M_1 \setminus M$. There exists a positive constant τ_1 such that, for all $t \in [0,\tau_1]$ and for arbitrary controls $u(s)$ and $v(s)$ given on the interval $0 \leqq s \leqq \tau_1$, the point $z(t)$—which is the solution of* (2)*—does not enter M. Further, there exists a constant w_1^1 such that, for $w^1 = w_1^1$, the following inequality holds for the first component of the curve* (12):

$$|\zeta^1(\tau_1)| \geqq (c/4)\tau_1^k. \tag{49}$$

Proof. The first part of this assertion is an obvious consequence of the compactness of the sets P and Q, and of the continuity of the function $f(u,v)$.

One can always assume that

$$\tau_1 \leqq c/(4c_2[1+\eta(0)]).$$

Further, we set

$$w_1^1 = \begin{cases} \dfrac{c}{2} & \text{if } \xi_0 + a_1^1\tau_1 + \cdots + a_k^1\tau_1^k \geqq 0, \\[2ex] -\dfrac{c}{2} & \text{if } \xi_0 + a_1^1\tau_1 + \cdots + a_k^1\tau_1^k < 0. \end{cases} \tag{50}$$

Then (see (13))

$$|\zeta^1(\tau_1)| \geqq [\tfrac{1}{2}c - |h(\tau_1)|]\tau_1^k \geqq [\tfrac{1}{2}c - c_2(1+\eta(0))\tau_1]\tau_1^k \geqq \frac{c}{4}\tau_1^k.$$

Assertion E has been proved. We note that, since τ_1 can be decreased as needed, one can assume that $|\pi z(\tau_1)| < \varepsilon$ (because $|\pi z(0)| = 0$).

2.1. Proof of the theorem on evading the encounter. Let the evasion conditions of § 1 be satisfied for the game (1). We choose a number ε in accordance with the assertions C and D, and denote by S the $(n-1)$-dimensional surface in the space R^n determined by the equation

$$\xi_1 = \varepsilon/(1+\eta)^k. \tag{51}$$

This surface subdivides the space R^n into two nonintersecting domains: the interior one S_-, which contains the subspace M_1, and the exterior one S_+. There are only three possibilities for the initial point z_0 of the game (1):

(i) $z_0 \in S_+$,
(ii) $z_0 \notin M_1, z_0 \in S_-$,
(iii) $z_0 \in M_1 \setminus M, z_0 \in S_-$.

If the initial point of the game belongs to the domain S_+, then a control $v(t)$ can be chosen arbitrarily during some period of time. Let t_0 be the first instant of time at which the point $z(t)$ hits the surface S. We take this time t_0 as the new beginning of time count, and the point $z(t_0)$ as the initial point of the game. Obviously, $\xi_{1,0} < \varepsilon$. Therefore, according to the rule in the discussions of Assertions B, C and D, the vector $w_0 = (w_0^1, w_0^2)$ can be assigned to this point, and then, according to Assertion A, also the special control $v_0(t)$ determined on the interval $0 \leqq t \leqq \tau_0$. This control will provide the following inequality for the trajectory of the game on this interval:

$$|\pi z(t)| = \xi_1(t) \geqq \varepsilon^k \frac{\gamma[\eta(t)]}{(1 + \eta(t))^{k^2}}. \tag{52}$$

For $t = \tau_0$, the point $z(\tau_0)$ will again appear in the domain S_+. After this, the process can be repeated.

Now, let the initial point z_0 of the game belong to the domain S_-, but let $z_0 \notin M_1$. Then the control $v_0(t)$ can be activated at once, which will provide, by virtue of Assertion D, the following inequality for $0 \leqq t \leqq \tau_0$:

$$|\pi z(t)| \geqq \xi_{10}^k \gamma[\eta(t)], \tag{53}$$

and, for $t = \tau_0$, the point $z(\tau_0)$ will again be in the domain S_+.

Now, let the initial point z_0 satisfy the condition $z_0 \in M_1 \setminus M$, $z_0 \in S_-$. In this case, we choose the constants τ_1 and w_1^1 (w_1^2 can be chosen in an arbitrary way). Then the point $z(t)$ will not enter M up to and including the time τ_1 (see (49)). By virtue of the remark made at the end of the proof of Assertion E, we have $|\pi z(\tau_1)| < \varepsilon$. We take the point $z(\tau_1)$ as the initial one. Obviously, $z(\tau_1) \notin M$, $z(\tau_1) \in S_-$. Therefore, the following inequality holds for $t \geqq \tau_1$:

$$\xi_1(t) \geqq \xi_1^k(\tau_1)\gamma[\eta(t)]. \tag{54}$$

The theorem has been proved.

3. Examples.

3.1. A control example. The motion laws of the pursuing and evading objects are given by the respective equations

$$\ddot{x} + \alpha\dot{x} = \rho u, \tag{55}$$

$$\ddot{y} + \beta\dot{y} = \sigma v. \tag{56}$$

Here, x, y, u and v are vectors of a Euclidean space E of dimension $\gamma \geqq 2$, x is the geometric position of the pursuing point, y is the geometric position of the evading point, u and v are control parameters with $|u| \leqq 1$, $|v| \leqq 1$, and α, β, σ

and ρ are positive numbers, where

$$\sigma > \rho. \tag{57}$$

The pursuit is assumed to be terminated at the time when $x = y$. The problem is posed whether the object (56) can evade the object (55).

We pass over to the corresponding differential game. To this end, we set

$$z = (z_1, z_2, z_3) = (x - y, \dot{x}, \dot{y}). \tag{58}$$

Thus, z is a vector of the three-dimensional Euclidean space $R = E \times E \times E$, and the game is given by the following system of differential equations:

$$\dot{z}_1 = z_2 - z_3, \quad \dot{z}_2 = -\alpha z_2 + \rho u, \quad \dot{z}_3 = -\beta z_3 + \sigma v. \tag{59}$$

The set P consists of all vectors $(0, -\rho u, 0)$, the set Q consists of all vectors $(0, 0, \sigma v)$, and the set M of game termination is given by the equation $z_1 = 0$. Any two-dimensional subspace of the space L can be taken as W. It can be verified directly that $\pi P = \pi Q = 0$. The set $\pi C Q$ is a two-dimensional ball of radius σ, the set $\pi C P$ is a two-dimensional ball of radius ρ. Therefore, the evasion conditions are satisfied for $\sigma > \rho$ with $k = 2$.

In view of the presence of friction $\alpha > 0$ and $\beta > 0$, the velocities $\dot{x}(t)$ and $\dot{y}(t)$ of the motions of objects (55) and (56) which start from the rest state ($\dot{x} = 0$ and $\dot{y} = 0$) are bounded:

$$|\dot{x}(t)| \leqq \rho/\alpha, \qquad |\dot{y}(t)| \leqq \sigma/\beta,$$

and the projection of the point $z(t)$ onto M remains within a compact subset of M for any t. Therefore, inequality (3) has a particularly simple form for the game (58), namely,

$$|\pi z(t)| \geqq c_4 |\pi z_0|^2, \tag{60}$$

where c_4 is a constant which depends only on the game.

3.2. The problem of a "boy" and an "alligator".[1] The motion laws of the pursuing and evading objects are given by the equations

$$\ddot{x} = u, \qquad \dot{y} = v, \tag{61}$$

where x, y, u and v are vectors of a Euclidean vector space E, $\nu \geqq 2$, u and v are the control parameters with $|u| \leqq 1$ and $|v| \leqq 1$, x is the geometric position of the pursuing object—the "alligator"— and y is the geometric position of the evading object—the "boy". The pursuit is assumed to be terminated when $x = y$.

Setting $z = (z_1, z_2, z_3) = (x, \dot{x}, y)$, we obtain the differential game with the matrix

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

[1] *Translator's note.* I translated the Russian "crocodile" as "alligator", because this problem was named after Prof. L. W. Neustadt's famous pet alligator, Cookie. [K.M.]

the termination set $M = \{z_1 = 0\}$ and the sets $P = \{(0, u, 0)\}$ and $Q = \{(0, 0, v)\}$. It can be verified directly that the assumptions of the theorem on evasion are satisfied also in this game, with $k = 1$, and that, therefore, the inequality

$$|\pi z(t)| \geqq |\pi z_0|\gamma[\eta(t)] \tag{62}$$

holds for any t.

REFERENCES

[1] L. S. Pontryagin and E. F. Mishchenko, *The problem of evading the encounter in linear differential games*, Differencial'nye Uravnenija, 7 (1971), pp. 436–445.

[2] L. S. Pontryagin, *On the linear differential game of evading*. Trudy MIAN, 112 (1971), Nauka.

[3] R. V. Gamkrelidze and G. L. Kharatshvili, *A differential game of evasion with nonlinear control*, this Journal, 12 (1974), pp. 332–349.

[4] A. F. Filippov, *On certain questions in the theory of optimal control*, Vestnik Moskov. Univ. Ser. I Mat. Meh., 25 (1959), no. 2; English transl., this Journal, 1 (1962), pp. 76–84.

SIAM J. Control
Vol. 12, No. 2, May 1974

THE CHARACTERIZATION OF THE WEAK* CLOSURE OF CERTAIN SETS OF INTEGRABLE FUNCTIONS[†]

CZESŁAW OLECH[‡]

Abstract. Let K be a set of integrable functions from a Hausdorff metrizable topological compact space T with a given regular Borel measure on it into an Euclidean space E. We assume that K has the following property (P): for each measurable subset A of T and any two functions u_1, u_2 from K, the function

$$\chi_A u_1 + \chi_{T\setminus A} u_2 \in K,$$

where χ_A is the characteristic function of the set A.

In this paper we give the characterization of the closure $\bar{K}$ of set K having property (P) in the weak topology of the conjugate space C^* of the space C of continuous functions from T into E. In particular, we also obtain conditions under which the set K is closed in the weak topology $w(L_1, C)$.

Introduction. The motivation to study the weak* closure of sets of L_1 having property (P) comes from the theory of optimal control, in particular from the existence theory of optimal solutions to some optimal control problems. In this theory one needs to know whenever a set K, which is the set of integrable selections of a given set-valued function, is closed in a weak topology. We refer the reader to the author's papers [4], [5] and [6] for examples. The result we present here is in fact a generalization of the lemma from [4] and [5] which the author has used to obtain some existence theorems for optimal solutions. This paper can also be considered to be a continuation of the work of A. Lasota and the author [1], [2], where the problem of closedness of the set of solutions to differential equations with multivalued right-hand sides were considered.

In a recent paper, R. T. Rockafellar [7] has given a characterization of the closure in question in a special case. This he obtains by deriving a formula for conjugate functionals to a given integral functional on $C(T, E)$, hence by a rather different method.

Finally, the result given here was applied by the author to obtain a necessary and sufficient condition for lower semicontinuity of integral functional on L_1 of the form

$$I(u) = \int_T f(t, u(t))\, dt,$$

again with respect to the weak topology $w(L_1, C)$ on L_1 (cf. [6]).

We will not discuss those applications in more detail. We hope that the problem itself may be of some interest to the reader.

1. Notations and assumptions. By T we denote a metrizable compact Hausdorff topological space. Let us fix a regular nonnegative Borel measure μ on T. We shall refer to this measure as dt and assume that it is complete nonatomic and

[†] Received by the editors October 18, 1973, and in final revised form January 29, 1974.

[‡] Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland.

that the support of it is the whole space T; that is, any open subset of T has positive measure. The latter assumption does not restrict the generality of our considerations. All terms like *measurability, integrability* and *absolute continuity* will always be meant with respect to this fixed measure unless an opposite is specifically stated.

E will stand for an Euclidean n-space and $\langle\cdot,\cdot\rangle$ will denote the scalar product in E. Let $L_1(T, E)$ and $C(T, E)$, or simply L_1 and C, be the spaces of integrable and continuous functions from T into E, respectively, with the usual norm. By C^*, as is customary, we denote the space of linear continuous functionals on C, that is, the space of E-valued regular Borel measures on T. Frequently we shall identify L_1 with a subspace of C^*, identifying an integrable function $u \in L_1$ with the functional $u(\varphi) = \int_T \langle u(t), \varphi(t)\rangle\, dt$ on C or with a regular absolutely continuous measure given by $m(A) = \int_A u(t)\, dt$ for each measurable $A \subset T$.

Let $K \subset L_1$ be nonempty and have the property

(P) $\quad \chi_A u_1 + \chi_{T\setminus A} u_2 \in K \quad$ if $u_1, u_2 \in K$ and $A \subset T$ is measurable.

For each $\varphi \in C$ we put

$$\psi_\varphi = \operatorname*{ess\,sup}_{u\in K} \langle u, \varphi\rangle. \tag{1}$$

That is, for each $u \in K$ we have the inequality

$$\langle u(t), \varphi(t)\rangle \leqq \psi_\varphi(t) \quad \text{a.e. in } T$$

and vice versa: if for some measurable $v: T \to R \cup \{+\infty\}$ we have the inequality $\langle u(t), \varphi(t)\rangle \leqq v(t)$ a.e. in T, for each $u \in K$, then $\psi_\varphi(t) \leqq v(t)$ a.e. in T.

The function ψ_φ can assume $+\infty$ values, is defined uniquely up to a set of measure zero, and is measurable in t.

For each fixed t, we put

$$D(t) = \{\varphi(t) | \varphi \in C \text{ and } \psi_\varphi \text{ is integrable on a neighborhood of } t\}. \tag{2}$$

Proposition 1. *$D(t)$ is a convex cone.*

Proof. If ψ_φ is locally integrable at t, then so is $\psi_{\lambda\varphi}$ for each $\lambda > 0$, hence $D(t)$ is a cone. If ψ_{φ_1} and ψ_{φ_2} are locally integrable at $t = t_0$, then by (1), for any λ between 0 and 1 we have the inequality

$$\psi_{\lambda\varphi_1 + (1-\lambda)\varphi_2}(t) \leqq \lambda\psi_{\varphi_1}(t) + (1-\lambda)\psi_{\varphi_2}(t),$$

which implies that $\psi_{\lambda\varphi_1+(1-\lambda)\varphi_2}$ is also integrable on a neighborhood of t_0. Therefore $\lambda\varphi_1(t_0) + (1-\lambda)\varphi_2(t_0)$ belongs to $D(t_0)$ also, which shows that $D(t)$ is convex for each t.

Let us recall that a set-valued function $D(t)$ from T into subsets of E is *lower semicontinuous* if for each open set $G \subset E$, the set

$$D^- G = \{t | D(t) \cap G \neq \varnothing\}$$

is an open subset of T. Let us recall at this point also that if $D^- G$ is measurable for each compact G, then D is called *measurable in t* and if it is closed for each compact G, then D is called *upper semicontinuous*.

Proposition 2. *$D(t)$ given by* (2) *is lower semicontinuous.*

Proof. Let $G \subset E$ be open and suppose $D(t_0) \cap G \neq \varnothing$; then there is a $\varphi \in C$ such that ψ_φ is integrable on a neighborhood N of t_0 and $\varphi(t_0) \in D(t_0) \cap G$.

But $\varphi(t) \in D(t)$ if $t \in N$, and since φ is continuous there is a neighborhood N_1 of t_0 such that $\varphi(t) \in G$ if $t \in N_1$. Hence we conclude that $N \cap N_1 \subset \{t | D(t) \cap G \neq \emptyset\}$, which proves the proposition.

2. Preliminary results. Denote by K the weak* closure of K in the weak* topology of the conjugate space C^* of the space C. That is, $\bar{K} \subset C^*$ is a set of regular E-valued Borel measures and a measure $m \in \bar{K}$ if and only if there is a generalized sequence $\{u_\alpha\} \subset K$ such that for each $\varphi \in C$ we have

$$\lim_\alpha \int_T \langle \varphi(t), u_\alpha(t) \rangle \, dt = \int_T \langle \varphi(t), dm \rangle . \tag{3}$$

Because of property (P) assumed for K, the closure $\bar{K}$ is a convex subset of C^*. This is a well-known fact, but for completeness of exposition we shall prove it now. Namely, we will prove that

$$\bar{K} = \bigcap_{\varphi \in C} \left\{ m \in C^* \,\middle|\, \int_T \langle \varphi(t), dm \rangle \leqq \int_T \psi_\varphi(t) \, dt \right\} . \tag{4}$$

Proof of (4). Denote the right-hand side of (4) by K_1. If m is a weak* limit of a net $\{u_\alpha\} \subset K$, then the inequality $\int \langle \varphi(t), dm \rangle \leqq \int \psi_\varphi(t) \, dt$ follows from (1) and (3). Thus $\bar{K} \subset K_1$. Suppose now that m belongs to K_1, And let $\varphi_1, \cdots, \varphi_k$ be any fixed finite sequence contained in C. Set

$$B = \left\{ \left[\int_T \langle \varphi_i(t), u(t) \rangle \, dt \right] \,\middle|\, u \in K \right\} \subset R^k.$$

Because of property (P), the set B is convex. This follows from a generalization of Lyapunov theorem; cf., for example, [3]. We shall prove that $\{\int_T \langle \varphi_i, dm \rangle\}$ belong to the closure $\bar{B}$ of B. Indeed,

$$\bar{B} = \bigcap_{a \in R^k} \left\{ x \in R^k \,\middle|\, \sum_{i=1}^k a_i x_i \leqq \sup_{b \in B} \sum_{i=1}^k a_i b_i \right\} .$$

But $\sup_{b \in B} \sum a_i b_i = \sup_{u \in K} \int_T \langle \sum_{i=1}^k a_i \varphi_i(t), u(t) \rangle \, dt = \int_T \psi_{\Sigma a_i \varphi_i}(t) \, dt$.

The latter equality follows from property (P). Hence for each $a \in R^k$, $\sum_{i=1}^k a_i \int \langle \varphi_i, dm \rangle \leqq \int_T \psi_{\Sigma a_i} \varphi_i(t) \, dt$, which proves that $\{\int_T \langle \varphi_i, dm \rangle\}$ belongs to $\bar{B}$. Thus for any $\varepsilon > 0$ there is a $u \in K$ such that

$$\left| \int_T \langle \varphi_i(t), u(t) \, dt - dm \rangle \right| < \varepsilon \quad \text{for each } i .$$

Therefore, we have proved that any neighborhood of $m \in K_1$ in the w^* topology of C^* contains a point from K, which proves that $K_1 \subset K$ and completes the proof of (4).

There exists a denumerable set $\{\varphi_n\} \subset C$ such that

$$\bar{K} = \bigcap_{n=1}^{\infty} \left\{ m \,\middle|\, \int_T \langle \varphi_n(t), dm \rangle \leqq \int_T \psi_{\varphi_n}(t) \, dt \right\} . \tag{5}$$

Indeed, let $B(N)$ denote the closed ball in C^* of radius N. The topology $w(C^*, C)$ restricted to $B(N)$ is metrizable. Therefore from (4) and the Lindelöf theorem, there exists a denumerable subset $\Phi_N \subset C$ such that

$$B(N) \cap \bar{K} = B(N) \cap \left(\bigcap_{\varphi \in \Phi_N} \left\{ m \,\middle|\, \int_T \langle \varphi, dm \rangle \leqq \int_T \psi_\varphi(t)\, dt \right\} \right)$$

Putting $\{\varphi_n\} = \bigcup_{N=1}^{\infty} \Phi_N$, we find that the latter implies (5).

We may, of course, assume that each ψ_{φ_n} in (5) is integrable on T. For the sequence $\{\varphi_n\}$ in (5), put

$$Q(t) = \bigcap_{n=1}^{\infty} \{x \in E \mid \langle \varphi_n(t), x \rangle \leqq \psi_{\varphi_n}(t)\} \tag{6}$$

and

$$C(t) = \{x \in E \mid \langle \varphi_n(t), x \rangle \leqq 0, n = 1, 2, \cdots\}. \tag{7}$$

PROPOSITION 3. *The set-valued function Q given by* (6) *is measurable in t, and the asymptotic cone of $Q(t)$ is equal a.e. in T to $C(t)$ given by* (7).

Proof. Let $G \subset E$ be compact. By (6),

$$Q^- G = \bigcap_{n=1}^{\infty} \{t \mid \min_{x \in G} \langle \varphi_n(t), x \rangle \leqq \psi_{\varphi_n}(t)\},$$

and therefore it is a measurable set, which proves that Q is measurable in t.

The asymptotic cone $C_{Q(t)}$ of a closed and convex set $Q(t)$ is the largest cone with the property that if $q \in Q(t)$, then $q + C_{Q(t)} \subset Q(t)$. By (6) and (7) it is clear that $C(t) = C_{Q(t)}$ if $\psi_{\varphi_n}(t)$ is finite for each n; thus the latter equality holds a.e. in T, which completes the proof.

PROPOSITION 4. *The cone-valued function C given by* (7) *is upper semicontinuous.*

Proof. Let $G \subset E$ be compact. Then $C^- G = \bigcap_{n=1}^{\infty} \{t \mid \min_{x \in G} \langle \varphi_n(t), x \rangle \leqq 0\}$. Since $\min_{x \in G} \langle \varphi_n(t), x \rangle$ is a continuous function of t, thus $C^- G$ is closed, which was to be proved.

Denote by K_Q the set of integrable selections of Q; that is,

$$K_Q = \{u \in L_1 \mid u(t) \in Q(t) \text{ a.e. in } T\}. \tag{8}$$

We have by (1) and (6) the inclusion $K \subset K_Q$. We shall need the following.

PROPOSITION 5. *If Q and Q_1 are two measurable, closed and convex set-valued functions on T and $K_Q = K_{Q_1}$, then $Q(t) = Q_1(t)$ a.e. in T.*

Proof. To prove this we shall use the following characterization (due to Rockafellar [8]) of a closed, convex set-valued measurable mapping. Namely, a mapping Q is measurable if and only if there exists a denumerable sequence of measurable selections $\{u_n\}$ of Q such that $\{u_n(t)\}$ is a dense subset of $Q(t)$ for each t. Let $\{u_n\}$ and $\{v_n\}$ be such sequences for $Q(t)$ and $Q_1(t)$ respectively. Since $v_n(t) \in Q(t)$ a.e. in T for each n, thus there is a set N of measure zero such that $v_n(t) \in Q(t)$ for each n and $t \in T \setminus N$. Since $Q_1(t) = \mathrm{cl}\{v_n(t)\}$, therefore $Q_1(t) \subset Q(t)$ a.e. in T. The opposite inclusion holds for the same reasons, so the proposition is proved.

3. The main result. Following Rockafellar, we say that a measure m is *C-valued*, where C is a cone-valued function, if for any scalar measure ρ such that m is absolutely continuous with respect to ρ, $(dm/d\rho)(t) \in C(t)$ ρ-a.e. in T.

The characterization of the closure $\bar{K}$ of K is given by the following.

THEOREM. *Suppose $K \subset L_1(T, E)$ satisfies property* (P). *Then there is a measurable set valued function P from T into closed and convex subsets of E, defined uniquely up to a set of measure zero, such that the asymptotic cone $C_P(t)$ is equal to the polar $D^0(t) = \{e | \langle e, d\rangle \leqq 0$ for each $d \in D(t)\}$* of *$D(t)$ given by* (2) *and such that $m \in \bar{K}$ if and only if*

$$\frac{dm_{\mathrm{a}}}{dt}(t) \in P(t) \quad \textit{a.e. in } T \tag{9}$$

and

$$m_{\mathrm{s}} \textit{ is } D^0\textit{-valued}, \tag{10}$$

where m_{a} and m_{s} are the absolutely continuous and singular parts of m, respectively.

Proof. First we will prove the existence of P by showing that (9) and (10) hold for Q and C given by (6) and (7). Let us notice that $K \subset K_Q$. If $m \in \bar{K}$, then there is a generalized sequence $\{u_\alpha\} \subset K$ such that for any $\varphi \in C$,

$$\int_T \langle \varphi(t), u_\alpha(t)\rangle\, dt \to \int_T \langle \varphi(t), dm\rangle. \tag{11}$$

Take φ such that ψ_φ is integrable on T, and put

$$\tilde{m}_\alpha(A) = \int_A (\langle \varphi(t), u_\alpha(t)\rangle - \psi_\varphi(t))\, dt,$$

where $A \subset T$ is measurable.

By (11), $\{\tilde{m}_\alpha\}$ is weak* convergent to $\tilde{m}$, where

$$\tilde{m}(A) = \int_A \langle \varphi(t), dm\rangle - \int_A \psi_\varphi(t)\, dt$$

for each measurable $A \subset T$.

Since each $\tilde{m}_\alpha$ is nonpositive, so is $\tilde{m}$. Therefore both the absolutely continuous part $\tilde{m}_{\mathrm{a}}$ and the singular part $\tilde{m}_{\mathrm{s}}$ of $\tilde{m}$ are nonnegative. But

$$\frac{d\tilde{m}_{\mathrm{a}}}{dt}(t) = \left\langle \varphi(t), \frac{dm_{\mathrm{a}}}{dt}(t)\right\rangle - \psi_\varphi(t) \leqq 0,$$

and for any nonnegative measure ρ such that m_{s} is absolutely continuous with respect to ρ,

$$\frac{d\tilde{m}_{\mathrm{s}}}{d\rho}(t) = \left\langle \varphi(t), \frac{dm_{\mathrm{s}}}{d\rho}(t)\right\rangle \leqq 0 \quad \rho\text{-a.e. in } T.$$

The first inequality applied to each φ_n in (6) gives (9) and the second gives (10). On the other hand, if m satisfies (9) and (10), then $\int_T \langle \varphi_n(t), dm\rangle \leqq \int_T \psi_{\varphi_n}(t)\, dt$ for each n, and thus $m \in \bar{K}$ by (5).

If (9) and (10) hold for another $P \neq Q$, then the sets K_P and K_Q are the same since each has to be equal $\bar{K} \cap L_1$. Therefore by Proposition 5, $P(t) = Q(t)$ a.e. in T. Hence $P(t)$ is uniquely defined up to a set of measure zero, and since $C(t)$ is the asymptotic cone of $Q(t)$ a.e. in t, thus $C(t) = C_P(t)$ a.e. in T.

It remains to show that $C(t) = D^0(t)$, where $D^0(t) = \{a | \langle a, d\rangle \leqq 0$ for each $d \in D(t)\}$ is the polar cone of $D(t)$. Manifestly we have the inclusion $D^0(t) \in C(t)$. To prove the opposite inclusion, take an arbitrary $d \in C(t_0)$, with $t_0 \in T$ fixed, and take any $u \in K$. Let $m \in C$ such that $dm_a/dt = u$ and $m_s = d\delta_{t_0}$, where δ_{t_0} is the Dirac measure concentrated at t_0. Take an arbitrary $e \in D(t_0)$; there is a $\varphi \in C$ such that $\varphi(t_0) = e$ and ψ_φ is integrable on a neighborhood N of t_0. Since T is a normal space, there is a continuous function λ such that $0 \leqq \lambda(t) \leqq 1$, $\lambda(t_0) = 1$ and $\lambda(t) = 0$ if $t \in T \setminus N$. It is clear, then, that $\psi_{\lambda\varphi}$ is integrable on T. Since the above-defined in m is an element of $\bar{K}$, thus we have by (4) the inequality

$$\begin{aligned}\int_T \langle \lambda(t)\varphi(t), dm\rangle &= \int_N \langle \lambda(t)\varphi(t), dm\rangle \\ &= \int_N \langle \lambda(t)\varphi(t), u(t)\rangle\, dt + \langle e, d\rangle \\ &\leqq \int_T \psi_{\lambda\varphi}(t)\, dt = \int_N \psi_{\lambda\varphi}(t)\, dt.\end{aligned}$$

Since this inequality holds for arbitrary N, it implies that $\langle e, d\rangle \leqq 0$. Since e was an arbitrary point of $D(t_0)$, therefore $d \in D^0(t)$. Hence $C(t) \subset D^0(t)$ and $C(t) = D^0(t)$. This finishes the proof of the theorem.

4. Closedness conditions for *K*. From the theorem we now deduce necessary and sufficient conditions for a set $K \subset L_1$ having property (P) to be weakly closed with respect to the topology $w(L_1, C)$. The first of them is an immediate consequence of the theorem and is stated in the following.

COROLLARY 1. *A set $K \subset L_1$ satisfying property* (P) *is closed in the topology $w(L_1, C)$ if and only if K is the set of integrable selections to a measurable set-valued mapping P from T into closed convex subsets of E such that P can be represented up to a set of measure zero by countable product*

$$\bigcap_{n=1}^{\infty} \{x | \langle x, \varphi_n(t)\rangle \leqq \psi_n(t)\}, \tag{12}$$

where the φ_n are continuous and the Ψ are integrable on T.

Proof. If K is $w(L_1, C)$-closed, then by the theorem there is a $P(t)$ such that $K = K_p$ and $P(t) = Q(t)$ a.e. in T, where Q is given by (6). This proves the necessity part. On the other hand, if $K = K_p$ and P is given by (12), then $K = \bigcap K_n$, where $K_n = \{u \in L_1(T, E) : \langle u(t), \varphi_n(t)\rangle \leqq \psi_n(t)\}$, and one can check directly that K_n is $w(L_1, C)$-closed; thus so is K and we are done.

Remark 1. In [2], the sufficiency part of the above result is essentially proved, though only in the case when $P(t)$ does not contain a line for each t. A different topology is considered there.

In the next corollary we shall give necessary and sufficient conditions for $K = K_S$ to be $w(L_1, C)$-closed in terms of the support function of S.

Let S be a mapping from T into closed and convex subset of E. The support function f of $S(t)$ is given by

$$f(t, p) = \sup \{\langle x, p\rangle | x \in S(t)\}. \tag{13}$$

For each fixed t, f is convex and lower semicontinuous in p. If $S(t)$ is not bounded then $f(t,\cdot)$ assumes $+\infty$ values. In this case, the set

$$A(t) = \{p \mid f(t,p) < +\infty\}$$

is a proper subset of E. $A(t)$ is called the *effective domain* of $f(t,\cdot)$. It is easy to check that $A(t)$ is a convex cone for each t. If $S(t)$ is measurable in t, then so is $f(\cdot,p)$ for each fixed p and A as a function of t. One can also see that for each continuous φ, $f(t,\varphi(t))$ is measurable as a function of t. We shall denote by intr. int. $A(t)$ the *intrinsic interior of* $A(t)$, that is, the interior of $A(t)$ relative to the smallest subspace of E containing $A(t)$.

COROLLARY 2. *Let S be a measurable mapping from T into closed and convex subsets of E.*

Then K_S is $w(L_1, C)$-closed if and only if the support function f of S satisfies the following condition:

(H) *there is a set $N \subset T$ of measure zero such that for each $t \in T\setminus N$ and each $a \in$ intr. int. $A(t)$, there is a $\varphi \in C$ such that $\varphi(t) = a$ and $f(t,\varphi(t))$ is integrable on a neighborhood of t.*

Proof. Sufficiency. Assume condition (H). Manifestly K_S satisfies property (P). Therefore, by our theorem (§ 3), there is a set-valued mapping P from T into closed and convex subsets of E and a sequence $\{\varphi_n\} \subset C$ such that K_p is the closure of K_S in $w(L_1, C)$ and

$$P(t) = \bigcap_{n=1}^{\infty} \{x \mid \langle x, \varphi_n(t)\rangle \leqq f(t,\varphi_n(t))\} \quad \text{a.e. in } T, \tag{14}$$

$$C_P(t) = D^0(t) = C(t) \quad \text{a.e. in } T, \tag{15}$$

where $f(t,\varphi_n(t))$ is integrable for each n, $C(t)$ is given by (7) and $D(t)$ is given by (2). Notice that in our case, in the definition of $D(t)$, $\psi_\varphi(t) = f(t,\varphi(t))$.

From (14) it follows that

$$S(t) \subset P(t) \quad \text{and} \quad C_S(t) \subset C_P(t) \quad \text{a.e. in } T, \tag{16}$$

and from the assumption (H) that

$$\text{intr. int. } A(t) \subset D(t) \quad \text{a.e. in } T. \tag{17}$$

But $C_S(t) = A^0(t)$, where $A^0(t)$ is the polar cone of $A(t)$, and (17) implies that $D^0(t) \subset A^0(t)$ a.e. in T, which together with (15) and the second inclusion of (16), gives that

$$A^0(t) = C_S(t) = C_P(t) = D^0(t) \quad \text{a.e. in } T. \tag{18}$$

Let Z be a denumerable dense subset of the positive cone of R^n, and put

$$\Phi = \{\varphi : \varphi = a_1\varphi_{n_i} + \cdots + a_n\varphi_{n_n}, \varphi_{n_i} \in \{\varphi_n\}, (a_1, \cdots, a_n) \in Z\}.$$

The set Φ is denumerable and for each $t \in T\setminus N_1$, $\{\varphi(t) : \varphi \in \Phi\}$ is a denumerable dense subset of $D(t)$ and (by (18)) also of $A(t)$, where N_1 is of measure zero.

By the convexity of $f(t,p)$ with respect to p, for each $\varphi \in \Phi$, $f(t,\varphi(t))$ is integrable on T. Set

$$Q(t) = \bigcap_{\varphi \in \Phi} \{x \mid \langle x, \varphi(t)\rangle \leqq f(t,\varphi(t))\}.$$

Since $\{\varphi(t) \mid \varphi \in \Phi\}$ is a dense subset of $A(t)$ if $t \in T \setminus N_1$, thus $Q(t) = S(t)$ a.e. in T, and by (14), $Q(t) \subset P(t)$ a.e. in T. By Corollary 1, K_Q is $w(L_1, C)$-closed. But $K_Q = K_S$ is $w(L_1, C)$ closed, which was to be proved.

Necessity. If K_S is closed, then $S(t) = P(t)$ a.e. in T, where P is given by (14). By (15), $A^0(t) = C_S(t) = D^0(t)$ a.e. in T. Therefore intr. int. $A(t) \subset D(t)$ a.e. in T, which together with the definition of $D(t)$, implies condition (H). This finishes the proof of the corollary.

Remark 2. If $S(t)$ in Corollary 2 does not contain a line for almost all t, which is equivalent to saying that $A(t)$ has nonempty interior a.e. in T, then the condition (H) can be replaced by the following: there is an $N \subset T$ of measure zero such that for each $t \in T \setminus N$ and each $a \in$ int. $A(t), f(t, a)$ is integrable on a neighborhood of t.

REFERENCES

[1] A. Lasota and C. Olech, *On the closedness of the set of trajectories of a control system*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 14 (1966), pp. 615–621.

[2] ———, *On Cesari's semicontinuity condition for set valued mappings*, Ibid., 16 (1968), pp. 711–716.

[3] C. Olech, *Lexicographical order, range of integrals and "bang-bang" principle*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 35–45.

[4] ———, *Existence theorems for optimal problems with vector valued cost functions*, Trans. Amer. Math. Soc., 136 (1969), pp. 157–180.

[5] ———, *Existence theorems for optimal problems involving multiple integrals*, J. Differential Equations, 6 (1969), pp. 512–526.

[6] ———, *A necessary and sufficient condition for lower semi-continuity of certain integral functionals*, to appear.

[7] R. T. Rockafellar, *Integrals which are convex functionals. II*, Pacific J. Math., 29 (1971), pp. 439–469.

[8] ———, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.

SIAM J. CONTROL
Vol. 12, No. 2, May 1974

AN EXISTENCE THEOREM WITHOUT CONVEXITY CONDITIONS*

LAMBERTO CESARI†

Abstract. We state and prove existence theorems for problems of optimal control which are linear in the state variables. As in previous work by L. W. Neustadt and C. Olech, no convexity condition is required. Examples are given.

Introduction. The present paper concerns problems of optimal control which admit a Mayer-type formulation as indicated below, linear in the state variables. By suitable combination of lower closure arguments as in previous papers by Cesari [1], [2], [3], and measure theoretical arguments as in previous papers by Neustadt [4] and Olech [5], [6], we obtain here an existence theorem for which, as in Neustadt's and Olech's existence statements, no convexity conditions are needed. Various corollaries follow for Lagrange problems, as well as for free problems of the calculus of variations. Examples are given.

1. We consider first the Mayer problem of the minimum of the functional

$$I[x, u] = g(t_1, x(t_1), t_2, x(t_2)), \tag{1.1}$$

with differential system

$$dx/dt = A(t)x(t) + C(t, u(t)), \tag{1.2}$$

$t \in [t_1, t_2]$, a.e., with boundary conditions

$$(t_1, x(t_1), t_2, x(t_2)) \in B \tag{1.3}$$

and constraints

$$u(t) \in U(t), \qquad t \in [t_1, t_2), \text{a.e.} \tag{1.4}$$

We denote by A a slab in the tx-space E_{n+1} defined by $A = [t_0, T] \times E_n, t_0, T$ finite. Thus, the state variable constraint usually written in the form $(t, x(t)) \in A$, $t_1 \leqq t \leqq t_2$, reduces here to the requirement $t_0 \leqq t_1 \leqq t_2 \leqq T$. Here we denote by B a given subset of the $t_1x_1t_2x_2$-space E_{2n+2}. For every $t \in [t_0, T]$, we denote by $U(t)$ a given subset of the u-space E_m, and then we denote by M_0 the set of all $(t, u) \in E_{m+1}$ with $t_0 \leqq t \leqq T, u \in U(t)$. We also denote by $A(t)$ a given $n \times n$ matrix with entries defined on $[t_0, T]$, and by $C(t, u)$ a given $n \times 1$ matrix with entries defined on M_0, so that the usual function $f(t, x, u) = (f_1, \cdots, f_n)$ of Mayer problems is here defined by $f(t, x, u) = A(t)x + C(t, u)$, $(t, x, u) \in M_0 \times E_n$. Finally, we denote by $g(t_1, x_1, t_2, x_2)$ a given scalar function defined on B.

A pair of functions $x(t) = (x^1, \cdots, x^n)$, $u(t) = (u^1, \cdots, u^m)$, $t_1 \leqq t \leqq t_2$, is said to be *admissible*, provided x is absolutely continuous (AC) in $[t_1, t_2]$, u is measurable in $[t_1, t_2], t_0 \leqq t_1 \leqq t_2 \leqq T$, and relations (1.2), (1.3) and (1.4) hold. For the sake of brevity, we may use the notation $\eta(x)$ to designate $\eta(x) = (t_1, x(t_1),$

* Received by the editors September 27, 1972.

† Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48104. This research was partially supported by the U.S. Air Force Office of Scientific Research under Project 71-2122.

$t_2, x(t_2))$, and then (1.1) and (1.3) take the form $I[x,u] = g(\eta(x))$, and $\eta(x) \in B$, respectively.

In both Neustadt's [4] and Olech's [6] work, the control space U was assumed to be a fixed compact subset of E_m. In Olech's work it was assumed that $|C(t,u)| \leqq \Phi(t)$ for all $(t,u) \in M_0$ and some fixed L-integrable function $\Phi(t) \geqq 0$ in $[t_0, T]$. No condition of this type will be required here, and the sets $U(t)$ will be assumed to be closed but not necessarily compact.

Given a set S of points $(x,y) \in E_h \times E_k$, we shall denote by $\{x\}_S$ the set $\{x\}_S = \{x \in E_h | (x,y) \in S \text{ for some } y \in E_k\}$.

2. Mayer problems with f linear in x. We proceed now to our existence theorem for Mayer problems (1.1–1.4).

THEOREM 2.1 (An existence theorem for Mayer problems with no convexity conditions). *Let $\alpha, n, 0 \leqq \alpha \leqq n$, be given integers, and for $x = (x^1, \cdots, x^n)$, let y, z denote $y = (x^1, \cdots, x^\alpha)$, $z = (x^{\alpha+1}, \cdots, x^n)$, so that $x = (y,z)$. For every $t \in [t_0, T]$, let $U(t)$ be a given nonempty closed subset of the u-space E_m, and let $M_0 \subset E_{m+1}$ be the set of all (t,u) with $t \in [t_0, T]$, $u \in U(t)$. Let us assume that M_0 is closed. Let $A(t) = [a_{ij}(t), i = 1, \cdots, n, j = 1, \cdots, \alpha]$ and $C(t,u) = [c_1(t,u), i = 1, \cdots, n]$ be $n \times \alpha$ and $n \times 1$ matrices with entries continuous in $[t_0, T]$ and M_0, respectively. Let $H(t,u)$ be a scalar continuous function on M_0, and let us assume that $H(t,u)$ and $c_{\alpha+1}, \cdots, c_n$ are not negative on M_0.*

Assumption (m$_1$): *For every $\varepsilon > 0$, there is a number $N \geqq 0$ such that $|c_i(t,u)| \leqq \varepsilon H(t,u), i = 1, \cdots, \alpha$, and $1 \leqq \varepsilon H(t,u)$ for all $(t,u) \in M_0$ with $|u| \geqq N$.*

Assumption (m$_2$): *There is an L-integrable scalar function $\psi(t) \geqq 0, t \in [t_0, T]$, and a constant $G \geqq 0$ such that $|c_i(t,u)| \leqq \psi(t) + GH(t,u), i = \alpha + 1, \cdots, n$, for all $(t,u) \in M_0$.*

Let B be a closed subset of the $t_1 y_1 z_1 t_2 y_2 z_2$-space E_{2n+2} such that the corresponding set $\{z_1\}_B$ is bounded and $\{z_2\}_B = E_{n-\alpha}$; hence B is certainly of the form $B = B_0 \times E_{n-\alpha}$, where B_0 is a closed subset of $E_{n+2+\alpha}$. Let $g(t_1, y_1, z_1, t_2, y_2, z_2)$ be a real-valued continuous function defined on B which is monotone nondecreasing with respect to each component $x_2^{\alpha+1}, \cdots, x_2^n$ of z_2. Let P be a compact subset of $A_0 = [t_0, T] \times E_\alpha$, and let Ω be the class of all admissible pairs $x(t), u(t), t_1 \leqq t \leqq t_2$, with $t_0 \leqq t_1 \leqq t_2 \leqq T$, whose trajectories x contain at least one point $(t^*, y(t^*)) \in P$, and for which $H(t, u(t))$ is L-integrable in $[t_1, t_2]$ and $\int_{t_1}^{t_2} H(t,u(t))\,dt \leqq M_1$ for some constant $M_1 \geqq 0$. We assume that Ω is not empty. Then the functional $I[x,u] = g(\eta(x))$ has an absolute minimum in Ω.

Note that in this existence theorem, we are concerned with the problem of the minimum of the functional

$$I[x,u] = g(t_1, x(t_1), t_2, x(t_2)), \tag{2.1}$$

with differential system

$$dx^i/dt = \sum_{j=1}^{\alpha} a_{ij}(t)x^j + c_i(t, u(t)), \qquad i = 1, \cdots, n, \tag{2.2}$$

with constraints

$$u(t) \in U(t), \quad \int_{t_1}^{t_2} H(t, u(t))\,dt \leqq M_1 \tag{2.3}$$

and boundary conditions

$$(2.4) \qquad (t_1, x(t_1), t_2, x(t_2)) \in B,$$

where $A = [t_0, T] \times E_n$, where $B = B_0 \times E_{n-\alpha}$ is independent of $x_2^{\alpha+1}, \cdots, x_2^n$, and where g is monotone not decreasing with respect to $x_2^{\alpha+1}, \cdots, x_2^n$. As we have seen in [2], the constraint in integral form is often implicit in the data. In Lagrange problems it may be the cost functional itself (cf. § 3). We denote below by $|y|$ the Euclidean norm, and by $\|A\|$ any norm of the matrix A such that $|Ay| \leqq \|A\| \, |y|$.

Here the sets $U(t)$ are closed but not necessarily compact. Under this hypothesis, even an Olech-type condition that there is an L-integrable function $\Phi(t) \geqq 0$, $t_0 \leqq t \leqq T$, such that

$$|c_i(t, u)| \leqq \Phi(t) \quad \text{for all } (t, u) \in M_0 \quad \text{and} \quad i = 1, \cdots, n,$$

alone does not guarantee the existence of the minimum. This can be seen by the following simple example. Take $n = m = 1$, $f = (1 + u^2)^{-1}$, $t_1 = 0$, $t_2 = 1$, $x(0) = 1$, $I = g = x(t_2)$, $H = \Phi = 1$. Then, $x(t_2) > 1$, $\inf x(t_2) = 1$, and the minimum does not exist. Note that the continuity requirements in Theorem 2.1 can be disregarded at the points of sets whose projections on the t-axis are finite.

Remark 1. As we shall see in the proof of Theorem 2.1, Assumption (m_1) will guarantee that the elements y of the class $\{y\}_\Omega$ are equiabsolutely continuous. Thus, the requirement $(t^*, y(t^*)) \in P$ will guarantee that the same elements y are equibounded. We shall show that Assumption (m_2) implies that the total variations of the elements z of the class $\{z\}_\Omega$ are equibounded. Then the requirement that the set $\{z_1\}_B$ be bounded will imply that the same elements z are equibounded.

Remark 2. Variable sets $R(t)$ of E_N, $t_0 \leqq t \leqq T$, are said to satisfy Property (Q) at a point $\bar{t} \in [t_0, T]$ provided $R(\bar{t}) = \bigcap_\varepsilon \mathrm{cl}\,\mathrm{co} \bigcup R(t)$, where $\bigcup$ is taken for all $t \in [t_0, T]$ with $|t - \bar{t}| \leqq \varepsilon$. Below, we shall say that the sets $R(t)$ have Property (Q) if this property holds at almost all $\bar{t} \in [t_0, T]$. For every $t \in [t_0, T]$ let $Q_H(t)$ denote the set of all $\tilde{z} = (z^0, z^1, \cdots, z^n) \in E_{n+1}$ such that $z^0 \geqq H(t, u)$, $z^i = c_i(t, u)$, $i = 1, \cdots, \alpha$, $z^i \geqq c_i(t, u)$, $i = \alpha + 1, \cdots, n$, for some $u \in U(t)$. We shall need these sets in the proof of Theorem 2.1. These sets need not be convex, but the sets $\mathrm{co}\, Q_H(t)$ are certainly convex. The closedness of M_0 and Assumption (m_1) will guarantee that the sets $\mathrm{co}\, Q_H(t)$ are closed and satisfy Property (Q) on $[t_0, T]$. Assumption (m_1) in Theorem 2.1 can be replaced by the following weaker set of assumptions.

Assumption (m_1'): Given $\varepsilon > 0$ there is a scalar L-integrable function $\psi_\varepsilon(t) \geqq 0$, $t \in [t_0, T]$, such that $|c_i(t, u)| \leqq \psi_\varepsilon(t) + \varepsilon H(t, u)$, $i = 1, \cdots, \alpha$, for all $(t, u) \in M_0$.

Assumption (m_1''): The sets $\mathrm{co}\, Q_H(t)$ satisfy Property (Q) at every point $t \in [t_0, T]$ with the exception perhaps of a set of measure zero. Growth condition (m_1') is known to be weaker than (m_1). Growth condition (m_1') alone still guarantees that the elements y of the class $\{y\}_\Omega$ are equiabsolutely continuous, but is too weak to imply Property (Q) for the sets $\mathrm{co}\, Q_H(t)$. We shall denote by $Q(t)$ the set of all $z = (z_1, \cdots, z^n)$ such that $z^i = c_i(t, u)$, $i = 1, \cdots, \alpha$, $z^i \geqq c_i(t, u)$, $i = \alpha + 1, \cdots, n$.

Whenever an Olech-type condition is verified, say, $|c_i(t, u)| \leqq \Phi(t)$ with $\Phi(t) \geqq 0$, $t_0 \leqq t \leqq T$, $\Phi \in L_1$, then we may well take $\alpha = n$, $x = y$, the functions $y \in \{y\}_\Omega$ are equibounded and equiabsolutely continuous, and we need only to know that the sets co $Q_H(t)$ have Property (Q). If no H is given, then we need only to know that the sets co $Q(t)$ have Property (Q).

Remark 3. Assumption (m_1) could well be replaced by the stronger assumption of the Tonelli–Nagumo type which is often easier to verify:

Assumption (m_1^*): There are constants $C, D \geqq 0$ and a continuous scalar function $\varphi(\zeta)$. $0 \leqq \zeta < +\infty$, such that $\varphi(\zeta)/\zeta \to +\infty$ as $\zeta \to +\infty$, $H(t,u) \geqq \varphi(|u|)$, $|c_i(t,u)| \leqq C + D|u|$, $i = 1, \cdots, \alpha$, for all $(t,u) \in M_0$.

Remark 4. In Theorem 2.1, the requirement $H(t,u) \geqq 0$, $c_i(t,u) \geqq 0$, $i = \alpha + 1, \cdots, n$, could be replaced by the weaker assumption: there is some scalar L-integrable function $\psi(t) \geqq 0$ such that $H(t,u) \geqq -\psi(t)$, $c_i(t,u) \geqq -\psi(t)$, $i = \alpha + 1, \cdots, n$, for all $(t,u) \in M_0$.

Proof of Theorem 2.1. As usual, we denote by $\{u\}_\Omega$ the class of all control functions u relative to admissible pairs $x(t), u(t), t_1 \leqq t \leqq t_2$ actually in the class Ω. Thus $\{c_i(t, u(t))\}_\Omega$ will denote the class of functions $c_i(t, u(t))$ for all $u \in \{u\}_\Omega$. Let us prove that the classes $\{c_i(t, u(t))\}_\Omega$, $i = 1, \cdots, \alpha$, are made up of functions which are equiabsolutely integrable in their interval of definition. Indeed, let $\varepsilon > 0$ be an arbitrary number, take $\varepsilon' = (2M_1)^{-1}\varepsilon$, and let $N \geqq 0$ be a number such that $|c_i(t,u)| \leqq \varepsilon' H(t,u)$, $i = 1, \cdots, \alpha$, for all $(t,u) \in M_0$ with $|u| \geqq N$. Then the set $M'_0 = \{(t,u) \in M_0 \mid |u| \leqq N\}$ is compact, and there is another constant $N' > 0$ such that $|c_i(t,u)| \leqq N'$ for all $(t,u) \in M'_0$, $i = 1, \cdots, \alpha$. Let $\delta = \varepsilon' M_1 (N')^{-1}$, and for any admissible pair $x(t), u(t), t_1 \leqq t \leqq t_2$, in Ω, let H denote any measurable subset of $[t_1, t_2]$ of measure $|H| \leqq \delta$. If H_1 denotes the subset $H_1 = \{t \in H \mid |u(t)| \leqq N\}$, and $H_2 = H - H_1$, then for every $i = 1, \cdots, \alpha$, we have

$$\int_H |c_i(t, u(t))|\, dt = \left(\int_{H_1} + \int_{H_2}\right) |c_i(t, u(t))|\, dt$$

$$\leqq \delta N' + \varepsilon' \int_{t_1}^{t_2} H(t, u(t))\, dt \leqq 2\varepsilon' M_1 = \varepsilon.$$

Let us prove that the vector functions $y(t) = (x^1, \cdots, x^\alpha)$, $t_1 \leqq t \leqq t_2$, of the class $\{y\}_\Omega$ are equiabsolutely continuous and equibounded in their intervals of definition. We know that $(t^*, y(t^*)) \in P$ for at least one point t^*, $t_0 \leqq t_1 \leqq t^* \leqq t_2 \leqq T$. Hence,

$$y(t) = X(t, t^*)\left[y(t^*) + \int_{t^*}^{t} X^{-1}(\tau, t^*) C_1(\tau, u(\tau))\, d\tau\right], \qquad t_1 \leqq t \leqq t_2, \tag{2.5}$$

where $y(t) = (x^1, \cdots, x^\alpha)$, $C_1 = (c_1, \cdots, c_\alpha)$, $A_1 = [a_{ij}(t), i, j = 1, \cdots, \alpha]$, and $X(t, t^*)$ is the fundamental $\alpha \times \alpha$ matrix of solutions of the homogeneous system $dy/dt = A_1(t)y$ with $X(t^*, t^*) = I$. If N_1, N'_1 denote the constants N, N' above for $\varepsilon = 1$, then we have $|c_i| \leqq N'_1$ for $|u| \leqq N_1$ and $|c_i| \leqq H$ for $|u| \geqq N_1$. In any case, we have $|c_i(t,u)| \leqq N'_1 + H(t,u)$ for all $(t,u) \in M_0$, $i = 1, \cdots, \alpha$. In (2.5), $(t^*, y(t^*)) \in P$ where P is compact, and the entries of $X(t, t^*)$, $X^{-1}(t, t^*)$ are certainly continuous, equiabsolutely continuous, and equibounded for $t_0 \leqq t, t^* \leqq T$. Hence, for some constant $M_2 > 0$, we have

$$\left|\int_{t^*}^{t} X^{-1}(\tau, t^*) C_1(\tau, u(\tau))\, d\tau\right| \leqq M_2 \alpha^2 [(T - t_0)N'_1 + M_1].$$

Thus, for any system of nonoverlapping subintervals $[\alpha_s, \beta_s], s = 1, \cdots, \mu$, of $[t_1, t_2]$ we have, for a suitable constant M_2,

$$\sum_s |y(\beta_s) - y(\alpha_s)| \leqq (M_2 + M_2\alpha^2((T - t_0)N_1' + M_1))\sum_s |X(\beta_s, t^*) - X(\alpha_s, t^*)|$$
$$+ M_2^2 \sum_s \sum_{i=1}^{\alpha} \int_{\alpha_s}^{\beta_s} |c_i(t, u(t))|\, dt.$$

We have proved that the vector functions y of the class $\{y\}_\Omega$ are equiabsolutely continuous in their interval of definition. Since $(t^*, y(t^*)) \in P$, where P is compact, and $t_0 \leqq t_1 \leqq t^* \leqq t_2 \leqq T$, we conclude that the same functions y are also equibounded.

Let us prove that the AC vector functions $z(t) = (x^{\alpha+1}, \cdots, x^n), t_1 \leqq t \leqq t_2$, of the class $\{z\}_\Omega$ are equibounded with uniformly bounded total variation. Indeed, for some constants G and M_2, we have $|a_{ij}(t)| \leqq M_2, |y(t)| \leqq M_2, |c_i(t, u)| \leqq \psi(t) + GH(t, u)$, and

$$dx^i/dt = \sum_{j=1}^{\alpha} a_{ij}(t)x^j + c_i(t, u(t)), \qquad i = \alpha + 1, \cdots, n.$$

Hence

$$|dx^i/dt| \leqq \alpha M_2^2 + \psi(t) + GH(t, u(t)), \qquad t_1 \leqq t \leqq t_2, \quad i = \alpha + 1, \cdots, n,$$
$$\int_{t_1}^{t_2} |dx^i/dt|\, dt \leqq \alpha M_2^2(T - t_0) + \int_{t_0}^{T} \psi(t)\, dt + G\int_{t_0}^{t_2} H(t, u(t))\, dt$$
$$\leqq \alpha M_2^2(T - t_0) + M_2 + GM_1.$$

On the other hand, the initial point $z(t_1) = (x^{\alpha+1}(t_1), \cdots, x^n(t_1))$ belongs to the set $\{z_1\}_B$, which is bounded by assumption. We have proved that the vector functions $z(t), t_1 \leqq t \leqq t_2$, of the class $\{z\}_\Omega$ are equibounded with equibounded total variation.

We shall now consider the generalized solutions $x(t) = (y(t), z(t)), p(t), u^*(t), t_1 \leqq t \leqq t_2$, of problem (2.1)–(2.4) under consideration, with the same functional $J(x, p, u^*) = g(\eta(x))$. Here $x(t) = (x^1, \cdots, x^n), p(t) = (p^1, \cdots, p^h), u^*(t) = (u^{(1)}, \cdots, u^{(h)}), t_1 \leqq t \leqq t_2$, we can take $h = n + 2$, and then $p(t) \in \Gamma = \{p_s \geqq 0, p_1 + \cdots + p_h = 1\}, u^{(s)}(t) \in U(t), s = 1, \cdots, h$,

$$dx^i/dt = \sum_{j=1}^{\alpha} a_{ij}(t)x^j(t) + c_i(t, p(t), u^*(t)), \qquad i = 1, \cdots, n,$$

$$c_i(t, p, u^*) = \sum_{s=1}^{h} p_s c_i(t, u^{(s)}),$$

$$H^*(t, p, u^*) = \sum_{s=1}^{h} p_s H(t, u^{(s)}).$$

We shall consider the class Ω^* of all generalized solutions $x(t) = (y, z), p(t), u^*(t), t_0 \leqq t_1 \leqq t \leqq t_2 \leqq T$, whose trajectories x possess at least one point $(t^*, y(t^*)) \in P$, with $(t_1, x(t_1), t_2, x(t_2)) \in B$, with $H^*(t, p(t), u^*(t))$ L-integrable in $[t_1, t_2]$ and

$\int_{t_1}^{t_2} H^*(t, p(t), u^*(t))\,dt \leqq M_1$. The argument above can be repeated, and we conclude that the vector functions $y(t) = (x^1, \cdots, x^\alpha)$, $t_1 \leqq t \leqq t_2$, of the class $\{y\}_{\Omega^*}$ are equiabsolutely continuous and equibounded, and that the vector functions $z(t) = (x^{\alpha+1}, \cdots, x^n)$, $t_1 \leqq t \leqq t_2$, of the class $\{z\}_{\Omega^*}$ are equibounded with equibounded total variation.

We conclude that the endpoints $\eta(x) = (t_1, x(t_1), t_2, x(t_2))$ of the generalized trajectories x form a bounded subset of B, that $\Omega \subset \Omega^*$, that the numbers $j = \inf_{\Omega^*} J(x, p, u^*)$, $i = \inf_\Omega I(x, u)$ are finite and that $j \leqq i$.

We shall now apply statement (2.3.iii) of [2]. We conclude that the sets $\operatorname{co} Q_H(t)$ are all closed and satisfy Property (Q) on $[t_0, T]$. We shall now apply a modified form of existence theorem (5.3A*) of [2]. Indeed, first we consider a minimizing sequence $x_k(t) = (y_k(t), z_k(t))$, $p_k(t)$, $u_k^*(t)$, $t_{1k} \leqq t \leqq t_{2k}$, $k = 1, 2, \cdots$, with $J(x_k, p_k, u_k^*) \to j$ as $k \to \infty$. By taking a suitable subsequence, we can well assume that $t_{1k} \to t_1$, $t_{2k} \to t_2$, $t_0 \leqq t_1 \leqq t_2 \leqq T$, that y_k converges in the uniform topology toward an AC function $y(t)$, $t_1 \leqq t \leqq t_2$, and that $y_k(t_{1k}) \to y(t_1)$, $y_k(t_{2k}) \to y(t_2)$, and also $z_k(t_{1k}) \to z_0$ for some $z_0 \in E_{n-\alpha}$. Now we take

$$y_k(t) = y_k(t_{1k}) + \int_{t_{1k}}^t A_1(\tau) y_k(\tau)\,d\tau + Y_k(t),$$

$$z_k(t) = z_k(t_{1k}) + \int_{t_{1k}}^t A_2(\tau) y_k(\tau)\,d\tau + Z_k(t),$$

$$Y_k(t) = \int_{t_{1k}}^t C_1(\tau, p_k(\tau), u_k^*(\tau))\,d\tau,$$

$$Z_k(t) = \int_{t_{1k}}^t C_2(\tau, p_k(\tau), u_k(\tau))\,d\tau.$$

Thus, y_k and z_k are decomposed into two parts, the first of which converges uniformly. It is enough now to repeat on the second parts Y_k, Z_k the same arguments of (5.3A*) of [2], in particular of closure theorem III of (3.1) of [2].

We conclude that the problem under consideration has a generalized optimal solution $x(t), p(t), u^*(t)$, $t_1 \leqq t \leqq t_2$, with

$$\begin{aligned} &J(x, p, u^*) = g(t_1, x(t_1), t_2, x(t_2)) = j, \\ &dx^i/dt = \sum_{j=1}^{\alpha} a_{ij}(t) x^j(t) + \sum_{s=1}^{h} p_s(t) c_i(t, u^{(s)}(t)), \qquad i = 1, \cdots, n, \\ &u^{(s)}(t) \in U(t), \qquad s = 1, \cdots, h, \\ &(t^*, y(t^*)) \in P, \qquad (t_1, x(t_1), t_2, x(t_2)) \in B, \\ &\sum_{s=1}^{h} \int_{t_1}^{t_2} p_s(t) H(t, u^{(s)}(t))\,dt \leqq M_1. \end{aligned} \tag{2.6}$$

Hence, for $i = 1, 2$, we have

$$y(t_i) = X(t_i, t^*)\left[y(t^*) + \sum_{s=1}^{h} \int_{t^*}^{t_i} p_s(\tau) X^{-1}(\tau, t^*) C_1(\tau, u^{(s)}(\tau))\,d\tau \right], \tag{2.7}$$

$$z(t_i) = z(t^*) + \int_{t^*}^{t_i} A_2(t)y(t)\,dt + \sum_{s=1}^{h} \int_{t^*}^{t_i} p_s(\tau)C_2(\tau, u^{(s)}(\tau))\,d\tau, \tag{2.8}$$

where $A_2(t) = [a_{ij}, i = \alpha + 1, \cdots, n, j = 1, \cdots, \alpha]$, $C_2(t) = [c_i, i = \alpha + 1, \cdots, n]$. Relations (2.8) can be written also in the form

$$\begin{aligned} z(t_i) = Z_i + \sum_{s=1}^{h} \int_{t^*}^{t_i} p_s(\tau)\Bigg\{ & C_2(\tau, u^{(s)}(\tau)) \\ & + \left(\int_{\tau}^{t_i} A_2(t)X(t, t^*)X^{-1}(\tau, t^*)\,dt\right)C_1(\tau, u^{(s)}(\tau)\Bigg\}\,d\tau, \end{aligned} \tag{2.9}$$

where

$$Z_i = z(t^*) + \int_{t^*}^{t_i} A_2(t)X(t, t^*)y(t^*)\,dt, \qquad i = 1, 2.$$

By applying (2.iv) of [3] to the sums with respect to s in (2.6), (2.7) and (2.9), we see that each interval $[t^*, t_i]$, $i = 1, 2$, can be decomposed into h disjoint measurable subsets $E_{i1}, \cdots, E_{ih}$ such that

$$\sum_{i=1}^{2} \sum_{s=1}^{h} \int_{E_{is}} H(t, u^{(s)}(t)\,dt \leqq M_1,$$

$$y(t_i) = X(t_i, t^*)\left[y(t^*) + \sum_{s=1}^{h} \int_{E_{is}} X^{-1}(\tau, t^*)C_1(\tau, u^{(s)}(\tau)\,d\tau\right],$$

$$z(t_i) = Z_i + \sum_{s=1}^{h} \int_{E_{is}} \left\{C_2(\tau, u^{(s)}(\tau)) + \left(\int_{\tau}^{t_i} A_2(t)X(t, t^*)X^{-1}(\tau, t^*)\,dt\right)C_1(\tau, u^{(s)}(\tau))\right\}\,d\tau.$$

Let us consider now the (usual) control function $\bar{u}(t)$, $t_1 \leqq t \leqq t_2$, defined by taking $\bar{u}(t) = u^{(s)}(t)$ for $t \in E_{is}$, $s = 1, \cdots, h$, $i = 1, 2$, and let us define the (usual) trajectory $\bar{x}(t)$, $t_1 \leqq t \leqq t_2$, by taking

$$\begin{aligned} d\bar{x}/dt &= A(t)\bar{x}(t) + C(t, \bar{u}(t)), \qquad t_1 \leqq t \leqq t_2, \text{a.e.}, \\ \bar{x}(t^*) &= x(t^*), \end{aligned}$$

that is, $\bar{x}(t) = (\bar{y}, \bar{z})$, $t_1 \leqq t \leqq t_2$, with $\bar{y}(t^*) = y(t^*)$, $\bar{z}(t^*) = z(t^*)$. We see immediately that $\bar{x}(t_i) = x(t_i)$, $i = 1, 2$, that is, $\bar{y}(t_i) = y(t_i)$, $\bar{z}(t_i) = z(t_i)$, $i = 1, 2$, and hence

$$i \leqq I[\bar{x}, \bar{u}] = g(t_1, \bar{x}(t_1), t_2, \bar{x}(t_2)) = g(t_1, x(t_1), t_2, x(t_2)) = J(x, p, u^*) = j.$$

Since $j \leqq i$, we conclude that $I[\bar{x}, \bar{u}] = i = j$, that is, $\bar{x}, \bar{u}$ is a usual optimal pair for the original problem (2.1)–(2.4).

Remark 5. The requirement $\{z_2\}_B = E_{n-\alpha}$ in Theorem 2.1 can be replaced by the following one, which is relevant in applications. Indeed, it is enough to know that the boundary conditions concerning $z_2 = z(t_2) = (x^{\alpha+1}(t_2), \cdots, x^n(t_2))$, are of the types:

$$\text{either} \quad z^i(t_2) \in E_1, \quad \text{or} \quad z^i(t_2) \leqq \zeta^i, \qquad -\infty < \zeta^i \leqq +\infty,$$

$i = \alpha + 1, \cdots, n$, where ζ^i are given numbers (or $+\infty$ as stated). This can be seen from the proof of (5.3.A*) in [2].

Remark 6. In many applications, $H(t, u)$ is one of the $c_i(t, x), i = \alpha + 1, \cdots, n$, say $H(t, u) = c_n(t, u), (t, u) \in M_0$, and it is not known a priori that $\int_{t_1}^{t_2} H(t, u(t))dt \leqq M_1$ for some constant M_1. In this situation, we may need the following lemma.

LEMMA 2.2. *Let $H(t, u) = c_n(t, u) \geqq 0$, and assume that the boundary conditions are so given that $x^n(t_1)$ can take only values $|x^n(t_1)| \leqq M_1$, and the only requirement on $x^n(t_2)$ is of the form $x^n(t_2) \leqq \zeta^n$, for some constants M_1 and ζ^n. Under the remaining conditions of Theorem* 2.1, *there is a constant M_3 such that $\int_{t_1}^{t_2} H(t, u(t))dt \leqq M_3$.*

Proof. Let $M_2 = |\zeta^n|$. Let $dx^n/dt = A_n(t)y + c_n(t, u(t))$ denote the nth equation (2.2), so that $A_n(t)$ is a $1 \times \alpha$ matrix. Let us choose constants $D, L_0, L > 0$ so that $T - t_0 = D$, the compact set P is completely contained in the set $\{(t, y) | t_0 \leqq t \leqq T, |y| \leqq L_0\}$, and

$$\|X(t, t^*)\|, \|X^{-1}(t, t^*)\|, |A_n(t)| \leqq L, \qquad t_0 \leqq t, t^* \leqq T.$$

Let us take a number $\varepsilon > 0$ so small that $nDL^3\varepsilon \leqq \frac{1}{2}$. Then there is a number $N(\varepsilon) > 0$ such that $t_0 \leqq t \leqq T$, $u \in U(t)$, $|u| \geqq N(\varepsilon)$ imply $|c_i(t, u)| \leqq \varepsilon c_n(t, u) = \varepsilon H(t, u)$, $i = 1, \cdots, \alpha$. Now the set $M_{0\varepsilon} = \{(t, u) | t_0 \leqq t \leqq T, u \in U(t), |u| \leqq N(\varepsilon)\}$ is compact. Hence, there is a constant $L_2 > 0$ such that $|c_i(t, u)| \leqq L_2$ for all $(t, u) \in M_{0\varepsilon}, i = 1, \cdots, \alpha$. Finally, for the vector $C_1(t, u) = (c_1, \cdots, c_\alpha)$, we have $|C_1(t, u)| \leqq \alpha L_2 \leqq nL_2$ for all $t_0 \leqq t \leqq T, u \in U(t), |u| \leqq N(\varepsilon)$, while we have $|C_1(t, u)| \leqq \alpha\varepsilon H(t, u) \leqq n\varepsilon H(t, u)$ for all $t_0 \leqq t \leqq T, u \in U(t), |u| \geqq N(\varepsilon)$. Note that (2.5) yields

$$|y(t)| \leqq \|X(t, t^*)\| \left[|y(t^*)| + \int_{t^*}^{t} \|X^{-1}(\tau, t^*)\| \, |C_1(\tau, u(\tau))| \, d\tau \right].$$

We denote by E' and E'' the subsets of $[t^*, t]$ where $|u(t)| \leqq N(\varepsilon)$ and $|u(t)| > N(\varepsilon)$, respectively. By splitting the integral above into integrals extended over E' and E'' and noting that $|E'|, |E''| \leqq D$, we have

$$|y(t)| \leqq L \left[L_0 + nDLL_2 + nL\varepsilon \int_{t_1}^{t_2} H(t, u(t))dt \right].$$

Finally, from the nth equation (2.2) we obtain, by integration on $[t_1, t_2]$,

$$\begin{aligned} \int_{t_1}^{t_2} H(t, u(t)) \, dt &= \int_{t_1}^{t_2} c_n(t, u(t)) \, dt \\ &= x^n(t_2) - x^n(t_1) - \int_{t_1}^{t_2} A_n(t) y(t) \, dt \\ &\leqq M_2 + M_1 + LD \left[LL_0 + nDL^2L_2 + nL^2\varepsilon \int_{t_1}^{t_2} H(t, u(t)) \, dt \right], \end{aligned}$$

where $nDL^3\varepsilon \leqq \frac{1}{2}$. Thus

$$\int_{t_1}^{t_2} H(t, u(t)dt \leqq 2(M_2 + M_1 + DL^2L_0 + nD^2L^3L_2) = M_3.$$

3. Lagrange problems with f_0 and f linear in x. We consider now the Lagrange problem of the minimum of a functional

$$I[x,u] = \int_{t_1}^{t_2} [A_0(t)x(t) + C_0(t,u(t))]\,dt, \tag{3.1}$$

with differential system

$$dx/dt = A(t)x(t) + C(t,u(t)), \tag{3.2}$$

constraints

$$u(t) \in U, \quad \text{or} \quad u(t) \in U(t), \tag{3.3}$$

and boundary conditions

$$(t_1, x(t_1), t_2, x(t_2)) \in B. \tag{3.4}$$

Here $x = (x^1, \cdots, x^n)$, $u = (u^1, \cdots, u^m)$ and, with the usual notations, $f_0(t,x,u) = A_0(t)x + C_0(t,u)$, $f(t,x,u) = A(t)x + C(t,u)$, where A_0, C_0, A, C are $1 \times n$, 1×1, $n \times n$, $n \times 1$ matrices whose entries are functions defined on some fixed interval $[t_0, T]$ or the set M_0, respectively.

By introducing the auxiliary variable x^{n+1} with differential equation $dx^{n+1}/dt = A_0(t)x + C_0(t,u)$, and initial data $x^{n+1}(t_1) = 0$, the problem above is reduced to a Mayer problem relative to the $n + 1$ state variables $\tilde{x} = (x^1, \cdots, x^n, x^{n+1}) = (x, x^{n+1})$, the same control variable $u = (u^1, \cdots, u^m)$, and functional $I[x,u] = g(\eta(\tilde{x})) = x^{n+1}(t_2)$.

A pair $x(t), u(t), t_1 \leqq t \leqq t_2$, is said to be *admissible for the problem* (3.1)–(3.4) provided x is AC, u is measurable, relations (3.2)–(3.4) are satisfied, and $C_0(t, u(t))$ is L-integrable in $[t_1, t_2]$. We shall assume below that A, C, A_0, C_0 are all continuous, $t_0 \leqq t_1 \leqq t_2 \leqq T$, for some fixed t_0, T, and we shall take for A the slab $A = [t_0, T] \times E_n$, (or $A = [t_0, T] \times E_{n+1}$ in the Mayer form).

We shall now apply our existence theorem (Theorem 2.1). In applying Theorem 2.1 we shall take $\alpha = n, H(t,u) = C_0(t,u)$. By M_0 we denote, as usual, the set of all $(t,u) \in E_{m+1}$ with $t \in [t_0, T], u \in U(t)$. In searching for the minimum of $I[x,u]$, we may well limit ourselves to the admissible pairs (x,u) with $I[x,u] \leqq L$ for some constant L. Thus, we have for $x^n(t), t_1 \leqq t \leqq t_2$, the boundary conditions $x^n(t_1) = 0, x^n(t_2) = I[x,u] \leqq L$. Because of Lemma 2.2, we conclude that for some constant $M_3 \geqq 0$, we have $\int_{t_1}^{t_2} H(t,u(t))\,dt \leqq M_3$ for all admissible pairs x, u with $I[x,u] \leqq L$. The following theorem is now a corollary of Theorem 2.1, and of Lemma 2.2.

THEOREM 3.1 (An existence theorem for Lagrange problems with no convexity conditions). *Let $A = [t_0, T] \times E_n, t_0, T$ finite, and for every $t \in [t_0, T]$, let $U(t)$ be a given subset of the u-space E_m. Let $f_0 = A_0(t)x + C_0(t,u), f = A(t)x + C(t,u)$, where $A_0(t), A(t)$, and $C_0(t,u), C(t,u)$ are $1 \times n, n \times n, 1 \times 1, n \times 1$ given matrices with entries continuous in $[t_0, T]$ and M_0, respectively. We require the set M_0 to be closed, and $C_0(t,u)$ to be nonnegative.*

Assumption $(\bar{\mathfrak{m}}_1)$: *For every $\varepsilon > 0$, there is an $N \geqq 0$ such that $|C(t,u)| \leqq \varepsilon C_0(t,u)$, and $1 \leqq \varepsilon C_0(t,u)$ for all $(t,u) \in M_0$ with $|u| \geqq N$.*

Let B be any given closed subset of the $t_1x_1t_2x_2$-space E_{2n+2}, let P be a compact subset of $A = [t_0, T] \times E_n$ and let Ω be the class, which we suppose is not empty, of all admissible pairs $x(t)$, $u(t)$, $t_1 \leqq t \leqq t_2$, whose trajectories x contain at least one point $(t^*, x(t^*)) \in P$. Then the functional (3.1) *has an absolute minimum in Ω.*

4. Examples.

Example 1. Let us consider the Mayer problem with $n = 3$, $m = 1$, system $x' = (1 - t)x + y$, $y' = u$, $z' = |1 - u^2|$, $U = E_1$, $t_1 = 0$, $t_2 = 1$, $x(0) = x_1$, $y(0) = y_1$, $z(0) = 0$, $x(1) = x_2$, $y(1) = y_2$, functional $I[x, y, z, u] = z(1)$. The sets $A = [0, 1] \times E_3$, $M = A \times E_1$, $B = (0, x_1, y_1, 0, 1, x_2, y_2) \times E_1$ are closed. Here $A = [1 - t, 1, 0; 0, 0, 0; 0, 0, 0]$, $C = (c_1, c_2, c_3)$, with $c_1 = 0$, $c_2 = u$, $c_3 = |1 - u^2|$. We take $\alpha = n - 1 = 2$, $H(t, u) = |1 - u^2|$. Here $|1 - u^2| \to +\infty$ $|1 - u^2|/|u| \to +\infty$ as $|u| \to +\infty$; hence, given $\varepsilon > 0$, there is some $N(\varepsilon) > 0$ such that $|u| \geqq N(\varepsilon)$ implies $1 \leqq \varepsilon H(t, u), |u| \leqq \varepsilon H(t, u)$, and consequently $|c_1|, |c_2| \leqq \varepsilon H(t, u)$. Thus, requirement (m_1) of Theorem 2.1 is satisfied. Since $c_3 = H(t, u)$, also requirement (m_2) of Theorem 2.1 is satisfied with $\psi = 0, G = 1$. Here B has the property required in Theorem 2.1, since $\{z_1\}_B = \{0\}$ is a fixed point and $\{z_2\}_B = E_1$. Finally, $g = z(1)$ increases with $z(1)$. Here c_3 and H are not negative, and $\int_0^1 H(t, u(t))\, dt = \int_0^1 c_3\, dt = z(1) = g$. Thus, in the search for the minimum of $I = z(1)$, we can limit ourselves to the class Ω of (usual) solutions with $\int_0^1 H\, dt = I \leqq L$ for some constant L. Moreover, every trajectory starts at the fixed point $(x_1, y_1, 0)$ at $t = t_1$. Thus, whenever we can satisfy the boundary conditions, then, by force of Theorem 3.1, the functional $I = z(1)$ has an absolute minimum under the constraints. Note that the sets $Q(t, x, y) = \{z^1 = (1 - t)x + y, z^2 = u, z^3 \geqq |1 - u^2|, u \in E_1\}$ are closed but certainly not convex.

Example 2. Let us consider the Mayer problem with $n = 2, m = 1$, system

$$x' = (1 - t)x + t^2 y + t^{-1/2} \cos u,$$

$$y' = -t^2 x + (1 + t)y + t^{-1/2} \sin u,$$

control space $U = E_1$, $t_1 = 0$, $t_2 \geqq 0$, $x(0) = a$, $y(0) = b$, and functional $I[x, y, u] = (x_2^2 + y_2^2)^{1/2} + t_2$, where $x_2 = x(t_2), y_2 = y(t_2)$. Since $I \to +\infty$ as $t_2 \to +\infty$ uniformly, we can limit ourselves to admissible pairs (usual solutions) defined on intervals $[0, t_2]$ with $0 \leqq t_2 \leqq T$ for some $T > 0$. Thus $A = [0, T] \times E_2, M = A \times E_1$, $B = (0, a, b) \times [0, T] \times E_2$, We have here $A = [1 - t, t^2; -t^2, 1 + t]$, $C(t, u) = (c_1, c_2)$. The sets $Q(t)$ (as defined in Remark 2) are certainly closed, but not convex. On the other hand, $|c_1|, |c_2| \leqq t^{-1/2}$ for all $0 \leqq t \leqq T, -\infty < u < +\infty$. Thus, Olech's assumption is satisfied with $\Phi(t) = t^{-1/2}$. The initial point $(0, x_1, y_1)$ is fixed, and the set $U = E_1$ is independent of t. By force of the last line in Remark 2, the functional $I = (x_2^2 + y_2^2)^{1/2} + t_2$ has an absolute minimum under the constraints.

Example 3. Let us consider the Mayer problem with $n = 2, m = 1$, system

$$x' = (1 - t)x + t^2 y + t(1 - t)u,$$

$$y' = -t^2 x + (1 + t)y + t(1 - t)|u|,$$

with $U = E_1, t_1 = 0, t_2 = 1, x(0) = a, y(0) = b$, constraint $\int_0^1 t^2 u^2\, dt \leqq M_1$, where a, b, M_1 are constants, and we assume that a, b, M_1 are so chosen that the system is compatible. Let $I[x, y, u] = x_2 + y_2 = x(1) + y(1)$ be the cost functional. We take here $H(t, u) = t^2 u^2, \alpha = 2$, the 2×2 matrix $A(t)$ has continuous entries, and $C(t, u) = (c_1, c_2)$, with $c_1 = t(1 - t)u$, $c_2 = t(1 - t)|u|$. The sets $Q(t)$ are here $Q(t) = \{(z^1, z^2)|z^1 = t(1 - t)u, z^2 = t(1 - t)|u|\}$, and, for $t \neq 0, 1$, these sets $Q(t)$

$= \{(z^1, z^2)|z^2 = |z^1|, -\infty < z^1 < +\infty\}$ are certainly not convex. We have now, for any $\varepsilon > 0$,

$$|c_1(t,u)| = |c_2(t,u)| \leqq t|u| = 2(2^{-1}\varepsilon^{-1/2})(\varepsilon^{1/2}t|u|)$$

$$\leqq 2^{-2}\varepsilon^{-1} + \varepsilon t^2u^2 = \psi_\varepsilon(t) + \varepsilon H(t,u),$$

where $\psi_\varepsilon(t)$ is the constant function $\psi_\varepsilon = 2^{-2}\varepsilon^{-1}$. Assumption ($\mathrm{m}_1'$) of Remark 2 is satisfied. Assumption (m_1'') is trivial, and Assumption (m_2) is empty here since $\alpha = n$. The initial point $(x(0), y(0)) = (a, b)$ is fixed, and we take $t^* = 0$, $P = \{(\mathrm{a}, b)\}$. Existence Theorem 2.1 can now be applied.

Example 4. Let us consider the Mayer problem with $n = 3$, $m = 1$, system

$$x' = (1-t)x + t^2y + t(1-t)u,$$

$$y' = -t^2x + (1+t)y + t(1-t)|u|,$$

$$z' = tx + (1-t)y + (t^2+1)u^2,$$

with $U = E_1, x(0) = y(0) = z(0) = 0, x^2(1) + y^2(1) \leqq 1, z(1) \leqq 1$, and cost functional $I[x,y,z,u] = 1 - (x^2(1) + y^2(1))^{1/2}$. We take here $\alpha = 2, H(t,u) = c_3 = (t^2+1)u^2$. As seen in Example 3, c_1, c_2 satisfy the growth condition with respect to $H = c_3$. By force of Lemma 2.2, there is some constant M_3 such that $\int_{t_1}^{t_2} H(t, u(t))\,dt \leqq M_3$ for all admissible pairs $x(t), u(t), 0 \leqq t \leqq 1$. We can now apply Theorem 2.1.

Example 5. Let us consider the Lagrange problem with $n = 2, m = 1$, system $x' = f_1 = tx + y, y' = f_2 = x - ty + u$, with $f_0 = (1+t)(1-u^2)^2, t_1 = 0, t_2 = 1$, $x(0) - a, y(0) - b$, given numbers, $x_2 = x(t_2), y_2 = y(t_2)$ both undetermined, and control space $U = E_1$. The sets $A = [0,1] \times E_2, M = A \times U, B = (1, a, b, 1) \times E_2$ are closed. We have here $A_0(t) = 0, A(t) = [t, 1; 1, -t], C_0(t,u) = (1+t)(1-u^2)^2$, $C(t,u) = (c_1, c_2), c_1 = 0, c_2 = u$, and obviously Assumption ($\overline{\mathrm{m}}_1$) of Theorem 3.1 is satisfied. The initial point is fixed. By force of Theorem 3.1, the functional $I = \int_0^1 (1+t)(1-u^2)^2\,dt$ has an absolute minimum under the constraints.

Example 6. Let us consider the Lagrange problem with $n = 2$, $m = 1$, same system as in Example 3, with $x(0) = y(0) = 0, x^2(1) + y^2(1) \leqq 1, U = E_1$, and cost functional

$$I[x,y,u] = \int_0^1 [tx + (1-t)y + (t^2+1)u^2]\,dt.$$

We can apply Existence Theorem 3.1.

5. Existence theorems for free problems. We are concerned here with the problem of the minimum of functionals of the type

$$I[x] = \int_{t_1}^{t_2} f_0(t, x(t), x'(t))\,dt, \tag{5.1}$$

with the only constraints

$$(5.2) \qquad (t, x(t)) \in A, \qquad (t_1, x(t_1), t_2, x(t_2)) \in \mathfrak{B}.$$

We need only consider admissible trajectories $x(t), t_1 \leqq t \leqq t_2$, that is, AC functions $x(t) = (x^1, \cdots, x^n), t \in [t_1, t_2]$, satisfying (5.2), and such that $f_0(t, x(t), x'(t))$ is L-integrable in $[t_1, t_2]$. As is well known, free problems can be thought of as Lagrange problems with $m = n, f = u, U = E_n$.

We assume here that $f_0(t, x, u)$ is of the form

$$(5.3) \qquad f_0(t, x, u) = A_0(t)x + H(t, u),$$

where now A_0 is a $1 \times n$ matrix with entries continuous in $[t_0, T]$, and $H(t, u)$ is a continuous scalar function in $[t_0, T] \times E_n$.

THEOREM 5.1 (An existence theorem for free problems with no convexity conditions). *Let $A = [t_0, T] \times E_n, t_0, T$ finite, let B be closed, let $f_0(t, x, u)$ be of the form* (5.3), *and let us assume that there is a scalar function $\varphi(\zeta), 0 \leqq \zeta < +\infty$, such that $\varphi(\zeta)/\zeta \to +\infty$ as $\zeta \to +\infty$, and $H(t, u) \geqq \varphi(|u|)$ for all $(t, u) \in [t_0, T] \times E_n$. Let P be a compact subset of A, and let Ω be the class of all admissible trajectories x, each of which possesses at least one point $(t^*, x(t^*)) \in P$. Then the functional* (5.1) *has an absolute minimum in Ω.*

This statement is a corollary of Theorem 3.1 and remarks.

Examples. The functions

$$f_0 = x + (1 - u^2)^2, \qquad n = 1,$$

$$f_0 = tx + (1 + t^2)y + (1 - u^2 - v^2)^2, \qquad n = 2,$$

$$f_0 = tx + t^{1/2}x + y + (2 + \cos t)(u - u^2)^2 + v^2, \qquad n = 2,$$

$$f_0 = x^2 + y^2 + (2 + t^2)(1 - u^2 - v^2)^2, \qquad n = 2,$$

satisfy the conditions on f_0 of Theorem 5.1.

Thus, if we take $m = 1$, and we consider the free problem of the minimum of

$$I[x] = \int_0^1 (1 - x'^2(t))^2 \, dt$$

with boundary conditions $x(0) = x(1) = 0$, existence Theorem 5.1 applies, and $I[x]$ attains an absolute minimum. Of course, it is easy to see here that $x_0(t) = 2^{-1} - |t - 2^{-1}|, 0 \leqq t \leqq 1$, is one of the infinitely many optimal solutions and $I[x_0] = 0$. Actually, any AC function $x(t), 0 \leqq t \leqq 1$, with $x(0) = x(1) = 0$, and $|x'(t)| = 1$ a.e. is an optimal solution. Let us consider for comparison's sake the corresponding generalized problem. It concerns the minimum of the functional

$$J[y, p, v] = \int_0^1 [p_1(t)(1 - v_1^2(t))^2 + p_2(t)(1 - v_2^2(t))^2] \, dt$$

with

$$dy/dt = p_1(t)v_1 + p_2(t)v_2(t),$$

$$y(0) = y(1) = 0,$$

where $p = (p_1, p_2)$, $v = (v_1, v_2)$, $0 \leqq p_i \leqq 1$, $i = 1, 2, p_1 + p_2 = 1$, $v \in E_2$. Then $y(t) = 0, p_1(t) = p_2(t) = \frac{1}{2}, v_1(t) = 1, v_2(t) = -1, 0 \leqq t \leqq 1$, is an optimal generalized solution, and $J = 0$. Here y can be approximated as close as we want by (optimal) usual trajectories x satisfying the same boundary data [3].

REFERENCES

[1] 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, pp. 413–429.

[2] ———, *Existence theorems for optimal controls of the Mayer type*, this Journal, 6 (1968), pp. 517–552.

[3] ———, *Convexity of the range of certain integrals*, this Journal, 13 (1975), to appear.

[4] L. W. Neustadt, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

[5] C. Olech, *Lexicographical order, range of integrals, and bang-bang principle*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 35–53.

[6] ———, *Extremal solutions of a control system*, J. Differential Equations, 2 (1966), pp. 74–101.

SIAM J. CONTROL
Vol. 12, No. 2, May 1974

A DIFFERENTIAL GAME OF EVASION WITH NONLINEAR CONTROL*

R. V. GAMKRELIDZE† AND G. L. KHARATISHVILI‡

Abstract. A differential game of evasion in which the state is governed by an nth order differential equation is studied. The equation is linear in the state and its derivatives. Control is exercised through a nonlinear forcing term. A sufficient condition for evasion is given and the evasion strategy is obtained as a generalized solution of a Volterra integral equation. The generalized solution is approximated by ordinary measurable functions by constructing a sliding regime.

1. Introduction. This paper contains the development of results on the theory of linear differential games of evasion (cf. [1]). In the paper we consider a differential game described by an nth order differential equation

$$z^{(n)} + A_1 z^{(n-1)} + \cdots + A_{n-1}z' + A_n z = f(u, v), \tag{1.1}$$
$$z = \begin{pmatrix} z^1 \\ \vdots \\ z^m \end{pmatrix} \in E^m, \qquad f = \begin{pmatrix} f^1 \\ \vdots \\ f^m \end{pmatrix} \in E^m,$$

where the A_i are constant square matrices, and $f(u, v)$ is an arbitrary continuous function of the point $(u, v) \in U \times V$, U and V being compact.

Equation (1.1) differs from that considered in [1] only in that $f(u, v)$ is an arbitrary nonlinear function of the control parameters, whereas in [1] f was assumed to be linear. However, this development now requires a completely new formulation of the sufficient condition for evasion (cf. § 4), rather than in [1], although the estimates obtained are the same. Moreover, in order to determine the strategy of evasion under the formulated sufficient condition, it is necessary to solve, in a certain generalized sense, an integral equation of Volterra type which is not soluble in the ordinary sense. In § 6 we indicate a method of successive approximations by which it is possible to calculate the generalized solution (measure). This generalized solution is then approximated by ordinary measurable functions by the method of constructing a sliding regime (cf. [2]).

2. Statement of the problem. In order to define a game, we must define, in addition to (1.1), a subspace M of E^m of dimension $\leqq m - 2$. The object of the player controlling the parameter $u \in U$ is to transfer the point z to the subspace M, while the object of the player controlling the parameter $v \in V$—and we shall identify ourselves with this player—is to choose the control $v(t)$ in conjunction with the control $u(t)$ so that the corresponding solution $z(t)$ of (1.1) does not meet M anywhere on the infinite time axis $0 \leqq t < \infty$ if $z(0) \notin M$. Thus, to form the value v at an arbitrary instant t, we may use the value u at the instant t or at all preceding times, but not at subsequent times.

* Received by the editors November 23, 1973. Translated into English by A. J. Lohwater, Case Western Reserve University, Cleveland, Ohio, 44106.

† Steklov Mathematical Institute, Academy of Sciences of the USSR, Moscow, USSR.

‡ Control Systems Institute, Academy of Sciences of the Georgian SSR, Tblisi, USSR.

For a precise formulation of the problem it is convenient to introduce the notion of strategy.

We shall define a *strategy* $v_u(t; Z_0)$ to be a mapping defined on the set of arbitrary measurable controls $u(\tau)$, $0 \leqq \tau < \infty$, $u(\tau) \in U$; this mapping depends on the time parameter $t \geqq 0$ and on the vector $Z_0 = (z_0, z_0', \cdots, z_0^{(n-1)})$ of initial conditions of (1.1), and possesses the following properties:

(i) For arbitrary measurable $u(\tau)$, $0 \leqq \tau < \infty$ and for arbitrary fixed Z_0, the mapping $v_u(t; Z_0)$ is measurable as a function of t and assumes values in V;

(ii) If $u_1(\tau)$ and $u_2(\tau)$, $0 \leqq \tau < \infty$, are two controls which coincide almost everywhere on $0 \leqq \tau \leqq \theta$, where θ is arbitrary, then $v_{u_1}(t; Z_0)$ and $v_{u_2}(t; Z_0)$, as functions of t, also coincide almost everywhere on the interval $0 \leqq t \leqq \theta$ for equal values of the parameter Z_0.

The second condition reflects the fact that the strategy $v_u(t; Z_0)$ is defined on $0 \leqq t \leqq \theta$ for given Z_0 only for the values $u(\tau)$, $0 \leqq \tau < \infty$, on the same interval $0 \leqq \tau \leqq \theta$, and does not depend on the values $u(\tau)$ for $\tau > \theta$.

Our problem may now be formulated in the following way. In the "configuration" space E^m of (1.1) we prescribe a subspace M of dimension $\leqq m - 2$. We are required to select a strategy $v_u(t; Z_0)$ such that the solution $z(t)$, $0 \leqq t < \infty$, of the equation

$$z^{(n)} + A_1 z^{(n-1)} + \cdots + A_n z = f(u(t), v(t; Z_0))$$

with the initial condition

$$(z(0), z'(0), \cdots, z^{(n-1)}(0)) = Z_0, \qquad z(0) \notin M,$$

does not intersect the manifold M for any $t \geqq 0$ for every control $u(t)$ and vector Z_0. For this it is still necessary to give a lower bound for the distance from $z(t)$ to M, for arbitrary $t \geqq 0$, as a function of $z(t)$ and its first $n - 1$ derivatives $z'(t), \cdots, z^{(n-1)}(t)$.

The solution of the problem consists in finding a sufficient condition for the existence of such a strategy (§ 4), which we shall call an *evasion strategy*, in determining a procedure allowing us to calculate the evasion strategy as a function of $u(\tau)$, $0 \leqq \tau < \infty$, and Z_0 (§ 6), and in obtaining the indicated bound (§ 5).

3. Use of the convolution symbolism. In order to formulate and prove the basic results, as well as to calculate the evasion strategy, it is convenient to use the convolution technique. Therefore we devote this section to the concepts and notation of the convolution symbolism which we shall find necessary.

Let K be the set of real functions on $0 \leqq t < \infty$ which are integrable on every finite interval. K becomes a commutative ring if we define the sum and difference of two arbitrary functions $x(t)$, $y(t)$ of K in the usual way and if we define their product as the convolution

$$x(t) * y(t) = \int_0^t x(t - \tau) y(\tau) \, d\tau.$$

The convolution of an arbitrary function $x(t)$ of K with the function which is identically equal to unity will be denoted in the sequel by S, and will be equivalent to the integration of $x(t)$:

$$S * x(t) = x(t) * S = \int_0^t x(\tau)\, d\tau .$$

The ring K obviously contains no unit elements, and moreover, no element inverse to S; however, since S is not a divisor of zero (that is, $S * x(t) = 0$ implies $x(t) = 0$), K can be embedded in a ring $\mathcal{M}$ containing a unit element and an element inverse to S. We may take for $\mathcal{M}$ the equivalence classes of all possible pairs (x, S^i), $x \in K$, i a positive integer, if we define in the well-known way the equivalence relation and the addition and multiplication of classes; the equivalence class of the pair (x, S^i) will be written as a fraction x/S^i. The subring in $\mathcal{M}$ consisting of elements of the form $S * x/S$, $x \in K$, will be identified with K, and the unit element in $\mathcal{M}$ will be $\delta = S/S$; the element inverse to S will be

$$D = S/S^2 = S^{-1}; \qquad D * S = S * D = \delta .$$

We shall call the ring $\mathcal{M}$ a *Mikusiński ring*. If $x(t)$ is n times differentiable, then we immediately have the fundamental formula

$$D^i * x(t) = x^{(i)}(t) + D^i * x(0) + D^{i-1} * x'(0) + \cdots + D * x^{(i-1)}(0), \quad 0 \leqq i \leqq n. \tag{3.1}$$

To every number λ corresponds the element

$$\hat{\lambda} = D * \lambda \in \mathcal{M},$$

and, in particular,

$$\hat{1} = D * 1 = D * S = \delta .$$

Numbers (i.e., constant functions) will be called *numerical constants*, and the elements of $\mathcal{M}$ corresponding to them will be called *constants of the ring* $\mathcal{M}$ or simply *constants*. Obviously,

$$\hat{\lambda} * x(t) = \lambda x(t), \qquad \hat{\lambda}_1 * \hat{\lambda}_2 = (\widehat{\lambda_1 \lambda_2}).$$

Every entire function $x(t)$ may be written in the form

$$x(t) = x_1 + \frac{t}{1!}x_2 + \frac{t^2}{2!}x_3 + \cdots = S * \hat{x}_1 + S^2 * \hat{x}_2 + S^3 * \hat{x}_3 + \cdots ,$$

where $x_1, x_2, \cdots$ are numerical constants and $\hat{x}_1, \hat{x}_2, \cdots$ are the corresponding constants of $\mathcal{M}$.

In the formula as written, the powers of the element S are to be clearly understood in the sense of ring multiplication, i.e., the convolution, and the power series in t is to be understood in the usual numerical sense.

We shall call an element of the form

$$x = \hat{x}_0 + S * \hat{x}_1 + S^2 * \hat{x}_2 + \cdots$$

an *entire element of the Mikusiński ring*, where

$$S * \hat{x}_1 + S^2 * \hat{x}_2 + \cdots = x_1 + t x_2 + \frac{t^2}{2!}x_3 + \cdots$$

is an entire numerical function of the numerical argument t.

We shall also consider *entire square matrices* of arbitrary order over $\mathscr{M}$, that is, square matrices whose elements are entire elements of $\mathscr{M}$. Every matrix of this type can clearly be written in the form

$$A = A(S) = \hat{A}_0 + S * \hat{A}_1 + S^2 * \hat{A}_2 + \cdots,$$

where the $\hat{A}_i$ are *constant square matrices* (whose elements are constants and which are in one-to-one correspondence with the numerical matrices A_i), and where the series

$$S * \hat{A}_1 + S^2 * \hat{A}_2 + \cdots = A_1 + tA_2 + \frac{t^2}{2!}A_3 + \cdots$$

is an entire matrix function of the variable t in the ordinary numerical sense.

The determinant of every entire matrix $A(S)$, calculated in the ordinary formal way using ring multiplication, is an entire element in $\mathscr{M}$ since the sum and product of two entire elements from $\mathscr{M}$ or of entire matrices is again an entire element or an entire matrix; we shall denote this determinant by $\overset{*}{\det} A(S)$.

Let $\hat{I}$ be the unit matrix, that is, a diagonal matrix of the corresponding dimensionality with the element δ along the diagonal. We shall show that an arbitrary entire matrix of the form

$$A(S) = \hat{I} + S * \hat{A}_1 + S^2 * \hat{A}_2 + \cdots$$

has the inverse entire matrix $A^{-1}(S)$:

$$A^{-1}(S) * A(S) = A(S) * A^{-1}(S) = \hat{I},$$

where

$$A^{-1}(S) = I + C(S) + C^2(S) + \cdots$$

and

$$C(S) = -(S * \hat{A}_1 + S^2 * \hat{A}_2 + \cdots).$$

The series

$$C(S) + C^2(S) + \cdots,$$

where the powers are clearly understood as being formed in the sense of the ring, will be considered as a series of entire matrix functions C^i of the variable t, where t will be allowed to assume complex values. We shall show that this series converges uniformly in an arbitrary disc Q_T of radius T with center at the origin in the complex t-plane, from which the assertion will then follow from the well-known theorem of Weierstrass. We denote by m_T the maximum of the numerical function

$$\|S * \hat{A}_1\| + \|S^2 * \hat{A}_2\| + \cdots$$

on the disc Q_T; here $\|\cdot\|$ is a certain matrix norm. Then, for $t \in Q_T$, we have

$$\begin{aligned}
&\|C(S)\| \leqq m_T,\\
&\|C^2(S)\| \leqq \int_0^t m_T^2 \, d\tau = m_T^2 t \leqq m_T^2 T,\\
&\quad\vdots\\
&\|C^i(S)\| \leqq \int_0^t \|C^{i-1}\| \, \|C\| \, d\tau \leqq m_T^i \frac{t^{i-1}}{(i-1)!} \leqq m_T^i \frac{T^{i-1}}{(i-1)!},\\
&\quad\vdots
\end{aligned}$$

and consequently, for $t \in Q_T$,

$$\|C(S) + C^2(S) + \cdots\| \leqq m_T\left(1 + Tm_T + \frac{(Tm_T)^2}{2!} + \cdots + \frac{(Tm_T)^i}{i!} + \cdots\right).$$

The bound which we obtained shows the uniform convergence of the series $C + C^2 + \cdots$ on the arbitrary disc Q_T. According to the Weierstrass theorem cited, the derivatives with respect to t of the entire function $C + C^2 + \cdots$ are equal to the series of the corresponding derivatives of $C, C^2, \cdots$, hence, in calculating the coefficients in the power series expansion of $C + C^2 + \cdots$, we must combine the coefficients for equal powers in the power series expansions of $C, C^2, \cdots$.

We shall now apply the results we have obtained to finding the solution of a linear differential equation with constant coefficients.

Let us rewrite (1.1) in the form

$$z^{(n)} + \hat{A}_1 * z^{(n-1)} + \cdots + \hat{A}_n * z = f(u, v),$$

and express the derivatives of z by means of (3.1); then we obtain

$$\begin{aligned}
&(D^n + \hat{A}_1 * D^{n-1} + \cdots + \hat{A}_{n-1} * D + \hat{A}_n) * z \\
&\quad = (D^n + \hat{A}_1 * D^{n-1} + \cdots + \hat{A}_{n-1} * D) * z_0 \\
&\qquad + (D^{n-1} + \hat{A}_1 * D^{n-2} + \cdots + \hat{A}_{n-2} * D) * z_0' \\
&\qquad + \cdots + (D^2 + \hat{A}_1 * D) * z_0^{(n-2)} + D * z_0^{(n-1)} + f(u, v).
\end{aligned}$$

If we multiply both sides of the equation by S^n and keep in mind that S is not a divisor of zero in $\mathscr{M}$, we obtain

$$\begin{aligned}
&(\hat{I} + S * \hat{A}_1 + \cdots + S^{n-1} * \hat{A}_{n-1} + S^n * \hat{A}_n) * z \\
&\quad = (\hat{I} + S * \hat{A}_1 + \cdots + S^{n-1} * \hat{A}_{n-1}) * z_0 \\
&\qquad + S * (\hat{I} + S * \hat{A}_1 + \cdots + S^{n-2} * \hat{A}_{n-2}) * z_0' \\
&\qquad + \cdots + S^{n-2} * (\hat{I} + S * \hat{A}_1) * z_0^{(n-2)} + S^{n-1} * z_0^{(n-1)} + S^n \\
&\qquad + S^n * f(u, v).
\end{aligned}$$

If we multiply both sides of this equation by the inverse matrix of $\hat{I} + S * \hat{A}_1 + S^2 * \hat{A}_2 + \cdots + S^n * \hat{A}_n$, the *resolvent of* (1.1), i.e.,

$$R(S) = I + C(S) + C^2(S) + \cdots,$$
$$C(S) = -(S * \hat{A}_1 + \cdots + S^n * \hat{A}_n),$$

and if we carry out the simple transformations, we obtain the formula for the solution z of (1.1):

$$\begin{aligned}
z &= z_0 + S * z_0' + \cdots + S^{n-1} * z_0^{(n-1)} \\
&\quad + S^n * (\varphi_0 * z_0 + \cdots + \varphi_{n-1} * z_0^{(n-1)}) + S^n * R(S) * f(u, v),
\end{aligned} \tag{3.2}$$

where $\varphi_0, \cdots, \varphi_{n-1}$ are certain entire matrices over $\mathscr{M}$ which may easily be expressed in terms of the $\hat{A}_i$, $i = 1, \cdots, n$, but whose form will not have any special significance in what follows.

As a conclusion to this section we shall show that every entire $k \times k$ matrix $A(S)$ *with nonvanishing determinant* has a representation of the form

$$A(S) = H^{(1)} * \Delta(l_1, \cdots, l_k) * H^{(2)}, \tag{3.3}$$

where

$$H^{(i)} = \hat{H}_0^{(i)} + S * \hat{H}_1^{(i)} + S^2 * \hat{H}_2^{(i)} + \cdots, \qquad i = 1, 2,$$

are entire invertible matrices, i.e., $\overset{*}{\det} \hat{H}_0^{(i)} \neq 0$ and $\Delta(l_1, \cdots, l_k)$ is the diagonal matrix

$$\begin{pmatrix} S^{l_1} & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ & & & & S^{l_k} \end{pmatrix}.$$

If we assume in this connection that $l_1 \leqq \cdots \leqq l_k$, then $l_1, \cdots, l_k$ forms a sequence of entire nonnegative numbers independent of the representation; we shall call these numbers the *indices* of the matrix $A(S)$. In particular, the *index of an entire element of the ring* $\mathcal{M}$ is that index of the first nonvanishing coefficient in the expansion into (nonnegative) powers of S.

For the proof we introduce the usual concept of *elementary transformations* of an entire matrix $A(S)$, where we mean by this the interchange of two rows (or columns) of $A(S)$, the addition to an arbitrary row (or column) any other row (or column) multiplied beforehand by an arbitrary entire element of $\mathcal{M}$, and finally the multiplication of a row (or column) by an arbitrary entire invertible element of $\mathcal{M}$, that is, by an element of the form

$$\hat{x}_0 + S * \hat{x}_1 + \cdots, \qquad \hat{x}_0 \neq 0.$$

We denote by m_r the least of the indices of the mth order minors of the determinant. It is easy to see that under the elementary operations on the matrices, the numbers m_r remain invariant. From this and from the fact that every elementary operation on rows (columns) is equivalent to multiplication on the left (on the right) by an entire invertible matrix, we easily obtain in the usual way the desired representation and the invariance of the indices for the matrix $A(S)$.

4. Formulation of the sufficient condition. Suppose that we are given the series

$$g_0(u, v) + S * g_1(u, v) + \cdots, \tag{4.1}$$

where $g_i(u, v) \in E^k$ are continuous functions of the point $(u, v) \in U \times V$ and have an entire majorant

$$\hat{\lambda}_0 + S * \hat{\lambda}_1 + S^2 * \hat{\lambda}_2 + \cdots, \qquad |g_i(u, v)| \leqq \lambda_i$$

for all $(u, v) \in U \times V$. By a *representation* of the original series we shall mean any description of the series in the form

$$g_0 + S * g_1 + \cdots = A(S) * (\xi_0(u, v) + S * \xi_1(u, v) + \cdots) + \alpha(t), \tag{4.2}$$

where the right-hand side satisfies the following conditions:

(i) $A(S) = \hat{A}_0 + S * \hat{A}_1 + \cdots$ is an entire matrix over the ring $\mathscr{M}$, and $\overset{*}{\det} A(S) \neq 0$.

(ii) The coefficients $\xi_i(u, v) \in E^k$ of the series $\xi_0 + S * \xi_1 + \cdots$ are continuous functions of the point $(u, v) \in U \times V$, and the series has an entire majorant in the sense indicated above.

(iii) The function $\alpha(t)$ does not depend on u, v. (It follows from the first two conditions that $\alpha(t)$ is entire.)

We denote by $[\xi_0(u, v)]$ the smallest linear manifold of E^k containing all the points $\xi_0(u, v)$, $(u, v) \in U \times V$. A representation will be called *canonical* if, in addition to the three conditions enumerated, it satisfies still a fourth condition:

(iv) Among all representation of the series, the linear manifold $[\xi_0(u, v)] \subset E^k$ has the largest possible dimension.

It is clear that the canonical representation is not completely defined uniquely by the original series.

Let $\operatorname{co}_v \xi_0(u, v)$ denote the convex hull of the set of all points $\xi_0(u, v)$, $v \in V$, for fixed $u \in U$. We give the following basic definition.

We shall say that the *parameter* v in (4.1) has complete *maneuverability* if there exists a representation (4.2) of this series such that the set

$$\bigcap_{u \in U} \operatorname{co}_v \xi_0(u, v) \subset E^k \tag{4.3}$$

contains interior points.

Every representation satisfying this condition is obviously canonical. It is not hard to show that if some canonical representation satisfies condition (4.3), any other canonical representation also satisfies it.

For convenience in formulating the sufficient condition of evasion, we shall adopt the following convention. By an *m-dimensional space* E^m we shall mean, in this paper, the arithmetic space of m-columns with the scalar square of the vector equal to the sum of the squares of its coordinates. Let L be an arbitrary subspace of E^m of dimension $k \leqq m$, and let P be the orthogonal projection of E^m onto L, that is, P is a symmetric matrix satisfying the equation $P^2 = P$ and the condition $PE^m = L$. Let B denote an orthogonal matrix satisfying the condition: $Z \in L$ if and only if the vector Bz has zero coordinates starting from the $(k + 1)$-st. Then the matrix BP gives an orthogonal projection onto L followed by an orthogonal transformation of coordinates in E^m in such a way that the image BPz of an arbitrary point $z \in E^m$ has zero coordinates in the last $m - k$ places. Consequently, every subspace $L \subset E^m$ defines, up to an orthogonal transformation, a linear mapping $\pi : E^m \to E^k$ corresponding to an orthogonal projection of E^m onto L. The fact that π is defined up to multiplication by an orthogonal matrix will not have any significance for us, since all of our assertions concerning π will be invariant under orthogonal transformations. We shall call the $k \times m$ matrix π the *projection of E^m onto E^k corresponding to the orthogonal projection of E^m onto L*, or simply a projection corresponding to the subspace L.

We return now to the game and consider the expression $R(S) * f(u, v)$ in the formula for the solution (3.2). Since the resolvent $R(S)$ is an entire $m \times m$ matrix over $\mathscr{M}$, i.e., $R(S) = \hat{I} + S * \hat{R}_1 + \cdots$, the expression

$$\begin{aligned} R(S) * f(u, v) &= f(u, v) + S * \hat{R}_1 * f(u, v) + \cdots \\ &= f_0(u, v) + S * f_1(u, v) + \cdots \end{aligned}$$

is a power series in S with an entire majorant:

$$\hat{\mu}_0 + S * \hat{\mu}_1 + \cdots; \qquad |f_i(u, v)| \leqq \hat{\mu}_i,$$

for all $(u, v) \in U \times V$ and $i = 0, 1, \cdots$. Moreover, if π is a projection of E^m onto E^k corresponding to some subspace $L \subset E^m$, then the expression

$$\begin{aligned} \pi(R(S) * f(u, v)) &= \hat{\pi} * R(S) * f(u, v) \\ &= \hat{\pi} * f_0(u, v) + S * \hat{\pi} * f_1(u, v) + \cdots \end{aligned}$$

is a series of the type (4.1), so that we may speak of representations of this series in the form (4.2), and in particular, we may speak of canonical representations.

THEOREM 4.1 (Sufficient condition for evasion). *Let L be a subspace of dimension $k \geqq 2$ lying in the orthogonal complement of $M \subset E^m$, and let π be a projection of E^m onto E^k corresponding to the subspace L. If the parameter v in the expression $\hat{\pi} * R(S) * f(u, v)$ has complete maneuverability, then there exists an evasion strategy.*

Thus an evasion strategy exists if it is possible to find a canonical representation

$$\hat{\pi} * R(S) * f(u, v) = H(S) * (\psi_0(u, v) + S * \psi_1(u, v) + \cdots) + x(t) \tag{4.4}$$

such that the set

$$\bigcap_{u \in U} \operatorname{co}_v \psi_0(u, v) \subset E^k \tag{4.5}$$

contains interior points.

The proof of this theorem is contained in the following two sections, where we construct the evasion strategy on the basis of the representation (4.4) satisfying (4.5) and where we give an estimate for the distance between $z(t)$ and M.

From what has been said, it is clear that if we know in advance that the parameter v in $\hat{\pi} * R * f(u, v)$ possesses complete maneuverability, then it is important for us to find the canonical representation of this expression. It is possible to describe a procedure for finding a canonical representation for an arbitrary prescribed series (4.1) in a finite number of steps. It would be desirable to have effective criteria by means of which it would be possible to infer, on the basis of the expression $R * f$, the existence of a projection π such that v possesses complete maneuverability in the expression $\hat{\pi} * R * f$, and, if such a projection exists, how to calculate it.

5. Derivation of the estimates. We assume the hypotheses and notation of Theorem 4.1, and we assume for convenience that the origin in the space E^k is an interior point of the set (4.5). We denote by Q the closed k-dimensional cube with center at the origin and with sides parallel to the axes, and contained in the interior of the set $\bigcap_{u \in U} \operatorname{co}_v \psi_0(u, v)$; the diameter of the cube can be arbitrarily small.

We denote by $\rho(z(t), M)$ the distance of the point $z(t)$ to the subspace M; we denote by P the orthogonal projection of E^m onto the orthogonal complement of M. Clearly,

$$\rho(z(t), M) = |Pz(t)| \geqq |\hat{\pi} * z(t)|.$$

If we take advantage of (3.2) and (4.4), we may write

$$\hat{\pi} * z(t) = \varphi(t; Z_0) + S^n * H(S) * [\psi_0(u, v) + S * \psi_1(u, v) + \cdots],$$

where

$$\varphi(t; Z_0) = \hat{\pi} * [z_0 + S * z_0' + \cdots + S^{n-1} * z_0^{(n-1)} + S^n * (\varphi_0 * z_0 + \cdots + \varphi_{n-1} * z_0^{(n-1)})] + \eta(t)$$

is an entire function of t which depends linearly on the coordinates of the vector Z_0 of initial conditions. Moreover, let the matrix $H(S)$ be represented in the form (see (3.3))

$$H(S) = H^{(1)}(S) * \Delta(l_1, \cdots, l_k) * H^{(2)}(S), \tag{5.1}$$

where $H^{(1)}(S)$ and $H^{(2)}(S)$ are entire invertible matrices and where $l_1 \leqq \cdots \leqq l_k = l$.

In the following section we shall show that for sufficiently small diameter of Q there exists a fixed interval $0 \leqq t \leqq T$ on which the integral equation in v

$$S^n * H(S) * [\psi_0(u, v) + S * \psi_1(u, v) + \cdots] = -t^{n+l}\zeta \tag{5.2}$$

may always be solved to any degree of precision for any point $\zeta \in Q$ and any measurable control $u(t) \in U$. In other words, for any $\varepsilon > 0$ there exists a measurable function $v(t) \in V$, $0 \leqq t \leqq T$, such that

$$|S^n * H(S) * (\psi_0(u, v) + S * \psi_1(u, v) + \cdots) + t^{n+l}\zeta| \leqq \varepsilon \tag{5.3}$$

for $0 \leqq t \leqq T$ and for arbitrary preassigned $u(t) \in U$ and $\zeta \in Q$. Thus, to calculate the solution $v(t)$ at the instant t, we use the point ζ and the values $u(\tau)$ on the interval $0 \leqq \tau \leqq t$.

For every $u(t) \in U$, $\zeta \in Q$ and $\varepsilon > 0$ we choose one of the possible functions $v = w(t; u, \zeta, \varepsilon)$ satisfying (5.3) and we call it an *evasive maneuver* corresponding to the control $u(t)$, the point $\zeta \in Q$ and the prescribed degree of accuracy $\varepsilon > 0$.

We now give two simple preliminary estimates.

Let $z(t)$ be a solution of (1.1) *with initial condition $Z_0 = (z_0, z_0', \cdots, z_0^{(n-1)})$, and denote by $|Z_0|$ the quantity $|z_0| + \cdots + |z_0^{(n-1)}|$. If $\rho(z_0, M) > 0$, there exists a λ so large that*

$$\rho(z(t), M) \geqq \frac{\rho(z_0, M)}{2} \quad \textit{for } 0 \leqq t \leqq \frac{\rho(z_0, M)}{\lambda(1 + |Z_0|)}. \tag{5.4}$$

The proof follows immediately if we remark that for sufficiently large λ, we have the bound

$$|Pz'(t)| \leqq \frac{\lambda}{2}(1 + |Z_0|), \qquad 0 \leqq t \leqq T.$$

We have

$$\begin{aligned}\rho(z(t), M) = |Pz(t)| &= \left| Pz_0 + \int_0^t Pz'(\tau)\, d\tau \right| \\ &\geqq \rho(z_0, M) - \left| \int_0^t Pz'\, d\tau \right| \geqq \rho(z_0, M) - \int_0^t \frac{\lambda}{2}(1 + |Z_0|)\, d\tau \\ &\geqq \rho(z_0, M) - \frac{\rho(z_0, M)}{2} = \frac{\rho(z_0, M)}{2}.\end{aligned}$$

The second estimate concludes with the following assertion.

If T is sufficiently small, then there exists a v such that, for an arbitrary vector Z_0 of initial conditions,

$$v(1 + |Z(t)|) \geqq 1 + |Z_0|, \qquad 0 \leqq t \leqq T, \tag{5.5}$$

where

$$Z(t) = (z(t), z'(t), \cdots, z^{(n-1)}(t)),$$
$$|Z(t)| = |z(t)| + |z'(t)| + \cdots + |z^{(n-1)}(t)|.$$

Proof. If we differentiate successively the left and right sides of the formula (3.2) for the solution, we obtain for $0 \leqq t \leqq T$ and for a sufficiently large constant χ,

$$|z(t)| \geqq |z_0| - t\chi(1 + |Z_0|),$$
$$|z'(t)| \geqq |z_0'| - t\chi(1 + |Z_0|),$$
$$|z^{(n-1)}(t)| \geqq |z_0^{(n-1)}| - t\chi(1 + |Z_0|),$$

whence

$$1 + |Z(t)| \geqq (1 - tn\chi)(1 + |Z_0|) \geqq (1 - Tn\chi)(1 + |Z_0|).$$

We obtain the inequality $1 - Tn\chi > 0$.

Finally, we formulate the following lemma (cf. [1]), a simple proof of which we give at the end of the section.

LEMMA 5.1. *There exists a $\theta > 0$ so small that for an arbitrary initial vector Z_0 there exists a point $\zeta(Z_0) \in Q$ satisfying the condition*

$$|\varphi(t; Z_0) - t^{n+l}\zeta(Z_0)| \geqq \theta t^{n+l}, \qquad 0 \leqq t \leqq T.$$

Now it is possible to proceed to the description of the conduct of the game itself and to the construction of the evasion strategy. This conduct of the game is identical to that described in [1].

We choose a positive number $\sigma > 0$ satisfying the following inequalities

$$\sigma < \tfrac{1}{2}\theta T^{n+l}, \quad \sigma < \lambda T, \quad \sigma/2 > \theta(\sigma/\lambda)^{n+l}. \tag{5.6}$$

The last inequality is compatible with the first two, since λ can be made arbitrarily large.

At the beginning of the game at the instant $t = 0$, let the distance between $z(0)$ and M be greater than σ. Then we assign to the control v an arbitrary value and we pay attention only to the distance between $z(t)$ and M. As soon as this distance becomes equal to σ, let us say for $t = t_1$, we define the control $v(t)$ on the interval $t_1 \leqq t \leqq t_1 + T$ by means of the "evasive maneuver" formula

$$v(t) = w(t - t_1; u, \zeta(Z(t_1)), \varepsilon),$$

where we now give an upper bound for ε, and the point $\zeta(Z(t_1)) \in Q$ satisfies the condition (Lemma 5.1)

$$|\varphi(t - t_1; Z(t_1)) - (t - t_1)^{n+l}\zeta(Z(t_1))| \geqq \theta(t - t_1)^{n+l}, \qquad t_1 \leqq t \leqq t_1 + T. \tag{5.7}$$

As a result of this "maneuver" $v(t)$, $t_1 \leqq t \leqq t_1 + T$, we have the following estimates:

$$\rho(z(t), M) \geqq \theta\left(\frac{\sigma}{\lambda}\right)^{n+l} \frac{1}{(1 + |Z(t_1)|)^{n+l}} - \varepsilon, \qquad t_1 \leqq t \leqq t_1 + T,$$

$$\rho(z(t_1 + T), M) \geqq \theta T^{n+l} - \varepsilon.$$

Indeed, for

$$0 \leqq t - t_1 \leqq \frac{\rho(z(t_1), M)}{\lambda(1 + |Z(t_1)|)} = \frac{\sigma}{\lambda(1 + |Z(t_1)|)},$$

we have, on the basis of (5.4) and (5.6),

$$\rho(z(t), M) \geqq \frac{\sigma}{2} \geqq \theta\left(\frac{\sigma}{\lambda}\right)^{n+l} \geqq \theta\left(\frac{\sigma}{\lambda}\right)^{n+l} \frac{1}{(1 + |Z(t_1)|)^{n+l}} - \varepsilon.$$

Moreover, for

$$\frac{\sigma}{\lambda(1 + |Z(t_1)|)} \leqq t - t_1 \leqq T,$$

we have, on the basis of (5.3) and (5.7),

$$\begin{aligned} \rho(z(t), M) &\geqq |\hat{\pi} * z(t)| = |\varphi(t - t_1 ; Z(t_1)) + S^n * H(S) * (\psi_0 + S * \psi_1 + \cdots)| \\ &\geqq |\varphi(t - t_1 ; Z(t_1)) - (t - t_1)^{n+l}\zeta(Z(t_1))| - \varepsilon \\ &\geqq \theta(t - t_1)^{n+l} - \varepsilon \geqq \theta\left(\frac{\sigma}{\lambda}\right)^{n+l} \frac{1}{(1 + |Z(t_1)|)^{n+l}} - \varepsilon; \end{aligned}$$

$$\rho(z(t_1 + T), M) \geqq \theta T^{n+l} - \varepsilon.$$

If we choose

$$\varepsilon \leqq \min\left(\tfrac{1}{2}\theta\left(\frac{\sigma}{\lambda}\right)^{n+l} \frac{1}{(1 + |Z(t_1)|)^{n+l}}, \tfrac{1}{2}\theta T^{n+l}\right),$$

we obtain, for $t_1 \leqq t \leqq t_1 + T$,

$$\rho(z(t), M) \geqq \tfrac{1}{2}\theta\left(\frac{\sigma}{\lambda}\right)^{n+l} \frac{1}{(1 + |Z(t_1)|)^{n+l}},$$

$$\rho(z(t_1 + T), M) \geqq \tfrac{1}{2}\,\theta T^{n+l} > \sigma.$$

Finally, if we take advantage of (5.5), we obtain the final estimates

$$\begin{aligned} \rho(z(t), M) &\geqq \tfrac{1}{2}\theta\left(\frac{\sigma}{\lambda\nu}\right)^{n+l} \frac{1}{(1 + |Z(t)|)^{n+l}} \\ &= \frac{\text{const.}}{(1 + |Z(t)|)^{n+l}}, \qquad t_1 \leqq t \leqq t_1 + T, \qquad (5.8) \\ \rho(z(t_1 + T), M) &> \sigma. \end{aligned}$$

Thus, in the course of the entire "maneuver", the first estimate is effective with a constant in the numerator which is independent of the method of conducting the game, but is determined only by the game itself, that is, by (1.1), by the sets U and V, and by the subspace M. At the end of the maneuver, the point z is outside the σ-neighborhood of the subspace M, so that we are in the same position that we were in at the beginning of the game at the instant $t = 0$, so that we can repeat the process described an infinite number of times. Since T is fixed, we can continue the game as long as we please, and, moreover, obtain a universal bound (5.8) for the distance $\rho(z(t), M)$. From the construction of the evasion control $v(t)$ described, it also follows that it is an evasion strategy in the sense defined in § 2 since the value of the solution at the instant t is expressed in terms of the initial condition and the values of the control $u(\tau)$ on the interval $0 \leqq \tau \leqq t$.

If $z(0)$ is inside a σ-neighborhood of M at the beginning of play, then if we conduct the game in the way described, we obtain the earlier estimate (5.8) for $t \geqq T$, and on the interval $0 \leqq t \leqq T$, we have

$$\rho(z(t), M) \geqq \tfrac{1}{2}\theta\left(\frac{\rho(z(0), M)}{\lambda v}\right)^{n+l} \frac{1}{(1 + |Z(t)|)^{n+l}}.$$

Proof of Lemma 5.1. It is enough to prove the lemma for the case $k = 2$: $\varphi = \begin{pmatrix}\varphi_1\\ \varphi_2\end{pmatrix}$. The derivatives

$$\left(\frac{\varphi^i}{t^{n+l}}\right)' = \frac{t^{n+l}(\varphi^i)' - (n + l)t^{n+l-1}\varphi^i}{t^{2(n+l)}}, \qquad i = 1, 2,$$

have a finite number of zeros on the interval $0 \leqq t \leqq T$, and these do not depend on Z_0 since the numerators consist of entire functions of t which depend linearly on the coordinates of Z_0. Consequently, for each Z_0 the segment $0 \leqq t \leqq T$ can be partitioned into N intervals (N does not depend on Z_0 but in general depends on T), on each of which $\varphi^i(t; Z_0)$, $i = 1, 2$, is monotone, and hence in the Nth order subdivision of Q into squares there exists a square into whose interior the curve φ/t^{n+l}, $0 \leqq t \leqq T$, does not pass. If we take $\zeta(Z_0)$ as the center of this square, then we have

$$\left|\frac{\varphi}{t^{n+l}} - \zeta(Z_0)\right| \geqq \frac{|Q|}{4^N},$$

where $|Q|$ is the length of the side of the square Q.

6. Solution of the integral equation. Since the diameter of the cube Q can be decreased arbitrarily, we can solve, instead of (5.2), the equation

$$S^n * H(S) * (\psi_0(u, v) + S * \psi_1(u, v) + \cdots) + S^{n+l} * \zeta = 0, \qquad \zeta \in Q.$$

We write the left-hand side in the form

$$S^{n-1} * H(S) * \{S * [(\psi_0 + S * \psi_1 + \cdots) + H^{-1} * S^l * \zeta]\},$$

where the expression

$$\omega(t) = -H^{-1}(S) * S^l * \zeta, \qquad \zeta \in Q,$$

is an entire function of t by the representation (5.1) and by the inequalities $l = l_k \geqq \cdots \geqq l_1$. For $0 \leqq t \leqq T$, the values of $\omega(t)$ lie in some closed k-dimensional sphere Q_1 of arbitrarily small diameter (for a sufficiently small cube Q) with center at the origin. Consequently, it is possible to assume that

$$Q_1 \subset \left[\bigcap_{u\in U} \operatorname{co}_v \psi_0(u, v)\right]^0, \tag{6.1}$$

where $[\,\cdot\,]^0$ denotes the interior of the set contained in the brackets. Therefore it is clear that instead of (5.3), it is sufficient to solve the inequality

$$|S * [\psi_0(u, v) + S * \psi_1(u, v) + \cdots) - \omega(t)]| < \varepsilon \tag{6.2}$$

for $0 \leqq t \leqq T$ and for arbitrary, prescribed, measurable $u(t) \in U$ and $\omega(t) \in Q_1$.

It is easy to see that the equation

$$S * [(\psi_0(u, v) + S * \psi_1(u, v) + \cdots) - \omega(t)] = 0,$$

or, equivalently, the equation

$$\psi_0(u(t), v) + S * \psi_1(u(t), v) + \cdots = \omega(t), \qquad u(t) \in U, \quad \omega(t) \in Q_1, \tag{6.3}$$

cannot be solved for v, since the set of points $\psi_0(u, v)$, $u \in U$, $v \in V$, generally lies far from the sphere Q_1 and hence, for sufficiently small ε,

$$|\psi_0(u, v) - \omega(t)| > |S * \psi_1 + \cdots|.$$

Moreover, it is clear from the same considerations that (6.3) cannot be solved approximately with the desired precision in the L_1-norm, that is, we cannot have

$$\int_0^T |(\psi_0(u(t), v) + S * \psi_1(u(t), v) + \cdots) - \omega(t)|\, dt \leqq \varepsilon$$

for sufficiently small ε. However, if we introduce a weaker norm $\|\cdot\|_w$ (in the sense of the induced topology) by the formula

$$\|p(t)\|_w = \sup_{t\in[0,T]} \left|\int_0^t p(\tau)\, d\tau\right|,$$

then, by an approximate solution of (6.3) to within ε in this norm we mean a solution on $0 \leqq t \leqq T$ of the inequality

$$\|(\psi_0(u(t), v) + S * \psi_1(u(t), v) + \cdots) - \omega(t)\|_w \leqq \varepsilon$$

for arbitrary prescribed measurable $u(t) \in U$, $\omega(t) \in Q_1$, that is, a solution of the inequality (6.2) to which we now turn our attention.

For every fixed value $u \in U$ there exists, by (6.1), a finite number of points $v_0, \cdots, v_r$ in V such that the convex hull of the points $\psi_0(u, v_0), \cdots, \psi_0(u, v_r)$ contains the sphere Q_1 in its interior. A simple argument using the compactness of U and the continuity of $\psi_0(u, v)$ shows that for every $u \in U$ it is possible to choose values $v_i(u) \in V$, $i = 0, 1, \cdots, r$, where r is independent of $u \in U$, such that the functions $v_i(u)$ are Borel measurable. If we set

$$p_0(u) = \psi_0(u, v_0(u)), \cdots, p_r(u) = \psi_0(u, v_r(u)),$$

then

$$\left[\bigcap_{u\in U} \operatorname{co}(p_0(u), \cdots, p_r(u))\right]^0 \supset Q_1. \tag{6.4}$$

We now take advantage of an approximation lemma by which we shall construct an optimal sliding regime (cf. [2]), and which we formulate in the following way.

APPROXIMATION LEMMA. *Let $q_0(t), \cdots, q_r(t), 0 \leqq t \leqq T$, be bounded measurable k-dimensional vector functions and $\mu_0(t), \cdots, \mu_r(t)$, $0 \leqq t \leqq T$, measurable scalar functions satisfying the conditions*

$$\sum_{i=0}^{r} \mu_i(t) \equiv 1, \qquad \mu_i(t) \geqq 0, \qquad i = 0, 1, \cdots, r. \tag{6.5}$$

Then, for arbitrary $\varepsilon > 0$, there exists a function $q_\mu(t)$, $0 \leqq t \leqq T$, depending on $\mu = \mu(t) = (\mu_0(t), \cdots, \mu_r(t))$, which assumes at each time instant t one of the values $q_0(t), \cdots, q_n(t)$ and which satisfies for all $t \in [0, T]$ the inequality

$$\left|\int_0^t \left[\sum_{i=0}^{r} \mu_i(\tau)q_i(\tau) - q_\mu(\tau)\right] d\tau\right| \leqq \varepsilon,$$

that is,

$$\left\|\sum_{i=0}^{r} \mu_i(t)q_i(t) - q_\mu(t)\right\|_w \leqq \varepsilon.$$

To construct $q_\mu(t)$, the interval $[0, T]$ must be partitioned into sufficiently fine subintervals I_α (the mesh depends on $q_i(t)$ but not on $\mu_i(t)$), and each of the intervals I_α must be subdivided into $r + 1$ mutually disjoint subintervals $I_{\alpha i}$, $i = 0, 1, \cdots, r$, of length

$$\text{meas.}\, I_{\alpha i} = \int_{I_\alpha} \mu_i(t)\, dt, \qquad i = 0, 1, \cdots, r.$$

The function $q_\mu(t)$ can then be defined by the relations

$$q_\mu(t) = q_i(t), \qquad t \in I_{\alpha i}, \qquad i = 0, 1, \cdots, r.$$

We consider now, instead of the insoluble equation (6.3), the following equation in the unknown functions $\mu_0(t), \mu_1(t), \cdots, \mu_r(t), 0 \leqq t \leqq T$:

$$\begin{gathered}\sum_{i=0}^{r} [\psi_0(u(t), v_i(u(t)))\mu_i + S * \psi_1(u(t), v_i(u(t)))\mu_i + \cdots] = \omega(t), \\ u(t) \in U, \quad \omega(t) \in Q_1,\end{gathered} \tag{6.6}$$

and we shall show that this equation may be solved on a sufficiently small interval $0 \leqq t \leqq T$, and that the solution $\mu(t) = (\mu_0(t), \cdots, \mu_r(t))$ will, moreover, satisfy (6.5). The solution can be thought of as a function of t, namely, a positive unit measure on the compact set V at the instant t, concentrated at the $r + 1$ points

$$v_0(t) = v_0(u(t)), \cdots, v_r(t) = v_r(u(t)),$$

and the sum

$$\sum_{i=0}^{r} \psi_j(u(t), v_i(t))\mu_i(t), \qquad j = 0, 1, 2, \cdots,$$

may be thought of as an average of the functions $\psi_j(u(t), v)$ of v with respect to the measure $\mu(t) = \mu_t$:

$$\sum_{i=0}^{r} \psi_j(u(t), v_i(t))\mu_i(t) = \int_V \psi_j(u(t), v)\, d\mu_t(v).$$

We shall call the measure $\mu(t)$ a *generalized solution of the integral equation* (6.3), which is insoluble in the ordinary sense.

Suppose that we are given a generalized solution $\mu(t) = (\mu_0(t), \cdots, \mu_r(t))$ and a number $\varepsilon > 0$. We choose N large enough so that, for $0 \leqq t \leqq T$,

$$(6.7) \qquad |S^{N+1} * \psi_{N+1} + S^{N+2} * \psi_{N+2} + \cdots | \leqq \varepsilon/3$$

independently of whatever controls we substitute in the function $\psi_{N+j}(u, v)$ in place of the parameters u and v. This is possible since the series $\psi_0 + S * \psi_1 + \cdots$ has a majorant. According to the approximation lemma, the interval $0 \leqq t \leqq T$ can be partitioned into such fine segments I_α that if each of the I_α is further subdivided into $r + 1$ subintervals $I_{\alpha i}$, $i = 0, 1, \cdots, r$, of length meas. $I_{\alpha i} = \int_{I_\alpha} \mu_i(t)\, dt$ and if the control $v(t)$ is defined by the formula

$$v(t) = v_i(t) = v_i(u(t)), \qquad t \in I_{\alpha i}, \quad i = 0, 1, \cdots, r,$$

then the inequality

$$\left\| \sum_{i=0}^{r} [\psi_0(u(t), v_i(t))\mu_i(t) + \cdots + S^N * \psi_N(u(t), v_i(t))\mu_i(t)] \right.$$
$$\left. - [\psi_0(u(t), v(t)) + \cdots + S^N * \psi_N(u(t), v(t))] \right\|_w \leqq \varepsilon/3$$

is satisfied. Consequently, if we combine this last inequality with (6.1), we obtain the estimate

$$\|[\psi_0(u(t), v(t)) + S * \psi_1(u(t), v(t)) + \cdots] - \omega(t)\|_w$$
$$= \left\| \sum_{i=0}^{r} [\psi_0(u(t), v_i(t))\mu_i(t) + S * \psi_1(u(t), v_i(t))\mu_i(t) + \cdots] \right.$$
$$\left. - [\psi_0(u(t), v(t)) + S * \psi_1(u(t), v(t)) + \cdots] \right\|_w \leqq \varepsilon,$$

which expresses the fact that (6.2) has a solution. We shall say that this *control $v(t)$ approximates the generalized solution $\mu(t)$ to within ε in the norm $\| \cdot \|_w$.*

Before we proceed to the construction of the generalized solution, we shall prove the following preparatory lemma.

LEMMA 6.1. *In the notation $p_i(t) = p_i(u(t)) = \psi_0(u(t), v_i(t))$, $i = 0, 1, \cdots, r$, let the closed sphere Q_2 contain Q_1 in its interior and be contained in*

$$[\bigcap_{t \in [0,T]} \operatorname{co}(p_0(t), \cdots, p_n(t))]^0$$

(cf. (6.4)). *There exist* $r+1$ *measurable scalar functions* $m_0(t,q),\cdots,m_r(t,q)$ *of the point* $(t,q)\in[0,T]\times Q_2$ *such that*

$$\text{(i) } m_i(t,q)\geqq 0,\qquad i=0,1,\cdots,r,$$

$$\sum_{i=0}^{r} m_i(t,q)\equiv 1,\quad \sum_{i=0}^{r} m_i(t,q)p_i(t)=q.$$

$$\text{(ii) } \sum_{i=0}^{r}|m_i(t,q')-m_i(t,q'')|\leqq C|q'-q''|,$$

where C *is independent of* $t\in[0,T]$ *and of* $q',q''\in Q_2$.

In other words, there is a certain regular method of making each point $q\in Q_2$ correspond to its barycentric coordinates with respect to the points $p_0(t),\cdots,p_r(t)$ (the convex hull of which contains Q_2 in its interior)—a fact expressed by condition (i) and by the measurability of the $m_i(t,q)$; condition (ii) asserts that these coordinates not only depend continuously on q, but also that they satisfy a Lipschitz condition in q with a constant C which is independent of the time.

Proof. We consider the fixed r-dimensional simplex $\Sigma=\operatorname{co}(\sigma_0,\cdots,\sigma_r)$ and make the affine mapping

$$P_t:\Sigma\to\operatorname{co}(p_0(t),\cdots,p_r(t))$$

by the equalities $P_i(\sigma_i)=p_i(t), i=0,1,\cdots,r$.

The preimage $P_t^{-1}(q)$ of an arbitrary point $q\in Q_2$ is a closed convex set in Σ. We denote by $\sigma(t,q)$ the center of gravity of a uniform distribution of unit mass on the entire set $P_t^{-1}(q)$. The barycentric coordinates $m_i(t,q)$, $i=0,1,\cdots,r$, of the point $\sigma(t,q)$ in the simplex Σ,

$$\sigma(t,q)=\sum_{i=0}^{r}m_i(t,q)\sigma_i, \tag{6.8}$$

satisfy all the required conditions. The first condition is automatically satisfied by the construction; to prove the second condition it is necessary to make a simple estimate of a multiple integral, and the measurability of the functions $m_i(t,q)$ is also easy to prove.

We proceed now to the construction of a solution of (6.6),

$$\mu(t)=(\mu_0(t),\cdots,\mu_r(t)),\qquad 0\leqq t\leqq T,$$

which satisfies (6.5). Let us rewrite (6.6) in the form

$$\sum_{i=0}^{r}p_i(t)\mu_i=\omega(t)-\sum_{j=1}^{\infty}S^j*\sum_{i=0}^{r}\Omega_{ji}(t)\mu_i,\qquad \omega(t)\in Q_1,$$

where

$$\Omega_{ji}(t)=\psi_j(u(t),v_i(t)),\quad j=1,2,\cdots;\quad i=0,1,\cdots,r.$$

We shall solve the equation by successive approximations, and as a zero-order approximation,

$$\mu^{(0)}(t)=(\mu_0^{(0)}(t),\cdots,\mu_r^{(0)}(t)),$$

we take the functions (cf. (6.8))

$$\mu_i^{(0)}(t)=m_i(t,\omega(t)),\qquad i=0,1,\cdots,r.$$

If we have constructed the $(l-1)$-st approximation,

$$\mu_{(t)}^{(l-1)} = (\mu_0^{(l-1)}(t), \cdots, \mu_r^{(l-1)}(t)),$$

the lth approximation is defined as follows. The point

$$q^{(l)}(t) = \omega(t) - \sum_{j=1}^{\infty} S^j * \sum_{i=0}^{r} \Omega_{ji}(t)\mu_i^{(l-1)}(t), \qquad 0 \leqq t \leqq T,$$

is contained in Q_2 for sufficiently small T since $\omega(t) \in Q_1 \subset [Q_2]^0$, and the series

$$\sum_{j=1}^{\infty} S^j * \sum_{i=0}^{r} |\Omega_{ji}(t)|$$

has the majorant

$$\sum_{j=1}^{\infty} S^j * \hat{\lambda}_j, \ \sum_{i=0}^{r} |\Omega_{ji}(t)| \leqq \lambda_j,$$

which is independent of $t \in [0, T]$. Hence, by Lemma 6.1, we can define the functions

$$\mu_i^{(l)}(t) = m_i(t, q^{(l)}(t)), \qquad i = 0, 1, \cdots, r,$$

which constitute the lth approximation $\mu^{(l)}(t) = (\mu_0^{(l)}(t), \cdots, \mu_r^{(l)}(t))$. We have

$$q^{(l)}(t) = \sum_{i=0}^{r} p_i(t)\mu_i^{(l)}(t) = \omega(t) - \sum_{j=1}^{\infty} S^j * \sum_{i=0}^{r} \Omega_{ji}(t)\mu_i^{(l-1)}(t), \quad l = 0, 1, 2, \cdots.$$

If we make use of the Lipschitz character of the functions $m_i(t, q^{(l)})$ in $q^{(l)}$, property (ii) in Lemma 6.1, we obtain

$$\begin{aligned}\sum_{i=0}^{r} |\mu_i^{(l)}(t) - \mu_i^{(l-1)}(t)| &\leqq C|q^{(l)}(t) - q^{(l-1)}(t)| \\ &\leqq C\left(\sum_{j=1}^{\infty} S^j * \sum_{i=0}^{r} \lambda_j|\mu_i^{(l-1)}(t) - \mu_i^{(l-2)}(t)|\right) \\ &= C\left(\sum_{j=1}^{\infty} S^j * \hat{\lambda}_j\right) * \sum_{i=0}^{r} |\mu_i^{(l-1)}(t) - \mu_i^{(l-2)}(t)| \\ &\leqq C_1\left(S * \sum_{i=0}^{r} |\mu_i^{(l-1)}(t) - \mu_i^{(l-2)}(t)|\right).\end{aligned}$$

This yields the final estimate

$$\sum_{i=0}^{r} |\mu_i^{(l)}(t) - \mu_i^{(l-1)}(t)| \leqq \frac{C_2^{l-1}}{(l-1)!}, \qquad 0 \leqq t \leqq T,$$

which guarantees the uniform convergence of the successive approximations

$$\mu^{(l)}(t) = (\mu_0^{(l)}(t), \cdots, \mu_r^{(l)}(t))$$

to the desired solution $\mu(t) = (\mu_0(t), \cdots, \mu_r(t))$ as $l \to \infty$.

We must call attention to the nonunique character of the generalized solution $\mu(t)$, which arises from the nonuniqueness of the barycentric coordinates of the point $q \in \text{co}\,(p_0, \cdots, p_r)$ relative to the points $p_0, \cdots, p_r$.

REFERENCES

[1] L. S. PONTRYAGIN, *A linear differential game of escape*, Trudy Mat. Inst. Steklov., 62 (1971), pp. 30–63.

[2] R. V. GAMKRELIDZE, *On some extremal problems in the theory of differential equations*, this journal, 3 (1965), pp. 106–128.

SIAM J. Control
Vol. 12, No. 3, August 1974

AN EXISTENCE THEOREM ON OPTIMAL CONTROL OF PARTIALLY OBSERVABLE DIFFUSIONS*

N. U. AHMED AND K. L. TEO†

Abstract. In this paper we consider the problem of the existence of optimal control systems described by the stochastic Ito differential equation. It is shown (Theorem 1) that Fleming's existence theorem [1, Thm. 3, p. 205] remains valid without the assumption that the drift coefficient f of the system is linear in the control variable. Further, it is shown that the control restraint set U can be taken as variable. Our result is based on the Fillippov technique (Himmelberg et al. [2, Thm. 3', p. 281]) rather than the lower semicontinuity arguments as used by Fleming [1, 1968, Appendix 3, p. 213]. However, our result does not contain his.

1. Introduction. Let us consider the system described by the family of stochastic Ito differential equations S:

$$\begin{aligned}&d\xi(t) = f(t, \xi(t), u(t, \hat{\xi}(t)))\,dt + g(t, \xi(t))\,dw(t),\\ &t \in [0, T], \quad 0 < T < \infty,\\ &\xi(t_0) = \xi_0 \quad (\pi_0\text{—initial probability measure}),\\ &u \in D,\end{aligned}$$

where for each $t \in [0, T]$,

$$\xi(t) = (\xi_1(t), \cdots, \xi_n(t)) \in R^n$$

is the dynamic state of the system,

$$\hat{\xi}(t) = (\xi_1(t), \cdots, \xi_l(t)) \in R^l, \qquad 0 \leqq l \leqq n,$$

is (without loss of generality) the first l components of this vector that the controller can observe;

$$\begin{aligned}&f: [0, T] \times R^n \times R^p \to R^n;\\ &u: [0, T] \times R^l \to R^p;\\ &g: [0, T] \times R^n \to n \times m \text{ matrices};\end{aligned}$$

$\{w(t),\ t \in [0, T]\}$ is the m-dimensional Wiener process independent of ξ_0; and D, to be defined later, is the set of admissible controls.

Let B be an open set in R^n (with compact closure) supporting the initial probability measure π_0, and let the boundary ∂B of B satisfy the following property: each point of ∂B has a neighborhood in which ∂B is locally representable by functions with Hölder continuous second order partial derivatives.

Given the data $\xi(s) = x$ with $x \in B$, $s \in (0, T)$, let us stop the process ξ at the first time $s \leqq \tau < T$ when $\xi(\tau) \in \partial B$; if $\xi(t) \in B$ for all $t \in [s, T)$, then we set $\tau = T$. The random variable τ is called the first passage time from the cylinder $Q \triangleq (0, T) \times B$.

* Received by the editors January 25, 1973, and in revised form May 7, 1973.

† Department of Electrical Engineering, University of Ottawa, Ottawa, Ontario, K1N 6N5, Canada.

Let us define the set of admissible controls on $\hat{Q}$ by

$$D \triangleq \{u : u \text{ measurable on } \hat{Q},\ u(t, \hat{x}) \in U(t, \hat{x}), (t, \hat{x}) \in \hat{Q}\},$$

where $\hat{Q}$ is the projection onto $(t, \hat{x})$ – space of the cylinder Q and $U(t, \hat{x})$ is a measurable set-valued function on $\hat{Q}$ with values that are nonempty compact convex subsets of a fixed compact set $\tilde{U} \subset R^p$. With the above preparations, we may state our problem precisely. Subject to the dynamic constraints S, find a control $u \in D$ that minimizes the cost functional

$$(1.1) \qquad J(u) = \mathscr{E}\left\{\int_0^\tau L(t, \xi(t), u(t, \hat{\xi}(t)))\, dt\right\},$$

where L is a real-valued continuous function defined on $\bar{Q} \times \tilde{U}$ and $\bar{Q}$ is the closure of the set Q. For convenience of further reference, this problem is called P.

In fact, the problem P can be reduced to the optimal control problem of an equivalent distributed parameter system with controls appearing in the coefficients of the differential operator. For this reduced problem, Fleming [1, 1968, Thm. 2, p. 203] presented interesting results on the necessary conditions of optimality (Pontryagin-type) under sufficiently general conditions. However, as far as the existence theorem [1, 1968, Thm. 3, p. 205] is concerned, the conditions imposed on the drift coefficient f were rather strong. In this paper it is shown that these conditions can be substantially relaxed at the expense of others. In fact, Fleming's existence theorem was based on the following assumptions:

(i) $a(t, x) \triangleq \frac{1}{2}(g \cdot g')(t, x)$ is continuous and bounded on $\bar{Q}$, where denotes matrix transpose;

(ii) there exists a number $c > 0$ such that

$$\sum_{i,j=1}^{n} a_{ij}(t, x) \cdot z_i z_j \geqq c|z|^2 \qquad \text{for all } z \in R^n$$

uniformly on $\bar{Q}$ (uniformly parabolic);

$$\text{(iii)} \quad \frac{|a(t, x) - a(t_1, x_1)|}{|t - t_1| + |x - x_1|} \leqq M,$$

where $t, t_1 \in [0, T]$, $x, x_1 \in \bar{B}$ and M is a constant;

(iv) f is a bounded measurable function on $\bar{Q}$ and is *linear in u on U* with U a *fixed* compact and convex subset of R^p;

(v) L is a real-valued continuous function defined on $\bar{Q} \times U$ and convex in u on U.

Our result is proved under the assumptions (i)–(iii) and condition A: f and L are bounded measurable on $\bar{Q}$ for each $u \in \tilde{U}$ and continuous on $\tilde{U}$ for all $(t, x) \in \bar{Q}$. Further, the set-valued function $\Gamma(t, x), (t, x) \in Q$, defined by $\Gamma(t, x) \triangleq \{\tilde{f}(t, x, u) : u \in U(t, \hat{x})\}$, is convex for each $(t, x) \in Q$, where $\tilde{f} \triangleq \begin{pmatrix} L \\ f \end{pmatrix}$ is the $(n + 1)$-vector constructed by adjoining L to the n-vector f, and $\hat{x}$ is the projection of x.

It is clear that condition A allows *nonlinearity* for the drift coefficient f.

2. Existence of optimal controls. For convenience, we denote by A^u the following differential operator:

$$A^u\psi \triangleq \left\{ \sum_{i=1}^{n} f_i(t, x, u(t, \hat{x})) \cdot \frac{\partial\psi}{\partial x_i} + \sum_{i,j=1}^{n} a_{ij}(t, x)\frac{\partial^2\psi}{\partial x_i \cdot \partial x_j} \right\},$$

where $u \in D$; $a_{ij}(t, x) \triangleq \frac{1}{2}(g \cdot g')_{ij}(t, x), i, j = 1, \cdots, n$; and f is *not* necessarily *linear* in u.

The proof of the existence of optimal controls for the problem P under the new conditions is based on several lemmas. These results, as presented below, are utilized to prove our main result (Theorem 1).

LEMMA 1. *Suppose that a satisfies conditions* (i), (ii) *and* (iii) *given in* § 1 *and that* f^u ($f^u(t, x) \triangleq f(t, x, u(t, \hat{x}))$) *is bounded measurable on* $\bar{Q}$ *for each* $u \in D$. *Then the system S has a unique solution* ξ^u *which is a strong Markov process.*

Proof. The proof follows from Corollary 3.2 and Theorem 6.2 of Stroock and Varadhan [3, pp. 366–367, 392].

Using Ito's lemma [3, Thm. 2.5, p. 352], it is not difficult to show that the stochastic optimal control problem reduces to the optimal control problem of an equivalent distributed parameter system as stated in the following lemma.

LEMMA 2. *The problem* P *reduces to an equivalent problem* P′ *which consists of the first boundary value problem S′ and the cost functional* (2.1):

$$S' \begin{cases} -\dfrac{\partial\phi^u(t, x)}{\partial t} = A^u\phi^u(t, x) + L(t, x, u), \quad (t, x) \in [0, T) \times B, \\ \phi^u(t, x)|_{\partial B} = 0 \quad \textit{for } t \in (0, T], \\ \phi^u(T, x) = 0 \quad \textit{for } x \in B, \\ u \in D, \end{cases}$$

$$\min_{u \in D} J(u) = \min_{u \in D} \int_B \phi^u(0, x)\pi_0(dx). \tag{2.1}$$

In the sequel, we need the following.

LEMMA 3. *Let* $\bar{f}(t, \hat{x}, \hat{\hat{x}}, v)(x \triangleq (\hat{x}, \hat{\hat{x}}))$ *be a measurable* $(n + 1)$*-vector-valued function of* $(t, \hat{x}, \hat{\hat{x}})$ *on* Q *for each* $v \in \tilde{U}$ *and a continuous function of* v *on* $\tilde{U}$ *for each* $(t, \hat{x}, \hat{\hat{x}}) \in Q$. *Then, if* $r(t, x) \in \Gamma(t, x)$ *for all* $(t, x) \in Q$ *is a measurable function, there exists a measurable function* v *of* $(t, \hat{x})$ *on* $\hat{Q}$ *with values in* $U(t, \hat{x})$ *such that* $r(t, x) = \bar{f}(t, x, v(t, \hat{x}))$ *for all* $(t, x) \in Q$.

Proof. Let $V(t, x) \triangleq \{(u \cdot \theta)(t, x)(\triangleq v(t, x)) : u$ measurable and $u(t, \hat{x}) \in U(t, \hat{x})\}$, where θ denotes the projection of (t, x)-space onto $(t, \hat{x})$-space and $(u \cdot \theta)(t, x) = u(\theta(t, x)) = u(t, \hat{x})$. Clearly the set-valued function V on Q is equivalent to the set-valued function U on $\hat{Q}$. Therefore the proof follows from the application of this projection and Theorem 3′ of [2, p. 281].

Based on the above lemmas, we have our main result.

THEOREM 1. *Consider the system S. Suppose that a satisfies assumptions* (i), (ii) *and* (iii) (§ 1) *and that f and L satisfy condition* A. *Then, there exists a control* $u^* \in D$ *that minimizes the cost functional J.*

Proof. In view of Lemma 2, it is sufficient to prove the existence of an optimal control for the problem P'. For each $u \in D$, the system S' has a unique solution, ϕ^u satisfying the following properties [1, Appendix 1, pp. 209–210]:

$$\text{(I)} \quad \int_Q \left[(\phi^u)^2 + \left(\frac{\partial \phi^u}{\partial t}\right)^2 + \sum_{i=1}^{n} \left(\frac{\partial \phi^u}{\partial x_i}\right)^2 + \sum_{i,j=1}^{n} \left(\frac{\partial^2 \phi^u}{\partial x_i \cdot \partial x_j}\right)^2 \right] dt\, dx \leqq N_1,$$

and

(II) for $0 < \delta < 1$ and $\gamma^u = \phi^u$ or $\partial \phi^u / \partial x$,

$$|\gamma^u(t, x)| + \frac{|\gamma^u(t', x') - \gamma^u(t, x)|}{[|t' - t| + |x' - x|^2]^{\delta/2}} \leqq N_2,$$

where N_1 and N_2 are constants independent of the choice of $u \in D$.

Define the set X by $X \triangleq \{\phi^u : u \in D\}$ and the functional I on $C(\bar{Q})$ by $I(\phi) = \int_B \phi(0, x)\, d\pi_0(x)$. Clearly, the functional I is a continuous linear functional on $C(\bar{Q})$. Therefore it is sufficient to prove that X is a compact subset of $C(\bar{Q})$, since this will guarantee the existence of a minimum for the functional h on X and this in turn would imply the existence of an optimal control $u^* \in D$. For this, let $\{\psi^n : n = 1, 2, \cdots\}$ be any sequence from X. Clearly, there exists a sequence $\{u_n : n = 1, 2, \cdots\} \subset D$ such that ψ^n is the unique solution of the first boundary value problem S' corresponding to the control u_n.

Let $\{\bar{f}^n(t, x)\} \triangleq \{\bar{f}(t, x, u_n(t, \hat{x}))\}$ and consider the set $\mathcal{N}$ defined by $\mathcal{N} \triangleq \{y : y$ measurable on Q and $y(t, x) \in \Gamma(t, x)$ for all $(t, x) \in Q\}$. Since by condition A, $\Gamma(t, x)$ is a compact convex set-valued function on Q and $\sup_{(t,x)\in Q} |y(t, x)| \leqq \beta < \infty$ for all $y \in \mathcal{N}$, it follows that $\mathcal{N}$ is a weak* compact (weak* closed) subset of $L_\infty(Q)$. Therefore, one can select a subsequence $\{\bar{f}^{n_s}\} \subset \{\bar{f}^n\}$ so that $\bar{f}^{n_s} \to y^0$ in the weak* topology with the limit y^0 belonging to $\mathcal{N}$. Thus, it follows from Lemma 3 that there exists a control $u_0 \in D$ so that $y^0(t, x) = \bar{f}(t, x, u_0(t, \hat{x}))$ for all $(t, x) \in Q$. Clearly, corresponding to the sequence $\{\bar{f}^{n_s}\}$ there exists a sequence $\{u_{n_s}\} \subset \{u_n\} \subset D$ which in turn generates a sequence $\{\psi^{n_s}\}$ contained in $\{\psi^n\}$. Then it follows from [1, Appendix 1, the last paragraph of p. 210] that $\psi^{n_s} \to \psi^0$ uniformly on $\bar{Q}$, where ψ^0 is the unique solution of the system S' corresponding to the coefficient $\bar{f} = y^0$ and consequently to the control $u_0 \in D$ obtained before. Thus $\psi^0 \in X$ and this implies that X is a compact subset of $C(\bar{Q})$. This completes the proof.

Remarks. In case L is independent of the control variable u, the convexity condition of Γ (condition A) reduces to that of the set $f(t, x, U(t, \hat{x})), (t, x) \in Q$. Under this situation our result contains Fleming's as a special case.

Conclusion. In this paper, we have presented an existence theorem for optimal controls for a class of stochastic systems in which the drift coefficient is not necessarily linear in the control variable and the control restraint set U is not necessarily fixed as assumed in Fleming [1, Thm. 3, p. 205]. However, it should be mentioned that our result does not contain his as shown by the example

$$\Gamma(t, x) = \left\{ \begin{pmatrix} u^2 \\ u \end{pmatrix} : |u| \leqq 1 \right\}$$

which fails to satisfy the convexity condition as assumed in our paper. This point was brought to our attention by one of the reviewers.

REFERENCES

[1] W. H. Fleming, *Optimal control of partially observable diffusions*, this Journal, 6 (1968), pp. 194–214.

[2] C. J. Himmelberg, M. Q. Jacobs and F. S. Van Vleck, *Measurable multifunctions, selectors, and Fillippov's implicit functions lemma*, J. Math. Anal. Appl., 25 (1969), pp. 276–284.

[3] D. W. Stroock and S. R. S. Varadhan, *Diffusion processes with continuous coefficients I*, Comm. Pure Appl. Math., 22 (1969), pp. 345–400.

SIAM J. Control
Vol. 12, No. 3, August 1974

NECESSITY AND EXISTENCE RESULTS ON CONSTRAINED OPTIMIZATION OF SEPARABLE FUNCTIONALS BY A MULTIPLIER RULE*

DANIEL H. WAGNER AND LAWRENCE D. STONE†

Abstract. Functionals E (extended real-valued) and C (vector-valued) are defined by $E(q) = \int_X e(x, q(x))\,d\mu x$ and $C(q) = \int_X c(x, q(x))\,d\mu x$, in a rather abstract setting, without differentiability assumptions. If, among other assumptions, either μ is a nonatomic measure or convexity conditions hold, then from Blackwell's generalization of Lyapunov's convexity theorem and a separating hyperplane theorem, it follows that for q^* to maximize E subject to an equality (inequality) constraint on C, it is necessary that there exist a vector λ (nonnegative vector λ) such that q^* maximizes $E - \lambda \cdot C$. For the latter to hold, by von Neumann's selection theorem under the principal condition that e and c are Borel functions, it is necessary that $q^*(x)$ maximize $e(x, \cdot) - \lambda \cdot c(x, \cdot)$ for a.e. $x \in X$. This much extends methods and results of Aumann and Perles. Existence results are derived from "upper closure" of the range of (C, E), under boundedness assumptions, as the range of a vector integral over selections from a set-valued function. This development utilizes and extends results and methods of Olech and Blackwell. The main result asserts existence of optimal functions under the principal conditions that the set-valued function is upper closed, e and c are Borel functions, C is bounded, and either μ is nonatomic or E is bounded above. Examples show that the results cannot be strengthened in various ways.

1. Introduction.

This paper is addressed to functions q^* which maximize a real-valued "effectiveness" functional E subject to, e.g., a closed convex set constraint on a k-vector-valued "cost" functional C. Both E and C are separable (sum of point-functions): $E(q) = \int_X e(x, q(x))\,d\mu x$ and $C(q) = \int_X c(x, q(x))\,d\mu x$ when $q(x) \in Y(x)$ for $x \in X$, with fixed X, Y, μ, e, and c. We give general conditions under which such optimal q^* exist. We further show under weak conditions that it is necessary that for some vector λ, such optimal q^* must maximize the Lagrangian $E - \lambda \cdot C$ (a functional multiplier rule), and that for this it is necessary that $q^*(x)$ maximize the Lagrangian $e(x, \cdot) - \lambda \cdot c(x, \cdot)$ for a.e. $x \in X$ (a pointwise multiplier rule). In § 2, the formal framework is given and the relevance of the nonlinear forms of E and C is noted.

We make no differentiability assumptions, although in some cases convexity-concavity assumptions on $c(x, \cdot)$, $e(x, \cdot)$ occur as alternative hypotheses, implying one-sided differentiability. In contrast to related papers, we admit infinite values of E and C, Theorem 2.2 being a key tool for this purpose, and we impose relatively few continuity or boundedness conditions.

We use rather abstract measure-theoretic and topological assumptions, since this adds little difficulty. This has the advantages of identifying with greater precision the properties that are needed to insure the conclusions of the theorems and of enabling application to broader categories of situations, as compared, for example, to using Euclidean space with Lebesgue measure or counting measure (discrete summation). Such more familiar structures are related to some of the

* Received by the editors February 15, 1972, and in revised form April 24, 1973. This work was supported in part by the Naval Analysis Programs, Office of Naval Research, under Contract N00014-69-C-0435.

† Daniel H. Wagner, Associates, Paoli, Pennsylvania 19301.

relatively abstract hypotheses by comments in the introduction to § 5 and Remarks 2.3, 3.4 and 6.4.

Section 6 deals with existence, usually assuming bounded C and σ-finite μ. Theorem 6.3 (ii) and from Theorem 6.10 to the end of the paper constitute new results. The main result is Theorem 6.13 (iii); we use most of the paper in its proof. The principal new features of our existence results are in permitting ∞ to be attained, constraint sets to be rather general, and for $x \in X$, $Z(x) = \{(c(x, y), e(x, y)) : y \in Y(x)\} \subset \mathscr{E}_{k+1}$ to be "upper closed" instead of closed. In circumstances of interest (see Remark 6.4), upper closure permits $e(x, \cdot)$ to be upper semicontinuous instead of continuous. With $Z(x)$ closed for $x \in X$, no topology on X is needed (Theorem 6.9), but with the change to upper closure, Borel conditioning must be imposed (see Remark 6.11). Theorem 6.13 then asserts that the range of (C, E) is upper closed (which insures existence) if E is bounded above, and that existence holds if *either* μ is nonatomic or E is bounded above; Remark 6.14 shows that existence may fail if the latter disjunction is violated. Theorems 6.3 (i), 6.6, and 6.9 depend on $Z(x)$ being closed for $x \in X$ and are adaptations, with modest generalization, of work of Blackwell [5] and Olech [31], except that 6.9 depends on 6.3 (ii). A more systematic summary of § 6 is given in its introduction.

Statements of necessity of the functional multiplier rule for optimality are given in § 3, without topologizing X. Necessity of the pointwise rule is given in § 5. The main necessity result is Corollary 5.2; all of the prior theorems are used in its proof. In particular, the von Neumann selection result, Theorem 4.1, is needed, requiring Borel conditions similar to those of Theorem 6.13. That necessity of the pointwise rule may fail if e is not a Borel function but is Lebesgue measurable is shown in Remark 5.5. Theorem 5.3 gives necessity of the pointwise rule under countable X and convexity conditions.

The development leading to Corollary 5.2 is largely comprised of a Lyapunov-type convexity theorem, a separating hyperplane argument, and von Neumann's selection theorem; this pattern of proofs has also been previously used by Aumann and Perles [3] to show necessity of a pointwise multiplier rule for optimality, but under more restrictive conditions (see Remark 5.6). Our necessity results are thus largely adaptations of prior work, in more general settings. Remark 5.5 is new, the inclusion of infinite integrability appears to be new, and the distinction between pointwise and functional rules is at least not usual. The generality of our necessity statements is, moreover, needed in § 6 and in [37] and [38].

Theorems 3.1 and 6.6 (ii) extend Blackwell's generalizations [5] to vector integrals of Lyapunov's theorem [23] on the convexity and compactness of the range of a vector measure, a history of which is given in [39]. For related results see, for example, [1], [2], [4], [14]–[16], [27]–[32], and [34].

In a succeeding article [42], we prove necessity of the pointwise rule and existence under a "coverability" condition pertaining to the concave envelope of $e(x, \cdot)$, under $c(x, y) = y$, $y \in Y(x)$, $x \in X$. This avoids Borel assumptions and, in existence statements, replaces the upper closure condition with an alternative. It is interesting that coverability has a property similar to the property of Borel functions that composition with a measurable function results in a measurable function (see [42, Lemma 4 (iv)]), and that either coverability or Borel assumptions yield necessity of the pointwise rule.

Our results on functionals also apply to optimization of separable set-functions, by identifying sets with indicator functions. In this way, one can obtain the necessity and existence results of Dantzig and Wald [9] on the Neyman–Pearson lemma, removing the finite integrability restriction from their necessity statements.

2. Preliminaries. Let ω be the set of positive integers and $\emptyset$ be the empty set. For $n \in \omega$, let $\mathscr{E}_n$ be Euclidean n-space and $0^n = (0, \cdots, 0) \in \mathscr{E}_n$. If $a, b \in \mathscr{E}_n$, we denote their inner product by $a \cdot b$; this is extended to vectors with $\pm\infty$ components in the obvious way, being undefined if $0 \cdot \infty$ or $\infty - \infty$ occurs. If A is a subset of extended $\mathscr{E}_n$, we denote **fin** $A = A \cap \mathscr{E}_n$ (the finite part of A). A subscript on a vector or a vector-valued function will always denote component. For $B \subset \mathscr{E}_n$, we denote the closure, interior, and boundary of B by **cl** B, **int** B, and **bdy** B. The range of a function f is denoted **rng** f, u.s.c. means upper semicontinuous, and for a, b in extended $\mathscr{E}_n$, $a \leqq b$ means $a_i \leqq b_i$ for $i = 1, \cdots, n$. Let $\mathscr{E}_n^+ = \mathscr{E}_n \cap \{a : a \geqq 0^n\}$.

We fix arbitrary nonvacuous sets X and $Y(x)$ for each $x \in X$. Defining

$$\Omega = \{(x, y) : x \in X \text{ and } y \in Y(x)\},$$

we fix $k \in \omega$ and real-valued functions $c_1, \cdots, c_k$ and e on Ω. We denote $c = (c_1, \cdots, c_k)$ and $(c, e) = (c_1, \cdots, c_k, e)$, and we regard $\mathscr{E}_{k+1} = \mathscr{E}_k \times \mathscr{E}_1$.

We fix a measure μ over X. *Measurability always refers to* μ. *Integrability* of $f : X \to \mathscr{E}_n$ means that $\int_X f_i \, d\mu$ exists (possibly $\pm\infty$) for $i = 1, \cdots, n$. A measurable set S is an *atom* of μ if $\mu(S) > 0$ and S is not a union of disjoint measurable sets having positive measure. We say μ is *nonatomic* if it has no atoms, is *purely atomic* if X is a union of atoms, or has *finite substance* if for each measurable $P \subset X$ for which $\mu(P) > 0$, there exists a measurable $Q \subset P$ such that $0 < \mu(Q) < \infty$ (e.g., if μ is σ-finite). Subsets of X having μ measure zero are ignored, i.e., if $P \subset X$ and $Q \subset X$, then "for $x \in P$" means "for μ a.e. $x \in P$", and "$P = Q$" means "$\mu(P - Q) = \mu(Q - P) = 0$."

For $x \in X$, we denote by $c(x, \cdot)$ the function mapping y into $c(x, y)$ for $y \in Y(x)$; if $q(x) \in Y(x)$ for $x \in X$, $c(\cdot, q(\cdot))$ is the function mapping x into $c(x, q(x))$ for $x \in X$, and similarly for e. We define

$$\Psi = \{q : q \text{ is a function on } X \text{ and } q(x) \in Y(x) \text{ for } x \in X\},$$

$$\Xi = \Psi \cap \{q : c(\cdot, q(\cdot)) \text{ and } e(\cdot, q(\cdot)) \text{ are measurable functions}\},$$

$$\Phi = \Xi \cap \{q : c(\cdot, q(\cdot)) \text{ and } e(\cdot, q(\cdot)) \text{ are integrable}\},$$

$$C_i(q) = \int_X c_i(x, q(x)) \, d\mu x \quad \text{for } i = 1, \cdots, k, \quad q \in \Phi,$$

$$E(q) = \int_X e(x, q(x)) \, d\mu x \quad \text{for } q \in \Phi.$$

Denoting $C = (C_1, \cdots, C_k)$ and $(C, E) = (C_1, \cdots, C_k, E)$, we say $q^* \in \Phi$ is *optimal* if[1]

$$E(q^*) = \max \{E(q) : C(q) = C(q^*)\}$$

[1] In this and similar usage, it is understood that $E(p) \in \{E(q) : C(q) = C(q^*)\}$ implies that $E(p)$ exists.

and we say q^* is *strongly optimal* if

$$E(q^*) = \max\{E(q): C(q) \leqq C(q^*)\}.$$

Since q^* appears on both sides of the formulas defining optimality, these definitions do not refer to a preassigned cost; optimality (strong optimality) of q^* merely means that there is no $q \in \Phi$ such that $C(q) = C(q^*)(C(q) \leqq C(q^*))$ and $E(q) > E(q^*)$. In statements on existence of optimal functions, we shall, however, preassign cost by requiring $C(q) \in N$, with closed N.

An equality constraint may be considered to be an inequality constraint by noting that for $a, b \in \mathscr{E}_n$, $a = b$ iff $(a \leqq b$ and $-a \leqq -b)$.

Suppose $\lambda \in \mathscr{E}_k$. We define the *pointwise Lagrangian*, l_λ, by

$$l_\lambda(x, y) = e(x, y) - \lambda \cdot c(x, y) \quad \text{for } (x, y) \in \Omega,$$

and the *functional Lagrangian*, L_λ, by

$$L_\lambda(q) = E(q) - \sum_{\lambda_i \neq 0} \lambda_i C_i(q) \quad \text{for } q \in \Phi \text{ such that } \infty - \infty \text{ does not occur}.$$

Note that if $L_\lambda(q)$ exists, then $L_\lambda(q) = \int_X l_\lambda(x, q(x))\, d\mu x$. Suppose $q^* \in \Phi$. We say that (q^*, λ) *satisfies the functional multiplier rule* if $L_\lambda(q^*)$ exists and

$$(2.1) \qquad L_\lambda(q^*) \geqq L_\lambda(q) \quad \text{whenever } L_\lambda(q) \text{ exists}.$$

We say that (q^*, λ) *satisfies the pointwise multiplier rule* if

$$(2.2) \qquad l_\lambda(x, q^*(x)) \geqq l_\lambda(x, y) \quad \text{for } y \in Y(x), \quad x \in X.$$

If either rule is satisfied with $\lambda \in \mathscr{E}_k^+$, we say the rule is *strongly* satisfied.

Traditionally, Lagrange multiplier rules are stated in terms of finding a stationary point of a Lagrangian, requiring differentiability assumptions in contrast to the formulations above. The usefulness of the present viewpoint, maximizing a Lagrangian without differentiability assumptions, was pointed up by Everett [10]. Earlier statements of multiplier rules in this vein were given by Slater [36] and by Lehmann as Lemma 3 in [21, § 3.6]. For more history, see [22], [12] and [41].

We now define an additional set-valued function Z by

$$Z(x) = \{(c(x, y), e(x, y)): y \in Y(x)\} \quad \text{for } x \in X.$$

Note that for $x \in X$, while $Z(x) \subset \mathscr{E}_{k+1}$, $Y(x)$ is an arbitrary set. We fix $\mathscr{F}$ as the set of integrable functions f on X such that $f(x) \in Z(x)$ for $x \in X$ and define

$$I_i(f) = \int_X f_i(x)\, d\mu x \quad \text{for } f \in \mathscr{F}, \qquad i = 1, \cdots, k + 1,$$

and

$$I = (I_1, \cdots, I_{k+1}).$$

By the Axiom of Choice, **rng** I = **rng** (C, E). The nonlinear (C, E) and arbitrarily-valued Y provide a more general framework than that of the linear I. We shall subsequently condition e and c in ways which yield results not obtainable from I alone.

THEOREM 2.1. *If $\lambda \in \mathscr{E}_k$, $q^* \in \Phi$, $E(q^*) > -\infty$, and $C(q^*)$ is finite, then in the following,* (i) *is sufficient for* (ii) *and* (ii) *is sufficient for* (iii):

(i) *(q^*, λ) satisfies (strongly satisfies) the pointwise multiplier rule*;

(ii) *(q^*, λ) satisfies (strongly satisfies) the functional multiplier rule*;

(iii) *q^* is optimal (strongly optimal).*

Proof. This is straightforward, with care to avoid $\infty - \infty$.

THEOREM 2.2. *Suppose μ has finite substance, $q^* \in \Phi$, $E(q^*)$ and $C(q^*)$ are finite, and $\lambda \in \mathscr{E}_k$. Then the following three conditions are equivalent*:

(i) *$L_\lambda(q^*) \geqq L_\lambda(q)$ whenever $L_\lambda(q)$ exists*;

(ii) *$l_\lambda(x, q^*(x)) \geqq l_\lambda(x, q(x))$ for $x \in X$, whenever $q \in \Xi$*;

(iii) *$L_\lambda(q^*) \geqq L_\lambda(q)$ whenever $E(q)$ and $C(q)$ are finite.*

Proof. Obviously (i) implies (iii), and (ii) implies (i). If (iii) holds and (ii) fails, then for some $q \in \Xi$, letting $P = \{x : l_\lambda(x, q(x)) > l_\lambda(x, q^*(x))\}$, we have $\mu(P) > 0$. For $j \in \omega$, let

$$P_j = P \cap \{x : |e(x, q(x))| \leqq j \text{ and } |c_i(x, q(x))| \leqq j, i = 1, \cdots, k\}.$$

Then $\bigcup_{j=1}^\infty P_j = P$, so for some $j_0, \mu(P_{j_0}) > 0$. Take a measurable $S \subset P_{j_0}$ such that $0 < \mu(S) < \infty$. Then $e(\cdot, q(\cdot))$ and $c(\cdot, q(\cdot))$ restricted to S are finitely integrable. Define $p(x) = q(x)$ for $x \in S$ and $p(x) = q^*(x)$ for $x \in X - S$. Then $p \in \Phi$ and $L_\lambda(p) > L_\lambda(q^*)$ in contradiction, which completes the proof.

Remark 2.3. In Remark 3.4, Theorem 5.3, and Theorem 6.3, X is countable and $\mu(\{x\}) = 1$ for $x \in X$, whence $E(q) = \sum_{x \in X} e(x, q(x))$, and similarly for C. These discrete summation byproducts of integration results also hold if summation is defined as a limit of partial sums.

3. Necessity of functional multiplier rule for optimality. Corollary 3.3 below gives the necessity of the (strong) functional multiplier rule for (strong) optimality. Theorem 3.2 asserts necessity of a Kuhn–Tucker [19] condition, (iii), for strong optimality for a given cost. These follow from a known consequence of a separating hyperplane theorem, applied with Theorem 2.2 and a generalized Lyapunov convexity statement, Theorem 3.1. Theorem 3.2 ((ii) iff (iv)) has been given by Meeks and Francis [26], [25] under stronger concavity and other conditions.

We say that the functionals *C and $-E$ are convex* if

$$\hat{\Phi} = \Phi \cap \{q : C_i(q) < \infty \text{ for } i = 1, \cdots, k \text{ and } E(q) > -\infty\}$$

is a convex subset of a real vector space and whenever $0 \leqq \alpha \leqq 1$ and $q, p \in \hat{\Phi}$, we have $C(\alpha q + (1 - \alpha)p) \leqq \alpha C(q) + (1 - \alpha)C(p)$ and $E(\alpha q + (1 - \alpha)p) \geqq \alpha E(q) + (1 - \alpha)E(p)$.

THEOREM 3.1 (Lyapunov and Blackwell). *If μ is nonatomic, then* **fin rng** *I, i.e.,* **fin rng** *(C, E), is convex.*

Proof. Apply the proof of Theorem 3 of [5].

THEOREM 3.2. *Suppose μ has finite substance, $v \in \mathscr{E}_k$, $u \in$* **rng** *C, $u_i < v_i$ for $i = 1, \cdots, k$, $q^* \in \Phi$, $E(q^*)$ and $C(q^*)$ are finite, and $C(q^*) \leqq v$. Suppose also that either* (a) *μ is nonatomic or* (b) *C and $-E$ are convex. Then the following are equivalent*:

(i) $E(q^*) = \max\{E(q) : C(q) \leqq v\}$;

(ii) $E(q^*) = \max\{E(q) : -\infty < C_i(q) \leqq v_i, i = 1, \cdots, k\}$;

(iii) *there exists $\lambda \in \mathscr{E}_k^+$ such that whenever $q \in \Phi$ and $L_\lambda(q)$ exists,*

$$L_\lambda(q) + \lambda \cdot v \leqq L_\lambda(q^*) + \lambda \cdot v \leqq L_\eta(q^*) + \eta \cdot v \quad \text{for } \eta \in \mathscr{E}_k^+, \tag{3.1a}$$

and

$$\lambda \cdot [v - C(q^*)] = 0; \tag{3.1b}$$

(iv) *there exists a* $\lambda \in \mathscr{E}_k^+$ *such that* (3.1) *holds when* $E(q)$ *and* $C(q)$ *are finite.*

Proof. Clearly (i) implies (ii); by Theorem 2.2 ((i) iff (iii)), we have (iii) iff (iv); and the proof that (iii) implies (i) is straightforward. That (ii) implies (iv) is a corollary (as pointed out by a referee) of Luenberger's Theorem 1 [22, p. 217] and its corollary by choosing the Ω of [22] to be **fin** $\{(w, d) : w \geqq C(q) \text{ and } d \leqq E(q) \text{ for some } q \in \Phi\}$; this is easily shown to be convex, using Theorem 3.1 in case (a).

Corollary 3.3. *Suppose* μ *has finite substance,* $q^* \in \Phi$, $|E(q^*)| < \infty$, $C(q^*) \in \textbf{int rng}\, C$, *and either* (a) μ *is nonatomic or* (b) C *and* $-E$ *are convex. Then for* q^* *to be optimal* (*strongly optimal*) *it is necessary and sufficient that for some* $\lambda \in \mathscr{E}_k$, (q^*, λ) *satisfy* (*strongly satisfy*) *the functional multiplier rule.*

Proof. This follows from Theorem 3.2 by setting $v = C(q^*)$ and noting that q^* is optimal iff $E(q^*) = \max\{E(q) : C(q) \leqq v \text{ and } -C(q) \leqq -v\}$.

Remark 3.4. Hypothesis (b) in Theorem 3.2 and Corollary 3.3 is satisfied if for $x \in X$, $Y(x)$ is a convex subset of a real vector space and $c(x, \cdot)$ and $-e(x, \cdot)$ are convex functions. This is useful when X is countable and $\mu(\{x\}) = 1$ for $x \in X$ (so integration is discrete summation; see Remark 2.3), since hypothesis (a) does not then apply. That necessity in Theorem 3.2 and Corollary 3.3 fails in this discrete case, in the absence of convexity of $-e(x, \cdot)$, $x \in X$, is shown by example in [37] and [43].

Remark 3.5. Suppose Φ_0 is a subset of Φ which is closed under "switching" or "exchange" in the sense that if $q, p \in \Phi_0$, P is measurable, $r(x) = p(x)$ for $x \in P$, and $r(x) = q(x)$ for $x \in X - P$, then $r \in \Phi_0$. Then Theorems 2.2, 3.1, and 3.2 and Corollary 3.3 hold if in the definitions and statements involving Φ, Φ_0 is used instead.

4. Von Neumann selection theorem. We now make topological assumptions, under which we give a generalized version of von Neumann's selection result, needed in § 5 and § 6. This is given as Lemma 5 of [40] under more special conditions, e.g., $X = \mathscr{E}_1$; however, von Neumann's method of proof suffices also for the present version. A new proof is given here, due to Professor Herbert Federer and based on the proof of § 2.2.12 of [11]. Professor J. C. Oxtoby has provided an alternative formulation of this proof, also outlined below.

If A and B are topological spaces and $f : A \to B$, we say that f is a *Borel function* if for each open $G \subset B$, $f^{-1}(G)$ is a Borel subset of A. If also g is a measurable function into A, then $f \circ g$ is measurable. Let $\mathscr{N} = \omega^\omega$, the set of infinite sequences of positive integers with the product topology formed from the discrete topology on each factor ω; $\mathscr{N}$ is homeomorphic to the irrational numbers. Any continuous image of $\mathscr{N}$ in a Hausdorff space is a Suslin set. Every Borel subset of a metric space is a Suslin set. If open sets are measurable, so are Suslin sets. The foregoing is found, for example, in § 2.2 of [11].

Theorem 4.1 (von Neumann). *Suppose* X *is a Hausdorff space, open subsets of* X *are* μ *measurable,* $\mathscr{S}$ *is a Suslin subset of a complete separable metric space* W, $h : \mathscr{S} \to X$ *is continuous,* $\mu(h(\mathscr{S})) < \infty$, *and* $\varepsilon > 0$. *Then there exist a compact* $D \subset h(\mathscr{S})$ *and a Borel function* $f : D \to \mathscr{S}$ *such that* $\mu(h(\mathscr{S}) - D) \leqq \varepsilon$, *and* $h(f(x)) = x$ *for* $x \in D$.

Proof (Federer). For $S \subset \mathscr{S}$, define $\gamma(S) = \mu(h(S))$. (This uses the measure foundations of [11]; if [17] is used, one replaces μ by its associated outer measure.) Let g map $\mathscr{N}$ continuously onto $\mathscr{S}$ (see § 2.2.10 of [11]).

Define $Z_0 = (W \times \mathscr{N}) \cap \{(w, \sigma) : w = g(\sigma)\}$. Then Z_0 is closed. For $(w, \sigma) \in W \times \mathscr{N}$ define $\eta(w, \sigma) = w$. Then $\eta(Z_0) = \mathscr{S}$. Corresponding to $i \in \omega$, we inductively choose $\tau_i \in \omega$ and closed sets $Z_i \subset Z_0$ so that

$$(4.1) \qquad Z_i = Z_{i-1} \cap \{(w, \sigma) : \sigma_i \leqq \tau_i\} \quad \text{and} \quad \gamma[\eta(Z_{i-1})] - \gamma[\eta(Z_i)] < \varepsilon 2^{-i};$$

this is possible because defining $A_{ij} = \eta(Z_{i-1} \cap \{(w, \sigma) : \sigma_i \leqq j\})$ for $i, j \in \omega$, we have

$$\bigcup_{j=1}^{\infty} A_{ij} = \eta(Z_{i-1}) \quad \text{and} \quad \gamma\left(\bigcup_{j=1}^{\infty} A_{ij}\right) = \mu\left(\bigcup_{j=1}^{\infty} h(A_{ij})\right) = \lim_{j \to \infty} \gamma(A_{ij}).$$

Here we have used the fact that a continuous image in W or X of a Suslin subset of a complete separable metric space is also a Suslin set (see § 2.2.10 of [11]); hence each A_{ij} and $h(A_{ij})$ is a Suslin set, and each $h(A_{ij})$ is μ measurable.

Let $K = \mathscr{N} \cap \{\sigma : \sigma_i \leqq \tau_i \text{ for } i \in \omega\}$. Then K is compact (by Tikhonov's theorem) and

$$\bigcap_{i=1}^{\infty} Z_i = Z_0 \cap (W \times K) = (W \times \mathscr{N}) \cap \{(w, \sigma) : \sigma \in K \text{ and } w = g(\sigma)\}.$$

Hence $g(K) = \eta(Z_0 \cap (W \times K)) \subset \mathscr{S}$. Let $D = h(g(K))$. Then $g(K)$ and D are compact.

There exists a Cantor set $\Gamma \subset [0, 1]$ and a continuous map β on Γ onto $g(K)$. Following [35, § 7.1, Chap IX], for $x \in D$, let $\alpha(x) = \min \{t : h(\beta(t)) = x\}$, whence $\alpha(x) \in \Gamma$. Let $f = \beta \circ \alpha$. For $t \in \Gamma$, $\{x : \alpha(x) \leqq t\} = h(\beta(\Gamma \cap [0, t]))$ which is a Borel set since $h \circ \beta$ is continuous. Thus, α is a Borel function and, therefore, so is f.

It remains to show that $\mu(h(\mathscr{S}) - h(g(K))) \leqq \varepsilon$. This will follow from (4.1) if we show that $g(K) \supset \bigcap_{i=1}^{\infty} \eta(Z_i)$ (the reverse inclusion is obvious), which may be proved by following exactly the argument in § 2.2.12 of [11] that $C \subset p[Z_0 \cap (X \times K)]$ (here W, η, and $\bigcap_{i=1}^{\infty} \eta(Z_i)$ correspond respectively to X, p, and C in [11]). This completes the proof.

An outline of Oxtoby's alternative formulation is as follows: take γ, g, and K as before, and choose $\tau \in \mathscr{N}$ and $\mathscr{N} = V_0 \supset V_1 \supset \cdots$ such that for $i \in \omega$,

$$V_i = V_{i-1} \cap \{\sigma : \sigma_i \leqq \tau_i\} \quad \text{and} \quad \gamma[g(V_{i-1})] - \gamma[g(V_i)] < \varepsilon 2^{-i};$$

then $K = \bigcap_{i=1}^{\infty} V_i$. Show that $g(K) \supset \bigcap_{i=1}^{\infty} g(V_i)$ by a diagonal selection argument similar to Sierpinski's, given in [35, § 5.3, Chap. II]. Then $D = h(g(K))$ serves as before.

Remark 4.2. If in Theorem 4.1 we require μ to be Borel regular (§ 2.2.3 of [11]) and X to be a metric space, then by Lusin's theorem (§ 2.3.5 of [11]) we can obtain f to be continuous. If μ is Borel regular and σ-finite, then any measurable function is a.e. equal to a Borel function (§ 2.3.6 of [11]); in that case, we can obtain f to be a Borel function defined a.e. on $h(\mathscr{S})$, without requiring $\mu(h(\mathscr{S})) < \infty$. For related results, see [6]–[8], [18], [20], [24], and [33].

5. Necessity of pointwise multiplier rule for optimality. We now apply the above results to prove the necessity of the pointwise multiplier rule for optimality, under some weak assumptions primarily of a Borel nature.

Defining $\pi(x, y) = x$ for $(x, y) \in \Omega$, we stipulate the following condition (also used in Lemma 6.12 and Theorem 6.13 below):

Condition (α): *X is a metric space, Ω is a Borel subset of a complete separable metric space, π is continuous, c and e are Borel functions, and μ is Borel regular.* Note that Condition (α) is satisfied if, in particular, Ω is a Borel subset of $\mathscr{E}_n$, π projects Ω into $\mathscr{E}_m$ with $m < n$, c and e are Borel functions, and μ is m-dimensional Lebesgue measure. It also holds if $X = \omega$ and for $x \in X$, $\mu(\{x\}) = 1$ and $Y(x)$ is a Borel subset of $\mathscr{E}_k$.

THEOREM 5.1. *Suppose Condition* (α) *holds, μ has finite substance, $q^* \in \Phi$, $E(q^*)$ and $C(q^*)$ are finite, and $e(\cdot, q^*(\cdot))$ and $c(\cdot, q^*(\cdot))$ are Borel functions. Then for $\lambda \in \mathscr{E}_k$, for (q^*, λ) to satisfy (strongly satisfy) the functional multiplier rule, it is necessary and sufficient that (q^*, λ) satisfy (strongly satisfy) the pointwise multiplier rule.*

Proof. Sufficiency follows from Theorem 2.1.

Suppose (2.1) holds but (2.2) fails. Let

$$\mathscr{B} = \{(x, y) : l_\lambda(x, q^*(x)) < l_\lambda(x, y)\}.$$

Then $\mathscr{B}$ is a Borel set, so $\pi(\mathscr{B})$ is measurable. Since (2.2) fails, $\mu(\pi(\mathscr{B})) > 0$. Choose a measurable $A \subset \pi(\mathscr{B})$ such that $0 < \mu(A) < \infty$. Since μ is Borel regular, there exists a Borel set $Q \supset A$ such that $\mu(Q) = \mu(A)$. Let $\mathscr{S} = \mathscr{B} \cap \pi^{-1}(Q)$. Then $\mathscr{S}$ is a Borel set and $0 < \mu(\pi(\mathscr{S})) < \infty$.

By Theorem 4.1 (with $h = \pi$), there exists a Borel set $P \subset \pi(\mathscr{B})$ and a Borel function $f: P \to \mathscr{B}$ such that $\pi(f(x)) = x$ for $x \in P$ and $\mu(P) > 0$. Let $\kappa(x, y) = y$ for $(x, y) \in \Omega$ and let $p = \kappa \circ f$. Then for $x \in P$, $(x, p(x)) \in \mathscr{B}$, i.e., $l_\lambda(x, p(x)) > l_\lambda(x, q^*(x))$. Let $\hat{q}(x) = p(x)$ for $x \in P$ and $\hat{q}(x) = q^*(x)$ for $x \in X - P$. Since p, e, and c are Borel functions, $\hat{q} \in \Xi$ and Theorem 2.2 (ii) fails with $q = \hat{q}$. Hence, Theorem 2.2 (i) fails, contrary to hypothesis.

COROLLARY 5.2. *Suppose the hypothesis of Theorem* 5.1 *holds, $C(q^*) \in$* **int rng** *C, and either* (a) *μ is nonatomic or* (b) *C and $-E$ are convex* (see § 3). *Then for q^* to be optimal (strongly optimal), it is necessary and sufficient that for some $\lambda \in \mathscr{E}_k$, (q^*, λ) satisfy (strongly satisfy) the pointwise multiplier rule.*

Proof. This follows from Corollary 3.3 and Theorem 5.1.

THEOREM 5.3. *Suppose X is countable and for $x \in X$, $\mu(\{x\}) = 1$, $Y(x) \subset \mathscr{E}_m$, $Y(x)$ is convex, and $c(x, \cdot)$ and $-e(x, \cdot)$ are convex functions. Suppose $q^* \in \Phi$, $|E(q^*)| < \infty$, and $C(q^*) \in$* **int rng** *C. Then the conclusion of Corollary* 5.2 *holds.*

Proof. Apply Corollary 3.3, Remark 3.4, and the method of proving Theorem 5.1 without resorting to Theorem 4.1.

Remark 5.4. If μ is σ-finite, then by Remark 4.2, Theorem 5.1 and Corollary 5.2 hold without assuming $e(\cdot, q^*(\cdot))$ and $c(\cdot, q^*(\cdot))$ are Borel functions.

Theorem 5.1 and Corollary 5.2 have *linear* functional corollaries which are easily formed by letting $Y = Z$, and for $(x, y) \in \Omega$, $(c(x, y), e(x, y)) = y$. The Borel condition on Ω becomes simply the condition that the "graph" of Z, i.e., $\{(x, z) : x \in X, z \in Z(x)\}$, be a Borel subset of $\bar{X} \times \mathscr{E}_{k+1}$, where $\bar{X}$ is a completion of X, a separable metric space. The latter condition is not implied by the hypothesis of Theorem 5. Thus, Theorem 5.1 is not a corollary of the linear functional statement.

Remark 5.5. We show by example that the assumption in Theorem 5.1 (and Corollary 5.2) that e is a Borel function may not be replaced by the assumption that e is a measurable function with respect to a well-behaved measure over Ω.

Let $X = [0, 1]$ and let μ be the Lebesgue measure restricted to X. For $x \in X$, let $Y(x) = [-2, 2]$; thus, $\Omega = [0, 1] \times [-2, 2]$. Let μ^* and μ_* be, respectively, outer and inner Lebesgue measure on X (under the foundations of [11], $\mu = \mu^*$). Choose $D \subset X$ such that $\mu^*(D) = 1$ and $\mu_*(D) = 0$; then $\mu^*(X - D) = 1$, $\mu_*(X - D) = 0$, and D is not measurable. Let $k = 1$, let $c(x, y) = y$ for $(x, y) \in \Omega$, and defining

$$\mathscr{A} = \left\{(x, y) : \left[x \in D \text{ and } |y| = 1 - \frac{1}{2i}\right] \text{ or } \right.$$
$$\left. \left[x \in X - D \text{ and } |y| = 1 - \frac{1}{1 + 2i}\right] \text{ for some } i \in \omega\right\},$$

let e be the indicator function of $\mathscr{A}$. Then e is a measurable function with respect to two-dimensional Lebesgue measure, since $(\mu^* \times \mu^*)(\mathscr{A}) = 0$. However, $\pi(\mathscr{A} \cap ([0, 1] \times \{\tfrac{1}{2}\})) = D$, so $\mathscr{A}$ is not a Borel set, and e is not a Borel function.

Suppose $q \in \Phi$ with $E(q) > 0$. Then $e(\cdot, q(\cdot))$ and q, i.e., $c(\cdot, q(\cdot))$, are measurable functions. Let $P = \{x : e(x, q(x)) > 0\}$, $A = \{q(x) : x \in P\}$, and $S_y = \{x : q(x) = y\}$ for $y \in A$. For $y \in A$, $S_y \subset D$ or $S_y \subset X - D$, and, therefore, $\mu_*(S_y) = 0$, whence $\mu(S_y) = 0$, since S_y must be measurable. Therefore, $\mu(P) = 0$, since A is countable, in contradiction to $E(q) > 0$. Thus for no $q \in \Phi$ is $E(q) > 0$.

Let $q^*(x) = 0$ for $x \in X$. Then $C(q^*) = 0 = E(q^*)$ and $E(q^*) = \max\{E(q) : C(q) = 0\}$. Thus q^* is optimal and $(q^*, 0)$ satisfies the functional multiplier rule. Clearly, (q^*, λ) does not satisfy the pointwise multiplier rule for any λ. Furthermore, there is no $(\hat{q}, \lambda)$ satisfying the pointwise multiplier rule with $C(\hat{q}) = C(q^*)$. Thus, Theorem 5.1 fails if e is not a Borel function, since all other hypotheses are satisfied.

If we redefine $e(x, -1) = e(x, 1) = 1$ for $x \in X$, then e is "coverable" as defined in [42] and the necessity conclusions of Theorem 5.1 and Corollary 5.2 are restored.

Remark 5.6. More restricted versions of Corollary 5.2 (with hypothesis (a)) have been obtained by Zahl [43], Aumann and Perles [3], and Meeks and Francis [26], [25], each using Lebesgue measure for μ. Although none of these is a corollary of the other two, among them the treatment with the most strength is in [3]. Indeed the pattern of our development leading to Corollary 5.2 is similar to the pattern used in [3], as noted in § 1. However, the version in [3] unnecessarily assumes that for $x \in X$, $Y(x) = \mathscr{E}_1^+$, and $e(x, \cdot)$ is nondecreasing and nonnegative.

Theorem 5.1 will be applied in proving our main result, Theorem 6.13 (iii). In [37] and [38], Stone applies Corollaries 5.2 and 5.3 to prove that, under weak hypotheses, incrementally optimal separable allocations are totally optimal.

6. Results on existence and rng I being upper closed. We now give results on existence of optimal functions, for cost constrained to a given closed set. Results of Olech [31], [32] and Blackwell [5] are of fundamental importance to this development; see Remark 6.7. Lemma 6.1, stated without proof, obviates explicit mention of the desired existence in most subsequent theorems by asserting that it follows from the finite part of the range of a separable vector functional, **fin rng** I (= **fin rng** (C, E)), being suitably bounded and "upper closed" (defined below).

We usually assume below that **rng** C is bounded and μ is σ-finite.

Upper closure results when μ is purely atomic, i.e., in effect when X is countable, are given in Theorem 6.3. Theorem 6.6 (ii) generalizes Lyapunov's compactness theorem and gives the desired existence when **rng** (C, E) is bounded and $Z(x)$ is closed for $x \in X$. Boundedness of **rng** E is weakened to **rng** E being bounded above in Theorem 6.9. Theorem 6.10 asserts that if for one p, $E(p) = \infty$, then, mainly under nonatomic μ, every interior C value is attained by a q with $E(q) = \infty$. For this much, no topology on X is needed.

Suppose for $x \in X$, $Z(x)$ is merely "upper closed" instead of closed. Then Theorems 6.6 and 6.9 fail, as shown by example in Remark 6.11. However, Theorem 6.13, which includes Condition (α) of § 5 (Borel and other topological assumptions), asserts existence if the constraint set is convex and either (a) μ is nonatomic or (b) **rng** E is bounded above. Remark 6.14 shows by example that existence may fail if both (a) and (b) fail. Theorem 6.13 also asserts that **fin rng** (C, E) is upper closed if (b) holds and that **fin rng** (C, E) contains the extreme points of its "upper boundary" if both (a) and (b) hold.

Suppose $A \subset \mathscr{E}_{k+1}$. We say A is *upper closed* if for $(w, d) \in$ **cl** A, there exists $b \geqq d$ such that $(w, b) \in A$. This specializes the concept given by Olech in [31] as lower closure with respect to a given closed convex proper cone, here taking the cone to be $\{0^k\} \times \{a : a \leqq 0\}$. We define the *upper boundary* of A, denoted **upbdy** A, by

$$\textbf{upbdy}\, A = \textbf{cl}\, A \cap \{(w, d) : d \geqq d' \text{ whenever } (w, d') \in \textbf{cl}\, A\}.$$

If $\{d : (w, d) \in A \text{ for some } w\}$ is bounded above, then **upbdy** A is the graph of an u.s.c. function on $\{w : (w, d) \in A \text{ for some } d\}$.

For $w \in$ **rng** C, we define

$$v(w) = \sup \{E(q) : C(q) = w\}.$$

This supremum is attained iff there exists an optimal function with cost w.

We say b is an *extreme point* of $S \subset \mathscr{E}_{k+1}$ if $b \in S$ and there exist no $a, d \in S - \{b\}$ and $0 < \alpha < 1$ such that $b = (1 - \alpha)a + \alpha d$. By **ext** S we mean the set of extreme points of S.

LEMMA 6.1. *If* **rng** C *is bounded*, **rng** E *is bounded above*, **fin rng** (C, E) *is upper closed*, $N \subset \mathscr{E}_k$ *is closed, and* $N \cap$ **rng** $C \neq \emptyset$, *then* v *is u.s.c., and there exists* $p^* \in \Phi$ *such that* $C(p^*) \in N$ *and* $E(p^*) = \max \{E(p) : C(p) \in N\}$.

LEMMA 6.2. *Let* $g : X \to \mathscr{E}_{k+1}$ *be a measurable function and* $g(x) \in Z(x)$ *for* $x \in X$. *Suppose* $1 \leqq j \leqq k + 1$, $I_j(h) < \infty$ *for* $h \in \mathscr{F}$, *and* **fin rng** $I \neq \phi$. *Then* $\int_X g_j \, d\mu < \infty$.

Proof. We may take $j = k + 1$. Suppose the conclusion fails, and choose $f \in \mathscr{F}$ such that $I(f)$ is finite. Let $h^0 = g$. For $i = 1, \cdots, k$, inductively let $P_i = \{x : h_i^{i-1}(x) \geqq 0\}$, choose a measurable Q_i such that $Q_i \subset P_i$ or $Q_i \subset X - P_i$ and $\int_{Q_i} h_{k+1}^{i-1} \, d\mu = \infty$, and define $h^i(x) = h^{i-1}(x)$ for $x \in Q_i$ and $h^i(x) = f(x)$ for $x \in X - Q_i$; then h_n^i is integrable, for $n = 1, \cdots, i$, and $\int_X h_{k+1}^i \, d\mu = \infty$. Thus $h^k \in \mathscr{F}$ and $I_{k+1}(h^k) = \infty$, which contradicts the hypothesis.

THEOREM 6.3. *Suppose μ is σ-finite and purely atomic. Then*:

(i) *if $Z(x)$ is closed for $x \in X$ and* **rng** I (=**rng** (C, E)) *is bounded, then* **rng** I *is compact* (due largely to Blackwell [5]);

(ii) *if $Z(x)$ is upper closed for $x \in X$,* **rng** C *is bounded, and* **rng** E *is bounded above, then* **fin rng** (C, E) *is upper closed.*

Proof. We may assume **fin rng** $I \neq \emptyset$, X is countable, and $\mu(\{x\}) = 1, x \in X$. Suppose the hypothesis of (ii) holds and $(w, d) \in$ **cl fin rng** (C, E). We must find $g \in \mathscr{F}$ such that $(I_1(g), \cdots, I_k(g)) = w$ and $I_{k+1}(g) \geqq d$. Choose $h^1, h^2, \cdots \in \mathscr{F}$ such that $I(h^n) \to (w, d)$. Now $(h^1, h^2, \cdots)$ is a pointwise bounded sequence since $(w, d) \in \mathscr{E}_{k+1}$, **rng** $(I_1, \cdots, I_k)$ is bounded, and **rng** I_{k+1} is bounded above; by diagonal selection, choose a subsequence $(h^{\beta_1}, h^{\beta_2}, \cdots)$ which converges pointwise, to h by definition. For $x \in X, h(x) \in$ **cl** $Z(x)$, so since $Z(x)$ is upper closed, we may choose $b(x) \geqq h_{k+1}(x)$ such that $g(x) \equiv (h_1(x), \cdots, h_k(x), b(x)) \in Z(x)$.

For $x \in X$, by the upper closure of $Z(x)$ and boundedness conditions, we choose $m(x) \in Z(x)$ such that $m_{k+1}(x) \geqq z_{k+1}$ for $z \in Z(x)$. By Lemma 6.2, $|\int_X m_{k+1}\, d\mu| < \infty$. By Fatou's lemma applied to $(m_{k+1} - h^{\beta_1}_{k+1}, m_{k+1} - h^{\beta_2}_{k+1}, \cdots)$,

$$d = \lim_{n \to \infty} \int_X h^{\beta_n}_{k+1}\, d\mu \leqq \int_X \limsup_{n \to \infty} h^{\beta_n}_{k+1}(x)\, d\mu x = \int_X h_{k+1}\, d\mu \leqq \int_X b\, d\mu.$$

By similar argument, $w_i \leqq \int_X h_i\, d\mu \leqq w_i$ for $i = 1, \cdots, k$, proving (ii).

By similar but easier argument, one proves (i) to complete the proof.

Remark 6.4. The hypothesis "$Z(x)$ is upper closed for $x \in X$" appearing in Theorem 6.3 (ii) (and also in Theorem 6.13 below) is obviously satisfied if for $x \in X$, $Y(x)$ is a compact space, $c(x, \cdot)$ is continuous, and $e(x, \cdot)$ is u.s.c.

LEMMA 6.5 (Olech [31], [32]). *Suppose μ is nonatomic and σ-finite, and $b \in$* **ext cl fin rng** *I. Then there exist $f^1, f^2, \cdots \in \mathscr{F}$ and $f: X \to \mathscr{E}_{k+1}$ such that $f^j(x) \to f(x)$ for $x \in X$ and $\int_X f\, d\mu = b$.*

THEOREM 6.6. *Suppose μ is σ-finite. Then* (recall **rng** I = **rng** (C, E)):

(i) *if μ is nonatomic and $Z(x)$ is closed for $x \in X$, then* **fin rng** I *contains* **ext cl fin rng** I (due to Olech [31], [32]);

(ii) *if $Z(x)$ is closed for $x \in X$ and* **rng** I *is bounded, then* **rng** I *is compact.*

Proof. From Lemma 6.5, (i) follows. One proves (ii) by partitioning X into purely atomic and nonatomic subsets as in [5], and applying (i) and Theorem 6.3 (i).

Remark 6.7. Theorems 6.3 (i) and 6.6 (ii) were given by Blackwell [5] for the case where $\mu(X) < \infty$ and $Z(x)$ is the same for each $x \in X$. Lemma 6.5 and Theorem 6.6 (i) are due to Olech as noted (see Lemma 2, Theorem 3, and related discussion in [31]). Also in [31] is a generalization of Theorems 6.6 (ii) and 6.9 below, when μ is nonatomic. (Proof of the latter generalization, Theorem 7, can be simplified by referring to [13] as in the proof of Lemma 6.8 (ii) below.)

LEMMA 6.8. *Suppose $\Gamma \subset \mathscr{E}_{k+1}$ is convex, $K =$* **cl** *Γ, $\{a : (u, a) \in K$ for some $u\}$ is bounded above, and $\{u : (u, a) \in K$ for some $a\}$ is bounded. Then*:

(i) **ext upbdy** K = (**ext** K) $\cap$ **upbdy** K;

(ii) *if* **ext upbdy** $K \subset \Gamma$, *then Γ is upper closed*;

(iii) *if $\Gamma = \Gamma^1 + \Gamma^2$ (vector sum) and Γ^1 and Γ^2 are upper closed, so is Γ.*

Proof. Denying (i), suppose $(w, d) \in$ **ext upbdy** K and $(w, d) \notin$ **ext** K. Choose $0 < \alpha < 1, (w', d') \in K -$ **upbdy** K, and $(w'', d'') \in K$ such that $(w, d) = \alpha(w', d')$

$+ (1 - \alpha)(w'', d'')$. Pick $(w', b) \in K$ such that $b > d'$. Let $a = \alpha b + (1 - \alpha)d''$. Then $(w, a) \in K$ and $a > d$ contrary to $(w, d) \in \mathbf{upbdy}\, K$, so (i) follows.

To prove (ii), because of the boundedness hypotheses, it suffices to suppose $(w, d) \in \mathbf{upbdy}\, K$ and to show $(w, d) \in \Gamma$. Let A be the convex hull of $\mathbf{ext}\, K$ and $A' = \{0^k\} \times \{a : a \leqq 0\}$. By Theorem 6 of § 2.5 of [13], $K \subset A + A'$, since under the boundedness hypotheses, either A' or $\{0^{k+1}\}$ is the characteristic cone of K. We have $(w, d) = (u, a) + (0^k, a')$ with $(u, a) \in A$ and $a' \leqq 0$. Then $a' = 0$ else $(w, d) \notin \mathbf{upbdy}\, K$. Hence $(w, d) = \sum_{i=1}^m \alpha_i(w^i, d^i)$ with $\alpha^i > 0$ and $(w^i, d^i) \in \mathbf{ext}\, K$ for $i = 1, \cdots, m$ and $\sum_{i=1}^m \alpha^i = 1$. If for some j, $(w^j, d^j) \notin \mathbf{upbdy}\, K$, take $(w^j, d') \in K$ with $d' > d^j$, whence $(w, d + \alpha^j(d' - d^j)) \in K$, contrary to $(w, d) \in \mathbf{upbdy}\, K$. Thus for $i = 1, \cdots, m$, $(w^i, d^i) \in \mathbf{ext\, upbdy}\, K$ by (i), so $(w^i, d^i) \in \Gamma$. Hence, $(w, d) \in \Gamma$.

To prove (iii), suppose $(w, d) \in K$. For $n \in \omega$ and $i = 1, 2$, choose $(w^{ni}, d^{ni}) \in \Gamma^i$ such that $(w^{n1} + w^{n2}, d^{n1} + d^{n2}) \to (w, d)$. Because of the boundedness hypotheses and the fact that $d^{n1} + d^{n2} \to d \in \mathscr{E}_1$, for $i = 1, 2$, $\{(w^{ni}, d^{ni}) : n \in \omega\}$ is bounded; let it have $(w^i, d^i) \in \mathbf{cl}\, \Gamma^i$ as a limit point and choose $(w^i, b^i) \in \Gamma^i$ such that $b^i \geqq d^i$. Then $(w^1, d^1) + (w^2, d^2) = (w, d)$, so $(w, b^1 + b^2) \in \Gamma$ and $b^1 + b^2 \geqq d$. This completes the proof.

THEOREM 6.9. *Suppose μ is σ-finite,* **rng** *C is bounded,* **rng** *E is bounded above, and $Z(x)$ is closed for $x \in X$. Then* **fin rng** *(C, E) is upper closed.*

Proof. If μ is nonatomic, the theorem follows from Lemma 6.8 (i) and (ii) (with $\Gamma = \mathbf{fin\, rng}\,(C, E)$) and Theorems 3.1 and 6.6 (i) (alternatively apply Theorem 7 of Olech [31]). Using this fact, Theorem 6.3 (ii), and the partitioning of X in the proof of Theorem 6.6 (ii), one obtains the theorem from Lemma 6.8 (iii).

THEOREM 6.10. *Let μ be nonatomic and $p \in \Phi$ with $C(p)$ finite and $E(p) = \infty$. Then*:

(i) *if $w \in$* **fin rng** *C, $v(w) > -\infty$, and $0 < \alpha < 1$, then there exists an $s \in \Phi$ such that $E(s) = \infty$ and $[C(s) = \alpha w + (1 - \alpha)C(p)$ or $C(s) = (1 - \alpha)w + \alpha C(p)]$*;

(ii) *if $C(p) \in \Lambda \subset$* **fin rng** *C, Λ is a line segment, and $v(u) > -\infty$ for all $u \in \Lambda$, then for all $v \in$* **int** *Λ there exists a* (*strongly optimal*) *$p^* \in \Phi$ such that $C(p^*) = v$ and $E(p^*) = \infty$*;

(iii) *if $v(u) > -\infty$ for all $u \in$* **int rng** *C, then for all $v \in$* **int rng** *C, there exists a* (*strongly optimal*) *$p^* \in \Phi$ such that $C(p^*) = v$ and $E(p^*) = \infty$.*

Proof. By Theorem 3.1 and Remark 3.5, if $v(u) > -\infty$ for $u \in \mathbf{int\, rng}\, C$, then **int rng** C is convex, so (iii) follows from (ii).

Let the hypothesis of (i) hold. Choose $r \in \Phi$ such that $C(r) = w$ and $E(r) > -\infty$. Again applying Theorem 3.1 and Remark 3.5, we choose a measurable $S \subset X$ and $s, \bar{s} \in \Phi$ such that $s(x) = p(x)$ for $x \in S$, $s(x) = r(x)$ for $x \in X - S$, $\bar{s}(x) = r(x)$ for $x \in S$, $\bar{s}(x) = p(x)$ for $x \in X - S$, $C(s) = (1 - \alpha)w + \alpha C(p)$, and $C(\bar{s}) = \alpha w + (1 - \alpha)C(p)$. Then $E(s) = \infty$ or $E(\bar{s}) = \infty$, else $E(p) \neq \infty$ in contradiction. This proves (i).

Suppose the hypothesis of (ii) holds. Since $v \in \mathbf{int}\, \Lambda$, we may choose $j \in \omega$ sufficiently large that $v + 2^{-j}[v - C(p)] \in \Lambda$. By inductive application of (i) with $\alpha = \frac{1}{2}$, one finds q such that $E(q) = \infty$ and $C(q) = v - 2^{-j}[v - C(p)]$. An additional application of (i) yields the p^* desired in (ii). This completes the proof.

Remark 6.11. We show by example that in Theorems 6.6 and 6.9, we may not assume that $Z(x)$ is upper closed, instead of closed, for $x \in X$.

Let μ, k, X, Y, Ω, D, and c be as in Remark 5.5. Define

$$e(x,y) = \begin{cases} 1 - \tfrac{1}{2}|y| & \text{for } y \neq 0, \quad (x,y) \in \Omega, \\ 2 & \text{for } y = 0, \quad x \in D, \\ 3 & \text{for } y = 0, \quad x \in X - D. \end{cases}$$

Then for $x \in X$, $Z(x)$ is upper closed but not closed (the other hypotheses of Theorems 6.6 and 6.9 are satisfied). There exists no optimal $q^* \in \Phi$ such that $C(q^*) = 0$, although $0 \in \mathbf{rng}\, C$; hence, $\mathbf{rng}\,(C, E)$ is not upper closed. The example also shows that although Theorems 6.6 and 6.9 hold if for $x \in X$, $Y(x)$ is compact and $e(x, \cdot)$ and $c(x, \cdot)$ are continuous, they fail if $e(x, \cdot)$ is merely u.s.c.

If we change Theorem 6.9 by letting $Z(x)$ be upper closed instead of closed and by adding Condition (α), we obtain a valid statement, viz., Theorem 6.13 (ii) below. In the hypothesis of Theorem 6.13, the Borel conditions are on Ω, e, and c, while the upper closure condition is on Z. This adds complications to the proof in going from the Y, C, E structure to the Z, I structure and vice versa. As in Remark 5.4, a linear functional corollary is easily formed and proved whose hypothesis is in terms of Z and I. An obvious corollary is also formed by noting Remark 6.4. Both of these corollaries avoid the mixed structure nature of the hypothesis mentioned above, and both are weaker statements than Theorem 6.13.

LEMMA 6.12. *Suppose Condition* (α) *holds* (see § 5), *μ is nonatomic and σ-finite, and for $x \in X$, $\varnothing \neq Z^0(x) \subset Z(x)$, $Z^0(x)$ is upper closed, $\{(z_1, \cdots, z_k) : z \in Z^0(x)\}$ is bounded, and $m(x) = \sup\{z_{k+1} : z \in Z^0(x)\}$. Suppose $\Delta \equiv \Omega \cap \{(x,y) : (c(x,y), e(x,y)) \in Z^0(x)\}$ is a Borel set. Then there exists $g(x) \in Z^0(x)$ for $x \in X$ such that g is a measurable function and $[g_{k+1} = m \text{ or } \int_X g_{k+1}\, d\mu = \infty]$.*

Proof. To see that m is a measurable function, note that for $a \in \mathscr{E}_1$,

$$X \cap \{x : m(x) > a\} = \pi(\Delta \cap \{(x,y) : e(x,y) > a\}),$$

which is a continuous image of a Borel set and thus measurable (§ 2.2.10 of [11]).

Let $P = \{x : m(x) = \infty\}$. Then P is measurable. Since μ is σ-finite and Borel regular, by Remark 4.2 we may assume m is a Borel function and P is a Borel set. Since μ is also nonatomic, we may choose disjoint Borel sets $P_1, P_2, \cdots$ such that $P = \bigcup_{n=1}^{\infty} P_n$ and such that if $\mu(P) > 0$, then $0 < \mu(P_n) < \infty$ for $n \in \omega$. Let $P_0 = X - P$,

$$\mathscr{A}_n = \Delta \cap \{(x,y) : x \in P_n \text{ and } e(x,y)\mu(P_n) \geqq n\} \quad \text{for } n \in \omega,$$

$$\mathscr{A}_0 = \Delta \cap \{(x,y) : x \in P_0 \text{ and } e(x,y) = m(x)\}.$$

Then, for $n = 0, 1, \cdots$, $\mathscr{A}_n$ is a Borel set, $\pi(\mathscr{A}_n) = P_n$, and we may apply Theorem 4.1 and Remark 4.2 to obtain a Borel function p_n on (almost all of) P_n such that $(x, p_n(x)) \in \mathscr{A}_n$ for $x \in P_n$. Let $g(x) = (c(x, p_n(x)), e(x, p_n(x)))$ for $x \in P_n$ and $n = 0, 1, \cdots$. Then g has the desired properties, which completes the proof.

THEOREM 6.13. *Suppose Condition* (α) *holds, μ is σ-finite,* $\mathbf{rng}\, C$ *is bounded, and for $x \in X$, $Z(x)$ is upper closed. Let $\Gamma =$* $\mathbf{fin\, rng}\, I$ ($= \mathbf{fin\, rng}\,(C, E)$). *Then*:

(i) *if μ is nonatomic and* $\mathbf{rng}\, E$ *is bounded above, then* $\mathbf{ext\, upbdy}\, \Gamma \subset \Gamma$;

(ii) *if* $\mathbf{rng}\, E$ *is bounded above, then Γ is upper closed*;

(iii) *if $N \subset \mathscr{E}_k$ is closed and convex, $N \cap$* $\mathbf{rng}\, C \neq \varnothing$, *and either μ is nonatomic or* $\mathbf{rng}\, E$ *is bounded above, then there exists $p^* \in \Phi$ such that $C(p^*) \in N$ and $E(p^*) = \max\{E(p) : C(p) \in N\}$.*

Proof. Since $\mathbf{rng}\, C$ is bounded, using § 4 and Lemma 6.2 one argues as in proving Lemma 6.12 to show, if $\mathscr{F} \neq \varnothing$, $\{(z_1, \cdots, z_k) : z \in Z(x)\}$ is bounded for $x \in X$.

To prove (i), we choose $b \in \mathbf{ext\, upbdy}\, \Gamma$. Noting Lemma 6.8 (i) and Remark 4.2, we obtain Borel functions $f^1, f^2, \cdots \in \mathscr{F}$ and f as given by Lemma 6.5.

For $x \in X$, define $Z^0(x) = Z(x) \cap \{z : z_i = f_i(x) \text{ for } i = 1, \cdots, k\}$, whence $Z^0(x)$ is upper closed and since $f(x) \in \mathbf{cl}\, Z(x)$, $\sup \{z_{k+1} : z \in Z^0(x)\} \geqq f_{k+1}(x)$. By Lemma 6.12, there exists $g(x) \in Z^0(x)$ for $x \in X$ such that $\int_X g_{k+1}\, d\mu \geqq \int_X f_{k+1}\, d\mu = b_{k+1}$. Since $g_i = f_i$ for $i = 1, \cdots, k$, $g \in \mathscr{F}$ and $I_i(g) = b_i$ for $i = 1, \cdots, k$. Also, $I_{k+1}(g) < \infty$, for $\mathbf{rng}\, E$ is bounded above. Since $b \in \mathbf{upbdy}\, \Gamma$, $I_{k+1}(g) = b_{k+1}$, i.e., $b \in \Gamma$, proving (i).

We prove (ii) by applying Theorems 3.1 and 6.3 (ii) and Lemma 6.8, as in the proof of Theorem 6.9, but using (i) instead of Theorem 6.6 (i).

If $\mathbf{rng}\, E$ is bounded above, then (iii) follows from (ii) and Lemma 6.1. To complete the proof, it suffices to show the following:

If μ is nonatomic, $\mathbf{rng}\, E$ *is not bounded above, $N \subset \mathscr{E}_k$ is closed and convex,*

(6.1) *and $N \cap$* $\mathbf{rng}\, C$ *$\neq \varnothing$, then there exists p^* such that*

$$C(p^*) \in N \text{ and } E(p^*) = \max \{E(p) : C(p) \in N\}.$$

We define $F = \mathbf{rng}\, C \cap \{w : v(w) > -\infty\}$. By Theorem 3.1 and Remark 3.5, F is convex. Hence, $N \cap F$ is convex. If $N \cap F = \varnothing$, (6.1) is trivial, so we assume

$$N \cap \mathbf{int}\, F \neq \varnothing \quad \text{or} \quad \varnothing \neq N \cap F \subset \mathbf{bdy}\, F. \tag{6.2}$$

Since $N \cap F$ is convex, if $N \cap F \subset \mathbf{bdy}\, F$, then $N \cap F$ is contained in a supporting hyperplane of F. Thus, we may choose $v^0 \in N \cap F$ and $\eta \in \mathscr{E}_k$ such that

$$\eta = 0^k \quad \text{and} \quad v^0 \in N \cap \mathbf{int}\, F, \qquad \text{if } N \cap \mathbf{int}\, F \neq \varnothing; \tag{6.3}$$

$$\eta \cdot v^0 \leqq \eta \cdot w \quad \text{for } w \in F \text{ with equality when } w \in N \cap F; \tag{6.4}$$

$$\eta_k \neq 0 \quad \text{if } N \cap F \subset \mathbf{bdy}\, F \quad \text{(reordering coordinates, if necessary)}. \tag{6.5}$$

Fix $p_0 \in \Phi$ such that $C(p_0) = v^0$, $E(p_0) > -\infty$, and $c(\cdot, p_0(\cdot))$ is a Borel function (see Remark 4.2). If $E(p_0) = \infty$, p_0 serves as p^*, so assume $|E(p_0)| < \infty$. We claim that

$$\eta \cdot c(x, p_0(x)) \leqq \eta \cdot c(x, y) \quad \text{for } y \in Y(x), \quad x \in X. \tag{6.6}$$

If $\eta = 0^k$, then (6.6) holds trivially. If $\eta_k > 0$, let $\lambda = (1/\eta_k)(\eta_1, \cdots, \eta_{k-1}, 0)$, $C^0 = (C_1, \cdots, C_{k-1}, E)$, and $E^0 = -C_k$. By (6.4),

$$E^0(p_0) - \sum_{\lambda_i \neq 0} \lambda_i C_i^0(p_0) \geqq E^0(s) - \sum_{\lambda_i \neq 0} \lambda_i C_i^0(s) \quad \text{whenever } C(s) \in F.$$

By Theorem 2.2 ((iii) implies (i)), we find that the preceding inequality holds for all $s \in \Phi$. Theorem 5.1 yields that for $x \in X$, $p_0(x)$ maximizes $-c_k(x, \cdot) - \sum_{i=1}^{k-1} \lambda_i c_i(x, \cdot)$ over $T(x)$, whence (6.6) follows. We argue similarly if $\eta_k < 0$.

For $x \in X$, let $Z'(x) = Z(x) \cap \{z : \eta \cdot (z_1, \cdots, z_k) = \eta \cdot c(x, p_0(x))\}$; then $Z'(x)$ is upper closed and $(c(x, p_0(x)), e(x, p_0(x))) \in Z'(x)$. With m and g given by Lemma 6.12, we take $\hat{q} \in \Xi$ such that $g = (c(\cdot, \hat{q}(\cdot)), e(\cdot, \hat{q}(\cdot)))$. If $g_{k+1} = m$, then $e(\cdot, \hat{q}(\cdot))$

$\geqq e(\cdot, p_0(\cdot))$; in any event, $e(\cdot, \hat{q}(\cdot))$ is integrable. Since **rng** C is bounded, by Lemma 6.2, $c(\cdot, \hat{q}(\cdot))$ is integrable, i.e., $\hat{q} \in \Phi$. We have

$$\begin{aligned}(6.7)\qquad & C(\hat{q}) \in F, \quad \eta \cdot c(\cdot, \hat{q}(\cdot)) = \eta \cdot c(\cdot, p_0(\cdot)), \quad \text{and} \\ & [E(\hat{q}) = \infty \text{ or } e(\cdot, \hat{q}(\cdot)) \geqq e(\cdot, q(\cdot)) \\ & \qquad \text{whenever } q \in \Phi \text{ with } \eta \cdot c(\cdot, q(\cdot)) = \eta \cdot c(\cdot, p_0(\cdot))].\end{aligned}$$

If $N \cap \mathbf{int}\, F \neq \varnothing$, then $\eta = 0^k$ by (6.3), so $E(\hat{q}) \geqq E(q)$ for $q \in \Phi$. Since **rng** E is not bounded above by hypothesis of (6.1), $E(\hat{q}) = \infty$. By Theorem 6.10 (ii), there exists $p \in \Phi$ such that $C(p) = v^0$ and $E(p) = \infty$; such p serves as p^*.

Hence by (6.2), we assume $\varnothing \neq N \cap F \subset \mathbf{bdy}\, F$. We prove (6.1) by induction on k, the number of one-dimensional constraints.

If $k = 1$, $\eta = (\eta_1) \neq (0)$ and $N \cap F$ is a singleton. Since $\eta \cdot C(\hat{q}) = \eta \cdot v^0$, $C(\hat{q}) = v^0$. Suppose $q \in \Phi$, $E(q) > -\infty$, and $C(q) \in N$. Then $C(q) \in N \cap F$, so $C(q) = v^0$. By (6.6) and (6.7), $\eta \cdot c(\cdot, \hat{q}(\cdot)) \leqq \eta \cdot c(\cdot, q(\cdot))$, so since $\eta \cdot C(\hat{q}) = \eta \cdot C(q)$, equality holds in this inequality. Hence, $E(\hat{q}) \geqq E(q)$ by (6.7) and $\hat{q}$ serves as p^*. Thus, (6.1) holds if $k = 1$.

Suppose $k > 1$ and (6.1) holds when there are $k - 1$ one-dimensional constraints. Let

$$\Omega' = \Omega \cap \{(x, y) : \eta \cdot c(x, y) = \eta \cdot c(x, p_0(x))\},$$

$$\Phi' = \Phi \cap \{q : \eta \cdot c(\cdot, q(\cdot)) = \eta \cdot c(\cdot, p_0(\cdot))\},$$

$c' = C|\Omega'$, $e' = e|\Omega'$, $C' = C|\Phi'$, $E' = E|\Phi'$, and $N' = N \cap \{w : \eta \cdot w = \eta \cdot v_0\}$. The hypotheses of (6.1) are satisfied by the primed replacements (were E' bounded above, (ii) would yield p^*), and $C'(q) \in N'$ is expressible as $k - 1$ one-dimensional constraints. By the induction hypothesis, there exists $p^* \in \Phi'$ such that

$$C'(p^*) \in N' \quad \text{and} \quad E'(p^*) = \max \{E'(q) : q \in \Phi' \text{ and } C'(q) \in N'\}.$$

Suppose $q \in \Phi$, $E(q) > -\infty$, and $C(q) \in N$; then $C(q) \in N \cap F$, so by (6.4), $\eta \cdot C(q) = \eta \cdot v^0$; thus, by (6.6), $\eta \cdot c(\cdot, q(\cdot)) = \eta \cdot c(\cdot, p_0(\cdot))$. Hence $q \in \Phi'$. Thus, $E(p^*) = \max \{E(q) : C(q) \in N\}$, proving (6.1) and hence the theorem.

Remark 6.14. We show by two examples, with a variation on each, that Theorem 6.13 and other results of this section cannot be extended in certain ways.

First, let $X = [0, \infty)$, μ be Lebesgue measure, $k = 1$, and for $x \in X$, $Y(x) = \{0, 1\}$, $e(x, 0) = c(x, 0) = 0$, $e(x, 1) = 1$, and $c(x, 1) = e^{-x}$. Then **rng** $C = [0, 1]$, $v(0) = 0$, $v(v) = \infty$ for $0 < v \leqq 1$, and for $0 \leqq v \leqq 1$, there exists $q \in \Phi$ such that $C(q) = v$ and $E(q) = v(v)$. Hence in Theorem 6.10 (iii) we cannot change $v \in \mathbf{int\,rng}\, C$ to $v \in \mathbf{rng}\, C$, since the conclusion fails with $v = 0$. Also, if **rng** E is not bounded above, then v need not be u.s.c., even if optimal functions exist for all costs, **rng** C is compact, and **fin rng** (C, E) is convex.

If we redefine $e(x, 0) = -1$ for $x \in X$, we have $v(0) = -\infty$ and $v(v) = \infty$ for $0 < v \leqq 1$.

For the second example, let $k = 1$, $X = \omega$, and $Y(x) = \{0, 1, 2\}$ and $\mu(\{x\}) = 1$ for $x \in X$. Define c by $c(x, 0) = 0$, $c(x, 1) = 3 \cdot 4^{-x}$, and $c(x, 2) = 4^{1-x}$ for $x \in X$. For $n \in \omega \cup \{\infty\}$, define $q_n \in \Phi$ by, for $x \in X$, $q_n(x) = 1$ if $x < n$, $q_n(x) = 2$ if $x = n$, and $q_n(x) = 0$ if $x > n$. Since $\int_{x \leqq a} c(x, 1)\, d\mu x = 1 - c(a + 1, 2)$ for $a \in \omega$, one may show that for $q \in \Phi$, $C(q) = 1$ iff $q = q_n$ for some $n \in \omega \cup \{\infty\}$.

Define $e(x,0) = 0$, $e(x,1) = -1$, and $e(x,2) = x - 2^{-x}$ for $x \in X$. Then for $n \in \omega$, $E(q_n) = 1 - 2^{-n}$ and $E(q_\infty) = -\infty$. Thus, $v(1) = 1$, but $E(q) < 1$ whenever $C(q) = 1$.

If we alternatively define $e(x,2) = 2x$ for $x \in X$, then $E(q_n) = n + 1$ for $n \in \omega$ and again $E(q_\infty) = -\infty$, whence $v(1) = \infty$, but $E(q) < \infty$ whenever $C(q) = 1$.

Now consider the existence corollaries to Theorems 6.3(ii), 6.9, and 6.13 (ii) formed from Lemma 6.1 with $N = \{v\}$. The example and its alternative show (with $v = 1$) that we cannot substitute either the hypothesis "$v(v) < \infty$" or the hypothesis "$v(v) = \infty$" for the hypothesis "**rng** E is bounded above" in any of these corollaries, nor may we substitute either for the hypothesis "μ is nonatomic or **rng** E is bounded above" appearing in Theorem 6.13 (iii).

The alternative definition of $e(\cdot, 2)$ above also shows that we may not delete the assumption that μ is nonatomic in Theorem 6.10. One can, however, easily show the following: if μ is purely atomic and σ-finite, $E(r) = \infty$ for some $r \in \Phi$, and **rng** C is bounded, then $\{C(p): E(p) = \infty\}$ is dense in **rng** C.

Acknowledgments. We wish to acknowledge some conversations with Professors Hubert Halkin and L. W. Neustadt and a referee's comments, which were extremely helpful in acquainting us with prior literature. As noted in § 4, the new proof of von Neumann's selection theorem given here is due to Professor Herbert Federer and a variation on this proof was provided by Professor J. C. Oxtoby, both communicated privately.

Added in proof. An extension of Theorem 6.13 (ii) above and Theorem 7 of [31] is announced in [44].

REFERENCES

[1] V. I. ARKIN, *An infinite-dimensional analog of non-convex programming problems*, Kibernetika (Kiev), 2 (1967), pp. 87–93; English transl., Cybernetics, 3 (1969), pp. 70–75.

[2] R. J. AUMANN, *Integrals of set-valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1–12.

[3] R. J. AUMANN AND M. PERLES, *A variational problem arising in economics*, J. Math. Anal. Appl., 11 (1965), pp. 488–503.

[4] H. T. BANKS AND M. Q. JACOBS, *The optimization of trajectories of linear functional differential equations*, this Journal, 8 (1970), pp. 461–488.

[5] D. BLACKWELL, *The range of certain vector integrals*, Proc. Amer. Math. Soc., 2 (1951), pp. 390–395.

[6] ———, *A Borel set not containing a graph*, Ann. Math. Statist., 39 (1968), pp. 1345–1347.

[7] C. CASTAING, *Sur les multi-applications mesurables*, Rev. Française Informat., Recherche Opérationnelle, 1 (1967), pp. 91–126.

[8] ———, *Sur les multi-applications mesurables*, Thèse présentée à la Faculté des Sciences de l'Université de Caen, 1967.

[9] G. B. DANTZIG AND A. WALD, *On the fundamental lemma of Neyman and Pearson*, Ann. Math. Statist., 22 (1951), pp. 87–93.

[10] H. EVERETT III, *Generalized Lagrange multiplier methods for solving problems of optimum allocation of resources*, Operations Res., 11 (1963), pp. 399–417.

[11] H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[12] F. J. GOULD AND J. W. TOLLE, *A necessary and sufficient qualification for constrained optimization*, SIAM J. Appl. Math., 20 (1971), pp. 164–172.

[13] B. GRUNBAUM, *Convex Polytopes*, Interscience, New York, 1967.

[14] H. HALKIN, *Some further generalizations of a theorem of Lyapunov*, Arch. Rational Mech. Anal., 17 (1964), pp. 272–277.

[15] ———, *Convexity and control theory*, Functional Analysis and Optimization, E. R. Caianiello, ed., Academic Press, New York, 1966.

[16] H. Halkin and E. C. Hendricks, *Subintegrals of set-valued functions with semianalytic graphs*, Proc. Nat. Acad. Sci. U.S.A., 59 (1968), pp. 365–367.

[17] P. R. Halmos, *Measure Theory*, Van Nostrand, Princeton, N.J., 1954.

[18] C. J. Himmelberg, M. Q. Jacobs and F. S. Van Vleck, *Measurable multifunctions, selectors, and Filippov's implicit functions lemma*, J. Math. Anal. Appl., 25 (1969), pp. 276–284.

[19] H. W. Kuhn and A. W. Tucker, *Nonlinear programming*, Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability, 1951, pp. 481–492.

[20] C. Kuratowski and C. Ryll-Nardzewski, *A general theorem on selectors*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 13 (1965), pp. 397–403.

[21] E. L. Lehmann, *Testing Statistical Hypotheses*, John Wiley, New York, 1959.

[22] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[23] A. Lyapunov, *Sur les fonctions-vecteurs completement additives*, Izv. Akad. Nauk SSSR Ser. Mat., 4 (1940), pp. 465–478.

[24] E. J. McShane and R. B. Warfield, Jr., *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.

[25] H. D. Meeks, *Duality relationships for a nonlinear version of the generalized Neyman–Pearson problem*, Ph.D. dissertation, Ohio State Univ., Columbus, 1970.

[26] H. D. Meeks and R. L. Francis, *Duality relationships for a non-linear version of the generalized Neyman–Pearson problem*, J. Optimization Theory Appl., 11 (1973), pp. 360–378.

[27] L. W. Neustadt, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

[28] C. Olech, *Extremal solutions of a control system*, J. Differential Equations, 2 (1966), pp. 74–101.

[29] ———, *Lexicographical order, range of integrals, and "bang-bang" principle*, Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967.

[30] ———, *On the range of an unbounded vector-valued measure*, Math. Systems Theory, 2 (1968), pp. 251–256.

[31] ———, *Integrals of set valued functions and linear optimal control problems*, IFAC, Fourth Congress of the International Federation of Automatic Control, Warsaw, 1969, pp. 22–35.

[32] ———, *Integrals of set-valued functions and linear optimal control problems*, Proc. of the Colloquium on Optimal Control Theory, Brussels, 1969, pp. 109–125.

[33] T. Parthasarathy, *Selection Theorems and their Applications*, Springer-Verlag, New York, 1972.

[34] H. Richter, *Verallgemeinerung eines in der Statistik benötigten Satzes der Masstheorie*, Math. Ann., 150 (1963), pp. 85–90.

[35] S. Saks, *Theory of the Integral*, Dover, New York, 1964.

[36] M. Slater, *Lagrange multipliers revised: A contribution to non-linear programming*, Cowles Commission Discussion Paper, Math. 403, November, 1950.

[37] L. D. Stone, *Total optimality of incrementally optimal allocations*, Naval Res. Logist. Quart., 20 (1973), pp. 419–430.

[38] ———, *Incremental and total optimization of separable functionals with constraints*, submitted for publication.

[39] J. J. Uhl, Jr., *The range of a vector-valued measure*, Proc. Amer. Math. Soc., 23 (1969), pp. 158–163.

[40] J. von Neumann, *On rings of operators. Reduction theory*, Ann. of Math., 50 (1949), pp. 401–485.

[41] D. H. Wagner, *Nonlinear functional versions of the Neyman–Pearson lemma*, SIAM Rev., 11 (1969), pp. 52–65.

[42] D. H. Wagner and L. D. Stone, *Optimization of allocations under a coverability condition*, this Journal, 12 (1974), pp. 373–379.

[43] S. Zahl, *An allocation problem with applications to operations research and statistics*, Operations Res., 11 (1963), pp. 426–441.

[44] D. H. Wagner, *The integral of a lower closed set-valued function*, Abstract 73T-B259, Notices Amer. Math. Soc., 20 (Oct., 1973).

SIAM J. Control
Vol. 12, No. 3, August 1974

OPTIMIZATION OF ALLOCATIONS UNDER A COVERABILITY CONDITION*

DANIEL H. WAGNER AND LAWRENCE D. STONE†

Abstract. Consider maximization of a real functional E given by $E(q) = \int_X e(x, q(x))\, d\mu x$, subject to equality or inequality constraint on $\int_X q\, d\mu$. It is proved that such extrema exist and necessitate satisfaction of a pointwise multiplier rule without assuming a topology on X, but assuming a condition called coverability of e, pertaining to the concave envelope of $e(x, \cdot)$, $x \in X$. Examples show that key hypotheses may not be omitted.

This paper relates to the results in [6] on the necessity of a pointwise multiplier rule for constrained maximization of a separable functional and on existence of such extrema. A simple constraint functional is used, but the objective functional is subjected to a "coverability" condition defined below in terms of concave envelopes. No Borel assumptions are made, in contrast to Corollary 5.2 and Theorem 6.13 of [6]; in further contrast to the latter, our assumption below differs from $Z(x)$ being upper closed. Necessity results are given as Theorem 5 and Corollary 6, and existence is given in Theorem 8. Remarks 7 and 9 show by examples that the results cannot be strengthened in various ways.

The usages below are consistent with [6], and are in some ways simpler. We fix an arbitrary nonvacuous set X (on which no topology is assumed). For $x \in X$, let $Y(x) \neq \emptyset$ be a real interval (not necessarily bounded or closed). For $x \in X$, define $T(x) = \inf Y(x)$ and $U(x) = \sup Y(x)$. Assume T and U are measurable (extended real-valued) functions. Let $\Omega = \{(x, y): x \in X \text{ and } y \in Y(x)\}$. Fix a measure μ over X, to which measurability and integrability refer unless stated otherwise. We ignore subsets of X having μ measure 0, e.g., "for $x \in X$" means "for μ a.e. $x \in X$." Denoting one-dimensional Lebesgue measure by $\mathscr{L}$, we use the product measure $\mu \times \mathscr{L}$ on Ω.

Let ω be the positive integers and $\mathscr{E}_1$ be the reals. Fixing $e: \Omega \to \mathscr{E}_1$, define

$$\Xi = \{q : q(x) \in Y(x) \text{ for } x \in X \text{ and } e(\cdot, q(\cdot)) \text{ and } q \text{ are measurable}\},$$

$$\Phi = \Xi \cap \{q : e(\cdot, q(\cdot)) \text{ and } q \text{ are integrable}\},$$

$$E(q) = \int_X e(x, q(x))\, d\mu x \quad \text{and} \quad C(q) = \int_X q\, d\mu \quad \text{for } q \in \Phi.$$

We say $q^* \in \Phi$ is optimal (strongly optimal) if

$$E(q^*) = \max\{E(q) : C(q) = C(q^*)\} \quad (E(q^*) = \max\{E(q) : C(q) \leqq C(q^*)\}).$$

Let $\lambda \in \mathscr{E}_1$. Define $l_\lambda(x, y) = e(x, y) - \lambda y$ for $(x, y) \in \Omega$. For $q \in \Phi$, if $\lambda = 0$ and $C(q) = \pm\infty$, define $L_\lambda(q) = E(q)$; otherwise, define $L_\lambda(q) = E(q) - \lambda C(q)$ if this exists. For $q^* \in \Phi$, we say (q^*, λ) *satisfies the functional multiplier rule* if

* Received by the editors February 15, 1972, and in revised form April 24, 1973. This work was supported in part by the Naval Analysis Programs, Office of Naval Research, under Contract N00014-69-C-0435.

† Daniel H. Wagner, Associates, Paoli, Pennsylvania 19301.

$L_\lambda(q^*) \geqq L_\lambda(q)$ whenever $L_\lambda(q)$ exists; (q^*, λ) *satisfies the pointwise multiplier rule* if $l_\lambda(x, q^*(x)) \geqq l_\lambda(x, y)$ for $y \in Y(x)$, $x \in X$. If either rule is satisfied and $\lambda \geqq 0$, we say it is *strongly* satisfied.

Differentiation is always with respect to the last component of the argument and is always one-sided. A superscript $+$ or $-$ on a function denotes right or left derivative respectively, e.g., $e^+(x, y) = \lim_{\delta \downarrow 0}[e(x, y + \delta) - e(x, y)]/\delta$.

Suppose F is a real interval, $f : F \to \mathscr{E}_1$, and for t, $u \in F$ and $0 \leqq \alpha \leqq 1$, $f(\alpha t + (1 - \alpha)u) \geqq \alpha f(t) + (1 - \alpha)f(u)$. We then say f is *concave*. On interior F, f is continuous, f^+ and f^- both exist, and except on a countable set $f^+ = f^-$; at endpoints of F, if any, f might be discontinuous. We say $y \in F$ is an *extremizing point* of f if there are no $z, w \in F$ and $0 < \beta < 1$ such that $z < w$ and $(y, f(y)) = \beta(z, f(z)) + (1 - \beta)(w, f(w))$.

If h and g are real-valued functions on F, we say that g is the *concave envelope* of h if (a) g is concave and continuous, (b) $g(y) \geqq h(y)$ for $y \in F$, and (c) whenever $\hat{g}$ is a continuous concave function on F such that $\hat{g}(y) \geqq h(y)$ for $y \in F$, we have $g(y) \leqq \hat{g}(y)$ for $y \in F$.

Assumption 1. Throughout we assume that e is *covered* by m, i.e., e is *coverable*, meaning that the following conditions are satisfied:

(i) Ω is $\mu \times \mathscr{L}$ measurable and $m : \Omega \to \mathscr{E}_1$ is a $\mu \times \mathscr{L}$ measurable function;

(ii) for $x \in X$, $m(x, \cdot)$ is the concave envelope of $e(x, \cdot)$;

(iii) for $x \in X$, $m(x, y) = e(x, y)$ whenever y is an extremizing point of $m(x, \cdot)$ (equivalently, $e(x, \cdot)$ is upper semicontinuous at each such y);

(iv) for $x \in X$ and $y \in Y(x)$, there exist extremizing points w and z of $m(x, \cdot)$ such that $w \leqq y \leqq z$.

For $(x, y) \in \Omega$, we define

$$\psi(x, y) = \inf\{z : z \geqq y \text{ and } z \text{ is an extremizing point of } m(x, \cdot)\}.$$

The following lemma, stated without proof, gives equivalent conditions on a point b, extendable to a function q on X by regarding $b = q(x)$ for a particular $x \in X$. As such, they provide alternative formulations of the pointwise multiplier rule; (iii) corresponds to a functional Neyman–Pearson condition as in [7].

LEMMA 2. *Let F be a real interval, $h : F \to \mathscr{E}_1$, g be the concave envelope of h, $b \in F$, and $\lambda \in \mathscr{E}_1$. Then the following six conditions are equivalent*:

(i) $h(b) - \lambda b = \max\{h(y) - \lambda y : y \in F\}$;

(ii) $h(b) = g(b)$ *and* $g(b) - \lambda b = \max\{g(y) - \lambda y : y \in F\}$;

(iii) $h(b) = g(b)$, $g^+(y) \geqq \lambda$ *for* $b > y \in F$, *and* $g^-(y) \leqq \lambda$ *for* $b < y \in F$;

(iv) $h(b) = g(b)$, $g^+(b) \leqq \lambda$ *if* $b < \sup F$, *and* $\lambda \leqq g^-(b)$ *if* $b > \inf F$;

(v) $h(b) = g(b)$ *and not* $[g^+(b) > \lambda$ *or* $g^-(b) < \lambda]$;

(vi) *the graph of h lies on or below the line through $(b, h(b))$ with slope λ.*

LEMMA 3. *For* $(x, y) \in \Omega$, (i) $\psi(x, y) \in Y(x)$ *and* (ii) $\psi(x, y)$ *is an extremizing point of* $m(x, \cdot)$.

Proof. From (iv) in Assumption 1, we obtain (i). If $\psi(x, y)$ were not an extremizing point of $m(x, \cdot)$, we could obtain $z, w \in Y(x)$ with $z < \psi(x, y) < w$ and $(\psi(x, y), m(x, \psi(x, y)))$ lying on the chord joining $(z, m(x, z))$ and $(w, m(x, w))$; but the interior of this chord must contain extremizing points of $m(x, \cdot)$ (by definition of ψ), in contradiction. Hence (ii) holds.

LEMMA 4. *Suppose $P \subset X$ is measurable, p is a measurable function on P, and $p(x) \in Y(x)$ for $x \in P$. Then:*

(i) *if we define $g(x, y) = p(x)$ for $x \in P$ and $y \in Y(x)$, then g is a $\mu \times \mathscr{L}$ measurable function;*

(ii) *if $q: P \to \mathscr{E}_1$, $A \equiv (P \times \mathscr{E}_1) \cap \{(x, y): y \leqq q(x)\}$ is $\mu \times \mathscr{L}$ measurable, and μ is σ-finite on P, then q is a measurable function;*

(iii) *m^+ and m^- are $\mu \times \mathscr{L}$ measurable functions;*

(iv) *if μ is σ-finite on P, then $m(\cdot, p(\cdot))$, $m^+(\cdot, p(\cdot))$, $m^-(\cdot, p(\cdot))$, and $\psi(\cdot, p(\cdot))$ are measurable functions.*

Proof. To prove (i), note that for $a \in \mathscr{E}_1$,

$$\{(x, y): g(x, y) > a\} = \Omega \cap [\{x: p(x) > a\} \times \mathscr{E}_1],$$

which is $\mu \times \mathscr{L}$ measurable. To prove (iii), define $g_0(x, y) = U(x), x \in X$. Then m^+ is defined on

$$\Omega \cap \{(x, y): y < U(x)\} = \Omega - (X \times \mathscr{E}_1) \cap \{(x, y): g_0(x, y) - y = 0\}$$

which is $\mu \times \mathscr{L}$ measurable by (i). Let $\delta \in \mathscr{E}_1$, and for $(x, y) \in \Omega$, let $\tau_\delta(x, y) = (x, y - \delta)$ and $m_\delta(x, y - \delta) = m(x, y)$. Then $\mu \times \mathscr{L}$ is invariant under τ_δ and for $a \in \mathscr{E}_1$, $\{(x, y): m_\delta(x, y) \geqq a\} = \tau_\delta(\{(x, y): m(x, y) \geqq a\})$. Thus, m_δ is $\mu \times \mathscr{L}$ measurable, so is m^+, being a limit of such, and similarly so is m^-, proving (iii).

Proof of (ii) primarily follows [3, exercises (5e) and (5f), § 34]. Suppose $a \in \mathscr{E}_1$. One may show that (under the measure foundations of [2] or [3]) since A is $\mu \times \mathscr{L}$ measurable, so is $D_n \equiv \{(x, y): (x, a + y/n) \in A\}$ for $n \in \omega$. Let $B = P \cap \{x: q(x) > a\}$. Then

$$B \times (0, 1] = \bigcup_{n=1}^{\infty} \{(x, y): x \in P, a + y/n \leqq q(x), \text{and } 0 < y \leqq 1\}$$

$$= \bigcup_{n=1}^{\infty} [D_n \cap (P \times (0, 1])].$$

Hence, $B \times (0, 1]$ is $\mu \times \mathscr{L}$ measurable. With μ σ-finite on P, Fubini's theorem applied to the indicator function of $B \times (0, 1]$ shows B is measurable, proving (ii).

To prove (iv), let $b \in \mathscr{E}_1$, $K = \{(x, y): m(x, y) > b \text{ and } x \in P\}$, and

$$r(x) = \inf\{y: (x, y) \in K \text{ or } m^-(x, y) < 0\},$$

$$s(x) = \sup\{y: (x, y) \in K \text{ or } m^+(x, y) > 0\} \quad \text{for } x \in P.$$

Since $m(x, \cdot)$ is concave and continuous for $x \in X$,

$$(P \times \mathscr{E}_1) \cap \{(x, y): y \leqq r(x)\}$$
$$= (P \times \mathscr{E}_1) \cap \{(x, y): [m(x, y) \leqq b \text{ and } m^-(x, y) \geqq 0] \text{ or } y \leqq T(x)\},$$

hence by (i), (iii), and (ii), r is a measurable function, and similarly so is s. Also, $\{x: m(x, p(x)) > b\} = \{x: r(x) < p(x) < s(x)\}$, so $m(\cdot, p(\cdot))$ is measurable.

For $\delta \geqq 0$, $p + \delta$ is a measurable function and by what we have just proved, so is $m(\cdot, p(\cdot) + \delta)$; hence, $m^+(\cdot, p(\cdot))$ is measurable and similarly so is $m^-(\cdot, p(\cdot))$. Also,

$$\{(x, y): p(x) \leqq y \leqq \psi(x, p(x))\}$$
$$= \{(x, y): [p(x) = y \text{ and not } m^-(x, p(x)) = m^+(x, p(x))] \text{ or}$$

$$[p(x) \leqq y \in Y(x),\ m(x,y) = m(x,p(x)) + [y - p(x)]m^+(x,p(x)),$$
$$\text{and } m^-(x,p(x)) = m^+(x,p(x))]\},$$

which is $\mu \times \mathscr{L}$ measurable by (i) and the proven part of (iv). Thus, $\psi(\cdot, p(\cdot))$ is measurable by (i) of Lemma 3, (i), and (ii), which completes the proof.

THEOREM 5. *Suppose $q^* \in \Phi$, $\lambda \in \mathscr{E}_1$, $|E(q^*)| < \infty$, $|C(q^*)| < \infty$, and μ has finite substance* (see [6, § 2]). *Then for* (q^*, λ) *to satisfy* (*strongly satisfy*) *the functional multiplier rule, it is necessary and sufficient that* (q^*, λ) *satisfy* (*strongly satisfy*) *the pointwise multiplier rule.*

Proof. Sufficiency follows from Theorem 2.1 of [6].

Suppose (q^*, λ) satisfies the functional multiplier rule. Let

$$S = \{x : m^+(x, q^*(x)) > \lambda\} \quad \text{and} \quad R = \{x : m^-(x, q^*(x)) < \lambda\}.$$

By Lemma 4 (iv), S and R are measurable. We shall show $\mu(S) = \mu(R) = 0$.

Supposing $\mu(S) > 0$, choose a measurable $P \subset S$ such that $0 < \mu(P) < \infty$ and suppose $x \in P$ (whence $q^*(x) < U(x)$ since $m^+(x, q^*(x))$ exists). Let

$$p(x) = \inf\{y : m^+(x,y) < \lambda \text{ or } [U(x) = \infty \text{ and } y = \psi(x, q^*(x) + 1)]$$
$$\text{or } [U(x) < \infty \text{ and } y = \psi(x, q^*(x) + \tfrac{1}{2}[U(x) - q^*(x)])]\}.$$

By Lemma 3(i), $q^*(x) < p(x) \in Y(x)$. If $p(x) = \inf\{y : m^+(x,y) < \lambda\}$, it is easily shown that $p(x)$ is an extremizing point of $m(x,\cdot)$, and otherwise this holds by Lemma 3(ii). Thus, $m(x, p(x)) = e(x, p(x))$ by Assumption 1(iii). Also, $m^+(x, y) \geqq \lambda$ for $q^*(x) \leqq y < p(x)$, and there exists z such that $q^*(x) < z < p(x)$ and $m^+(x, y) > \lambda$ for $q^*(x) \leqq y \leqq z$.

By Lemma 4(i), (iii), and (iv), $\{(x, y) : q^*(x) \leqq y \leqq p(x)\}$ is $\mu \times \mathscr{L}$ measurable. Hence by Lemma 4(i), (ii), and (iv), p and $m(\cdot, p(\cdot))$ are measurable. Thus for $x \in P$, since $m(x, \cdot)$ is absolutely continuous on $[q^*(x), p(x)]$ by § 6.3 of [4],

$$\begin{aligned} l_\lambda(x, p(x)) &= m(x, p(x)) - \lambda p(x) \\ &= \int_{q^*(x)}^{p(x)} [m^+(x,y) - \lambda]\, dy + m(x, q^*(x)) - \lambda q^*(x) \qquad (1) \\ &> m(x, q^*(x)) - \lambda q^*(x) \geqq l_\lambda(x, q^*(x)). \end{aligned}$$

Define $q(x) = p(x)$ for $x \in P$ and $q(x) = q^*(x)$ for $x \in X - P$. Then $q \in \Xi$ and by (1) and Theorem 2.2((i) implies (ii)) of [6], (q^*, λ) does not satisfy the functional multiplier rule, in contradiction. Thus $\mu(S) = 0$. Similarly (symmetrizing the definition of ψ and the statement of Lemmas 3 and 4), $\mu(R) = 0$. Thus

$$\text{(2)} \qquad \text{not } [m^+(x, q^*(x)) > \lambda \text{ or } \lambda > m^-(x, q^*(x))] \quad \text{for } x \in X.$$

Let $Q = \{x : m(x, q^*(x)) > e(x, q^*(x))\}$. Then Q is measurable. For $x \in Q$, $q^*(x)$ is not an extremizing point of $m(x, \cdot)$, so by (2), $m^+(x, q^*(x)) = m^-(x, q^*(x)) = \lambda$. Define $q_0(x) = \psi(x, q^*(x))$ for $x \in Q$ and $q_0(x) = q^*(x)$ for $x \in X - Q$. Then for $x \in Q$, $q_0(x) \in Y(x)$ and $m(x, q_0(x)) = e(x, q_0(x))$ by Lemma 3, so

$$\begin{aligned} l_\lambda(x, q_0(x)) &= m(x, q_0(x)) - \lambda q_0(x) \\ &= m(x, q^*(x)) + [q_0(x) - q^*(x)]m^+(x, q^*(x)) - \lambda q_0(x) \\ &= m(x, q^*(x)) - \lambda q^*(x) > e(x, q^*(x)) - \lambda q^*(x) = l_\lambda(x, q^*(x)). \end{aligned}$$

Also, $q_0 \in \Xi$ by Lemma 4(iv). Hence by Theorem 2.2((i) implies (ii)) of [6], $\mu(Q) = 0$, so $m(\cdot, q^*(\cdot)) = e(\cdot, q^*(\cdot))$. From this, (2), and Lemma 2((v) implies (i)), we have $l_\lambda(x, q^*(x)) \geqq l_\lambda(x, y)$ for $y \in Y(x)$, $x \in X$, proving the theorem.

COROLLARY 6. *If the hypothesis of Theorem 5 holds, μ is nonatomic, and $C(T) < C(q^*) < C(U)$, then for q^* to be optimal (strongly optimal), it is necessary and sufficient that for some $\lambda \in \mathscr{E}_1$, (q^*, λ) satisfy (strongly satisfy) the pointwise multiplier rule.*

Proof. This follows from Corollary 3.3 of [6] and Theorem 5.

Remark 7. We show by examples that Theorem 5 and Corollary 6 cannot be strengthened in certain ways. Referring to the example of Remark 5.5 of [6], let

$$m(x, y) = 1 \quad \text{if } |y| \leqq 1 \quad \text{and} \quad m(x, y) = 2 - |y| \quad \text{if } 1 \leqq |y| \leqq 2 \qquad \text{for } (x, y) \in \Omega;$$

then $e(x, \cdot)$ has the concave envelope $m(x, \cdot)$ (whose graph is an isoceles trapezoid) for $x \in X$, and m is continuous. However, e is not coverable, since $m(x, 1) = m(x - 1) = 1 \neq 0 = e(x, 1) = e(x, -1)$ for $x \in X$, so (iii) of Assumption 1 fails. All other hypotheses of Theorem 5 are satisfied, but as noted in [6], the necessity conclusions fail. Therefore, in Theorem 5 one may not replace (iii) of Assumption 1 by the condition that e is measurable and m is continuous. By redefining $e(x, -1) = e(x, 1) = 1$ for $x \in X$, e becomes coverable, and Theorem 5 applies. Incidentally, if $r(x) = \frac{1}{2}$ for $x \in X$, then $e(\cdot, r(\cdot))$ is not measurable, but $m(\cdot, r(\cdot))$ and r are integrable.

To see that Theorem 5 fails if (iv) is omitted from Assumption 1, in Remark 5.5 of [6] redefine Ω to be $[0, 1] \times \mathscr{E}_1$ and $\mathscr{A}$ to be

$$\{(x, y) : [x \in D \text{ and } |y| = 2i] \text{ or } [x \in X - D \text{ and } |y| = 2i + 1] \text{ for some } i \in \omega\} \\ - X \times \{0\},$$

let $e(x, 0) = \frac{1}{2}$ for $x \in X$, elsewhere on Ω let e be the indicator function of $\mathscr{A}$, and proceed as before.

In Theorem 5 we assume $\mu \times \mathscr{L}$ measurability of m but not of e. To see that the latter would not insure the former, redefine $\mathscr{A} = D \times \{0\}$ and let e be the indicator function of $\mathscr{A}$.

THEOREM 8. *Suppose* (i) *$Y(x)$ is compact for $x \in X$,* (ii) *μ is nonatomic,* (iii) *$T \in \Phi$, $U \in \Phi$,* (iv) *$-\infty < C(T) \leqq C(U) < \infty$, and* (v) *$C(T) \leqq v \leqq C(U)$. Then there exist an optimal $q^* \in \Phi$ such that $C(q^*) = v$ and $p^* \in \Phi$ such that $C(p^*) \leqq v$ and $E(p^*) = \max\{E(p) : C(p) \leqq v\}$.*

Proof. For $x \in X$ and $\lambda \in \mathscr{E}_1$, define

$$\varphi_u(x, \lambda) = \sup\{y : y = T(x) \text{ or } m^+(x, y) \geqq \lambda\},$$

$$\varphi_l(x, \lambda) = \inf\{y : y = U(x) \text{ or } m^+(x, y) \leqq \lambda\}.$$

Then by (i),

$$(3) \qquad -\infty < T(x) \leqq \varphi_l(x, \lambda) \leqq \varphi_u(x, \lambda) \leqq U(x) < \infty \quad \text{for } x \in X, \quad \lambda \in \mathscr{E}_1.$$

Suppose $\lambda \in \mathscr{E}_1$. By (iv), μ is σ-finite over $\{x : U(x) > T(x)\}$. We have

$$\{(x, y) : y \leqq \varphi_u(x, \lambda)\} = \{(x, y) : y \leqq T(x) \text{ or } m^+(x, y) \geqq \lambda\},$$

which is $\mu \times \mathscr{L}$ measurable by Lemma 4(iii) and (i), so by Lemma 4(ii), $\varphi_u(\cdot, \lambda)$ is measurable and, by (iii), (iv), and (3), integrable; similarly, so is $\varphi_l(\cdot, \lambda)$. Define

$$I_u(\lambda) = \int_X \varphi_u(x, \lambda)\, d\mu x \quad \text{and} \quad I_l(\lambda) = \int_X \varphi_l(x, \lambda)\, d\mu x.$$

For $x \in X$, $\varphi_l(x, \cdot)$ is right continuous and $\varphi_u(x, \cdot)$ is left continuous. Thus, I_l is right continuous and I_u is left continuous by the monotone convergence theorem. By (3),

$$C(T) \leqq I_l(\lambda) \leqq I_u(\lambda) \leqq C(U). \tag{4}$$

Since $\lim_{\lambda \to \infty} \varphi_u(x, \lambda) = T(x)$ for $x \in X$, by the dominated convergence theorem and (4) we have $\lim_{\lambda \to \infty} I_u(\lambda) = C(T)$, hence $\lim_{\lambda \to \infty} I_l(\lambda) = C(T)$. Similarly, $\lim_{\lambda \to -\infty} I_l(\lambda) = \lim_{\lambda \to -\infty} I_u(\lambda) = C(U)$.

Obviously, if $v = C(T)$ or $v = C(U)$, then T or U would respectively serve for q^*. Therefore, we assume that $C(T) < v < C(U)$. There exists a λ_0 such that

$$I_u(\lambda_0) = \lim_{\lambda \uparrow \lambda_0} I_u(\lambda) \geqq v \geqq \lim_{\lambda \downarrow \lambda_0} I_u(\lambda).$$

By the right continuity of I_l, $I_l(\lambda_0) \leqq v$ so for some $0 \leqq \alpha \leqq 1$, $v = \alpha I_u(\lambda_0) + (1 - \alpha) I_l(\lambda_0)$.

By Theorem 3.1 and Remark 3.5 of [6], (iv), and (3), we obtain a measurable $P \subset X$ and an integrable q^* such that $q^*(x) = \varphi_u(x, \lambda_0)$ for $x \in P$, $q^*(x) = \varphi_l(x, \lambda_0)$ for $x \in X - P$, and

$$C(q^*(x)) = \alpha C(\varphi_u(\cdot, \lambda_0)) + (1 - \alpha) C(\varphi_l(\cdot, \lambda_0)) = \alpha I_u(\lambda_0) + (1 - \alpha) I_l(\lambda_0) = v.$$

For $x \in X$, by the definitions of φ_l and φ_u, $q^*(x)$ is an extremizing point of $m(x, \cdot)$; thus by Assumption 1 (iii) and the same definitions,

$$e(x, q^*(x)) = m(x, q^*(x)) \text{ and not } [m^+(x, q^*(x)) > \lambda_0 \text{ or } m^-(x, q^*(x)) < \lambda_0]. \tag{5}$$

By Lemma 4 (iv), $m(\cdot, q^*(\cdot))$, i.e., $e(\cdot, q^*(\cdot))$, is a measurable function. We may assume there exists $\hat{q} \in \Phi$ such that $E(\hat{q}) > -\infty$. By virtue of (iv), $l_{\lambda_0}(\cdot, \hat{q}(\cdot))$ is integrable and $L_{\lambda_0}(\hat{q}) > -\infty$. Thus, $l_{\lambda_0}(\cdot, q^*(\cdot))$ is integrable since it dominates $l_{\lambda_0}(\cdot, \hat{q}(\cdot))$. Hence by (iv), $e(\cdot, q^*(\cdot))$ is integrable, i.e., $q^* \in \Phi$. Since $-\infty < L_{\lambda_0}(\hat{q}) \leqq L_{\lambda_0}(q^*)$, $E(q^*) > -\infty$. Hence by (5), Lemma (2) ((v) implies (i)), and Theorem 2.1 of [6], q^* is optimal, as desired.

To obtain p^* define $s(x) = \inf\{y : \text{not } [m^-(x, y) < 0 \text{ or } 0 < m^+(x, y)]\}$ for $x \in X$. Then $m(x, s(x)) \geqq m(x, y)$ for $y \in Y(x)$, $x \in X$, and $m(\cdot, s(\cdot)) = e(\cdot, s(\cdot))$ by (iii) of Assumption 1. It is easily shown by Lemma 4 (iii) and (ii) that s is measurable, whence by Lemma 4 (iv), so is $m(\cdot, s(\cdot))$. Since $m(\cdot, s(\cdot)) \geqq e(\cdot, \hat{q}(\cdot))$ and $E(\hat{q})$. $> -\infty$, $m(\cdot, s(\cdot))$ is integrable. Thus, $E(s) = \int_X m(x, s(x))\, d\mu x \geqq E(q)$ for $q \in \Phi$.

Now let $a = \sup\{E(q) : C(q) \leqq v\}$, $v(w) = \sup\{E(q) : C(q) = w\}$ for $w \in$ range C, and $F =$ range $C \cap \{w : v(w) > -\infty\}$. By Theorem 3.1 and Remark 3.5 of [6], F and $\mathscr{E}_2 \cap$ range (C, E) are convex. We have shown that for $w \in F$, $w = C(q)$ and $v(w) = E(q)$ for some q. It follows from Theorem 6.10 of [6] that v is concave on interior F or ∞ on this set.

If $v \geqq \sup F$, s serves as p^*; $v \leqq \inf F$ is trivial. We may assume v is in interior F, on which v is concave. Choose $p, q_1, q_2, \cdots \in \Phi$ such that $C(q_j) \leqq v$ for $j \in \omega$, $E(q_j) \to a$ and $C(q_j) \to C(p) \leqq v$. If $C(p) > \inf F$, then v is continuous at $C(p)$, so $a = v(C(p))$ and, as we have shown, there exists p^* such that $E(p^*) = v(C(p)) = a$ and $C(p^*) = C(p)$. If $C(p) = \inf F$, since $a \geqq v(v)$ and v is concave, we have $v^-(v) \leqq 0$ whence $C(s) \leqq v$. Again, s serves as p^*, which completes the proof.

Remark 9. In [5], Stone provides an existence result, Theorem 3.3, under conditions different from those of Theorem 8; he replaces the condition $C(U) < \infty$ (thereby permitting $U = \infty$) by the condition that $E(T)$ and $E(U)$ are finite and $e(x, \cdot)$ is nondecreasing for $x \in X$. Example 3.4 of [5] shows that neither $C(U) < \infty$ nor $C(T) > -\infty$ may be omitted from Theorem 8.

The following example demonstrates that in Theorem 8, Assumption 1 (iii) may not be replaced by the condition that e is a Borel function and $e(x, \cdot)$ is nondecreasing for $x \in X$. Let $X = [0, 1]$, and for $x \in X$, let $T(x) = 0$ and $U(x) = 2$. For $(x, y) \in \Omega$, let $e(x, y) = 0$ for $0 \leqq y \leqq 1$ and $e(x, y) = 1$ for $1 < y \leqq 2$. Note that if $C(q) = 1$, then $E(q) < 1$. However, $\sup \{E(q) : C(q) = 1\} = 1$.

Defining concave envelopes without requiring continuity at endpoints would invalidate Theorem 8. To see this, let $\Omega = [0, 1] \times [0, 1]$, and for $(\mathrm{x}, \mathrm{y}) \in \Omega$, let $e(x, y) = y^2$ and $m(x, y) = y$ for $y \in (0, 1]$ and $e(x, 0) = m(x, 0) = -1$. Then m would cover e since the only extremizing points of $m(x, \cdot)$ are 0 and 1, for $x \in [0, 1]$. However, there is no optimal $q^* \in \Phi$ such that $\mathrm{C}(q^*) = \frac{1}{2}$.

Remark 10. Our use of concave envelopes was originally motivated in part by Arkin [1]. In [1], $e(x, \cdot)$ is assumed to be a probability distribution function. However, we point out that the proof of necessity in [1] for the case where $e(x, \cdot)$ is not concave is incomplete in that it is shown merely that (our notation) if q^* is optimal with respect to E and C and (defining $M(q) = \int_X m(x, q(x))\, d\mu x$ when this exists) *if* there exists a function r^* which is optimal with respect to M and C with $C(r^*) = C(q^*)$, then there is an r^{**} such that $C(r^{**}) = C(r^*) = C(q^*)$ and $M(r^*) = M(r^{**}) = E(r^{**}) = E(q^*)$, and hence that r^{**} is optimal with respect to E and C. It remains to show that such r^* exists and that $q^*(x)$ maximizes a Lagrangian for $x \in X$.

REFERENCES

[1] V. L. Arkin, *Uniformly optimal strategies in search problems*, Theor. Probability Appl., 2 (1964), pp. 674–680.

[2] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[3] P. R. Halmos, *Measure Theory*, Van Nostrand, Princeton, N.J., 1954.

[4] G. H. Hardy, J. E. Littlewood and G. Pólya, *Inequalities*, Cambridge Univ. Press, Cambridge, 1964.

[5] L. D. Stone, *Total optimality of incrementally optimal allocations*, Naval Res. Logist. Quart., 20 (September, 1973).

[6] D. H. Wagner and L. D. Stone, *Necessity and existence results on constrained optimization of separable functionals by a multiplier rule*, this Journal, 12 (1974), pp. 356–372.

[7] D. H. Wagner, *Nonlinear functional versions of the Neyman–Pearson lemma*, SIAM Rev., 11 (1969), pp. 52–65.

SIAM J. Control
Vol. 12, No. 3, August 1974

SMALL NOISE OPEN LOOP CONTROL*

CHARLES J. HOLLAND†

Abstract. A truncated expansion of the optimal cost in powers of the noise coefficient is established for a class of fixed stopping time small noise open loop control problems. These problems arise by adding a white noise term with a small coefficient to the system equations in the deterministic control problem of Pontryagin.

1. Introduction. In this paper we establish a truncated expansion (Theorem 4.3) of the optimal cost in powers of the noise coefficient for a class of fixed stopping time small noise open loop control problems. These problems arise by adding an additive white noise term with a small noise coefficient to the system equations in the deterministic control problem. The theorem shows that if the noise coefficient is small, then the optimal open loop deterministic control U^0 is approximately optimal in the stochastic problem. The theorem also gives an approximation to both the optimal open loop stochastic cost and the cost of using U^0 in the stochastic problem. See the remarks following the proof of Theorem 4.3 for additional interpretations. Fleming established the corresponding result for the completely observable case in [2], and his work has influenced the approach taken here although our methods of proof are quite different.

In [2] Fleming was able to establish an expansion for the optimal feedback control. We were unable to establish an expansion for the open loop control for the class of problems considered here. In [9], however, such an expansion was established under different assumptions, among which is the one that each open loop control generate a Gaussian process.

Other approaches to the open loop control problem include the stochastic programming work of Wets and Van Slyke [12] and the work of Mortensen [11].

2. The problem. Suppose that the state $\xi(t)$ evolves according to the stochastic differential equations

$$d\xi = f(t, \xi(t), U(t))\,dt + \sigma(t)\,dw, \tag{1}$$

where w is an n-dimensional Brownian motion, and with initial condition $\xi(s_0) = x_0$, x_0 a constant in R^n. In (1), $U(t)$ is a control with values in the control set K. We seek to minimize

$$J(U) = E\left\{\int_{s_0}^{T} L(t, \xi(t), U(t))\,dt \,|\, \xi(s_0) = x_0\right\} \tag{2}$$

over the class of open loop controls $\mathscr{U}$. An *open loop control* $U \in \mathscr{U}$ is a Borel measurable function on $[s_0, T]$ with values in K.

Let $Q = [s_0, T] \times R^n$. Throughout we assume the following:

(i) The initial point (s_0, x_0) is a fixed constant in R^{n+1}, is known to the controller, and is a regular point.

* Received by the editors January 25, 1973, and in revised form April 16, 1973.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 47906. This research was supported in part by the National Science Foundation and in part by the Division of Applied Mathematics at Brown University.

(3)
(ii) K is a compact, convex subset of R^k.
(iii) $f(t, x, u) = A(t, x) + B(t, x)u$.
(iv) The functions f, L are C^∞-functions on $Q \times K$.
(v) f, L and their first, second and third order partial derivatives are bounded on $Q \times K$.
(vi) There exists $C_0 > 0$ such that $v'L_{uu}(s, x, u)v \geqq C_0 v'v$ for all nonzero $v \in R^k$ and for all $(s, x, u) \in Q \times K$.
(vii) $\sigma(t) = (2\varepsilon)^{1/2} I$, where I is the $n \times n$ identity matrix.

We call (s_0, x_0) a *regular point* if there exists a unique optimal open loop control for the deterministic control problem (1), (2) with $\varepsilon = 0$.

For each $\varepsilon \geqq 0$ in (1), let $\lambda^\varepsilon = \inf_{U \in \mathscr{U}} J(U)$. We shall show that $\lambda^\varepsilon = \lambda^0 + \varepsilon\gamma + o(\varepsilon)$ for some constant γ.

With any $U \in \mathscr{U}$ in (1) the Ito conditions are satisfied, and there exists a solution $\xi(t)$ to (1) which is unique in the sense that any two solutions agree on $[s_0, T]$ with probability one. Let us use the expanded notation $\xi(t, U, \varepsilon, s, x)$ for the solution to (1) to indicate the value of ε, the control U, and the initial condition (s, x). For each $\varepsilon \geqq 0$ there exists an optimal control for the minimization problem (1), (2) which will be denoted by U^ε. The existence of U^ε follows from Theorem 3 in [5]. Finally, let $\xi^\varepsilon(t) = \xi(t, U^\varepsilon, \varepsilon, s, x)$, $\eta(t, U, s, x) = \xi(t, U, 0, s, x)$, $\eta^\varepsilon(t) = \eta(t, U^\varepsilon, s, x)$.

Kushner [10] developed a necessary condition for a control to be optimal which will be useful later in establishing regularity properties of the optimal control. Let p' denote the transpose of p.

LEMMA 2.1. *Let $p^\varepsilon(t)$ satisfy*

$$dp^\varepsilon(t) = \{-f_x'(t, \xi^\varepsilon(t), U^\varepsilon(t))p^\varepsilon(t) - L_x(t, \xi^\varepsilon(t), U^\varepsilon(t))\}\,dt \tag{4}$$

with $p^\varepsilon(T) = 0$ and $\theta^\varepsilon(t, u) = E\{L(t, \xi^\varepsilon(t), u) + p^{\varepsilon\prime}(t)f(t, \xi^\varepsilon(t), u)\}$. Then $\theta^\varepsilon(t, u)$ is minimum on K for $u = U^\varepsilon(t)$ for almost all t in $[s_0, T]$.

This section is concluded with some standard estimates relating to the solution of stochastic differential equations. Let

$$\|g(t)\|_s = \sup_{s \leqq t \leqq T} |g(t)|, \qquad \|g(t)\| = \sup_{s_0 \leqq s \leqq T} |g(t)|,$$

and M_1, M_2 be upper bounds on $|f_x|$, $|f_u|$ respectively on $Q \times K$. From the generalized Gronwall's inequality [7, p. 83], one obtains

$$\begin{aligned} &\|\xi(t, U, \varepsilon, s, x) - \eta(t, U^*, s, x)\|_s \\ &\leqq \left\{M_2 \int_{s_0}^T |U(t) - U^*(t)|\,dt + (2\varepsilon)^{1/2}\|w(t) - w(s)\|_s\right\} \exp M_1(T - s_0) \end{aligned} \tag{5}$$

for all $U, U^* \in \mathscr{U}$, $\varepsilon \geqq 0$, $s_0 \leqq s \leqq t \leqq T$. From [1, p. 392], we recall that

$$\Pr\{\|w(t) - w(s)\| \geqq \lambda\} \leqq C \exp(-r\lambda^2) \tag{6}$$

for $s_0 \leqq s$, $t \leqq T$ and appropriate positive constants C, r. Finally, there exists a positive constant C_5 depending only on M_1, $T - s_0$, ε_0, m, and the compact control set K such that

$$E\|\xi(t, U, \varepsilon, s, x)\|_s^n < C_5 \quad \text{for } n = 1, \cdots, m, \tag{7}$$

and

$$(8) \qquad E|\xi(t, U, \varepsilon, s, x) - \xi(t^*, U, \varepsilon, s, x)| \leqq C_5|t - t^*|^{1/2}$$

for $0 \leqq \varepsilon \leqq \varepsilon_0, s_0 \leqq s \leqq t, t^* \leqq T$. See [6, pp. 392–400].

3. Uniform convergence of U^ε to U^0. In this section uniform convergence of U^ε to U^0 as $\varepsilon \to 0$ is established. It is first shown that $U^\varepsilon \to U^0$ in measure as $\varepsilon \to 0$ using the same technique as Fleming applied in the completely observable case [2]. The first lemma, a modification of Lemma 3.1 in [2], is concerned with deterministic control problems. The proof uses the special form of f and the assumption that (s_0, x_0) is a regular point.

LEMMA 3.1. *Given $a > 0$ there exists $\delta > 0$ with the property that if $U \in \mathscr{U}$ and*

$$\int_{s_0}^T L(t, \eta(t, U, s_0, x_0), U(t))\, dt < \lambda^0 + \delta,$$

then

$$\|\eta - \xi^0\| < a \quad \textit{and} \quad \int_{s_0}^T |U(t) - U^0(t)|\, dt < a.$$

LEMMA 3.2. *There exists $C > 0$ such that*

$$|\lambda^\varepsilon - \lambda^0| < C\varepsilon^{1/2} \quad \textit{for all } \varepsilon \geqq 0.$$

Proof. U^ε is at best optimal in the deterministic problem, hence

$$\lambda^0 - \lambda^\varepsilon \leqq E \int_{s_0}^T L(t, \eta^\varepsilon(t), U^\varepsilon(t)) - L(t, \xi^\varepsilon(t), U^\varepsilon(t))\, dt$$

$$\leqq M_3(T - s_0)E\|\eta^\varepsilon - \xi^\varepsilon\|,$$

where M_3 is a bound on $|L_x|$ on $Q \times K$. From equations (5) and (6) the inequality $\lambda^0 - \lambda^\varepsilon \leqq C\varepsilon^{1/2}$ follows. The opposite inequality is obtained similarly by using U^0 in the stochastic problem.

LEMMA 3.3.

(i) $|U^\varepsilon - U^0| \to 0$ *in measure on* $[s_0, T]$ *as* $\varepsilon \to 0$.

(ii) $\|\xi^\varepsilon - \xi^0\| \to 0$ *as* $\varepsilon \to 0$ *almost surely.*

(iii) $\|p^\varepsilon - p^0\| \to 0$ *as* $\varepsilon \to 0$ *almost surely.*

Proof. (i) As shown in Lemma 3.2,

$$\int_{s_0}^T L(t, \eta^\varepsilon(t), U^\varepsilon(t)) = E \int_{s_0}^T L(t, \eta^\varepsilon(t), U^\varepsilon(t)) - L(t, \xi^\varepsilon(t), U^\varepsilon(t))\, dt + \lambda^\varepsilon$$

$$\leqq C\varepsilon^{1/2} + \lambda^\varepsilon \leqq 2C\varepsilon^{1/2} + \lambda^0.$$

For each $a > 0$ in Lemma 3.1, pick ε^* so that $2C\varepsilon^{1/2} < \delta$ for $0 \leqq \varepsilon \leqq \varepsilon^*$. Then $\int_{s_0}^T |U^\varepsilon(t) - U^0(t)|\, dt < a$ and hence $|U^\varepsilon(t) - U^0(t)| \to 0$ in measure on $[s_0, T]$ as $\varepsilon \to 0$.

(ii) Let $\{\varepsilon_m\}$ be any sequence with $\varepsilon_m \to 0$ as $m \to \infty$. It suffices to show for every $\gamma > 0$,

$$\lim_{\varepsilon_m \to 0} \Pr\left\{ \bigcup_{0 \leqq \varepsilon \leqq \varepsilon_m} \|\xi^\varepsilon - \xi\| > \gamma \right\} = 0.$$

See [8, p. 91]. Fix $\gamma > 0$. Now $\|\xi^\varepsilon - \xi^0\| \leqq \|\xi^\varepsilon - \eta^\varepsilon\| + \|\eta^\varepsilon - \xi^0\|$. From Lemmas 3.1 and 3.2, pick $\varepsilon_1 > 0$ so that $\|\eta^\varepsilon - \xi^0\| \leqq \gamma/2$ if $0 < \varepsilon < \varepsilon_1$. Therefore $\Pr\{\|\xi^\varepsilon - \xi^0\| > \gamma\} \leqq \Pr\{\|\xi^\varepsilon - \eta^\varepsilon\| > \gamma/2\}$. From (5) with $U = U^*$ if $\|\xi^{\varepsilon\gamma} - \eta^{\varepsilon\gamma}\| < \gamma/2$, then $\|\xi^\varepsilon - \eta^\varepsilon\| < \gamma/2$ for $0 \leqq \varepsilon \leqq \varepsilon_\gamma$. Therefore

$$0 \leqq \lim_{\varepsilon_m \to 0} \Pr\left\{\bigcup_{0 \leqq \varepsilon \leqq \varepsilon_m} \|\xi^\varepsilon - \xi^0\| > \gamma\right\} \leqq \lim_{\varepsilon_m \to 0} \Pr\{\|\xi^{\varepsilon_m} - \eta^{\varepsilon_m}\| > \gamma/2\}$$

$$\leqq \lim_{\varepsilon_m \to 0} \Pr\{(2\varepsilon_m)^{1/2}\|w(t) - w(s)\| \exp M_1(T - s_0) > \gamma/2\} = 0$$

and (ii) is established.

(iii) Recall (4). Since $|f_x|$ and $|L_x|$ are bounded on $Q \times K$, a simple Gronwall estimate shows that $\|p^\varepsilon(t)\| \leqq C$ for $\varepsilon \geqq 0$ for some constant C. Let $v = T - t$ and use the generalized Gronwall's inequality to obtain for $0 \leqq v \leqq T - s_0$,

$$|p^\varepsilon(v) - p^0(v)| \leqq \int_{s_0}^T \left(\exp \int_{s_0}^T |A^0(u)|\,du\right)\{|A^0(s) - A^\varepsilon(s)|\,|p^\varepsilon(s)| + |h^\varepsilon(s) - h^0(s)|\}\,ds,$$

where $A^\varepsilon(t) = -f'_x(t, \xi^\varepsilon(t), U^\varepsilon(t))$ and $h^\varepsilon(t) = -L_x(t, \xi^\varepsilon(t), U^\varepsilon(t))$. Almost surely the integrand tends to zero in measure on $[s_0, T]$. Since $\|p^\varepsilon\|$, $\|A^\varepsilon\|$, $\|h^\varepsilon\|$ are uniformly bounded, then by the Lebesgue dominated convergence theorem, $\|p^\varepsilon - p^0\| \to 0$ almost surely.

To prove that $U^\varepsilon \to U^0$ uniformly on $[s_0, T]$, some a priori estimates on the smoothness of the optimal controls U^ε are established using Lemma 2.1 and Lemma 3.4, which is a modification of Lemma 2.1 in [4].

LEMMA 3.4. *Let $\theta: [s_0, T] \times K \to R$ be such that*

(i) *θ is C^2 in u for each t;*

(ii) *θ_u satisfies a uniform Hölder condition in s on $[s_0, T]$ with constants C^*, α^*, i.e., for all $s, t \in [s_0, T]$, $u \in K$,*

$$|\theta_u(s, u) - \theta_u(t, u)| \leqq C^*|t - s|^{\alpha^*};$$

(iii) *$v'\theta_{uu}v \geqq \gamma^*|v|^2$ for all nonzero $v \in R^K$.*

Let $U^*(s)$ be the unique $u \in K$ such that $\theta(s, U^*(s)) = \min_{u\in K} \theta(s, u)$. Then U^* is uniformly Hölder continuous on $[s_0, T]$ and the Hölder constants C, α depend only on C^*, α^* and γ^*.

THEOREM 3.5. $\|U^\varepsilon - U^0\| \to 0$ *as* $\varepsilon \to 0$.

Proof. Since $U^\varepsilon \to U^0$ in measure as $\varepsilon \to 0$, it suffices to show that given $\varepsilon_0 \geqq 0$, there exist constants C, α such that $|U^\varepsilon(t) - U^\varepsilon(s)| \leqq C|t - s|^\alpha$ for all $0 \leqq \varepsilon \leqq \varepsilon_0$, $s_0 \leqq s, t \leqq T$. Let θ^ε be defined as in Lemma 2.1. θ^ε is clearly C^2 in u for fixed t in $[s_0, T]$; also $v'\theta^\varepsilon_{uu}v \geqq C_0|v|^2$ for all nonzero $v \in R^K$.'Thus (i) and (iii) are satisfied. Now

$$\begin{aligned}\theta^\varepsilon_u(t, u) - \theta^\varepsilon_u(s, u) = E\{&L_u(t, \xi^\varepsilon(t), u) - L_u(s, \xi^\varepsilon(s), u)\\ &+ p^\varepsilon(t)[B(t, \xi^\varepsilon(t)) - B(s, \xi^\varepsilon(s))]\\ &+ B(s, \xi^\varepsilon(s))[p^\varepsilon(t) - p^\varepsilon(s)]\}.\end{aligned}$$

Since $|dp^\varepsilon/dt|$ is uniformly bounded, say by M_4, then $|p^\varepsilon(t) - p^\varepsilon(s)| \leqq M_4|t - s|$.

Hence using the bounds on L_{ux}, L_{ut}, B_t, B_x, $\|p^\varepsilon\|$, and equations (7) and (8), one obtains the validity of (ii).

4. The expansion. In the previous section it was established that $U^\varepsilon \in \mathscr{H}$, where $\mathscr{H}$ denotes the class of Hölder continuous functions on $[s_0, T]$ with values in K. For each $U \in \mathscr{H}$, $\varepsilon \geqq 0$, consider the partial differential equation

$$\text{(9)} \qquad \varepsilon\Delta_x\Psi(t,x) + \Psi_x(t,x)f(t,x,U(t)) + \Psi_t(t,x) + L(t,x,U(t)) = 0$$

with boundary condition $\Psi(T,x) = 0$, where Δ_x denotes the Laplacian in the variables x. Consider Ψ also as a function of U and ε. Then $\Psi(s, x, U, \varepsilon)$ is related to the open loop control problem by

$$\text{(10)} \qquad \Psi(s,x,U,\varepsilon) = E\int_s^T L(t,\xi(t,U,\varepsilon,s,x),U(t))\,dt.$$

Since $U \in \mathscr{H}$, for $\varepsilon \geqq 0$ the solution Ψ to (9) is such that all partial derivatives of Ψ with at most one t-differentiation are continuous. For $\varepsilon > 0$ this follows from differentiating (9) with respect to x_i and the Hölder continuity of the coefficients. For $\varepsilon = 0$ use the method of characteristics. Define $\chi(s, x, U, \varepsilon)$ for $\varepsilon \geqq 0, s \leqq T$, as the solution to the equation

$$\text{(11)} \qquad \begin{aligned} &\varepsilon\Delta_x\chi(s,x,U,\varepsilon) + \chi_x(s,x,U,\varepsilon)f(s,x,U(s)) + \chi_t(s,x,U,\varepsilon) \\ &\qquad + \Delta_x\Psi(s,x,U,0) = 0 \end{aligned}$$

with boundary condition $\chi(T, x, U, \varepsilon) = 0$. Recall (9). Then for $\varepsilon > 0$, $\varepsilon\chi(s,x,U,\varepsilon) = \Psi(s,x,U,\varepsilon) - \Psi(s,x,U,0)$. For the initial point (s_0, x_0),

$$\text{(12)} \qquad \begin{aligned} \lambda^\varepsilon - \Psi(s_0,x_0,U^\varepsilon,0) &\leqq \lambda^\varepsilon - \lambda^0 \\ &\leqq \Psi(s_0,x_0,U^0,\varepsilon) - \lambda^0. \end{aligned}$$

Hence in order to establish the expansion, it suffices to divide (12) by ε and show that all limits as $\varepsilon \to 0$ are equal. It is convenient to break up the proof into two lemmas.

LEMMA 4.1. *$\Psi(s_0, x_0, U, \varepsilon) = \Psi(s_0, x_0, U, 0) + \varepsilon\chi(s_0, x_0, U, 0) + o(\varepsilon)$ uniformly in $U \in \mathscr{H}$, where $o(\varepsilon)/\varepsilon \to 0$ as $\varepsilon \to 0$.*

Proof. It must be shown for all $U \in \mathscr{H}$ that there exists $g(\varepsilon)$ such that $|\chi(s_0, x_0, U, \varepsilon) - \chi(s_0, x_0, U, 0)| < g(\varepsilon)$, where $g(\varepsilon) \to 0$ as $\varepsilon \to 0$. By Dynkin's formula,

$$\chi(s_0,x_0,U,\varepsilon) = E\int_{s_0}^T \Delta_x\Psi(t,\xi(t,U,\varepsilon,s_0,x_0),U,0)\,dt,$$

while the method of characteristics yields

$$\chi(s_0,x_0,U,0) = \int_{s_0}^T \Delta_x\Psi(t,\eta(t,U,s_0,x_0),U,0)\,dt.$$

From (5), (6) we have that $\|\xi(t, U, \varepsilon, s_0, x_0) - \eta(t, U, s_0, x_0)\| \to 0$ almost surely as $\varepsilon \to 0$ uniformly in $U \in \mathscr{H}$. Hence it is sufficient to establish a uniform bound on the partial derivatives $\Psi_{x_ix_ix_j}$, i, j positive integers with $1 \leqq i, j \leqq n$, for all $(s, x) \in Q \times K$ and $U \in \mathscr{H}$.

Denote $b = (t, U, s, x)$ and $q = (t, \eta(b), U(t))$. Since

$$\Psi(s, x, U, 0) = \int_s^T L(q)\, dt,$$

then

$$\begin{aligned} &\Psi_{x_i}(s, x, U, 0) = \int_s^T \sum_{l=1}^n L_{x_l}(q) \frac{\partial \eta_l(b)}{\partial x_i}\, dt, \\ (13)\qquad &\Psi_{x_i x_i}(s, x, U, 0) = \int_s^T \left\{ \sum_{l=1}^n \left\{ \sum_{k=1}^n L_{x_l x_k}(q) \frac{\partial \eta_l(b)}{\partial x_i} \frac{\partial \eta_k(b)}{\partial x_i} \right\}\right. \\ &\qquad\qquad\qquad\qquad \left. + L_{x_l}(q) \frac{\partial^2 \eta_l(b)}{\partial x_i^2} \right\} dt \end{aligned}$$

and

$$\begin{aligned} \Psi_{x_i x_i x_j}(s, x, U, 0) = \int_s^T &\left\{ \sum_{l=1}^n \left\{ \sum_{k=1}^n \left\{ \sum_{r=1}^n L_{x_l x_k x_r}(q) \frac{\partial \eta_l(b)}{\partial x_j} \frac{\partial \eta_k(b)}{\partial x_i} \frac{\partial \eta_r(b)}{\partial x_j} \right\}\right.\right. \\ &\left. + L_{x_l x_k}(q) \left[\frac{\partial^2 \eta_l(b)}{\partial x_i \partial x_j} \frac{\partial \eta(b)}{\partial x_i} + \frac{\partial \eta_l(b)}{\partial x_i} \frac{\partial^2 \eta_k(b)}{\partial x_i x_j} \right] \right\} \\ &\left. + L_{x_j}(q) \frac{\partial^3 \eta_l(b)}{\partial x_i^2 \partial x_j} + \sum_{r=1}^n L_{x_j x_r}(q) \frac{\partial^2 \eta_l(b)}{\partial x_i^2} \frac{\partial \eta_r(b)}{\partial x_j} \right\} dt. \end{aligned}$$

By assumption, the partials L_{x_l}, $L_{x_l x_k}$, $L_{x_l x_k x_r}$ are bounded on $Q \times K$. Hence it suffices to show that $\partial \eta_k(b)/\partial x_i$, $\partial^2 \eta_j(b)/\partial x_i \partial x_l$, and $\partial^3 \eta_j(b)/\partial x_i^2 \partial x_l$ are uniformly bounded for all b such that $s \leqq t \leqq T$, $(s, x) \in Q$ and $U \in \mathscr{H}$. The vector $\partial \eta(b)/\partial x_i$ satisfies

$$(14)\qquad \frac{d}{dt}\left(\frac{\partial \eta(b)}{\partial x_i} \right) = f_x(t, \eta(t, U, s, x), U(t)) \frac{\partial \eta(b)}{\partial x_i}$$

with initial condition $\partial \eta(b)/\partial x_i = (0, \cdots, 0, 1_i, 0 \cdots, 0)'$. Using the bound on f_x, one obtains

$$\frac{\partial \eta}{\partial x_i}(t, U, s, x) \leqq \exp M_1(T - s_0) \quad \text{for } i = 1, \cdots, n,$$

$s, t \in [s_0, T]$, $x \in R^n$ and $U \in \mathscr{H}$. Similarly, $\partial^2 \eta_r(b)/\partial x_i \partial x_j$ satisfies

$$\frac{d}{dt}\left(\frac{\partial^2 \eta_r(b)}{\partial x_i \partial x_j} \right) = \sum_{r=1}^n f_{r_{x_l}}(q) \frac{\partial^2 \eta_l(b)}{\partial x_i \partial x_j} + \sum_{l=1}^n \sum_{k=1}^n f_{r_{x_l}}(q) \frac{\partial \eta_k(b)}{\partial x_i} \frac{\partial \eta_l(b)}{\partial x_j}$$

with initial condition $(\partial^2 \eta_r/\partial x_i \partial x_j)(s, U, s, x) = 0$.

From the above estimates, $\partial \eta_k(b)/\partial x_i$ and $\partial \eta_l(b)/\partial x_j$ are uniformly bounded. By assumption the matrices f_x, f_{xx} are uniformly bounded; hence one obtains that

$$\frac{\partial^2 \eta_r}{\partial x_i \partial x_j}(t, U, s, x) \leqq C_1 \exp C_2(T - s_0)$$

for suitable positive constants C_1, C_2 for all integers i, j, r; $1 \leqq i, j, r \leqq n$, $s \leqq t \leqq T$,

$(s, x) \in Q$ and $U \in \mathscr{H}$. One can show in a similar manner that the values of $|\partial^3 \eta_r(b)/\partial^2 x_i\, \partial x_j|$ are uniformly bounded. Thus Lemma 4.1 is proved.

LEMMA 4.2. $\chi(s_0, x_0, U^\varepsilon, 0) \to \chi(s_0, x_0, U^0, 0)$ *as* $\varepsilon \to 0$.

Proof. It must be shown that

$$\int_{s_0}^{T} \Delta_x \Psi(t, \eta^\varepsilon(t), U^\varepsilon, 0) - \Delta_x \Psi(t, \xi^0(t), U^0, 0)\, dt \to 0 \quad \text{as } \varepsilon \to 0.$$

Since $U^\varepsilon \to U^0$ in measure, then from (5), (6), $\|\eta^\varepsilon - \xi^0\| \to 0$ as $\varepsilon \to 0$. Let B be a compact subset of R^n which contains the points $\xi^0(t)$, $\eta^\varepsilon(t)$ for $s_0 \leqq t \leqq T$, $\varepsilon \geqq 0$. Then it suffices to show that for each integer i, $1 \leqq i \leqq n$,

$$\Psi_{x_i x_i}(t, x, U^\varepsilon, 0) \to \Psi_{x_i x_i}(t, x, U^0, 0)$$

uniformly on $[s_0, T] \times B$ as $\varepsilon \to 0$. From (13) we have

$$\Psi_{x_i x_i}(s, x, U^\varepsilon, 0) = \int_s^T \left\{ \sum_{l=1}^n \left\{ \sum_{k=1}^n L_{x_l x_k}(q^\varepsilon) \frac{\partial \eta_l(b^\varepsilon)}{\partial x_i} \frac{\partial \eta_k(b^\varepsilon)}{\partial x_i} \right\} \right.$$
$$\left. + L_{x_l}(q^\varepsilon) \frac{\partial^2 \eta_l(b^\varepsilon)}{\partial x_i^2} \right\} dt, \tag{15}$$

where $q^\varepsilon = (t, \eta(b^\varepsilon), U^\varepsilon(t))$ and $b^\varepsilon = (t, U^\varepsilon, s, x)$. Since each of the terms in the integrand is uniformly bounded for all t such that $s_0 \leqq t \leqq T$, it is only necessary to show that each of the terms $|L_{x_l x_k}(q^\varepsilon) - L_{x_l x_k}(q^0)|$, $|L_{x_l}(q^\varepsilon) - L_{x_l}(q^0)|$,

$$\left| \frac{\partial \eta_l(b^\varepsilon)}{\partial x_i} - \frac{\partial \eta_l(b^0)}{\partial x_i} \right|, \quad \left| \frac{\partial^2 \eta_l(b^\varepsilon)}{\partial x_i^2} - \frac{\partial^2 \eta_l(b^0)}{\partial x_i^2} \right|$$

has limit 0 as $\varepsilon \to 0$ uniformly in b^ε and q^ε.

Consider the terms involving $|L_{x_l} x_m(q^\varepsilon) - L_{x_l x_m}(q^0)|$. First

$$|L_{x_l x_m}(t, \eta(t, U^\varepsilon, s, x) U^\varepsilon(t)) - L_{x_l x_m}(t, \eta(t, U^0, s, x), U^0(t))|$$
$$\leqq n \sup_{Q \times K, r = 1, \cdots, n} |L_{x_l x_m x_r}(t, x, u)|\, |\eta(t, U^\varepsilon, s, x) - \eta(t, U^0, s, x)| \tag{16}$$
$$+ n \sup_{Q \times K, r = 1, \cdots, n} |L_{x_l x_m u_r}(t, x, u)|\, |U^\varepsilon(t) - U^0(t)|.$$

By Theorem 3.5, $U^\varepsilon - U^0 \to 0$ uniformly on $[s_0, T]$ as $\varepsilon \to 0$, and from (5),

$$\|\eta(t, U^\varepsilon, s, x) - \eta(t, U^0, s, x)\|_s \to 0 \quad \text{as } \varepsilon \to 0 \tag{17}$$

uniformly for $(s, x) \in [s_0, T] \times B$. Hence the uniformity for $L_{x_l x_m}$ is established. The verification for terms L_{x_l} is similar. Note that (17) is equivalent to $|q^\varepsilon - q^0| \to 0$ as $\varepsilon \to 0$ uniformly in s, t, x, $(s, x) \in [s_0, T] \times B$, $s \leqq t \leqq T$.

Let

$$v^\varepsilon(t) = \frac{\partial \eta_l(b^\varepsilon)}{\partial x_i} - \frac{\partial \eta_l(b^0)}{\partial x_i}.$$

From (14) we have that

$$|v^\varepsilon(t)| \leqq \int_s^T |f_x(q^0)|\, |v^\varepsilon(t)|\, dt + \int_s^T |f_x(q^\varepsilon) - f_x(q^0)| \left| \frac{\partial \eta_l(b^\varepsilon)}{\partial x_i} \right| dt$$

for $s \leqq t \leqq T$. Since $|\partial\eta_l(b^\varepsilon)/\partial x_i|$ is uniformly bounded, and $|q^\varepsilon - q^0| \to 0$ uniformly as $\varepsilon \to 0$, using the generalized Gronwall inequality one obtains that $\|v^\varepsilon(t)\|_s \to 0$ as $\varepsilon \to 0$ uniformly with respect to $(s, x) \in [s_0, T] \times B$, $s \leqq t \leqq T$. The verification for the remaining term follows in a similar but more tedious manner. Thus Lemma 4.2 is established.

THEOREM 4.3. *$\lambda^\varepsilon = \lambda^0 + \varepsilon\chi + o(\varepsilon)$, where*

$$\chi = \chi(s_0, x_0, U^0, 0) = \int_{s_0}^{T} \Delta_x \Psi(t, \xi^0(t), U, 0)\, dt.$$

Proof. Using Lemma 4.1 in (12), one obtains

$$\begin{aligned}\varepsilon\chi(s_0, x_0, U^\varepsilon, 0) + o(\varepsilon) &\leqq \Psi^\varepsilon(s_0, x_0) - \Psi^0(s_0, x_0)\\ &\leqq \varepsilon\chi(s_0, x_0, U^0, 0) + o(\varepsilon).\end{aligned}$$

The result follows by dividing by $\varepsilon > 0$ and using Lemmas 4.1 and 4.2.

Recall that $\Psi(s_0, s_0, U^0, \varepsilon)$ is the cost of using U^0 in the ε-problem. We have shown that

$$\Psi^\varepsilon(s_0, x_0) = \lambda^0 + \varepsilon\chi(s_0, x_0, U^0, 0) + o(\varepsilon)$$

and $\Psi(s_0, x_0, U^0, \varepsilon) = \lambda^0 + \varepsilon\chi(s_0, x_0, U^0, 0) + o(\varepsilon)$. Thus to first order in ε the costs of using U^ε and U^0 in the open loop stochastic control problem are the same. Thus U^0 is approximately optimal in the stochastic control problem for small ε and the approximate cost can be found by dropping the $o(\varepsilon)$ term in the previous equation.

Let us consider the open loop stochastic linear regulator problem. Let $f(t, x, u) = A(t)x + B(t)u$ and $L(t, x, u) = x'Mx + u'Nu$ with $N > 0$, $M \geqq 0$. As is well known, since each open loop control generates a Gaussian process, the stochastic control problem can be converted to a deterministic control problem with the same equations as the original deterministic control problem and modified cost functional

$$L(t, x, u) = x'Mx + u'Nu + \operatorname{trace} \int_{s_0}^{T} MQ(t)\, dt. \tag{18}$$

$Q(t)$, the covariance matrix of the uncontrolled stochastic process, satisfies the equation

$$\frac{dQ}{dt} = A(t)Q + QA'(t) + (2\varepsilon)I, \qquad Q(s_0) = 0. \tag{19}$$

Therefore $U^\varepsilon = U^0$ and

$$\Psi^\varepsilon(s_0, x_0) = \lambda^0 + \operatorname{trace} \int_{s_0}^{T} M(t)Q(t)\, dt$$

in case either $K = R^k$ or K is compact and convex. Defining $S(t)$ as the solution to

(19) with $\varepsilon = 1/2$, one can write

$$\Psi^\varepsilon(s_0, x_0) = \lambda^0 + 2\varepsilon \operatorname{trace} \int_{s_0}^{T} M(t)S(t)\,dt$$

which is in the form in Theorem 4.3 with $o(\varepsilon) = 0$.

Consider again the general class of problems covered by the assumptions (3), and let $\phi^\varepsilon(s_0, x_0)$ be the optimal cost of the corresponding completely observable stochastic control problem. Then Fleming [2] showed that

$$\phi^\varepsilon(s_0, x_0) = \lambda^0 + \varepsilon\theta(s_0, x_0) + o(\varepsilon)$$

for some constant $\theta(s_0, x_0)$. Let $\Phi^\varepsilon(s_0, x_0)$ be the cost of using the optimal deterministic feedback control in the ε-completely observable problem. Then he obtained

$$\Phi^\varepsilon(s_0, x_0) = \lambda^0 + \varepsilon\theta(s_0, x_0) + o(\varepsilon).$$

Thus the costs $\Phi^\varepsilon(s_0, x_0)$ and $\phi^\varepsilon(s_0, x_0)$ are the same to first order in ε. Finally, as shown in [2, p. 502],

$$\theta(s_0, x_0) \leqq \chi(s_0, x_0).$$

Acknowledgment. This research is a portion of a Ph.D. thesis written under the direction of Wendell Fleming at Brown University. The author wishes to express his appreciation to him.

REFERENCES

[1] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[2] W. H. Fleming, *Stochastic control for small noise intensities*, this Journal, 9 (1971), pp. 483–517.

[3] ———, *Optimal control of partially observable diffusions*, this Journal, 6 (1968), pp. 194–214.

[4] ———, *Duality and a priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., 16 (1966), pp. 254–279; Erratum, Ibid., 19 (1966), p. 204.

[5] W. H. Fleming and M. Nisio, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777–794.

[6] I. I. Gikhman and A. V. Skorokhod, *Introduction to the Theory of Random Processes*, W. B. Saunders, Philadelphia, 1969.

[7] J. K. Hale, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.

[8] P. Halmos, *Measure Theory*, Van Nostrand, Princeton, N.J., 1950.

[9] C. J. Holland, *Small noise open loop control problems*, thesis, Brown University, Providence, 1972.

[10] H. J. Kushner, *On the stochastic maximum principle: Fixed time of control*, J. Math. Anal. Appl., 11 (1965), pp. 78–92.

[11] R. E. Mortensen, *Stochastic control with noisy observations*, Internat. J. Control, 4 (1966), pp. 455–464.

[12] R. VanSlyke and R. Wets, *Programming under uncertainty and stochastic optimal control*, this Journal, 4 (1966), pp. 179–193.

SIAM J. CONTROL
Vol. 12, No. 3, August 1974

CONTROLLABILITY FOR PARTIAL DIFFERENTIAL EQUATIONS OF PARABOLIC TYPE*

YOSHIYUKI SAKAWA†

Abstract. The purpose of this paper is to study questions regarding controllability for the distributed-parameter systems described by partial differential equations of parabolic type. Fattorini [2]–[4] studied controllability by finitely many functions of time. We also consider this type of controllability for the distributed-parameter systems with control functions appearing in the differential equations as well as in the boundary conditions. Necessary and sufficient conditions for controllability are presented. We give some useful results for designing controllers which make systems controllable. Several examples are worked out.

1. Introduction. We consider in this paper control systems described by the linear partial differential equations of parabolic type. Two cases are treated where the control inputs appear in the partial differential equations as distributed inputs and/or they appear in the boundary conditions as boundary inputs. The latter form of control is easier to realize physically. We seek controllability conditions which ensure that the system can be steered from the zero initial state to an arbitrary small vicinity of a final state.

Fattorini [2]–[4] studied controllability by finitely many functions of time; this notion is technically significant from the view point of realization of controls. He obtained his main results by applying the concept of ordered representation of a Hilbert space with respect to a self-adjoint operator. In this paper we make use of the results of Ito [9], by which the solutions of the parabolic equations with boundary conditions of general type can be expressed in terms of the fundamental solutions, and the fundamental solutions can be further expressed in terms of the eigenvalues and eigenfunctions of the corresponding elliptic equations. Using these results we shall obtain the finite controllability conditions of Fattorini. The results here also show us how to construct the controllers which make systems controllable. The results are then applied to several boundary control problems.

2. Preliminary results. Let D be a bounded domain of an r-dimensional Euclidean space, and let S, the boundary of D, consist of a finite number of $(r - 1)$-dimensional hypersurfaces of class C^3.[1] The spatial coordinate vector will be denoted by $x = (x_1, x_2, \cdots, x_r) \in D$. Consider a linear parabolic partial differential equation

$$\partial u(t,x)/\partial t = \Delta u(t,x) - q(x)u(t,x) + \sum_{i=1}^{n} g_i(x)f_i(t), \tag{1}$$

where t is the time and Δ denotes the Laplacian given by

$$\Delta = \partial^2/(\partial x_1)^2 + \cdots + \partial^2/(\partial x_r)^2.$$

* Received by the editors November 3, 1972, and in revised form April 10, 1973.

† Faculty of Engineering Science, Osaka University, Toyonaka, Osaka, Japan.

[1] In general, C^n denotes the set of all functions having n continuous derivatives.

It is assumed that $q(x)$ and $g_i(x)$ are Hölder continuous on the compact domain $\bar{D}(\bar{D} = D \cup S$, the upper bar denotes the closure), and that $f_i(t)$ are Hölder continuous on $[0, T]$, where T is an arbitrarv time.

The boundary condition is given by

$$\alpha(\xi)u(t, \xi) + (1 - \alpha(\xi))\partial u(t, \xi)/\partial n = 0, \tag{2}$$

where $\xi \in S$, n is the exterior normal to the surface S at a point $\xi \in S$, and $\alpha(\xi)$ is a function of class C^2 on S satisfying

$$0 \leqq \alpha(\xi) \leqq 1.$$

The initial condition is given by

$$\lim_{t \to 0} u(t, x) = u_0(x) \quad \text{in } L_2(D), \tag{3}$$

where $u_0(x) \in L_2(D)$, and $L_2(D)$ denotes the Hilbert space of all square integrable real-valued functions $u(x)$ with the norm

$$\|u\| = \left[\int_D u(x)^2\, dx\right]^{1/2}.$$

Equations (1) and (2) can be written together as a differential equation in $L_2(D)$:

$$du(t)/dt = Au + \sum_{i=1}^{n} g_i f_i(t), \tag{4}$$

where g_i are the elements of $L_2(D)$ corresponding to the $g_i(x)$ mentioned above, the domain $D(A)$ of the operator A is given by

$$D(A) = \{u : \Delta u \in L_2(D), \alpha(\xi)u(\xi) + (1 - \alpha(\xi))\partial u(\xi)/\partial n = 0, \xi \in S\},$$

and

$$Au = \Delta u - q(x)u \quad \text{if } u \in D(A).$$

Clearly the operator A is self-adjoint. Applying Green's formula, we obtain

$$\begin{aligned}(Au, u) &= (\Delta u, u) - (qu, u) \\ &= \int_S \frac{\partial u(\xi)}{\partial n} u(\xi)\, dS_\xi - \sum_{j=1}^{r} \left(\frac{\partial u}{\partial x_j}, \frac{\partial u}{\partial x_j}\right) - (qu, u) \\ &\leqq -[\min_{x \in \bar{D}} q(x)](u, u),\end{aligned} \tag{5}$$

where $(\cdot, \cdot)$ denotes the inner product of $L_2(D)$, and dS_ξ is the surface element of S at $\xi \in S$. Therefore, the operator A is semibounded from above.

It is shown by Ito [9] that there exists a function $U(t, x, y)$ $(0 < t; x, y \in \bar{D})$ which is of class C^1 in t, of class C^2 in x and y in $\bar{D}$, and satisfies the following conditions (6)–(9):

$$\partial U(t, x, y)/\partial t = \Delta_x U(t, x, y) - q(x)U(t, x, y), \tag{6}$$

$$\alpha(\xi)U(t, \xi, y) + (1 - \alpha(\xi))\partial U(t, \xi, y)/\partial n_\xi = 0, \qquad \xi \in S, \tag{7}$$

$$\lim_{t\to 0}\int_D U(t,x,y)u_0(y)\,dy = u_0(x) \quad \text{in } L_2(D), \tag{8}$$

$$U(t,x,y) = U(t,y,x), \tag{9}$$

where (6) and (7) hold for all $y \in \bar{D}$ and (8) holds for all $u_0(x) \in L_2(D)$. This function $U(t,x,y)$ $(0 < t;\, x, y \in \bar{D})$ is called a *fundamental solution* of the system described by (1) and (2).

Ito [9] obtained the following theorems.

THEOREM 1. (Ito) *There exists a unique solution to the initial-boundary value problem described by* (1), (2), *and* (3), *and the solution is given by*

$$\begin{aligned} u(t,x) &= \int_D U(t,x,y)u_0(y)\,dy \\ &\quad + \int_0^t d\tau \int_D U(t-\tau,x,y)\left(\sum_{i=1}^n g_i(y)f_i(\tau)\right) dy, \end{aligned} \tag{10}$$

where $0 < t < \infty$, $x \in D$.

THEOREM 2. (Ito) *There exists a sequence* $\{\lambda_i, \phi_{ij}; j = 1, \cdots, m_i, i = 1, 2, \cdots\}$ *of eigenvalues and eigenfunctions satisfying the following conditions*:

(i) *Where* $C = \min_{x \in \bar{D}} q(x)$,

$$C \leqq \lambda_1 < \lambda_2 < \cdots < \lambda_i < \cdots, \quad \lim_{i\to\infty} \lambda_i = \infty. \tag{11}$$

(ii) $\{\phi_{ij}(x); j = 1, \cdots, m_i, i = 1, 2, \cdots\}$ *is a complete orthonormal system in* $L_2(D)$, *where the positive integers* m_i *are finite for any* $i < \infty$.

(iii) *Each* $\phi_{ij}(x)$ *satisfies the following equations*

$$\int_D U(t,x,y)\phi_{ij}(y)\,dy = e^{-\lambda_i t}\phi_{ij}(x), \tag{12}$$

$$\Delta\phi_{ij}(x) - q(x)\phi_{ij}(x) = -\lambda_i\phi_{ij}(x), \tag{13}$$

and the boundary condition (2).

(iv) *The fundamental solution is expressed as*

$$U(t,x,y) = \sum_{i=1}^{\infty} e^{-\lambda_i t} \sum_{j=1}^{m_i} \phi_{ij}(x)\phi_{ij}(y), \tag{14}$$

where the sequence in the right side converges uniformly on $[\delta, \infty] \times \bar{D} \times \bar{D}$ *for arbitrary* $\delta > 0$.

(v) *For arbitrary* $h \in L_2(D)$ *given by*

$$h(x) = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} h_{ij}\phi_{ij}(x),$$

we have

$$(U_t h)(x) = \sum_{i=1}^{\infty} e^{-\lambda_i t} \sum_{j=1}^{m_i} h_{ij}\phi_{ij}(x), \tag{15}$$

where

$$(U_t h)(x) \equiv \int_D U(t, x, y)h(y)\,dy, \tag{16}$$

and the sequence converges uniformly on $[\delta, \infty) \times \bar{D}$ *for arbitrary* $\delta > 0$.

Now, since the function $(U_t h)(x)$ is continuous on $\bar{D}$, the operator U_t, called the *semigroup associated with* (1), maps $L_2(D)$ into itself. Using this operator, we can rewrite the solution (10) as

$$u(t) = U_t u_0 + \sum_{i=1}^{n} \int_0^t U_{t-\tau} g_i f_i(\tau)\,d\tau. \tag{17}$$

The positive integers m_i are called the *multiplicity* of the eigenvalues λ_i. If $\sup \{m_i\} = m < \infty$, we shall say that A has *multiplicity* m; if $\sup \{m_i\} = \infty$, A is said to have *infinite multiplicity.* In the following, controllability of the system described by (1) and (2) will be considered for the case where A has finite multiplicity.

3. Controllability by distributed inputs. In the partial differential equation (1), the inputs or controls $f_i(t), i = 1, \cdots, n$, are taken in the class, denoted F, of functions Hölder continuous on $[0, T]$, T being an arbitrary time. Following Fattorini [4], we shall say that the distributed parameter system is *completely controllable* if, for any $u_0 \in L_2(D)$, $v \in L_2(D)$, and $\varepsilon > 0$, there exists a set of controls $f = (f_1, f_2, \cdots, f_n), f_i \in F, i = 1, \cdots, n$, such that the solution of (1)–(3) satisfies

$$\|u(t_1) - v\| \leqq \varepsilon$$

for some t_1 depending in general on u_0, v and ε. If t_1 can be chosen independently of u_0, v, and ε, then we shall say that the system is *completely controllable in time* t_1.

If the initial state u_0 is taken to be the null element of $L_2(D)$, then we shall say that the system is *null controllable* (*null controllable in time* t_1). Let us define the *attainable set* $K(t)$ in $L_2(D)$ by

$$K(t) = \left\{ u = \sum_{i=1}^{n} \int_0^t U_{t-\tau} g_i f_i(\tau)\,d\tau : f_i \in F, i = 1, \cdots, n \right\}. \tag{18}$$

Then it is clear that $K(t)$ is a linear manifold and that the system is null controllable if and only if $\bar{K} = L_2(D)$, where $K = \bigcup_{t>0} K(t)$ and $\bar{K}$ denotes the closure of K. It is null controllable in time t_1 if and only if $\bar{K}(t_1) = L_2(D)$. Fattorini [4] proved that if the system described by (1) and (2) (or (4)) is null controllable in time t_1, then it is completely controllable in time t_1.

Now, if $u^* \in K^\perp$, then it follows that

$$\sum_{k=1}^{n} \int_0^t (U_{t-\tau} g_k, u^*) f_k(\tau)\,d\tau = 0 \tag{19}$$

for all $f_k \in F$, $k = 1, \cdots, n$, and for all $t > 0$, where $(\cdot, \cdot)$ denotes the inner product in $L_2(D)$. From (19) we see that $u^* \in K^\perp (u^* \in K(t_1)^\perp)$ if and only if

$$(U_t g_k, u^*) = 0, \qquad k = 1, \cdots, n, \tag{20}$$

for all $t > 0\,(0 < t \leqq t_1)$. On the other hand, the necessary and sufficient condition for the system to be null controllable (in time t_1) is that $u^* \in K^{\perp}$ ($u^* \in K(t_1)^{\perp}$) implies $u^* = 0$. Therefore, the system is null controllable (in time t_1) if and only if equality (20) holding for all $t > 0$ $(0 < t \leqq t_1)$ implies $u^* = 0$.

From Theorem 2, we see that

$$(U_t g_k, u^*) = \sum_{i=1}^{\infty} e^{-\lambda_i t} \sum_{j=1}^{m_i} g_{ij}^k u_{ij}, \tag{21}$$

where

$$g_{ij}^k = (g_k, \phi_{ij}), \qquad u_{ij} = (u^*, \phi_{ij}). \tag{22}$$

Since the function given by (21) is analytic in $t \in (0, \infty)$, (21) can vanish for $0 < t \leqq t_1$ $(t_1 > 0)$ only if it vanishes for all $t > 0$. Consequently the system (4) is null controllable if and only if it is null controllable in any time $t_1 > 0$.

The following result is partly included in that of Fattorini [3], who proved the theorem using the ordered representation of a Hilbert space with respect to a self-adjoint operator. Here, however, the proof will be made by an elementary calculus. The result here is an extension of a result of Tsujioka [10].

THEOREM 3. *Suppose that A has finite multiplicity m. Let us define $n \times m_i$ matrices G_i by*

$$G_i = \begin{bmatrix} g_{i1}^1 & g_{i2}^1 & \cdots & g_{im_i}^1 \\ g_{i1}^2 & g_{i2}^2 & & g_{im_i}^2 \\ \vdots & \vdots & & \vdots \\ g_{i1}^n & g_{i2}^n & \cdots & g_{im_i}^n \end{bmatrix}, \tag{23}$$

where g_{ij}^k $(j = 1, \cdots, m_i; k = 1, \cdots, n)$ are defined by (22). *The control system described by* (1) *and* (2) *is null controllable in any finite time if and only if $n \geqq m = \max\{m_i\}$ and*

$$\operatorname{rank} G_i = m_i \quad \textit{for all } i = 1, 2, \cdots. \tag{24}$$

Proof. Let $u^* \in K^{\perp}$; then (20) holds for all $t > 0$. For any complex number λ with $\operatorname{Re}\lambda < \lambda_1$, we see from (11) and (21) that

$$\begin{aligned} 0 &= \sum_{i=1}^{\infty} \int_0^{\infty} e^{-(\lambda_i - \lambda)t} \left(\sum_{j=1}^{m_i} g_{ij}^k u_{ij} \right) dt \\ &= \sum_{i=1}^{\infty} \frac{1}{\lambda_i - \lambda} \left(\sum_{j=1}^{m_i} g_{ij}^k u_{ij} \right), \qquad k = 1, \cdots, n. \end{aligned} \tag{25}$$

By analytic continuation, we see that (25) holds for all λ such that $\lambda \neq \lambda_i$, $i = 1, 2, \cdots$. Let C_i be a circle in the complex plane of radius ε_i with λ_i as center, where ε_i is such that

$$0 < \varepsilon_i < \min(\lambda_i - \lambda_{i-1}, \lambda_{i+1} - \lambda_i).$$

Then we have

$$\sum_{j=1}^{m_i} g_{ij}^k u_{ij} = \frac{1}{2\pi i}\int_{C_i}\sum_{i=1}^{\infty}\frac{1}{\lambda-\lambda_i}\left(\sum_{j=1}^{m_i} g_{ij}^k u_{ij}\right)d\lambda = 0, \tag{26}$$
$$k = 1,\cdots,n; \quad i = 1,2,\cdots.$$

If $n \geqq m_i$ and rank $G_i = m_i, i = 1,2,\cdots$, then (26) implies $u_{i1} = \cdots = u_{im_i} = 0$, $i = 1,2,\cdots$. Thus we have $K^\perp = \{0\}$ which implies that the system is null controllable.

Conversely, suppose that rank $G_i < m_i$ for some i. Then there exists a nonzero m_i-vector $u_i = (u_{i1},\cdots,u_{im_i})$ satisfying (26). This means that $\bar{K} \neq L_2(D)$ and the system is not null controllable, and the theorem is complete.

Now, the following corollary is technically important when we construct a controller.

COROLLARY 3.1. *Suppose that A has finite multiplicity m. Let a sequence $\{a_i\}$ of numbers be such that*

$$\sum_{i=1}^{\infty} a_i^2 < \infty, \qquad a_i \neq 0, \quad i = 1,2,\cdots.$$

If $n \geqq m$ and the functions $g_k(x), k = 1,\cdots,n$, are such that

$$g_k(x) = \sum_{i=1}^{\infty} a_i\phi_{ik}(x), \qquad k = 1,\cdots,n, \tag{27}$$

then the control system described by (1) *and* (2) *is null controllable, where $\phi_{ik}(x) \equiv 0$ if $k > m_i$.*

Proof. From the orthogonality of the eigenfunctions, we see that

$$g_{ij}^k = (g_k,\phi_{ij}) = a_i\delta_{jk},$$

where δ_{jk} is the Dirac δ-function. Therefore it follows that

$$G_i = a_i\begin{pmatrix} I_{m_i\times m_i} \\ O_{(n-m_i)\times m_i}\end{pmatrix}$$

which implies that rank $G_i = m_i, i = 1,2,\cdots$, thus completing the corollary.

Remark 1. If A has infinite multiplicity, then it is clear that the system (4) is not controllable by finitely many functions of time.

4. Controllability by boundary inputs. In this section, we consider the same system as (1) and (2), except that the control inputs are given through the boundary condition; that is, we consider the following system:

$$\partial u(t,x)/\partial t = \Delta u(t,x) - q(x)u(t,x), \tag{28}$$

with the boundary condition

$$\alpha(\xi)u(t,\xi) + (1-\alpha(\xi))(\partial u(t,\xi)/\partial n)) = \sum_{i=1}^{n} h_i(\xi)f_i(t), \tag{29}$$

where $\xi \in S$, $h_i(\xi)$ are of class C^2 on S, and $f_i(t)$, the boundary control inputs, are such that $df_i(t)/dt$ are Hölder continuous on $[0, T]$, T being an arbitrary time. Since null controllability is considered, the initial condition is assumed to be zero.

In [9], S. Ito proved that the unique solution of (28) and (29) with zero initial condition exists and it is given by

$$u(t,x) = \int_0^t d\tau \int_S \left\{U(t-\tau,x,\xi) - \frac{\partial U(t-\tau,x,\xi)}{\partial n_\xi}\right\}\left(\sum_{i=1}^n h_i(\xi)f_i(\tau)\right) dS_\xi. \tag{30}$$

In analogy with the preceding section, if $u^*(x) \in K^\perp$, then it follows that

$$\int_D u^*(x)\,dx \int_S \left\{U(t,x,\xi) - \frac{\partial U(t,x,\xi)}{\partial n_\xi}\right\} h_i(\xi)\,dS_\xi = 0, \tag{31}$$
$$i = 1,\cdots,n; \qquad t > 0.$$

If we put

$$\begin{aligned} h_{ij}^k &= \int_S \left\{\phi_{ij}(\xi) - \frac{\partial\phi_{ij}(\xi)}{\partial n}\right\} h_k(\xi)\,dS_\xi, \\ u_{ij} &= \int_D \phi_{ij}(x)u^*(x)\,dx, \end{aligned} \tag{32}$$

then from Theorem 2 and (31), we obtain

$$\sum_{i=1}^{\infty} e^{-\lambda_i t} \sum_{j=1}^{m_i} h_{ij}^k u_{ij} = 0, \qquad k = 1,\cdots,n. \tag{33}$$

In the same way as before, we see that

$$\sum_{j=1}^{m_i} h_{ij}^k u_{ij} = 0, \qquad k = 1,\cdots,n; \quad i = 1,2,\cdots. \tag{34}$$

Thus we obtain the following theorem.

THEOREM 4. *Suppose that A has finite multiplicity m. Let us define $n \times m_i$ matrices H_i by*

$$H_i = \begin{bmatrix} h_{i1}^1 & h_{i2}^1 & \cdots & h_{im_i}^1 \\ h_{i1}^2 & h_{i2}^2 & & h_{im_i}^2 \\ \vdots & \vdots & & \vdots \\ h_{i1}^n & h_{i2}^n & \cdots & h_{im_i}^n \end{bmatrix}. \tag{35}$$

The boundary control system described by (28) *and* (29) *is null controllable in any finite time if and only if $n \geqq m = \max\{m_i\}$ and*

$$\operatorname{rank} H_i = m_i, \qquad i = 1,2,\cdots.$$

It seems difficult to obtain a result analogous to Corollary 3.1. However, we obtain the following corollary which can be directly derived from Theorem 4.

COROLLARY 4.1. *If the multiplicity of A is one, that is, if $m_1 = m_2 = \cdots = 1$, then the boundary control system described by* (28) *and* (29) *is null controllable if and only if*

$$h_i = (h_i^1, h_i^2, \cdots, h_i^n) \neq 0 \tag{36}$$

for all $i = 1, 2, \cdots$, where

$$h_i^k = \int_S \left\{\phi_i(\xi) - \frac{\partial \phi_i(\xi)}{\partial n}\right\} h_k(\xi)\, dS_\xi. \tag{37}$$

Now, let us consider several examples of boundary control systems.

Example 1. We consider the following boundary control system:

$$\partial u/\partial t = \partial^2 u/\partial x^2 - q(x)u, \qquad a_1 < x < a_2, \tag{38}$$

$$\begin{aligned} &\alpha_1 u(t, a_1) - (1 - \alpha_1)[\partial u(t, a_1)/\partial x] = f(t), \\ &\alpha_2 u(t, a_2) + (1 - \alpha_2)[\partial u(t, a_2)/\partial x] = 0, \end{aligned} \tag{39}$$

where $0 \leqq \alpha_1, \alpha_2 \leqq 1$, $q(x)$ is Hölder continuous on $[a_1, a_2]$, and $f(t)$, the control input, is such that $df(t)/dt$ is Hölder continuous on $[0, \infty)$.

Since the eigenfunctions associated with the eigenvalue λ_i must satisfy the second order ordinary differential equation

$$d^2\phi/dx^2 - (q(x) - \lambda_i)\phi = 0, \tag{40}$$

the eigenfunctions can be expressed as a linear combination of two independent solutions of (40), the coefficients of which should be determined from the boundary conditions at $x = a_1$ and $x = a_2$. Therefore we see that the multiplicity in this case is one.

Since $h(a_1) = 1$, $h(a_2) = 0$, we see from (37) that

$$h_i = \phi_i(a_1) + \partial\phi_i(a_1)/\partial x.$$

On the other hand, the eigenfunction $\phi_i(x)$ satisfies the boundary condition

$$\alpha_1 \phi_i(a_1) - (1 - \alpha_1)[\partial\phi_i(a_1)/\partial x] = 0,$$

where $0 \leqq \alpha_1 \leqq 1$. From this it is easily seen that $h_i \neq 0$, except the case where $\phi_i(a_1) = \partial\phi_i(a_1)/\partial x = 0$. But this special case implies that $\phi_i(x) \equiv 0$, a contradiction. Hence $h_i \neq 0$, $i = 1, 2, \cdots$, and from Corollary 4.1, the null controllability of (38) and (39) can be concluded.

Example 2. We next consider the heat equation on a rectangle domain $0 < x_1 < \pi/a_1, 0 < x_2 < \pi/a_2$:

$$\partial u/\partial t = \partial^2 u/\partial x_1^2 + \partial^2 u/\partial x_2^2, \tag{41}$$

with the boundary condition

$$u(t, \xi) = h(\xi) f(t), \qquad \xi \in S. \tag{42}$$

This example was treated by Fattorini [4]. However, we apply our result to this problem.

It is easily seen that the eigenvalues are given by $\lambda_{n_1n_2} = (n_1^2a_1^2 + n_2^2a_2^2)$, $n_1, n_2 = 1, 2, \cdots$, and the corresponding eigenfunctions are [1]

$$\phi_{n_1n_2}(x) = \sin n_1a_1x_1 \sin n_2a_2x_2, \qquad n_1, n_2 = 1, 2, \cdots .$$

Assume a_1^2, a_2^2 are linearly independent over the integers; then the multiplicity in this case is one. Since the eigenfunctions satisfy the boundary condition $\phi_{n_1n_2}(\xi) = 0$, it follows from Corollary 4.1 that the control system described by (41) and (42) is null controllable if and only if

$$h_{n_1n_2} = \int_S \frac{\partial \phi_{n_1n_2}(\xi)}{\partial n} h(\xi)\, dS_\xi \neq 0,$$

for $n_1, n_2 = 1, 2, \cdots$, or

$$\begin{aligned} &- n_2a_2 \int_0^{\pi/a_1} h(x_1, 0) \sin n_1a_1x_1\, dx_1 \\ &- (-1)^{n_2}n_2a_2 \int_0^{\pi/a_1} h\left(x_1, \frac{\pi}{a_2}\right) \sin n_1a_1x_1\, dx_1 \\ &+ n_1a_1 \int_0^{\pi/a_2} h(0, x_2) \sin n_2a_2x_2\, dx_2 \\ &+ (-1)^{n_1}n_1a_1 \int_0^{\pi/a_2} h\left(\frac{\pi}{a_1}, x_2\right) \sin n_2a_2x_2\, dx_2 \neq 0, \\ &\qquad\qquad\qquad\qquad n_1, n_2 = 1, 2, \cdots. \end{aligned} \tag{43}$$

For example, the function $h(\xi)$ given by

$$h(x_1, 0) = \sum_{n=1}^{\infty} b_n \sin na_1x_1,$$

$$h(x_1, \pi/a_2) = h(0, x_2) = h(\pi/a_1, x_2) \equiv 0,$$

where $b_n \neq 0$, for all $n = 1, 2, \cdots$, satisfies (43) and makes the system controllable.

If a_1^2, a_2^2 are not linearly independent over the integers, i.e., if there exist some integers l and m such that $a_1^2 = (l/m)a_2^2$, then A can have infinite multiplicity. If $l = m$, then A does have infinite multiplicity. In this case, the system (41) is not controllable by finitely many functions of time.

Example 3. We examine the heat equation in the polar coordinate form defined on a unit circle:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}, \qquad 0 \leqq r < 1, \quad 0 \leqq \theta < 2\pi, \tag{44}$$

with the boundary condition

$$u|_{r=1} = h_1(\theta)f_1(t) + h_2(\theta)f_2(t). \tag{45}$$

It is easily calculated [1] that the eigenvalues of (44) with the homogeneous boundary condition $u|_{r=1} = 0$ are given by

$$\lambda_{nm} = \beta_{nm}^2, \qquad n = 0, 1, \cdots, \quad m = 1, 2, \cdots, \tag{46}$$

where β_{nm} are the nonzero real roots of the Bessel functions $J_n(\cdot)$ of nth order, i.e.,

$$J_n(\beta_{nm}) = 0, \qquad n = 0, 1, \cdots, \quad m = 1, 2, \cdots,$$

and that the corresponding eigenfunctions are given by

$$\phi_{0m}(r, \theta) = J_0(\beta_{0m}r), \qquad m = 1, 2, \cdots, \tag{47}$$

$$\begin{aligned} \phi_{nm1}(r, \theta) &= J_n(\beta_{nm}r) \cos n\theta, \\ \phi_{nm2}(r, \theta) &= J_n(\beta_{nm}r) \sin n\theta, \qquad n, m = 1, 2, \cdots. \end{aligned} \tag{48}$$

It is well known that each of the Bessel functions $J_n(x)$ has infinitely many real zeros and they are all different except $x = 0$. Clearly the multiplicities of the eigenvalues λ_{0m} are one, but the multiplicities of $\lambda_{nm}(n \geqq 1)$ are two. For that reason, we assumed two dimensional control functions in (45).

We apply Theorem 4 to this case. From Theorem 4, the control system described by (44) and (45) is null controllable if and only if

$$|H_{nm}| = \begin{vmatrix} h^1_{nm1} & h^1_{nm2} \\ h^2_{nm1} & h^2_{nm2} \end{vmatrix} \neq 0 \tag{49}$$

for all $n = 1, 2, \cdots$, and $m = 1, 2, \cdots$, and

$$|h^1_{0m}| + |h^2_{0m}| \neq 0 \tag{50}$$

for all $m = 1, 2, \cdots$, where

$$\begin{aligned} h^j_{nmi} &= \int_0^{2\pi} \frac{\partial \phi_{nmi}(1, \theta)}{\partial r} h_j(\theta)\, d\theta, \\ h^j_{0m} &= \frac{\partial \phi_{0m}(1)}{\partial r} \int_0^{2\pi} h_j(\theta)\, d\theta, \qquad i, j = 1, 2. \end{aligned} \tag{51}$$

It is easily seen that

$$\begin{aligned} |H_{nm}| = \beta_{nm}J_n'(\beta_{nm}) \Bigg[\int_0^{2\pi} h_1(\theta) \cos n\theta\, d\theta \int_0^{2\pi} h_2(\theta) \sin n\theta\, d\theta \\ - \int_0^{2\pi} h_1(\theta) \sin n\theta\, d\theta \int_0^{2\pi} h_2(\theta) \cos n\theta\, d\theta \Bigg], \end{aligned}$$

$$|h^1_{0m}| + |h^2_{0m}| = \beta_{0m}|J_0'(\beta_{0m})| \left[\left| \int_0^{2\pi} h_1(\theta)\, d\theta \right| + \left| \int_0^{2\pi} h_2(\theta)\, d\theta \right| \right].$$

Since the functions $h_j(\theta)$, $j = 1, 2$, are periodic, they can be expressed in the form of Fourier series:

$$h_j(\theta) = \tfrac{1}{2}a_{j0} + \sum_{n=1}^{\infty} (a_{jn} \cos n\theta + b_{jn} \sin n\theta). \tag{52}$$

Because all the values $J_n'(\beta_{nm})$ do not vanish, a necessary and sufficient condition for the null controllability is given by

$$\begin{aligned} a_{1n}b_{2n} - b_{1n}a_{2n} &\neq 0, \qquad n = 1, 2, \cdots, \\ |a_{10}| + |a_{20}| &\neq 0. \end{aligned} \tag{53}$$

For example, the following pair of functions

$$
\begin{aligned}
h_1(\theta) &= \tfrac{1}{2}a_{10} + \sum_{n=1}^{\infty} a_{1n}\cos n\theta, \\
h_2(\theta) &= \sum_{n=1}^{\infty} b_{2n}\sin n\theta,
\end{aligned} \tag{54}
$$

where $a_{10} \neq 0$, $a_{1n} \neq 0$, $b_{2n} \neq 0$, $n = 1, 2, \cdots$, satisfies the null controllability condition (53).

Remark 2. It is easily seen that the Laplacian in an n-dimensional sphere has infinite multiplicity if $n \geqq 3$. Therefore, it is important to note that the heat equation on the n-dimensional sphere ($n \geqq 3$) is not controllable by finitely many functions of time.

Remark 3. We have so far treated the heat equations as typical parabolic equations for avoiding complexity. However, since it is shown by Ito [8] that Theorems 1 and 2 still hold even in the case where, instead of the Laplacian, the elliptic differential operator A is self-adjoint and is given by

$$Au = \sum_{i=1}^{r} \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{r} a_{ij}(x)\frac{\partial u}{\partial x_j}\right) - q(x)u,$$

where the matrix $(a_{ij}(x))$ of the coefficient functions is strictly positive definite for all $x \in \bar{D}$, we can obtain the same results concerning controllability for self-adjoint parabolic equations of the general type.

Remark 4. Recently, Fattorini and Russell [5]–[7] considered controllability conditions which ensure that the system can be steered from the zero initial state to an arbitrary final state exactly. They gave a precise description of the attainable set. Of course, the exact controllability conditions in this mean are much more restrictive than those given here.

Acknowledgment. The author wishes to thank the referees for their helpful comments on an earlier manuscript.

REFERENCES

[1] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, vol. 1, Interscience, New York, 1966.

[2] H. O. Fattorini, *Some remarks on complete controllability*, this Journal, 4 (1966), pp. 686–694.

[3] ———, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.

[4] ———, *Boundary control systems*, this Journal, 6 (1968), pp. 349–385.

[5] H. O. Fattorini and D. L. Russell, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Rational Mech. Anal., 43 (1971), pp. 272–292.

[6] ———, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Quart. Appl. Math., to appear.

[7] D. L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Applied Mathematics, Society for Industrial and Appl. Math., Philadelphia, to appear.

[8] S. Ito, *Fundamental solutions of parabolic differential equations and boundary value problems*, Japan J. Math., 27 (1957), pp. 55–102.

[9] ———, *Partial Differential Equations*, Baifukan, Tokyo, 1966. (In Japanese.)

[10] K. Tsujioka, *Remarks on controllability of second order evolution equations in Hilbert spaces*, this Journal, 8 (1970), pp. 90–99.

SIAM J. CONTROL
Vol. 12, No. 3, August 1974

AN EXAMPLE OF OPTIMAL STOCHASTIC CONTROL WITH CONSTRAINTS*

JEAN-MICHEL BISMUT†

Abstract. The purpose of this paper is to apply a very simple form of the Hahn–Banach theorem to a problem of stochastic optimization with a supply constraint. The Lagrange multiplier associated with the constraint defines a stochastic process, the properties of which are extensively studied.

The purpose of this paper is to solve an apparently very simple problem of optimal stochastic control. More specifically, we wish to find an explicit solution for a problem of minimization of a linear functional with a quadratic supply constraint on the control.

To solve the problem, a stochastic Lagrange multiplier is introduced, in the space of additive measures. It is then proved that this multiplier defines a local martingale which appears in the Ito–Watanabe–Meyer multiplicative decomposition of the right-continuous supermartingale defined by the dynamic programming process (see [9]). The solution of the problem is then expressed very simply by means of this local martingale.

The procedure is presented in a more general duality framework for optimal stochastic control in [3]. In particular, by transforming very easily the problem which we solve into a very simple problem of optimal stochastic control, we show in [3] that a certain type of state constraint changes the Lagrange multiplier associated with some problems of optimal stochastic control, which is in many cases a semimartingale, into a local semimartingale. This can be compared with the results of deterministic control given in [11], where the Lagrange multiplier, which is generally an absolutely continuous function, is changed into a right-continuous bounded variation function, when state constraints are introduced.

To avoid excessive notational difficulties, we give none of these generalizations, and we refer to [3] for a more general treatment of this type of constraint. This will also enable us to concentrate on the probabilistic difficulties of the problem which are far more serious here.

All the basic results in probability theory which we will use can be found in [5], [6], [7] and [8].

1. The problem. Let V be a finite-dimensional space, let $(\Omega, \mathscr{F}, P)$ be a complete probability space, and let $\{\mathscr{F}_t\}_{t \in R^+}$ be an increasing right-continuous sequence of complete sub-σ-fields of $\mathscr{F}$ [5, Chap. IV, p. 30].

Let $\mathscr{T}$ be the σ-field in $\Omega \times [0, +\infty[$ of the well-measurable sets [5, Chap. VIII, D14] and $\mathscr{T}^*$ the completion of $\mathscr{T}$ for the measure $dP \otimes dt$. For our purpose,

* Received by the editors September 29, 1972, and in final revised form May 4, 1973.

† Department of Mathematics, Faculté des Sciences de Paris, Paris, France. This paper is based on the 7th chapter of a thesis to be submitted in Pure Mathematics at the Faculté des Sciences of Paris. It has been supported in part by I.R.I.A. (Institut de Recherches en Informatique et Automatique Domaine de Voluceau, 78 Rocquencourt, France).

and using the modification theorems of Meyer, we could have used for $\mathscr{T}$ the σ-field of measurable adapted sets [5, Chap. IV, D45], progressive sets [5, Chap. IV, D45], or predictable sets [7, no. 203], which have the same completion for $dP \otimes dt$ by [7, nos. 210–212–214 and 215].

Let X_t be a $\mathscr{T}^*$-measurable process with values in V, such that

$$E \int_0^{+\infty} |X_t|^2 \, dt < +\infty. \tag{1.1}$$

DEFINITION 1.1. L_{22} is the space of $dP \otimes dt$ classes of $\mathscr{T}^*$-measurable functions u with values in V, such that

$$E \int_0^{+\infty} |u_t|^2 \, dt < +\infty.$$

A norm is defined in L_{22} by

$$\|u\|_{22} = \left(E \int_0^{+\infty} |u_t|^2 \, dt \right)^{1/2}. \tag{1.2}$$

DEFINITION 1.2. H is the space of $dP \otimes dt$ classes of $\mathscr{T}^*$-measurable functions, with values in V, such that

$$\int_0^{+\infty} |u_t|^2 \, dt \quad \text{is in } L_\infty.$$[1]

DEFINITION 1.3. K is the subset of the elements u of H such that

$$\int_0^{+\infty} |u_t|^2 \, dt \leqq 1 \quad \text{a.s.} \tag{1.3}$$

H and K are obviously subsets of L_{22}.

Let I be the linear functional defined on L_{22} by

$$u \to E \int_0^{+\infty} \langle X_t, u_t \rangle \, dt. \tag{1.4}$$

DEFINITION 1.4. The problem of control which we want to solve is the minimization of I on K.

This problem is trivial in the deterministic case. The problem comes entirely from the stochastic nature of the problem, and more particularly from the information constraint.

2. Existence—Introduction of a Lagrange multiplier. The problem of existence is easy to solve.

PROPOSITION 2.1. *I has a minimum on K.*

Proof. L_{22} is a Hilbert space. K is nonempty (0 is in K), convex, bounded and closed in L_{22}. K is necessarily weakly compact. Since I is weakly continuous it has a minimum on K.

[1] L_∞ is the space of real random variables, which are essentially bounded.

THEOREM 2.1. *The problem of the minimization of I on K is equivalent to the search for a saddle point on*[2] $H \times (L_\infty)^*_+$ *of the functional*:

$$(2.1) \qquad (u, \mu) \underset{\mathscr{J}}{\to} I(u) + \left\langle \mu, \int_0^{+\infty} |u_t|^2 \, dt - 1 \right\rangle .$$

Proof. Let B be the mapping defined on H, with values in L_∞, by

$$(2.2) \qquad u \to \int_0^{+\infty} |u_t|^2 \, dt - 1.$$

If we give to L_∞ its natural order, B is convex, in the sense that for t in $[0, 1]$ and (u_1, u_2) in H, then

$$(2.3) \qquad B(tu_1 + (1 - t)u_2) \leqq tB(u_1) + (1 - t)B(u_2).$$

Moreover, K is the set of elements u in H such that

$$(2.4) \qquad B(u) \leqq 0.$$

Besides, we have

$$(2.5) \qquad B(0) < 0$$

because -1 is the interior of the cone of negative elements of L_∞.

We apply then the result given in [1]: since the generalized Slater hypothesis is verified, the initial problem is equivalent to the search for a saddle point of $\mathscr{J}$ on $H \times (L_\infty)^*_+$.

COROLLARY. (u^0, μ) *is a saddle point of* $\mathscr{J}$ *on* $H \times (L_\infty)^*_+$ *if and only if*

(i) u^0 *is in* K;

(ii)

$$(2.6) \qquad \left\langle \mu, \int_0^{+\infty} |u_t^0|^2 \, dt - 1 \quad = 0; \right\rangle$$

(iii) *for any* u *in* H, *one has*

$$(2.7) \qquad I(u^0) \leqq I(u) + \left\langle \mu, \int_0^{+\infty} |u_t|^2 \, dt - 1 \right\rangle .$$

Proof. This is obvious from the definition of a saddle point.

Remark. If u^1 is also an optimum of I on K, one has: $I(u^0) = I(u^1)$. Inequality (2.7) proves then that:

$$(2.8) \qquad \left\langle \mu, \int_0^{+\infty} |u_t^1|^2 - 1 \right\rangle = 0.$$

(u^1, μ) is then also a saddle point of $\mathscr{J}$. This proves that the set of saddle points of $\mathscr{J}$ may be written as $K' \times M'$, K' being the set of minimums of I on K, and M' being the corresponding set of Lagrange multipliers.

[2] $(L_\infty)^*_+$ is the set of positive elements of the strong dual of L_∞, which is the set of the finitely additive measures on $(\Omega, \mathscr{F}, P)$.

3. Local martingale associated with a Lagrange multiplier. Let (u^0, μ) be a saddle point of $\mathscr{J}$. σ_t^0 is the continuous process defined by

$$\sigma_t^0 = \int_0^t |u_s^0|^2 \, ds. \tag{3.1}$$

Let S^0 be the real-valued function defined on Ω by

$$S^0 = \inf\{t : \sigma_t^0 = 1\}. \tag{3.2}$$

PROPOSITION 3.1. *S^0 is a predictable stopping time* [7, *p.* 145, *no.* 105].

Proof. S^0 is a stopping time by [5, Chap. IV, p. 53]. Let $\{S_n^0\}_{n \in N}$ be the sequence of stopping times defined by

$$S_n^0 = \inf\{t : \sigma_t^0 \geqq 1 - 1/n\}. \tag{3.3}$$

Since σ^0 is continuous, one has

$$P(\lim_n S_n^0 = S^0) = 1; \tag{3.4}$$

and moreover on $(S^0 < +\infty)$ for any n, $S_n^0 < S^0$; S^0 is then predictable.

We keep the definition of the family $\{S_n^0\}_{n \in N}$ given in the proof of Proposition 3.1.

THEOREM 3.1. *For any stopping time S, the restriction of*[3]

$$(1 - \sigma_S^0)\mu$$

to $\mathscr{F}_S$ is defined by an element of $L_1^{\mathscr{F}_S}$.[4]

Proof. Let A be an $\mathscr{F}_S$-measurable set of Ω. Let u^A be defined by

$$u^A = \begin{cases} u^0, & 0 \leqq s < S, \\ 1_{CA} u^0, & S \leqq s < +\infty, \end{cases} \tag{3.5}$$

where 1_{CA} takes value 1 on CA and 0 on A. Then u^A is necessarily $\mathscr{T}^*$-measurable.

Moreover, the following relation holds:

$$\int_0^{+\infty} |u_t^A|^2 \, dt = 1_A \sigma_S^0 + 1_{CA} \sigma_\infty^0. \tag{3.6}$$

u_A is then in H. By using inequality (2.7) in the corollary of Theorem 2.1, we have

$$I(u^0) \leqq I(u^A) + \langle \mu, 1_A \sigma_S^0 + 1_{CA} \sigma_\infty^0 - 1 \rangle. \tag{3.7}$$

But we know that

$$\sigma_\infty^0 \leqq 1 \quad \text{a.s.} \tag{3.8}$$

Since μ is a positive additive measure, we have

$$I(u^0) \leqq I(u^A) + \langle \mu, 1_A \sigma_S^0 + 1_{CA} - 1 \rangle, \tag{3.9}$$

[3] This represents the linear functional defined on L_∞ by

$$x \to \langle \mu, (1 - \sigma_{t \wedge S_n^0}^0) x \rangle.$$

[4] If $\mathscr{F}'$ is a sub-sigma-algebra of $\mathscr{F}$, $L_1^{\mathscr{F}'}$ is the set of the integrable $\mathscr{F}'$-measurable random variables.

or equivalently,

$$I(u^0) \leqq I(u^A) + \langle \mu, 1_A(\sigma_S^0 - 1)\rangle. \tag{3.10}$$

Inequality (3.10) may be written

$$\langle (1 - \sigma_S^0)\mu, 1_A\rangle \leqq I(u_A) - I(u_0). \tag{3.11}$$

But then

$$I(u^A) - I(u^0) = -E1_A \int_S^{+\infty} \langle X_s, u_s^0\rangle\, ds. \tag{3.12}$$

Moreover,

$$\int_S^{+\infty} \langle X_s, u_s^0\rangle\, ds$$

is an integrable random variable.

Inequality (3.11) proves then that $(1 - \sigma_S^0)\mu$ is countably additive and absolutely continuous with respect to P when restricted to $\mathscr{F}_S$. Moreover,

$$\langle (1 - \sigma_S^0)\mu, 1\rangle \leqq \langle \mu, 1\rangle < +\infty. \tag{3.13}$$

By the Radon–Nikodym theorem, the theorem follows.

Let f_t^n be the Radon–Nikodym density of the restriction of $(1 - \sigma_{t\wedge S_n^0}^0)\mu$ to $\mathscr{F}_{t\wedge S_n^0}$. We have necessarily:

$$1 - \sigma_{t\wedge S_n^0}^0 \geqq 1/n. \tag{3.14}$$

Let us then define g_t^n by

$$g_t^n = f_t^n/(1 - \sigma_{t\wedge S_n^0}^0).$$

By (3.14), g_t^n is the restriction of μ to $\mathscr{F}_{t\wedge S_n^0}$, and is necessarily positive and integrable. In the same way, we define $\tilde{f}^n$, the restriction of $(1 - \sigma_{S_n^0}^0)\mu$ to $\mathscr{F}_{S_n^0}$, and $\tilde{g}^n$ by

$$\tilde{g}^n = \tilde{f}^n/(1 - \sigma_{S_n^0}^0).$$

Remark. It is essential to note that μ generally has no density relative to P, but is only an additive measure on $\mathscr{F}$. However, we have proved that its restriction to a family of sub-sigma-algebras of $\mathscr{F}$ has a density relative to P. We will use this fact to do calculations which would be trivial if the density of μ existed on $\mathscr{F}$. This is the main justification for introducing $\{S_n^0\}_{n\in N}$ and $(\tilde{g}^n\}_{n\in N}$.

PROPOSITION 3.2. *For any n in N, g_t^n is a positive uniformly integrable martingale, stopped at S_n^0.*

Proof. We need only to prove that for any t in $[0, +\infty)$,

$$g_t^n = E^{\mathscr{F}_{t\wedge S_n^0}}\tilde{g}^n. \tag{3.15}$$

Let A be $\mathscr{F}_{t\wedge S_n^0}$-measurable. Then,

$$E(1_A\tilde{g}^n) = \langle \mu, 1_A\rangle \tag{3.16}$$

and

$$\langle \mu, 1_A\rangle = E(1_A g_t^n). \tag{3.17}$$

Since g^n_t is $\mathscr{F}_{t \wedge S^0_n}$-measurable, (3.16) and (3.17) prove that (3.15) is true.
g^n_t is then a positive uniformly integrable martingale. This completes the proof.

We consider only the right-continuous version of the martingales which we study, which exist by [5, Chap. VI, T4].

THEOREM 3.2. *One can find one and only one local martingale g_t stopped at S^0, and continuous at S^0, such that*:

$$g_{t \wedge S^0_n} = g^n_t. \tag{3.18}$$

Proof. Let us prove first the following relation:

$$g^{n+1}_{t \wedge S^0_n} = g^n_t. \tag{3.19}$$

To prove (3.19), we need only to prove the following:

$$g^n_{S^0_n} = E^{\mathscr{F}_{S^0_n}} g^{n+1}_{S^0_{n+1}}. \tag{3.20}$$

Let A be an $\mathscr{F}_{S^0_n}$-measurable set of Ω. Then,

$$\langle \mu, 1_A \rangle = E(1_A g^{n+1}_{S^0_{n+1}}) = E(1_A g^n_{S^0_n}). \tag{3.21}$$

Thus (3.20) is proved.

We now define g_t on $[0, S^0_n[$ by

$$g_t = g^n_t. \tag{3.22}$$

Equation (3.19) proves that this definition is consistent on $t < S^0$, because S^0_n increases to S^0 as $n \to +\infty$.

g is then right-continuous on $[0, S^0[$. Let us prove that g_t has a limit when t increases to S^0.

Let h^n_t be defined by

$$h^n_t = 1_{\{t < S^0_n\}} g_t. \tag{3.23}$$

We have by definition

$$h^n_t = 1_{\{t < S^0_n\}} g^n_t. \tag{3.24}$$

Since g^n_t is right-continuous, h^n_t is right-continuous. Moreover, h^n_t is a supermartingale, because g^n_t is a martingale and $1_{\{t < S^0_n\}}$ is a decreasing process.

Then, for $t' \geqq t$,

$$E^{\mathscr{F}_t} h^n_{t'} \leqq h^n_t. \tag{3.25}$$

Since S^0_n increases to S^0, h^n_t converges a.s. to $1_{\{t < S^0\}} g_t$, and by applying Fatou's lemma to the increasing sequence h^n_t, we deduce from (3.25):

$$E^{\mathscr{F}_t} 1_{\{t' < S^0\}} g_{t'} \leqq 1_{\{t < S^0\}} g_t. \tag{3.26}$$

Inequality (3.26) proves that $1_{\{t < S^0\}} g_t$ is a positive supermartingale. Moreover it is right-continuous. By applying T3 of [5, Chap. VI], its trajectories have a.s. no oscillatory discontinuities. g_t has then a.s. a left-hand limit when t increases to S^0. We call g_{S^0} this limit, and for $t \geqq S^0$ we define g_t by

$$g_t = g_{S^0}. \tag{3.27}$$

Since $g_{t \wedge S^0_n}$ is a martingale for any n, one has for $t \leqq t'$:

$$E^{\mathscr{F}_t} g_{t' \wedge S^0_n} = g_{t \wedge S^0_n}. \tag{3.28}$$

But we know that by definition,

$$g_t = \lim_{n \to +\infty} g_{t \wedge S_n^0}. \tag{3.29}$$

Then by applying Fatou's lemma to (3.28), one gets

$$E^{\mathscr{F}_t} g_{t'} \leqq g_t. \tag{3.30}$$

Inequality (3.30) proves that g is a supermartingale, which is right-continuous.

By the Meyer–Ito–Watanabe decomposition of right-continuous supermartingales ([5, Chap. VII] and [4]), one can find an increasing natural process A and a local martingale M such that

$$g_t = M_t - A_t. \tag{3.31}$$

But

$$g_t = g_{t \wedge S^0}. \tag{3.32}$$

Then,

$$g_t = M_{t \wedge S^0} - A_{t \wedge S^0}. \tag{3.33}$$

Since $M_{t \wedge S^0}$ is a local martingale, and since $A_{t \wedge S^0}$ is an increasing natural process, the uniqueness of the decomposition (3.31) proves that M and A are stopped at S^0. But $g_{t \wedge S_n^0}$ is a martingale. Then $A_{S_n^0}$ is necessarily equal to 0. Since A is stopped at S^0, A can only consist of a jump at S^0. But A_{S^0} is measurable with respect to the σ-algebra $\mathscr{A}$ defined by

$$\mathscr{A} = \bigvee_{1}^{+\infty} \mathscr{F}_{S_n^0} \tag{3.34}$$

(see [5, Chap. VII, T49]). g being left-continuous at S^0 is $\mathscr{A}$-measurable. M_{S^0} is then necessarily $\mathscr{A}$-measurable. Then—the proof can be deduced from the method used in the proof of T47 in [5, Chap. VII]—M is left-continuous at S^0. A is then left-continuous at S^0. Then necessarily:

$$A_{S^0} = 0. \tag{3.35}$$

Since A is stopped at S^0, A is the null process. Then

$$g = M \tag{3.36}$$

and g is a local martingale.

Remark. The end of the proof is inspired by the proof given by Meyer in [9], for the multiplicative decomposition of right-continuous positive supermartingales, although the origin of the problem is completely different. The relationship with this decomposition will be proved in the sequel.

Let us notice also that, even if μ is not directly related to a particular u^0, S^0 depending on u^0, g depends, at least at first sight, on u^0 (see Corollary 1 of Theorem 4.3).

PROPOSITION 3.3. *For any n and for any u in H, one has*

$$I(u^0) \leqq I(1_{\{t < S_n^0\}} u) + E \int_0^{S_n^0} g_t |u_t|^2 \, dt - E(g_0). \tag{3.37}$$

Proof. Let us apply (2.7) to $1_{\{t<S_n^0\}}u$:

$$I(u^0) \leqq E\int_0^{S_n^0} \langle X_t, u_t\rangle\, dt + \left\langle \mu, \int_0^{S_n^0} |u_t|^2\, dt - 1\right\rangle. \tag{3.38}$$

But the restriction of μ to $\mathscr{F}_{S_n^0}$ is $g_{S_n^0}$ by Theorem 3.1 and Proposition 3.2. Then,

$$\left\langle \mu, \int_0^{S_n^0} |u_t|^2\, dt - 1\right\rangle = E\left(g_{S_n^0}\left(\int_0^{S_n^0} |u_t|^2\, dt - 1\right)\right). \tag{3.39}$$

By using T16 of [5, Chap. VII], one has, because of the uniform integrability of the martingale $g_{t\wedge S_n^0}$,

$$E\left(g_{S_n^0}\int_0^{S_n^0} |u_t|^2\, dt - 1\right) = E\int_0^{S_n^0} g_t|u_t|^2\, dt - E(g_0). \tag{3.40}$$

Relation (3.37) follows from (3.38), (3.39) and (3.40).

THEOREM 3.3. *For u in L_{22}, the following relation holds*:

$$I(u^0) \leqq I(1_{\{t<S^0\}}u) + E\int_0^{S^0} g_t|u_t|^2\, dt - E(g_0). \tag{3.41}$$

Moreover,

$$E\int_0^{+\infty} g_t|u_t^0|^2\, dt = E(g_0). \tag{3.42}$$

Proof. Let R_n be the sequence of stopping times defined by

$$R_n = \inf\left\{t\colon \int_0^t |u_t|^2\, dt = n\right\}. \tag{3.43}$$

Then $1_{\{t<R_n\}}u$ is in H. We apply then (3.37) to $1_{\{t<R_n\}}u$.

$$I(u^0) \leqq I(1_{\{t<R_n\wedge S_n^0\}}u) + E\int_0^{R_n\wedge S_n^0} g_t|u_t|^2\, dt - E(g_0). \tag{3.44}$$

When n increases to infinity, by taking the limit in (3.44), one gets

$$I(u^0) \leqq I(1_{\{t<S^0\}}u) + E\int_0^{S^0} g_t|u_t|^2\, dt - E(g_0). \tag{3.45}$$

Moreover, since $g_{t\wedge S_n^0}$ is a uniformly integrable martingale,

$$\begin{aligned} \langle \mu, \sigma^0_{S_n^0}\rangle &= E\left(g_{S_n^0}\int_0^{S_n^0} |u_t^0|^2\, dt\right) \\ &= E\int_0^{S_n^0} g_t|u_t^0|^2\, dt. \end{aligned} \tag{3.46}$$

But by definition, the following relation holds:

$$|\sigma^0_{S^0} - \sigma^0_{S_n^0}| \leqq 1/n \quad \text{a.s.} \tag{3.47}$$

$\sigma^0_{S^0_n}$ converges to $\sigma^0_{S^0}$ in L_∞. We can then take the limit in (3.46), and get[5]

$$\langle \mu, \sigma^0_{S^0} \rangle = E \int_0^{S^0} g_t |u^0_t|^2 \, dt. \tag{3.48}$$

But we have also:

$$\langle \mu, \sigma^0_{S^0} \rangle = \langle \mu, 1 \rangle = E(g_0). \tag{3.49}$$

Then since u^0 is null for $t \geqq S^0$,

$$E \int_0^{+\infty} g_t |u^0_t|^2 \, dt = E(g_0). \tag{3.50}$$

4. Expression of an optimal control.

PROPOSITION 4.1.

$$\begin{aligned} X_t + 2g_t u^0_t &= 0 \quad \textit{for } t < S^0, \\ u^0_t &= 0 \quad \textit{for } t \geqq S^0. \end{aligned} \tag{4.1}$$

Proof. (3.41) says that the functional defined on L_{22},

$$u \underset{I'}{\rightarrow} E \int_0^{S^0} \langle X_t, u_t \rangle \, dt + E \int_0^{S^0} g_t |u_t|^2 \, dt, \tag{4.2}$$

has a minimum at u^0.

By using the result given in [10] on the subdifferential of convex integrands one gets immediately

$$1_{\{t < S^0\}}(X_t + 2g_t u^0_t) = 0. \tag{4.3}$$

For $t \geqq S^0$, u^0 is necessarily null. This completes the proof.

We now have two difficulties. First we do not know if in (4.1), u^0_t can be written as $-X_t/(2g_t)$, because we have still no idea of the times when g_t is null. Second, even if μ does not depend on a particular optimal u^0, g_t depends on u^0 through S^0. If we want to find unambiguously all the optimal controls, we must improve our results on g_t, and we have to study carefully its zeros. To do this, we introduce new processes associated with the optimization problem.

Let Z^0_t be the process defined by

$$Z^0_t = -\frac{1}{2} E^{\mathscr{F}_t} \int_t^{+\infty} \langle X_s, u^0_s \rangle \, ds. \tag{4.4}$$

By writing

$$Z^0_t = -\frac{1}{2} E^{\mathscr{F}_t} \int_0^{+\infty} \langle X_s, u^0_s \rangle \, ds + \frac{1}{2} \int_0^t \langle X_s, u^0_s \rangle \, ds, \tag{4.5}$$

and by noticing that by (4.1),

$$\langle X_s, u^0_s \rangle \leqq 0, \tag{4.6}$$

Z^0_t is then a right-continuous supermartingale.

[5] The argument on the limit of the left-hand member of (3.46) is crucial. The proof can however use relations (3.11) and (3.12).

PROPOSITION 4.2. *For any n, we can identify*[6] *the processes* $Z^0_{t\wedge S^0_n}$ *and*

$$(1 - \sigma^0_{t\wedge S^0_n})g_{t\wedge S^0_n}$$

Proof. Proposition 4.1 proves that

$$\langle X_s, u^0_s\rangle + 2g_s|u^0_s|^2 = 0. \tag{4.7}$$

From (4.7), we deduce

$$E^{\mathscr{F}_{t\wedge S^0_n}}\int_{t\wedge S^0_n}^{+\infty}\langle X_s, u^0_s\rangle\, ds + 2E^{\mathscr{F}_{t\wedge S^0_n}}\int_{t\wedge S^0_n}^{+\infty} g_s|u_s|^2\, ds = 0. \tag{4.8}$$

By using (4.5) and [5, Chap. VI, T13 and R14a], we see that

$$Z^0_{t\wedge S^0_n} = -\frac{1}{2}E^{\mathscr{F}_{t\wedge S^0_n}}\int_{t\wedge S^0_n}^{+\infty}\langle X_s, u^0_s\rangle\, ds. \tag{4.9}$$

Then (4.8) and (4.9) prove that

$$Z^0_{t\wedge S^0_n} = E^{\mathscr{F}_{t\wedge S^0_n}}\int_{t\wedge S^0_n}^{+\infty} g_s|u^0_s|^2\, ds. \tag{4.10}$$

Let A be $\mathscr{F}_{t\wedge S^0_n}$-measurable. Using (2.6), we have

$$0 \leqq \langle\mu, 1_A(1 - \sigma^0_{S^0})\rangle \leqq \langle\mu, 1 - \sigma^0_{S^0}\rangle = 0. \tag{4.11}$$

The following relation holds:

$$E(1_A g_{t\wedge S^0_n}(1 - \sigma^0_{t\wedge S^0_n})) = \langle\mu, 1_A(1 - \sigma^0_{t\wedge S^0_n})\rangle. \tag{4.12}$$

By using (4.11), we have

$$E(1_A g_{t\wedge S^0_n}(1 - \sigma^0_{t\wedge S^0_n})) = \langle\mu, 1_A(\sigma^0_{S^0} - \sigma^0_{t\wedge S^0_n})\rangle. \tag{4.13}$$

By definition, we know that

$$|\sigma^0_{S^0} - \sigma^0_{S^0_m}| \leqq 1/m. \tag{4.14}$$

We deduce then, because of the continuity of the linear functional determined by μ with respect to the L_∞-topology,

$$\langle\mu, 1_A(\sigma^0_{S^0} - \sigma^0_{t\wedge S^0_n})\rangle = \lim_{\substack{m\to+\infty\\ m\geqq n}}\langle\mu, 1_A(\sigma^0_{S^0_m} - \sigma^0_{t\wedge S^0_n})\rangle. \tag{4.15}$$

But by Theorem 3.1, for $m \geqq n$,

$$\langle\mu, 1_A(\sigma^0_{S^0_m} - \sigma^0_{t\wedge S^0_n})\rangle = E(1_A g_{S^0_m}(\sigma^0_{S^0_m} - \sigma^0_{t\wedge S^0_n})). \tag{4.16}$$

Proposition 3.2 and [5, Chap. VII, T16] prove that

$$E(1_A g_{S^0_m}(\sigma^0_{S^0_m} - \sigma^0_{t\wedge S^0_n})) = E\left(1_A\int_{t\wedge S^0_n}^{S^0_m} g_s|u^0_s|^2\, ds\right). \tag{4.17}$$

[6] We identify two processes when they have a.s. the same trajectories.

Moreover, by Fatou's lemma, one has

$$E\left(1_A \int_{t\wedge S_n^0}^{S^0} g_s|u_s^0|^2\,ds\right) = \lim_{\substack{m\to+\infty\\ m\geqq n}} E\left(1_A \int_{t\wedge S_n^0}^{S_m^0} g_s|u_s^0|^2\,ds\right). \tag{4.18}$$

By comparing (4.13), (4.15), (4.17) and (4.18), we see that

$$E(1_A g_{t\wedge S_n^0}(1 - \sigma^0_{t\wedge S_n^0})) = E\left(1_A \int_{t\wedge S_n^0}^{S^0} g_s|u_s^0|^2\,ds\right). \tag{4.19}$$

Equation (4.19) proves then that

$$g_{t\wedge S_n^0}(1 - \sigma^0_{t\wedge S_n^0}) = E^{\mathscr{F}_{t\wedge S_n^0}} \int_{t\wedge S_n^0}^{S^0} g_s|u_s^0|^2\,ds. \tag{4.20}$$

Since u_t^0 is null for $t \geqq S^0$, and comparing (4.10) and (4.20), we have

$$Z^0_{t\wedge S_n^0} = g_{t\wedge S_n^0}(1 - \sigma^0_{t\wedge S_n^0}). \tag{4.21}$$

The result follows from the right-continuity of the two processes.

THEOREM 4.1. *We can identify the processes* Z_t^0 *and* $(1 - \sigma_t^0)g_t$.

Proof. For any n, we know that, by Proposition 4.2,

$$Z^0_{S_n^0} = (1 - \sigma^0_{S_n^0})g_{S_n^0}. \tag{4.22}$$

On $(S^0 < +\infty)$, we have necessarily

$$(1 - \sigma^0_{S_n^0}) \to 0. \tag{4.23}$$

Since g has a left-hand limit at S^0, one deduces, for $(S^0 < +\infty)$,

$$Z^0_{S^0-} = 0. \tag{4.24}$$

By [5, Chap. VII, T15], Z^0 is necessarily stopped at S^0. The theorem is proved.

The relation given in Theorem 4.1 is not completely satisfactory, because Z^0 depends, at least apparently on a particular optimal solution u^0. But the following result can be proved.

LEMMA 4.1. Z_0^0 *does not depend on a particular optimal solution.*

Proof. The proof is elementary, and is left to the reader.

u^{0t} is now defined in the following way:

u^{0t} minimizes

$$u \to E\int_t^{+\infty} \langle X_s, u_s\rangle\,ds \tag{4.25}$$

on the set of elements u of L_{22} such that:

$$\int_t^{+\infty} |u_s|^2\,ds \leqq 1 \quad \text{a.s.} \tag{4.26}$$

We define Z_t^1 similarly by

$$Z_t^1 = -\frac{1}{2}E^{\mathscr{F}_t}\int_t^{+\infty} \langle X_s, u_s^{0t}\rangle\,ds. \tag{4.27}$$

Z_t^2 and Z_t^3 are then defined by

$$Z_t^2 = \frac{1}{2} E^{\mathscr{F}_t}\left(\int_t^{+\infty} |X_s|^2 \, ds \right)^{1/2}, \tag{4.28}$$

$$Z_t^3 = \frac{1}{2} E^{\mathscr{F}_t} \int_t^{+\infty} e^{-s} |X_s| \, ds. \tag{4.29}$$

PROPOSITION 4.3. *Z_t^1, Z_t^2 and Z_t^3 are right-continuous positive supermartingales, converging to* 0 *when $t \to +\infty$. Moreover,*

$$Z_t^3 \leqq Z_t^1 \leqq Z_t^2, \tag{4.30}$$

Proof. We prove here only that Z_t^1 is a right-continuous supermartingale. If we define $u^{0t'}$ on $[t, t'[$ by giving to it the value 0 on this interval, we have necessarily

$$\int_t^{+\infty} |u_s^{0t'}|^2 \, ds \leqq 1 \quad \text{a.s.} \tag{4.31}$$

Then we will have

$$E^{\mathscr{F}_t} \int_t^{+\infty} \langle X_s, u_s^{0t} \rangle \, ds \leqq E^{\mathscr{F}_t} \int_t^{+\infty} \langle X_s, u_s^{0t'} \rangle \, ds, \tag{4.32}$$

or equivalently,

$$Z_t^1 \geqq E^{\mathscr{F}_t} Z_{t'}^1. \tag{4.33}$$

By [5, Chap. VI, T4], we have to prove that $t \to E(Z_t^1)$ is right-continuous. We have necessarily

$$E(Z_t^1) \geqq E(Z_t^1)^+ . \tag{4.34}$$

Moreover, by the Lebesgue theorem, we know that

$$E \int_t^{+\infty} \langle X_s, u_s^{0t} \rangle \, ds = \lim_{\substack{t' > t \\ t' \to t}} \downarrow E \int_{t'}^{+\infty} \langle X_s, u_s^{0t} \rangle \, ds. \tag{4.35}$$

But necessarily,

$$\int_{t'}^{+\infty} |u_s^{0t}|^2 \, ds \leqq 1. \tag{4.36}$$

Then,

$$E \int_{t'}^{+\infty} \langle X_s, u_s^{0t'} \rangle \, ds \leqq E \int_{t'}^{+\infty} \langle X_s, u_s^{0t} \rangle \, ds. \tag{4.37}$$

By using (4.35) and (4.37), we get

$$E(Z_t^1) \leqq E(Z_t^1)^+ . \tag{4.38}$$

By comparing (4.34) and (4.38), $E(Z_t^1)$ is right-continuous. Moreover,

$$Z_t^1 \leqq E^{\mathscr{F}_t} \left\{ \left(\int_t^{+\infty} |u_s^{0t}|^2 \, ds \right)^{1/2} \left(\int_t^{+\infty} |X_s|^2 \, ds \right)^{1/2} \right\}, \tag{4.39}$$

and then necessarily,

$$Z_t^1 \leqq Z_t^2. \tag{4.40}$$

Moreover, let $\tilde{u}$ be defined by

$$\tilde{u}_s = -(\operatorname{sgn} X_s)\, e^{-s}. \tag{4.41}$$

Then $\tilde{u}$ is in L_{22}, and moreover,

$$\int_0^{+\infty} |\tilde{u}_s|^2\, ds \leqq 1 \quad \text{a.s.} \tag{4.42}$$

We see then immediately that

$$Z_t^3 \leqq Z_t^1. \tag{4.43}$$

Remark. The main reason why we introduce the process Z_t^1 is because it does not depend explicitly on any particular choice of an optimal control. Moreover, if X_t is a Markov process, Z_t^1 is the natural dynamic programming function.

We are going to see how g_t and Z_t^1 are associated. Since Z_t^1 does not depend on a given optimal control, we will be able to show that g_t is a process which does not depend on u^0.

PROPOSITION 4.4. *We can identify the processes Z_t^0 and $(1 - \sigma_t^0)^{1/2} Z_t^1$.*

Proof. Necessarily,

$$\int_t^{+\infty} |u_s^0|^2\, ds \leqq 1 - \sigma_t^0. \tag{4.44}$$

If $\tilde{u}^t$ is defined by

$$\tilde{u}_s^t = \begin{cases} u_s^0(1 - \sigma_t^0)^{-1/2} & \text{on } t \leqq s < S^0, \\ 0 & \text{on } s \geqq S^0, \end{cases} \tag{4.45}$$

necessarily,

$$\int_t^{+\infty} |\tilde{u}_s^t|^2\, ds \leqq 1 \quad \text{a.s.}; \tag{4.46}$$

then

$$E^{\mathscr{F}_t} \int_t^{+\infty} \langle X_s, u_s^{0t} \rangle\, ds \leqq E^{\mathscr{F}_t} \int_t^{+\infty} \langle X_s, \tilde{u}_s^t \rangle\, ds \tag{4.47}$$

and (4.47) may be written, because Z_t^0 is null when $(1 - \sigma_t^0)$ is null,

$$Z_t^0 \leqq (1 - \sigma_t^0)^{1/2} Z_t^1. \tag{4.48}$$

Let us suppose that (4.49) holds with strict inequality on a nonnegligible $\mathscr{F}_t$-measurable set A. Let u^1 be defined by

$$u_s^1 = \begin{cases} u_s^0 \quad \text{for } 0 \leqq s < t, \\ (1 - \sigma_t^0)^{1/2} u_s^{0t} \quad \text{on } A, t \leqq s < +\infty, \\ u_s^0 \quad \text{on } CA. \end{cases} \tag{4.49}$$

Then,

$$\int_0^{+\infty} |u_s^1|^2\,ds \leqq 1 \quad \text{a.s.} \tag{4.50}$$

Moreover, by (4.48) and (4.49), we have

$$\begin{aligned} E\int_0^t \langle X_s, u_s^1\rangle\,ds &= E\int_0^t \langle X_s, u_s^0\rangle\,ds, \\ E\int_t^{+\infty} \langle X_s, u_s^1\rangle\,ds &< E\int_t^{+\infty} \langle X_s, u_s^0\rangle\,ds. \end{aligned} \tag{4.51}$$

Then one would have

$$I(u^1) < I(u^0). \tag{4.52}$$

This is a contradiction. The result follows by the right-continuity of the processes which we have considered.

THEOREM 4.2. *We can identify the processes* Z_t^1 *and* $(1 - \sigma_t^0)^{1/2}g_t$.

Proof. By Theorem 4.1 and Proposition 4.4, one has

$$\begin{aligned} Z^0_{t\wedge S_n^0} &= (1 - \sigma^0_{t\wedge S_n^0})g_{t\wedge S_n^0}, \\ Z^0_{t\wedge S_n^0} &= (1 - \sigma^0_{t\wedge S_n^0})^{1/2}Z^1_{t\wedge S_n^0}. \end{aligned} \tag{4.53}$$

Since $1 - \sigma^0_{t\wedge S_n^0}$ is larger than $1/n$, one then has

$$Z^1_{t\wedge S_n^0} = (1 - \sigma^0_{t\wedge S_n^0})^{1/2}g_t. \tag{4.54}$$

The proof is completed as the proof of Theorem 4.1.

Now let T be the stopping time defined by

$$T = \inf\{t : Z_t^1 = 0\}. \tag{4.55}$$

PROPOSITION 4.5. $1_{\{t \geqq T\}}X_t = 0\, dP \otimes dt$ *a.s.*

Proof. Inequality (4.30) proves that $Z_T^3 = 0$. But

$$Z_T^3 = \frac{1}{2}E^{\mathcal{F}_T}\int_T^{+\infty} e^{-s}|X_s|\,ds. \tag{4.56}$$

The result follows.

Remark. Conversely, (4.30) proves that T is the first time at which one is sure that the function $1_{\{t\geqq T\}}X_t$ is null: this follows from

$$Z_t^1 \leqq Z_t^2. \tag{4.57}$$

The result of Ito–Watanabe–Meyer given in [9] says then that the right-continuous positive supermartingale Z_t^1 can be decomposed in a unique way into the product of M and B, M and B being such that:

- M is a local martingale stopped at T, such that if the additive decomposition of Z^1 is written

$$Z_t^1 = N_t - A_t, \tag{4.58}$$

N_t being a local martingale and A_t an increasing natural process, then if we define[7]

$$\begin{aligned} &v_c = \inf\{t : Z^1_{t-} = 0\}, \\ (4.59) \qquad &v_1 = \inf\{t : Z^1_{t-} - \Delta A_t = 0, \Delta A_t > 0\}, \\ &v' = v_c \wedge v_1, \end{aligned}$$

M is continuous at v'.

- B is a decreasing natural right-continuous positive process such that $B_0 = 1$, stopped at T, continuous at v_c.

Moreover, if $M'_t B'_t$ is any other decomposition of Z^1_t into the product of a local martingale M' and a decreasing right-continuous natural process B' such that $B'_0 = 1$, on $t < T$, $B'_t = B_t$, $M'_t = M_t$.

But Theorem 4.2 gives precisely such a decomposition.

We then have the following theorem.

THEOREM 4.3. *For u^0 to minimize I on K, it is necessary and sufficient that u^0 satisfy the following properties*:

(a) *on* $0 \leqq t < T$, $u^0_t = -X_t/(2M_t)$;

(b) *on* $t \geqq T$, u_t *can take any family of values compatible with the constraint*

$$\int_0^{+\infty} |u^0_t|^2\, ds \leqq 1.$$

Moreover, $1_{\{t \geqq T\}} X_t = 0\, dP \otimes dt$ *a.s.*

In particular, it is possible to choose u^0 such that if $Z_{T-} \neq 0$,

$$\int_0^{+\infty} |u^0_t|^2\, dt < 1.$$

Proof. Theorem 4.2 implies that $T \leqq S^0$. Moreover, the previous result states that $g_t = M_t$ on $t < T$. By Proposition 4.1, for $t < T$, we have

$$X_t + 2M_t u^0_t = 0. \tag{4.60}$$

Moreover, M_t is nonnull for $t < T$, and we can write

$$u^0_t = -X_t/(2M_t). \tag{4.61}$$

Besides, by Proposition 4.5, $1_{\{t \geqq T\}} X_t = 0$. For $t \geqq T$, we can then choose any u^0 compatible with the constraints, without changing the value of the criteria. In particular, we can take u^0 null for $t \geqq T$. For such a choice of u^0, if $Z_{T-} \neq 0$, necessarily

$$\int_0^{+\infty} |u^0_t|^2\, dt < 1. \tag{4.62}$$

COROLLARY 1. $g = M$.

Proof. g is stopped at T. To prove this, we know that $T \leqq S^0$. If $T = S^0$, g being stopped at S^0 is also stopped at T. If $T < S^0$, $(1 - \sigma^0_T)^{1/2} > 0$. Theorem 4.2 proves then that $g_T = 0$. Sine g is a right-continuous positive supermartingale, by [5, Chap. VII, T15], g is stopped at T.

[7] ΔA_t is equal to $A_t - A_{t-}$.

Using the formula for the change of variables given in [8], since $B'_t = (1 - \sigma_t^0)^{1/2}$ is a continuous decreasing process, one has

$$Z_t^1 = Z_0^1 + \int_0^t B'_s \, dg_s + \int_0^t g_s \, dB'_s. \tag{4.63}$$

But this is precisely the additive decomposition of Z^1, with

$$A_t = - \int_0^t g_s \, dB'_s. \tag{4.64}$$

A is then continuous, and $\nu_1 = +\infty$. To prove that g is continuous at ν', we will prove only that it is continuous at ν_c.

If $T = S^0$, $\nu_c = S^0$, and by construction, g is continuous at S^0.

If $T < S^0$, let us suppose that $Z_{T-} = 0$. Then, because $1 - \sigma_T > 0$, g_{T-} is necessarily null.

Moreover, by writing

$$B''_t = B'_{t \wedge T}, \tag{4.65}$$

we have, since g is stopped at T,

$$Z_t^1 = g_t B''_t. \tag{4.66}$$

g and B'' have all the properties of M and B. By the uniqueness result, they are equal respectively to M and B.

Remark. B'' is precisely associated with the particular u^0 defined in the proof of Theorem 4.3. Moreover, this corollary gives the important result that g, which could have depended on the particular u^0, was actually a fixed process. Finally, let us note that in general T and S^0 are not equal: in particular, if T is totally inaccessible [5, Chap. VII, D42], $T < S^0$ a.s.

COROLLARY 2. *An expression for M is*

$$M_t = \begin{cases} Z_t^1 \exp \dfrac{1}{8} \displaystyle\int_0^t \left| \dfrac{X_s}{Z_s^1} \right|^2 ds & \textit{for } 0 \leqq t < T, \\ M_{T-} \quad \textit{for } t \geqq T \textit{ if } Z_{T-}^1 = 0, \\ 0 \qquad \textit{for } t \geqq T \textit{ if } Z_{T-}^1 \neq 0. \end{cases} \tag{4.67}$$

Proof. This result can be deduced from [9]. Let us prove directly. For $t < T$, we have

$$\sigma_t^0 = \frac{1}{4} \int_0^t \left| \frac{X_s}{M_s} \right|^2 ds. \tag{4.68}$$

But we know that

$$(Z_t^1)^2 = M_t^2 (1 - \sigma_t^0). \tag{4.69}$$

Then, on $s < T$, the following relation holds:

$$\left| \frac{X_s}{M_s} \right|^2 = \left| \frac{X_s}{Z_s^1} \right|^2 (1 - \sigma_s^0). \tag{4.70}$$

By using (4.70) in (4.68), we see that

$$\frac{\dot{\sigma}_t^0}{1 - \sigma_t^0} = \frac{1}{4}\left|\frac{X_t}{Z_t^1}\right|^2. \tag{4.71}$$

From (4.71), one gets for $t < T$:

$$1 - \sigma_t^0 = \exp -\frac{1}{4}\int_0^t \left|\frac{X_s}{Z_s^1}\right|^2 ds. \tag{4.72}$$

Then, for $t < T$, from Theorem 4.2 and Corollary 1 of Theorem 4.3, we obtain (4.67). If $Z_{T^-} = 0$, M is continuous at T. If $Z_{T^-} > 0$, $1 - \sigma_T^0 > 0$ and $M_T = 0$.

Remark. Formula (4.72) allows us to give an intuitive explanation of some of the results.

If $Z_{T^-} > 0$, it is easily proved that each trajectory has a strictly positive lower bound, and

$$1 - \sigma_T^0 > 0.$$

But even if $Z_{T^-} = 0$, it may happen that $1 - \sigma_T^0 > 0$, in particular in the case where X_s is a.e. equal to zero on a left-hand neighborhood of T.

This corresponds to the case where, although the predictions were "optimistic" ($Z_s > 0$ for $s < T$), X_s has in fact taken null values before T.

There is intuitively (and mathematically) a basic difference between these two cases: in the first case, the "bad luck" was unpredictable or was just instantaneous betting. In the second case, the facts have contradicted optimistic predictions. In the two cases, there are idle resources left, or useless resources.

Conclusion. We notice that the first three sections are almost completely independent of the linear nature of the criteria, and of the quadratic nature of the constraint.

In § 4, on the contrary, these properties are constantly used. One of the most striking features of the problem is its close relationship with the multiplicative decomposition of positive right-continuous supermartingales, the properties of which are very much used.

We refer for applications to [3].

Acknowledgment. The author is indebted to a referee for very helpful comments and suggestions.

REFERENCES

[1] A. Bensoussan, J.-L. Lions and R. Temam, *Décomposition des problemes d'optimisation*, to appear.

[2] J. M. Bismut, *Conjugate convex functions in optimal stochastic control*, J. Math. Anal. Appl., to appear.

[3] ———, *Analyse convexe et probabilités*, Doctoral dissertation, Faculté des Sciences de Paris, to appear.

[4] K. Ito and S. Watanabe, *Transformation of Markov processes by multiplicative functionals*, Ann. Inst. Fourier Grenoble, 15 (1965), pp. 13–30.

[5] P. A. Meyer, *Probabilités et Potentiels*, Hermann, Paris, 1966; English transl., Blaisdell, Boston, 1969.

[6] ———, *Intégrales stochastiques*, *Séminaire de probabilités* 1, Lecture Notes in Mathematics, no. 39, Springer-Verlag, Berlin, 1967, pp. 72–162.

[7] ———, *Guide détaillé de la théorie générale des processus. Séminaire de probabilités* 2, Lecture Notes in Mathematics, no. 51, Springer-Verlag, Berlin, 1968, pp. 140–165.

[8] P. A. MEYER AND DADE C. DOLÉANS, *Intégrales stochastiques par rapport aux martingales locales*, Séminaire de probabilités, no. 4, Lecture Notes in Mathematics, no. 124, Springer-Verlag, Berlin, 1970, pp. 77–107.

[9] ———, *On the multiplicative decomposition of positive supermartingales*, Markov Processes and Potential Theory, J. Chover, ed., John Wiley, New York, 1967, pp. 103–116.

[10] R. T. ROCKAFELLAR, *Integrals which are convex functionals. I*, Pacific J. Math., 24 (1968), pp. 525–539.

[11] ———, *State constraints in convex problems of Bolza*, to appear.

SIAM J. CONTROL
Vol. 12, No. 3, August 1974

N-PERSON LINEAR-QUADRATIC DIFFERENTIAL GAMES WITH CONSTRAINTS*

RICHARD C. SCALZO†

Abstract. Since 1970, it has been known that open loop Nash equilibrium points exist for N-person differential games when there are integral bounds on the control functions. These results were obtained using various fixed-point theorems, and required that the duration of the game be "small".

In this paper it is assumed that the controls are constrained to take values in compact, convex subsets of R^{q_i}. A fixed-point theorem due to Tychonov is used, and as a result there are no restrictions on the duration of the game.

1. Introduction. The approach to differential games used in this paper is that of A. Friedman, and the notation is taken from [1].

In this paper it is shown that open loop Nash equilibrium points exist for N-person linear-quadratic differential games with the controls constrained to take values in compact, convex subsets of R^{q_i}. The duration of the game considered is arbitrary.

In the case of integral bounds on the controls, open loop Nash equilibrium points are already known to exist; see [5], [6]. In all these papers the duration of the game is "small".

2. Statement of the main result. Consider the N-person differential game with dynamics

$$\frac{dx}{dt} = A(t)x + \sum_{i=1}^{N} B_i(t)u_i + f(t), \qquad t \in [0, T], \quad x \in R^m, \tag{2.1}$$

$$x(0) = x_0, \tag{2.2}$$

where $A(t)$ is a continuous $m \times m$ matrix, the $B_i(t)$ are continuous $m \times q_i$ matrices, and $f(t)$ is a given continuous function. The control functions $u_i(t)$ are measurable functions satisfying

$$u_i(t) \in U_i, \quad \text{a.e. for } t \in [0, T], \qquad i = 1, 2, \cdots, N, \tag{2.3}$$

where $U_i \subset R^{q_i}$ are compact and convex. The cost functionals are given by

$$\begin{aligned} J_i(u_1, \cdots, u_N) &= \langle x(T) - \xi_i, W_i[x(T) - \xi_i]\rangle \\ &\quad + \int_0^T \langle z_i(t) - C_i(t)x(t), Q_i(t)[z_i(t) - C_i(t)x(t)]\rangle\, dt \\ &\quad + \int_0^T \langle u_i(t), R_i(t)u_i(t)\rangle\, dt, \qquad i = 1, 2, \cdots, N, \end{aligned} \tag{2.4}$$

where $\xi_i \in R^m$ are fixed points, $z_i(t)$ are given continuous functions, $C_i(t)$ are continuous matrices, $Q_i(t)$, $R_i(t)$ are continuous and positive semidefinite, and W_i are positive semidefinite.

* Received by the editors March 1, 1973.

† Department of Industrial Engineering and Management Science, Northwestern University, Evanston, Illinois 60201.

Denote by X_T the space of trajectories $x(t)$ satisfying (2.1), (2.2) when $u_i(t)$, $i = 1, 2, \cdots, N$, vary over all control functions. This is a compact convex subset of $C^m[0, T] = \times_{i=1}^{m} C[0, T]$ (cf. [1, Thm. 2.4.1, p. 43]).

Clearly the controls $u_i(t)$ belong to $L^{p,q_i}(0, T)$ for any $1 \leqq p \leqq \infty$. For each i, denote by $\mathbf{U}_i$ the set of all controls $u_i(t)$ for the ith player. Unless otherwise stated, $\mathbf{U}_i$ will be considered as a subset of $L^{2,q_i}(0, T)$ with the weak topology. It follows from (2.3) that $\mathbf{U}_i$ is weakly closed and convex (see [1, Lemma 2.4.1, p. 43]). Since $L^{2,q_i}(0, T)$, with the norm topology, is a Hilbert space and since $\mathbf{U}_i$ is bounded in $L^{2,q_i}(0, T)$ with the norm topology, it follows that $\mathbf{U}_i$ is weakly sequentially compact. Denote by $\mathbf{U}$ the set $\mathbf{U}_1 \times \cdots \times \mathbf{U}_N$ with the product of weak topologies, and by $\hat{\mathbf{U}}_i$ the set $\mathbf{U}_1 \times \cdots \times \mathbf{U}_{i-1} \times \mathbf{U}_{i+1} \times \cdots \times \mathbf{U}_N$ with the product of weak topologies. It is clear that $\mathbf{U}$ and $\hat{\mathbf{U}}_i$ are also convex and weakly sequentially compact.

We now state the main result of the paper.

THEOREM 2.1. *The N-person differential game associated with* (2.1)–(2.4) *has an open loop Nash equilibrium point.*

3. Auxiliary lemmas. We shall need the following lemmas.

LEMMA 3.1. *Let* $\hat{u}'_1 = (u'_1, \cdots, u'_{i-1}, u'_{i+1}, \cdots, u'_N)$ *be a fixed element of* $\hat{\mathbf{U}}_i$. *Then the functional*

$$J_i(u_i, \hat{u}'_i) = J_1(u'_1, \cdots, u'_{i-1}, u_i, u'_{i+1}, \cdots, u'_N) \tag{3.1}$$

is convex and lower-semicontinuous for $u_i \in \mathbf{U}_i$.

Proof. The convexity follows easily from (2.1)–(2.4). To prove the lower-semicontinuity, note that the first two terms of (2.4) are continuous in u_i. Thus it remains to show that the last term in (2.4) is lower-semicontinuous in u_i.

For each real number r, consider the set

$$\mathbf{U}_i(r) = \left\{ u_i \in \mathbf{U}_i : \int_0^T \langle u_i(t), R_i(t) u_i(t) \rangle \, dt \leqq r \right\}.$$

Since $R_i(t)$ is positive semidefinite, $\mathbf{U}_i(r)$ is convex. It is easily seen that $\mathbf{U}_i(r)$ is strongly closed. But as is well known, a strongly closed, convex subset of a Hilbert space is weakly closed. Hence $\mathbf{U}_i(r)$ is weakly closed for any real number r. But this is equivalent to the property of lower-semicontinuity in u_i; see [2, p. 89].

LEMMA 3.2. *Consider the family of functionals* $J_i(u_i, \hat{u}_i)$ *on* $\hat{\mathbf{U}}_i$, *where* u_i *is the parameter of the family and ranges over* $\mathbf{U}_i$. *This family is equicontinuous on* $\hat{\mathbf{U}}_i$.

Note that $\times_{j \neq i} L^{2,q_j}(0, T)$, with the product of weak topologies, is a locally convex linear topological space. Hence its topology is completely determined by a base of neighborhoods of the origin. The equicontinuity of Lemma 3.2 means that for every $\varepsilon > 0$ there is a neighborhood of the origin, $\hat{H}$, such that if $\hat{u}'_i \in \hat{\mathbf{U}}_i$ and if $\hat{u}_i \in (\hat{H} + \hat{u}'_i) \cap \hat{\mathbf{U}}_i$, then

$$|J_i(u_i, \hat{u}_i) - J_i(u_i, \hat{u}'_i)| \leqq \varepsilon \quad \text{for all } u_i \in \mathbf{U}_i. \tag{3.2}$$

Proof of Lemma 3.2. Let $\varepsilon > 0$ be given, and let $u_i \in \mathbf{U}_i$ be given. Let $\hat{u}'_i = (u'_1, \cdots, u'_{i-1}, u'_{i+1}, \cdots, u'_N)$ be any element of $\hat{\mathbf{U}}_i$, and choose $\hat{H}$ as follows:

$$\begin{aligned} \hat{H} = \Bigg\{ \hat{u}_i \in \times_{j \neq i} L^{2,q_j}(0, T); \ & \left| \int_{t_k}^{t_{k+1}} \Phi^{-1}(t) B_j(t) u_j(t) \, dt \right| \\ & \leqq \frac{\delta(\varepsilon)}{3m(N-1)\|\Phi\|}, j \neq i, 0 = t_0 < t_1 < \cdots < t_m = T \Bigg\}, \end{aligned} \tag{3.3}$$

where $\Phi(t)$ is a fundamental solution of (2.1), and $\|\Phi\| = \max_{0 \leqq t \leqq T} |\Phi(t)|$, where $t_{k+1} - t_k = T/m$, and m is chosen so that if $|t - t'| \leqq T/m$, then for any $x(t)$, $|x(t) - x(t')| \leqq \delta(\varepsilon)/3$. That this can be done follows from a standard theorem on the equicontinuity of the set of solutions of (2.1), (2.2). Finally, $\delta(\varepsilon)$ is chosen so that if $\max_{0 \leqq t \leqq T} |x(t) - \bar{x}(t)| \leqq \delta(\varepsilon)$, then

$$|\langle x(T) - \xi_i, W_i[x(T) - \xi_i]\rangle - \langle \bar{x}(T) - \xi_i, W_i[\bar{x}(T) - \xi_i]\rangle| \leqq \varepsilon/2 \tag{3.4}$$

and

$$\begin{aligned} \max_{0 \leqq t \leqq T} |\langle z_i(t) - C_i(t)x(t), Q_i(t)[z_i(t) - C_i(t)x(t)]\rangle \\ - \langle z_i(t) - C_i(t)\bar{x}(t), Q_i(t)[z_i(t) - C_i(t)\bar{x}(t)]\rangle| \leqq \varepsilon/(2T), \end{aligned} \tag{3.5}$$

where $x(t)$ is the trajectory corresponding to $(u_i, \hat{u}_i)$ with $\hat{u}_i \in \hat{H}$ and $\bar{x}(t)$ is the trajectory corresponding to $(u_i, \hat{0})$.

Now note that if $\hat{u}'_i \in \hat{\mathbf{U}}_i$, and if $\hat{u}_i \in (\hat{H} + \hat{u}'_i) \cap \hat{\mathbf{U}}_i$, then

$$|x(t) - x'(t)| \leqq |x(t) - x(t_k)| + |x(t_k) - x'(t_k)| + |x'(t_k) - x'(t)|, \tag{3.6}$$

where $x(t)$ is the trajectory corresponding to $(u_i, \hat{u}_i)$, $x'(t)$ is the trajectory corresponding to $(u_i, \hat{u}'_i)$ and k is the largest index such that $t_k < t$. Thus

$$|x(t) - x(t_k)| \leqq \delta(\varepsilon)/3, \qquad |x'(t_k) - x'(t)| \leqq \delta(\varepsilon)/3, \tag{3.7}$$

since $|t - t_k| < T/m$. But we also have

$$\begin{aligned} |x(t_k) - x'(t_k)| &= |\Phi(t_k)| \cdot \left| \int_0^{t_k} \sum_{j \neq i} \Phi^{-1}(t) B_j(t) u''_j(t)\, dt \right| \\ &\leqq \|\Phi\| \sum_{l=0}^{k-1} \sum_{j \neq i} \left| \int_{t_l}^{t_{l+1}} \Phi^{-1}(t) B_j(t) u''_j(t)\, dt \right| \\ &\leqq \|\Phi\| \cdot k \cdot \frac{\delta(\varepsilon)}{3m(N-1)\|\Phi\|}(N-1) \\ &\leqq \delta(\varepsilon)/3, \quad \text{since } \hat{u}''_i = \hat{u}_i - \hat{u}'_i \in \hat{H}. \end{aligned} \tag{3.8}$$

Next note that

$$\begin{aligned} &|J_i(u_i; \hat{u}_i) - J_i(u_i, \hat{u}'_i)| \\ &\quad \leqq |\langle x(T) - \xi_i, W_i[x(T) - \xi_i]\rangle - \langle x'(T) - \xi_i, W_i[x'(T) - \xi_i]\rangle| \\ &\quad + \left| \int_0^T \langle z_i(t) - C_i(t)x(t), Q_i(t)[z_i(t) - C_i(t)x(t)]\rangle\, dt \right. \\ &\quad \left. - \int_0^T \langle z_i(t) - C_i(t)x'(t), Q_i(t)[z_i(t) - C_i(t)x'(t)]\rangle\, dt \right|. \end{aligned} \tag{3.9}$$

The result follows from (3.3)–(3.9).

LEMMA 3.3. *For each $\hat{u}_i \in \hat{\mathbf{U}}_i$ there is a $\bar{u}_i \in \mathbf{U}_i$ such that* (i)

$$J_i(\bar{u}_i, \hat{u}_i) = \inf_{u_i} J_i(u_i, \hat{u}_i), \tag{3.10}$$

and (ii)

$$\inf_{u_i} J_i(u_i, \hat{u}_i) \tag{3.11}$$

is continuous on $\hat{\mathbf{U}}_i$.

Proof. (i) follows by noting that for fixed $\hat{u}_i$, $J_i(u_i, \hat{u}_i)$ is weakly lower-semicontinuous on $\mathbf{U}_i$ and that $\mathbf{U}_i$ is weakly sequentially compact.

To prove (ii), let $\varepsilon > 0$ be given. Choose $\hat{H}$ as in Lemma 3.2. Then for each $\hat{u}'_i \in \hat{\mathbf{U}}_i$ and $\hat{u}_i \in (\hat{H} + \hat{u}'_i) \cap \hat{\mathbf{U}}_i$,

$$J_i(u_i, \hat{u}_i) \leqq J_i(u_i, \hat{u}'_i) + \varepsilon \quad \text{for all } u_i \in \mathbf{U}_i$$

and

$$J_i(u_i, \hat{u}'_i) \leqq J_i(u_i, \hat{u}_i) + \varepsilon \quad \text{for all } u_i \in \mathbf{U}_i.$$

Taking the infinum of both sides of each of the above inequalities yields the result.

We shall require the following well-known result on partitions of unity; see [3, p. 171].

LEMMA 3.4. *Let X be a normal topological space and let $\mathcal{O}$ be a locally finite open covering of X. Then there exists a family of continuous real-valued functions f_G, $G \in \mathcal{O}$, such that $f_G = 0$ on $X - G$, $0 \leqq f_G \leqq 1$ elsewhere, and*

$$\sum_{G \in \mathcal{O}} f_G(x) = 1 \quad \textit{for all } x \in X.$$

Since $\hat{\mathbf{U}}_i$ is a compact subset of $\bigtimes_{j \neq i} L^{2,q_j}(0, T)$ with the product of weak topologies, and since $\hat{\mathbf{U}}_i$ is clearly Hausdorff, $\hat{\mathbf{U}}_i$ is a normal topological space. Furthermore, every open covering of $\hat{\mathbf{U}}_i$ has a finite subcovering (hence is locally finite). Thus by Lemma 3.4 there is a partition of unity for any open covering of $\hat{\mathbf{U}}_i$.

For each $\hat{u}_i \in \hat{\mathbf{U}}_i$ and each $\varepsilon > 0$, consider the sets

$$\mathbf{U}_i(\hat{u}_i, \varepsilon) = \{u_i \in \mathbf{U}_i; J_i(u_i, \hat{u}_i) \leqq \inf_{u'_i} J_i(u'_i, \hat{u}_i) + \varepsilon\}. \tag{3.12}$$

LEMMA 3.5. *For each $\varepsilon > 0$ there is a continuous function $\varphi_i^\varepsilon : \hat{\mathbf{U}}_i \to \mathbf{U}_i$ with the property that $\varphi_i^\varepsilon(\hat{u}_i) \in \mathbf{U}_i(\hat{u}_i, \varepsilon)$ for each $\hat{u}_i \in \hat{\mathbf{U}}_i$.*

Proof. Let $\varepsilon > 0$ be given. By Lemma 3.2 for each $\hat{u}'_i$ there is a neighborhood, $H(\hat{u}'_i)$, of $\hat{u}'_i$ such that

$$|J_i(u_i, \hat{u}_i) - J_i(u_i, \hat{u}'_i)| \leqq \varepsilon/4, \tag{3.13}$$

for all u_i and for all $\hat{u}_i \in H(\hat{u}'_i)$. Then

$$\{H(\hat{u}_i); \hat{u}_i \in \hat{\mathbf{U}}_i\}$$

is an open covering of $\hat{\mathbf{U}}_i$. By the paragraph following Lemma 3.4 there is a finite subcovering

$$\{H(\hat{u}_{i,j})\}_{j=1}^n.$$

For each j, choose a fixed $u_{ij} \in \mathbf{U}_i(\hat{u}_{i,j}, \varepsilon/2)$. Set

$$\psi_j(\hat{u}_i) = u_{i,j} \quad \text{for all } \hat{u}_i \in \hat{\mathbf{U}}_i. \tag{3.14}$$

We next wish to show that if $\hat{u}_i \in H(\hat{u}_{i,j})$, then $\psi_j(\hat{u}_i) \equiv u_{i,j} \in \mathbf{U}_i(\hat{u}_i, \varepsilon)$. This means that if $\hat{u}_i \in \hat{H}(\hat{u}_{i,j})$, then

$$(3.15) \qquad J_i(u_{i,j}, \hat{u}_i) \leqq \inf_{u_i} J_i(u_i, \hat{u}_i) + \varepsilon .$$

But since $\hat{u}_i \in H(\hat{u}_{i,j})$, we have by applying (3.13) with $\hat{u}_i' = \hat{u}_{i,j}$, and with $u_i = u_{i,j}$, that

$$(3.16) \qquad J_i(u_{i,j}, \hat{u}_i) \leqq J_i(u_{i,j}, \hat{u}_{i,j}) + \varepsilon/4 .$$

Next, since $u_{i,j} \in \mathbf{U}_i(\hat{u}_{i,j}, \varepsilon/2)$,

$$(3.17) \qquad J_i(u_{i,j}, \hat{u}_{i,j}) \leqq \inf_{u_i} J(u_i, \hat{u}_{i,j}) + \varepsilon/2 .$$

From (3.13) with $\hat{u}_i' = \hat{u}_{i,j}$, we get

$$(3.18) \qquad \inf_{u_i} J_i(u_i, \hat{u}_{i,j}) \leqq \inf_{u_i} J_i(u_i, \hat{u}_i) + \varepsilon/4 .$$

Combining (3.16)–(3.18) yields (3.15).

The maps $\psi_j : \hat{\mathbf{U}}_i \to U_i$ are constant, and hence continuous. Let $\{\theta_j\}_{j=1}^n$ be a partition of unity subordinate to the covering $\{H(\hat{u}_{i,j})\}_{j=1}^n$, and set

$$(3.19) \qquad \varphi_i^\varepsilon = \sum_{j=1}^n \theta_j \psi_j .$$

Clearly $\varphi_i^\varepsilon : \hat{\mathbf{U}}_i \to L^{2,q_i}(0, T)$ is continuous, and since $\mathbf{U}_i$ is convex, $\varphi_i^\varepsilon : \hat{\mathbf{U}}_i \to \mathbf{U}_i$. It remains only to show that $\varphi_i^\varepsilon(\hat{u}_i) \in \mathbf{U}_i(\hat{u}_i, \varepsilon)$ for each $\hat{u}_i$. But

$$(3.20) \qquad \begin{aligned} J_i(\varphi_i^\varepsilon(\hat{u}_i), \hat{u}_i) &= J_i\left(\sum_{j=1}^n \theta_j(\hat{u}_i) \cdot \psi_j(\hat{u}_i), \hat{u}_i \right) \\ &\leqq \sum_{j=1}^n \theta_j(\hat{u}_i) \cdot J_i(\psi_j(\hat{u}_i), \hat{u}_i) \\ &\leqq \sum_{j=1}^n \theta_j(\hat{u}_i) \cdot [\inf_{u_i} J_i(u_i, \hat{u}_i) + \varepsilon] \\ &= \inf_{u_i} J_i(u_i, \hat{u}_i) + \varepsilon ; \end{aligned}$$

the first inequality follows by the convexity of $J_i(u_i, \hat{u}_i)$ in u_i, the second inequality by definition of ψ_j and θ_j, and the last by a property of the θ_j.

Thus $\varphi_i^\varepsilon(\hat{u}_i) \in \mathbf{U}_i(\hat{u}_i, \varepsilon)$ for each $\hat{u}_i$, and this completes the proof of Lemma 3.5.

Finally, we require a fixed-point theorem of Tychonov; see [4, p. 456].

LEMMA 3.6. *Let A be a compact, convex subset of a locally convex linear topological space. Then every continuous map from A into itself has a fixed point.*

Since $\mathbf{U}$ is a compact, convex subset of $\bigtimes_{j=1}^N L^{2,q_j}(0, T)$, we can apply the Tychonov theorem with $A = \mathbf{U}$.

4. Proof of Theorem 2.1. Note first that each of the continuous maps $\varphi_i^\varepsilon : \hat{\mathbf{U}}_i \to \mathbf{U}_i$ may be considered as a continuous map $\tilde{\varphi}_i^\varepsilon : \mathbf{U} \to \mathbf{U}_i$ by setting $\tilde{\varphi}_i^\varepsilon(u_i, \hat{u}_i) = \varphi_i^\varepsilon(\hat{u}_i)$.

For each $\varepsilon > 0$, let $\tilde{\varphi}^\varepsilon : \mathbf{U} \to \mathbf{U}$ be given by

$$\tilde{\varphi}^\varepsilon(u) = (\tilde{\varphi}_1^\varepsilon(u), \cdots, \tilde{\varphi}_N^\varepsilon(u)), \tag{4.1}$$

where $u = (u_1, \cdots, u_N)$.

Since $\tilde{\varphi}^\varepsilon$ is continuous, the Tychonov theorem applies. We conclude that $\tilde{\varphi}^\varepsilon$ has a fixed point, $u^\varepsilon = (u_1^\varepsilon, \cdots, u_N^\varepsilon)$. By definition of the maps $\tilde{\varphi}_i^\varepsilon$ we have

$$\begin{aligned} J_i(u^\varepsilon) &\leqq \inf_{u_i} J_i(u_i, \hat{u}_i^\varepsilon) + \varepsilon \\ &\leqq J_i(u_i, \hat{u}_i^\varepsilon) + \varepsilon \quad \text{for all } u_i \in \mathbf{U}_i, \qquad i = 1, 2, \cdots, N. \end{aligned} \tag{4.2}$$

For $n = 1, 2, \cdots$, set $\varepsilon = 1/n$ and consider the sequence $\{u^{1/n}\} = \{(u_1^{1/n}, \cdots, u_N^{1/n})\}$. Since $\mathbf{U}$ is weakly sequentially compact, $\{u^{1/n}\}$ has a convergent subsequence, which we again denote by $\{u^{1/n}\}$. Let $\bar{u} = (\bar{u}_1, \cdots, \bar{u}_N)$ be the weak limit of $\{u^{1/n}\}$. As in the proof of Lemma 3.1, it can be seen that $J_i(u)$ is lower-semicontinuous on $\mathbf{U}$. Hence by (4.2), we have

$$\begin{aligned} J_i(\bar{u}) &\leqq \underline{\lim}_{n \to \infty} J_i(u^{1/n}) \\ &\leqq \underline{\lim}_{n \to \infty} [J_i(u^i, \hat{u}^{1/n}) + 1/n] \\ &= \underline{\lim}_{n \to \infty} J_i(u_i, \hat{u}_i^{1/n}) \\ &= J_i(u_i, \hat{\bar{u}}_i) \end{aligned} \tag{4.3}$$

for all $u_i \in \mathbf{U}_i$, $i = 1, 2, \cdots, N$, where $\hat{\bar{u}} = (\bar{u}_1, \cdots, \bar{u}_{i-1}, \bar{u}_{i+1}, \cdots, \bar{u}_N)$ and where the last step follows by Lemma 3.2.

But (4.3) is just the inequality required to show that $\bar{u}$ is an open loop Nash equilibrium point, and the proof is complete.

5. Remarks. For differential games associated with (2.1)–(2.3) and costs for which Lemmas 3.1–3.3 remain valid, Theorem 2.1 remains valid. One such example is a class of games considered by Varaiya [6], where the costs have the form

$$\begin{aligned} J_i(u_1, \cdots, u_N) = g_i(x(T_0)) &+ \int_0^T h_i(t, x(t))\, dt \\ &+ \int_0^T f_i(t, u_i(t))\, dt, \qquad i = 1, 2, \cdots, N. \end{aligned} \tag{5.1}$$

We must, of course, replace his assumptions on g_i, h_i, f_i by the following:

(5.2)

(a) $g_i(x)$ is convex in x,

(b) $h_i(t, x)$ is continuous in (t, x) and convex in x,

(c) $f_i(t, u_i)$ is continuous in t and the Hessian $(\partial^2 f_i / \partial u_i^2)$ is continuous and positive semidefinite.

Also, it should be noted that the above methods apply equally well to the games with integral bounds on the controls.

REFERENCES

[1] A. Friedman, *Differential Games*, Wiley-Interscience, New York, 1971.

[2] E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag, New York, 1965.

[3] J. L. Kelley, *General Topology*, Van Nostrand, New York, 1955.

[4] N. Dunford and J. Schwartz, *Linear Operators*, vol. I, Wiley-Interscience, New York, 1958.

[5] D. L. Lukes and D. L. Russell, *A global theory for linear quadratic differential games*, J. Math. Anal. Appl., 33 (1971), pp. 96–123.

[6] P. Varaiya, *N-person nonzero sum differential games with linear dynamics*, this Journal, 8 (1970), pp. 441–449.

SIAM J. CONTROL
Vol. 12, No. 3, August 1974

DIFFERENTIAL GAMES WITH RESTRICTED PHASE COORDINATES*

RICHARD C. SCALZO†

Abstract. In 1970, Avner Friedman proved the existence of value and saddle points for games with restricted phase coordinates. He assumed that $x_0 \in \operatorname{int} X$, where X is the phase set and x_0 is the initial condition of the game. He also assumed, in case there is a penalty boundary for each player, that the duration of the game is so small that no trajectories can intersect both penalty boundaries.

The purpose of this paper is to remove the above restrictions under some suitably mild assumptions. In the first section, existence of value is proved in case $x_0 \in \partial X$. The second section is devoted to showing that the value is continuous for $x \in X$ and that the duration of the game need not be restricted so as not to allow intersection of trajectories with both penalty boundaries.

1. Introduction.

The approach to differential game theory used in this paper is that of A. Friedman, and the notation is taken from [1].

Friedman has given conditions under which a differential game with restricted phase coordinates has value if $x_0 \in \operatorname{int} X$ and if $T_0 - t_0 \leqq T(\bar{\varepsilon})$; see [1, pp. 207–230]. The purpose of this paper is to remove, under some mild additional assumptions, the restriction on the duration of such games. To do this we give, in §2, conditions which extend the existence of value to the case where $x_0 \in \partial X$. In this section we use the restriction $T_0 - t_0 \leqq T(\bar{\varepsilon})$. Section 3 is devoted to showing that under the conditions used in § 2 a game has value when the duration, $T_0 - t_0$, is arbitrary.

2. Existence of value when $x_0 \in \partial X$.

Consider a differential game, G, associated with a system of m differential equations

$$\frac{dx}{dt} = f(t, x, y, z), \qquad t_0 \leqq t \leqq T_0, \tag{2.1}$$

$$x(t_0) = x_0 \tag{2.2}$$

(with controls $y(t), z(t)$), where $y(t)$, $z(t)$ are measurable functions with values in Y, Z respectively, a.e. $[t_0, T_0]$, and $Y \subseteq R^p$, $Z \subseteq R^q$ are compact sets.

Let

$$\begin{aligned} X &= \{x \in R^m | x_{m-1} \geqq 0, x_m \geqq 0\}, \\ \partial X_y &= \{x \in R^m | x_{m-1} > \mu, x_m = 0\}, \\ \partial X_z &= \{x \in R^m | x_m > \mu, x_{m-1} = 0\}, \\ \partial X_0 &= \{x \in R^m | x_m = 0, 0 \leqq x_{m-1} \leqq \mu\} \cup \{x \in R^m | x_{m-1} = 0, \\ &\qquad\qquad\qquad\qquad 0 \leqq x_{m-1} \leqq \mu\}, \end{aligned} \tag{2.3}$$

* Received by the editors December 22, 1972.

† Department of Industrial Engineering and Management Science, Northwestern University, Evanston, Illinois 60201.

where $\mu > 0$ is constant and ∂X_0 is such that no solution of (2.1), (2.2) with $x_0 \in X$ can leave X by crossing ∂X_0. Let

$$(2.4) \qquad P(y,z) = \begin{cases} g(x(T_0)) + \int_{t_0}^{T_0} h(t, x(t), y(t), z(t))\, dt & \\ \qquad\qquad\qquad \text{if } x(t) \in X, t_0 \leqq t \leqq T_0, & \\ -M \qquad \text{if } x(t) \text{ leaves } X \text{ for the first time by crossing } \partial X_y, & \\ +M \qquad \text{if } x(t) \text{ leaves } X \text{ for the first time by crossing } \partial X_z, & \end{cases}$$

where

$$M > M_0 = \sup_{y(t)} \sup_{z(t)} \left| g(x(T_0)) + \int_{t_0}^{T_0} h(t, x(t), y(t), z(t))\, dt \right|$$

and the suprema range over all controls $y(t), z(t)$.

We shall need the following assumptions:

(A_1) $f(t, x, y, z)$ is continuous on $[t_0, T_0] \times R^m \times Y \times Z$.

(A_2) There is a nonnegative function $k(t)$ on $[t_0, T_0]$ with $\int_{t_0}^{T_0} k(t)\, dt < \infty$, such that

$$|f(t,x,y,z)| \leqq k(t)(1 + |x|)$$

on $[t_0, T_0] \times R^m \times Y \times Z$.

(A_3) For each $R > 0$ there is a nonnegative function $k_R(t)$, such that $\int_{t_0}^{T_0} k_R(t)\, dt < \infty$, with

$$|f(t,x,y,z) - f(t,\bar{x},y,z)| \leqq k_R(t)|x - \bar{x}|$$

for all $t \in [t_0, T_0]$, $y \in Y$, $z \in Z$, $|x| \leqq R$, $|\bar{x}| \leqq R$.

Note. (A_1)–(A_3) guarantee the existence and uniqueness of solutions $x(t)$ to (2.1), (2.2) for every pair of controls $y(t), z(t)$, and further that $|x(t)| \leqq C$, where C is a constant depending only on the system (2.1), (2.2), consequently $M_0 < \infty$.

(A_4) $f(t,x,y,z) = f^1(t,x,y) + f^2(t,x,z)$ on $[t_0, T_0] \times R^m \times Y \times Z$.

(B_1) $h(t,x,y,z)$ is continuous on $[t_0, T_0] \times R^m \times Y \times Z$.

(B_2') $g(x)$ is continuous on R^m.

(B_3) $h(t,x,y,z) = h^1(t,x,y) + h^2(t,x,z)$ on $[t_0, T_0] \times R^m \times Y \times Z$.

($\bar{P}_2'$)(a) $Y = Y_1 \times Y_2, Y_1 \subseteq R^{p_1}, Y_2 \subseteq R^{p_2}, p_1 + p_2 = p$;

$Z = Z_1 \times Z_2, Z_1 \subseteq R^{q_1}, Z_2 \subseteq R^{q_2}, q_1 + q_2 = q$,

where some of the sets Y_1, Y_2, Z_1, Z_2 may be empty.

(b) Write $y = (y', y'')$, $z = (z', z'')$, where $y \in Y$, $y' \in Y_1$, $y'' \in Y_2$, $z \in Z$, $z' \in Z_1$, $z'' \in Z_2$. Also write $x' = (x_1, \cdots, x_{m-2})$. Then for any $t \in [t_0, T_0]$, $x' \in R^{m-2}$, $y'' \in Y_2$, $z'' \in Z_2$,

$$(2.5) \qquad \max_{y' \in Y_1} \min_{z \in Z} f_m(t, x', x_{m-1}, 0, y', y'', z) > 0 \quad \text{if } x_{m-1} \geqq \mu,$$

$$(2.6) \qquad \max_{z' \in Z_1} \min_{y \in Y} f_{m-1}(t, x', 0, x_m, y, z', z'') > 0 \quad \text{if } x_m \geqq \mu.$$

(c) For any $\varepsilon > 0$, let

$$\partial X_y^\varepsilon = \{x \in R^m | 0 \leqq x_m \leqq \varepsilon, x_{m-1} \geqq \mu\},$$

$$\partial X_z^\varepsilon = \{x \in R^m | 0 \leqq x_{m-1} \leqq \varepsilon, x_m \geqq \mu\}.$$

There is an $\varepsilon^* > 0$ such that the following relations hold for all $t \in [t_0, T_0]$, $(y', y'') \in Y, (z', z'') \in Z, \bar{y}' \in Y_1, \bar{z}' \in Z_1$:

$$(2.7) \quad f_i(t, x, y', y'', z) = f_i(t, x, \bar{y}', y'', z) \quad \text{for all } x \in \partial X_y^{\varepsilon^*}, \quad i = 1, 2, \cdots, m-1,$$

$$(2.8) \quad f_i(t, x, y, z', z'') = f_i(t, x, y, \bar{z}', z'') \quad \text{for all } x \in \partial X_z^{\varepsilon^*},$$

$$i = 1, 2, \cdots, m-3, m-2, m.$$

(d) There is an $\bar{\varepsilon}^* > 0$ such that the following relations hold for all $t \in [t_0, T_0]$, $(y', y'') \in Y, (z', z'') \in Z$:

$$(2.9) \qquad \max_{y' \in Y_1} f_m(t, x, y', y'', z) = \max_{y' \in Y_1} f_m(t, \bar{x}, y', y'', z)$$

if $x \in \partial X_y^{\bar{\varepsilon}^*}$, $\bar{x} \in \partial X_y^{\bar{\varepsilon}^*}$, where the maximum in (2.9) is attained independently of x, y'';

$$(2.10) \qquad \max_{z' \in Z_1} f_{m-1}(t, x, y, z', z'') = \max_{z' \in Z_1} f_m(t, \bar{x}, y, z', z'')$$

if $x \in \partial X_z^{\bar{\varepsilon}^*}$, $\bar{x} \in \partial X_z^{\bar{\varepsilon}^*}$, where the maximum in (2.10) is attained independently of x, z''.

Remark 1. $(\bar{\mathrm{P}}_2')$(d) is the only additional assumption. See [1, Thm. 6.3.2, p. 226].

Remark 2. It follows easily from (2.9), (A_4) that if $x_0 \in \partial X_y$ and $[t_0, t_1]$ is sufficiently small, then there is a control $y'_*(t)$ on $[t_0, t_1]$ such that if $y(t) = (y'(t), y''(t))$ is any control for y and $z(t)$ is any control for z, then

$$(2.11) \qquad x_m^*(t) \geqq x_m(t), \qquad t_0 \leqq t \leqq t_1,$$

where $x_m^*(t)$ is the trajectory corresponding to $((y'_*(t), y''(t)), z(t))$ and $x(t)$ is the trajectory corresponding to $(y(t), z(t))$.

$(\bar{\mathrm{Q}}_2)$(a) There is an $\varepsilon_0 > 0$ such that $(\partial f_i / \partial x_j)(t, x, y, z)$ exists and is continuous for $1 \leqq i, j \leqq m$ if either $x_m \geqq 0, 0 \leqq x_{m-1} \leqq \varepsilon_0$ or $x_{m-1} \geqq 0, 0 \leqq x_m \leqq \varepsilon_0$.

(b) For any control functions $y(t)$, $z(t)$ for which the corresponding trajectory $x(t)$ satisfies $0 \leqq x_m(t) \leqq \varepsilon_0$, $x_{m-1}(t) \geqq \mu$ or $0 \leqq x_{m-1}(t) \leqq \varepsilon_0$, $x_m(t) \geqq \mu$ in some subinterval $(\bar{t}, \bar{\bar{t}})$ of (t_0, T_0), set

$$(2.12) \qquad \begin{aligned} b_{ij}(t) &= \frac{\partial f_i}{\partial x_j}(t, x'(t), x_{m-1}(t), 0, y(t), z(t)), \\ \hat{b}_{ij}(t) &= \frac{\partial f_i}{\partial x_j}(t, x'(t), 0, x_m(t), y(t), z(t)), \end{aligned}$$

for $1 \leqq i, j \leqq m$, $\bar{t} \leqq t \leqq \bar{\bar{t}}$, where $x(t) = (x'(t), x_{m-1}(t), x_m(t))$.

Consider the linear differential systems

$$(2.13) \qquad \frac{dZ_i}{dt} = \sum_{j=1}^m b_{ij}(t) Z_j, \qquad \frac{d\hat{Z}_i}{dt} = \sum_{j=1}^m \hat{b}_{ij}(t) \hat{Z}_j,$$

for $1 \leqq i \leqq m$, $\bar{t} \leqq t \leqq \bar{\bar{t}}$, with initial conditions

$$(2.14) \qquad Z_m(\bar{t}) = 1, \qquad Z_i(\bar{t}) = 0, \qquad i = 1, 2, \cdots, m-1,$$

$$(2.15) \qquad \hat{Z}_{m-1}(\bar{t}) = 1, \qquad \hat{Z}_i(\bar{t}) = 0, \qquad i = 1, 2, \cdots, m-3, m-2, m.$$

Then there is $\theta_0 > 0$, independent of $y(t), z(t), \bar{t}, \bar{\bar{t}}$ such that

$$(2.16) \qquad Z_m(t) \geqq \theta_0, \qquad \hat{Z}_{m-1}(t) \geqq \theta_0 \quad \text{for } \bar{t} \leqq t \leqq \bar{\bar{t}}.$$

(T) Let $\bar{\varepsilon}$ be any positive number. Denote by $T(\bar{\varepsilon})$ a number such that no trajectory $x(t)$ of (2.1), (2.2) can enter into both $\partial X_y^{\bar{\varepsilon}}$ and $\partial X_z^{\bar{\varepsilon}}$. Then $T_0 - t_0 \leqq T(\bar{\varepsilon})$.

For each $\Gamma^\delta = (\Gamma^{\delta,1}, \cdots, \Gamma^{\delta,n})$ write $\Gamma^{\delta,j} = (\Gamma_1^{\delta,j}, \Gamma_2^{\delta,j})$, where $\Gamma^{\delta,j}(z_1, y_1', y_1'', \cdots, z_j) = (y_j', y_j'')$, $\Gamma_1^{\delta,j}(z_1, y_1', y_1'', \cdots, z_j) = y_j'$, $\Gamma_2^{\delta,j}(z_1, y_1', y_1'', \cdots, z_j) = y_j''$. Treat Δ^δ similarly.

THEOREM 2.1. *Consider a game, G, associated with* (2.1)–(2.4). *Assume that* (A_1)–(A_4), (B_1), (B_2'), (B_3), $(\bar{\mathrm{P}}_2')$, $(\bar{\mathrm{Q}}_2)$ *and* (T) *hold. If $x_0 \in X$, then G has value.*

Proof. If $x_0 \in \operatorname{int} X$, then Theorem 2.1 follows from Theorem 6.3.2 [1, p. 226]. Thus it suffices to assume that $x_0 \in \partial X_y$. The proof is attained by a modification of the proof of Theorem 6.3.2.

From $(\bar{P}_2')$ it follows that

$$(2.17) \qquad -M < V_\delta \leqq V^\delta < M \quad \text{if } \delta \leqq \bar{\delta} \text{ is sufficiently small.}$$

For any $\gamma > 0$ there is an upper δ-strategy $\bar{\Gamma}^\delta$ for y such that

$$(2.18) \qquad V^\delta \leqq P(\bar{\Gamma}^\delta(z), z) + \gamma \quad \text{for all controls } z.$$

LEMMA 2.1. *There exist $\varepsilon > 0$, $\eta_0 > 0$, and $\delta_0(\eta, \varepsilon) > 0$ sufficiently small such that for every $0 < \eta \leqq \eta_0$, $\delta \leqq \delta_0$ there is an upper δ-strategy $\tilde{\Gamma}^\delta$ for y such that*

$$(2.19) \qquad V^\delta \leqq P(\tilde{\Gamma}^\delta(z), z) + \gamma + \alpha(\eta) \quad \textit{for all controls } z,$$

where $\alpha(\eta) \to 0$ if $\eta \to 0$,

$$(2.20) \qquad \tilde{\Gamma}_1^{\delta,j}(z_1, y_1', y_1'', \cdots, z_j)(t) = y_{*,j}'(t),$$

$1 \leqq j \leqq \sigma_0, \sigma_0 = [\eta\varepsilon/\delta]$, *where $y_{*,j}'(t)$ is the restriction to I_j of a fixed control $y_*(t)$ chosen as in Remark* 2.

Proof. Let $\varepsilon < \varepsilon_0/2$, where ε_0 is as in $(\bar{\mathrm{Q}}_2)$. Denote by $\bar{x}(t)$ the trajectory corresponding to $(\bar{\Gamma}^\delta(z), z)$ and by $\tilde{x}(t)$ the trajectory corresponding to $(\tilde{\Gamma}^\delta(z), z)$, where $z = z(t)$ is any control for z.

Let $z(t)$ be any control for z. For $1 \leqq j \leqq \sigma_0$, set

$$(2.21) \qquad \tilde{\Gamma}^{\delta,j}(z_1, \tilde{y}_1, \cdots, z_j)(t) = \tilde{y}_j(t) = (y_{*,j}'(t), \bar{y}_j''(t)),$$

where $y_{*,j}'(t)$ is the restriction to I_j of a fixed control chosen so that Remark 2 holds on I_j and $\bar{y}_j''(t)$ coincides with the control function obtained from $\bar{\Gamma}_2^{\delta,j}, z, \bar{y}$. Thus (2.20) holds. It remains to prove that (2.19) holds.

By $(\bar{\mathrm{P}}_2')$, there is a $\gamma_0 > 0$ such that

$$(2.22) \qquad 0 < \gamma_0 \leqq \max_{y' \in Y_1} f_m(t, \bar{x}(t), y', \bar{y}''(t), z(t)) \equiv \lambda(t), \qquad t_0 \leqq t \leqq T_0,$$

for all controls $z(t)$. There are two cases to conside .

Case 1. $\tilde{x}_m(t_0 + \sigma_0\delta) - \bar{x}_m(t_0 + \sigma_0\delta) \geqq \frac{1}{6}\gamma_0\eta\varepsilon$. In this case, set

$$\begin{aligned}\tilde{\Gamma}^{\delta,j}(z_1, \tilde{y}_1, \cdots, z_j)(t) &= \tilde{y}_j(t)\\ &= \bar{y}_j(t) = \Gamma^{\delta,j}(z_1, \bar{y}_1, \cdots, z_j)(t)\end{aligned} \tag{2.23}$$

for $j > \sigma_0$, until there is a $j \leqq n$ such that $\bar{x}_m(t) > \varepsilon$ for some $t \in I_j$.

If there is no $j < n$ such that $\bar{x}_m(t) > \varepsilon$ for some $t \in I_j$, we may apply $(\bar{Q}_2)$ in a manner similar to the proof of Theorem 6.2.1 [1, p. 217]. We get

$$\tilde{x}_m(t) \geqq \bar{c}\eta\varepsilon \quad \text{on } [t_0 + \sigma_0\delta, T_0], \tag{2.24}$$

where $\bar{c}$ is a constant depending only on the system (2.1). Since $\tilde{y}(t) = \bar{y}(t)$ except on a set of measure at most $\eta\varepsilon$, we see as in Theorem 6.2.1 [1, p. 220], that (2.19) holds.

If there is a $j < n$ such that for some $t \in I_j$, $\bar{x}_m(t) > \varepsilon$, then we may proceed exactly as in the proof of Theorem 6.2.1 [1, pp. 214–220]. We get

$$\tilde{x}_m(t) > c^*\eta \quad \text{on } [t_0 + \sigma_0\delta, T_0]. \tag{2.25}$$

Since $\tilde{y}(t) = \bar{y}(t)$, except on a set of measure at most $K\eta$, where K is a constant depending only on the system (2.1), we again have as in Theorem 6.2.1 [1, p. 220], that (2.19) holds.

Case 2. $0 \leqq \tilde{x}_m(t_0 + \sigma_0\delta) - \bar{x}_m(t_0 + \sigma_0\delta) < \frac{1}{6}\gamma_0\eta\varepsilon$. Set $s'_1 = t_0 + \sigma_0\delta$, and define $s'_2, I_{k_1}, I_{k_2}, \bar{s}$ in a manner similar to the proof of Theorem 6.2.1 [1, p. 213], and proceed to modify Γ^δ into $\tilde{\Gamma}^\delta$ as there. There are two cases to consider, according as $\bar{s} = s'_2$, or $\bar{s} < s'_2$. (Here $\bar{s}$ is the right-hand endpoint of the last interval of modification and $\bar{s} = s'_2$ corresponds to case (b), p. 215; $\bar{s} < s'_2$ corresponds to case (a), p. 215.) In either case, it is easily seen that $\tilde{x}_m(t) > 0$ on $[s'_1, \bar{s}]$ if $\eta\varepsilon$ is small enough.

If $\bar{s} = s'_2$, we then have that $\tilde{x}_m(t) > 0$ on $[s'_1, s'_2]$, and that $\tilde{x}_m(s'_2) \geqq \bar{x}_m(s'_2) - \bar{K}\eta\varepsilon$, since $\tilde{y} = \bar{y}$ except on a set of measure at most $2\eta\varepsilon$. By definition of s'_2, if $\delta(\varepsilon)$ is small enough, then $\bar{x}_m(s'_2) \geqq \frac{3}{4}\varepsilon$, hence $\tilde{x}_m(s'_2) \geqq \varepsilon[\frac{3}{4} - \bar{K}\eta] > \frac{1}{2}\varepsilon$ if η_0 is sufficiently small. We may now finish the construction of $\tilde{\Gamma}^\delta$ from Γ^δ as in the proof of Theorem 6.2.1 [1], with ε replaced by $\varepsilon/2$.

In case $\bar{s} < s'_2$, it must be shown that

$$\tilde{x}_m(\bar{s}) - \bar{x}_m(\bar{s}) \geqq \tfrac{1}{6}\gamma_0\eta\varepsilon. \tag{2.26}$$

But since $\tilde{x}_m(s'_1) \geqq \bar{x}_m(s'_1)$ and $\tilde{x}_m(t) > 0$ on $[s'_1, \bar{s}]$, (2.26) follows using exactly the same arguments used in the proof of (6.2.32), p. 217. Since (2.26) holds, we may now finish the construction of $\tilde{\Gamma}^\delta$ from Γ^δ as in Case 1. This completes the proof of the lemma.

We now have the following situation.

There is a δ-strategy $\tilde{\Gamma}^\delta$ for y, such that the $\tilde{\Gamma}^{\delta,j}$ are constant maps, for $1 \leqq j \leqq \sigma_0$, and such that if $z(t)$ is any control for z, $y(t)$ any control for y, $\tilde{y}(t) = \tilde{\Gamma}^\delta(z(t))$, then

$$\tilde{x}_m(t) \geqq x_m(t), \qquad t_0 \leqq t \leqq t_0 + \sigma_0\delta, \tag{2.27}$$

$$\tilde{x}_m(t_0 + \sigma_0\delta) \geqq \tfrac{1}{2}\gamma_0\eta\varepsilon \equiv \varepsilon' \tag{2.28}$$

and (2.19) holds.

We can now construct from $\tilde{\Gamma}^{\delta}$, a δ-strategy $\tilde{\tilde{\Gamma}}^{\delta}$ exactly as in the proof of Theorem 6.2.1 [1], with ε, η, t_0 replaced by $\varepsilon', \eta', t_0 + \sigma_0\delta$ for any η' sufficiently small. We get

$$\tilde{\tilde{x}}_m(t) > c^*\eta' > 0 \quad \text{on } [t_0 + \sigma_0\delta, T_0], \tag{2.29}$$

$$V^{\delta} \leqq P(\tilde{\tilde{\Gamma}}^{\delta}(z), z) + \gamma + \alpha(\eta) + \beta(\eta'), \tag{2.30}$$

where $\beta(\eta') \to 0$ if $\eta' \to 0$, so that (2.20) holds.

Since $\tilde{\tilde{\Gamma}}_1^{\delta,j}$, $1 \leqq j \leqq \sigma_0$, are constant maps, such that if $\tilde{\tilde{x}}(t)$ is the trajectory corresponding to $(\tilde{\tilde{\Gamma}}^{\delta}(z), z)$, where z is any control for player z, $\tilde{\tilde{x}}_m(t) \geqq \frac{1}{2}\gamma_0(t - t_0)$ for $t_0 \leqq t \leqq t_0 + \sigma_0\delta$, we may complete the proof in the following manner.

Let $\tilde{\Delta}^{\delta}$ be any δ-strategy for z such that

$$V_{\delta} \geqq P[\tilde{\Delta}^{\delta}, \Gamma_{\delta}] - \gamma \quad \text{for all } \Gamma_{\delta}. \tag{2.31}$$

Set $y_1 = (y'_{*,1}(t), y''_1(t))$, where $\{y'_{*,1}(t)\} = \text{range } \tilde{\tilde{\Gamma}}_1^{\delta,1}$ and $y''_1(t)$ is an arbitrary control for y with values in Y_2. Set $z_1 = \tilde{\Delta}^{\delta,1}(y_1)$. For $1 \leqq j, k \leqq \sigma_0$, set $\{y'_{*,j}(t)\} = \text{range } \tilde{\tilde{\Gamma}}_1^{\delta,j}$ and $y_k''^{\tau}(t) = y''_k(t + \delta)$ in I_{k-1}. We take for $t \in I_j$, $1 \leqq j \leqq \sigma_0$,

$$\begin{aligned} y'_j(t) &= y'_{*,j}(t), \\ y''_j(t) &= \tilde{\tilde{\Gamma}}_2^{\delta,j-1}(z_1, (y'_{*,1}, y_2''^{\tau}), \cdots, (y'_{*,j-2}, y_j''^{\tau}), z_{j-1})(t - \delta), \\ z_j(t) &= \tilde{\Delta}^{\delta,j}(y_1, z_1, \cdots, z_{j-1}, y_j)(t). \end{aligned} \tag{2.32}$$

For $k > \sigma_0$, set $y_k'^{\tau}(t) = y'_k(t + \delta)$ in I_{k-1}, and for $j > \sigma_0$, $t \in I_j$, we take

$$\begin{aligned} y'_j(t) &= \tilde{\tilde{\Gamma}}_1^{\delta,j-1}(z_1, (y_{*,1}, y_2''^{\tau}), \cdots, z_{j-2}, (y_{j-2}'^{\tau}, y_{j-2}''^{\tau}), z_{j-1})(t - \delta), \\ y''_j(t) &= \tilde{\tilde{\Gamma}}_2^{\delta,j-1}(z_1, (y_{*,1}, y_2''^{\tau}), \cdots, z_{j-2}, (y_{j-2}'^{\tau}, z_{j-2}'^{\tau}), z_{j-1})(t - \delta), \\ z_j(t) &= \tilde{\Delta}^{\delta,j}(z_1, y_1, \cdots, z_{j-1}, y_j)(t). \end{aligned} \tag{2.33}$$

Using a slight modification of Lemma 2.3.1 [1, p. 40], with $\lambda = \delta$, and

$$y_{\delta}(t) = \tilde{\tilde{y}}_{\delta}(t) = (\tilde{\tilde{y}}'_{\delta}(t), \tilde{\tilde{y}}''_{\delta}(t)), \tag{2.34}$$

$$\hat{y}_{\delta}(t) = \begin{cases} \tilde{\tilde{y}}'_{\delta}(t), & t_0 \leqq t \leqq t_0 + \sigma_0\delta, \\ \tilde{\tilde{y}}''_{\delta}(t - \delta), & t_0 + \delta \leqq t \leqq T_0, \\ \tilde{\tilde{y}}(t - \delta), & t_0 + (\sigma_0 + 1)\delta \leqq t \leqq T_0, \end{cases} \tag{2.35}$$

the result follows as in Theorem 6.2.1 [1, p. 221].

Remark 3. It follows upon consideration of the above proof that Theorem 2.1 remains valid whenever $(\bar{\mathrm{P}}'_2)$(d) is replaced by any condition guaranteeing the existence of a control $y'_*(t)$ in Y_1 such that for any controls $y''(t) \in Y_2, z(t)$,

$$f_m(t, x^*(t), y'_*(t), y''(t), z(t)) > c_0 \tag{2.36}$$

for all $t \in [t_0, t_0 + \sigma_0\delta]$, if $x^*(t) \in \partial X_y^{\bar{\varepsilon}^*}$, where $c_0 > 0$ is a constant depending on the system (2.1), but independent of $y''(t)$, $z(t)$, η, ε, δ.

Remark 4. It is clear that the construction of $\tilde{\Gamma}^{\delta}$ as in Lemma 2.1 and the construction of $\tilde{\tilde{\Gamma}}^{\delta}$ as in Theorem 2.1 can be carried out whenever $x_0 \in \text{int } X$ and $0 < x_m^0 \leqq \bar{\varepsilon}_0$, where $\bar{\varepsilon}_0$ is sufficiently small. Hence, Lemma 2.1 and Theorem 2.1 are valid whenever $x_0 \in X$ and $0 \leqq x_m^0 \leqq \bar{\varepsilon}_0$, where $\bar{\varepsilon}_0$ is sufficiently small.

3. Elimination of the condition (T). Denote by $V(\tau, \xi)$ the value of a game associated with (2.1),

$$x(\tau) = \xi, \tag{3.1}$$

(2.3), (2.4) and the payoff

$$P_{\tau,\xi}(y, z) = g(x(T_0)) + \int_{\tau}^{T_0} h(t, x(t), y(t), z(t))\, dt \tag{3.2}$$

with penalties $\pm M$.

LEMMA 3.1. *Let G be a game associated with* (2.1), (2.3), (2.4), (3.1), (3.2) *satisfying the conditions of Theorem* 2.1. *Then $V(\tau, \xi)$ is continuous on* $[t_0, T_0) \times X$.

Proof. For $\xi \in \operatorname{int} X$, Theorem 3.1 follows from Theorem 6.6.1 [1, p. 236]. Assume, therefore, that $\xi \in \partial X_y$. We begin by choosing $\tilde{\Gamma}^\delta$ as in Lemma 2.1. The proof proceeds as in Theorem 6.6.1 except that now we must also show that $(\psi\tilde{\bar{x}})_m(t) > 0$ for $\bar{\tau} \leqq t \leqq \bar{\tau} + \sigma_0\bar{\delta}$. But this follows easily from the continuity of $f(t, x, y, z)$ in t. We now have that

$$V^\delta(\tau, \xi) \leqq V^{\bar{\delta}}(\bar{\tau}, \bar{\xi}) + \eta_0(|\tau - \bar{\tau}| + |\xi - \bar{\xi}|),$$

where $\eta_0(s) \to 0$ as $s \to 0$.

Upon considering Remark 4 of § 2 and the proof of the above inequality, it is clear that we may reverse the roles of (τ, ξ) and $(\bar{\tau}, \bar{\xi})$ to get the reverse inequality.

Let

$$W^n(\tau, \xi) = V^\delta(\tau, \xi), \quad W_n(\tau, \xi) = V_\delta(\tau, \xi), \qquad \delta = (T_0 - t_0)/n. \tag{3.3}$$

LEMMA 3.2. *Let G be a game as in Lemma* 3.1. *Then*

$$\lim_{n\to\infty} W^n(\tau, \xi) = V(\tau, \xi), \qquad \lim_{n\to\infty} W_n(\tau, \xi) = V(\tau, \xi) \tag{3.4}$$

uniformly on compact subsets of $[t_0, T_0) \times X$.

Proof. The proof is the same as that of Theorem 6.6.3 [1, p. 238].

Note. The value of a game also depends on T_0, and so we may write $V(\tau, \xi, T_0)$ instead of $V(\tau, \xi)$. Also, it is clear that under the assumptions of Lemma 3.1, $V(\tau, \xi, T_0)$ is continuous in T_0.

LEMMA 3.3. *Let $\{G_n\}$ be a sequence of games associated with* (2.1), (2.3), (2.4), (3.1) *and*

$$P_n^{\tau,\xi}(y, z) = g_n(x(T_0)) + \int_{\tau}^{T_0} h_n(t, x(t), y(t), z(t))\, dt \tag{3.5}$$

with common penalties $\pm M$. Then if

(i) $g_n(x) \to g(x)$, $h_n(t, x, y, z) \to h(t, x, y, z)$ *uniformly on compact subsets of* X, $[t_0, T_0] \times X \times Y \times Z$, *respectively*

(ii) *the game G associated with* (2.1), (2.3), (2.4), (3.1) *and*

$$P_{\tau,\xi}(y, z) = g(x(T_0)) + \int_{\tau}^{T_0} h(t, x(t), y(t), z(t))\, dt \tag{3.6}$$

with penalties $\pm M$ has value $V(\tau, \xi)$, and

(iii) *if $V_n(\tau, \xi)$ denotes the value of G_n*, *then*

(3.7) $$V_n(\tau, \xi) \to V(\tau, \xi)$$

uniformly on compact subsets of $[t_0, T_0] \times X$.

Proof. It is clear that $|P_n^{\tau,\xi}(y, z) - P_{\tau,\xi}(y, z)|$ is small, uniformly in y, z when there are no penalties associated with the controls y, z. If there is a penalty $\pm M$, then the payoffs are the same.

We are now in a position to eliminate condition (T).

THEOREM 3.1. *Let G be a game associated with* (2.1)–(2.4). *Assume* (A_1)–(A_4), (B_1), (B_2'), (B_3), $(\bar{P}_2')$ *and* $(\bar{Q}_2)$ *hold. Then if* $x_0 \in X$, *G has value.*

Proof. Assume first that $T_0 - t_0 < 2T(\bar{\varepsilon})$.

Let Π_n denote the partition of $[t_0, T_0]$ given by $\{t_0 + j\delta | \delta = (T_0 - t_0)/n, j = 0, 1, \cdots, n\}$. Set

$$\Pi_n^1 : t_0 < t_1 < \cdots < t_{[n/2]} = \bar{t}_1,$$

$$\Pi_n^2 : \bar{t}_1 < t_{[n/2]+1} < \cdots < t_n = T_0.$$

Next, consider the following sequence of games, each with the dynamics of the game G, and with payoffs:

$$\hat{G}_1: \quad P_1(y, z) = \int_{t_0}^{\bar{t}_1} h(t, x(t), y(t), z(t))\, dt$$

with penalties $\pm M$;

$$G_2: \quad P_2(y, z) = g(x(T_0)) + \int_{\bar{t}_1}^{T_0} h(t, x(t), y(t), z(t))\, dt$$

with penalties $\pm M$;

$$G_n: \quad P_n(y, z) = V_2^\delta(\bar{t}_1, x_1) + \int_{t_0}^{\bar{t}_1} h(t, x(t), y(t), z(t))\, dt$$

with penalties $\pm M$, where $x(t)$ is the trajectory corresponding to $y, z, x_1 = x(\bar{t}_1)$ and $V_2^\delta(\bar{t}_1, x_1)$ is the upper δ-value of G_2 for $\delta = (T_0 - t_0)/n$.

Denote by $V_n^\delta(t_0, x_0, \bar{t}_1)$ the upper δ-value of G_n for the partitions Π_n^1.

$$G_3: \quad P_3(y, z) = V_2(\bar{t}_1, x_1) + \int_{t_0}^{\bar{t}_1} h(t, x(t), y(t), z(t))\, dt$$

with penalties $\pm M$, where $V_2(\bar{t}_1, x_1)$ is the value of G_2 at $(\bar{t}_1, x_1), x_1$ as in G_n. Denote the value of G_3 by $V_3(t_0, x_0, \bar{t}_1)$; the upper δ-value by $V_3^\delta(t_0, x_0, \bar{t}_1)$.

Note that $\bar{t}_1$ depends on n; $\bar{t}_1 \to (T_0 + t_0)/2$ if $n \to \infty$. Also note that the penalties $\pm M$ may be chosen to be the same and that there is a number n_0 such that for $n \geqq n_0, G_1, G_2, G_3, G_n$ all have value. Denote by $V^{\delta,n}(t_0, x_0)$ the upper δ-value of G for $\delta = (T_0 - t_0)/n$.

Finally, denote by $V(t_0, x_0)$ the value of the game G. We shall show that $V(t_0, x_0)$ exists and

(3.8) $$V(t_0, x_0) = V_3(t_0, x_0, (T_0 + t_0)/2).$$

Clearly it suffices to show that

$$\lim_{\delta\to 0} V^{\delta}(t_0, x_0) = V_3(t_0, x_0, (T_0 + t_0)/2), \tag{3.9}$$

$$\lim_{\delta\to 0} V_{\delta}(t_0, x_0) = V_3(t_0, x_0, (T_0 + t_0)/2). \tag{3.10}$$

We prove only (3.9), the proof of (3.10) being similar. Write

$$\Delta_\delta = (\Delta_{\delta,1}, \cdots, \Delta_{\delta,[n/2]}, \Delta_{\delta,[n/2]+1}, \cdots, \Delta_{\delta,n}) = (\Delta_\delta^1, \Delta_\delta^2),$$

where $\Delta_\delta^1 = (\Delta_{\delta,1}, \cdots, \Delta_{\delta,[n/2]})$, $\Delta_\delta^2 = (\Delta_{\delta,[n/2]}, \cdots, \Delta_{\delta,n})$.

Define $\Gamma^\delta = (\Gamma_1^\delta, \Gamma_2^\delta)$ similarly.

For $n = 2m$, we have

$$V^{\delta,2m}(t_0, x_0) = \inf_{\Delta_\delta^1} \sup_{\Gamma_1^\delta} [P_1[\Delta_\delta^1, \Gamma_1^\delta] + V_2^\delta(\bar{t}_1, x_1)] = V_{2m}^\delta(t_0, x_0), \tag{3.11}$$

where $\delta = (T_0 - t_0)/2m$. Also

$$V_3^\delta(t_0, x_0, \bar{t}_1) = \inf_{\Delta_\delta^1} \sup_{\Gamma_1^\delta} [P_1[\Delta_\delta^1, \Gamma_1^\delta] + V_2(\bar{t}_1, x_1)]. \tag{3.12}$$

Note that in (3.11), (3.12), $\bar{t}_1 = (T_0 + t_0)/2$.

By Lemma 3.3,

$$V_{2m}^\delta(t_0, x_0, (T_0 + t_0)/2) \to V_3(t_0, x_0, (T_0 + t_0)/2) \tag{3.13}$$

as $m \to \infty$; thus by (3.11),

$$V^{\delta,2m}(t_0, x_0) \to V_3(t_0, x_0, (T_0 + t_0)/2) \quad \text{as } m \to \infty. \tag{3.14}$$

It follows from the continuity of $V_3(t_0, x_0, \bar{t}_1)$ in $\bar{t}_1$ and from Lemma 3.2 that

$$|V^{\delta,2m+1}(t_0, x_0) - V^{\bar{\delta},2m}(t_0, x_0)| \to 0 \quad \text{as } m \to \infty. \tag{3.15}$$

This gives (3.9) and completes the proof in case $T_0 - t_0 < 2T(\bar{\varepsilon})$. We now proceed step by step to the case $T_0 - t_0 < kT(\bar{\varepsilon})$; this then yields the result for $[t_0, T_0]$ of arbitrary length whenever (A_1)–(A_4), (B_1), (B_2'), (B_3), $(\bar{P}_2')$, $(\bar{Q}_2)$ hold on $[t_0, T_0]$.

Remark 5. It is easily seen that all of the results on continuity of $V(t_0, x_0)$ now hold without the assumption (T).

Remark 6. The results of this paper clearly extend to more general domains X (see [1, Chap. 6]) if suitable modifications of assumptions (R_1)–(R_8) are made.

Remark 7. The following well-known games satisfy the conditions of Theorem 2.1:

(a) war of attrition and attack;
(b) modified war of attrition and attack;
(c) Battle of Bunker Hill.

See [1, pp. 231–234].

Acknowledgment. I wish to thank Avner Friedman for his helpful suggestions in making the proofs more lucid and also for first suggesting the argument presented in Theorem 3.1.

REFERENCE

[1] A. FRIEDMAN, *Differential Games*, Wiley-Interscience, New York, 1971.

SIAM J. CONTROL
Vol. 12, No. 3, August 1974

SYMMETRIES OF AUTONOMOUS LINEAR CONTROL SYSTEMS*

ETHELBERT NWAKUCHE CHUKWU†

Abstract. This paper is concerned with symmetry and identification problems for time-optimal linear autonomous control systems in Euclidean space. A symmetry is a nonsingular linear transformation which reproduces each of the reachable sets. Under rather general conditions all the symmetries of a control system are described constructively (Theorem 4). In an analogous situation, it is determined when two control systems have the same reachable sets: the identification problem (Theorem 6). Both of these are special cases of a wider result (Theorem 3) which may well be amenable to further generalization.

1. Introduction. The main object of this work is to study the symmetries and the identification problems of autonomous linear control systems described by the equations

$$(\mathrm{S}_i) \qquad \dot{x}_i(t) = A_i x_i(t) - p_i(t), \qquad i = 1, 2,$$

where A_i is a real constant $n_i \times n_i$ matrix and $p_i : [0, t] \to P_i$ is summable. It is assumed as basic that P_i is a fixed compact and convex subset of R^{n_i} containing 0; we do not require that P_i be a parallelepiped, or even that it be symmetric.

In this section the basic definitions of symmetries of control systems (S_i) and of their reachable sets are stated. The general framework of our study is then formulated. Our main result is proved in § 2. This specializes in § 3 to a constructive description of all symmetries of a control system. The effectiveness of the method is tested on a special class of systems. Section 4 studies the relation between systems having the same performance, i.e., having coinciding reachable sets. It is shown that, in a sense, systems with the same performance are almost identical. The last section indicates a number of intriguing unresolved questions connected with our research.

The basic definitions and properties of linear time-optimal control systems may be found in Hermes and LaSalle [8] and in Lee and Markus [10]. General references for the properties of convex sets are Eggleston [1] and Stoer and Witzgall [12]. The terminology and notations of [8] will be maintained.

Consider the control system

$$(\mathrm{S}) \qquad \dot{x}(t) = Ax(t) - p(t);$$

here x is an n-vector and $p : [0, t] \to P$ is summable with values constrained to lie in a fixed compact and convex set P, and the matrix A is of type n, i.e., A is an $n \times n$ constant matrix.

Remarks. Let r^n be the space of compact subsets of R^n with the Hausdorff metric topology. The fixed compact set P can be considered as the constant set-valued map $P : [0, t] \to r^n$, which is convex.

* Received by the editors February 22, 1973.

† Department of Mathematics, Cleveland State University, Cleveland, Ohio 44115.

We use $-p(t)$ instead of $p(t)$ in order to obtain a nice representation for the reachable set of (S) when the set-valued map $P:[0,t] \to r^n$ is not symmetric.

DEFINITION 1. The *reachable set* $\mathscr{R}(t)$ of (S) is

$$\mathscr{R}(t) = \left\{ \int_0^t e^{-As}p(s)\,ds : p:[0,t] \to P \text{ summable} \right\}.$$

The following properties of the reachable set are well known: For each $t \geqq 0$, $\mathscr{R}(t)$ is compact and convex whenever the constraint set P is such. Also, $\mathscr{R}(t)$ is symmetric whenever P is symmetric [8, p. 46].

DEFINITION 2. A nonsingular n-square matrix M is called a *symmetry of the reachable sets of* (S) if $M\mathscr{R}(t) = \mathscr{R}(t)$ for each $t \geqq 0$.

Also, M is called a *transformation of the reachable sets of* (S_1) *and* (S_2) if

(1) $$M\mathscr{R}_1(t) = \mathscr{R}_2(t)$$

for each $t \geqq 0$, and if it is a monomorphism (that is, $MX = 0$ if and only if $X = 0$).

DEFINITION 3. A nonsingular n-square matrix M is a *symmetry of the control system* (S) if

(2) $$AM = MA,$$

(3) $$MP = P.$$

Also, M is called a *transformation between the systems* (S_1) *and* (S_2) if it is a monomorphism, and

(4) $$MA_1 = A_2M,$$

(5) $$MP_1 = P_2.$$

The motivation for Definition 3 is that a nonsingular linear transformation $x \to y = Mx$ takes the system (S) into the system

$(\hat{\mathrm{S}})$ $$\dot{y}(t) = MAM^{-1}y - Mp(t).$$

A complete description of the symmetries of the reachable sets $\mathscr{R}(t)$ of (S) is the symmetry problem. Intimately related to this is the identification problem. When two different systems in n-space have the same performance [4, p. 349] or coinciding reachable sets, then a transformation of the reachable sets is $M = I$, where I is the identity matrix. It is then natural to ask what other properties the two systems must have. A complete answer to this problem is treated in § 4.

PROPOSITION 1 (Addition formula for reachable sets). *If in* (S), $0 \in P$ *and* $\mathscr{R}(\cdot)$ *is the reachable set, then for each* $t \geqq 0$, $s \geqq 0$,

(6) $$\mathscr{R}(t) + e^{-At}\mathscr{R}(s) = \mathscr{R}(t+s).$$

Both inclusions needed here are immediate. One of these is a remark in [3, p. 348].

LEMMA 1. *Let* r^n *be the space of compact subsets of a ball* B_k^n *of radius* k *in* R^n, *endowed with the Hausdorff metric topology. Let* $P:[0,T] \to r^n$ *be continuous. Then*

$$\lim_{h \to \infty} \frac{1}{h}\int_t^{t+h} P(\tau)\,d\tau = \operatorname{co} P(t) \quad \textit{for all } t \in [0,T].$$

Proof. The proof of this lemma is almost word for word the proof given by Hermes [7, Lemma 1.3] for a similar result when P is Borel measurable. Indeed,

$$\lim_{h\to\infty} \frac{1}{h} \int_t^{t+h} P(\tau)\, d\tau = \operatorname{co} P(t)$$

(where $\operatorname{co} P$ is the set-valued function with values the convex hull of $P(t)$) for all points of approximate continuity of P, which are almost every $t \in [0, T]$. Because P is continuous, every point $t \in [0, T]$ is a point of approximate continuity. The result is now immediate.

LEMMA 2. *Let* $P_i : [0, T] \to r^n$ *be continuous. Let*

$$\mathscr{A}_i(t) = \int_0^t P_i(\tau)\, d\tau, \qquad i = 1, 2.$$

Then

$$\mathscr{A}_1(t) = \mathscr{A}_2(t) \quad \textit{for all } t \in [0, T]$$

if and only if

$$\operatorname{co} P_1(t) = \operatorname{co} P_2(t) \quad \textit{for all } t \in [0, T],$$

where $\operatorname{co} P_i$ *is the set-valued function with values the convex hull of* $P_i(t)$.

Proof. Let $\mathscr{A}_1(t) = \mathscr{A}_2(t)$ for all $t \in [0, T]$. Because P_i is (at least) Borel measurable, it follows from Hermes [7, Thm. 1.1] that there exists a dense subset D of $[0, T] \equiv I$ with the following property: The Lebesque measure of the complement of D in I is zero; and

$$\operatorname{co} P_1(\theta) = \operatorname{co} P_2(\theta) \quad \text{for all } \theta \in D.$$

For any arbitrary $t \in I$, there exists a sequence $\{\theta_n\}$ of points of D such that $\theta_n \to t$ as $n \to \infty$, and

$$\operatorname{co} P_1(\theta_n) = \operatorname{co} P_2(\theta_n).$$

Because P_i is continuous so is $\operatorname{co} P_i$. Indeed, if $\{P_u\}$ is a sequence of set-valued maps such that

$$P_u \to P \quad \text{as } u \to \infty,$$

then

$$\operatorname{co} \overline{\lim}\, P_u \subseteq \overline{\lim} \operatorname{co} P_u \subseteq \operatorname{co} \underline{\lim}\, P_u,$$

where $\underline{\lim}\, P_u = \liminf P_u$, $\overline{\lim}\, P_u = \limsup P_u$. Consequently,

$$\operatorname{co} P_i(\theta_n) \to \operatorname{co} P_i(t) \quad \text{as } n \to \infty,$$

whenever $\theta_n \to t$. Because r^n is Hausdorff it follows from the uniqueness of limits that

$$\operatorname{co} P_1(t) = \operatorname{co} P_2(t).$$

This completes the proof since t is arbitrary; and the converse is trivial.

The reachable set $\mathscr{R}(t)$ of (S) can be considered as a set-valued mapping $\mathscr{R} : [0, \infty) \to r^n$, where r^n is the set of compact subsets of $B_k^n \subset R^n$, with the metric topology.

THEOREM 1 (Hájek). *In* (S) *let P be a compact set in R^n such that $0 \in P$. Then*

$$\lim_{t\to 0+} \frac{\mathscr{R}(t)}{t} = \operatorname{co} P,$$

where the limit is taken in r^n.

Proof. A direct proof is contained in [5]. It also follows immediately from Lemma 1.

Because P is fixed and e^{-As} continuous, the function

$$e^{-A(\cdot)}P:[0, \infty) \to r^n$$

is continuous. Therefore,

$$\lim_{h\to 0} \frac{1}{h}\int_0^h e^{-A\sigma}P\,d\sigma = \operatorname{co} P, \tag{7}$$

where we have used Lemma 1.

DEFINITION 4. The *controllability space* $\mathscr{C}(t)$ of (S), for each $t > 0$, is the least linear space containing $\mathscr{R}(t)$.

It is a consequence of Theorem 2 that $\mathscr{C}(t)$ is independent of t (and will, therefore, be denoted by $\mathscr{C}$).

DEFINITION 5. The system (S) in n-space is controllable if and only if the controllability space is R^n.

The next result is well known. Its present formulation is contained in [5], where a detailed proof is given.

THEOREM 2. *The following are equivalent descriptions of* $\mathscr{C}$:

(8) *linear span of* $\mathscr{R}(t)$;

(9) *linear span of* $P, AP, \cdots, A^{n-1}P$;

(10) *least linear space containing P and invariant under all e^{-As}, $s \in R^1$*;

(11) *least linear space containing P invariant under A.*

COROLLARY 1. *If* (S) *is controllable, then R^n coincides with the linear span of P, $AP, \cdots, A^{n-1}P$.*

By the linear span of P, AP, $\cdots$, $A^{A-1}P$, we mean the span of the union of these sets.

If the convexity assumption for P is removed, it can be verified that

$$\mathscr{C} = \operatorname{span} \bigcup_{i=0}^{n-1} A^i \operatorname{co} P.$$

Thus $\mathscr{C}$ is the least linear space containing co P invariant under A.

In what follows we shall abbreviate linear span to span.

The next proposition is probably well known, but its statement as well as its proof is difficult to locate in the literature. It is formulated in its present form by Hájek. For some other important properties of convex sets the general references are [1] and [12].

PROPOSITION 2. *Suppose A, B, C are subsets of R^n such that*

$$A + C \subseteq B + C.$$

Suppose B is closed and convex, C nonempty, compact and convex. Then

$$A \subseteq B. \tag{12}$$

Proof. Obviously we may assume $B \neq \varnothing$. Suppose $A \nsubseteq B$, and aim at a contradiction. Then there exists some $a_0 \in A \setminus B$. Since B is nonvoid, closed and convex, there exists a point $b_0 \in B$ closest to a_0; furthermore, $d \equiv a_0 - b_0 \neq 0$, and d is an exterior normal to B at b_0, that is,

$$(d, b - b_0) \leqq 0 \tag{13}$$

for all $b \in B$ [8, p. 33 ff.].

Because C is a nonempty, compact and convex set, there exists a point $c_0 \in C$ at which d is an exterior normal to C:

$$(d, c - d_0) \leqq 0 \quad \text{for all } c \in C. \tag{14}$$

Inequalities (13) and (14) yield

$$(d, b + c - (b_0 + c_0)) \leqq 0$$

for all $b \in B, c \in C$. It follows that d is an exterior normal to $B + C$ at $b_0 + c_0$.

Now consider the point $a_0 + c_0 \in A + C$. Clearly $a_0 + c_0 \in B + C$. But then

$$0 \geqq (d, a_0 + c_0 - (b_0 + c_0)) = (d, a_0 - b_0) \equiv (d, d) = \|d\|^2.$$

This implies that $d = 0$, a contradiction.

Remark. The compactness assumption of C cannot be omitted. Consider, for example, $C = R^1, B = \{0\}, A = \{2\}$. Then $A + C \subseteq B + C$, but $A \nsubseteq B$.

COROLLARY 2. *Suppose A, B, C are nonvoid, compact and convex subsets of R^n such that*

$$A + B = A + C.$$

Then

$$B = C.$$

2. Main theorem.

Consider two autonomous linear control systems

$$\dot{x}_1(t) = A_1 x_1(t) - P_1, \qquad P_1(\cdot) \in P_1 \subseteq R^{n_1}, \tag{S^1}$$

$$\dot{x}_2(t) = A_2 x_2(t) - P_2, \qquad P_2(\cdot) \in P_2 \subseteq R^{n_2}, \tag{S^2}$$

where A_i is a real n_i-square matrix and $P_i : [0, t] \to P_i$ is summable for $i = 1, 2$. It is assumed that P_i is a compact subset of R^{n_i}.

THEOREM 3. *For $i = 1, 2$ assume that $0 \in P_i$, and that the set Q_i of extreme points of* co P_i *is countable. Let M* be a matrix monomorphism $R^{n_1} \to R^{n_2}$. *Then*

$$M\mathscr{R}_1(t) = \mathscr{R}_2(t) \tag{15}$$

for each $t \geqq 0$, *if and only if*

$$M \operatorname{co} P_1 = \operatorname{co} P_2, \tag{16}$$

$$MA_1 = A_2 M \quad \textit{in } \mathscr{C}_1 \tag{17}$$

(*in the sense that* $MA_1x = A_2Mx$ *for all* $x \in \mathscr{C}_1$, *the controllability space of* (S_1)).

Proof. Assume (15); that is,

$$\int_0^t M\, e^{-A_1 s} P_1\, ds = \int_0^t e^{-A_2 s} P_2\, ds. \tag{18}$$

Then from Lemma 2 it is necessary and sufficient that

$$M\, e^{-A_1 s} \operatorname{co} P_1 = e^{-A_2 s} \operatorname{co} P_2 \tag{19}$$

for all S. In (19) if $s = 0$, then

$$M \operatorname{co} P_1 = \operatorname{co} P_2, \tag{20}$$

condition (16). The set-valued functions in (19) are defined on $(0, t)$ and both $M\, e^{-A_1 s}$ and $e^{-A_2 s}$ do not change sign in $(0, t)$. Then using the ideas of Banks and Jacobs [13, p. 250], we can π-differentiate the set functions with respect to s; and on setting $s = 0$ in the derivative, one deduces that

$$MA_1^k \operatorname{co} P_1 = A_2^k M \operatorname{co} P_1, \qquad k = 0, 1, 2, \cdots. \tag{21}$$

Using the remarks following Corollary 1 and the uniqueness of solutions of differential equations, one obtains the desired result:

$$MA_1 = A_2 M \quad \text{in } \mathscr{C}. \tag{22}$$

This approach seems to lead, however, to some obscurities in the crucial role of the assumption of countability of extreme points of co P_i. Therefore a more direct proof involving differentiation of point-valued functions will be given.

In (19), since $M\, e^{-A_1 s}, e^{-A_2 s}$ define monic linear mappings, the sets Q_i of extreme points satisfy

$$M\, e^{-A_1 s} Q_1 = e^{-A_2 s} Q_2. \tag{23}$$

Denote by $Q_i = \{q_{ik}; k = 1, 2, \cdots\}$ for $i = 1, 2$ the countable set of extreme points of co P_i. Pick any extreme point $q \in Q_i$. Then for each $s \geqq 0$, there is an index $j = j(s)$ such that

$$M\, e^{-A_1 s} q = e^{-A_2 s}{}_{2j(s)}. \tag{24}$$

For each $n = 1, 2, \cdots$, consider the set $A_n = \{s : j(s) = n\}$. Evidently $[0, \infty) = \bigcup_{n=1}^{\infty} A_n$. Since the right-hand side is a countable union of sets, and the left an uncountable set, there is an index n and an uncountable set of real numbers s such that $j(s) = n$. Thus

$$M\, e^{-A_1 s} q = e^{-A_2 s} q_{2n} \tag{25}$$

for an uncountable set of real numbers $s \in [0, \infty)$. Since both sides of (25) define analytic functions of s, (25) actually holds for all $s \geqq 0$.

On differentiating at $s = 0$ we obtain

$$MA_1^k q = A_2^k q_{2n} \tag{26}$$

for all k. In particular ($k = 0$), $Mq = q_{2n}$, and so

$$MA_1^k q = A_2^k Mq \quad (\text{all } k, \text{all } q \in Q_1). \tag{27}$$

Since $\mathscr{C}_i$ is the linear span of

$$\operatorname{co} P_i, A_i \operatorname{co} P_i, \cdots, A_i^{n_i-1} \operatorname{co} P_i$$

(Theorem 2), and hence also of $Q_1, A_iQ_i, \cdots, A_i^{n_i-1}Q_i$, every point $x \in \mathscr{C}_1$ can be expressed as a linear combination

$$x = \sum_k \alpha_k A_1^{r_k} q_{1r_k}.$$

Then

$$MA_1x = \sum_k \alpha_k MA_1^{r_k+1} q_{1r_k} = \sum_k \alpha_k A_2^{r_k+1} q_2 r_k,$$

by (26); and also

$$A_2Mx = A_2 \sum_k \alpha_k MA_1^{r_k} q_{1r_k} = A_2 \sum_k \alpha_k A_2^{r_k+1} q_{2r_k}.$$

Thus indeed

$$MA_1x = A_2Mx \quad \text{for all } x \in \mathscr{C}_1.$$

It follows from the uniqueness of solutions of differential equations that $MA_1 = A_2M$ if and only if

$$M\, e^{-A_1s} = e^{-A_2s}M \tag{28}$$

for all $s \geqq 0$. Hence in the controllability space of (S_1) conditions (17) and (16) are necessary and sufficient for (19) and hence for (18). This completes the proof.

Remark. If P is assumed compact and convex in R^n, then P is the convex hull of its extreme points. The following result now follows.

PROPOSITION 3. *For $i = 1, 2$, assume that $0 \in P_i$, and that the set Q_i of extreme points of P_i is countable. Let M be a matrix monomorphism*

$$R^{n_1} \to R^{n_2}.$$

Then,

$$M\mathscr{R}_1(t) = \mathscr{R}_2(t)$$

for each $t \geqq 0$ if and only if

$$MP_1 = P_2,$$

$$MA_1 = A_2M \quad \mathit{in}\ \mathscr{C}_1$$

(*in the sense that $MA_1x = A_2Mx$ for all $x \in \mathscr{C}_1$*).

When $n_1 = n_2$ in Proposition 3, then M is an invertible n-square matrix. Conditions

$$MA_1 = A_2M \quad \text{or} \quad MA_1M^{-1} = A_2 \quad \text{and} \quad MP_1 = P_2$$

imply that the systems (S_1) and (S_2) corresponding respectively to A_1, P_1 and A_2, P_2 are algebraically equivalent in the sense of Kalman [9]. This equivalence means that (S_1) and (S_2) are the same linear system only written relative to different bases for R^n.

Thus whenever $n_1 = n_2$, Proposition 3 states that

$$M\mathscr{R}_1(t) = \mathscr{R}_2(t), \qquad t \geqq 0,$$

if and only if the systems (S_1) and (S_2) are merely different realizations of the same linear system, i.e., describe the same linear system relative to different bases for R^n.

3. Symmetries of control systems. The following is the special case $(S_1) = (S_2)$ of Theorem 3.

THEOREM 4. *Suppose that the constraint set P of* (S) *is compact, convex, contains* 0, *and that its set of extreme points is countable. Let M define a nonsingular linear mapping of the state space into itself. Then*

$$M\mathscr{R}(t) = \mathscr{R}(t) \tag{29}$$

for each $t \geqq 0$ if and only if

$$MP = P, \tag{30}$$

$$MA = AM \quad in\, C, \tag{31}$$

in the sense that $MAx = AMx$ for all $x \in \mathscr{C}$.

In other terms, the symmetries of system (S) are precisely the symmetries of the reachable sets of (S) (in the sense of Definitions 2 and 3). It is clear from this that the symmetry problem for (S) is the same as that of $\mathscr{R}$(t). It is solved when the elements M are explicitly determined. This determination is sometimes facilitated by the following theorem.

THEOREM 5. *Let Q be the set of extreme points of P in the system* (S). *Then*

$$MQ = Q \quad and \quad MA = AM \tag{32}$$

whenever M is a symmetry of (S).

Proof. Let M be a symmetry of (S). Then from (2), $AM = MA$ and $MP = P$. Because M is linear and P compact, every extreme point of MP is the image of an extreme point of $P : Q \subseteq MQ$. Because M is nonsingular, Mq is an extreme point if $q \in Q : MQ \subseteq Q$. Hence $MQ = Q$. The proof is complete.

Examples.

$$\begin{aligned} \dot{x}_1 &= -2x_1 + u_1 + u_2, \\ \dot{x}_2 &= -x_2 + u_1 + 2u_2. \end{aligned} \tag{33}$$

Here

$$A = \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \qquad P = B\mathscr{U},$$

$$Q = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}.$$

The necessary conditions of Theorem 5,

$$AM = MA \quad \text{and} \quad B^{-1}MBQ = Q,$$

yield that the set of symmetries of (33) consists of I, $-I$ only.

$$\begin{aligned} \dot{x}_1 &= x_2 + u_1, \\ \dot{x}_2 &= -x_1 + u_2. \end{aligned} \tag{34}$$

Here

$$\begin{gathered} A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad P = B\mathscr{U}, \\ Q = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}. \end{gathered} \tag{35}$$

Theorem 5 yields

$$\left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \right\}$$

as the set of symmetries of (34).

The first result helps to remove a little obscurity in [8, p. 83 ff.] on the number and shape of the switching curves of the system (33). Indeed, the motivation for this paper comes from the attempt to gain some insight into the difficult problem of constructing the switching locus and the optimal feedback control for multidimensional control systems (S). The second is consistent with the investigation of Pontryagin [11, p. 42 ff.], where the synthesis problem of the time-optimal control system (33) is studied. It would be very interesting to prove Theorem 3 without the hypothesis of "a countable number of extreme points," but there does not seem to be an easy proof.

4. Identification of control systems. The problem studied here is how two systems with the same performance are related. It will be shown that under reasonable conditions such systems have coinciding constraint sets, solutions and optimal solutions.

THEOREM 6. *Given two systems* (S_1) *and* (S_2) *in n-space, suppose that the constraint set P_i of* (S_i) *is compact, convex, contains* 0, *and that its set of extreme points is countable. Then*

$$\mathscr{R}_1(t) = \mathscr{R}_2(t) \tag{36}$$

for each $t \geqq 0$ if and only if $\mathscr{C}_1 = \mathscr{C}_2$, and

$$P_1 = P_2, \tag{37}$$

$$A_1x = A_2x \quad \textit{for all } x \in \mathscr{C}_1 = \mathscr{C}_2. \tag{38}$$

Proof. This is immediate from Proposition 3 on noting that here $M = I$, the identity matrix.

There is an important special case of Theorem 6 dealing with the equation

$$(\mathrm{L}_i) \qquad \dot{x}_1 = A_i x_i + B_i u_i, \qquad i = 1, 2,$$

where $u_i : [0, t] \to \mathscr{U}_i$ is summable, with values in a fixed unit cube $\mathscr{U}_i$ in $R^m i$ and B_i a constant $n_i \times m_i$ matrix. Here

$$\mathscr{U}_i = \{u \in R^m i; |u_j| \leqq 1, j = 1, 2, \cdots, n_i\}$$

has a finite set Q_i of extreme points, namely, the points all of whose coordinates are ± 1.

COROLLARY 3. *Given systems* $(\mathrm{L}_1), (\mathrm{L}_2)$ *in n-space, where* B_i *is an* $n \times m_i$ *matrix such that* rank $B_i = m_i \leqq n$; *then*

$$\mathscr{R}_1(t) = \mathscr{R}_2(t)$$

for each $t \geqq 0$ *if and only if* $\mathscr{C}_1 = \mathscr{C}_2, m_1 = m_2$,

$$(39) \qquad B_1 = B_2 T$$

where T *is a permutation of the set of extreme points* (*in the sense that* $B_1 q_1 = B_2 T(q_1)$ *for each* $q_1 \in Q_1$), *and*

$$(40) \qquad A_1 x = A_2 x \quad \textit{for all } x \in \mathscr{C}_1 = \mathscr{C}_2.$$

Proof. First observe that (L_i) is the special case of (S_i), where $P_i = B_i U_i$; and P_i is compact, convex and symmetric. It also has a finite set of extreme points. Indeed, every extreme point of P_i is the image of an extreme point of $\mathscr{U}_i$, and the set of extreme points of $\mathscr{U}_i$ is finite. Theorem 6 applies:

$$\mathscr{R}_1(t) = \mathscr{R}_2(t)$$

for each $t \geqq 0$ if and only if

$$\mathscr{C}_1 = \mathscr{C}_2, \qquad B_1\mathscr{U}_1 = B_2\mathscr{U}_2$$

and

$$A_1 x = A_2 x \quad \text{for all } x \in \mathscr{C}_1 = \mathscr{C}_2.$$

Consider the condition $B_1\mathscr{U}_1 = B_2\mathscr{U}_2$. Since $\mathscr{U}_i$ has a nonvoid interior (in R^{m_i}) and B_i has rank m_i, the linear span of $B_i\mathscr{U}_i$ has dimension m_i. It now follows that $m_1 = m_2$, so that $Q_1 = Q_2$. Also $B_1 Q_1 = B_2 Q_2$ since rank $B_i = m_i \leqq n$. Hence for each $q_1 \in Q_1$ there is some $q_2 \in Q_2 = Q_1$ such that $B_1 q_1 = B_2 q_2$. Furthermore (again from the rank condition), q_2 is determined uniquely, and the q_2's obtained in this fashion on varying the q_1's exhaust Q_2. Thus there exists a one-to-one onto map $T: Q_1 \to Q_2$ such that $B_1 q_1 = B_2 T(q_1)$ for each $q_1 \in Q_1$, and since $Q_1 = Q_2$, T is a permutation of the extreme points.

We observe that if (L_1) has one-dimensional controls such that $|u| \leqq 1$, if $u \in \mathscr{U}$, then Q_1 has two elements ± 1. Thus T is either the identity map I, or $-I$. Hence $B_1 = \pm B_2$.

Remarks. Corollary 3 contains the special result of Hájek [4, Thm. 1] for the one-dimensional control system

$$\dot{x}(t) = Ax + bu, \qquad |u| \leqq 1.$$

However, his method differs from that of this investigation. He studies the particular shapes of the reachable sets by using the boundary points of these sets, a study that rests on the construction of terminal manifolds. Such a construction for the multidimensional system studied here is not only very difficult, but also, to the author's knowledge, an open problem.

Corollary 3 is also analogous to the result of Hautus and Olsder for the system

$$\text{(L)} \qquad \dot{x} = Ax(t) + Bu(t),$$

where $u(t)$ is constrained to lie in a set Ω_p defined as the set of all measurable vector functions $v(t)$ with $\|v\|_p \leqq 1$, $1 \leqq p \leqq \infty$, and $p \neq 2$, where $\|\cdot\|_p$ denotes the l_p-norm.

5. Further developments. The study of symmetries and identification of linear control systems whose foundation is the basis of this paper and the basis of [4] and [6] is important in its own right. Also it promises some insight into the problem of constructing the switching locus and optimal feedback control for multidimensional control systems. In this section some brief comments are made on several unresolved problems connected with our research.

I. In our investigation, symmetries are defined by nonsingular matrices. Also, the proof of the main theorem depends heavily on the monomorphism of M. For example, the equation

$$(23) \qquad M e^{-A_1 t} Q_1 = e^{-A_2 t} Q_2$$

required this strong condition on M. Can this be relaxed? Of course (23) no longer holds when M is not monic. What does hold is the relation

$$M e^{-A_1 t} Q_1 \subseteq e^{-A_2 t} Q_2 .$$

Thus when M is arbitrary, subsequent arguments in this paper no longer hold. Though it is plausible that the theorem is valid, there is no obvious proof.

II. Suppose M is rectangular. If $x_1 \in R^{n_1}$, $x_2 \in R^{n_2}$ and $x_2 = Mx_1$, then M defines a linear subspace

$$L = \{(x_1, x_2) : x_2 = Mx_1\}$$

of the product space $R^{n_1} \times R^{n_2}$.

Now let L be a linear subspace of the product space $R^{n_1} \times R^{n_2}$. Let $M_1 \subseteq R^{n_1}$, $M_2 \subseteq R^{n_2}$ and $L \subseteq R^{n_1} \times R^{n_2}$.

DEFINITION 6. M_1, M_2 are said to be in relation L if and only if for each $x_1 \in M_1$ there exists $x_2 \in M_2$ such that $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in L$, and for each $x_2 \in M_2$ there exists some $x_1 \in M_1$ such that $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in L$.

In the usual notation of relations the conditions on M_1, M_2 are that $M_1 \subseteq L[M_2]$, $M_2 \subseteq L^{-1}[M_1]$. If L is a linear subspace, there must exist matrices

U_1, U_2 such that $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in L$ if and only if

$$U_1 x_1 = U_2 x_2$$

(and conversely, given U_1 and U_2 this condition defines L). The problem of interest here is as follows. Suppose $\mathscr{R}_1(t), \mathscr{R}_2(t)$ are in relation L for each $t \geqq 0$. What is the appropriate conclusion? This should specialize to Theorem 3 in case $U_2 = I$ and U_1 defines a nonsingular linear map $x_1 \to x_2$. The symmetry problem will then extend to a "generalized" symmetry problem where the generalized symmetry is defined by the linear subspace L.

Our investigation would yield appropriate conclusions whenever there are arbitrary matrices U_1, U_2 such that

$$U_1 \mathscr{R}_1(t) = U_2 \mathscr{R}_2(t)$$

for each $t \geqq 0$. The special case $U_2 = 0$ of this situation is obviously connected with some extensions of the problem of controllability and observability of the system (S).

III. An obvious extension of the above research would be to find immediate parallels of our string of results in the case of nonautonomous control systems

$$(\mathrm{N}_i) \qquad \dot{x}_i(t) \in A_i(t)x_i - Q_i(t),$$

where $Q_i : [0, t] \to \Omega^{n_i}$ is a measurable set-valued function with values in Ω^{n_i}, the space of compact subsets of R^{n_i} with the Hausdorff topology, and A_i has square summable components.

Active research is being conducted in all these three areas of investigation.

Acknowledgment. This work is part of the author's doctoral dissertation written at Case Western Reserve University, Cleveland, Ohio. The author is very grateful to Professor Otomar Hájek for his unfailing inspiration and patient guidance of this research.

The author is indebted to the referee for suggesting the more direct proof of the main theorem which gives a more general result than the original version; and for pointing out the relevance of Proposition 3 to the work of Kalman [9].

REFERENCES

[1] H. G. Eggleston, *Convexity*, Cambridge Tracts in Mathematics, No. 47, Cambridge University Press, New York, 1958.

[2] F. R. Gantmacher, *The Theory of Matrices*, vol. 2, Chelsea, New York, 1959.

[3] O. Hájek, *Geometric theory of time-optimal control*, this Journal, 9 (1971), pp. 339–350.

[4] ———, *Identification of control systems by performance*, Math. Systems Theory, 5 (1971), pp. 349–352.

[5] ———, *Mathematical Control Theory Seminar Lecture Notes*, Case Western Reserve University, Cleveland, Ohio, 1972.

[6] M. L. J. Hautus and G. J. Olsder, *Uniqueness theorem for linear control systems with coinciding reachable sets*, to appear.

[7] H. Hermes, *Calculus of set valued functions and control*, J. Math. Mech., 18 (1968), pp. 47–59.

[8] H. Hermes and J. P. LaSalle, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[9] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

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

[11] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Wiley-Interscience, New York, 1962.

[12] J. STOER AND C. WITZGALL, *Convexity and Optimization in Finite Dimensions*, vol. I, Springer-Verlag, New York, 1970.

[13] H. T. BANKS AND M. Q. JACOBS, *A differential calculus for multi-functions*, J. Math. Anal. Appl., 29 (1970), pp. 246–271.

SIAM J. CONTROL
Vol. 12, No. 3, August 1974

GEOMETRICALLY CONSTRAINED OBSERVABILITY*

ROBERT F. BRAMMER†

Abstract. This paper deals with observed processes in situations in which observations are available only when the state vector lies in certain regions. For linear autonomous observed processes, necessary and sufficient conditions are obtained for half-space observation regions. These results are shown to contain a theorem dual to a controllability result proved by the author for a linear autonomous control system whose control restraint set does not contain the origin as an interior point. Observability results relating to continuous observation systems and sampled data systems are presented, and an example of observing the state of an electrical network is given.

Recently, the problem of controllability in linear autonomous systems has been considered without the assumption that the origin is interior to the range of the set of admissible controllers. In [1] the following theorem is proved.

THEOREM A. *Consider the control system in R^n:*

$$\dot{x} = Ax + Bu, \qquad u \in \Omega \subset R^m, \tag{L}$$

which has the properties:

(a) *The set Ω contains a vector in the kernel of B (i.e., there exists $u \in \Omega$ satisfying $Bu = 0$).*

(b) *The convex hull of Ω, $\mathrm{CH}(\Omega)$, has nonempty interior in R^m.*

The following conditions are necessary and sufficient for the null controllability of (L):

(c) *The controllability matrix $C[A, B]$ has rank n. (The matrix $C[A, B]$ is defined by the columns of $(B, AB, \cdots, A^{n-1}B)$.)*

(d) *There is no eigenvector, v, of A^T satisfying $(v, Bu) \leqq 0$ for all $u \in \Omega$.*

It is the purpose of this paper to obtain observability results dual to the above controllability theorem. This leads to the notion of geometrically constrained observations. An observation process is geometrically constrained if the observations of the state are available only when the state vector lies in certain regions. This situation will be shown to be dual to that of a control system whose control restraint set does not contain the origin as an interior point.

The process

$$\dot{x} = f(x, u), \qquad y = h(x), \tag{1}$$

where $f: R^n \times R^m \to R^n, u: R^1_+ \to \Omega \subset R^m$, and $h: R^n \to R^p$ is called an *observed process* in R^n. We assume that the state dynamics are described by the control system $\dot{x} = f(x, u)$, that the control function $u(\cdot)$ is known, but that the only information available concerning the state vector is the observation vector, $y = h(x)$. We shall restrict our consideration to functions f and h that are continuously differentiable and to measurable control functions $u(\cdot)$, and we shall assume that solutions to (1), $x(\cdot, x_0, u(\cdot))$, exist and are unique for all x_0, $u(\cdot)$,

* Received by the editors May 11, 1972, and in revised form January 31, 1973.

† Goddard Space Flight Center, Greenbelt, Maryland 20771. This research was supported in part by the National Science Foundation under Grant GP-31386X.

and $t \geqq 0$. We shall assume $0 \in \Omega$ and that $f(0,0) = 0$ and $h(0) = 0$. Thus, each initial condition, x_0, and controller, $u(\cdot)$, leads to a unique solution, $x(\cdot, x_0, u(\cdot))$, and, consequently, an observation history, $y(\cdot, x_0, u(\cdot)) = h(x(\cdot, x_0, u(\cdot)))$. The process (1) is *observable* if there exists a finite interval $[0, T]$ such that for any x_1 and x_2 if $y(t, x_1, y(\cdot)) = y(t, x_2, u(\cdot))$ for some controller $u(\cdot)$ and for all $t \in [0, T]$, then $x_1 = x_2$. Thus, if a process is observable, the initial condition, and hence the entire solution is uniquely defined by the observations over some finite interval. The process (1) is also applicable to situations in which there is no control. In such a case, let $\Omega = \{0\} \in R^m$, and define $f(x, 0) = f(x)$.

For linear systems we have the following well-known theorem [4, p. 111].

THEOREM 1. *The autonomous linear observed process*

$$\dot{x} = Ax + Bu, \quad y = Hx, \qquad u \in \Omega \subset R^m, \tag{2}$$

is observable if and only if the dual control system

$$\dot{x} = A^T x + H^T u \tag{3}$$

is null-controllable.

Remark 2. The proof of this theorem assumes that the origin is interior to Ω, and that the observations are available for all values of the state vector.

In many physical systems observations are available only when the state vector is in certain regions of the state space. For example, a radar can measure the position of a space vehicle only when the vehicle is above the local horizon at the radar. A simple example involving an autonomous linear process is given by the electrical circuit in Fig. 1. The state of the system is defined by the current, x_1, in the loop, and the voltage, x_2, across the capacitor. The measurement available concerning the state is the voltage, y, across the output terminal. Due to the presence of the diode, D, the output will be zero if x_2 is negative and will be x_2 if x_2 is positive. Thus, we have the following system:

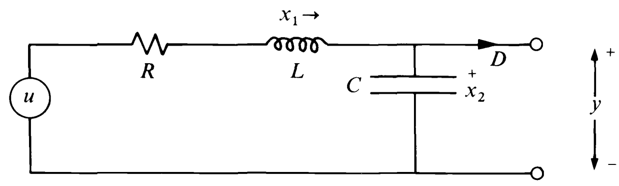

FIG. 1

$$\dot{x} = Ax + Bu, \quad y = Hx \quad \text{when } Dx > 0, \qquad u \in [-1, 1],$$

where

$$A = \begin{pmatrix} -R/L & -1/L \\ 1/C & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad H = D = (0, 1).$$

Assume that the matrix A is oscillatory (i.e., $R < 2\sqrt{L/C}$). There are two important features to be noted in this example. First, for certain initial conditions and controllers observations will never be available. Let $x_1(0) = 0, |x_2(0) + \varepsilon| < \varepsilon/2$. Choose $u(\cdot) = -\varepsilon$. It is easy to see that the capacitor will charge and $x_2(t)$ will oscillate about the value $-\varepsilon$ with amplitude less than $\varepsilon/2$ and, consequently, never become positive. Thus, this process is not observable. Second, there is a control, the zero control, for which the voltage, x_2, becomes positive over some subinterval of $[0, T]$, where

$$T = 2\pi/\sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}}$$

for any initial condition $(x_1(0), x_2(0))$. Furthermore, at least one such subinterval has length greater than or equal to $T/4$.

Using this example as a guide, we define an *observed process with constrained observations* by adding to (1) a *constraint* function $d: R^n \to R^l$ so that the observation vector $y = h(x)$ is available if and only if at least one component of the l-vector $d(x)$ is positive. The function d is continuously differentiable and $d(0) = 0$. Let L denote the positive unit l-cube in R^l (i.e., $L = \prod_1^l [0, 1]$). At least one component of the vector $d(x)$ is positive if and only if there exists some $w \in L$ satisfying $(w, d(x)) > 0$. Thus, we have the process

$$(4) \qquad \dot{x} = f(x, u), \quad y = h(x) \quad \text{with constraint function } d.$$

The observed process with constrained observations is *observable* if the following conditions are satisfied:

(5) There exists a controller, $u_0(\cdot)$, and positive numbers, γ and T, such that for all $x_0 \in R^n$ the observations $y(\cdot, x_0, u_0(\cdot))$ are available on some open subinterval of $[0, T]$ having length greater than or equal to γ.

(6) If $y(t, x_1, u(\cdot)) = y(t, x_2, u(\cdot))$ for all $t \in [0, T]$ and for any admissible $u(\cdot)$, and if observations are available for an interval of length γ, then $x_1 = x_2$.

The process is observable if there is some controller for which the availability of observations is assured, and if the observations of a solution, when available, are sufficient to distinguish solutions of (4).

For the autonomous linear problem we have the process

$$(7) \qquad \dot{x} = Ax + Bu, \quad y = Hx \quad \text{with constraint matrix } D.$$

In this process A, B, H, and D are real constant $n \times n$, $n \times m$, $p \times n$, and $l \times n$ matrices respectively. Let $d_1, \cdots, d_l$ be the row n-vectors of the matrix D. The condition on the availability of the observations is that at least one of the scalar products $(d_1, x), \cdots, (d_l, x)$ is positive. Geometrically this means that the state vector is in at least one of the half-spaces for which some d_j is an inward normal at the origin. This process (7) includes the electrical example considered earlier. Before proving an observability result which will demonstrate that the electrical circuit example defines an observable process, we require two preliminary lemmas.

LEMMA 3. *Let A and H be as in* (7), *and assume that*

$$(8) \qquad C(A^T, H^T) \quad \textit{has rank } n.$$

Let $\delta > 0$, $t \in [0, T]$, and define the matrix function $M_\delta(t)$ as follows:

$$M_\delta(t) = \int_t^{t+\delta} e^{A^T s} H^T H \, e^{As} \, ds.$$

Then there exists $\alpha(\delta) > 0$ satisfying $(x, M_\delta(t)x) \geqq \alpha(\delta)(x, x)$.

Proof. It is easy to see that if $C(A^T, H^T)$ has rank n, then $M_\delta(t)$ is a positive definite symmetric matrix [2, p. 29]. Hence, the scalar product $(x, M_\delta(t)x)$ has a positive lower bound, $\alpha(\delta, t)$, as x varies over the unit sphere, S^{n-1}. Thus,

$$(x, M_\delta(t)x) \geqq \alpha(\delta, t)(x, x)$$

for all $x \in R^n$. Since $M_\delta(t)$ is continuous in t and the scalar product is continuous in x and t, $\alpha(\delta, t)$ is continuous in t. Thus, for $t \in [0, T]$, $\alpha(\delta, t) \geqq \alpha(\delta) > 0$, where $\alpha(\delta) = \inf_{t\in[0,T]} \alpha(\delta, t)$.

LEMMA 4. *Let $S_+^{l-1} = L \cap S^{l-1}$, and let A and D be as in* (7). *Define the function $\beta(T, x, w): R^1 \times S^{n-1} \times S_+^{l-1} \to R^1$ to be the following*:

$$\beta(T, x, w) = \sup_{t\in[0,T]} (w, D\, e^{At}x).$$

Let $\sigma(T)$ be the infimum of $\beta(T, x, w)$, where the infimum is taken over all x and w in the domain of β. Then there exists $T > 0$ satisfying $\sigma(T) > 0$ if and only if the following two conditions are satisfied:

(9) *$C(A^T, D^T)$ has rank n.*

(10) *There is no eigenvector, v, of A satisfying $(v, D^T w) \leqq 0$ for all $w \in L$.*

Proof. Since $(w, D\, e^{At}x) = (x, e^{A^Tt}D^Tw)$, it follows from the analysis of Lemma 2.4 and Theorem 1.4 of [1] that (9) and (10) are necessary and sufficient for the scalar product $(x, e^{A^Tt}D^Tw)$ to attain a positive value at some time $t(x, w) > 0$ for each fixed x and w in the domain of the function β. From the compactness of the domains of x and w it follows that there exists a finite number, T, for which $\beta(T, x, w) > 0$. By the continuity of $\beta(T, \cdot, \cdot)$ and the compactness of the domains, it also follows that $\sigma(T) > 0$.

Remark 5. Furthermore, if $0 < \lambda < 1$, for each x and w there is an open subinterval of $[0, T]$ during which $(w, D\, e^{At}x) > \lambda\sigma(T)$. Let $(t_1(x, w), t_2(x, w))$ denote an interval of maximal length during which this inequality is satisfied. Again, by the compactness of the domains of x and w, there is a positive lower bound, $\gamma(\lambda)$, to the length of these maximal intervals.

Using these lemmas, we can establish the following observability result which establishes necessary and sufficient conditions for the observability of the process (7). This theorem will prove that the electrical example is observable, and the theorem will be specialized to a dual theorem to Theorem A in the same sense that Theorem 1 is dual to Theorem 1.1 of [1].

THEOREM 6. *Conditions* (8), (9) *and* (10) *are necessary and sufficient for the observability of the linear autonomous process* (7).

Proof. Lemma 4 shows that (9) and (10) are necessary and sufficient to ensure that observations of any solution of (7) with $u_0 \equiv 0$ will be available at least during some subinterval of $[0, T]$ having length γ, where T and γ are determined as in Lemma 4. If (8) is not satisfied, there is a nonzero vector v satisfying $H\, e^{At}v = 0$. Thus, the solution $x(\cdot, v, 0)$ cannot be distinguished from $x(\cdot, 0, 0)$ even though observations are available. Consequently, the process is not observable. If (8) is satisfied, then since $y(t, x_0, u(\cdot))$ is given by the expression

$$y(t, x_0, u(\cdot)) = H\, e^{At}x_0 + \int_0^t H\, e^{A(t-s)}Bu(s)\, ds,$$

we have

$$y(t, x_1, u(\cdot)) - y(t, x_2, u(\cdot)) = H\, e^{At}(x_1 - x_2).$$

Thus if $y(t, x_1, u(\cdot)) = y(t, x_2, u(\cdot))$ over any open interval, $(t, t + \gamma)$, we have

$$0 = \int_t^{t+\gamma} |y(s, x_1, u(\cdot)) - y(s, x_2, u(\cdot))|^2\, ds = (x_3, M_\gamma(t)x_3),$$

where $x_3 = x_1 - x_2$. By Lemma 3, $x_3 = 0$. Therefore, $x_1 = x_2$ and the system is observable. The following corollary is the observability dual to Theorem A.

COROLLARY 7. *In the process* (7), *let* $l = p$ *and* $D = H$. *The process* (7) *is observable if and only if the dual control process*

$$\dot{x} = A^T x + H^T u, \qquad u \in L \subset R^l, \tag{11}$$

is null-controllable.

Remark 8. Setting $D = H$ means that observations are available if and only if *one* of the observation vector components is positive. This is dual to the control situation in which *all* the control components are positive since $u \in L$.

In the electrical example if $R < 2\sqrt{L/C}$, it is easy to verify that the conditions of Theorem 6 are satisfied. Consequently, the process is observable.

This analysis can be extended to produce a local observability result in the neighborhood of the origin for the process (4). First, we define as in [4, p. 378] local observability for the process (1). The process (1) is said to be *locally observable near the origin* if there exists $\varepsilon > 0$ such that for $|u(t)| < \varepsilon, |x_1| < \varepsilon, |x_2| < \varepsilon$, the process is observable according to the definition for the process (1). The norm symbol $|\cdot|$ denotes the usual Euclidean norm in a finite-dimensional Euclidean space and the norm symbol $\|\cdot\|$ denotes the L^2-norm for functions. The following result appears in [4, p. 378].

THEOREM 9. *For the process* (1), *let* $A = f_x(0, 0)$ *and* $H = h_x(0)$. *If* $C(A^T, H^T)$ *has rank n, then* (1) *is locally observable near the origin.*

We define local observability for observed processes with constrained observations as follows: the process (4) is locally observable near the origin if there exists $\varepsilon > 0$ such that (5) and (6) are satisfied for controllers on $[0, T]$ satisfying $\|u(\cdot)\| < \varepsilon$ and initial conditions x_0 satisfying $|x_0| < \varepsilon$. The following theorem gives a local observability result for process (4).

THEOREM 10. *For the process* (4), *let* $A = f_x(0, 0), H = h_x(0)$, *and* $D = d_x(0)$. *The process* (4) *is locally observable near the origin if conditions* (8), (9) *and* (10) *are satisfied.*

Proof. To show that (5) is satisfied, we will produce $T > 0$ and $\varepsilon_1 > 0$ such that if $|x_0| < \varepsilon_1$ then observations of any nonzero solution $x(t, x_0, 0)$ are available on some subinterval of $[0, T]$. To show that (6) is satisfied, we will produce $\varepsilon_2 > 0$ such that if $|x_1| < \varepsilon_2$, $|x_2| < \varepsilon_2, |u(t)| < \varepsilon_2$ for $t \in [0, T]$, and if $y(t, x_1, u(\cdot)) = y(t, x_2, u(\cdot))$ over any open subinterval of $[0, T]$, then $x_1 = x_2$. Therefore, by taking $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$, we will prove the local observability near the origin of the process (4) by showing that observations are both available and equal.

By hypothesis, $f(x, u) = Ax + Bu + g_1(x, u), d(x) = Dx + g_2(x), h(x) = Hx + g_3(x)$ and for every $\varepsilon > 0$ there exists $\delta(\varepsilon) > 0$ such that $|x| < \delta(\varepsilon)$ and $|u| < \delta(\varepsilon)$ implies $|g_1(x, u)| < \varepsilon(|x| + |u|), |g_2(x)| < \varepsilon|x|, |g_3(x)| < \varepsilon|x|$. Since f is a C^1-function, the solution $x(t, x_0, u(\cdot))$ is differentiable with respect to x_0, and since $x(t, 0, 0) \equiv 0$, it is easy to see that on any finite interval, $[0, T]$, there exist $\delta_1(T) > 0$ and $M(T) > 0$ such that if $|x_0| + \|u(\cdot)\| < \delta_1(T)$, then $|x(t, x_0, u(\cdot))| < M(T)(|x_0| + \|u(\cdot)\|)$.

The solution $x(t, x_0, 0)$ has the form

$$x(t, x_0, 0) = e^{At}x_0 + G(t, x_0, 0),$$

where $G(t, x_0, u(\cdot))$ is given by

$$G(t, x_0, u(\cdot)) = \int_0^t e^{A(t-s)}g_1(x(s, x_0, u(\cdot), u(s)))\, ds.$$

Define $N(T) = \sup_{[0,T]} |e^{At}|$. Let $w \in L$, and we have

$$\begin{aligned}(w, d(x(t, x_0, 0))) &= (w, Dx(t, x_0, 0)) + (w, g_2(x(t, x_0, 0)))\\ &= (w, D\, e^{At}x_0) + (w, DG(t, x_0, 0)) + (w, g_2(x(t, x_0, 0))).\end{aligned}$$

Apply Lemma 4 to A and D to get $T > 0$ satisfying $\sigma(T) > 0$. Using this value for T, we obtain $\delta_1(T)$ and $M(T)$. Thus, choosing $\varepsilon > 0$ to satisfy

$$\varepsilon < \min\left(\frac{\sigma(T)}{3|D|N(T)M(T)T}, \frac{\sigma(T)}{3M(T)}\right),$$

we have, for x_0 satisfying $|x_0| \leqq \min(\delta_1(T), \delta(\varepsilon)) = \delta$,

$$\begin{aligned}|(w, DG(t, x_0))| &\leqq \int_0^T |w|\,|D|\,|e^{A(t-s)}|\,|g_1(x(s, x_0, 0), 0)|\, ds\\ &\leqq |w|\,|D|N(T)M(T)T\varepsilon|x_0|\\ &< \frac{\sigma(T)}{3}|w|\,|x_0|,\end{aligned}$$

and

$$|(w, g_2(x(t, x_0, 0)))| < \frac{\sigma(T)}{3}|w|\,|x_0|.$$

Therefore, if $(w, d(x(t, x_0, 0))) \leqq 0$ for all $t \in [0, T]$, then

$$0 \geqq (w, d(x(t, x_0))) \geqq (w, D\, e^{At}x_0) - \frac{2\sigma(T)}{3}|w|\,|x_0|.$$

If $|x_0|\,|w| \neq 0$, it follows from Remark 5 with $\lambda = 2/3$ that the above inequality cannot be satisfied. Thus, if $|x_0|$ is sufficiently small, and if no observations are available on $[0, T]$, then $x_0 = 0$. Otherwise, observations must become available on some subinterval of $[0, T]$.

By the inequality

$$(w, d, (x(t, x_0, 0))) \geqq |w|\,|x_0|\left(\left(\frac{w}{|w|}, D\, e^{At}\frac{x_0}{|x_0|}\right) - \frac{2\sigma(T)}{3}\right),$$

it follows from Lemma 4 and Remark 5 that observations of each nonzero solution are available for an interval of length at least γ. Thus, we have satisfied (5).

The solution $x(t, x_0, u(\cdot))$ has the form

$$x(t, x_0, u(\cdot)) = e^{At}x_0 + \int_0^t e^{A(t-s)}Bu(s)\,ds + G(t, x_0, u(\cdot)).$$

Assume that $y(t, x_1, u(\cdot)) = y(t, x_2, u(\cdot))$ over some subinterval of $[0, T]$, $(t_1, t_1 + \gamma)$. Therefore,

$$\int_{t_1}^{t_1+\gamma} |y(t, x_1, u(\cdot)) - y(t, x_2, u(\cdot))|^2\,dt = 0.$$

The observations are given by

$$y(t, x_0, u(\cdot)) = H\,e^{At}x_0 + \int_0^t H\,e^{A(t-s)}Bu(s)\,ds + HG(t, x_0, u(\cdot))$$
$$+ g_3(x(t, x_0, u(\cdot))).$$

Therefore,

$$0 = \int_{t_1}^{t_1+\gamma} |y(t, x_1, u(\cdot)) - y(t, x_2, u(\cdot))|^2\,dt$$
$$= \int_{t_1}^{t_1+\gamma} |H\,e^{At}(x_1 - x_2) + H(G(t, x_1, u(\cdot)) - G(t, x_2, u(\cdot)))$$
$$+ g_3(x(t, x_1, u(\cdot))) - g_3(x(t, x_2, u(\cdot)))|^2\,dt.$$

Let $x_3 = x_1 - x_2$, and we have

$$0 = (x_3, M_\gamma(t_1)x_3) + \int_{t_1}^{t_1+\gamma} [2(H\,e^{At}x_3, G_{12}(t, x_1, x_2, u(\cdot)))$$
$$+ 2(H\,e^{At}x_3, g_{12}(t, x_1, x_2, u(\cdot)))$$
$$+ 2(G_{12}(t, x_1, x_2, u(\cdot)), g_{12}(t, x_1, x_2, u(\cdot)))$$
$$+ |G_{12}(t, x_1, x_2, u(\cdot))|^2 + |g_{12}(t, x_1, x_2, u(\cdot))|^2]\,dt,$$

where $G_{12}(t, x_1, x_2, u(\cdot)) = H(G(t, x_1, u(\cdot)) - G(t, x_2, u(\cdot)))$, and $g_{12}(t, x_1, x_2, u(\cdot)) = g_3(x(t, x_1, u(\cdot))) - g_3(x(t, x_2, u(\cdot)))$. Denote the above integral by I. We have

$$|G_{12}(t, x_1, x_2, u(\cdot))| \leqq |H|[N(T)M(T)\varepsilon(|x_1| - |x_2|)]$$
$$\leqq |H|N(T)M(T)\varepsilon|x_3|,$$

and

$$|g_{12}(t, x_1, x_2, u(\cdot))| < \varepsilon M(T)|x_3|$$

for $|x_1| + \|u(\cdot)\| < \delta(\varepsilon)$ and $|x_2| + \|u(\cdot)\| < \delta(\varepsilon)$. Thus, we have the result that the integral, I, satisfies the inequality

$$|I| < \varepsilon P(T)|x_3|^2,$$

where $P(T)$ is a positive number given by

$$P(T) = \gamma[2|H|^2N^2(T)M(T) + 2|H|N(T)M(T)$$
$$+ 2|H|N(T)M^2(T)\varepsilon + \varepsilon|H|^2N^2(T)M^2(T) + \varepsilon M^2(T)].$$

Therefore,

$$0 = (x_3, M_\gamma(t)x_3) + I \geqq (\alpha(\gamma) - \varepsilon P(T))|x_3|^2.$$

By choosing ε sufficiently small, we conclude that $x_3 = 0$ or $x_1 = x_2$. Thus we have satisfied (6), and the process (4) is locally observable near the origin.

Remark 11. The techniques used to prove Theorem 6 and Theorem 10 will suffice to prove Theorem 1 and Theorem 9 (situations without constrained observations). The proof of Theorem 10 given here differs from the proof of Theorem 9 given in [4, p. 378] by being a finite-dimensional proof and making no use of an implicit function theorem in a Banach space.

The case of observations made at discrete times is considered in [4, p. 381]. Here we consider this problem for constrained observations and present a new proof which includes the result in [4].

For the case of observations available only at discrete times (sampled data systems), we require first the following lemma.

LEMMA 12. *Let $M(t)$ be a real analytic $p \times n$ matrix function defined on R^1. Assume that for any a and b, $\int_a^b M^T(t)M(t)\,dt$ is nonsingular. Then, except for a closed set of measure zero in R^n, the function*

$$\det\left(\sum_1^n M^T(t_i)M(t_i)\right) \neq 0.$$

Proof. Let (a, b) be any open interval. Let S be the linear subspace generated by vectors of $M(t)$ for all $t \in (a, b)$. If $S \neq R^n$, then there exists $x \neq 0 \in R^n$ satisfying $M(t)x = 0$ for all $t \in (a, b)$. Therefore,

$$\int_a^b (M(t)x, M(t)x)\,dt = 0 = \int_a^b (x, M^T(t)M(t)x)\,dt = \left(x, \int_a^b M^T(t)M(t)\,dtx\right)$$

which is impossible since $\int_a^b M^T(t)M(t)\,dt$ is nonsingular. (For a matrix of this form, nonsingular implies positive definite symmetric.) Thus, there exist $t_1, \cdots, t_n$ in (a, b) so that the matrix $M(t_1, \cdots, t_n)$ defined by

$$M(t_1, \cdots, t_n) = \begin{bmatrix} M(t_1) \\ \vdots \\ M(t_n) \end{bmatrix}$$

has rank n. Therefore, $M^T(t_1, \cdots, t_n)M(t_1, \cdots, t_n)$ has rank n, and since

$$M^T(t_1, \cdots, t_n)M(t_1, \cdots, t_n) = \sum_1^n M^T(t_i)M(t_i),$$

$\det(\sum_1^n M^T(t_i)M(t_i)) > 0$. Thus, $\det(\sum_1^n M^T(t_i)M(t_i))$ is a real analytic entire function which is not identically zero. The set on which a nonconstant real analytic function vanishes is called an analytic variety and is a countable union of analytic manifolds of dimension less than or equal to $n - 1$, each of which has measure zero in R^n. See, for example, [5, p. 361] and [3, p. 205].

For the autonomous linear process (7) the observations have the form

$$y(t, x_0, u(\cdot)) = H\, e^{At}x_0 + H \int_0^t e^{A(t-s)}Bu(s)\, ds.$$

Assume that (7) satisfies conditions (8), (9) and (10). If $u(\cdot) = 0$, then the analysis of Theorem 6 shows that there is a positive number γ and a finite interval $[0, T]$ such that for any $x_0 \in R^n$ there is an open subinterval, $J(x_0)$, of $[0, T]$ of length greater than or equal to γ during which the observations $y(t, x_0, 0)$ are available. Assume that the sampling occurs with a rate of at least $n + 1$ samples per γ time units. Thus, in any open interval of length at least γ we have at least n samples, and, consequently, we will have at least n measurements of each solution. Lemma 3 and Lemma 12 imply that for all but an exceptional set (an exceptional set is a closed set of measure zero) of times the matrix $\sum_1^n e^{A^Tt_i}H^TH\, e^{At_i}$ is invertible. We have

$$\sum_1^n e^{A^Tt_i}H^Ty(t_i, x_0, 0) = \left(\sum_1^n e^{A^Tt_i}H^TH\, e^{At_i}\right)x_0,$$

and therefore for all but an exceptional set of measurement sequences, the uncontrolled solution is observable. This satisfies condition (5). Similarly, if $y(t_i, x_1, u(\cdot)) = y(t_i, x_2, u(\cdot))$ for n samples and $\sum_1^n e^{A^Tt_i}H^TH\, e^{At_i}$ is invertible, then $x_1 = x_2$, thus satisfying condition (6). Furthermore, it is clear from the proof of Theorem 6 that (8), (9), and (10) are necessary for observability. Thus, we have proved the following theorem.

THEOREM 13. *Conditions* (8), (9) *and* (10) *are sufficient to ensure that there exists a finite interval* $[0, T]$ *and a sampling rate, S, such that for all but an exceptional set of measurement sequences of rate at least S the process* (7) *is observable. Furthermore, if* (8), (9) *and* (10) *are not all satisfied, then* (7) *is not observable for any measurement sequence.*

The local linearization techniques of Theorem 10 can be adapted to the problem of observability using discrete-time measurements in the following way. Using the notation of Theorem 10 and assuming that conditions (8), (9) and (10) are satisfied, we note that the existence of the positive number γ required by (5) is guaranteed by the results of Theorem 10. Let $\{t_1, \cdots, t_q\}$ be any measurement sequence with rate at least $(n + 1)/\gamma$. For all but an exceptional set of measurement sequences, the matrix $\sum_{j+1}^{j+n} e^{A^Tt_i}H^TH\, e^{At_i}$ is nonsingular where $0 \leqq j \leqq q - n$. Thus, for any open subinterval of length γ we have at least n measurements and $\sum_{j+1}^{j+n} e^{A^Tt_i}H^TH\, e^{At_i}$ is nonsingular. Since we have

$$y(t_i, x_0, u(\cdot)) = H\, e^{At_i}x_0 + \int_0^{t_i} H\, e^{A(t_i-s)}Bu(s)\, ds + HG(t_i, x_0, u(\cdot)) + g_3(x(t_i, x_0, u(\cdot))),$$

then, if the measurements are available at $t_{j+1}, \cdots, t_{j+n}$,

$$x_0 = \left(\sum_{j+1}^{j+n} e^{A^Tt_i}H^TH\, e^{At_i}\right)^{-1} \cdot \left[\sum_{j+1}^{n} e^{A^Tt_i}H^Ty(t_i, x_0, u(\cdot)) + \sum_{j+1}^{j+n} e^{A^Tt_i}H^TH \int_0^{t_i} e^{A(t_i-s)}Bu(s)\, ds + o(|x_0| + \|u\|)\right].$$

It follows from the implicit function theorem that for $|x_0| + \|u(\cdot)\|$ sufficiently small that x_0 is uniquely determined from the observations. (Recall that the vectors $y(t_i, x_0, u(\cdot))$ are known.) By taking the smallest radius for $|x_0| + \|u\|$ from among the $q - n$ consecutive subsets of $\{t_1, \cdots, t_q\}$, we obtain the following theorem.

THEOREM 14. *Conditions* (8), (9) *and* (10) *are sufficient to ensure that there exist a finite interval and a sampling rate S such that for all but an exceptional set of measurement sequences of rate at least S, taken from this interval, the process* (4) *is locally observable.*

Finally, we wish to consider the possibility of decreasing the measurement rate while retaining observability. It is clear that n measurements during any subinterval may be more than necessary to determine the initial state vector. For example, if H is the $n \times n$ identity matrix, then only one measurement is needed. In the general case let $M(t) = H\,e^{At}$. From the proof of Lemma 12 and the subsequent proofs of Theorems 13 and 14 it is easily seen that if there are k values of t satisfying rank $[M^T(t_1), \cdots, M^T(t_k)] = n$, then $\sum_1^k e^{A^T t_i} H^T H\, e^{At_i}$ is nonzero almost everywhere and a rate of $(k + 1)/\gamma$ will suffice in the above theorems. The following theorem identifies the smallest possible k.

THEOREM 15. *Let k be the smallest integer such that* rank $[H^T, A^T H^T, \cdots, (A^T)^{k-1} H^T] = n$. *If* (7) *is observable, a measurement rate of* $(k + 1)/\gamma$ *will be sufficient for observability. Furthermore, no smaller value of k is sufficient.*

Proof. Let k be defined as above, and let $\varepsilon_1, \cdots, \varepsilon_k$ be small unequal positive numbers. The matrix

$$\begin{aligned}&[e^{A^T\varepsilon_1}H^T, \cdots, e^{A^T\varepsilon_k}H^T]\\ &\qquad = [H^T, \cdots, (A^T)^{k-1}H^T]\end{aligned}\begin{bmatrix} I_{p\times p} & & I_{p\times p} \\ \varepsilon_1 I_{p\times p} & \cdots & \varepsilon_k I_{p\times p} \\ \vdots & & \vdots \\ \dfrac{\varepsilon_1^{k-1}}{(k-1)!} + o(\varepsilon_1^k)I_{p\times p} & \cdots & \dfrac{\varepsilon_k^{k-1}}{(k-1)!} + o(\varepsilon_k^k). \end{bmatrix}.$$

For the ε_j unequal, the Vandermonde determinant

$$\begin{bmatrix} I & & I \\ \varepsilon_1 I_{p\times p} & & \varepsilon_k I_{p\times p} \\ \vdots & & \vdots \\ \dfrac{\varepsilon_1^{k-1}}{(k-1)!}I_{p\times p} & \cdots & \dfrac{\varepsilon_k^{k-1}}{(k-1)!}I_{p\times p} \end{bmatrix} \quad \left(\begin{matrix} I_{p\times p} \text{ denotes the } p\text{-dimensional} \\ \text{identity matrix} \end{matrix}\right)$$

is nonzero, and thus for the ε_j small and unequal the matrix $[e^{A^T\varepsilon_1}H^T, \cdots, e^{A^T\varepsilon_k}H^T]$ is the product of a nonsingular matrix and a matrix of rank n, and, consequently, must have rank n. Thus by Lemma 12, $\sum_1^k e^{A^T t_i} H^T H\, e^{At_i}$ is nonsingular almost everywhere, and it follows that measurement rates of $(k + 1)/\gamma$ will give observability whenever $(n + 1)/\gamma$ does.

For any integer r less than k, the matrix $\sum_1^r e^{A^T t_i} H^T H\, e^{A t_i}$ is singular since $[e^{A^T \varepsilon_1} H^T, \cdots, e^{A^T \varepsilon_r} H^T]$ has rank less than n over an open subset (the ε_j's can be varied about a small open set), and thus the real analytic function det $\sum_1^r e^{A^T t_i} H^T H \cdot e^{A t_i}$ vanishes on an open set and, hence, everywhere. Thus, for any solution for which only r measurements are available, it is readily seen from the analysis of Theorem 13 that we do not have observability.

Acknowledgment. The author would like to thank the referee for his careful and constructive remarks.

REFERENCES

[1] R. F. Brammer, *Controllability in linear autonomous systems with positive controllers*, this Journal, 10 (1972), pp. 339–353.

[2] R. Bucy and P. Joseph, *Filtering for Stochastic Processes with Applications to Guidance*, Interscience, New York, 1968.

[3] S. Cairns, *Differential and Combinatorial Topology*, Princeton University Press, Princeton, N.J., 1965.

[4] E. Lee and L. Markus, *Foundations of Optimal Theory*, John Wiley, New York, 1967.

[5] S. Lefschetz, *Topology*, 2nd ed., Chelsea, New York, 1956.

SIAM J. CONTROL
Vol. 12, No. 3, August 1974

POINTS DE NASH DANS LE CAS DE FONCTIONNELLES QUADRATIQUES ET JEUX DIFFERENTIELS LINEAIRES A N PERSONNES*

A. BENSOUSSAN†

Abstract. In this article, we consider the problem of Nash points for quadratic functionals in Hilbert spaces. Necessary and sufficient conditions of optimality are given. We apply the results to N-person differential games. Open-loop and closed-loop solutions are considered. When there are no constraints, two types of Riccati equations are obtained. We completely develop the parabolic case, generalizing results of A. Friedman [2] who considered finite-dimensional systems. The infinite-dimensional systems that are modeled are the ones which represent distributed parameter systems as in Lions [4].

Introduction. On étudie dans cet article le problème des points de Nash pour fonctionnelles quadratiques du type suivant:

$$J_i(v_i, \ldots, v_n) = \sum_{j=1}^{n} a_{ij}(v_j, v_j) + 2 \sum_{j>k} b_{ijk}(v_j, v_k) - 2 \sum_{j=1}^{n} L_{ij}(v_j).$$

On commence par énoncer des conditions nécessaires et suffisantes (ou simplement suffisantes) afin qu'il existe un point $u_1, \cdots, u_n$ tel que:

$$J_i(u_1, \cdots, u_n) \leqq J_i(u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_n)$$

pour tout v_i vérifiant les contraintes.

On étudie ensuite le problème des jeux différentiels à n personnes avec critère quadratique. Par rapport au problème de contrôle optimal, intervient ici la situation particulière que les stratégies en boucle fermée ou en boucle ouverte ne sont pas équivalentes. L'équation du système étant décrite par:

$$\Lambda y = f + \sum_{i=1}^{n} B_i v_i$$

on considérera les fonctionnelles:

$$J_i(v_1, \cdots, v_n) = \sum_{j=1}^{n} \alpha_{ij} \|M_j y\|^2 + \sum_{j=1}^{n} \beta_{ij}(N_j v_j, v_j).$$

Les stratégies en boucle ouverte correspondent au problème suivant: Trouver un point de Nash pour les fonctionnelles J_i. Les stratégies en boucle fermée sont des applications $\mathscr{L}_i(y)$ à valeurs dans les espaces de contrôle.

On considère alors les fonctionnelles:

$$\mathscr{T}_i(\mathscr{L}_1, \mathscr{L}_2, \cdots, \mathscr{L}_n) = J_i(\mathscr{L}_1(y), \mathscr{L}_2(y), \cdots, \mathscr{L}_n(y))$$

et on cherche un point de Nash $\mathscr{L}_1, \cdots, \mathscr{L}_n$ (voir aussi D. Lukes [7]).

L'existence de points de Nash en boucle ouverte ou en boucle fermée est liée à l'étude de systèmes d'équations de Riccati.

* Received by the editors September 29, 1972.

† I.R.I.A., Informatique Numérique (Batiment 12), Domaine de Voluceau, 78-Rocquencourt, France.

Toutefois, en boucle fermée, les équations de Riccati prennant une importance trés grande, cas la résolution de ces équations est le seul moyen permettant de prouver l'existence de stratégies optimales, alors que dans le cas "boucle ouverte" des méthodes directes sont possibles.

On fait alors une étude complète du cas évolutif et on montre l'existence de solutions pour un système d'équations de Riccati, sur un intervalle de temps fini assez petit. Cela généralise un résultat de A. Friedman [2] où l'existence locale pour les équations différentielles en dimension finie est utilisée pour démontrer l'existence de stratégies optimales.

Le modèle général étudié est en dimension infinie, ce qui permet de traiter les systèmes décrits par des équations différentielles aussi bien que les systémes décrits par des équations aux dérivées partielles.

Le plan est ci-après:

1. Conditions d'existence d'un point Nash pour fonctionnelles quadratiques
    1.1. Données—Notations
    1.2. Conditions nécessaires et suffisantes d'existence d'un point de Nash
    1.3. Condition suffisante d'existence et d'unicité

2. Jeux différentiels à N personnes en boucle ouverte
    2.1. Données—Notations
    2.2. Condition nécessaire et suffisante d'existence d'un point de Nash en boucle ouverte
    2.3. Découplage

3. Jeux différentiels à N personnes en boucle fermée
    3.1. Le problème
    3.2. Point de Nash en boucle fermée

4. Etude du cas parabolique
    4.1. Notations
    4.2. Point de Nash en boucle fermée
    4.3. Etude des équations de Riccati

1. Conditions d'existence d'un point de Nash pour fonctionnelles quadratiques.

1.1. Données—Notations. Soient $J_1, \cdots, \mathscr{U}_n$ n espaces de Hilbert et $\mathscr{U}_i^{\mathrm{ad}}, \cdots, \mathscr{U}_n^{\mathrm{ad}} n$ sous-ensembles convexes fermés de $\mathscr{U}_1, \cdots, \mathscr{U}_n$ respectivement. On posera: $v = (v_i, \cdots, v_n) \in \mathscr{U}_1 \times \cdots \times \mathscr{U}_n$ et on définit n fonctionnelles $J_1(v), \cdots, J_n(v)$ par:

$$J_i(v) = \sum_{j=1}^{n} a_{ij}(v_j, v_j) + 2 \sum_{j>k} b_{ijk}(v_j, v_k) - 2 \sum_{j=1}^{n} L_{ij}(v_j), \tag{1.1}$$

où

(1.2) $a_{ij}(v_j, w_j), i, j = 1, \cdots, n,$ sont des fonctionnelles bilinéaires continues symétriques sur $\mathscr{U}_j$,

(1.3) $b_{ijk}(v_j, v_k), i, j, k = 1, \cdots, n, j > k,$ sont des fonctionnelles bilinéaires continues sur $\mathscr{U}_j \times \mathscr{U}_k$, (1.4) $L_{ij}(v_j), i, j = 1, \cdots, n$, sont des fonctionnelles linéaires continues sur $\mathcal{U}_j$.

On rappelle qu'un point $u = (u_1, \cdots, u_n)$ est un point de Nash si les inégalités suivantes sont vérifiées:

$$J_i(u_1, u_i, u_n) \leqq J_i(u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_n) \quad \text{pour tout} \quad v_i \in K_i, \quad i = 1, \cdots, n. \tag{1.5}$$

1.2. Conditions nécessaires et suffisantes d'existence d'un point de Nash. Considérons la fonctionnelle:

$$\tilde{J}_i(v_i) = J_i(u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_n) \tag{1.6}$$

considérée comme fonction de la seule variable v_i. D'après (1.1), on a aussitôt:

$$\tilde{J}_i(v_i) = a_{ii}(v_i, v_i) + 2 \sum_{j=i+1}^{n} b_{iji}(u_j, v_i) + 2 \sum_{j=1}^{i-1} b_{iij}(v_i, u_j) - 2L_{ii}(v_i) + Cte \tag{1.7}$$

et u$_i$ réalise le minimum de $\tilde{J}_i(v_i)$ sur K_i.

Il est classique (cf. J. L. Lions [4]) que si a$_{ii}(v_i, v_i)$ est $\geqq 0$, alors pour que $\tilde{J}_i(v_i)$ atteigne son minimum sur K_i, il est nécessaire et suffisant que l'on ait:

$$\left(\frac{\partial \tilde{J}_i}{\partial v_i}(u_i), v_i - u_i\right) \geqq 0 \quad \text{pour tout} \quad v_i \in K_i. \tag{1.8}$$

On posera pour simplifier les notations:

$$\begin{aligned} a_i(v_i, v_i) &= a_{ii}(v_i, v_i), \\ b_{ij}(v_j, v_i) &= \begin{cases} b_{iji}(v_j, v_i) & si \quad j > i, \\ b_{iij}(v_i, v_j) & si \quad j < i, \end{cases} \\ L_i(v_i) &= L_{ii}(v_i). \end{aligned} \tag{1.9}$$

On déduit aussitôt de (1.8) le théorème suivant.

THÉORÈME 1.1. *Sous les hypothèses et notations du* § 1.1, *et* $a_i(v_i, v_i) \geqq 0$, *alors la C.N.S. pour que* $u = (u_1, \cdots, u_n)$ *soit un point de Nash pour les fonctionnelles* $J_1, \cdots, J_n$ *est que l'on ait*:

$$a_i(u_i, v_i - u_i) + \sum_{j \neq i} b_{ij}(u_j, v_i - u_i) - L_i(v_i - u_i) \geqq 0 \quad \textit{pour tout} \quad v_i \in K_i, \quad i = 1, \cdots, n, \tag{1.10}$$

ou encore

$$\sum_{i=1}^{n} a_i(u_i, v_i - u_i) + \sum_{i \neq j = 1}^{n} b_{ij}(u_j, v_i - u_i) - \sum_{i=1}^{n} L_i(v_i - u_i) \geqq 0 \quad \textit{pour tout} \quad v_i \in K_i, \quad i = 1, \cdots, n. \tag{1.11}$$

Démonstration. Les conditions (1.10) ne sont autres que la traduction des conditions (1.8). Quant à l'équivalence de (1.10) et (1.11) c'est évident.

1.3. Condition suffisante d'existence et d'unicité. Sous la forme (1.11), la C.N.S. d'existence d'un point de Nash se ramène donc à l'étude d'une inéquation variationnelle sur l'espace de Hilbert $\mathcal{U} = \mathcal{U}_1 \times \cdots \times \mathcal{U}_n$.

Plus précisèment posons:

$$a(v, w) = \sum_{i=1}^{n} a_i(v_i, w_i) + \sum_{i \neq j = 1}^{n} b_{ij}(v_j, w_i) \tag{1.12}$$

et

$$L(v) = \sum_{i=1}^{n} L_i(v_i); \tag{1.13}$$

alors (1.11) s'écrit:

$$(1.11)_{\text{bis}} \qquad a(u, v - u) - L(v - u) \geqq 0 \quad \text{pour tout} \quad v \in K_1 \times \cdots \times K_n.$$

On a alors d'après un résultat classique de Lions et Stampacchia [6]: Si l'hypothèse suivante est vérifiée:

$$a(v, v) \geqq \alpha \|v\|^2_{\mathcal{U}}, \tag{1.14}$$

il *existe u unique* $\in K$ tel que $(1.11)_{\text{bis}}$ soit vérifiée. La condition *suffisante* d'existence et d'unicité $(1.11)_{\text{bis}}$ s'explicite donc de la maniére suivante:

$$\text{Il existe} \quad \alpha > 0 \quad \text{tel que} \quad \sum_{i=1}^{n} a_i(v_i, v_i) + \sum_{i \neq j = 1}^{n} b_{ij}(v_j, v_i) \geqq \alpha \sum_{i=1}^{n} \|v_i\|^2. \tag{1.15}$$

Remarque 1.1. Considérons à titre d'exemple le problème du point-selle étudié par Lemaire [3]. On prend $n = 2$, et

$$\begin{aligned} a_{12}(v_2, v_2) &= -a_2(v_2, v_2), \\ a_{21}(v_1, v_1) &= -a_1(v_1, v_1), \\ b_{121}(v_2, v_1) &\doteq b(v_2, v_1), \\ b_{221}(v_2, v_1) &= -b(v_2, v_1), \\ L_{12}(v_2) &= -L_2(v_2), \\ L_{21}(v_1) &= -L_1(v_1). \end{aligned} \tag{1.16}$$

Alors (1.9) donne aussitôt

$$\begin{aligned} b_{12}(v_2, v_1) &= b_{121}(v_2, v_1) = b(v_2, v_1), \\ b_{21}(v_1, v_2) &= b_{221}(v_2, v_1) = -b(v_2, v_1), \end{aligned} \tag{1.17}$$

et donc,

$$b_{12}(v_2, v_1) + b_{21}(v_1, v_2) = 0$$

par conséquent (1.15) devient

$$a_1(v_1, v_1) + a_2(v_2, v_2) \geqq \alpha(\|v_1\|^2 + \|v_2\|^2), \tag{1.18}$$

ce qui est évidemment équivalent à:

$$\begin{aligned} a_1(v_1, v_1) &\geqq \alpha \|v_1\|^2, \\ a_2(v_2, v_2) &\geqq \alpha \|v_2\|^2. \end{aligned} \tag{1.19}$$

Remarque 1.2. *Le cas des contraintes couplées.* Le cas où décisions sont astreintes à des contraintes couplées ne se ramène pas du tout malheureusement à des inéquations variationnelles. En effet, notons pour simplifier $\bar{u}_i$ l'ensemble des coordonnées $u_1, \cdots, u_{i-1}, \cdots, u_n$, alors (1.5) doit étre remplacée par:

$$(1.5') \qquad J_i(u) \leqq J_i(v_i, \bar{u}_i), \quad \text{pour tout } v_i \text{ tel que} \quad (v_i, \bar{u}_i) \in K \subset \prod_{i=1}^{n} \mathcal{U}_i.$$

Par conséquent la contrainte $(v_i, \bar{u}_i) \in K$ peut s'écrire sous la forme:

$$(1.5'') \qquad v_i \in K_i(\bar{u}_i) \subset \mathcal{U}_i$$

bien entendu $u_i \in K_i(\bar{u}_i)$. Il en résulte que la C.N.S. d'existence d'un point de Nash s'écrira:

$$(1.11') \qquad \sum_{i=1}^{n} a_i(u_i, v_i - u_i) + \sum_{i \neq j = 1}^{n} b_{ij}(u_j, v_i - u_i) - \sum_{i=1}^{n} L_i(v_i - u_i) \geqq 0 \qquad \text{pour tout} \quad v_i \in K_i(\bar{u}_i),$$

ce qui s'écrit donc aussi:

$$(1.11'') \qquad a(u, v - u) - L(v - u) \geqq 0 \quad \text{pour tout} \quad v \in K(u)$$

où

$$(1.20) \qquad \begin{aligned} K(u) &= \prod_{i=1}^{n} K_i(\bar{u}_i) \\ &= \{|(v_1 \cdots v_n)| \text{ pour tout } i, (u_1, \cdots, u_{i-1}, v_i, u_{i+1}, \cdots, u_n) \in K\}. \end{aligned}$$

On voit ainsi que (1.11)″ n'est pas une inéquation variationnelle classique et pose en fait des problèmes d'un type nouveau.

Remarque 1.3. Considérons le cas où $K_i = \mathcal{U}_i$, alors pour que $\tilde{J}_i(v_i)$ atteigne son minimum, il est nécessaire que $a_{ii}(v_i, v_i) \geqq 0$ pour tout v_i. De sorte que l'on peut énoncer le corollaire suivant.

COROLLAIRE 1.1. *Sous les hypothèses et notations du* § 1.1, *la C.N.S. pour que les fonctionnelles $J_i(v)$ possèdent un point de Nash sur tout l'espace est que*

$$(1.21) \qquad a_i(v_i, v_i) \geqq 0 \quad \textit{pour tout} \quad v_i \in \mathcal{U}_i$$

et qu'il existe $u_1, \cdots, u_n$ tel que

$$(1.22) \qquad \sum_{i=1}^{n} a_i(u_i, v_i) + \sum_{i \neq j = 1}^{n} b_{ij}(u_j, v_i) = \sum_{i=1}^{n} L_i(v_i) \qquad \textit{pour tout} \quad v_i, \qquad i = 1, \cdots, n.$$

Le point de Nash est unique si et seulement si la solution de (1.22) *est unique.*

COROLLAIRE 1.2. *Si* (1.15) *et* (1.21) *sont vérifiées, il résulte du Théorème de Lax-Milgram* (*cas particulier du Théorème de Lions—Stampacchia*) (cf., par example, Yosida [8]) *que les fonctionnelles $J_i(v)$ possèdent un point de Nash unique sur tout l'espace.*

2. Jeux différentiels à N personnes en boucle ouverte.

2.1. Données—Notations. Soit $\mathscr{H}$ un espace de Hilbert. Considérons sur $\mathscr{H}$ un opérateur non borné Λ, dont le domaine $D(\Lambda)$ est dense dans $\mathscr{H}$. On munit $D(\Lambda)$ de la norme du graphe, c'est-à-dire

$$\|\varphi\|^2_{D(\Lambda)} = \|\varphi\|^2_{\mathscr{H}} + \|\Lambda\varphi\|^2_{\mathscr{H}} \tag{2.1}$$

et on suppose

$$\begin{aligned}&\Lambda \text{ est un isomorphisme de } \mathscr{H} \text{ dans } D(\Lambda)\\ &\text{et} \quad (\Lambda\varphi, \varphi) \geqq 0 \quad \text{pour tout} \quad \varphi \in D(\Lambda).\end{aligned} \tag{2.2}$$

On considère alors un système décrit par l'équation

$$\Lambda y = f + \sum_{i=1}^{n} B_i v_i, \tag{2.3}$$

où $v_i \in \mathscr{U}_i$ (espaces de Hilbert) et

$$f \in \mathscr{H}, \quad B_1 \in \mathscr{L}(\mathscr{U}_1; \mathscr{H}), \cdots, B_n \in (\mathscr{U}_n; \mathscr{H}). \tag{2.4}$$

On se donne ensuite les fonctionnelles

$$J_i(v) = \sum_{j=1}^{n} \alpha_{ij}\|M_j y\|^2 + \sum_{j=1}^{n} \beta_{ij}(N_j v_j, v_j), \qquad i = 1, \cdots, n, \tag{2.5}$$

où les α_{ij} et β_{ij} sont des coefficients réels et

$$\begin{cases} M_j \in \mathscr{L}(\mathscr{H}; F), & F \text{ espace de Hilbert},\\ N_j \in \mathscr{L}(\mathscr{U}_j; \mathscr{U}_j) & \text{autoadjoint} \quad \geqq 0 \quad \text{et inversible}. \end{cases} \tag{2.6}$$

On pose ensuite

$$R_i = \sum_{l=1}^{n} \alpha_{il} M_l^* M_l. \tag{2.7}$$

Les R_i sont symétriques (mais pas forcément $\geqq 0$). Un calcul rapide simple montre alors que l'on a:

$$\begin{aligned} a_{ij}(v_j, v_j) &= \beta_{ij}(N_j v_j, v_j) + (R_i\Lambda^{-1}B_j v_j, \Lambda^{-1}B_j v_j),\\ b_{ijk}(v_j, v_k) &= (R_i\Lambda^{-1}B_j v_j, \Lambda^{-1}B_k v_k), \qquad j > k,\\ L_{ij}(v_j) &= -(R_i\bar{y}, \Lambda^{-1}B_j v_j). \end{aligned} \tag{2.8}$$

Le problème consiste à trouver un point de Nash dans tout l'espace, pour les fonctionnelles $J_i(v)$ (problème dit en "boucle ouverte" par analogie avec le contrôle optimal).

2.2. Condition nécessaire et suffisante d'existence d'un point de Nash en boucle ouverte. On peut appliquer le Corollaire 1.1. La condition (1.21) donne aussitôt:

$$\beta_{ii}(N_i v_i, v_i) + (R_i\Lambda^{-1}B_i v_i, \Lambda^{-1}B_i v_i) \geqq 0 \quad \text{pour tout } v_i, \quad \text{pour tout } i, \tag{2.9}$$

et (1.22) donne

$$\text{(2.10)}\quad \begin{aligned}&\sum_{i=1}^{n} \beta_{ii}(N_iu_i, v_i) + (R_i\Lambda^{-1}B_iu_i, \Lambda^{-1}B_iv_i)\\ &\quad + \sum_{i\neq j=1}^{n} (R_iy_j(u_j), y_i(v_i)) = -\sum_{i=1}^{n} (R_i\bar{y}, y_i(v_i)),^1\end{aligned}$$

ce qui s'ècrit encore

$$\text{(2.11)}\quad \sum_{i=1}^{n} \beta_{ii}(N_iu_i, v_i) + \sum_{i=1}^{n} (R_iy(u), \Lambda^{-1}B_iv_i) = 0 \qquad \text{pour tout} \quad v_1, \cdots, v_n.$$

On introduit alors $p_i \in D(\Lambda^*)$ solution de

$$\text{(2.12)}\qquad \Lambda^*p_i = R_iy(u).$$

On déduit aussitôt de (2.11) que si

$$\text{(2.13)}\qquad \beta_{ii} \neq 0,$$

alors on obtient

$$\text{(2.14)}\qquad u_i = \frac{1}{\beta_{ii}}N_i^{-1}B_i^*p_i$$

de sorte que l'on obtient le théorème suivant.

THÉORÈME 2.1. *Sous les hypothèses et notations de* § 2.1, *une C.N.S. d'existence de points de Nash pour les fonctionnelles $J_i(v)$ est que*

$$\text{(2.15)}\quad \beta_{ii}(N_iv_i, v_i) + (R_i\Lambda^{-1}B_iv_i, \Lambda^{-1}B_iv_i) \geqq 0 \quad \textit{pour tout} \quad v_i, \qquad i = 1, \cdots, n,$$

et qu'il existe $y, p_1, \cdots, p_n$ solutions du système

$$\text{(2.16)}\quad \begin{aligned}&\Lambda y + \sum_{i=1}^{n} \frac{B_iN_i^{-1}B_i^*}{\beta_{ii}}p_i = f,\\ &\Lambda^*p_i = R_iy, \qquad\qquad i = 1, \cdots, n,\\ &y \in D(\Lambda), \quad p_i \in D(\Lambda^*).\end{aligned}$$

Il y a unicité si et seulement si la solution de (2.16) *est unique.*

COROLLAIRE 2.1. *Une C.N.S. pour que les fonctionnelles $J_i(v)$ admettent un point de Nash unique est que* (2.15) *soit vérifiée et que*

$$\text{(2.17)}\qquad X = \Lambda + \sum_{i=1}^{n} \frac{B_iN_i^{-1}B_i^*}{\beta_{ii}}\Lambda^{*-1}R_i$$

soit un isomorphisme de $D(\Lambda) \to \mathscr{H}$.

Démonstration. En effet, l'existence d'une solution unique de (2.16) est équivalente au fait que X est un isomorphisme de $D(\Lambda) \to \mathscr{H}$ (puisque f est quelconque).

[1] Où on a posé $y_j(v_j) = \Lambda^{-1}B_jv_j$.

Donnons maintenant deux conditions suffisantes pour que X soit un isomorphisme de $D(\Lambda) \to \mathscr{H}$.

PROPOSITION 2.1. *Si l'une ou l'autre des conditions suivantes est satisfaite:*

$$\begin{aligned}&\sum_{i=1}^{n} \beta_{ii}(N_i u_i, u_i) + \sum_{i,j=1}^{n} (R_i\Lambda^{-1}B_i u_i, \Lambda^{-1}B_j u_j)\\ &\geqq \alpha \sum_{i=1}^{n} \|u_i\|^2 \quad \textit{pour tout} \quad u_i, \cdots, u_n,\end{aligned} \tag{2.18}$$

ou bien

$$\begin{aligned}&R_i = S_i T \quad \textit{où} \quad T \in \mathscr{L}(\mathscr{H}, \mathscr{H}), \quad T \geqq 0 \quad \textit{autoadjoint} \quad S_i h \in D(\Lambda^*) \textit{ si } h \in D(\Lambda^*),\\ &\qquad S_i \in \mathscr{L}(\mathscr{H}, \mathscr{H}),\end{aligned} \tag{2.19}$$

et

$$\Lambda^* S_i = S_i \Lambda^*,$$

$$\sum_{i=1}^{n} \frac{B_i N_i^{-1} B_i^*}{\beta_{ii}} S_i \geqq 0;$$

alors X est un isomorphisme de $\mathscr{H}$ dans $D(\Lambda)$.

Démonstration. Si la condition (2.18) est satisfaite, on vérifie par un calcul simple que la condition (1.15) est satisfaite et donc d'après le théorème de Lions et Stampacchia (cf. § 1.3), il existe un point de Nash unique pour les fonctionnelles $J_i(v)$, ce qui d'après le Corollaire 2.1, implique bien que X est un isomorphisme de $D(\Lambda)$ dans $\mathscr{H}$.

Par ailleurs si les conditions (2.19) sont vérifiées, alors on remarque aussitôt que l'on a

$$\Lambda^{*-1} S_i = S_i \Lambda^{*-1}, \tag{2.20}$$

de sorte que le système (2.16) s'ècrit

$$\begin{aligned}&\Lambda y + \sum_{i=1}^{n} \frac{B_i N_i^{-1} B_i^*}{\beta_{ii}} p_i = f,\\ &\Lambda p_i = S_i T y, \qquad\qquad i = 1, \cdots, n,\\ &y \in D(\Lambda), \quad p_i \in D(\Lambda^*).\end{aligned} \tag{2.21}$$

Tenant compte de (2.20), la deuxième relation (2.21) s'écrira

$$p_i = S_i \Lambda^{*-1} T y. \tag{2.22}$$

On peut donc poser $p = \Lambda^{*-1} T y$ et (2.21) conduit au système en (y, p):

$$\begin{aligned}&\Lambda y + \sum_{i=1}^{n} \frac{B_i N_i^{-1} B_i^* S_i}{\beta_{ii}} p = f,\\ &\Lambda^* = T y,\\ &y \in D(\Lambda), \quad p \in D(\Lambda^*).\end{aligned} \tag{2.23}$$

Posons

$$\mathscr{D} = \sum_{i=1}^{n} \frac{B_i N_i^{-1} B_i^* S_i}{\beta_{ii}}. \tag{2.24}$$

Alors $T \geqq 0$ et $\mathscr{D} \geqq 0$ impliquent que (2.23) est le système des équations d'Euler du problème de contrôle optimal suivant: Le système est gouverné par l'equation

$$\Lambda z = f + \mathscr{D}^{1/2} v \tag{2.25}$$

et le critère est

$$J(v) = (Tz, z) + \|v\|^2. \tag{2.26}$$

Par conséquent (2.23) possède une solution unique, ce qui implique bien que (2.16) possède une solution unique, d'où le résultat.

Prenons un exemple. Le modèle est le suivant: L'équation d'état est donnée par (2.3). On a ensuite:

$$\begin{aligned} &\alpha_{ij} = 0 \quad \text{si} \quad j \neq i, \\ &\alpha_{ii} = \varepsilon_i = \pm 1, \\ &\beta_{ii} = 1, \\ &\beta_{ij} = \varepsilon_i \varepsilon_j \quad \text{si} \quad i \neq j, \\ &M_j = M. \end{aligned} \tag{2.27}$$

Par conséquent on obtient

$$R_i = \varepsilon_i M^* M \tag{2.28}$$

et donc, on peut prendre $S_i = \varepsilon_i I$. Si donc

$$\sum_{i=1}^{n} \varepsilon_i B_i N_i^{-1} B_i \geqq 0 \tag{2.29}$$

il résulte de la Proposition 2.1 et du Théorème 2.1 que si

$$N_i + \varepsilon_i B_i^* \Lambda^{-1*} M^* M \Lambda^{-1} B_i \geqq 0 \tag{2.30}$$

et (2.29) sont vérifiées, alors les fonctionnelles $J_i(v)$ admettent un point de Nash. Le problème ainsi traité se reformule aisément de la manière suivante:

$$\cdots \min_{v_i(\varepsilon_i = +1)} \cdots \max_{v_j(\varepsilon_j = -1)} \cdots J(v) = \|My\|^2 + \sum_{i=1}^{n} \varepsilon_i (N_i v_i, v_i).$$

En particulier pour $n = 2$, on retrouve le classique problème du minimax pour les jeux différentiels en boucle ouverte (cf. Lemaire [3], Bensoussan [1]).

2.3. Découplage. On introduit les systèmes d'équations (du type Riccati)

$$P_i \Lambda + \Lambda^* P_i + P_i \sum_{j=1}^{n} B_j N_j^{-1} B_j^* P_j = \frac{R_i}{\beta_{ii}}, \qquad i = 1, \cdots, n. \tag{2.31}$$

Nous allons supposer ici que ce système possède une solution. Donnons un cas où cette condition sera satisfaite.

PROPOSITION 2.2. *Si l'hypothèse* (2.19) *est satisfaite, alors le système* (2.31) *possède une solution donnée par*

$$P_i = \frac{S_i P}{\beta_{ii}}, \tag{2.32}$$

où P est la solution unique de l'équation de Riccati

$$P\Lambda + \Lambda^* P + P \sum_{j=1}^{n} \frac{B_j N_j^{-1} B_j^* S_j}{\beta_{jj}} P = T. \tag{2.33}$$

Démonstration. Le fait que (2.33) possède une solution unique $\in \mathscr{L}(\mathscr{H}, \mathscr{H})$ résulte de la théorie du découplage (Lions [4], [5]).

On déduit alors de (2.33),

$$\frac{S_i P \Lambda}{\beta_{ii}} + \frac{S_i \Lambda^* P}{\beta_{ii}} + \frac{S_i P}{\beta_{ii}} \sum_{j=1}^{n} \frac{B_j N_j^{-1} B_j^* S_j P}{\beta_{jj}} = \frac{S_i}{\beta_{ii}} T \tag{2.34}$$

et tenant compte de (2.19) on obtient

$$\frac{S_i P \Lambda}{\beta_{ii}} + \frac{\Lambda^* S_i P}{\beta_{ii}} + \frac{S_i P}{\beta_{ii}} \sum_{j=1}^{n} \frac{B_j N_j^{-1} B_j^* S_j P}{\beta_{jj}} = \frac{R_i}{\beta_{ii}}. \tag{2.35}$$

Il est alors évident que $P_i = S_i P/\beta_{ii}$ est solution de (2.31).

On a ensuite la proposition suivante.

PROPOSITION 2.3. *Lorsque* (2.31) *possède une solution, alors l'ensemble* (y, p_i) *solution de* (2.21) *satisfait aux relations*

$$p_i = (P_i y + \rho_i)\beta_{ii}, \tag{2.36}$$

où les ρ_i *sont solutions des équations*

$$\Lambda^* \rho_i + P_i \sum_{j=1}^{n} B_j N_j^{-1} B_j^* \rho_j = P_i f, \qquad i = 1, \cdots, n. \tag{2.37}$$

Démonstration. On définit ρ_i à l'aide des relations (2.36). On obtient ensuite

$$\begin{aligned} \Lambda^* \rho_i &= \frac{\Lambda^* p_i}{\beta_{ii}} - \Lambda^* P_i y \\ &= \frac{R_i y}{\beta_{ii}} - \Lambda^* P_i y \\ &= P_i \Lambda y + P_i \sum_{j=1}^{n} B_j N_j^{-1} B_j^* P_j y. \end{aligned} \tag{2.38}$$

Or de la première relation (2.21) on déduit

$$\Lambda y + \sum_{i=1}^{n} B_i N_i^{-1} B_i^* P_i y + \sum_{i=1}^{n} B_i N_i^{-1} B_i^* \rho_i = f \tag{2.39}$$

de sorte que (2.38) conduit aussitôt à (2.37).

COROLLAIRE 2.2. *Si le système* (2.31) *possède une solution, et si la C.N.S. d'existence d'un point de Nash unique* (cf. Théorème 2.1) *est satisfaite, alors celui-ci* *est donné par la formule*

$$u = -N_i^{-1}B_i^* p_i y - N_i^{-1}B_i^*\rho_i, \qquad i = 1, \cdots, n. \tag{2.40}$$

Remarque 2.1. Posons

$$\mathscr{L}_i^*(y) = -N_i^{-1}B_i^*P_i y - N_i^{-1}B_i^*\rho_i, \tag{2.41}$$

i.e., $\mathscr{L}_i^*(y)$ est la fonction affine de y définie par (2.41). Alors (2.40) montre que $u_i = \mathscr{L}_i^*(y)$, pour y solution de (2.16). On exprime ce fait en disant que les contrôles u_i sont obtenus par l'intermédiaire d'un feedback (ou stratégie). Ceci ne prouve pas que les stratégies $\mathscr{L}_i^*$ constituent un point de Nash dans l'ensemble des stratégies, problème que nous allons examiner maintenant.

3. Jeux différentiels à N personnes en boucle fermée.

3.1. Le problème. On considère des applications $\mathscr{L}_i$, $i = 1, \cdots, n$ de $\mathscr{H} \to \mathscr{U}_i$. On dira que $\mathscr{L}_i \equiv \mathscr{L}_i(z)$ est une stratégie pour le joueur n° i. Les joueurs peuvent choisir leurs stratégies avec la seule restriction que l'équation

$$\Lambda z = f + \sum_{i=1}^{n} B_i \mathscr{L}_i(z) \tag{3.1}$$

possède une solution *unique*.

On introduit alors les fonctionnelles $\mathscr{T}_i(\mathscr{L}_1, \mathscr{L}_2, \cdots, \mathscr{L}_n)$ définies par

$$\mathscr{T}_i(\mathscr{L}_1, \mathscr{L}_2, \cdots, \mathscr{L}_n) = J_i(\mathscr{L}_1(z), \mathscr{L}_2(z), \cdots, \mathscr{L}_n(z)), \tag{3.2}$$

où z est solution de (3.1).

Le problème consiste à trouver un point de Nash pour les fonctionnelles $\mathscr{T}_i$, c'est-à-dire des stratégies $\hat{\mathscr{L}}_i(y)$ telles que

$$\mathscr{T}_i(\hat{\mathscr{L}}_1, \cdots, \hat{\mathscr{L}}_i, \cdots, \hat{\mathscr{L}}_n) \leqq \mathscr{J}_i(\hat{\mathscr{L}}_1, \cdots, \hat{\mathscr{L}}_{i-1}, \mathscr{L}_i, \hat{\mathscr{L}}_{i+1}, \cdots, \hat{\mathscr{L}}_n) \tag{3.3}$$
$$\text{pour toute stratégie } \mathscr{L}_i.$$

On dira aussi que (3.3) consiste à chercher un point de Nash en boucle fermée.

3.2. Point de Nash en boucle formée.

On introduit les systèmes d'equations de Riccati

$$\begin{aligned}&\left(\Lambda^* + \sum_{j \neq i} Q_j^* B_j N_j^{-1} B_j^*\right) Q_i + Q_i^*\left(\Lambda + \sum_{j \neq i} B_j N_j^{-1} B_j^* Q_j\right) + Q_i^* B_i N_i^{-1} B_i^* Q_i \\ &\quad = \frac{R_i}{\beta_{ii}} + \sum_{j \neq i} \frac{\beta_{ij} Q_j^* B_j N_j^{-1} B_j^* Q_j}{\beta_{ii}}, \qquad i = 1, \cdots, n,\end{aligned} \tag{3.4}$$

et on fait l'hypothèse que ce système d'équations possède une solution. On définit ensuite les $\sigma_i \in D(\Lambda^*)$ solutions du système d'équations:

$$\Lambda^* \sigma_i + \left(\sum_{j=1}^{n} Q_j^* B_j N_j^{-1} B_j^*\right)\sigma_i = Q_i^* f + \sum_{j \neq i}\left(\frac{\beta_{ij} Q_j^*}{\beta_{ii}} - Q_i^*\right) B_j N_j^{-1} B_j^* \sigma_j. \tag{3.5}$$

On suppose que les équations (3.5) possèdent une solution. On admet également que l'équation

$$\Lambda y + \sum_{j=1}^{n} B_j N_j^{-1} B_j^* Q_j y = f - \sum_{j=1}^{n} B_j N_j^{-1} B_j^* \sigma_j \tag{3.6}$$

possède une solution unique $y \in D(\Lambda)$.

On se restreint enfin aux $\beta_{ii} > 0$.

On a alors le théorème suivant.

THÉORÈME 3.1. *Sous les hypothèses que* (3.4), (3.5) *possèdent des solutions, que* (3.6) *possède une solution unique et que* $\beta_{ii} > 0$, *il existe un point de Nash en boucle fermée unique défini par*

$$\hat{\mathscr{L}}_i(z) = -N_i^{-1}B_i^*Q_iz - N_i^{-1}B_i^*\sigma_i. \tag{3.7}$$

Démonstration. Il s'agit de démontrer la propriété (3.3). Cela revient à montrer que $\hat{\mathscr{L}}_i$ est solution du problème de contrôle optimal en boucle fermée suivant: le système est gouverné par l'équation

$$\Lambda z + \sum_{j\neq i} B_jN_j^{-1}B_j^*Q_jz = f - \sum_{j\neq i} B_jN_j^{-1}B_j^*\sigma_j + B_i\mathscr{L}_i(z), \tag{3.8}$$

où $\mathscr{L}_i(z)$ est une application de $\mathscr{H} \to \mathscr{U}_i$ telle que (3.8) possède une solution unique et le critère est

$$\begin{aligned} J_i(\mathscr{L}_i) = (R_iz, z) + \beta_{ii}(N_i\mathscr{L}_i(z), \mathscr{L}_i(z)) + \sum_{j\neq i} \beta_{ij} \\ \cdot (N_j(-N_j^{-1}B_j^*Q_jz - N_j^{-1}B_j^*\sigma_j), -N_j^{-1}B_j^*Q_jz - N_j^{-1}B_j^*\sigma_j). \end{aligned} \tag{3.9}$$

A toute stratégie $\mathscr{L}_i$ est associée la fonction $w_i(z)$ définie par la formule

$$w_i(z) = \mathscr{L}_i(z) + N_i^{-1}B_i^*Q_iz + N_i^{-1}B_i^*\sigma_i. \tag{3.10}$$

Puisque l'èquation (3.8) possède une solution unique, la fonction w_i a donc la propriété que l'équation

$$\Lambda z + \sum_{j=1}^{n} B_jN_j^{-1}B_j^*Q_jz = f - \sum_{j=i}^{n} B_jN_j^{-1}B_j^*\sigma_j + B_iw_i(z) \tag{3.11}$$

possède une solution unique.

Notons que $\hat{\mathscr{L}}_i$ est une stratégie puisque (3.6) possède une solution unique. A $\hat{\mathscr{L}}_i$ correspond la fonction $w_i = 0$. Enfin si on pose

$$z = y + \tilde{z}, \tag{3.12}$$

alors de (3.6) et (3.11) résulte aussitôt

$$\Lambda\tilde{z} + \sum_{j=1} B_jN_j^{-1}B_j^*Q_j\tilde{z} = B_iw_i(z). \tag{3.13}$$

En utilisant (3.10), on transforme (3.9) de la manière suivante:

$$\begin{aligned} J_i(\mathscr{L}_i) = \bar{J}_i(w_i) &= (R_iz, z) + \beta_{ii}(N_i(w_i - N_i^{-1}B_i^*Q_iz - N_i^{-1}B_i^*\sigma_i), \\ &\qquad w_i - N_i^{-1}B_i^*Q_iz - N_i^{-1}B_i^*\sigma_i) \\ &\quad + \sum_{j\neq i} \beta_{ij}(B_j^*Q_jz + B_j^*\sigma_j, N_j^{-1}B_j^*Q_jz + N_j^{-1}B_j^*\sigma_j) \\ &= (R_iz, z) + \beta_{ii}(N_iw_i, w_i) \\ &\quad + \sum_{j=1}^{n} \beta_{ij}(B_j^*Q_jz + B_j^*\sigma_j, N_j^{-1}B_j^*Q_jz + N_j^{-1}B_j^*\sigma_j) \\ &\qquad - 2\beta_{ii}(w_i, B_i^*Q_iz + B_i^*\sigma_i). \end{aligned} \tag{3.14}$$

Par ailleurs, en utilisant (3.12) on obtient encore

$$
\begin{aligned}
\bar{\bar{J}}_i(w_i) &= (R_i\tilde{z}, \tilde{z}) + \beta_{ii}(N_iw_i, w_i) \\
&\quad + (R_iy, y) + 2(R_iy, \tilde{z}) \\
&\quad + \sum_{j=1}^{n} \beta_{ij}(B_j^*Q_jy + B_j^*\sigma_j, N_j^{-1}B_j^*Q_jy + N_j^{-1}B_j^*\sigma_j) \\
&\quad + \sum_{j=1}^{n} \beta_{ij}(B_j^*Q_jz, N_j^{-1}B_j^*Q_j\tilde{z}) \\
&\quad + 2\sum_{j=1}^{n} \beta_{ij}(B_j^*Q_j\tilde{z}, N_j^{-1}B_j^*Q_jy + N_j^{-1}B_j^*\sigma_j) \\
&\quad - 2\beta_{ii}(w_i, B_i^*Q_iy + B_i^*\sigma_i) \\
&\quad - 2\beta_{ii}(w_i, B_i^*Q_i\tilde{z}).
\end{aligned}
\tag{3.15}
$$

On a ensuite, d'après (3.13)

$$
(B_iw_i, Q_i\tilde{z}) = \left(Q_i\tilde{z}, \Lambda\tilde{z} + \sum_{j=1}^{n} B_jN_j^{-1}B_j^*Q_j\tilde{z}\right).
\tag{3.16}
$$

On remarque que les èquations (3.4) peuvent encore s'écrire

$$
\begin{aligned}
&\left(\Lambda^* + \sum_{j=1}^{n} Q_j^*B_jN_j^{-1}B_j^*\right)Q_i + Q_i^*\left(\Lambda + \sum_{j=1}^{n} B_jN_j^{-1}B_j^*Q_j\right) \\
&\quad = \frac{R_i}{\beta_{ii}} + \sum_{j=1}^{n} \beta_{ij}\frac{Q_j^*B_jN_j^{-1}B_j^*Q_j}{\beta_{ii}}.
\end{aligned}
\tag{3.4'}
$$

Donc (3.16) donne

$$
\begin{aligned}
(B_iw_i, Q_i\tilde{z}) &= \left(\tilde{z}, \frac{R_i}{\beta_{ii}}\tilde{z} + \sum_{j=1}^{n} \frac{\beta_{ij}Q_j^*B_jN_j^{-1}B_j^*Q_j}{\beta_{ii}}\tilde{z} \right. \\
&\qquad \left. - \Lambda^*Q_i\tilde{z} - \sum_{j=1}^{n} Q_j^*B_jN_j^{-1}B_j^*Q_i\tilde{z}\right) \\
&= \frac{1}{\beta_{ii}}(R_i\tilde{z}, \tilde{z}) + \frac{1}{\beta_{ii}}\left(\tilde{z}, \sum_{j=1}^{n} \beta_{ij}Q_j^*B_jN_j^{-1}B_j^*Q_j\tilde{z}\right) \\
&\qquad - (\Lambda\tilde{z}, Q_i\tilde{z}) - \left(\sum_{j=1}^{n} B_jN_j^{-1}B_j^*Q_j\tilde{z}, Q_i\tilde{z}\right) \\
&= \frac{1}{\beta_{ii}}(R_i\tilde{z}, \tilde{z}) + \frac{1}{\beta_{ii}}\left(\tilde{z}, \sum_{j=1}^{n} \beta_{ij}Q_j^*B_jN_j^{-1}B_j^*Q_j\tilde{z}\right) \\
&\qquad - (Q_i\tilde{z}, B_iw_i),
\end{aligned}
\tag{3.17}
$$

d'où

$$
2\beta_{ii}(Q_i\tilde{z}, B_iz_i) = (R_i\tilde{z}, \tilde{z}) + \left(\tilde{z}, \sum_{j=1} \beta_{ij}Q_j^*B_jN_j^{-1}B_j^*Q_j\tilde{z}\right).
\tag{3.18}
$$

Par ailleurs on a

$$
\begin{aligned}
&(B_iw_i, Q_iy) + (B_iw_i, \sigma_i) \\
(3.19)\qquad &= \left(Q_iy, \Lambda\tilde{z} + \sum_{j=1}^{n} B_jN_j^{-1}B_j^*Q_j\tilde{z}\right) + \left(\sigma_i, \Lambda\tilde{z} + \sum_{j=1}^{n} B_jN_j^{-1}B_j^*Q_j\tilde{z}\right) \\
&= \left(\tilde{z}, \left(\Lambda^*Q_i + \sum_{j=1}^{n} Q_j^*B_jN_j^{-1}B_j^*Q_i\right)y\right) \\
&\quad + \left(\tilde{z}, \Lambda^*\sigma_i + \sum_{j=1}^{n} Q_j^*B_jN_j^{-1}B_j^*\sigma_i\right).
\end{aligned}
$$

En utilisant (3.4′) et (3.5) on obtient

$$
\begin{aligned}
(B_iw_i, Q_iy + \sigma_i) &= \left(\tilde{z}, \frac{R_i}{\beta_{ii}}y\right) + \left(\tilde{z}, \sum_{j=1}^{n} \frac{\beta_{ij}Q_j^*B_jN_j^{-1}B_j^*Q_j}{\beta_{ii}}y\right) \\
&\quad - \left(\tilde{z}, Q_i^*\left(\Lambda + \sum_{j=1}^{n} B_jN_j^{-1}B_j^*Q_j\right)y\right) \\
(3.20)\qquad &\quad + (\tilde{z}, Q_i^*f) + \left(\tilde{z}, \sum_{j=1}^{n} \frac{\beta_{ij}Q_j^*B_jN_j^{-1}B_j^*\sigma_j}{\beta_{ii}}\right) \\
&\quad - \left(\tilde{z}, \sum_{j=1}^{n} Q_i^*B_jN_j^{-1}B_j^*\sigma_j\right).
\end{aligned}
$$

Et tenant compte de (3.6) il vient encore

$$
\begin{aligned}
(B_iw_i, Q_iy + \sigma_i) &= \frac{1}{\beta_{ii}}(\tilde{z}, R_iy) + \left(\tilde{z}, \sum_{j=1}^{n} \frac{\beta_{ij}Q_j^*B_jN_j^{-1}B_j^*Q_j}{\beta_{ii}}y\right) \\
&\quad - \left(Q_i\tilde{z}, f - \sum_{j=1}^{n} B_jN_j^{-1}B_j^*\sigma_j\right) \\
(3.21)\qquad &\quad + (\tilde{z}, Q_i^*f) + \left(\tilde{z}, \sum_{j=1}^{n} \frac{\beta_{ij}Q_j^*B_jN_j^{-1}B_j\sigma_j)}{\beta_{ii}}\right) \\
&\quad - \left(\tilde{z}, \sum_{j=1}^{n} Q_i^*B_jN_j^{-1}B_j^*\sigma_j\right).
\end{aligned}
$$

Soit

$$
(3.22)\qquad 2\beta_{ii}(B_iw_i, Q_iy + \sigma_i) = 2(\tilde{z}, R_iy) + 2\left(\tilde{z}, \sum_{j=1}^{n} \beta_{ij}Q_j^*B_jN_j^{-1}B_j^*(Q_jy + \sigma_j)\right).
$$

En utilisant (3.18) et (3.22), l'expression (3.15) donne

$$
(3.23)\qquad \bar{\bar{J}}_i(w_i) = \bar{\bar{J}}_i(0) + \beta_{ii}(N_iw_i, w_i).
$$

Comme $\beta_{ii} > 0$, on a donc

$$
(3.24)\qquad \bar{\bar{J}}_i(w_i) \geqq \bar{\bar{J}}_i(0).
$$

Ce qui signifie

$$
\bar{J}_i(\mathscr{L}_i) \geqq \bar{J}_i(\mathscr{L}_i)
$$

pour toute stratégie $\mathscr{L}_i$. D'autre part

$$\bar{\bar{J}}_i(w_i) = \bar{\bar{J}}_i(0) \quad \text{implique} \quad w_i(z) = 0$$

soit $z = y$ en raison de l'unicité de la solution de (3.6). Puis

$$\mathscr{L}_i(y) = \hat{\mathscr{L}}_i(y).[2] \tag{3.25}$$

Remarque 3.1. Il est intéressant de noter que $\mathscr{L}_i$ est différent de $\mathscr{L}_i^*$. Vérifions cependant que dans le cas de l'exemple (2.27), il y a coïncidence entre Q_i et P_i. En effet, d'après la Proposition 2.2, on a

$$P_i = \varepsilon_i P,$$

où P est solution de

$$P\Lambda + \Lambda^* P + P \sum_{j=1}^{n} \varepsilon_j B_j N_j^{-1} B_j^* P = M^* M. \tag{3.26}$$

L'equation (3.4') s'écrit

$$\begin{aligned} &\left(\Lambda^* + \sum_{j=1}^{n} Q_j^* B_j N_j^{-1} B_j^*\right) Q_i + Q_i^* \left(\Lambda + \sum_{j=1}^{n} B_j N_j^{-1} B_j^* Q_j\right) \\ &\quad = \varepsilon_i M^* M + \sum_{j=1}^{n} \varepsilon_i \varepsilon_j Q_j^* B_j N_j^{-1} B_j^* Q_j. \end{aligned} \tag{3.27}$$

Vérifions que $\varepsilon_i P$ est solution de (3.27). En effet, on doit avoir

$$\begin{aligned} &\left(\Lambda^* + \sum_{j=1}^{n} \varepsilon_j P B_j N_j^{-1} B_j^*\right) \varepsilon_i P + \varepsilon_i P \left(\Lambda + \sum_{j=1}^{n} \varepsilon_j B_j N_j^{-1} B_j^* P\right) \\ &\quad = \varepsilon_i M^* M + \sum_{j=1}^{n} \varepsilon_i \varepsilon_j P B_j N_j^{-1} B_j^* P, \end{aligned}$$

ce qui est évidemment vérifié grâce à (3.26).

On a ensuite $\rho_i = \sigma_i$. En effet $\rho_i = \varepsilon_i \rho$, où ρ est solution de

$$\Lambda^* \rho + P \sum_{j=1}^{n} \varepsilon_j B_j N_j^{-1} B_j^* \rho = Pf. \tag{3.28}$$

Or σ_i est solution de (cf. (3.5))

$$\Lambda^* \sigma_i + P \sum_{j=1}^{n} \varepsilon_j B_j N_j^{-1} B_j^* \sigma_i = \varepsilon_i Pf, \tag{3.29}$$

d'où $\sigma_i = \varepsilon_i \rho$.

Par conséquent dans ce cas $\hat{\mathscr{L}}_i = \mathscr{L}_i^*$, mais il n'y a pas égalité en général.

4. Etude du cas parabolique.

4.1. Notations. Le système que l'on considère est le suivant: soient V et H deux espaces de Hilbert avec

$$V \subset H \tag{4.1}$$

[2] Théoriquement $\mathscr{L}_i$ peut-être différent de $\hat{\mathscr{L}}_i$ mais coïncide avec $\hat{\mathscr{L}}_i$ au point y. Par conséquent $\mathscr{L}_i$ ne peut être distinguée de $\hat{\mathscr{L}}_i$, puisqu'elles conduisent au même état et au même contrôle. On peut dire que $\mathscr{L}_i$ est équivalent à $\hat{\mathscr{L}}_i$. Il y a donc unicité à une équivalence près.

V dense dans H et injection compacte. On identifie H et son dual et on note V' le dual de V. Alors

$$V \subset H \subset V'. \tag{4.2}$$

Chaque espace étant dense dans le suivant, avec injection continue. Soit $t \in [0, T]$.

On se donne une famille $A(t) \in L^\infty(0, T, \mathscr{L}(V, V'))$ et vérifiant: p.p. $D(A(t)) \supset D$ dense dans H,

$$\begin{aligned}&\text{il existe} \quad \lambda \geqq 0 \quad \text{tel que} \quad \langle A(t)z, z\rangle + \lambda|z|^2 \geqq \alpha\|z\|,^3\\ &\qquad\qquad\qquad \alpha > 0, \quad \text{pour tout} \quad z \in V, \quad \text{p.p.}\, t.\end{aligned} \tag{4.3}$$

On considère alors le système décrit par l'équation d'évolution:

$$\begin{aligned}&\frac{dy}{dt} + A(t)y = f(t) + \sum_{i=1}^{n} B_i(t)v_i(t),\\ &y(0) = y_0 \in H,\end{aligned} \tag{4.4}$$

où

$$\begin{aligned}&f \in L^2(0, T, H), \quad B_i \in L^\infty(0, T, \mathscr{L}(\mathscr{U}_i, H))\\ &v_i(t) \ \text{le contrôle} \in L^2(0, T, \mathscr{U}_i).\end{aligned} \tag{4.5}$$

Les fonctionnelles de Coût sont définies par les formules:[4]

$$J_i(v) = \int_0^T (R_i(t)y(t), y(t))\,dt + \sum_{j=1}^{n} \int_0^T \beta_{ij}(t)(N_j(t)v_j(t), v_j(t))\,dt, \tag{4.6}$$

où $R_i(t) \in L^\infty(0, T, \mathscr{L}(H, H))$, *symétrique* (mais pas forcément $\geqq 0$). Les $N_j(t) \in L^\infty(0, T, \mathscr{L}(\mathscr{U}_j, \mathscr{U}_j))$ et sont autoadjoints, définis >0 et inversibles (pour presque tout t). On prendra

$$\beta_{ii}(t) = 1. \tag{4.7}$$

4.2. Point de Nash en boucle fermée. On considère le problème de la recherche d'un point de Nash pour les fonctionnelles $J_i(v)$. L'analogue des équations (3.4), (3.5), (3.6) devient ici

$$\begin{aligned}&-\frac{d}{dt}(Q_i\varphi) + Q_i^*(t)\left(\frac{d\varphi}{dt} + A(t)\varphi + \sum_{j\neq i} B_j(t)N_j^{-1}(t)B_j^*(t)Q_j(t)\varphi\right)\\ &+ (A^*(t) + \sum_{j\neq i} Q_j^*(t)B_j(t)N_j^{-1}(t)B_j^*(t))Q_i(t)\varphi(t)\\ &+ Q_i^*(t)B_i(t)N_i^{-1}(t)B_i^*(t)Q_i(t)\varphi(t)\\ &\qquad = R_i(t)\varphi + \sum_{j\neq i} \beta_{ij}Q_j^*(t)B_j(t)N_j^{-1}(t)B_j^*(t)Q_j(t)\varphi(t)\\ &\qquad \text{pour tout} \quad \varphi \in W = \left\{\varphi \in L^2(0, T, V)\left|\frac{d\varphi}{dt} + A(t)\varphi \in L^2(0, T, H)\right.\right\},\\ &\qquad\qquad\qquad Q_i(T) = 0,\end{aligned} \tag{4.8}$$

[3] $\langle, \rangle$ désigne la dualité V, V', $|\cdot|$ désigne la norme dans H et $\|\cdot\|$désigne la norme dans V. On note $W(0, T) = \{\varphi \in L^2(0, T, V)|d\varphi/dt \in L^2(0, T, H)\}$.

[4] Où v désigne le vecteur $v_1(\cdot), \cdots, v_n(\cdot)$.

$$\tag{4.9}\begin{aligned}&-\frac{d\sigma_i}{dt} + \left(A^*(t) + \sum_{j=1}^{n} Q_j^*(t)B_j(t)N_j^{-1}(t)B_j^*(t)\right)\sigma_i\\&\qquad = Q_i^*(t)f(t) + \sum_{j\neq i}(\beta_{ij}(t)Q_j^*(t) - Q_i^*(t))B_j(t)N_j^{-1}(t)B_j^*(t)\sigma_j(t),\\&\sigma_i(T) = 0,\\&\sigma_i \in L^2(T, 0; V),\end{aligned}$$

$$\tag{4.10}\begin{aligned}&\frac{dy}{dt} + \left(A(t) + \sum_{j=1}^{n} B_j(t)N_j^{-1}(t)B_j^*(t)Q_j(t)\right)y\\&\qquad = f(t) - \sum_{j=1}^{n} B_j(t)N_j^{-1}(t)B_j^*(t)\sigma_j(t),\\&\qquad y(0) = y_0,\\&\qquad y \in L^2(0, T; V).\end{aligned}$$

On peut alors obtenir un analogue du Théorème 3.1.

THÉORÈME 4.1. *Sous l'hypothèse que les équations de Riccati* (4.8) *possèdent une solution et* $Q_i \in L^\infty(0, T, \mathscr{L}(H, H))$ *il existe un point de Nash unique en boucle fermée pour les fonctionnelles* J_i *définies en* (4.6).

Démonstration. Tout d'abord le système (4.9) possède une solution unique. On peut en effet le réécrire sous forme vectorielle en posant

$$\tag{4.11}\boldsymbol{\sigma} = \begin{pmatrix}\sigma_1\\ \vdots\\ \sigma_n\end{pmatrix}, \quad \mathbf{H} = H^n, \quad \mathbf{V} = V^n, \quad \mathbf{V}' = V'^n,$$

puis $\mathbf{A}(t) \in \mathscr{L}(\mathbf{V}, \mathbf{V}')$ defini par

$$\tag{4.12}\mathbf{A}(t)\boldsymbol{\sigma} = \begin{pmatrix}A(t)\sigma_1\\ \vdots\\ A(t)\sigma_n\end{pmatrix},$$

$\mathbf{Q}^*(t) \in \mathscr{L}(\mathbf{H}, \mathbf{H})$ défini par

$$\tag{4.13}\mathbf{Q}(t)\boldsymbol{\sigma} = \begin{pmatrix}\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ \sum_{j=1}^{n} Q_j^*B_jN_j^{-1}B_j^*\sigma_i + \sum_{j\neq i}^{n}(Q_i^* - \beta_{ij}Q_j^*)B_jN_j^{-1}B_j^*\sigma_j\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\end{pmatrix},$$

$$\mathbf{f}(t) = \begin{pmatrix}Q_1^*f\\ \vdots\\ Q_n^*f\end{pmatrix}.$$

Alors (4.9) peut s'écrire

$$\begin{aligned} &-\frac{d\boldsymbol{\sigma}}{dt} + (\mathbf{A}^*(t) + Q^*(t))\boldsymbol{\sigma} = \mathbf{f}, \\ &\boldsymbol{\sigma}(T) = 0. \end{aligned} \tag{4.14}$$

On vérifie aisèment que l'on peut appliquer à (4.14) la théorie générale des équations différentielles opérationnelles de Lions [4], d'où existence et unicité de $\boldsymbol{\sigma} \in L^2(0, T; \mathrm{V})$ (donc $d\boldsymbol{\sigma}/dt \in L^2(0, T; \mathrm{V}')$, ce qui implique $\boldsymbol{\sigma} \in \subset ([0, T]; \mathbf{H})$.

Il en résulte aussitôt que l'équation (4.10) possède une solution unique. Le reste de la démonstration se fait comme pour le Théorème 3.1.

4.3. Etude des équations de Riccati.

Orientation. Pour les points de Nash en boucle ouverte, un moyen de justifier les équations de Riccati que l'on trouve (cf. (2.31)) est d'utiliser la méthodologie du découplage mise au point par J. L. Lions [4] dans le cas du contrôle optimal. Cela suppose que soit satisfaite la C.N.S. d'existence du point de Nash (cf. Théorème 2.1).

Par contre, pour les points de Nash en boucle fermée, il est indispensable de pouvoir étudier directement les équations de Riccati (4.8). Nous connaissons déjà un cas où les équations (4.8) possèdent une solution, c'est le cas de l'exemple (2.27) (cf. Remarque 3.1). Dans ce cas, la solution "boucle ouverte découplée" et la solution "boucle fermée" coïncident (mais cela suppose que soit satisfaite la condition (2.29)).

On va s'intéresser maintenant à un autre cas où on peut étudier directement les équations (4.8), ce qui prouvera l'existence d'un point de Nash en boucle fermée, sans pour autant qu'il existe un point de Nash en boucle ouverte.

4.3.1. Hypothèses—Enoncé du Théorème. On prend

$$R_i \geqq 0 \quad \text{pour tout } i, \tag{4.15}$$

$$\beta_{ij} = 0 \quad \text{si} \quad j \neq i. \tag{4.16}$$

On a alors le théorème suivant.

THÉORÈME 4.2. *Si* (4.15) *et* (4.16) *sont satisfaites, alors il existe pour T assez petit une solution pour les équations* (4.8) *vérifiant*:

$$\begin{aligned} &Q_i \in L^\infty(0, T, \mathscr{L}(H, H)), \\ &Q_i^*(t) = Q_i(t) \quad p.p., \\ &Q_i(t) \geqq 0, \end{aligned} \tag{4.17}$$

$$\begin{aligned} &si \quad \varphi \in L^2(0, T; V) \left| \frac{d\varphi}{dt} + A(t)\varphi \in L^2(0, T; H), \right. \quad alors \\ &Q_i\varphi \in L^2(0, T; V), \quad \frac{dQ_i\varphi}{dt} \in L^2(0, T; V'). \end{aligned} \tag{4.18}$$

Remarque 4.1. Les conditions (4.17) et (4.18) font que l'équation (4.8) a un sens, p.p. t en tant qu'égalité de deux éléments de V'.

4.3.2. Démonstration du Théorème 4.2. On généralise tres légèrement en introduisant une condition initiale. On se place sur un intervalle (T_0, T_1) (au lieu de $(0, T)$). On pose $l = T_1 - T_0$, et on se donne

$$Q_i(T_1) = Q_i \in \mathscr{L}(H; H) \quad \text{autoadjoint} \quad \geqq 0. \tag{4.19}$$

On va déterminer une borne supérieure de l au dessous de laquelle le théorème d'existence est vérifié.

On construit alors une suite d'approximations de Q_i par le procédé suivant: On part de $Q_i^0(t) = Q_i$ et on définit Q_i^{k+1} en fonction de Q_i^k par l'équation de Riccati:

$$\begin{gathered}
-\frac{d}{dt}(Q_i^{k+1}\varphi) + Q_i^{k+1}\left(\frac{d\varphi}{dt} + A\varphi + \sum_{j\neq i} B_jN_j^{-1}B_j^*Q_j^k\varphi\right) \\
+\left(A^* + \sum_{j\neq i} Q_j^kB_jN_j^{-1}B_j^*\right)Q_i^{k+1} + Q_i^{k+1}B_iN_i^{-1}B_i^*Q_i^{k+1}\varphi = R_i\varphi \qquad (4.20)\\
\text{pour tout} \quad \varphi \in L^2(T_0, T_1; V) \left| \frac{d\varphi}{dt} + A(t)\varphi \in L^2(T_0, T_1; H),\right. \\
Q_i^{k+1}(T_1) = Q_i.
\end{gathered}$$

En effet (4.20) est une équation de Riccati de type classique (où A est changé en $A + \sum_{j\neq i} B_jN_j^{-1}B_j^*Q_j^k$). Elle définit Q_i^{k+1} de manière unique vérifiant (4.17) (sur (T_0, T_1)). Par ailleurs (4.8) est liée à la solution du problème de contrôle optimal suivant:

Le système est régi par

$$\begin{gathered}
\frac{dz}{dt} + \left(A + \sum_{j\neq i} B_jN_j^{-1}B_j^*Q_j^k\right)z = B_iv_i, \\
z(s) = h \in H, \quad s \in [T_0, T_1],
\end{gathered} \tag{4.21}$$

et le critère est donné par

$$J_k^i(s, h, v_i) = \int_s^{T_1} (R_iz, z)\,dt + \int_s^{T_1} (N_iv_i, v_i)\,dt + (\overline{Q}_iz(T_1), z(T_1)). \tag{4.22}$$

Alors l'équation d'Euler correspondant à la minimisation de (4.22) conduit au système d'équations:

$$\begin{gathered}
\frac{d\varphi_i^{k+1}}{dt} + \left(A + \sum_{j\neq i} B_jN_j^{-1}B_j^*Q_j^k\right)\varphi_i^{k+1} + B_iN_i^{-1}B_i^*\psi_i^{k+1} = 0, \\
t \in]s, T_1[, \\
-\frac{d\psi_i^{k+1}}{dt} + \left(A^* + \sum_{j\neq i} Q_j^kB_jN_j^{-1}B_j^*\right)\psi_i^{k+1} - R_i\varphi_i^{k+1} = 0, \qquad (4.23)\\
\psi_i^{k+1}(T_1) = \overline{Q}_i\varphi_i^{k+1}(T_1), \\
\varphi_i^{k+1}(s) = h,
\end{gathered}$$

et on a

$$\psi_i^{k+1}(s) = Q_i^{k+1}(s)h, \tag{4.24}$$

ce qui définit les opérateurs $Q_i^{k+1}(s)$ en tout point s. Plus généralement si on considère le système:

$$\begin{aligned} &\frac{d\beta_i^{k+1}}{dt} + \left(A + \sum_{j\neq i} B_jN_j^{-1}B_j^*Q_j^k\right)\beta_i^{k+1} + B_iN_i^{-1}B_i^*\gamma_i^{k+1} = g_1, \\ &-\frac{d\gamma_i^{k+1}}{dt} + \left(A^* + \sum_{j\neq i} Q_j^kB_jN_j^{-1}B_j^*\right)\gamma_i^{k+1} - R_i\beta_i^{k+1} = g_2, \\ &\gamma_i^{k+1}(T_1) = \overline{Q}_i\beta_i^{k+1}(T_1), \\ &\beta_i^{k+1}(s) = h, \end{aligned} \tag{4.25}$$

alors on a

$$\gamma_i^{k+1}(t) = Q_i^{k+1}(t)\beta_i^{k+1}(t) + r_i^{k+1}(t) \quad \text{pour tout} \quad t\in[s, T_1], \tag{4.26}$$

où $r_i^{k+1}(t)$ est solution de

$$\begin{aligned} &-\frac{dr_i^{k+1}}{dt} + \left(A^* + \sum_{j\neq i} Q_j^kB_jN_j^{-1}B_j^* + Q_i^{k+1}B_iN_i^{-1}B_i^*\right)r_i^{k+1} = g_2 + Q_i^{k+1}g_1, \\ &r_i^{k+1}(T_1) = 0 \end{aligned} \tag{4.27}$$

(pour les détails, cf. J. L. Lions [4]).

On a alors le lemme suivant.

LEMME 4.1. *On peut trouver un nombre* $K > 0$ *tel que*

$$\|Q_i^k(t)\|_{\mathscr{L}(H,H)} \leqq K \Rightarrow \|Q_i^{k+1}(t)\|_{\mathscr{L}(H,H)} \leqq K \ p.p.t, \tag{4.28}$$

le nombre K *étant indépendant de* k *et* i.

Démonstration. Un calcul classique (cf. J. L. Lions [4]) montre que l'on a

$$\inf_{v_i} J_k^i(s, h, v_i) = (Q_i^{k+1}(s)h, h). \tag{4.29}$$

Par conséquent si on considère l'équation

$$\begin{aligned} &\frac{d\xi}{dt} + \left(A + \sum_{j\neq i} B_jN_j^{-1}B_j^*Q_j^k\right)\xi = 0, \\ &\xi(s) = h, \end{aligned} \tag{4.30}$$

alors on a

$$(Q_i^{k+1}(s)h, h) \leqq J_k^i(s, h, o) = \int_s^{T_1} (R_i\xi, \xi)\,dt + (\overline{Q}_i\xi(T_1), \xi(T_1)). \tag{4.31}$$

Posons

$$\begin{aligned} C_1 &= \sum_j \frac{\|B_j\|^2}{v_j} \quad (\text{où} \quad N_j \geqq v_jI), \\ C_2 &= \max_i \|R_i\|, \\ C_3 &= \|\overline{Q}_i\|. \end{aligned} \tag{4.32}$$

Soit θ un nombre à déterminer. Introduisons

$$\zeta(t) = e^{-\theta t}\xi(t) \tag{4.33}$$

de sorte que ζ est solution de l'equation

$$\begin{gathered}\frac{d\zeta}{dt} + \left(A + \theta I + \sum_{j\neq i} B_j N_j^{-1} B_j^* Q_j^k\right)\zeta = 0,\\ \zeta(s) = h\, e^{-\theta s}.\end{gathered} \tag{4.34}$$

Posons

$$\alpha = \alpha_1 + \alpha_2. \tag{4.35}$$

On déduit de (4.35) la classique égalité de l'énergie

$$\begin{gathered}|\zeta(T_1)|^2 + \int_s^{T_1} \langle A\zeta, \zeta\rangle\, dt + \theta \int_s^{T_1} |\zeta|^2\, dt\\ + \int_s^{T_1} \left(\zeta, \sum_{j\neq i} B_j N_j^{-1} B_j^* Q_j^k \zeta\right) dt = e^{-2\theta s}|h|^2,\end{gathered} \tag{4.36}$$

d'où on déduit si $\|Q_j^k\| \leqq K$,

$$|\zeta(T_1)|^2 + \alpha_1 \int_s^{T_1} \|\zeta\|^2\, dt + (\alpha_2 + \theta - KC_1) \int_s^{T_1} |\zeta|^2\, dt \leqq e^{-2\theta s}|h|^2. \tag{4.37}$$

Posons

$$\varepsilon = \alpha_2 + \theta - KC_1, \quad \varepsilon' = KC_1, \tag{4.38}$$

et on impose comme première contrainte $\varepsilon > 0, \varepsilon' > 0$. On déduit de (4.37),

$$\int_s^{T_1} |\zeta|^2\, dt \leqq \frac{e^{-2\theta s}}{\varepsilon}|h|^2, \qquad |\zeta(T_1)|^2 \leqq e^{-2\theta s}|h|^2. \tag{4.39}$$

Mais alors (4.31) donne

$$\begin{aligned}(Q_i^{k+1}(s)h, h) &\leqq C_2 \int_s^{T_1} e^{2\theta t}|\zeta|^2\, dt + C_3\, e^{2\theta T_1}|\zeta(T_1)|^2\\ &\leqq \max\left(e^{2\theta T_0}, e^{2\theta T_1}\right)\left\{\frac{C_2\, e^{-2\theta s}}{\varepsilon} + C_3\, e^{-2\theta s}\right\}|h|^2\\ &\leqq \max\left(e^{2\theta T_0}, e^{2\theta T_1}\right) \max\left(e^{-2\theta T_0}, e^{-2\theta T_1}\right)\left(\frac{C_2}{\varepsilon} + C_3\right)|h|^2\\ &= \left(\frac{C_2}{\varepsilon} + C_3\right) e^{2|\theta|(T_1 - T_0)}|h|^2.\end{aligned} \tag{4.40}$$

Par conséquent, grâce au fait que $Q_i^{k+1}(s)$ est $\geqq 0$ autoadjoint ou déduit de (4.40)

$$\|Q_i^{k+1}(s)\| \leqq \left(\frac{C_2}{\varepsilon} + C_3\right) e^{2|\varepsilon + \varepsilon' - \alpha_2|l}. \tag{4.41}$$

Il s'agit donc de montrer que l'estimation (4.41) peut être rendue $\leqq K$. Autrement dit, il faut montrer que l'on peut trouver $\varepsilon, \varepsilon' > 0$ et l (qui est à notre disposition) tels que l'on ait:

$$C_1\left(\frac{C_2}{\varepsilon} + C_3\right) e^{2|\varepsilon+\varepsilon'-\alpha_2|l} \leqq \varepsilon'. \tag{4.42}$$

Posons

$$x = \frac{\varepsilon'}{\varepsilon}, \quad x > 0 \quad \textit{arbitraire},$$

et choisissons ε de façon que

$$\varepsilon^2 x - \varepsilon C_1 C_3 - C_1 C_2 > 0. \tag{4.43}$$

Il suffit de choisir $\varepsilon >$ racine positive de l'équation du second degré (4.43). Ayant ainsi choisi ε, on choisit α_2, tel que

$$0 < \alpha_2 < \min(\alpha, \varepsilon(1 + x)). \tag{4.44}$$

Alors $\varepsilon + \varepsilon' - \alpha_2 > 0$. On choisit ensuite l de façon que

$$0 < l < \frac{1}{2[\varepsilon(1 + x) - \alpha_2]} \log \frac{\varepsilon^2 x}{C_1 C_2 + C_1 C_3 \varepsilon}. \tag{4.45}$$

Grâce à (4.43) l'argument du log est >1 de sorte que (4.45) est possible. Mais alors, on voit immédiatement que (4.41) est vérifiée. On remarque aussi que

$$\|Q_i^0(t)\| \leqq \|\overline{Q}_i\| = C_3 < \left(\frac{C_2}{\varepsilon} + C_3\right) e^{2[\varepsilon(1+x)-\alpha_2]} \leqq \frac{\varepsilon'}{C_1} = K.$$

Ce qui achève la démonstration du Lemme 4.1.

Démontrons également le lemme suivant.

LEMME 4.2. *Pour*

$$\varphi \in \left\{ z \in L^2(T_0, T_1, V) \,\middle|\, \frac{dz}{dt} + Az \in L^2(T_0, T_1, H) \right\}$$

la suite $Q_i^k(t)\varphi(t)$ demeure dans un borné de

$$W(T_0, T_1) = \left\{ z \in L^2(T_0, T_1, V) \,\middle|\, \frac{dz}{dt} \in L^2(T_0, T_1, V') \right\}.$$

Démonstration. Considérons le système d'équations:

$$\begin{aligned} &\frac{d\varphi}{dt} + \left(A(t) + \sum_{j \neq i} B_j(t) N_j^{-1}(t) B_j^* Q_j^k\right)\varphi + B_i N_i^{-1} B_i^* \psi_i^{k+1} = g_{k+1,i}^{\varphi}, \\ &-\frac{d\psi_i^{k+1}}{dt} + \left(A^* + \sum_{j \neq i} Q_j^k B_j N_j^{-1}(t) B_j^*\right)\psi_i^{k+1} - R_i\varphi = 0, \\ &\psi_i^{k+1}(T_1) = \overline{Q}_i \varphi(T_1), \\ &\varphi(T_0) = \varphi(T_0), \end{aligned} \tag{4.46}$$

que l'on doit interpréter de la manière suivante: φ est donné, on calcule ensuite ψ_i^{k+1} en résolvant les deuxième et troisième équations (4.46). On définit ensuite

$g^{\varphi}_{k+1,i} \in L^2(T_0, T_1, H)$ par la première équation (4.46). En comparant (4.46) et (4.25) on voit aussitôt que la formule (4.42) s'applique, d'où

$$\psi_i^{k+1}(t) = r_i^{k+1}(t) + Q_i^{k+1}(t)\varphi(t) \quad \text{pour tout} \quad t \in [T_0, T_1], \tag{4.47}$$

où $r_i^{k+1}(t)$ est solution de

$$\begin{aligned} -\frac{dr_i^{k+1}}{dt} + \left(A^* + \sum_{j\neq i} Q_j^k B_j N_j^{-1} B_j^* + Q_i^{k+1} B_i N_i^{-1} B_i^*\right) r_i^{k+1} \\ = Q_i^{k+1}(t) g^{\varphi}_{k+1,i}, \\ r_i^{k+1}(T_1) = 0. \end{aligned} \tag{4.48}$$

Pour montrer que $Q_i^{k+1}\varphi$ demeure dans un borné de $W(T_0, T_1)$, il suffit d'après (4.47) de montrer φ_i^{k+1} et r_i^{k+1} demeurent effectivement dans des bornés de $W(T_0, T_1)$. On pose alors

$$\tilde{\psi}_i^{k+1}(t) = e^{\theta t}\psi_i^{k+1}(t), \tag{4.49}$$

où θ a la même valeur que dans le Lemme 4.1. Par conséquent $\tilde{\psi}_i^{k+1}(t)$ est solution de

$$\begin{aligned} -\frac{d\tilde{\psi}_i^{k+1}}{dt} + \left(A^* + \theta I + \sum_{j\neq i} Q_j^k B_j N_j^{-1} B_j^*\right)\tilde{\psi}_i^{k+1} = e^{\theta t} R_i \varphi, \\ \tilde{\psi}_i^{k+1}(T_1) = e^{\theta T_1}\overline{Q}_i \varphi(T_1). \end{aligned} \tag{4.50}$$

On en déduit par intégration par parties,

$$\begin{aligned} |\tilde{\psi}_i^{k+1}(T_0)|^2 + \int_{T_0}^{T_1} \langle A^*\tilde{\psi}_i^{k+1}, \tilde{\psi}_i^{k+1}\rangle\, dt \\ + \theta\int_{T_0}^{T_1} |\tilde{\psi}_i^{k+1}(t)|^2\, dt + \int_{T_0}^{T_1}\left(\sum_{j\neq i} Q_j^k B_j N_j^{-1} B_j^* \tilde{\psi}_i^{k+1}, \tilde{\psi}_i^{k+1}\right) dt \\ = e^{2\theta T_1}|\overline{Q}_i\varphi(T_1)|^2 + \int_{T_0}^{T_1} e^{\theta t}(R_i\varphi, \tilde{\psi}_i^{k+1})\, dt, \end{aligned} \tag{4.51}$$

d'où

$$\begin{aligned} \alpha_1 \int_{T_0}^{T_1} \|\tilde{\psi}_i^{k+1}(t)\|^2\, dt + (\alpha_2 + \theta - KC_1)\int_{T_0}^{T_1} |\tilde{\psi}_i^{k+1}(t)|^2\, dt \\ \leqq e^{2\theta T_1}|\overline{Q}_i\varphi(T_1)|^2 + \sqrt{\int_{T_0}^{T_1} e^{2\theta t}|R_i\varphi|^2\, dt}\;\sqrt{\int_{T_0}^{T_1} |\tilde{\psi}_i^{k+1}(t)|^2\, dt}\,. \end{aligned} \tag{4.52}$$

Tenant compte des valeurs (4.38), (4.45), on obtient

$$\begin{aligned} \alpha_1 \int_{T_0}^{T_1} \|\tilde{\psi}_i^{k+1}(t)\|^2\, dt + \frac{1}{2l}\int_{T_0}^{T_1} |\tilde{\psi}_i^{k+1}(t)|^2\, dt \\ \leqq e^{2T_1(1/l-\alpha_2)} C_3 |\varphi(T_1)|^2 \\ + e^{T_1(1/l-\alpha_2)} C_2 \sqrt{\int_{T_0}^{T_1} |\varphi|^2\, dt}\;\sqrt{\int_{T_0}^{T_1} |\tilde{\psi}^{k+1}(t)|^2\, dt}, \end{aligned} \tag{4.53}$$

d'où il résulte

$$\sqrt{\int_{T_0}^{T_1} |\tilde{\psi}^{k+1}(t)|^2\,dt} \leqq e^{T_1(1/l-\alpha_2)}\left\{lC_2\sqrt{\int_{T_0}^{T_1} |\varphi|^2\,dt} + \sqrt{l^2C_2^2\int_{T_0}^{T_1} |\varphi|^2\,dt + 2lC_3|\varphi(T_1)|^2}\right\} \tag{4.54}$$

et

$$\alpha_1\int_{T_0}^{T_1} \|\tilde{\psi}_i^{k+1}(t)\|^2\,dt \leqq e^{2T_1(1/l-\alpha_2)}\left\{C_3|\varphi(T_1)|^2 + lC_2^2\int_{T_0}^{T_1} |\varphi|^2\,dt + C_2\sqrt{\int_{T_0}^{T_1} |\varphi|^2\,dt}\sqrt{l^2C_2^2\int_{T_0}^{T_1} |\varphi|^2\,dt + 2lC_3|\varphi(T_1)|^2}\right\}. \tag{4.55}$$

Mais (4.49) donne puisque $\theta > 0$,

$$\int_{T_0}^{T_1} \|\psi_i^{k+1}(t)\|_V^2\,dt \leqq \int_{T_0}^{T_1} \|\tilde{\psi}_i^{k+1}(t)\|_V^2\,dt, \tag{4.56}$$

ce qui avec (4.55) montre bien que ψ_i^{k+1} demeure dans un borné de $L^2(T_0, T_1, V)$. La deuxième équation (4.46) donne alors, en tenant compte du Lemme 4.1,

$$\left\|\frac{d\psi_i^{k+1}}{dt}\right\|_{V'} = \left\| - R_i\varphi + A^*\psi_i^{k+1} + \sum_{j\neq i} Q_j^kB_jN_j^{-1}B_j^*\psi_i^{k+1}\right\|_{V'}$$
$$\leqq C_2|\varphi|_H + \|A\|_{\mathscr{L}(V,V')}\|\psi_i^{k+1}\|_V + KC_1|\psi_i^{k+1}|_H.$$

Donc

$$\int_{T_0}^{T_1}\left\|\frac{d\psi_i^{k+1}}{dt}\right\|_{V'}^2 dt \leqq 3\left(C_2^2\int_{T_0}^{T_1} |\varphi|^2\,dt + \|A\|^2\int_{T_0}^{T_1} \|\varphi_i^{k+1}\|^2\,dt + \frac{1}{l}\int_{T_0}^{T_1} |\psi_i^{k+1}|_H^2\,dt\right). \tag{4.57}$$

Par conséquent $d\psi_i^{k+1}/dt$ demeure bien dans un borné de $L^2(T_0, T_1, V')$ et donc ψ_i^{k+1} demeure dans un borné de $W(T_0, T_1)$. Par ailleurs de la première équation (4.46) on déduit

$$\int_{T_0}^{T_1} |g_{k+1,i}^{\varphi}|_H^2\,dt \leqq 3\int_{T_0}^{T_1} \left|\frac{d\varphi}{dt} + A(t)\varphi\right|_H^2 dt + 3KC_1\int_{T_0}^{T_1} |\varphi|^2\,dt + 3C_1\int_{T_0}^{T_1} |\psi_i^{k+1}|^2\,dt \tag{4.58}$$

et donc $g_{k+1,i}^{\varphi}$ demeure dans un borné de $L^2(T_0, T_1, H)$. Il en est de même pour $Q_i^{k+1}(t)g_{k+1,i}^{\varphi}(t)$ puisque $\|Q_i^{k+1}(t)\| \leqq K$. Par des calculs analogues à ceux faits pour ψ_i^{k+1}, on montre que r_i^{k+1} appartient également à un borné de $W(T_0, T_1)$, ce qui achève la démonstration du Lemme 4.2.

4.3.3. Propositions préliminaires. Avant de passer à l'étude la différence $\Sigma_i^k = Q_i^{k+1} = Q_i^k$, nous allons donner quelques résultats utiles concernant les équations de Riccati dont le second membre n'est pas défini positif. Ces résultats nous seront utiles par la suite.

Considérons le système (notations standard):

$$\frac{dz}{dt} + \mathscr{A}(t)z = Bv \quad \text{sur} \quad]T_0, T_1[, \qquad z(T_0) = z_0, \tag{4.59}$$

et le critère

$$J(v) = \int_{T_0}^{T_1} (Nv, v)\, dt + \int_{T_0}^{T_1} (Sz, z)\, dt. \tag{4.60}$$

On suppose que S est symétrique mais non $\geqq 0$ et que $\mathscr{A}(t) \in \mathscr{L}(V, V')$ vérifie:

(4.61) il existe λ et β tels que $\langle \mathscr{A}(t)z, z\rangle + \lambda|z|^2 \geqq \beta\|z\|^2$ pour tout $z \in V$.

On posera

$$\beta = \beta_1 + \beta_2. \tag{4.62}$$

En général il n'existe pas de contrôle optimal pour le problème (4.59), (4.60). Il est nécessaire et suffisant pour cela que la forme quadratique $J(v)$ soit coercive. Posons

$$z = \bar{z} + \tilde{z} \tag{4.63}$$

ou

$$\frac{d\bar{z}}{dt} + \mathscr{A}(t)\bar{z} = 0,$$

$$\bar{z}(T_0) = z_0,$$

$$\frac{d\tilde{z}}{dt} + \mathscr{A}(t)z = Bv,$$

$$\tilde{z}(T_0) = 0.$$

Alors on a

$$J(v) = \int_{T_0}^{T_1} (Nv, v)\, dt + \int_{T_0}^{T_1} (S\tilde{z}, \tilde{z})\, dt + 2\int_{T_0}^{T_1} (S\bar{z}, \tilde{z})\, dt + \int_{T_0}^{T_1} (S\bar{z}, \bar{z})\, dt.$$

Cherchons un minorant du terme de plus haut degré en v dans $J(v)$. Dans (4.59) on fait le changement de fonction

$$\tilde{z} = \tilde{\zeta}\, e^{\tau t},$$

d'où

$$\frac{d\tilde{\zeta}}{dt} + (\tau + \mathscr{A}(t))\tilde{\zeta} = e^{-\tau t}Bv,$$

$$\tilde{\zeta}(T_0) = 0,$$

et par intégration par parties (utilisant le fait que

$$\int_{T_0}^{T_1} \langle \mathscr{A}(t)\tilde{\zeta}, \tilde{\zeta} \rangle \, dt \geqq \beta_1 \int_{T_0}^{T_1} \|\tilde{\zeta}\|^2 \, dt + \beta_2 \int_{T_0}^{T_1} |\tilde{\zeta}|^2 \, dt - \lambda \int_{T_0}^{T_1} |\tilde{\zeta}|^2 \, dt)$$

on obtient

$$\begin{aligned} |\tilde{\zeta}(T_1)|^2 + \beta_1 \int_{T_0}^{T_1} \|\tilde{\zeta}\|^2 + (\beta_2 + \tau - \lambda) \int_{T_0}^{T_1} |\tilde{\zeta}|^2 \, dt & \\ \leqq \int_{T_0}^{T_1} (e^{-\tau t} Bv, \tilde{\zeta}) \, dt & \\ \leqq \sqrt{\int_{T_0}^{T_1} |\tilde{\zeta}(t)|^2 \, dt} \sqrt{\int_{T_0}^{T_1} e^{-2\tau t} |Bv|^2 \, dt}. & \end{aligned}$$

En posant

$$\eta = \beta_2 + \tau - \lambda, \qquad \eta > 0, \tag{4.64}$$

on obtient

$$\int_{T_0}^{T_1} |\tilde{\zeta}(t)|^2 \, dt \leqq \frac{1}{\eta^2} \max(e^{-2\tau T_0}, e^{-2\tau T_1}) \|B\|^2 \int_{T_0}^{T_1} |v|^2 \, dt. \tag{4.65}$$

Par conséquent

$$\begin{aligned} \int_{T_0}^{T_1} |(S\tilde{z}, \tilde{z})| \, dt \leqq \|S\| \int_{T_0}^{T_1} e^{2\tau t} |\tilde{\zeta}(t)|^2 \, dt & \\ \leqq \|S\| \max(e^{2\tau T_0}, e^{2\tau T_1}) \int_{T_0}^{T_1} |\tilde{\zeta}(t)|^2 \, dt, & \end{aligned}$$

d'où finalement

$$\int_{T_0}^{T_1} |(S\tilde{z}, \tilde{z})| \, dt \leqq \|S\| \, e^{2l|\tau|} \frac{\|B\|^2}{\eta^2} \int_{T_0}^{T_1} |v|^2 \, dt$$

Il en résulte si $N \geqq \nu I$,

$$\int_{T_0}^{T_1} (Nv, v) \, dt + \int_{T_0}^{T_1} (S\tilde{z}, \tilde{z}) \, dt \geqq \left(\nu - \frac{\|S\| \, \|B\|^2 \, e^{2l|\tau|}}{\eta^2} \right) \int_{T_0}^{T_1} |v|^2 \, dt$$

et donc on obtient la proposition suivante.

PROPOSITION 4.1. *Si la condition*

$$\rho = \nu - \frac{\|S\| \, \|B\|^2 \, e^{2l|\tau|}}{\eta^2} > 0 \tag{4.66}$$

est vérifiée, alors le problème (4.59), (4.60) *possède une solution unique.*

On peut alors écrire les équations d'Euler relatives à la minimisation de (4.60). On obtient (calcul classique, cf. Lions [4])

$$\begin{gathered}\frac{dz}{dt} + \mathscr{A}(t)z + BN^{-1}B^*q = 0,\\ -\frac{dq}{dt} + \mathscr{A}^*(t)q - Sz = 0,\\ q(T_1) = 0,\\ z(T_0) = z_0,\end{gathered} \tag{4.67}$$

et on peut appliquer la méthodologie du découplage. Pour cela on introduit une famille $\Pi(t) \in \mathscr{L}(H, H)$ définie de la manière suivante. On considère la famille de problèmes de contrôle (incluant (4.59) et (4.60)):

$$\begin{gathered}\frac{dz}{dt} + \mathscr{A}(t)z = Bv \quad \text{sur} \quad]s, T_1[,\\ z(s) = h,\end{gathered} \tag{4.68}$$

$$J_s^h(v) = \int_s^{T_1} (Nv, v)\,dt + \int_s^{T_1} (Sz, z)\,dt; \tag{4.69}$$

alors on a (équations d'Euler analogues à (4.67)):

$$\begin{gathered}\frac{d\varphi}{dt} + \mathscr{A}(t)\varphi + BN^{-1}B^*\psi = 0,\\ -\frac{d\psi}{dt} + \mathscr{A}^*(t)\psi - S\varphi = 0,\\ \psi(T_1) = 0,\\ \varphi(s) = h\end{gathered} \tag{4.70}$$

et

$$\psi(s) = \Pi(s)h. \tag{4.71}$$

On a alors

$$(\Pi(s)h, h) = \inf_v J_s^h(v). \tag{4.72}$$

Notons que $\Pi(s)$ est autoadjoint, mais pas forcément $\geqq 0$. Il nous sera utile de majorer $|(\Pi(s)h, h)|$. On a tout d'abord

$$(\Pi(s)h, h) \leqq J_s^h(0) \int_s^T (S\bar{\varphi}, \bar{\varphi})\,dt, \tag{4.73}$$

où φ est solution de

$$\begin{gathered}\frac{d\bar{\varphi}}{dt} + \mathscr{A}(t)\bar{\varphi} = 0 \quad \text{sur} \quad]s, T_1[,\\ \bar{\varphi}(s) = h.\end{gathered} \tag{4.74}$$

Par ailleurs, on a

$$\begin{aligned} J_s^h(v) = \int_s^{T_1} (Nv, v)\,dt + \int_s^{T_1} (S\bar{\varphi}, \bar{\varphi})\,dt + 2\int_s^{T_1} (S\bar{\varphi}, \tilde{\varphi})\,dt \\ + \int_s^{T_1} (S\tilde{\varphi}, \tilde{\varphi})\,dt, \end{aligned} \tag{4.75}$$

où $\tilde{\varphi} = \varphi - \bar{\varphi}$ (partie homogène de $v \to \varphi$). En utilisant la propriété (4.66), on déduit de (4.75),

$$J_s^h(v) \geqq \rho \int_s^{T_1} |v|^2\,dt + 2\int_s^{T_1} (S\bar{\varphi}, \tilde{\varphi})\,dt + \int_s^{T_1} (S\bar{\varphi}, \bar{\varphi})\,dt. \tag{4.76}$$

En introduisant $\bar{\psi}$ solution de

$$\begin{gathered} -\frac{d\bar{\psi}}{dt} + \mathscr{A}^*(t)\bar{\psi} = S\bar{\varphi}, \\ \bar{\psi}(T_1) = 0, \end{gathered} \tag{4.77}$$

on peut écrire

$$\int_s^{T_1} (S\bar{\varphi}, \tilde{\varphi})\,dt = \int_s^{T_1} \left(-\frac{d\bar{\psi}}{dt} + \mathscr{A}^*(t)\bar{\psi}, \tilde{\varphi} \right) dt = \int_s^{T_1} (\bar{\psi}, Bv)\,dt$$

et donc (4.75) donne

$$\begin{aligned} J_s^h(v) &\geqq \rho \int_s^{T_1} |v|^2\,dt + 2\int_s^{T_1} (B^*\bar{\psi}, v)\,dt + \int_s^{T_1} (S\bar{\varphi}, \bar{\varphi})\,dt \\ &= \int_s^{T_1} \left| \sqrt{\rho}v + \frac{B^*\bar{\psi}}{\sqrt{\rho}} \right|^2 dt + \int_s^{T_1} (S\bar{\varphi}, \bar{\varphi})\,dt - \frac{1}{\rho}\int_s^{T_1} |B^*\bar{\psi}|^2\,dt. \end{aligned} \tag{4.78}$$

Par conséquent on a

$$\inf_v J_s^h(v) = (\Pi(s)h, h) \geqq \int_s^{T_1} (S\bar{\varphi}, \bar{\varphi})\,dt - \frac{1}{\rho}\int_s^{T_1} |B^*\bar{\psi}|^2\,dt. \tag{4.79}$$

De (4.73) et (4.79) résulte aussitôt :

$$\begin{aligned} |(\Pi(s)h, h)| &\leqq \int_s^{T_1} (S\bar{\varphi}, \bar{\varphi})\,dt + \frac{1}{\rho}\int_s^{T_1} |B^*\bar{\psi}|^2\,dt \\ &\leqq \int_s^{T_1} (S\bar{\varphi}, \bar{\varphi})\,dt + \frac{\|B\|^2}{\rho}\int_s^{T_1} |\bar{\psi}|^2\,dt. \end{aligned} \tag{4.80}$$

Dans (4.77) on fait le changement de fonction

$$\bar{\psi} = \bar{\chi}\,e^{-\tau t},$$

d'où $\bar{\chi}$ est solution de

$$-\frac{d\bar{\chi}}{dt} + \tau\bar{\chi} + \mathscr{A}^*(t)\bar{\chi} = e^{\tau t}S\bar{\varphi},$$

$$\bar{\chi}(T_1) = 0,$$

d'où on déduit, par intégration par parties, la majoration

$$\eta \int_s^{T_1} |\bar{\chi}|^2 \, dt \leqq \int_s^{T_1} (e^{\tau t} S \bar{\varphi}, \bar{\chi}) \, dt$$

et donc

$$\int_s^{T_1} |\bar{\chi}|^2 \, dt \leqq \frac{1}{\eta^2} \max (e^{2\tau T_0}, e^{2\tau T_1}) \int_s^{T_1} |S\bar{\varphi}|^2 \, dt,$$

$$\int_s^{T_1} |\bar{\varphi}|^2 \, dt = \int_s^{T_1} e^{-2\tau t} |\bar{\chi}|^2 \, dt \leqq \max (e^{-2\tau T_0}, e^{-2\tau T_1}) \int_s^{T_1} |\bar{\chi}|^2 \, dt$$

$$\leqq \frac{e^{2|\tau| l}}{\eta^2} \int_s^{T_1} |S\bar{\varphi}|^2 \, dt.$$

Par conséquent (4.80) conduit à

$$\begin{aligned} |(\Pi(s)h, h)| &\leqq \|S\| \int_s^{T_1} |\bar{\varphi}|^2 \, dt + \frac{\|B\|^2 \|S\|^2}{\rho} \frac{e^{2\tau l}}{\eta^2} \int_s^{T_1} |\bar{\varphi}|^2 \, dt \\ &= \frac{\nu \|S\|}{\rho} \int_s^{T_1} |\bar{\varphi}|^2 \, dt. \end{aligned} \tag{4.81}$$

Si enfin dans (4.74) on fait le changement de fonction

$$\bar{\varphi} = \bar{\delta} \, e^{\tau t},$$

alors $\bar{\delta}$ est solution de

$$\frac{d\bar{\delta}}{dt} + (\tau + \mathscr{A}(t))\bar{\delta} = 0,$$

$$\bar{\delta}(s) = h \, e^{-\tau s},$$

d'où, en intégrant par parties,

$$|\bar{\delta}(T_1)|^2 + \tau \int_s^{T_1} |\bar{\delta}|^2 \, dt + \int_s^{T_1} \langle \mathscr{A}\bar{\delta}, \bar{\delta} \rangle \, dt = |\bar{\delta}(s)|^2.$$

Par conséquent, on a la majoration

$$\eta \int_s^{T_1} |\bar{\delta}|^2 \, dt \leqq e^{-2\tau s} |h|^2 \leqq \max (e^{2\tau T_1}, e^{-2\tau T_0}) \, |h|^2,$$

d'où

$$\int_s^{T_1} |\bar{\varphi}|^2 \, dt = \int_s^{T_1} e^{2\tau t} |\bar{\delta}|^2 \, dt \leqq \max (e^{2\tau T_1}, e^{2\tau T_0}) \int_s^{T_1} |\bar{\delta}|^2 \, dt \frac{e^{2|\tau| l} |h|^2}{\eta},$$

d'où finalement on obtient la proposition suivante.

PROPOSITION 4.2. *Si* (4.66) *est vérifiée, on a la majoration*

$$|(\Pi(s)h, h)| \leqq \frac{\nu \|S\|}{\rho \eta} e^{2|\tau| l} |h|^2. \tag{4.82}$$

4.3.4. Etude de Σ_i^k. On pose donc $\Sigma_i^k = Q_i^{k+1} - Q_i^k$. Alors d'après (4.20) Σ_i^k est solution de l'équation

$$
\begin{aligned}
&-\frac{d}{dt}(\Sigma_i^k\varphi) + \Sigma_i^k\left(-\frac{d\varphi}{dt} + A\varphi\right) + A^*\Sigma_i^k\varphi \\
&\quad + Q_i^{k+1} \sum_{j\neq i} B_jN_j^{-1}B_j^*Q_j^k\varphi - Q_i^k \sum_{j\neq i} B_jN^{j-1}B_j^*Q_j^{k-1}\varphi \\
&\quad + \sum_{j\neq i} Q_j^kB_jN_j^{-1}B_j^*Q_i^{k+1}\varphi - \sum_{j\neq i} Q_j^{k-1}B_jN_j^{-1}B_j^*Q_i^k\varphi \\
&\quad + Q_i^{k+1}B_iN_i^{-1}B_i^*Q_i^{k+1}\varphi - Q_i^kB_iN_i^{-1}B_i^*Q_i^k\varphi = 0,
\end{aligned}
\tag{4.83}
$$

$$\Sigma_i^k(T_1) = 0 \quad \text{pour tout} \quad \varphi \in L^2(T_0, T_1, V), \quad \frac{d\varphi}{dt} + A\varphi \in L^2(T_0, T_1, H).$$

Un calcul algébrique simple montre alors que (4.83) se réécrit sous la forme

$$
\begin{aligned}
&-\frac{d}{dt}(\Sigma_i^k\varphi) + \Sigma_i^k\left(\frac{d\varphi}{dt} + A\varphi + \sum_{j=1}^{n} B_jN_j^{-1}B_j^*Q_j^k\varphi\right) \\
&\quad + \left(A^* + \sum_{j=1}^{n} Q_j^kB_jN_j^{-1}B_j^*\right)\Sigma_i^k\varphi + \Sigma_i^kB_iN_i^{-1}B_i^*\Sigma_i^k\varphi \\
&\quad = -Q_i^k \sum_{j\neq i} B_jN_j^{-1}B_j^*\Sigma_j^{k-1}\varphi - \sum_{j=i} \Sigma_j^{k-1}B_jN_j^{-1}B_j^*Q_i^k\varphi, \\
&\qquad\qquad \Sigma_i^k(T_1) = 0,
\end{aligned}
\tag{4.84}
$$

et

$$\Sigma_i^0(t) = Q_i^1(t) - Q_i^0(t) = Q_i^1(t) - \bar{Q}_i. \tag{4.85}$$

La forme (4.84) montre que Σ_i^k est solution d'une équation de Riccati, dont le second membre est

$$S_i^k = -Q_i^k \sum_{j\neq i} B_jN_j^{-1}B_j^*\Sigma_j^{k-1} - \sum_{j\neq i} \Sigma_j^{k-1}B_jN_j^{-1}B_j^*Q_i^k, \tag{4.86}$$

qui est symétrique, mais non $\geqq 0$ (en général). On est dans le cadre du § 4.3.3 avec

$$\mathscr{A}(t) = A + \sum_{j=1}^{n} B_jN_j^{-1}B_j^*Q_j^k$$

donc $\lambda = KC_1 = \varepsilon'$, $\beta = \alpha$, $\beta_1 = \alpha_1$, $\beta_2 = \alpha_2$, $\eta = \varepsilon$, $\tau = \theta$ (donc >0). La condition (4.66) s'écrit

$$\rho_i^k = \nu_i - \frac{\|S_i^k\|\;\|B_i\|^2\, e^{2l\theta}}{\varepsilon^2} > 0. \tag{4.87}$$

On a alors le lemme suivant.

LEMME 4.3. *On choisit* $0 < x < x_0 = (1 + \sqrt{5})/4$, *puis* ε *comme en* (4.45), α_2 *comme en* (4.46) *et*

$$0 < l < \min\left\{\frac{\log[\varepsilon^2x/(C_1C_2 + C_1C_3\varepsilon)]}{2[\varepsilon(1 + x) - \alpha_2]}, \frac{\log[1/(4x^2 + 2x)]}{2\varepsilon(1 + x) - \alpha_2]}\right\}, \tag{4.88}$$

alors $\|\Sigma_i^k(t)\|_{\mathscr{L}(H,H)} \to 0$ *lorsque* $k \to$ *uniformément en* t *sur* $[T_0, T_1]$.

Démonstration. Remarquons que (4.88) a un sens car $4x^2 + 2x < 1$ et (4.47) est respectée. Dans ces conditions nous allons montrer que l'on a

$$\|\Sigma^k\| \leqq \frac{2x\, e^{2l[\varepsilon(1+x)-\alpha_2]}}{1 - 4x\, e^{2l[\varepsilon(1+x)-\alpha_2]}} \|\Sigma^{k-1}\|, \qquad k \geqq 1. \tag{4.89}$$

Si (4.89) est vérifiée, alors d'après (4.88)

$$e^{2l[\varepsilon(1+x)-\alpha_2]} < \frac{1}{4x^2 + 2x},$$

d'où

$$2x\, e^{2l[\varepsilon(1+x)-\alpha_2]} < 1 - 4x^2\, e^{2l[\varepsilon(1+x)-\alpha_2]}$$

et donc si

$$a = \frac{2x\, e^{2l[\varepsilon(1+x)-\alpha_2]}}{1 - 4x^2\, e^{2l[\varepsilon(1+x)-\alpha_2]}},$$

on a $0 < a < 1$. Par conséquent

$$\|\Sigma_i^k(t)\| \leqq a^{k-1} \|\Sigma_i^0(t)\| \leqq 2Ka^{k-1} \tag{4.90}$$

ce qui implique aussitôt le résultat. Tout revient donc à démontrer (4.89). Vérifions (4.89) pour $k = 1$. Mais Σ_i^1 est solution de (4.84) avec un second membre S_i^1 (cf. (4.86)) tel que

$$\|S_i^1\| \leqq 4K^2 C_1 = \frac{4\varepsilon'^2}{C_1}.$$

Mais alors

$$\frac{\|S_i^1\|\, \|B_i\|^2\, e^{2l\theta}}{\varepsilon^2} \leqq \frac{4\varepsilon'^2}{C_1 \varepsilon^2} \|B_i\|^2\, e^{2l\theta} \leqq 4v_i x^2\, e^{2l\theta}$$

et comme d'après ci-dessus

$$4x^2\, e^{2l\theta} = 4x^2\, e^{2l[\varepsilon(1+x)-\alpha_2]} < 1,$$

on a bien (4.87). Par conséquent on peut appliquer à (4.84) (pour $k = 1$) ce qui a été fait au § 4.3.3, et en particulier la Proposition 4.2. Il en résulte[5]

$$\begin{aligned} \|\Sigma_i^1(s)\| &= \sup_h \frac{|(\Sigma_i^1(s)h, h)|}{\|h\|^2} \leqq \frac{v^i \|S_i^1\|}{\rho_i^1 \varepsilon} e^{2\theta l} \\ &\leqq \frac{1}{1 - 4x^2\, e^{2l\theta}} \frac{e^{2\theta l}}{\varepsilon} 2C_1 K \|\Sigma_i^0\| \\ &= \frac{2x\, e^{2\theta l}}{1 - 4x^2\, e^{2l\theta}} \|\Sigma_i^0\|, \end{aligned} \tag{4.91}$$

ce qui n'est autre que (4.89). Supposons (4.89) vraie pour $k - 1$ et démontrons la pour k, $k \geqq 2$. D'après (4.90) on a

$$\|\Sigma_i^{k-1}(t)\| \leqq 2Ka^{k-2} \leqq 2K$$

[5] Car Σ_i^k est symétrique.

de sorte que l'on a

$$\|S_i^k\| \leqq 4K^2C_1.$$

Par conséquent (4.87) est vérifiée. On peut appliquer à (4.84) les résultats du § 4.3.3, ce qui donne

$$\begin{aligned}\|\Sigma_i^k\| &\leqq \frac{v_i\|S_i^k\|}{\rho_i^k\varepsilon}e^{2\theta l}\\ &\leqq \frac{1}{1-4x^2e^{2l\theta}}\frac{e^{2\theta l}}{\varepsilon}2C_1K\|\Sigma_i^{k-1}\|,\end{aligned}$$

c'est-à-dire encore (4.89). Le lemme est ainsi démontré.

4.3.5. Fin de la démonstration du Théorème 4.2. Soit $\psi \in W(T_0, T_1)$, on déduit de (4.21) l'expression

$$\begin{aligned}\int_{T_0}^{T_1}(R_i\varphi,\psi)\,dt &= \int_{T_0}^{T_1}\left\langle -\frac{d}{dt}Q_i^{k+1}\varphi,\psi\right\rangle\,dt\\ &+\int_{T_0}^{T_1}\left(Q_i^{k+1}\left(\frac{d\varphi}{dt}+A\varphi+\sum_{j\neq i}B_jN_j^{-1}B_j^*Q_j^k\varphi\right),\psi\right)dt\\ &+\int_{T_0}^{T_1}\left\langle\left(A^*+\sum_{j\neq i}Q_j^kB_jN_j^{-1}B_j^*\right)Q_i^{k+1}\varphi,\psi\right\rangle\,dt\\ &+\int_{T_0}^{T_1}(Q_i^{k+1}B_iN_i^{k-1}B_i^*Q_i^{k+1}\varphi,\psi)\,dt\\ &= \int_{T_0}^{T_1}\left\langle -\frac{d}{dt}Q_i^{k+1}\varphi,\psi\right\rangle\,dt\\ &+\int_{T_0}^{T_1}\left(\frac{d\varphi}{dt}+A\varphi+\sum_{j\neq i}B_jN_j^{-1}B_j^*Q_j^k\varphi, Q_i^{k+1}\psi\right)dt\\ &+\int_{T_0}^{T_1}\left\langle Q_i^{k+1}\varphi, A\psi\right\rangle\,dt\\ &+\int_{T_0}^{T_1}\left(Q_i^{k+1}\varphi,\sum_{j\neq i}B_jN_j^{-1}B_j^*Q_j^k\psi\right)dt\\ &+\int_{T_0}^{T_1}(B_iN_i^{-1}B_i^*Q_i^{k+1}\varphi, Q_i^{k+1}\psi)\,dt.\end{aligned} \tag{4.92}$$

D'après le Lemme 4.1, la suite Q_i^k demeure dans un borné de $L_s^\infty(T_0, T_1, \mathscr{L}(H, H))$.[6] D'après le Théorème de Dunford Pettis, $L_s^\infty(T_0, T_1, \mathscr{L}(H, H))$ et le dual de $L^1(T_0, T_1, H_1)$ espace des fonctions de $t:]T_0, T_1[\to$ l'espace H_1 des opérateurs *nucléaires* sur H.

On rappelle que H_1 est caractérisé de la manière suivante. Un opérateur $\Gamma \in \mathscr{L}(H, H)$ est nucléaire s'il peut s'écrire sous la forme

$$\Gamma h = \sum_{i=1}^{\infty}\lambda_i(h, e_i)f_i, \tag{4.93}$$

[6] Fonctions scalairement measurables et bornées.

où e_i et f_i sont des systèmes orthonormés de H (donnés), et si $\Sigma\lambda_i < +\infty$. On sait qu'alors les λ_i sont >0 et H_1 est un espace de Banach pour la norme

$$\|\Gamma\|_{H_1} = \Sigma\lambda_i. \tag{4.94}$$

Si $\Sigma \in \mathscr{L}(H, H)$ alors $\Sigma\Gamma$ est nucléaire et $\mathscr{L}(H, H)$ est le dual de H_1 pour la dualité

$$\langle \Sigma, \Gamma \rangle = \|\Sigma\Gamma\|_{H_1}. \tag{4.95}$$

Comme la suite Q_i^k demeure dans une boule du dual d'un espace de Banach, on peut extraire un sous suite $Q_i^{k_1}$ convergente vers Q_i dans

$$\sigma(L_s^\infty(T_0, T_1, \mathscr{L}(H, H)); L^1(T_0, T_1, H_1)),$$

$$\text{i.e.,} \quad \int_{T_0}^{T_1} \|(Q_i^{k_1}(t) - Q_i(t))\Gamma(t)\|_{H_1}\, dt \to 0 \quad \text{quand} \quad k_1 \to \infty, \tag{4.96}$$

$$\text{pour tout} \quad \Gamma \in L^1(T_0, T_1, H_1).$$

Soit $\beta(t) \in L^1(T_0, T_1)$ et $h \in H$. On prend $\Gamma_h(t)v = ((h, v)h/|h|^2)|\beta(t)|$ qui est bien un élément de $L^1(T_0, T_1, H_1)$ car $\|\Gamma_h(t)\|_{h_1} = \|\beta(t)|$. Mais alors

$$(Q_i^{k_1}(t) - Q_i(t))\Gamma_h(t)v = \frac{(h, v)(Q_i^{k_1}(t) - Q_i(t))h}{\|h\|^2}|\beta(t)|$$

$$= \frac{(h, v)}{|h|}\, \frac{(Q_i^{k_1}(t) - Q_i(t))h}{|(Q_i^k(t) - Q_i(t))h|}\, \frac{|\beta(t)|\,|(Q_i^{k_1}(t) - Q_i(t))h|}{|h|},$$

d'où

$$\|(Q_i^{k_1}(t) - Q_i(t))\Gamma_h(t)\|_{H_1} = |\beta(t)|\frac{|(Q_i^{k_1}(t) - Q_i(t)h)|}{|h|}.$$

Par conséquent il résulte de (4.96)

$$\int_{T_0}^{T_1} |\beta(t)|\,|(Q_i^{k_1}(t) - Q_i(t)h)| \to 0 \quad \text{pour tout} \quad h \in H, \qquad k_1 \to \infty.$$

Comme β est quelconque dans L^1, il en résulte

$$Q_i^{k_1}(t)h \to Q_i(t)h \text{ dans } H \text{ p.p. } t, \quad \text{pour tout} \quad h \in H, \qquad k_1 \to 0 + \infty. \tag{4.97}$$

Il résulte aussitôt de (4.97) que $Q_i(t)$ est p.p. $\geqq 0$ autoadjoint. Soit maintenant $\varphi \in W(T_0, T_1)$.

On définit (noter que $\varphi \in C(T_0, T_1, H)$)

$$\Gamma\varphi(t)v = \frac{(\varphi(t), v)\varphi(t)}{|\varphi(t)|_H}$$

qu'est bien un élément de $L^1(T_0, T_1, H_1)$ car $\|\Gamma\varphi(t)\|_{H_1} = |\varphi(t)|$ pour tout t. Donc

$$(Q_i^{k_1}(t) - Q_i(t))\Gamma\varphi(t)v = \frac{(\varphi(t), v)(Q_i^{k_1}(t) - Q_i(t))\varphi(t)}{|\varphi(t)|}$$

$$= \frac{(\varphi(t), v)}{|\varphi(t)|}\, \frac{(Q_i^{k_1}(t) - Q_i(t))\varphi(t)}{|(Q_i^{k_1}(t) - Q_i(t))\varphi(t)|}|(Q_i^{k_1}(t) - Q_i(t))\varphi(t)|.$$

Par conséquent,

$$\|(Q_i^{k_1}(t) - Q_i(t))\Gamma\varphi(t)\|_{H_1} = |(Q_i^{k_1}(t) - Q_i(t))\varphi(t)|$$

et (4.96) implique

$$(4.98) \qquad \int_{T_0}^{T_1} |(Q_i^{k_1}(t) - Q_i(t))\varphi(t)|_H \, dt \to 0, \quad \text{lorsque} \quad k \to \varphi + \infty.$$

En particulier, on peut prendre $\varphi \in L^2(T_0, T_1, V) | d\varphi/dt + A\varphi \in L^2(T_0, T_1, H)$. Or, d'après le Lemme 4.2, on peut, quitte à extraire une nouvelle sous suite supposer que $Q_i^{k_1}\varphi$ converge dans $W(T_0, T_1)$ faible vers un élément $\psi \in W(T_0, T_1)$. La convergence a lieu aussi dans $L^2(T_0, T_1, H)$ faible (en fait fort car l'injection de $W(T_0, T_1)$ dans $L^2(T_0, T_1, H)$ est compacte). Mais (4.98) montre aussitôt que $\psi = Q_i\varphi$, d'où

$$(4.99) \qquad \begin{aligned} &Q_i^{k_1}(t)\varphi(t) \to Q_i(t)\varphi(t) \quad \text{dans } W(T_0, T_1) \text{ faible} \\ &\qquad \text{pour tout} \quad \varphi \in L^2(T_0, T_1, V) \left| \frac{d\varphi}{dt} + A\varphi \in L^2(T_0, T_1, H), \right. \\ &Q_i^{k_1}(t)\psi(t) \to Q_i(t)\psi(t) \quad \text{dans } L^2(T_0, T_1, H) \text{ fort} \\ &\qquad \text{pour tout} \quad \psi \in W(T_0, T_1). \end{aligned}$$

Considérons maintenant la suite $Q_i^{k_1+1}(t)$. A priori elle ne converge pas, mais on peut extraire de k_1 une sous suite k_2 telle que $Q_i^{k_2+1}(t) \to Q_i'(t)$ dans $\sigma(L^\infty(T_0 T_1, \mathscr{L}(H, H)), L^1(T_0, T_1, H_1))$. De manière analogue à (4.97) on a pour tout h,

$$\begin{aligned} Q_i^{k_2+1}(t)h &\to Q_i'(t)h \quad \text{dans } H \text{ p.p. } t, \qquad & k_2 \to \infty, \\ Q_i^{k_2}(t)h &\to Q_i(t)h \quad \text{dans } H \text{ p.p. } t, \qquad & k_2 \to \infty. \end{aligned}$$

Mais d'après le Lemme 4.3 on a

$$\|Q_i^{k_2+1}(t) - Q_i^{k_2}(t)\| = \|\Sigma_i^{k_2}(t)\| \to 0$$

et donc

$$Q_i(t) = Q_i'(t) \text{ p.p. } t.$$

On a donc la propriété

$$(4.100) \qquad \begin{aligned} &\left.\begin{aligned} Q_i^{k_2}(t)\varphi(t) &\to Q_i(t)\varphi(t) \\ Q_i^{k_2+1}(t)\varphi(t) &\to Q_i(t)\varphi(t) \end{aligned}\right\} \begin{aligned} &\text{dans } W(T_0, T_1) \text{ faible} \quad \text{pour tout} \\ &\quad \varphi \in L^2(T_0, T_1, V) \left| \frac{d\varphi}{dt} + A\varphi \in L^2(T_0, T_1, H), \right. \end{aligned} \\ &\left.\begin{aligned} Q_i^{k_2}(t)\psi(t) &\to Q_i(t)\psi(t) \\ Q_i^{k_2+1}(t)\psi(t) &\to Q_i(t)\psi(t) \end{aligned}\right\} \begin{aligned} &\text{dans } L^2(T_0, T_1, H) \text{ fort} \\ &\qquad\qquad \text{pour tout} \quad \psi \in W(T_0, T_1). \end{aligned} \end{aligned}$$

On peut alors passer à la limite dans (4.92) écrite pour $k = k_2$. On obtient

$$(4.101) \qquad \begin{aligned} \int_{T_0}^{T_1} (R_i\varphi, \psi) \, dt = &\int_{T_0}^{T_1} \left\langle -\frac{d}{dt} Q_i\varphi, \psi \right\rangle dt \\ &+ \int_{T_0}^{T_1} \left(\frac{d\varphi}{dt} + A\varphi + \sum_{j \neq i} B_j N_j^{-1} B_j^* Q_j\varphi, Q_i\psi \right) dt \\ &+ \int_{T_0}^{T_1} \langle Q_i\varphi, A\psi \rangle \, dt + \int_{T_0}^{T_1} \left(Q_i\varphi, \sum_{j \neq i} B_j N_j^{-1} B_j^* Q_j\psi \right) dt \\ &+ \int_{T_0}^{T_1} (B_i N_i^{-1} B_i^* Q_i\varphi, Q_i\psi) \, dt \end{aligned}$$

soit encore

$$\begin{aligned}&\int_{T_0}^{T_1}\left\langle -\frac{dQ_i\varphi}{dt},\psi\right\rangle dt+\int_{T_0}^{T_1}\left(Q_i\left(\frac{d\varphi}{dt}+A\varphi+\sum_{j\neq i}B_jN_j^{-1}B_j^*Q_j\varphi\right),\psi\right)dt\\ (4.102)\quad &+\int_{T_0}^{T_1}\langle A^*Q_i\varphi,\psi\rangle\,dt+\int_{T_0}^{T_1}\left(\sum_{j\neq i}Q_jB_jN_j^{-1}B_j^*Q_i\varphi,\psi\right)dt\\ &+\int_{T_0}^{T_1}(Q_iB_iN_i^{-1}B_i^*Q_i\varphi,\psi)\,dt=\int_{T_0}^{T_1}(R_i\varphi,\psi)\,dt\end{aligned}$$

pour tout $\psi\in W(T_0,T_1)$,

ce qui prouve bien que

$$\begin{aligned}&-\frac{dQ_i\varphi}{dt}+Q_i\left(\frac{d\varphi}{dt}+A\varphi+\sum_{j\neq i}B_jN_j^{-1}B_j^*Q_j\varphi\right)+A^*Q_i\varphi\\ &+\sum_{j\neq i}Q_jB_jN_j^{-1}B_j^*Q_i\varphi+Q_iB_iN_i^{-1}B_i^*Q_i\varphi=R_i\varphi\end{aligned}$$

égalité de deux éléments de $L^2(T_0,T_1,V)$.

Par ailleurs de (4.92) on déduit en intégrant par parties,

$$\begin{aligned}&\int_{T_0}^{T_1}(R_i\varphi,\psi)\,dt=-(\bar{Q}_i\varphi(T_1),\psi(T_1))+(Q_i^{k_2+1}(T_0)\varphi(T_0),\psi(T_0))\\ (4.103)\quad &\qquad+\int_{T_0}^{T_1}\left\langle Q_i^{k_2+1}\varphi,\frac{d\psi}{dt}\right\rangle dt+\text{reste inchangé.}\end{aligned}$$

De (4.101) on déduit

$$\begin{aligned}&\int_{T_0}^{T_1}(R_i\varphi,\psi)\,dt=-(Q_i(T_1)\varphi(T_1),\psi(T_1))+(Q_i(T_0)\varphi(T_0),\psi(T_0)\\ (4.104)\quad &\qquad+\int_{T_0}^{T_1}\left\langle Q_i\varphi,\frac{d\psi}{dt}\right\rangle dt+\text{reste inchangé.}\end{aligned}$$

On note que

$$\int_{T_0}^{T_1}\left\langle Q_i^{k_2+1}\varphi,\frac{d\psi}{dt}\right\rangle dt\to\int_{T_0}^{T_1}\left\langle Q_i\varphi,\frac{d\psi}{dt}\right\rangle dt.$$

Si donc on prend $\psi\in W(T_0,T_1)$ tel que $\psi(T_0)=0$, on voit

$$(4.105)\qquad (Q_i(T_1)\varphi(T_1),\psi(T_1))=(\bar{Q}_i\varphi(T_1),\psi(T_1)).$$

Prenons $\psi=\gamma(t)v$, où $v\in V$ et $\gamma(t)\in H^1(T_0,T_1)$ telle que $\gamma(T_0)=0$. Il résulte aussitôt de (4.105)

$$(4.106)\qquad (Q_i(T_1)\varphi(T_1),v)=(\bar{Q}_i\varphi(T_1),v)\quad\text{pour tout}\quad v\in V,$$

et comme V est dense dans H, on obtient

$$\begin{aligned}&Q_i(T_1)\varphi(T_1)=\bar{Q}_i\varphi(T_1)\\ (4.107)\quad &\text{pour tout}\quad \varphi\in L^2(T_0,T_1,V)\left|\frac{d\varphi}{dt}+A(t)\varphi\in L^2(T_0,T_1,H).\right.\end{aligned}$$

Prenons en particulier φ solution de

$$\frac{d\varphi}{dt} + A(t)\varphi = 0, \tag{4.108}$$
$$\varphi(T_0) = a, \qquad a \in H.$$

Lorsque a varie dans $H, \varphi(T_1, a)$ parcourt un sous ensemble dense dans H. Comme $Q_i(T_1)$ et $\bar{Q}_i \in \mathscr{L}(H, H)$ et coincident sur un sous ensemble dense, ils sont égaux ce qui prouve bien que $Q_i(T_1) = \bar{Q}_i$ et achève la démonstration du Théroème 4.2.

4.3.6. Etude d'un cas particulier. On étudie maintenant le cas B_i = identité, N_i = identité; on suppose de plus que $A(t)$ est indépendant de t et autoadjoint.[7] L'équation s'écrit:

$$-\frac{d}{dt}(Q_i\varphi) + Q_i(t)\left(A\varphi + \sum_{j \neq i} Q_j\varphi\right) + \left(A + \sum_{j \neq i} Q_j\right)Q_i\varphi + Q_i^2\varphi = R_i\varphi \tag{4.109}$$
$$\text{pour tout} \quad \varphi \in D(A),$$
$$Q_i(T) = \bar{Q}_i.$$

On a alors le théorème suivant.

Théorème 4.3. *Outre les hypothèse du Théorème* 4.2, *on fait les hypothèses précédentes, alors la conclusion du Théorème* 4.2 *est vraie pour tout intervalle fini T.*

Démonstration. On prend comme norme de V

$$\|v\|_V^2 = \langle Av, v\rangle.$$

Ce qui est possible, d'après les hypothèses sur A. On considère alors une base orthonormée spéciale de V, $w_1, \cdots, w_i, \cdots$ orthogonale dans H, et telle que:

$$\langle Aw_i, v\rangle = \lambda_i(w_i, v), \quad \lambda_i > 0, \quad \text{pour tout} \quad v \in V.^8$$

On note ϖ_m le projecteur sur V_m = espace engendré par $w_1, \cdots, w_m$. On peut prolonger ϖ_m à H en posant

$$\varpi_m h = \sum_{i=1}^{m} \lambda_i(w_i, h)w_i.$$

On considère alors le système d'équations de Riccati:

$$-\frac{d}{dt}(Q_i^m w_h, w_k) + \left\langle\left(A + \sum_{j \neq i} Q_j^m(t)\right)w_h, Q_i^m(t)w_k\right\rangle$$
$$+ \left\langle\left(A + \sum_{j \neq i} Q_j^m(t)\right)Q_i^m(t)w_h, w_k\right\rangle \tag{4.110}$$
$$+(Q_i^m(t)w_h, Q_i^m(t)w_k) = (R_i w_h, w_k), \qquad h, k = 1, \cdots, m,$$
$$(Q_i^m(T)w_h, w_k) = (\bar{Q}_i w_h, w_k).$$

[7] Hypothèse trés probablement inutile.

[8] Il suffit de prendre les vecteurs propres de A^{-1} qui est compact de $H \to H$.

Comme le système (4.110) est un système d'équations différentielles, il posède une solution sur (t_m, T).

Nous avons alors le lemme suivant.

LEMME 4.4. *La suite Q_i^m vérifie la propriété suivante: Pour $h_m \in V_m$, $|Q_i^m(s)h_m| \leqq K|h_m|$, où K est une constante indépendante de s, i, m.*

Démonstration. Pour $s \in]t_m, T]$, on considère le problème de contrôle optimal

$$\begin{aligned}&\left(\frac{dz_m}{dt}, w_k\right) + \left\langle\left(A + \sum_{j\neq i} Q_j(t)\right)z_m, w_k\right\rangle = (v_i(t), w_k), \qquad h = 1, \cdots, m,\\ &z_m(s) = h_m,\end{aligned} \tag{4.111}$$

$$J_m^i(s, h_m; v_i) = \int_s^T (R_i z_m, z_m)\,dt + \int_s^T |v_i|^2\,dt + (\bar{Q}_i z_m(T), z_m(T)), \tag{4.112}$$

ce qui implique (car (4.110) peut s'interpréter comme l'équation de Riccati associée à (4.111), (4.112))

$$(Q_i^m(s)h_m, h_m) = \inf_{v_i} J_m^i(s, h_m; v_i). \tag{4.113}$$

Par conséquent, en introduisant ξ_m correspondant à $v_i = 0$, soit

$$\begin{aligned}&\left(\frac{d\xi_m}{dt}, w_k\right) + \left\langle\left(A + \sum_{j\neq i} Q_j^m(t)\right)\xi_m, w_k\right\rangle = 0,\\ &\xi_m(s) = h_m,\end{aligned} \tag{4.114}$$

on a

$$(Q_i^m(s)h_m, h_m) \leqq \int_s^T (R_i\xi_m, \xi_m)\,dt + (\bar{Q}_i\xi_m(T), \xi_m(T)).$$

Mais on déduit de (4.114)

$$\tfrac{1}{2}|\xi_m(T)|^2 + \alpha\int_s^T \|\xi_m(t)\|^2\,dt \leqq \tfrac{1}{2}|h_m|^2,$$

d'où

$$(Q_i^m(s)h_m, h_m) \leqq \left(\frac{C_2}{2\alpha} + C_3\right)|h_m|^2,$$

où

$$C_2 = \max_i \|R_i\|,$$

$$C_3 = \max_i \|\bar{Q}_i\|,$$

ce qui, tenant compte du fait que $Q_i^m(s)$ est $\geqq 0$, autoadjoint, implique aussitôt le lemme avec

$$K = \frac{C_2}{2\alpha} + C_3.$$

Il résulte du Lemme 4.4 que la solution de (4.110) peut se prolonger au delà de t_m et en fait est définie sur $(0, T)$.

Soit maintenant $\varphi \in D(A)$ et $\beta(t) \in H^1(0, T)$, on a alors le lemme suivant.

LEMME 4.5. *La suite $Q_i^m(t)\varpi_m\varphi\beta(t)$ demeure dans un borné de $W(0, T)$.*

Démonstration. On considère le système d'équations $(\varphi_m(t) = \varpi_m\varphi\beta(t))$:

$$\left(\frac{d\varphi_m}{dt}, w_k\right) + \left\langle\left(A + \sum_{j\neq i} Q_j^m(t)\right)\varphi_m, w_k\right\rangle + (\psi_i^m(t), w_k) = (g_m^{(t)}, w_k),$$

$$k = 1, \cdots, m,$$

$$(4.115)\qquad -\left(\frac{d}{dt}\psi_i^m, w_k\right) + \left\langle\left(A + \sum_{j\neq i} Q_j^m(t)\right)\psi_i^m, w_k\right\rangle - (R_i\varphi_m, w_k) = 0,$$

$$k = 1, \cdots, m,$$

$$(\psi_i^m(T), w_k) = (\bar{Q}_i\varphi_m(T), w_k), \quad \varphi_m(0) = \varphi_m(0),$$

que l'on interpréte comme pour la démonstration du Lemme 4.2. On définira g_m par

$$g_m(t) = \beta'(t)\varphi + \beta(t)A\varphi + \beta(t)\sum_{j\neq i} Q_j^m(t)\varpi_m\varphi + \psi_i^m(t).$$

On a alors

$$(4.116)\qquad \psi_i^m(t) = r_i^m(t) + Q_i^m(t)\varphi_m(t),$$

où $r_i^m(t)$ est solution de l'équation

$$-\left(\frac{dr_i^m}{dt}, w_k\right) + \left\langle\left(A + \sum_{j=1}^{n} Q_j^m(t)\right)r_i^m(t), w_k\right\rangle = (Q_i^m(t)w_k, g_m(t))$$

$$(4.117)\qquad \text{pour tout}\quad k = 1, \cdots, m,$$

$$r_i^m(T) = 0.$$

On vérifie aisément que ψ_i^m demeure dans un borné de $W(0, T)$. A partir du Lemme 4.4 on vérifie que $g_m(t)$ demeure dans un borné de $L^2(0, T; H)$ (noter que $|\varpi_m|_{\mathscr{L}(H;H)} \leqq 1$). Puis à partir de (4.117) on vérifie que r_i^m demeure dans un borné de $W(0, T)$, d'où le résultat à partir de (4.116).

Posons

$$(4.118)\qquad \tilde{Q}_i^m(t) = Q_i^m(t)\varpi_m.$$

Il résulte du Lemme 4.4 que $\tilde{Q}_i^m(t)$ demeure dans un borné de $L_s^\infty(0, T; \mathscr{L}(H; H))$. On peut donc extraire une sous suite $\tilde{Q}_i^{m'}(t)$ qui converge vers $Q_i(t)$ dans

$$\sigma(L_s^\infty(0, T; \mathscr{L}(H; H)); L^1(0, T; H_1)).$$

En raisonnant comme pour la fin de la démonstration du théorème 4.2 ((4.98) et (4.99)) on obtient (on écrit m au lieu de m')

$$(4.119)\qquad \begin{aligned} &Q_i^m(t)\varpi_m\varphi\beta(t) \to Q_i(t)\varphi\beta(t) \quad \text{dans } W(0, T) \text{ faible},\\ &Q_i^m(t)\varpi_m\varphi\beta(t) \to Q_i(t)\varphi\beta(t) \quad \text{dans } L^2(0, T; H) \text{ fort}.\end{aligned}$$

On multiplie (4.110) par $((\varphi, w_h))$ et on somme pour $h = 1, \cdots, m$. En remarquant que

$$\langle A\varpi_m\varphi, v\rangle = \langle A\varphi, \varpi_m v\rangle \quad \text{pour tout} \quad v \in V,$$

on obtient

$$\begin{aligned}
&-\frac{d}{dt}(Q_i^m(t)\varpi_m\varphi, w_k) + \Big(A\varphi + \sum_{j\neq i} Q_j^m(t)\varpi_m\varphi, Q_i^m(t)w_k\Big) \\
(4.120)\qquad &+ \Big\langle\Big(A + \sum_{j\neq i} Q_j^m(t)\Big)Q_i^m(t)\varpi_m\varphi, w_k\Big\rangle + (Q_i^m(t)\varpi_m\varphi, Q_i^m(t)w_k\Big) \\
&= (R_i\varpi_m\varphi, w_k).
\end{aligned}$$

Soit $\psi \in D(A)$. En multipliant (4.120) par $((\psi, w_k))$ et en additionant en k, on obtient

$$\begin{aligned}
&-\frac{d}{dt}(Q_i^m(t)\varpi_m\varphi, \varpi_m\psi) + (A\varphi, Q_i^m(t)\varpi_m\psi) + (Q_i^m(t)\varpi_m\varphi, A\psi) \\
(4.121)\qquad &+ \Big(Q_i^m(t)\varpi_m\psi, \sum_{j\neq i} Q_j^m(t)\varpi_m\varphi\Big) + \Big(Q_i^m(t)\varpi_m\varphi, \sum_{j\neq i} Q_j^m(t)\varpi_m\psi\Big) \\
&+ (Q_i^m(t)\varpi_m\varphi, Q_i^m(t)\varpi_m\psi) = (R_i\varpi_m\varphi, \varpi_m\psi).
\end{aligned}$$

En multipliant par $\beta(t) \in \mathscr{D}(]0, T[)$ et intégrant, il vient

$$\begin{aligned}
&\int_0^T (Q_i^m(t)\varpi_m\varphi, \varpi_m\psi)\beta'(t)\,dt + \int_0^T (A\varphi, Q_i^m(t)\varpi_m\psi)\beta(t)\,dt \\
&+ \int_0^T (A\psi, Q_i^m(t)\varpi_m\varphi)\beta(t)\,dt + \int_0^T (Q_i^m(t)\varpi_m\psi, \sum_{j\neq i} Q_j^m(t)\varpi_m\varphi)\beta(t)\,dt \\
&+ \int_0^T (Q_i^m(t)\varpi_m\varphi, \sum_{j\neq i} Q_j^m(t)\varpi_m\psi)\beta(t)\,dt + \int_0^T (Q_i^m(t)\varpi_m\varphi, Q_i^m(t)\varpi_m\psi)\beta(t)\,dt \\
&\qquad = \int_0^T (R_i\varpi_m\varphi, \varpi_m\psi)\beta(t)\,dt.
\end{aligned}$$

En passant à la limite en m, on obtient:

$$\begin{aligned}
&\int_0^T (Q_i(t)\varphi, \psi)\beta'(t)\,dt + \int_0^T (A\varphi, Q_i(t)\psi)\beta(t)\,dt + \int_0^T (A\psi, Q_i(t)\varphi)\beta(t)\,dt \\
&+ \int_0^T \Big(Q_i(t)\psi, \sum_{j\neq i} Q_j(t)\varphi\Big)\beta(t)\,dt + \int_0^T \Big(Q_i(t)\varphi, \sum_{j\neq i} Q_j(t)\psi\Big)\beta(t)\,dt \\
&+ \int_0^T (Q_i(t)\varphi, Q_i(t)\psi)\beta(t)\,dt = \int_0^T (R_i\varphi, \psi)\beta(t)\,dt,
\end{aligned}$$

d'où on déduit aisément l'èquation différentielle (4.109). La condition initiale se démontre comme pour le Théorème 4.2, (4.105).

BIBLIOGRAPHIE

[1] A. BENSOUSSAN, *Differential games for systems governed by partial differential equations*, Proc. 1st Conference on Differential Games, Amherst, Mass., 1969.

[2] A. FRIEDMAN, *Differential Games*, John Wiley, New York, 1971.

[3] B. LEMAIRE, Thesis, Paris, 1971.

[4] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[5] ———, *Quelques méthodes de résolution de problémes aux limites non linéaires*, Dunod, Paris, 1969.

[6] J. L. LIONS ET G. STAMPACCHIA, *Variational inéqualitiés*, Comm. Pure Appl. Math., 20 (1967), pp. 493–519.

[7] D. L. LUKES, *Equilibrium feedback control in linear games with quadratic costs*, this Journal, 9 (1971), pp. 234–252.

[8] K. YOSIDA, *Functional Analysis*, Springer Verlag, Berlin, 1965.

SIAM J. CONTROL
Vol. 12, No. 3, August 1974

A NOTE ON COMPLETE CONTROLLABILITY AND STABILIZABILITY FOR LINEAR CONTROL SYSTEMS IN HILBERT SPACE*

MARSHALL SLEMROD†

Abstract. We consider the linear control system $\dot{x} = Ax + Bu$. Here A is the infinitesimal generator of a strongly continuous semigroup of bounded linear operators $T(t)$ on a Hilbert space E, and B is a bounded linear operator from a Hilbert space H to E. We give sufficient conditions for the existence of a bounded linear operator K from E to H so that the control system with feedback control law $u(t) = Kx(t)$ has the zero solution asymptotically stable. In particular, we study the relationship between the concept of complete controllability and the existence of K.

1. Introduction. In this note we shall consider the linear control system given by

$$\dot{x}(t) = Ax(t) + Bu(t). \tag{1.1}$$

Here A is the infinitesimal generator of a C_0 semigroup of bounded linear operators $T(t)$, $t \geqq 0$, on a Hilbert space E, and B is a bounded linear operator from a Hilbert space H to E. We pose the problem of determining conditions for the existence of a bounded linear operator K from E to H so that the control system with feedback control law $u(t) = Kx(t)$ has the zero solution asymptotically stable.

This problem was considered in [1] where we obtained weaker results than those given here. In particular, in [1] we required that $T(t)$ be a C_0 group on E and satisfy the inequality

$$\int_0^{\varepsilon} \|B^*T^*(-t)x\|_H^2\,dt \geqq \delta\|x\|_E^2 \tag{1.2}$$

for some $\varepsilon > 0$, $\delta > 0$ and all x in E. This inequality is equivalent to complete controllability in the case when E, H are finite-dimensional, but is a stronger condition than complete controllability when E is infinite-dimensional. In this note we first show that for systems satisfying (1.2) we can obtain stronger results than those given in [1]. Specifically, we show that under hypothesis (1.2), systems of the form (1.1) can be stabilized with exponential decay rates, whereas in [1] we obtained a weaker type of convergence. Second and more importantly, we show that the weaker hypothesis of complete controllability will imply a type of stabilizability if some compactness of trajectories is assumed and $(Ax, x) \leqq 0$ for all $x \in D(A)$. The method of analysis is a Lyapunov stability approach exploiting a generalization of the invariance principle of LaSalle [5]. The invariance principle will be stated and then applied to system (1.1) where our stabilizing feedback is $u(t) = -B^*x(t)$. An example illustrating the results when the uncontrolled system is the wave equation will also be given.

* Received by the editors November 16, 1972, and in revised form June 11, 1973.

† Division of Applied Mathematics, Center for Dynamical Systems, Brown University, Providence, Rhode Island. Now at Department of Mathematics, Rensselaer Polytechnic Institute, Troy, New York 12181. This research was supported in part by the Office of Naval Research under Grant NONR-N00014-67-A-0191-0009 and by the United States Army Research Office—Durham under Grant DA-ARO-D-31-124-71-G12S2.

Notation. If X is a Hilbert space,[1] we denote its inner product by $(\cdot, \cdot)_X$ and norm by $\|\cdot\|_X$. If Y is also a Hilbert space, we denote the space of bounded linear operators from X to Y by $\mathscr{L}(X, Y)$ with operator norm $\|\cdot\|_{\mathscr{L}(X,Y)}$.

2. Definitions, assumptions and preliminary results. Consider the linear control system given by (1.1).

DEFINITION 2.1. We say that (1.1) is *stabilizable* (*weakly stabilizable*) if there exists $K \in \mathscr{L}(E, H)$ so that the C_0 semigroup $S(t)$ on E generated by $A + BK$ is such that $S(t)y \to 0$ ($S(t)y \to$ weakly) as $t \to \infty$ for all $y \in E$.

DEFINITION 2.2. We say that (1.1) is *completely controllable* if for $y \in E$, $B^*T^*(t)y = 0$ for all $t \geqq 0$ implies $y = 0$.

Definition 2.2 is equivalent to the standard notions of complete controllability. Furthermore, Fattorini [2], [3] has given a theory of completely controllable systems in Hilbert space and numerous examples.

Now let us note two assumptions which will be used at various times.

Assumption 1. Let A_1 be the infinitesimal generator of a C_0 group $T_1(t)$ of bounded linear operators on E. We say A_1 satisfies Assumption 1 if for some $\varepsilon > 0$, there exists $\delta > 0$ so that

$$\int_0^{\varepsilon} \|B^*T_1^*(-t)y\|_H^2\, dt \geqq \delta \|y\|_E^2 \quad \text{for all } y \text{ in } E.$$

Assumption 2. Let A_1 be the infinitesimal generator of a C_0 semigroup $T_1(t)$ of bounded linear operators on E with domain $D(A_1)$. We say A_1 satisfies Assumption 2 if there exists a constant λ_0 so that, for $\lambda > \lambda_0$, λ is in the resolvent set of A_1 and $(A_1 + \lambda I)^{-1}$ is a compact operator on E.

Remark 2.1. Assumption 1 is equivalent to complete controllability of (1.1) when $A = A_1$ and E, H are finite-dimensional.

Remark 2.2. Assumption 2 reflects the fact that in applications we wish to consider the case when A_1 is a linear differential operator. In this situation the domain of A_1 is smoother than E and hence we expect compactness of the resolvent operator.

Consider, for example, the case when $E = \mathring{H}_1(\Omega) \oplus H_0(\Omega)$, Ω some bounded sufficiently smooth spatial domain in R^n, and

$$A_1 = \begin{bmatrix} 0 & 1 \\ \Delta & 0 \end{bmatrix}, \quad \Delta \text{ the Laplacian.}^2$$

In this case $D(A_1) = \{H_2(\Omega) \cap \mathring{H}_1(\Omega)\} \oplus \mathring{H}_1(\Omega)$ and $\dot{x} = A_1 x$ represents the *wave equation with homogeneous Dirichlet boundary conditions.*

Let λ be on the resolvent set of A_1, $f_1 \in \mathring{H}_1(\Omega)$, $f_2 \in H_0(\Omega)$, and let $f = [f_1, f_2] \in E$. Then $A_1 x + \lambda x = f$ may be written as

$$x_2 + \lambda x_1 = f_1,$$

$$\Delta x_1 + \lambda x_2 = f_2,$$

[1] All Hilbert spaces will be assumed to be real.

[2] In examples using the wave equation we endow E with the equivalent energy norm $\|(u, v)\|_E^2 = \int_\Omega |\nabla u|^2 + |v|^2\, dx$

where $x_1 \in H_2(\Omega) \cap \mathring{H}_1(\Omega)$, $x_2 \in \mathring{H}_1(\Omega)$ and $x = [x_1, x_2] \in D(A_1)$. (Since λ is in the resolvent set of A_1, solutions x_1, x_2 must exist.) We therefore know that

$$\Delta x_1 - \lambda^2 x_1 = f_2 - \lambda f_1.$$

It thus follows that $\|\Delta x_1 - \lambda^2 x_1\|_{H_0(\Omega)} = \|f_2 - \lambda f_1\|_{H_0(\Omega)}$ which in turn implies that $\|x_1\|_{H_2(\Omega)} \leqq \text{const.} \|f_2 - \lambda f_1\|_{H_0(\Omega)}$.[3] Thus for $\{[f_{1n}, f_{2n}]\}$ a bounded sequence in E we see that the corresponding solutions $\{[x_{1n}, x_{2n}]\}$ are such that $\{x_{1n}\}$ is in a bounded set of $H_2(\Omega)$ and x_{2n} will lie in a bounded set of $H_1(\Omega)$. Hence by the Sobolev imbedding theorem, $\{[x_{1n}, x_{2n}]\}$ has a convergence subsequence in E.[4] Thus $(A_1 + \lambda I)^{-1}$ is a compact operator on E for any λ in the resolvent set of A_1.

The same argument as given above will be valid when Δ is replaced by any linear elliptic operator. In this case, instead of $\dot{x} = A_1 x$ representing the wave equation, it may also represent more general second order evolution equations of both the hyperbolic and vibrating-plate nature.

In applications when using Assumption 2 we often wish to use perturbations of the operator A_1 and not A_1 itself. This presents no difficulties, as Assumption 2 induces a type of stability under perturbations, as seen in the following lemma.

LEMMA 2.1. *Let $P \in \mathscr{L}(E, E)$ and assume Assumption* 2 *holds for A_1. Then there exists a λ_1 in the resolvent set of A_1 so that $(A_1 + P + \lambda I)^{-1}$ is a compact operator on E for all $\lambda > \lambda_1$.*

Proof. We know, since A_1 is the infinitesimal generator of a C_0 semigroup on E, that the Hille–Yosida–Phillips theorem implies $\|(A_1 + \lambda I)^{-1}\|_{\mathscr{L}(E,E)} \leqq M(\lambda - \omega)^{-1}$ for $\lambda > \omega$, where ω, M are constants. Thus for λ sufficiently large we have the inequality $\|P\|_{\mathscr{L}(E,E)} \|(A_1 + \lambda I)^{-1}\|_{\mathscr{L}(E,E)} < 1$. Applying the standard perturbation theorem on bounded invertibility [8, p. 196] yields the desired result.

We are now able to pass to some results on stabilization. The first theorem is an improvement of [1, Thm. 4.1] and follows from a suggestion by Prof. F. Massey. The second theorem incorporates Assumption 2 and yields stronger results than the first theorem.

THEOREM 2.1. *If A satisfies Assumption* 1, *then* (1.1) *is stabilizable. Furthermore, given $\lambda > 0$, there exists $K \in \mathscr{L}(E, H)$ (depending on λ) so that the C_0 group $S(t)$ generated by $A + BK$ satisfies $\|S(t)\|_{\mathscr{L}(E,E)} \leqq Me^{-\lambda t}$, $t \geqq 0$.*

Proof. Define

$$D_{\varepsilon,\lambda} x = \int_0^{\varepsilon} e^{-2\lambda t} T(-t) BB^* T^*(-t) x \, dt.$$

Then $D_{\varepsilon,\lambda} \in \mathscr{L}(E, E)$ is self-adjoint and, by Assumption 1, possesses a bounded inverse. Now consider the control system given by

$$\dot{y}(t) = (A + \lambda I) y(t) + Bu(t). \tag{2.1}$$

Substituting $u(t) = -B^* D_{\varepsilon,\lambda}^{-1} y(t)$ into (2.1), we see that $(A + \lambda I) - BB^* D_{\varepsilon,\lambda}^{-1}$ is the infinitesimal generator of a C_0 group $S_1(t)$ on E. A simple computation[5] shows

[3] This follows from the well-known a priori elliptic estimate. See [10, p. 170].

[4] For the Sobolev imbedding theorem, see [10, p. 169].

[5] The computation essentially parallels a similar computation made for the finite-dimension stabilization problem in [11]. See Appendix.

that for $y_0 \in D(A^*)$,

$$(2.2)\quad \frac{d}{dt}(S_1^*(t)y_0, D_{\varepsilon,\lambda}S_1^*(t)y_0)_E = -\|e^{-\lambda\varepsilon}B^*T^*(-\varepsilon)S_1^*(t)y_0\|_H^2 - \|B^*S_1^*(t)y_0\|_H^2.$$

Hence by the denseness of $D(A^*)$ in E we have $\|S_1^*(t)\|_{\mathscr{L}(E,E)} \leqq M$ for $t \geqq 0$, M a positive constant. Since $S_1(t) = e^{\lambda t}S(t)$, it follows that $\|S(t)\|_{\mathscr{L}(E,E)} \leqq Me^{-\lambda t}$ when $K = -B^*D_{\varepsilon,\lambda}^{-1}$.

THEOREM 2.2. *If A satisfies Assumption* 1 *and A^* satisfies Assumption* 2, *then, given $\lambda > 0$, the $K \in \mathscr{L}(E, H)$ specified in Theorem* 2.1 *is such that the group $S(t)$ generated by $A + BK$ satisfies $e^{\lambda t}S(t)y \to 0$ weakly as $t \to \infty$ for all y in E.*

Proof. Again, consider the control system given by (2.1) with feedback control $u(t) = -B^*D_{\varepsilon,\lambda}^{-1}y(t)$. Denote $S_1(t)$ as the C_0 group on E generated by $(A + \lambda I) - BB^*D_{\varepsilon,\lambda}^{-1}$. From equality (2.2) we know $\|S_1^*(t)y_0\|_E \leqq L\|y_0\|_E$ for some constant L and all y_0 in $D(A^*)$. Using the semigroup property of $S_1^*(t)$ we also see that $\|(A^* + \lambda I - BB^*D_{\varepsilon,\lambda}^{-1})S_1^*(t)y_0\|_E \leqq L\|(A^* + \lambda I - BB^*D_{\varepsilon,\lambda}^{-1})y_0\|_E$ for all y_0 in $D(A^*)$. Now let λ_1 be a number given by Lemma 2.1 which makes $(A^* + P + \lambda_1 I)^{-1}$ compact, where we identify $P = \lambda I - BB^*D_{\varepsilon,\lambda}^{-1}$. It then follows from the triangle inequality that

$$\|(A^* + \lambda I - BB^*D_{\varepsilon,\lambda}^{-1} + \lambda_1 I)S_1^*(t)y_0\|_E \leqq L\|(A^* + \lambda I - BB^*D_{\varepsilon,\lambda}^{-1})y_0\|_E + |\lambda_1|L\|y_0\|_E.$$

It then follows that for $y_0 \in D(A^*)$, $S_1^*(t)y_0$ lies in a compact set of E for $t \geqq 0$. Since $D(A^*)$ is dense in E we have from [6, Prop. 3.4] that for $y_0 \in E$, $S_1^*(t)y_0$ lies in a compact set of E.

At this point if we make the identification $(A + \lambda I)$ with the operator A of Theorem 4.1 of [1] and apply Theorem 4.1 of [1] we can conclude that $S_1(t)y \to 0$ weakly as $t \to \infty$ for all y in E. Since $S_1(t) = e^{\lambda t}S(t)$ the theorem is proved.

3. Relationship between complete controllability and stabilizability. In this section we will consider the relationship between complete controllability and stabilizability. In particular, we will restrict ourselves to the case when A is the infinitesimal generator of a C_0 *contraction* semigroup $T(t)$ on E, i.e., $\|T(t)\|_{\mathscr{L}(E,E)} \leqq 1$ for $t \geqq 0$. For this class of generators it will be shown that Assumption 2 and complete controllability of (1.1) imply weak stabilizability of (1.1). Furthermore, in the special case when $A = -A^*$ we will obtain the stronger result that Assumption 2 and complete controllability of (1.1) imply stabilizability of (1.1).

The principal tool in our analysis will be Hale's generalization [4] of the well-known invariance principle of LaSalle [5]. For completeness we will state a simplified linear version of the invariance principle. For the general nonlinear theory the reader is referred to [4], [6]. Also, it should be noted that questions similar to those studied here have been investigated in a similar invariance-principle spirit by Dafermos in [7].

Let $R = (-\infty, \infty)$, $R^+ = [0, \infty)$ and $\mathscr{B}$ be a Banach space with norm $\|\phi\|_{\mathscr{B}}$ for $\phi \in \mathscr{B}$.

Definition 3.1. Let $U:R^+ \times \mathscr{B} \to \mathscr{B}$ be a C_0 semigroup of bounded linear operators. The *positive orbit* $O^+(\phi)$ through $\phi \in \mathscr{B}$ is defined as

$$O^+(\phi) = \bigcup_{t \geqq 0} U(t)\phi.$$

Definition 3.2. A set M^+ in $\mathscr{B}$ is a *positively invariant set* for the semigroup U if for each $\phi \in M^+$, $O^+(\phi) \subset M^+$.

Definition 3.3. If U is also a C_0 group on $\mathscr{B}$, we will say a set M in $\mathscr{B}$ is *invariant* for the group U if for each $\phi \in M$, $U(t)\phi \in M$ for $t \in R$.

Remark 3.1. It is obvious that for $U(t)$ a group invariance implies positive invariance.

Definition 3.4. If V is a continuous scalar function on $\mathscr{B}$, define the functional

$$\dot{V}(\phi) = \overline{\lim_{t \to 0^+}} \frac{1}{t}[V(U(t)\phi) - V(\phi)].$$

Definition 3.5. $V:\mathscr{B} \to R$ is said to be a *Lyapunov functional* on a set G in $\mathscr{B}$ if V is continuous on $\bar{G}$, the closure of G, and if $\dot{V}(\phi) \leqq 0$ for ϕ in G. Furthermore, denote

$$S = \{\phi \text{ in } G;\ \dot{V}(\phi) = 0\}.$$

With these definitions it is possible to prove the following theorem.

Theorem 3.1. *Let U be a C_0 semigroup (group) on $\mathscr{B}$. If V is a Lyapunov functional on G and a positive orbit $O^+(\phi)$ belongs to G and is in a compact set of $\mathscr{B}$, then $U(t)\phi \to M^+$ $(U(t)\phi \to M)$ as $t \to \infty$, where*

$$M^+ = \textit{the largest positive invariant set in } S$$

$$(M = \textit{the largest invariant set in } S).$$

It is self-evident that in applications to the problem of asymptotic stability the objective is to find a suitable functional V and show $M^+ = \{0\}$ in the case of semigroups and $M = \{0\}$ in the case of groups. (Of course, in the group case, since $M^+ \supset M$, $M^+ = \{0\}$ implies $M = \{0\}$.)

We will now indicate our stabilization results. First let us choose our stabilizing feedback as $u(t) = -B^*x(t)$ so that our stabilized control system is given by

$$\dot{x}(t) = Ax(t) - BB^*x(t). \tag{3.1}$$

Since $BB^* \in \mathscr{L}(E, E)$, $A - BB^*$ is the infinitesimal generator of a C_0 semigroup (group) on E when A is the infinitesimal generator of a C_0 semigroup (group) on E. Let us denote this semigroup (group) by $S(t)$. Our aim is to find conditions so that $S(t)y \to 0$ as $t \to \infty$ for $y \in E$.

Theorem 3.2. *Let A be the infinitesimal generator of a C_0 contraction semigroup $T(t)$ on E for $t \geqq 0$. If* (i) *for every $y \in E$, $S^*(t)y$ remains in some compact set of E (which may depend on y) for $t \geqq 0$,* (ii) (1.1) *is completely controllable, then $S(t)y \to 0$ weakly as $t \to \infty$ for all $y \in E$, i.e.,* (1.1) *is weakly stabilizable.*

Proof. Let $C = A - BB^*$. Then C^* is the infinitesimal generator of $S^*(t)$, where $D(C^*) = D(A^*)$. Consider the functional on E given by $V(x) = \frac{1}{2}\|x\|_E^2$. Since $\|T(t)\|_{\mathscr{L}(E,E)} \leqq 1$ we have $(x, A^*x)_E \leqq 0$ for all $x \in D(A^*)$ and thus a simple computation shows $\dot{V}(S^*(t)x_0) \leqq -\|B^*S^*(t)x_0\|_H^2$ for $x_0 \in D(C^*)$. Now let $y \in E$. Since $D(C^*)$ is dense in E there exists $\{y_n\} \subset D(C^*)$ so that $y_n \to y$ in E. Then

$$\begin{aligned} \dot{V}(y) &= \overline{\lim_{t\to 0^+}} \frac{1}{t}[V(S^*(t)y) - V(y)] = \overline{\lim_{t\to 0^+}}\ \overline{\lim_{n\to\infty}} \frac{1}{t}[V(S^*(t)y_n) - V(y_n)] \\ &= \overline{\lim_{t\to 0^+}} \lim_{n\to\infty} \frac{1}{t}\left[\int_0^t \dot{V}(S^*(s)y_n)\,ds\right] \\ &\leqq \overline{\lim_{t\to 0^+}} \lim_{n\to\infty} \frac{1}{t}\left[-\int_0^t \|B^*S^*(s)y_n\|_H^2\,ds\right] \\ &\leqq -\|B^*y\|_H^2 \quad \text{for } y \text{ in } E. \end{aligned} \tag{3.2}$$

Now we are ready to apply Theorem 3.1. We identify the semigroup U of Definition 3.1 as $S^*(t)$ and the set G in Definition 3.4 as E. V is then a Lyapunov functional on E. Furthermore, hypothesis (i) implies that for the semigroup $S^*(t)$, $O^+(y)$ belongs to a compact set of E. Theorem 3.1 then implies that $S^*(t)y \to M^+$ as $t \to \infty$, where M^+ is the largest positive invariant set in $\{y \in E;\ \dot{V}(y) = 0\}$.

Our task is now to show that $M^+ = \{0\}$. Let $m \in M^+$ and define $z(t) = \int_0^t S^*(s)m\,ds$. Since C^* is closed it follows that $z(t) \in D(C^*)$ and[6]

$$\dot{z}(t) = A^*z(t) - BB^*z(t) + m \quad \text{for } t \geqq 0. \tag{3.3}$$

From the definition of M^+ it follows that $B^*S^*(t)m = 0$ for all $t \geqq 0$. Thus $B^*z(t) = 0$ for $t \geqq 0$ and we see, using (3.3), that this implies by the standard variation of constants formula and $z(0) = 0$ that $z(t) = \int_0^t T^*(s)m\,ds$. But $B^*z(t) = 0$ so

$$\int_0^t B^*T^*(s)m\,ds = 0 \quad \text{for } t \geqq 0.$$

This in turn implies $B^*T^*(t)m = 0$ for all $t \geqq 0$ and, employing the complete controllability hypothesis, we see that $m = 0$. Applying Theorem 3.1, we then conclude that $S^*(t)y \to 0$ in E as $t \to \infty$ for $y \in E$. Thus $S(t)y \to 0$ weakly in E as $t \to \infty$ and we have proved weak stabilizability.

THEOREM 3.3. *Let A be the infinitesimal generator of a C_0 contraction group $T(t)$ on E for $t \geqq 0$. If* (i) *for every $y \in E$, $S(t)y$ remains in a compact set of E (which may depend on y) for $t \geqq 0$,* (ii) (1.1) *is completely controllable,* (iii) $A = -A^*$, *then $S(t)y \to 0$ as $t \to \infty$ for all $y \in E$, i.e.,* (1.1) *is stabilizable.*

Proof. Again, let $C = A - BB^*$. Consider the functional on E given by $V(x) = \frac{1}{2}\|x\|_E^2$. Since $\|T(t)\|_{\mathscr{L}(E,E)} = 1$ we have $(x, Ax)_E = 0$ for all $x \in D(A)$ and thus a simple computation shows that $\dot{V}(S(t)x_0) \leqq -\|B^*S(t)x_0\|_E^2$ for $x_0 \in D(C)\ (= D(A))$. Following the same argument as in the proof of Theorem 3.2, we can conclude that

$$\dot{V}(y) \leqq -\|B^*y\|_H^2 \quad \text{for } y \text{ in } E. \tag{3.4}$$

[6] See the Appendix.

Again, we wish to apply Theorem 3.1. We identify the semigroup U of Definition 3.1 as $S(t)$ and the set G in Definition 3.4 as E. Furthermore, hypothesis (i) implies that for the group $S(t)$, $O^+(y)$ belongs to a compact set of E. Theorem 3.1 then implies that $S(t)y \to M$ as $t \to \infty$, where M is the largest invariant set in $\{y \in E;\, \dot{V}(y) = 0\}$.

Let $m \in M$ and define $z(t) = \int_0^t S(s)m\,ds$. Since C is closed it follows that $z(t) \in D(C)$ and

$$\dot{z}(t) = Az(t) - BB^*z(t) + m \quad \text{for } -\infty < t < \infty. \tag{3.5}$$

From the definition of M it follows that $B^*S(t)m = 0$ for all $-\infty < t < \infty$. Thus $B^*z(t) = 0$ for all $-\infty < t < \infty$ and we see, using (3.5), that $B^*T(t)m = 0$ for all $-\infty < t < \infty$. Since $A = -A^*$ we know that $T(t) = T^*(-t)$ and in particular this implies $B^*T^*(t)m = 0$ for all $t \geqq 0$. Thus $m = 0$ and, applying Theorem 3.1, we can conclude $S(t)y \to 0$ as $t \to \infty$.

Remark 3.2. In the preceding two theorems we have made compactness hypotheses on the trajectories of the semigroups $S(t)$ and $S^*(t)$ in order to apply Theorem 3.1. There is a special case where these hypotheses are satisfied without any further assumptions. This is the case when $T(t)$ is a compact operator for all $t > 0$. This is the situation, for example, when *A is the infinitesimal generator of a solution of a linear parabolic equation.* When $T(t)$ is compact for all $t > 0$ it then follows that $S(t)$ and $S^*(t)$ are compact for all $t > 0$. Hence, for any y in E, $S(t)y$ and $S^*(t)y$ will lie in compact sets of E for $t \geqq 0$. It should also be noted that if $S(t)$ is compact for all $t > 0$ then $S(t)y \to 0$ weakly as $t \to \infty$ in E implies $S(t)y \to 0$ as $t \to \infty$ in E which, by a result of Datko [9, Thm. 3] implies that $\|S(t)\|_{\mathscr{L}(E,E)} \leqq Me^{-\omega t}$ for $M \geqq 1$ and $\omega > 0$. These results are summarized in the following corollary.

COROLLARY 3.1. *Let A be the infinitesimal generator of a C_0 contraction semigroup $T(t)$ on E. If* (i) *$T(t)$ is a compact operator for every $t > 0$,* (*ii*) (1.1) *is completely controllable, then $\|S(t)\|_{\mathscr{L}(E,E)} \leqq Me^{-\omega t}$ for $M \geqq 1$ and $\omega > 0$.*

The above corollary is of little usefulness in applications since compact semigroups are usually associated with linear parabolic equations which are inherently asymptotically stable. The more interesting case is the one motivated by linear hyperbolic equations, for example, the wave equation, where the uncontrolled system is stable but not asymptotically stable. Furthermore, for linear hyperbolic equations $S(t)$ is never compact for any t and we must verify the compactness hypothesis of Theorems 3.2 and 3.3 by the type of technique used in Theorem 2.2. This is illustrated in the following result, where we show that Assumption 2 will yield the desired compactness properties.

THEOREM 3.4. (i) *If A is the infinitesimal generator of a C_0 contraction semigroup $T(t)$ on E and A^* satisfies Assumption* 2, *then for y in E, $S^*(t)y$ remains in a compact set of E for $t \geqq 0$.*

(ii) *If A is the infinitesimal generator of a C_0 contraction semigroup $T(t)$ on E and A satisfies Assumption* 2, *then for y in E, $S(t)y$ remains in a compact set of E for $t \geqq 0$.*

Proof. (i) Let $P = -BB^*$ and let λ_1 be a value of λ in Lemma 2.1 so that $(A^* - BB^* + \lambda_1 I)^{-1}$ is compact. We know from (3.2) that for $y_0 \in D(A^*)$, $t \geqq 0$, $\|S^*(t)y_0\|_E \leqq \|y_0\|_E$. Using the semigroup property of $S^*(t)$ we then have

$\|(A^* - BB^*)S^*(t)y_0\|_E \leqq \|(A^* - BB^*)y_0\|_E$. This in turn implies, by the triangle inequality, that $\|(A^* - BB^* + \lambda_1 I)S^*(t)y_0\|_E \leqq \|(A^* - BB^*)y_0\|_E + |\lambda_1|\,\|y_0\|_E$. Since $(A^* - BB^* + \lambda_1 I)^{-1}$ is compact, we therefore know that for $y_0 \in D(A^*)$, $S^*(t)y_0$ lies in a compact set of E for $t \geqq 0$. Since $D(A^*)$ is dense in E we can apply Proposition 3.4 of [6] to conclude that for $y \in E$, $S^*(t)y$ lies in a compact set of E for $t \geqq 0$.

(ii) The proof is exactly the same as in (i).

Theorems 3.2, 3.3 and 3.4 can be combined to give the following results.

THEOREM 3.5. *Let A be the infinitesimal generator of a C_0 contraction semigroup $T(t)$ on E for $t \geqq 0$. If* (i) *A^* satisfies Assumption* 2, *and* (ii) (1.1) *is completely controllable, then $S(t)y \to 0$ weakly as $t \to \infty$ for all y in E, i.e.,* (1.1) *is weakly stabilizable.*

THEOREM 3.6. *Let A be the infinitesimal generator of a C_0 contraction group $T(t)$ on E. If* (i) *A satisfies Assumption* 2, (ii) (1.1) *is completely controllable, and* (iii) *$A = -A^*$, then $S(t)y \to 0$ as $t \to \infty$ for all y in E, i.e.,* (1.1) *is stabilizable.*

An example: perturbed wave equation. We consider as a model case the problem of stabilizing the perturbed wave equation. Let

$$A = \begin{bmatrix} 0 & 1 \\ \Delta & 0 \end{bmatrix} + P, \qquad P \in \mathscr{L}(E, E), \tag{3.6}$$

where $E = \overset{\circ}{H}_1(\Omega) \oplus H_0(\Omega)$, Ω some bounded sufficiently smooth spatial domain in R^n, $D(A) = \{H_2(\Omega) \cap \overset{\circ}{H}_1(\Omega)\} \oplus \overset{\circ}{H}_1(\Omega)$, and $(Px, x) \leqq 0$ for all x in E. In this case $\dot{x} = Ax$ represents some stable perturbation of the wave equation and A is the infinitesimal generator of a C_0 semigroup of contractions on E. It follows from Remark 2.2 and Lemma 2.1 that A and A^* satisfy Assumption 2. We can then apply Theorem 3.5 to conclude that for A given by (3.6), complete controllability implies weak stabilizability. Furthermore, in the case $P = 0$ we have from Theorem 3.6 that complete controllability implies stabilizability.

It should be noted that Remark 2.2 implies that the same results as given for the perturbed wave equation can be obtained for more general second order evolution equations as long as the uncontrolled system defines a contraction semigroup.

Appendix. Computation of equation (2.2). Let $w(t) = S_1^*(t)y_0$. Then $w(t) \in D(A^*)$ and

$$\begin{aligned}
&\frac{d}{dt}(w(t), D_{\varepsilon,\lambda}w(t))_E \\
&\quad = 2(D_{\varepsilon,\lambda}A^*w(t), w(t))_E + 2\lambda(D_{\varepsilon,\lambda}w(t), w(t))_E - 2\|B^*w(t)\|_H^2 \\
&\quad = 2\int_0^{\varepsilon} e^{-2\lambda s}(T(-s)BB^*T^*(-s)A^*w(t), w(t))_E\, ds \\
&\qquad + 2\lambda(D_{\varepsilon,\lambda}w(t), w(t))_E - 2\|B^*w(t)\|_H^2 \\
&\quad = -2\int_0^{\varepsilon} e^{-2\lambda s}\left(T(-s)BB^*\frac{d}{ds}T^*(-s)w(t), w(t)\right)_E ds
\end{aligned}$$

(cont.)

$$
\begin{aligned}
&+ 2\lambda(D_{\varepsilon,\lambda}w(t), w(t))_E - 2\|B^*w(t))\|_H^2 \\
&= -2\int_0^\varepsilon e^{-2\lambda s}\left(\frac{d}{ds}B^*T^*(-s)w(t), B^*T^*(-s)w(t)\right)_H ds \\
&+ 2\lambda(D_{\varepsilon,\lambda}w(t), w(t))_E - 2\|B^*w(t)\|_H^2 \\
&= -\int_0^\varepsilon \frac{d}{ds}\|e^{-\lambda s}B^*T^*(-s)w(t)\|_H^2\,ds - 2\|B^*w(t)\|_H^2 \\
&= -\|e^{-\lambda\varepsilon}B^*T^*(-\varepsilon)S_1^*(t)y_0\|_H^2 - \|B^*S_1^*(t)y_0\|_H^2.
\end{aligned}
$$

Computation of equation (3.3). For $x \in D(A^*)$ we know

$$\frac{d}{dt}S^*(t)x = A^*S^*(t)x - BB^*S^*(t)x.$$

Integrating with respect to t from 0 to t yields

$$S^*(t)x - x = A^*\int_0^t S^*(s)x\,ds - BB^*\int_0^t S^*(s)x\,ds,$$

where the fact that C^* is closed is used to move $A^* - BB^*$ outside the integral. Since $D(A^*)$ is dense in E and A^* is closed the above equation holds for all x in E and in particular for $x = m$. Equation (3.3) follows immediately since $z(t) = \int_0^t S^*(s)m\,ds$.

Acknowledgment. The author would like to thank Prof. D. Henry of the University of Kentucky for pointing out an observation by Prof. F. Massey which yielded the improvements over the results given in [1]. Also, the author would like to thank Prof. C. Dafermos of Brown University for several stimulating conversations.

REFERENCES

[1] M. SLEMROD, *The linear stabilization problem in Hilbert space*, J. Functional Analysis, to appear.

[2] H. O. FATTORINI, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.

[3] ———, *Controllability of higher-order linear systems*, Mathematical Theory of Control, A. V. Balakrishnan and L. Neustadt, eds., Academic Press, New York, 1967, pp. 301–311.

[4] J. K. HALE, *Dynamical systems and stability*, J. Math. Anal. Appl., 26 (1969), pp. 39–59.

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

[6] C. M. DAFERMOS, *Uniform processes and semicontinuous Lyapunov functionals*, Ibid., 11 (1972), pp. 401–415.

[7] ———, *Applications of the invariance principle for compact processes. I. Asymptotically dynamical systems*, Ibid., 9 (1971), pp. 291–299.

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

[9] R. DATKO, *Uniform asymptotic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428–445.

[10] L. BERS, F. JOHN and M. SCHECTER, *Partial Differential Equations*, Interscience, New York, 1964.

[11] D. L. LUKES, *Stabilizability and optimal control*, Funkcial. Ekvac., 11 (1968), pp. 39–50.

SIAM J. CONTROL
Vol. 12, No. 3, August 1974

AN EXISTENCE THEOREM FOR PENALTY FUNCTION THEORY*

J. P. EVANS† AND F. J. GOULD‡

Abstract. A necessary condition is presented for the existence of global maxima of penalty functions without a barrier. This condition with one additional restriction is proved to be sufficient.

1. Introduction. Consider the nonlinear programming problem

$$\text{maximize} \quad f(x)$$
$$\text{subject to} \quad g_j(x) \leqq 0, \qquad j = 1, \cdots, m,$$

where $f: R^n \to R$, $g_j: R^n \to R$, $j = 1, \cdots, m$, and all functions are assumed continuous on R^n. A frequently employed solution technique is to compose, using f and the g's, a penalty function which often contains a scalar parameter, say δ. The penalty function $P(x, \delta)$ is formed in such a way that, hopefully, the following properties are obtained:

(i) For each parameter value δ_k the function $P(x, \delta_k)$ has a global maximum in x, say at x^k.

(ii) As $\delta_k \to \infty$ the points x^k (or, more typically, any convergent subsequence of these points) tend to a solution to the nonlinear program.

Numerous classes of penalty functions have appeared in the literature and considerable research has been devoted to both the theoretic and the computational aspects of their usage. The two most heavily studied classes are barrier functions and exterior functions. A third less studied class is the exponential functions. An excellent survey of the history of penalty function development is found in the book by Fiacco and McCormick [6].

At this time most of the theory is sufficiently well developed that the thrust of interest is in the computational sphere. However, there is an important loose end concerning the existence of the unconstrained maxima x^k. For the barrier methods the existence is guaranteed by the method of construction of the function P. For the classes described as exponential and exterior the existence is not guaranteed and it is a trivial matter to construct examples where $P(x, \delta)$ has no global maximum in x. In previous works on these latter classes any mention of this problem has usually been limited to the assumption that the maximum does indeed exist. The purpose of this paper is to clarify those circumstances under which such an assumption is valid. In particular, for a general class of penalty functions a global growth rate condition is shown to be necessary for the existence of global maxima; with an additional mild restriction this growth rate is sufficient. The class of penalty functions will include the above-mentioned exponential and exterior functions.

* Received by the editors March 15, 1973, and in revised form June 22, 1973. This work was supported in part by the Office of Naval Research under Contract N000-14-67-A-0321-0003(NR-047-095).

† School of Business Administration and Curriculum in Operations Research and Systems Analysis, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514.

‡ Graduate School of Business, University of Chicago, Chicago, Illinois 60637.

In certain favorable circumstances our condition is operationally verifiable, though in practice such verifications are rarely attempted and the result is presented mainly for the sake of theoretic completeness.

2. A class of multiplier functions.

In this section we introduce a class of multiplier functions on which the subsequent discussion of existence results will be based. A representative list of examples indicates roughly the generality of the class.

Let $\lambda : R \times A \to RU\{+\infty\}$, where $A = \{\alpha \in R : \alpha \geqq 0\}$. Define the *class* M of multiplier functions to be that set of functions λ satisfying the following properties:

1. $\lambda(\xi, \alpha)$ is continuous on $R \times A$, nondecreasing in $\xi \leqq 0$, increasing in $\xi > 0$ and $\lambda(0, \alpha) \leqq 1$ for all $\alpha \in A$.
2. $\xi > 0, \alpha > 0$ implies $\lambda(\xi, \alpha) > 0$.
3. For each $\xi > 0$, $\lambda(\xi, \alpha)$ is increasing in α, and $\lambda(\xi, \alpha) \to \infty$ as $\alpha \to \infty$.
4. For each $\xi \in R$, $\lambda(\xi, \alpha)$ is convex in α.
5. $\lambda(\xi, \alpha) \geqq -1$ for all $\xi \in R$, $\alpha \in A$.

Examples.

$$\lambda(\xi, \alpha) = e^{\alpha\xi} - 1, \tag{2.1a}$$

$$\lambda(\xi, \alpha) = e^{\alpha\xi}; \tag{2.1b}$$

$$\lambda(\xi, \alpha) = \begin{cases} 0, & \xi \leqq 0, \\ \alpha\xi^p, & \xi > 0, \end{cases} \tag{2.2}$$

where p is a fixed positive integer;

$$\lambda(\xi, \alpha) = \begin{cases} 0, & \xi \leqq 0, \\ e^{\alpha\xi} - 1, & \xi > 0. \end{cases} \tag{2.3}$$

Next we define the penalty function $P: R \times A \times A^+ \to RU\{-\infty\}$ by

$$P(x; \alpha, \beta) = f(x) - \beta \sum_{j=1}^{m} \lambda(g_j(x), \alpha),$$

where $\lambda \in M, \beta \in A^+ = \{\beta \in R : \beta > 0\}$. If λ is chosen as in example (2.1b) and $\beta = 1$, we have the exponential penalty function studied by Allran and Johnsen [2]. Again taking λ as in (2.1b) but now allowing β to vary in A^+ subject to the restriction $\alpha \geqq 1/\beta \geqq 1$ we have the more general exponential function investigated by Murphy [7] and Evans and Gould [4]. The λ of example (2.2) in conjunction with $\beta = 1$ is the potential function of Pietrzykowski [8] and includes as a special case the quadratic loss function investigated by Ablow and Brigham [1], Camp [3], and Zangwill [9]. Examples (2.2) and (2.3) give rise to penalty functions of the exterior type. That is, no penalty is assessed in the feasible region, but a penalty, which increases with the extent of constraint violation, is levied on each infeasible point.

Observe that example (2.1b) assigns a finite positive penalty to each point at which a constraint is satisfied; in particular, this applies to points in the boundary of the feasible region. Intuitively, this places such functions in an intermediate position between barrier functions and the exterior penalty function mentioned above. Section 4 contains a more precise description of the properties of this class and shows that under favorable conditions a weak barrier property holds; namely, the unconstrained maximizers are *eventually* all feasible, i.e., within the constraint set $\{x \in R^n : g_j(x) \leqq 0, j = 1, \cdots, m\}$.

3. Existence of a penalty function maximizer. The major step in penalty methods is the unconstrained maximization of the penalty function. For compact constraint sets this maximum is guaranteed to exist for barrier functions by the assignment of an infinite penalty to boundary and infeasible points. However, for the class of functions of interest in this paper it is easy to construct programming problems such that the resultant penalty function has no unconstrained maximum.[1] In this section we develop necessary and sufficient conditions for $P(x; \alpha, \beta)$ to have an unconstrained maximum.

All results in this section are for penalty functions constructed with multiplier functions from the class M defined in § 2. In addition we shall employ a growth rate condition the satisfaction of which depends upon the programming problem under consideration. First define

$$S_\delta = \{x \in R^n : g_j(x) \leqq \delta, j = 1, \cdots, m\},$$

$$\tilde{g}(x) = \max \{g_j(x) : j = 1, \cdots, m\}.$$

The growth rate property to be exploited is:

C1. For each $\delta > 0$ there exist scalars $\bar{\alpha} > 0$ and $K > 0$ such that $f(x) \leqq K\lambda(\tilde{g}(x), \bar{\alpha})$ for all $x \in R^n - S_\delta$.

Observe that if condition C1 holds for $\bar{\alpha} \in A$, property 3 of λ implies that C1 holds for each $\alpha \geqq \bar{\alpha}$. Moreover, if C1 holds for $K = \bar{K}$ it also holds for $K \geqq \bar{K}$.

THEOREM 3.1. *Suppose condition* C1 *is satisfied and that for some* $\delta > 0$ *the set* S_δ *is compact. Then for each* $\beta > 0$ *there is an* $\alpha^* \in A$ *such that* $\alpha \geqq \alpha^*$ *implies that* $P(x; \alpha, \beta)$ *has a global maximum and* each *global maximizer belongs to* S_δ.

Proof. Let $\beta > 0$ be given. Let $\bar{x} \in S_0 = \{x \in R^n : g_j(x) \leqq 0, j = 1, \cdots, m\}$. Then from property 1 of λ,

$$\sum_{j=1}^{m} \lambda(g_j(\bar{x}), \alpha) \leqq m \quad \text{for all } \alpha \in A.$$

Hence

$$\max [P(x; \alpha, \beta) : x \in S_\delta] \geqq f(\bar{x}) - m\beta \quad \text{for all } \alpha \in A.$$

Now suppose $x \in R^n - S_\delta$; thus $\tilde{g}(x) > \delta$. From condition C1 there exist constants $\bar{\alpha} \in A$ and $K > 0$ such that $f(x) \leqq K\lambda(\tilde{g}(x), \bar{\alpha})$ for each $x \in R^n - S_\delta$. By property 4 of λ for $p \geqq 1$ we have

$$p\lambda(\xi, \alpha) \leqq \lambda(\xi, p\alpha) + (p - 1)\lambda(\xi, 0).$$

Without loss of generality we can assume $2K > \beta$. Thus for $\alpha \geqq \hat{\alpha} = 2K\bar{\alpha}/\beta$

[1] For example, take $\lambda(\cdot)$ as in (2.2) with $p = 2, f(x) = x^5$ and $g(x) = x^2 - 1$, where $x \in R$.

we have for $x \in R^n - S_\delta$,

$$\begin{aligned} f(x) \leqq K\lambda(\tilde{g}(x), \bar{\alpha}) &= \frac{\beta}{2}\left(\frac{2K}{\beta}\right)\lambda(\tilde{g}(x), \bar{\alpha}) \\ &\leqq \frac{\beta}{2}\lambda(\tilde{g}(x), \hat{\alpha}) + \frac{\beta}{2}\left(\frac{2K-\beta}{\beta}\right)\lambda(\tilde{g}(x), 0) \qquad (3.1) \\ &\leqq \frac{\beta}{2}\lambda(\tilde{g}(x), \alpha) + \left(\frac{2K-\beta}{2}\right)\lambda(\tilde{g}(x), 0) \end{aligned}$$

in which the last inequality follows from property 3 of λ. By this same property we can choose $\tilde{\alpha}$ such that $\alpha \geqq \tilde{\alpha}$ implies

$$(3.2) \qquad -f(\bar{x}) + (2m-1)\beta \leqq \frac{\beta}{2}\lambda(\tilde{g}(x), \alpha) - \left(\frac{2K-\beta}{2}\right)\lambda(\tilde{g}(x), 0).$$

Then for $\alpha \geqq \alpha^* = \max[\alpha, \tilde{\alpha}]$, (3.1) and (3.2) yield

$$f(x) - f(\bar{x}) + (2m-1)\beta \leqq \beta\lambda(\tilde{g}(x), \alpha).$$

By property 5 of λ, $\lambda(g_j(x), \alpha) \geqq -1$, $j = 1, \cdots, m$, for each $x \in R^n$. Thus for $x \in R^n - S_\delta$ we have

$$\lambda(\tilde{g}(x), \alpha) \leqq \sum_{j=1}^{m} \lambda(g_j(x), \alpha) + (m-1).$$

Thus $x \in R^n - S_\delta$ and $\alpha \geqq \alpha^*$ yields

$$P(x; \alpha, \beta) = f(x) - \beta \sum_{j=1}^{m} \lambda(g_j(x), \alpha) \leqq f(\bar{x}) - m\beta \leqq \max[P(x; \alpha, \beta) : x \in S_\delta].$$

The conclusion follows.

The following result is a close relative of Theorem 3.1 and can be established from the conclusion of that theorem.

THEOREM 3.2. *Suppose the conditions of Theorem* 3.1 *hold. Let* $\beta > 0$ *be given and suppose* $\alpha^* \in A$ *is such that each maximizer of* $P(x; \alpha^*, \beta)$ *belongs to* S_δ. *Let* $\hat{P}(x)$ *be any continuous function such that*

(i) $\hat{P}(x) \geqq P(x; \alpha^*, \beta)$, $x \in S_\delta$,

(ii) $\hat{P}(x) \leqq P(x; \alpha^*, \beta)$, $x \in R^n - S_\delta$.

Then $\hat{P}(x)$ *has a global maximizer in* R^n *and each such maximizer belongs to* S_δ.

We now show that if $P(x; \alpha, \beta)$ has a global maximum on R^n then condition C1 must be valid.

THEOREM 3.3. *Let* $\alpha_0 \in A$, $\beta_0 > 0$ *be given and suppose* $P(x; \alpha_0, \beta_0)$ *has a global maximum on* R^n. *Then condition* C1 *is satisfied.*

Proof. Let $\delta > 0$ be given; we must show that there exist positive scalars α, K such that

$$f(x) \leqq K\lambda(\tilde{g}(x), \alpha) \quad \text{for all } x \in R^n - S_\delta.$$

Let

$$\tilde{v} = \max[P(x; \alpha_0, \beta_0) : x \in R^n] \geqq f(x) - \beta_0 \sum_{j=1}^{m} \lambda(g_j(x), \alpha_0)$$

for each $x \in R^n$. Since, for each j, $g_j(x) \leqq \tilde{g}(x)$, we obtain, from property 1 of λ,

$$\sum_{j=1}^{m} \lambda(g_j(x), \alpha) \leqq m\lambda(\tilde{g}(x), \alpha) \quad \text{for all } \alpha \in A;$$

hence

$$P(x; \alpha, \beta_0) \geqq f(x) - m\beta_0\lambda(\tilde{g}(x), \alpha) \quad \text{for all } \alpha \in A.$$

Now we can choose $K > 2m\beta_0$ and $\bar{\alpha} > \alpha_0$ such that

$$K\lambda(\delta, \bar{\alpha}) \geqq 2\tilde{v}$$

by property 3 of λ. Then for $x \in R^n - S_\delta$,

$$f(x) \leqq \tilde{v} + m\beta_0\lambda(\tilde{g}(x), \alpha_0) \leqq \frac{K}{2}\lambda(\delta, \bar{\alpha}) + m\beta_0\lambda(\tilde{g}(x), \alpha_0)$$

$$\leqq K\lambda(\tilde{g}(x), \bar{\alpha}),$$

where the last inequality holds by properties 1, 2, and 3 of λ. Since δ was arbitrary, the theorem is proved.

It should be noted that Theorem 3.1 shows that conditions C1 *and* the compactness of S_δ, for some $\delta > 0$, is a sufficient condition for the existence of a penalty function maximizer. Theorem 3.3 shows that C1 alone is necessary, without regard to compactness of S_δ. In this sense C1 is necessary and nearly sufficient. It should be clear that verification of condition C1 for any specific problem requires a certain amount of knowledge of the functions involved. In various cases this verification may be more or less difficult or even impossible. We now show that if the λ functions are convex in the first argument, then C1 will always hold for the functions in a concave programming problem with a compact constraint set.

THEOREM 3.4. *Suppose that the multiplier function λ satisfies properties* 1 *through* 5 *of* § 2 *and the following additional property.*

6. *For each $\alpha \geqq 0$, $\lambda(\xi, \alpha)$ is convex in ξ over $\{\xi : \xi \geqq 0\}$. Assume the nonlinear program is concave with f continuously differentiable and that the constraint set S_0 is compact. Then for each $\delta > 0$ there exist positive scalars $\bar{\alpha}$, K such that*

$$f(x) \leqq K\lambda(\tilde{g}(x), \bar{\alpha}) \quad \textit{for all } x \in R^n - S_\delta.$$

Proof. Let $\delta > 0$ be given. Since the constraint functions are convex and since S_0 is compact it follows that, for each $\delta > 0$, the set S_δ is also compact. Since $S_0 \neq \varnothing$ it follows that, for some x^0, $\tilde{g}(x^0) \leqq 0$, and hence

$$g_j(x^0) < \delta, \qquad j = 1, \cdots, m.$$

It then follows from Lemma 3.1 of [5] that there is an $\varepsilon > 0$ such that for each $x \in R^n - S_\delta$ there is a $z \in \partial S_\delta$ and a constraint active at z (i.e., $g_j(z) = \delta$) such that

$$g_j(x) \geqq \varepsilon\|x - z\|, \tag{3.3}$$

where z and j may depend on x, but ε does not. Also note that since S_δ is compact there exist constants p_1 and p_2 such that for each $z \in \partial S_\delta$,

$$f(z) \leqq p_1 \quad \text{and} \quad \|\nabla f(z)\| \leqq p_2. \tag{3.4}$$

Now let $x \in R^n - S_\delta$ and let z be as in (3.3). Then

$$\begin{aligned} f(x) &\leqq f(z) + (x - z)^T \nabla f(z) && \text{(by concavity of } f\text{)} \\ &\leqq p_1 + p_2 \|x - z\| && \text{(by (3.4))} \\ &\leqq p_1 + p_2 g_j(x)/\varepsilon && \text{(by (3.3))} \\ &\leqq p_1 + p_2 \tilde{g}(x)/\varepsilon. \end{aligned}$$

Hence by property 3 of λ there is an $\bar{\alpha} > 0$ such that

$$f(x) \leqq \lambda(\delta, \bar{\alpha}) + \lambda(\delta, \bar{\alpha}) \tilde{g}(x). \tag{3.5}$$

Now since $\lambda(\xi, \alpha)$ is convex in ξ for $\alpha \geqq 0$, for any $\xi \geqq \delta$ we have

$$\lambda(\delta, \alpha) \leqq \frac{\delta}{\xi} \lambda(\xi, \alpha) + \frac{\xi - \delta}{\xi} \lambda(0, \alpha)$$

which can be rearranged in the form

$$\xi[\lambda(\delta, \alpha) - \lambda(0, \alpha)] \leqq \delta[\lambda(\xi, \alpha) - \lambda(0, \alpha)]$$

in which both terms in square brackets are positive by property 1 of λ. Hence $\xi \geqq \delta$ implies that

$$\xi \lambda(\delta, \alpha) \leqq \frac{\delta \lambda(\delta, \alpha)}{\lambda(\delta, \alpha) - \lambda(0, \alpha)} \lambda(\xi, \alpha) - \frac{\delta \lambda(0, \alpha)}{\lambda(\delta, \alpha) - \lambda(0, \alpha)} \lambda(\delta, \alpha)$$

in which the second term on the right is nonnegative by the properties of λ. Thus letting $K' = \delta \lambda(\delta, \bar{\alpha}) / [\lambda(\delta, \bar{\alpha}) - \lambda(0, \bar{\alpha})]$ yields

$$\tilde{g}(x) \lambda(\delta, \bar{\alpha}) \leqq K' \lambda(\tilde{g}(x), \bar{\alpha}) \quad \text{for all } x \in R^n - S_\delta.$$

Now if we let $K = K' + 1$, substituting this last inequality in (3.5) yields

$$\begin{aligned} f(x) &\leqq \lambda(\delta, \hat{\alpha}) + K' \lambda(\tilde{g}(x), \bar{\alpha}) \\ &\leqq \lambda(\tilde{g}(x), \bar{\alpha}) + K' \lambda(\tilde{g}(x), \bar{\alpha}) = K \lambda(\tilde{g}(x), \bar{\alpha}) \end{aligned}$$

for each $x \in R^n - S_\delta$, which is the desired result.

Remark. Note that the Slater condition is not required.

4. Application of the existence theorem. In this section several specific relations to previous works are identified. Let

$$I = \{x \in R^n : g_j(x) < 0, \ j = 1, \cdots, m\}$$

and as before

$$S_\delta = \{x \in R^n : g_j(x) \leqq \delta, \ j = 1, \cdots, m\}.$$

It will be assumed that I is nonempty and that the constraint set S_0 is compact.

A. *Exponential penalty functions.* These functions, exemplified by the first example in § 2, have been investigated by Allran and Johnsen [2] and Murphy [7]. In 1970, Allran and Johnsen studied the exponential penalty function determined by selecting λ as in example (2.1b). That is,

$$P(x; \alpha, 1) = f(x) - \sum_{j=1}^{m} e^{\alpha g_j(x)}.$$

Though the work involved in their primary results is substantially in error, correct arguments can be provided to show that without convexity assumptions:

(i) For all α sufficiently large, $\max_{x\in I} P(x;\alpha,1)$ is attained for some x^α in I. This implies that, in particular, since I is open, $P(x;\alpha,1)$ has a local maximum in I for all α sufficiently large.

(ii) As $\alpha \to \infty$ any convergent subsequence of x^α (and there must be at least one since $x^\alpha \in I \subset S_0$, a compact set) converges to a solution of the problem $\max\,[f(x):x \in \text{closure } I]$.

(iii) Consequently if closure $I = S_0$ the subsequence converges to a nonlinear programming solution.

The above results reduce the original constrained problem to another constrained problem ($\max_{x\in I} P(x;\alpha,1)$) which in general may be no more convenient to solve than the original problem. However, in the special case where $P(x;\alpha,1)$ is strictly concave in x it follows immediately that for α sufficiently large, $\max_{x\in R^n} P(x;\alpha,1)$ exists at some unique x^α which is in I. This then facilitates matters by allowing for the use of sequential unconstrained methods on the penalty function.

In a 1971 investigation by Murphy, again taking λ as in example (2.1b) but allowing β to vary subject to the restriction $\alpha \geqq 1/\beta \geqq 1$, it has been shown that:

(iv) In the general (not necessarily concave) case, if closure $I = S_0$ and a sequence of penalty function maximizers exists, and is eventually in a compact set, then each convergent subsequence has a limit point which solves the original problem.

(v) In the differentiable concave case (but without strict concavity), when the product $\alpha\beta$ is sufficiently large, $\max_{x\in R^n} P(x;\alpha,\beta)$ exists and moreover all of the maximizing values of x are feasible (though not necessarily in I as in (iii) above).

Our existence Theorem 3.1 supplements this development by providing conditions directly related to the structure of the original problem which are sufficient for the existence of the penalty function maximizers in a compact set as required in (iv) above.

B. *Exterior methods.* In this section the multiplier functions λ are selected as in examples (2.2) and (2.3), and we suppose that for some $\delta > 0$ the set S_δ is compact. Then without further assumptions the following results are true.

(vi) Select any γ such that $0 < \gamma < \delta$. Then for all α sufficiently large, $\max_{x\in S_\delta} P(x;\alpha,\beta)$ is achieved at some $x^\alpha = S_\gamma$.

(vii) As $\alpha \to \infty$ any convergent subsequence of the x^α's (there is at least one) converges to a solution to the nonlinear program.

Thus far, again, this result merely replaces one constrained problem with another. However, Theorem 3.1 is directly applicable. If the growth rate assumption C1 is satisfied, then the operation $\max_{x\in S_\delta}$ can be replaced with $\max_{x\in R^n}$, the above same results (vi) and (vii) hold, and unconstrained optimization can be used on the penalty function. In most expositions on exterior methods maximization over R^n is required, but nowhere, to our knowledge, is there any existence justification in nonconcave problems.

REFERENCES

[1] C. M. Ablow and G. Brigham, *An analog solution of programming problems*, Operations Research, 3 (1955), pp. 388–394.

[2] R. R. Allran and S. E. V. Johnsen, *An algorithm for solving nonlinear programming problems subject to nonlinear inequality constraints*, Computer J., 13 (1970), pp. 171–177.

[3] G. D. Camp, *Inequality constrained stationary value problems*, Operations Research, 3 (1955), pp. 548–550.

[4] J. P. Evans and F. J. Gould, *Stability and exponential penalty function techniques in nonlinear programming*, Institute of Statistics Mimeo Series No. 723, Department of Statistics, University of North Carolina at Chapel Hill, 1970.

[5] J. P. Evans, F. J. Gould and J. W. Tolle, *Exact penalty functions in nonlinear programming*, Department of Statistics, University of North Carolina at Chapel Hill, 1972; Mathematical Programming, 4 (1973), pp. 72–97.

[6] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[7] F. H. Murphy, *A class of exponential penalty functions*, Tech. Rep., Graduate School of Business, Rutgers University, Newark, N.J., 1971.

[8] T. Pietrzykowski, *Application of the steepest descent method to concave programming*, Proc. IFIPS Congress (Munich), North-Holland, Amsterdam, 1962.

[9] W. I. Zangwill, *Nonlinear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.

SIAM J. CONTROL
Vol. 12, No. 3, August 1974

REALIZATION AND STRUCTURE THEORY OF BILINEAR DYNAMICAL SYSTEMS*

PAOLO D'ALESSANDRO,† ALBERTO ISIDORI† AND ANTONIO RUBERTI‡

Abstract. Starting from the description of the system provided by the series expansion of the zero-state response, this paper develops the realization theory for bilinear systems. It is shown that the condition of realizability corresponds to that of the factorizability of the kernels of this expansion. The paper then analyzes the properties of the minimal and nonminimal factorizations and provides a solution of the problem of characterizing the minimal realizations. Subsequently, developing the structure analysis of the state space, it is shown that there exists a canonical form of the equations and the results of the realization theory are interpreted on this basis. Lastly, the bases are laid for developing realization procedures, and it is shown that the sequence of kernels of a bilinear system is completely identified by a finite number of these kernels.

1. Introduction. In this paper we present a complete realization theory of bilinear systems, a class of nonlinear systems which has raised considerable interest in the last few years (see, for example, [8]).

The system model considered in this theory is the sequence of kernels characterizing the Volterra series expansion of the zero-state response. It is shown first that a necessary and sufficient condition for the existence of bilinear realizations can be expressed as a sort of factorizability property (§ 3). Then we present quite a complete analysis of the sequences that satisfy this condition (§ 4), in order to derive the tools for studying minimal bilinear realizations. A complete characterization of these realizations reveals the interest in considering, besides the dimension of the state space, also the least number of multipliers needed to implement the nonlinear map. The analysis shows that this number assumes its minimum just over the minimal realizations. These, moreover, are a single equivalence class (§ 5).

As for linear systems, the minimal realization theory can be connected with the structure analysis of the state space. This latter is developed here on the basis of reachability (from the origin) and unobservability. As a result, we obtain a canonical form of the equations and we show that the kernels of the Volterra series expansion of the zero-state response depend only on one of the subsystems identified in the decomposition. This subsystem provides, modulo an equivalence, a minimal bilinear realization (§ 6). Finally, we present an outline of the procedures for constructing minimal bilinear realizations (§ 7).

For the sake of brevity and notational simplicity, the analysis is presented here for single-input, single-output systems; the extension to the multidimensional case can be found in [4]. Some results of the present paper (statement (c) of Theorem 4 and Theorem 8) have been proved, simultaneously and independently, by R. W. Brockett [1].

* Received by the editors June 8, 1972, and in revised form May 21, 1973. This work was supported in part by the National Council of Research of Italy.

† Istituto di Automatica, University of Rome, 00184 Rome, Italy.

‡ Istituto di Automatica, University of Rome, 00184 Rome, Italy, and Fondazione U. Bordoni, Rome, Italy.

2. Bilinear equations and their solutions. The input-state-output equations of a single-input single-output bilinear system are usually written in the form

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

where $u(t) \in R$ is the input, $y(t) \in R$ is the output, and $x(t) \in R^n$ is the state at time t. The matrices A, N, B, C are constant matrices of suitable dimensions.

If the system described by equations (1) is in the initial state x_0 at time $t = 0$, the response $y(t)$ can be decomposed into the sum

$$y(t) = y_x(t) + y_u(t) + y_{xu}(t), \tag{2}$$

in which the first two terms respectively represent the zero-input response and the zero-state response.

The expression for $y_x(t)$ can be obtained immediately by putting $u(t) = 0$ into equations (1), i.e.,

$$y_x(t) = C\,e^{At}x_0. \tag{3}$$

If the input belongs to the class of functions bounded on finite intervals [2], the response $y_u(t)$ can be expanded in a Volterra series with symmetrical kernels according to the expression

$$y_u(t) = \sum_{i=1}^{\infty} \frac{1}{i!} \int_0^t \cdots \int_0^t w_i(t_1, \cdots, t_i) \left[\prod_{k=1}^{i} u(t - t_k) \right] dt_1 \cdots dt_i. \tag{4}$$

The symmetrical kernels of series (4) can be calculated from

$$w_i(t_1, \cdots, t_i) = \sum_{\text{per}} v_i(t_1, \cdots, t_i), \tag{5}$$

where the summation is carried out over all the $i!$ permutations of the variables $t_1, \cdots, t_i$. The functions $v_i(t_1, \cdots, t_i)$ are given by

$$v_1(t_1) = C\,e^{At_1}B, \tag{6a}$$

$$\begin{aligned} v_i(t_1, \cdots, t_i) = {} & C\,e^{At_i}N\,e^{A(t_{i-1}-t_i)} \cdots N\,e^{A(t_1-t_2)} \\ & \cdot B\left[\prod_{k=0}^{i-2} \delta_{-1}(t_{k+1} - t_{k+2})\right], \qquad i > 1, \end{aligned} \tag{6b}$$

where $\delta_{-1}(t)$ denotes the unit step function.

On the basis of the same assumptions, the response $y_{xu}(t)$ can likewise be expanded in a Volterra series according to the expression

$$y_{xu}(t) = \sum_{i=1}^{\infty} \frac{1}{i!} \int_0^t \cdots \int_0^t z_i(t_1, \cdots, t_i\; N)e^{A(t-t_1)}x_0 \left[\prod_{k=1}^{i} u(t - t_k) \right] dt_1 \cdots dt_i. \tag{7}$$

Here the expressions for the symmetrical kernels $z_i(t_1, \cdots, t_i)$, $i = 1, \cdots, \infty$, can be determined from the expressions (5) and (6) of the kernels $w_i(t_1, \cdots, t_i)$, by replacing the matrix B by an $n \times n$ identity matrix.

3. Realizability conditions. Starting from expression (4) of the zero-state response of system (1), it is natural to assume the following definition of realizability.

Definition 1. A sequence $\{w_i(t_1, \cdots, t_i)\}_1^\infty$ of symmetrical kernels of a Volterra series expansion is realizable by means of a constant bilinear dynamical system with finite-dimensional state space (briefly: is bilinearly realizable) if there exist four constant matrices A, N, B, C, respectively $n \times n$, $n \times n$, $n \times 1$, $1 \times n$, such that the following relations are satisfied:

$$C\, e^{At_1}B = w_1(t_1) \quad \text{for all } t_1, \tag{8a}$$

$$\begin{aligned} &\sum_{\text{per}} C\, e^{At_i}N\, e^{A(t_{i-1}-t_i)} \cdots N\, e^{A(t_1-t_2)}B\left[\prod_{k=0}^{i-2} \delta_{-1}(t_{k+1} - t_{k+2})\right] \\ &\quad = w_i(t_1, \cdots, t_i) \quad \text{for all } t_k, \quad k = 1, \cdots, i, \quad \text{for all } i > 1. \end{aligned} \tag{8b}$$

The first result concerns a condition of realizability, and is expressed by the following theorem.

Theorem 1. *A necessary and sufficient condition for a sequence $\{w_i(t_1, \cdots, t_i)\}_1^\infty$ of symmetrical kernels of a Volterra series expansion to be bilinearly realizable is*:

(a) *that $w_1(t_1)$ have a proper rational Laplace transform*;

(b) *that there exist three matrices $F(t)$, $G(t)$, $H(t)$, respectively $m \times m$, $m \times 1$, $1 \times m$, of functions with proper rational Laplace transforms, such that the following relations are satisfied*:

$$w_i(t_1, \cdots, t_i) = H(t_i)F(t_{i-1} - t_i) \cdots F(t_2 - t_3)G(t_1 - t_2) \tag{9a}$$

on

$$S_i = \{(t_1, \cdots, t_i): t_1 > t_2 > \cdots > t_i\} \quad \textit{for all } i > 1. \tag{9b}$$

Remark 1. The above condition also holds if the variables $t_1, \cdots, t_i$ in (9) are permutated in any way.

Proof. Necessity. If the sequence $\{w_i(t_1, \cdots, t_i)\}_1^\infty$ is realizable, then, by hypothesis, there exist four matrices A, N, B, C such that equations (8) are satisfied. The first of these implies the condition (a). Equation (8b), considered over the sets S_i defined above, reduces to

$$C\, e^{At_i}N\, e^{A(t_{i-1}-t_i)} \cdots N\, e^{A(t_1-t_2)}B = w_i(t_1, \cdots, t_i) \quad \text{for all } i > 1. \tag{10}$$

If the matrix N is now factorized in the form

$$N = N'N'' \tag{11}$$

where N' is $n \times m$ and N'' is $m \times n$, it is immediately seen that (10) can be transformed into the form (9), with the substitutions

$$C\, e^{At}N' = H(t), \qquad N''\, e^{At}N' = F(t), \qquad N''\, e^{At}B = G(t). \tag{12}$$

Since all these functions have proper rational Laplace transforms, it follows that condition (b) is also satisfied.

Sufficiency. Suppose that (a) and (b) are both satisfied, and consider the matrix

$$L(t) = \begin{pmatrix} w_1(t) & H(t) \\ G(t) & F(t) \end{pmatrix}. \tag{13}$$

Since all the elements of $L(t)$ have proper rational Laplace transforms, this may be interpreted as the weighting pattern of a constant linear system of finite order with $m + 1$ outputs and $m + 1$ inputs. Consequently, there must exist three matrices A, R, S, respectively $n \times n$, $n \times (m + 1)$, $(m + 1) \times n$, such that

$$S\,e^{At}R = L(t). \tag{14}$$

By partitioning S and R in the form

$$S = \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}, \qquad R = (R_1 \quad R_2), \tag{15}$$

where S_1 is a $1 \times n$ matrix and R_1 an $n \times 1$ matrix, one obtains

$$\begin{aligned} w_1(t) &= S_1\,e^{At}R_1, & H(t) &= S_1\,e^{At}R_2, \\ G(t) &= S_2\,e^{At}R_1, & F(t) &= S_2\,e^{At}R_2. \end{aligned} \tag{16}$$

Now substitute (16) in (9) and let

$$B = R_1, \quad C = S_1, \quad N = R_2S_2. \tag{17}$$

It follows that the bilinear system characterized by the matrix A and the matrices N, B, C defined in (17) verifies the equations in (8) over the sets S_i.

On the other hand, since $w_i(t_1, \cdots, t_i)$ is symmetrical by definition, equations (8) are satisfied for all values of the variables $t_1, \cdots, t_i$. This completes the proof.

4. Factorizable sequences of kernels. The result established in Theorem 1 naturally leads to the consideration of those subsequences $\{w_i(t_1, \cdots, t_i)\}_2^\infty$ which can be expressed in the form (9). In the sequel such sequences will be called *factorizable* and the triplet $\{F(t), G(t), H(t)\}$ a *factorization.*

In order to analyze the properties of these sequences, it is convenient to introduce the following matrices (note that $F(t)$ is $m \times m$, $G(t)$ is $m \times 1$ and $H(t)$ is $1 \times m$):

$$P_k[F, G](t_1, \cdots, t_k) = [G(t_1) \quad F(t_2)G(t_1) \cdots F(t_k)F(t_{k-1}) \cdots G(t_1)], \tag{18}$$

$$Q_k[F, H](t_1, \cdots, t_k) = \begin{pmatrix} H(t_1) \\ H(t_1)F(t_2) \\ \vdots \\ H(t_1) \cdots F(t_{k-1})F(t_k) \end{pmatrix}. \tag{19}$$

We shall be concerned with the problem of testing whether the m rows of $P_k[F, G]$ (the m columns of $Q_k[F, H]$) are linearly independent functions or not; since $F(t)$, $G(t)$ and $H(t)$, by hypothesis, are functions with proper rational Laplace transform (see Theorem 1), it will be sufficient to test the linear independence over some subinterval of R^k, e.g., over $\Delta_k = \{(t_1, \cdots, t_k): 0 \leqq t_i \leqq 1, \text{ for all } i\}$. This leads to introducing the Gramian matrices

$$\mathscr{P}_k[F, G] = \int_{\Delta_k} P_k P_k^* \, dt_1 \cdots dt_k, \tag{20}$$

$$\mathscr{Q}_k[F, H] = \int_{\Delta_k} Q_k^* Q_k \, dt_1 \cdots dt_k. \tag{21}$$

The first basic results concerning matrices (20) and (18) are expressed by the following lemmas.

LEMMA 1. *There exists an integer k' such that*

$$\mathscr{R}[\mathscr{P}_k] \subset \mathscr{R}[\mathscr{P}_{k+1}] \quad \textit{for all } k < k', \tag{22}$$

$$\mathscr{R}[\mathscr{P}_k] = \mathscr{R}[\mathscr{P}_{k+1}] \quad \textit{for all } k \geqq k', \tag{23}$$

and, furthermore,

$$k' \leqq m. \tag{24}$$

Proof. Relations (22) and (23) can be proved by first showing that

$$\mathscr{R}[\mathscr{P}_k] \subseteq \mathscr{R}[\mathscr{P}_{k+1}] \quad \text{for all } k \tag{25}$$

and that

$$\mathscr{R}[\mathscr{P}_{k-1}] = \mathscr{R}[\mathscr{P}_k] \quad \text{for some } k \Rightarrow \mathscr{R}[\mathscr{P}_k] = \mathscr{R}[\mathscr{P}_{k+1}] \tag{26}$$

whose validity is a direct consequence of definitions (18) and (20). Lastly, (24) follows from (22), (23) and

$$\mathscr{R}[\mathscr{P}_k] \subseteq R^m \quad \text{for all } k. \tag{27}$$

LEMMA 2. *If only $\tilde{m} < m$ rows of $P_m[F, G]$ $(t_1, \cdots, t_m)$ are linearly independent over Δ_m, then there exists a constant nonsingular $m \times m$ matrix T such that*

$$TF(t)T^{-1} = \begin{pmatrix} F_{11}(t) & F_{12}(t) \\ 0 & F_{22}(t) \end{pmatrix}, \qquad TG(t) = \begin{pmatrix} G_1(t) \\ 0 \end{pmatrix}, \tag{28}$$

where the matrices $F_{11}(t)$ and $G_1(t)$ are respectively $\tilde{m} \times \tilde{m}$ and $\tilde{m} \times 1$ and $P_{\tilde{m}}[F_{11}, G_1]$ $\cdot (t_1, \cdots, t_{\tilde{m}})$ has its $\tilde{m}$ rows linearly independent over $\Delta_{\tilde{m}}$.

Proof. If only $\tilde{m}$ rows of P_m $[F, G]$ are linearly independent over Δ_m, there exists a constant nonsingular $m \times m$ matrix T such that

$$TP_m[F, G](t_1, \cdots, t_m) = \begin{pmatrix} \tilde{P}(t_1, \cdots, t_m) \\ 0 \end{pmatrix} \tag{29}$$

with the $\tilde{m}$ rows of $\tilde{P}(t_1, \cdots, t_m)$ linearly independent over Δ_m.

Bearing in mind definition (18), this proves the second equation of (28) and, after the partition

$$TF(t)T^{-1} = \begin{pmatrix} F_{11}(t) & F_{12}(t) \\ F_{21}(t) & F_{22}(t) \end{pmatrix}, \tag{30}$$

also gives

$$\tilde{P}(t_1, \cdots, t_m) = P_m[F_{11}, G_1](t_1, \cdots, t_m), \tag{31}$$

$$F_{21}(t)P_{m-1}[F_{11}, G_1](t_1, \cdots, t_{m-1}) = 0. \tag{32}$$

It follows from (31) that $\mathscr{P}_m[F_{11}, G_1]$ has rank $\tilde{m}$; therefore, on the basis of Lemma 1, $\mathscr{P}_{\tilde{m}}[F_{11}, G_1]$ also has rank $\tilde{m}$, or, what is the same, the $\tilde{m}$ rows of $P_{\tilde{m}}[F_{11}, G_1](t_1, \cdots, t_{\tilde{m}})$ are linearly independent. From this and from (32) it follows that $F_{21}(t) = 0$ and this completes the proof.

Similar results can be proved for matrices (21) and (19).

LEMMA 3. *There exists an integer k'' such that*

$$\mathscr{R}[\mathscr{Q}_k] \subset \mathscr{R}[\mathscr{Q}_{k+1}] \quad \textit{for all } k < k'', \tag{33}$$

$$\mathscr{R}[\mathscr{Q}_k] = \mathscr{R}[\mathscr{Q}_{k+1}] \quad \textit{for all } k \geqq k'', \tag{34}$$

and, furthermore,

$$k'' \leqq m. \tag{35}$$

LEMMA 4. *If only $\tilde{m} < m$ columns of $Q_m[F, H](t_1, \cdots, t_m)$ are linearly independent over Δ_m, then there exists a constant nonsingular $m \times m$ matrix T such that*

$$TF(t)T^{-1} = \begin{pmatrix} F_{11}(t) & 0 \\ F_{21}(t) & F_{22}(t) \end{pmatrix}, \qquad H(t)T^{-1} = (H_1(t) \quad 0), \tag{36}$$

where the matrices $F_{11}(t)$ and $H_1(t)$ are respectively $\tilde{m} \times \tilde{m}$ and $1 \times \tilde{m}$, and $Q_{\tilde{m}}[F_{11}, H_1](t_1, \cdots, t_{\tilde{m}})$ has its $\tilde{m}$ columns linearly independent over $\Delta_{\tilde{m}}$.

On the basis of these preliminary results, it is possible to examine the properties of the factorizable sequences of kernels and, particularly, to describe the set of all factorizations. In this analysis the integer m will be called the *dimension* of a factorization and, thus, a factorization will be *minimal* when its dimension assumes the smallest value over the set of all factorizations of a given sequence.

The properties of the minimal factorizations are expressed by the following theorem.

THEOREM 2. *An m-dimensional factorization $\{F(t), G(t), H(t)\}$ of a factorizable sequence of kernels is minimal if and only if the rows of $P_m[F, G]$ and the columns of $Q_m[F, H]$ are linearly independent over Δ_m. The minimal factorizations are a single equivalence class modulo the relation*

$$\{F_1(t), G_1(t), H_1(t)\} \sim \{F_2(t), G_2(t), H_2(t)\} \Leftrightarrow \begin{cases} F_1(t) = TF_2(t)T^{-1} \\ G_1(t) = TG_2(t) \\ H_1(t) = H_2(t)T^{-1}, \end{cases} \tag{37}$$

where T is a constant nonsingular $m \times m$ matrix.

Proof. The proof will be divided into two parts, the first concerning the minimality and the second the equivalence.

Part 1. *Necessity.* Suppose, for example, that the matrix $P_m[F, G]$, corresponding to a minimal factorization, has only $\tilde{m} < m$ linearly independent rows. As a consequence of Lemma 2, there will then exist a constant nonsingular $m \times m$

matrix T such that

$$TF(t)T^{-1} = \begin{pmatrix} F_{11}(t) & F_{12}(t) \\ 0 & F_{22}(t) \end{pmatrix}, \qquad TG(t) = \begin{pmatrix} G_1(t) \\ 0 \end{pmatrix}. \tag{38}$$

Considering also the partition

$$H(t)T^{-1} = (H_1(t) \quad H_2(t)), \tag{39}$$

it is immediately seen that the triplet $\{F_{11}(t), G_1(t), H_1(t)\}$, which has a dimension $\tilde{m} < m$, is also a factorization of the sequence $\{w_i(t_1, \cdots, t_i)\}_2^\infty$ and this contradicts the hypothesis of minimality. A similar proof, based on Lemma 4, can be given for the matrix $Q_m[F, H]$.

Part 1. *Sufficiency*. Suppose that the factorization $\{F(t), G(t), H(t)\}$, for which the matrices $P_m[F, G]$ and $Q_m[F, H]$ have, respectively, their rows and their columns linearly independent, is not a minimal one and that, consequently, there exists a factorization $\{\tilde{F}(t), \tilde{G}(t), \tilde{H}(t)\}$ of dimension $\tilde{m} < m$ for the same sequence $\{w_i(t_1, \cdots, t_i)\}_2^\infty$. Considering now the matrices $P_m[\tilde{F}, \tilde{G}]$ and $Q_m[\tilde{F}, \tilde{H}]$, it is readily seen that (see (9))

$$\begin{aligned} &Q_m[F, H](t_1, \cdots, t_m)P_m[F, G](t_{m+1}, \cdots, t_{2m}) \\ &\qquad = Q_m[\tilde{F}, \tilde{H}](t_1, \cdots, t_m)P_m[\tilde{F}, \tilde{G}](t_{m+1}, \cdots, t_{2m}). \end{aligned} \tag{40}$$

Premultiplying both sides of (40) by $Q_m^*[F, H](t_1, \cdots, t_m)$, postmultiplying by $P_m^*[F, G](t_1, \cdots, t_m)$ and then integrating, one obtains

$$\begin{aligned} \mathcal{Q}_m \cdot \mathcal{P}_m = &\int_{\Delta_m} Q_m^*[F, H]Q_m[\tilde{F}, \tilde{H}]\, dt_1 \cdots dt_m \\ &\cdot \int_{\Delta_m} P_m[\tilde{F}, \tilde{G}]P_m^*[F, G]\, dt_{m+1} \cdots dt_{2m}. \end{aligned} \tag{41}$$

The rank of the matrix on the L.H.S. is equal to m by hypothesis, while the rank of the matrix on the R.H.S. does not exceed $\tilde{m}$; it follows that

$$m \leqq \tilde{m}. \tag{42}$$

This contradicts the hypothesis that $\tilde{m} < m$ and completes the proof of the first part.

Part 2. If the triplet $\{F(t), G(t), H(t)\}$ is a minimal factorization of the given sequence of kernels, it can be verified immediately that the triplet $\{TF(t)T^{-1}, TG(t), H(t)T^{-1}\}$ is also a minimal factorization (see (9)). Vice versa, let $\{F_1(t), G_1(t), H_1(t)\}$ and $\{F_2(t), G_2(t), H_2(t)\}$ be any two minimal factorizations of the given sequence of kernels. Considering the matrices $Q_m[F_1, H_1]$, $P_m[F_1, G_1]$, $Q_m[F_2, H_2]$ and $P_m[F_2, G_2]$, one finds that

$$\begin{aligned} &Q_m[F_1, H_1](t_1, \cdots, t_m) \cdot P_m[F_1, G_1](t_{m+1}, \cdots, t_{2m}) \\ &\qquad = Q_m[F_2, H_2](t_1, \cdots, t_m) \cdot P_m[F_2, G_2](t_{m+1}, \cdots, t_{2m}). \end{aligned} \tag{43}$$

By virtue of the results proved in the first part, the m columns of $Q_m[F_1, H_1]$ and $Q_m[F_2, H_2]$ and the m rows of $P_m[F_1, G_1]$ and $P_m[F_2, G_2]$ are linearly independent over Δ_m. Consequently, generalizing the methods and results established by D. C. Youla in [9] with regard to functions defined in R^1, one finds that

$$\begin{aligned} Q_m[F_1, H_1] &= Q_m[F_2, H_2]T^{-1}, \\ P_m[F_1, G_1] &= TP_m[F_2, G_2], \end{aligned} \tag{44}$$

where T is a constant nonsingular matrix. Bearing in mind the expressions (18) and (19), it follows that

$$H_1(t) = H_2(t)T^{-1}, \tag{45}$$

$$G_1(t) = TG_2(t). \tag{46}$$

As concerns the relation between $F_1(t)$ and $F_2(t)$, one can first of all write

$$\begin{aligned} &Q_m[F_1, H_1](t_1, \cdots, t_m) \cdot F_1(t_{m+1}) \cdot P_m[F_1, G_1](t_{m+2}, \cdots, t_{2m+1}) \\ &\quad = Q_m[F_2, H_2](t_1, \cdots, t_m) \cdot F_2(t_{m+1}) \cdot P_m[F_2, G_2](t_{m+2}, \cdots, t_{2m+1}). \end{aligned} \tag{47}$$

Substituting the expressions (44) in the L.H.S. of (47), and bearing in mind the linear independence of the columns of $Q_m[F_2, H_2]$ and of the rows of $P_m[F_2, G_2]$, one finally obtains

$$F_1(t) = TF_2(t)T^{-1}, \tag{48}$$

thus concluding the proof of the second part.

In addition to the result expressed by the preceding theorem, it is also convenient to find a relation between any factorization and a minimal one. This is expressed by the following theorem.

THEOREM 3. *Any factorization of a factorizable sequence of kernels can be written in the form*

$$\begin{aligned} F(t) &= T\begin{pmatrix} F_0(t) & 0 & F_{13}(t) & 0 \\ F_{21}(t) & F_{22}(t) & F_{23}(t) & F_{24}(t) \\ 0 & 0 & F_{33}(t) & 0 \\ 0 & 0 & F_{43}(t) & F_{44}(t) \end{pmatrix}T^{-1}, \\ G(t) &= T\begin{pmatrix} G_0(t) \\ G_2(t) \\ 0 \\ 0 \end{pmatrix}, \qquad H(t) = (H_0(t) \quad 0 \quad H_3(t) \quad 0)T^{-1}, \end{aligned} \tag{49}$$

where T is a constant nonsingular matrix, and the triplet $\{F_0(t), G_0(t), H_0(t)\}$ is a minimal factorization of the same sequence.

Proof. It is a direct consequence of Lemmas 2 and 4 and of Theorem 2.

We observe that this result also provides a rule for computing m_0.

COROLLARY 1. *The dimension m_0 of a minimal factorization of a factorizable sequence of kernels is given by*

$$m_0 = \text{rank}\,\{\mathcal{Q}_m[F, H]\mathcal{P}_m[F, G]\}, \tag{50}$$

where $\{F(t), G(t), H(t)\}$ is any given factorization of the same sequence.

Before concluding this section it is important to stress that all the results proved here are valid even if the hypothesis that $F(t)$, $G(t)$ and $H(t)$ are functions with proper rational Laplace transform is relaxed. In fact, the only requirement was the possibility of testing the linear independence of the rows of $P_k[F, G]$ (columns of $Q_k\,[F, H]$) only on the interval Δ_k; for this, for example, it is sufficient to require analyticity on $F(t)$, $G(t)$ and $H(t)$.

5. Minimal realizations. We are now in a position to prove a basic result concerning minimal bilinear realizations. Consistently with the case of linear systems, we shall assume the following.

DEFINITION 2. A bilinear realization $\{A, N, B, C\}$ is *minimal* if the dimension of the state space assumes the minimum value over the set of all realizations of a given sequence of kernels.

Denoting by $\delta\{L(t)\}$ the order of the weighting pattern $L(t)$, we then have the following theorem.

THEOREM 4. *Let $\{w_i(t_1, \cdots, t_i)\}_1^\infty$ be a sequence of bilinearly realizable kernels, and let $\{F_0(t), G_0(t), H_0(t)\}$ be a minimal factorization of the subsequence $\{w_i(t_1, \cdots, t_i)\}_2^\infty$. The minimal bilinear realizations are such that*

(a) *the state space dimension is given by*

$$n_0 = \delta\left\{\begin{pmatrix} w_1(t) & H_0(t) \\ G_0(t) & F_0(t) \end{pmatrix}\right\}; \tag{51}$$

(b) *the matrix N has minimum rank, given by*

$$r_0 = \dim\{F_0(t), G_0(t), H_0(t)\} \triangleq m_0; \tag{52}$$

(c) *they are a single equivalence class modulo the relation*

$$\{A_1, N_1, B_1, C_1\} \sim \{A_2, N_2, B_2, C_2\} \Leftrightarrow \begin{cases} A_1 = TA_2T^{-1} \\ N_1 = TN_2T^{-1} \\ B_1 = TB_2 \\ C_1 = C_2T^{-1}, \end{cases} \tag{53}$$

where T is a constant nonsingular matrix.

Proof. Let $\{F(t), G(t), H(t)\}$ denote any factorization of $\{w_i(t_1, \cdots, t_i)\}_2^\infty$ and consider the matrix

$$L(t) = \begin{pmatrix} w_1(t) & H(t) \\ G(t) & F(t) \end{pmatrix}. \tag{54}$$

From the proof of sufficiency in Theorem 1, it follows that every linear realization of the matrix $L(t)$ identifies a bilinear realization of $\{w_i(t_1, \cdots, t_i)\}_1^\infty$; conversely, from any bilinear realization it is possible to construct a factorization and, consequently, a matrix such as (54). One can therefore conclude that all bilinear realizations can be obtained by taking all linear realizations of all matrices $L(t)$ associated with the given sequence, and then applying (14), (15) and (17). It follows that the dimension of the minimal bilinear realizations is equal to

$$n_0 = \min_{L(t)} \delta\{L(t)\}. \tag{55}$$

If now the forms (49) are substituted in (54) and the matrix $\begin{bmatrix} 1 & 0 \\ 0 & T \end{bmatrix}$ is factored on the left, and its inverse on the right, it is seen that

$$(56) \qquad \delta\{L(t)\} \geqq \delta\{L_0(t)\} \quad \text{for all } L(t),$$

where $L_0(t)$ denotes the matrix

$$(57) \qquad L_0(t) = \begin{pmatrix} w_1(t) & H_0(t) \\ G_0(t) & F_0(t) \end{pmatrix}.$$

This, since $L_0(t)$ is an element of the set of matrices $L(t)$, proves statement (a).

In order to prove statement (b), let $\{A, N, B, C\}$ denote a minimal bilinear realization and let

$$(58) \qquad N'N'' = N$$

be a factorization of N such that N' and N'' have full rank $r = \operatorname{rank}(N)$. Clearly, the triplet

$$(59) \qquad F(t) = N''e^{At}N', \quad G(t) = N''e^{At}B, \quad H(t) = Ce^{At}N'$$

is an r-dimensional factorization. From this it follows that $r \geqq m_0$ because, by assumption, the dimension of any factorization cannot be lower than m_0. We shall also prove that the inequality $r > m_0$ yields a contradiction. In this case, in fact, the triplet (59) would be a nonminimal factorization and therefore, according to the results achieved in the preceding section, reducible. Assume, for instance, that (59) is not minimal because only $r_1 < r$ rows of $P_r[N''e^{At}N', N''e^{At}B]$ are linearly independent; then it is possible to find a nonsingular $r \times r$ matrix T that reduces (59) to the r_1-dimensional triplet (see (38) and (39) in the proof of Theorem 2)

$$(60) \qquad F_{11}(t) = \tilde{N}_1''e^{At}\tilde{N}_1', \quad G_1(t) = \tilde{N}_1''e^{At}B, \quad H(t) = Ce^{At}\tilde{N}_1',$$

where $\tilde{N}_1''$ and $\tilde{N}_1'$ (respectively $r_1 \times n$ and $n \times r_1$) are defined by

$$(61) \qquad TN'' = \begin{pmatrix} \tilde{N}_1'' \\ \tilde{N}_2'' \end{pmatrix}, \qquad N'T^{-1} = (\tilde{N}_1' \quad \tilde{N}_2').$$

Furthermore, this matrix T is such that

$$(62) \qquad \tilde{N}_2''e^{At}B = 0, \qquad \tilde{N}_2''e^{At}\tilde{N}_1' = 0.$$

The triplet (60) is still a factorization and, therefore, the quadruplet $\{A, \tilde{N}_1'\tilde{N}_1'', B, C\}$ is still a bilinear realization, minimal because its dimension is unchanged. From this it follows that the triplet $\left\{A, (B \quad \tilde{N}_1'), \begin{pmatrix} C \\ \tilde{N}_1'' \end{pmatrix}\right\}$ is a minimal linear realization of

$$(63) \qquad L_1(t) = \begin{pmatrix} w_1(t) & H_1(t) \\ G_1(t) & F_{11}(t) \end{pmatrix}$$

(since, otherwise, the quadruplet $\{A, \tilde{N}_1'\tilde{N}_1'', B, C\}$ would not be a minimal bilinear realization). As a consequence, the n rows of the matrix $e^{At}(B \quad \tilde{N}_1')$ are linearly independent and this, applied to (62), implies $\tilde{N}_2'' = 0$. But this is in contradiction with the assumption that N'' has full rank $r > m_0$ and thus we conclude that $r = m_0$.

As concerns statement (c), it is immediately verified that two quadruplets of matrices equivalent according to (53) yield the same sequence $\{w_i(t_1, \cdots, t_i)\}_1^\infty$. Consider now any two minimal bilinear realizations of the same sequence $\{A_1, N_1, B_1, C_1\}$ and $\{A_2, N_2, B_2, C_2\}$. It has been proved (statement (b)) that

$$\text{rank}\,\{N_1\} = \text{rank}\,\{N_2\} = m_0 \tag{64}$$

and it is therefore always possible to construct from these realizations, according to (12), two m_0-dimensional (i.e., minimal) factorizations $\{F_1(t), G_1(t), H_1(t)\}$ and $\{F_2(t), G_2(t), H_2(t)\}$. These factorizations are equivalent by Theorem 2; therefore, constructing for both factorizations the matrices (54), we have

$$\begin{pmatrix} w_1(t) & H_1(t) \\ G_1(t) & F_1(t) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & M \end{pmatrix} \begin{pmatrix} w_1(t) & H_2(t) \\ G_2(t) & F_2(t) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & M^{-1} \end{pmatrix}, \tag{65}$$

where M is a nonsingular $m_0 \times m_0$ matrix.

On the other hand, bearing in mind (59), the above expression (65) becomes

$$\begin{pmatrix} C_1 \\ N_1'' \end{pmatrix} e^{A_1 t}(B_1 \quad N_1') = \begin{pmatrix} 1 & 0 \\ 0 & M \end{pmatrix} \begin{pmatrix} C_2 \\ N_2'' \end{pmatrix} e^{A_2 t}(B_2 \quad N_2') \begin{pmatrix} 1 & 0 \\ 0 & M^{-1} \end{pmatrix}. \tag{66}$$

In (66) there appear two minimal linear realizations of the same matrix. These must therefore be related by

$$\begin{aligned} \begin{pmatrix} C_1 \\ N_1'' \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ 0 & M \end{pmatrix} \begin{pmatrix} C_2 \\ N_2'' \end{pmatrix} T^{-1}, \\ (B_1 \quad N_1') &= T(B_2 \quad N_2') \begin{pmatrix} 1 & 0 \\ 0 & M^{-1} \end{pmatrix}, \\ A_1 &= TA_2T^{-1}, \end{aligned} \tag{67}$$

where T is a nonsingular $n \times n$ matrix. It follows from this that the two arbitrarily chosen minimal realizations are equivalent modulo the relation (53) and this completes the proof of the theorem.

The relevance of the result proved in part (b) of this theorem will become clear from the inspection of the simulation diagram of Fig. 1. Here the rank of N assumes the meaning of the least number of multipliers needed for the simulation by means of this diagram (note also that there does not exist any other diagram providing a simulation with a lower number of multipliers). This consideration suggests the interest in characterizing the simulation of a bilinear realization through the number of integrators and of two input multipliers; from this point of view it is nice to verify that, rather unexpectedly, the minimization of the first one implies that of the second.

6. Structure analysis. The preceding sections have been concerned with the problem of realizing the input-output description of bilinear systems. In analogy

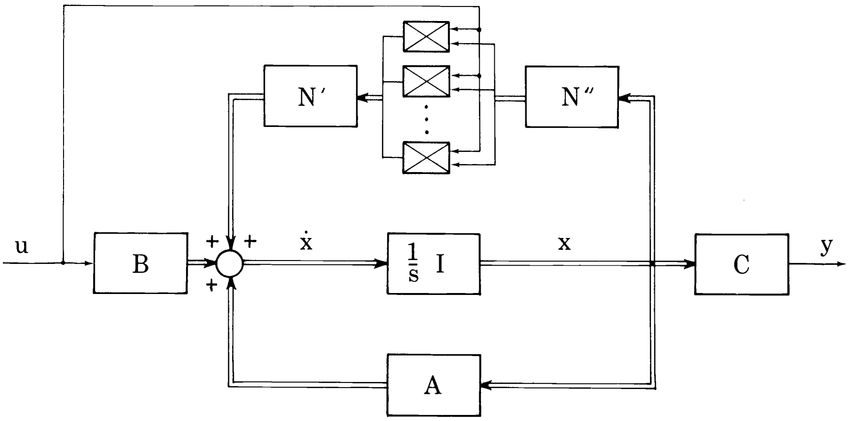

FIG. 1

with the theory of linear systems, and for the same reasons, it is now interesting to develop a structure analysis from the state space description. As a result, we shall show that it is possible to effect a decomposition of the state space and to prove the existence of a canonical form of the equations, thereby laying the foundations for a complete analysis of the correlation existing between the minimal realizations and one of the subsystems resulting from the decomposition of the state space.

This analysis will be based on suitable properties of input-state and, respectively, state-output interaction. As concerns the former it is convenient to assume the following.

DEFINITION 3. A state x of system (1) is *reachable from the origin* if there exists an admissible input function that maps the origin of the state space into the state x in a finite interval of time.

Due to the nonlinearity of system (1), the subset of all states reachable from the origin is not a linear subspace. However, a state-space decomposition is still possible if this subset is embedded into a suitable (i.e., the least one) linear subspace. For the purpose of evaluating this subspace, we introduce the sequence of matrices

$$\bar{P}_1 = B, \tag{68a}$$

$$\bar{P}_i = (A\bar{P}_{i-1} \quad N\bar{P}_{i-1}), \qquad i = 2, 3, \cdots, \tag{68b}$$

and we then have the following theorem.

THEOREM 5. *The subset of all states of system* (1) *reachable from the origin spans a subspace* $\mathscr{X}_p$ *of* R^n *which*:

(a) *is the least subspace invariant under* A *and* N *and containing* $\mathscr{R}[B]$;

(b) *can be expressed as*

$$\mathscr{X}_p = \mathscr{R}[\bar{P}_1 \quad \bar{P}_2 \cdots \bar{P}_n]. \tag{69}$$

Proof. The proof of statement (a) is an extension of that used in the case of linear systems. It is readily observed that, for any given subspace $\mathscr{X}$ of R^n,

$$x(t) \in \mathscr{X} \quad \text{for all } t \in [0, T] \Rightarrow \dot{x}(t) \in \mathscr{X} \quad \text{for all } t \in [0, T]. \tag{70}$$

By hypothesis, at least a basis $\{x_1, \cdots, x_r\}$ of $\mathscr{X}_p$ is reachable from the origin. Therefore, for (70),

$$(A + Nux_j + Bu) \in \mathscr{X}_p \quad \text{for all } u \in R \quad \text{and for all } j. \tag{71}$$

But, again for (70),

$$Bu \in \mathscr{X}_p \quad \text{for all } u \in R \tag{72}$$

and, hence, from this and (71),

$$(A + Nu)\mathscr{X}_p \subseteq \mathscr{X}_p \quad \text{for all } u \in R. \tag{73}$$

From (72) and (73) it follows that $\mathscr{X}_p$ contains $\mathscr{R}[B]$ and is invariant under A and N. Moreover, observing that any trajectory starting from the origin belongs to a subspace invariant under A and N and containing $\mathscr{R}[B]$, it is easy to prove, by contradiction, that $\mathscr{X}_p$ is the least subspace with these properties.

As concerns statement (b), we easily verify that

$$\mathscr{R}[\bar{P}_1 \cdots \bar{P}_{i-1}] \subseteq \mathscr{R}[\bar{P}_1 \cdots \bar{P}_i] \subseteq R^n \tag{74}$$

and that

$$\begin{aligned} &\mathscr{R}[\bar{P}_1 \cdots \bar{P}_{k-1}] = \mathscr{R}[\bar{P}_1 \cdots \bar{P}_k] \quad \text{for some } k \\ &\qquad \Rightarrow \mathscr{R}[\bar{P}_1 \cdots \bar{P}_k] = \mathscr{R}[\bar{P}_1 \cdots \bar{P}_{k+1}], \end{aligned} \tag{75}$$

from which it follows that $\mathscr{R}[\bar{P}_1 \cdots \bar{P}_n]$ is invariant under A and N and contains $\mathscr{R}[B]$. To prove that this is the least subspace with these properties, observe that any subspace $\mathscr{X}$ of this type must satisfy

$$\mathscr{X} \supseteq \mathscr{R}[B] = \mathscr{R}[\bar{P}_1], \tag{76a}$$

$$\mathscr{X} \supseteq A\mathscr{X} \supseteq A\mathscr{R}[\bar{P}_1], \quad \mathscr{X} \supseteq N\mathscr{X} \supseteq N\mathscr{R}[\bar{P}_1] \tag{76b}$$

and, then,

$$\mathscr{X} \supseteq \mathscr{R}[\bar{P}_1 \quad \bar{P}_2]. \tag{76c}$$

By iterating this construction it follows that any subspace invariant under A and N and containing $\mathscr{R}[B]$ contains, in turn, $\mathscr{R}[\bar{P}_1 \cdots \bar{P}_n]$ which, consequently, is the least of them. This completes the proof.

The state-output interaction property considered in the present analysis is expressed by the following definition.

Definition 4. A state x of systems (1) is *unobservable* if the component of the response depending on the initial state is identically zero for every admissible input function.

From (2), (3) and (7) it immediately follows that the set of all unobservable states is a subspace. Introducing the matrix sequences

$$\bar{Q}_1 = C, \tag{77a}$$

$$\bar{Q}_i = \begin{pmatrix} \bar{Q}_{i-1}A \\ \bar{Q}_{i-1}N \end{pmatrix}, \qquad i = 2, 3, \cdots, \tag{77b}$$

we have the following theorem.

THEOREM 6. *The subset of all the unobservable states of system* (1) *is a subspace* $\mathscr{X}_q$ *of* R^n *which*

(a) *is the largest subspace invariant under* A *and* N *and contained in* $\mathscr{N}[G]$;

(b) *can be expressed as*

$$\mathscr{X}_q = \mathscr{N}\begin{Bmatrix}\bar{Q}_1\\ \bar{Q}_2\\ \vdots\\ \bar{Q}_n\end{Bmatrix}. \tag{78}$$

Proof. The proof can be developed along the same lines as that of Theorem 5.

Referring to the properties of reachability from the origin and unobservability, or, more exactly, to the subspaces $\mathscr{X}_p$ and $\mathscr{X}_q$, it is possible to effect the decomposition of the state space $\mathscr{X}$ of the system (1) into the direct sum of four subspaces $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ and $\mathscr{D}$, following the procedure adopted in the case of linear systems, i.e.,

$$\begin{aligned}\mathscr{A} &= \mathscr{X}_p \cap \mathscr{X}_q,\\ \mathscr{X}_p &= \mathscr{A} \oplus \mathscr{B},\\ \mathscr{X}_q &= \mathscr{A} \oplus \mathscr{C},\\ \mathscr{X} &= \mathscr{A} \oplus \mathscr{B} \oplus \mathscr{C} \oplus \mathscr{D}.\end{aligned} \tag{79}$$

On the basis of a decomposition of this type, which will be called a canonical decomposition, it is possible to prove the following theorem.

THEOREM 7. *Assuming as basis in the state space the union of bases of the four subspaces* $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ *and* $\mathscr{D}$ *of a canonical decomposition, equations* (1) *assume the form*

$$\begin{aligned}\begin{pmatrix}\dot{x}_a(t)\\ \dot{x}_b(t)\\ \dot{x}_c(t)\\ \dot{x}_d(t)\end{pmatrix} &= \begin{pmatrix}A_{aa} & A_{ab} & A_{ac} & A_{ad}\\ 0 & A_{bb} & 0 & A_{bd}\\ 0 & 0 & A_{cc} & A_{cd}\\ 0 & 0 & 0 & A_{dd}\end{pmatrix}\begin{pmatrix}x_a(t)\\ x_b(t)\\ x_c(t)\\ x_d(t)\end{pmatrix}\\ &\quad + \begin{pmatrix}N_{aa} & N_{ab} & N_{ac} & N_{ad}\\ 0 & N_{bb} & 0 & N_{bd}\\ 0 & 0 & N_{cc} & N_{cd}\\ 0 & 0 & 0 & N_{dd}\end{pmatrix}\begin{pmatrix}x_a(t)\\ x_b(t)\\ x_c(t)\\ x_d(t)\end{pmatrix} u(t) + \begin{pmatrix}B_a\\ B_b\\ 0\\ 0\end{pmatrix} u(t),\\ y(t) &= (0 \quad C_b \quad 0 \quad C_d)\begin{pmatrix}x_a(t)\\ x_b(t)\\ x_c(t)\\ x_d(t)\end{pmatrix},\end{aligned} \tag{80}$$

where $(x_a^T \; 0 \; 0 \; 0)^T$, $(0 \; x_b^T \; 0 \; 0)^T$, $(0 \; 0 \; x_c^T \; 0)^T$ *and* $(0 \; 0 \; 0 \; x_d^T)^T$ *are coordinates of vectors belonging respectively to the subspaces* $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$ *and* $\mathscr{D}$ *(canonical form).*

Moreover, within the equivalence relation induced by a constant coordinate transformation in the state space, the quadruplets $\{A, N, B, C\}$ *which assume the form associated with* (80) *are an equivalence class modulo the transformation defined by a matrix of the type*

$$T = \begin{pmatrix} T_{aa} & T_{ab} & T_{ac} & T_{ad} \\ 0 & T_{bb} & 0 & T_{bd} \\ 0 & 0 & T_{cc} & T_{cd} \\ 0 & 0 & 0 & T_{dd} \end{pmatrix}, \tag{81}$$

where the partitions are consistent with those of the state vector.

Proof. The proof of (80) is a standard consequence of statements (a) in Theorems 5 and 6. The proof of the equivalence between canonical forms can be deduced directly from that given by the authors in [3] for a similar result in the case of linear systems.

A further direct analogy with well-known results in the case of linear systems can be obtained by calculating the kernels $w_i(t_1, \cdots, t_i)$ of the Volterra series expansion for the zero-state response. In fact, one finds that these kernels depend only on the matrices A_{bb}, N_{bb}, B_b, C_b associated with the "part b" resulting from the canonical decomposition.

Furthermore, always on the basis of the analogy with the case of linear systems, it becomes natural to establish a connection between the results of the structure analysis and the realization theory, characterizing the minimality by means of the properties that identify the "part b".

To this end it is convenient to show that the subspaces $\mathscr{X}_p$ and $\mathscr{X}_q$, considered in the structure analysis, can also be characterized in terms of the matrices (20) and (21) on which the realization theory was based. In fact, we have

$$\mathscr{X}_p = \mathscr{R}\{\mathscr{P}_n[e^{At}N, e^{At}B]\}, \tag{82}$$

$$\mathscr{X}_q = \mathscr{N}\{\mathscr{Q}_n[Ne^{At}, Ce^{At}]\} \tag{83}$$

that are alternative to the ones provided by (69) and (78). To prove this result, observe first that, from definition (20), we have

$$\mathscr{P}_k[e^{At}N, e^{At}B] = \mathscr{P}_1[e^{At}N, e^{At}B] + \int_0^1 e^{At}N\mathscr{P}_{k-1}[e^{At}N, e^{At}B]N^*e^{A^*t}\,dt. \tag{84}$$

Again, by definition,

$$\mathscr{R}\{\mathscr{P}_1[e^{At}N, e^{At}B]\} = \mathscr{R}[B \quad AB \cdots A^{n-1}B] \tag{85}$$

from which (84) supplies

$$\begin{aligned} \mathscr{R}\{\mathscr{P}_2[e^{At}N, e^{At}B]\} &= \mathscr{R}[B \quad AB \cdots A^{n-1}B] \\ &+ \mathscr{R}[N(B \quad AB \cdots A^{n-1}B) \cdots A^{n-1}N(B \quad AB \cdots A^{n-1}B)]. \end{aligned} \tag{86}$$

By iterating this construction, and taking into account (68) and (69), it is possible to arrive at

$$\mathscr{R}\{\mathscr{P}_n[e^{At}N, e^{At}B]\} = \mathscr{R}[\bar{P}_1 \quad \bar{P}_2 \cdots \bar{P}_n] = \mathscr{X}_p. \tag{87}$$

Equation (83) is proved in a similar way.

We are now in a position to prove the following theorem.

THEOREM 8. *A realization $\{A, N, B, C\}$ of a bilinearly realizable sequence of kernels is minimal if and only if its state space is both observable and spanned by the states reachable from the origin.*

Proof. Clearly, minimality implies the structure properties of the statement; otherwise, it would be possible to find a realization with a lower dimension (i.e., the one defined by the quadruplet $\{A_{bb}, N_{bb}, B_b, C_b\}$ of the canonical form (80)). The converse may be proved in the following way. Let $\{A, N, B, C\}$ and $\{\bar{A}, \bar{N}, \bar{B}, \bar{C}\}$ be any two realizations with their own state space reachable from the origin and observable; let $n, \bar{n}$ denote their dimensions. On the basis of definitions (18) and (19) and from factorization (9), denoting max $(n, \bar{n})$ with $\tilde{n}$, it can be seen that

$$\begin{aligned} &Q_{\tilde{n}}[Ne^{At}, Ce^{At}](t_1, \cdots, t_{\tilde{n}})P_{\tilde{n}}[e^{At}N, e^{At}B](t_{\tilde{n}+1}, \cdots, t_{2\tilde{n}}) \\ &\quad = Q_{\tilde{n}}[\bar{N}_{e^{\bar{A}t}}, \bar{C}e^{\bar{A}t}](t_1, \cdots, t_{\tilde{n}})P_{\tilde{n}}[e^{\bar{A}t}\bar{N}, e^{\bar{A}t}\bar{B}](t_{\tilde{n}+1}, \cdots, t_{2\tilde{n}}). \end{aligned} \tag{88}$$

The hypothesis that the state space of the quadruplet $\{A, N, B, C\}$ is spanned by the states reachable from the origin and observable implies, thanks to (82) and (83), that the n columns of $Q_{\tilde{n}}$ and the n rows of $P_{\tilde{n}}$ in the L.H.S. of (88) are linearly independent. The same can be said for the $\bar{n}$ columns of $Q_{\tilde{n}}$ and the $\bar{n}$ rows of $P_{\tilde{n}}$ in the R.H.S. of (88). From this we conclude that $n = \bar{n}$. This, together with the result proved in the first part of the proof, implies that any realization with the aforesaid structure properties is minimal.

Remark 2. Since, by construction, the matrix quadruplet $\{A_{bb}, N_{bb}, B_b, C_b\}$ identified by the canonical decomposition satisfies the conditions of Theorem 8, it can be considered as a minimal realization of the sequence of kernels associated with the Volterra series expansion of the zero-state response of system (1).

We have also the following corollary.

COROLLARY 2. *The dimension n_0 of a minimal realization of a realizable sequence of kernels is given by*

$$n_0 = \operatorname{rank}\{\mathscr{Q}_n[Ne^{At}, Ce^{At}]\mathscr{P}_n[e^{At}N, e^{At}B]\}, \tag{89}$$

where $\{A, N, B, C\}$ is any given realization of the same sequence.

7. Outline of realization methods. In the proof of the condition of bilinear realizability for a sequence of kernels (Theorem 1), in the analysis of the properties of the factorizations of such sequences (Theorem 2), and also in the study of the connection between the structure analysis and the realization theory (Theorem 8), there have emerged various procedures that can be considered as steps on the way to the construction of minimal realizations, when a triplet $\{F(t), G(t), H(t)\}$ is available. In order to get an overall view of the various possibilities, it might perhaps be useful to list the partial computational procedures which, as mentioned, emerge from the previous analysis:

(a) Reduction of a factorization, i.e., construction of a minimal factorization $\{F_0(t), G_0(t), H_0(t)\}$ from any given factorization $\{F(t), G(t), H(t)\}$ (see proof of Theorem 3).

(b) Construction of a linear realization from the matrix

$$L(t) = \begin{pmatrix} w_1(t) & H(t) \\ G(t) & F(t) \end{pmatrix}$$

using one of the many existing procedures (see proof of Theorem 1).

(c) Reduction of a bilinear realization, i.e., construction of the quadruplet $\{A_{bb}, N_{bb}, B_b, C_b\}$ from any realization $\{A, N, B, C\}$ (see Theorem 7 and Remark 2).

Combining these procedures, various methods for the construction of minimal bilinear realizations can be developed. For example, one could apply procedure (b), without seeking a minimal realization, and subsequently procedure (c). Alternatively, one could successively apply procedures (a) and (b), being careful to ensure that the latter leads to a minimal realization.

It seems worthwhile to stress the considerable analogies that exist between these procedures and the ones that are used in the realization theory of time-invariant linear systems.

The outlined methods presuppose the knowledge of a triplet $\{F(t), G(t), H(t)\}$ satisfying the factorizability condition (9) for all integers i. An interesting problem, at this point, would be to determine whether the triplet may be computed from a finite sequence of kernels, or not. This problem is not considered in full in the present paper; we only observe that its actual significance relies upon the result expressed by the following theorem.

THEOREM 9. *A factorizable sequence of kernels* $\{w_i(t_1, \cdots, t_i\}_2^\infty$ *is uniquely specified by the sequence* $\{w_i(t_1, \cdots, t_i)\}_2^{2m_0+1}$, *where* m_0 *is the dimension of its minimal factorizations.*

Proof. The proof will be constructive; in other words, it will consist of a procedure for determining uniquely the kernel w_{2m_0+2} from $\{w_i(t_1, \cdots, t_i)\}_2^{2m_0+1}$.

Consider a minimal factorization $\{F_0(t), G_0(t), H_0(t)\}$ of the given sequence and put

$$(90) \qquad \mathscr{S}_{k,h}(t_1, \cdots, t_{k+h}) = Q_k[F_0, H_0](t_1, \cdots, t_k)P_h[F_0, G_0](t_{k+1}, \cdots, t_{k+h}),$$

where (see (9)) the element in position (i, j) is equal to the kernel of order $i + j$ calculated for suitable arguments.

Proceed now by constructing:

(a) A factorization

$$(91) \qquad \mathscr{S}_{m_0,m_0}(t_1, \cdots, t_{2m_0}) = R(t_1, \cdots, t_{m_0})S(t_{m_0+1}, \cdots, t_{2m_0})$$

such that the columns of $R(t_1, \cdots, t_{m_0})$ and the rows of $S(t_1, \cdots, t_{m_0})$ are linearly independent. By comparing this with (90), written for $k = h = m_0$, and recalling Theorem 2, it follows (see [9]) that

$$(92) \qquad \begin{aligned} R &= Q_{m_0}[F_0, H_0]M, \\ S &= M^{-1}P_{m_0}[F_0, G_0], \end{aligned}$$

where M is a constant nonsingular matrix.

(b) The nonsingular matrices

$$U = \int_{\Delta_{m_0}} R^*R \, dt_1 \cdots dt_{m_0}, \tag{93}$$

$$V = \int_{\Delta_{m_0}} SS^* \, dt_1 \cdots dt_{m_0}. \tag{94}$$

(c) The matrices

$$\begin{aligned} \tilde{R}(t_1, \cdots, t_{m_0+1}) = \int_{\Delta_{m_0}} & \mathscr{S}_{m_0+1,m_0}(t_1, \cdots, t_{2m_0+1}) \\ & \cdot S^*(t_{m_0+2}, \cdots, t_{2m_0+1}) \, dt_{m_0+2} \cdots dt_{2m_0+1} V^{-1}, \end{aligned} \tag{95}$$

$$\begin{aligned} \tilde{S}(t_{m_0+2}, \cdots, t_{2m_0+2}) = U^{-1} \int_{\Delta_{m_0}} & R^*(t_2, \cdots, t_{m_0+1}) \\ & \cdot \mathscr{S}_{m_0,m_0+1}(t_2, \cdots, t_{2m_0+2}) \, dt_2 \cdots dt_{m_0+1}. \end{aligned} \tag{96}$$

By direct substitution it is possible to verify that

$$\tilde{R}(t_1, \cdots, t_{m_0+1})\tilde{S}(t_{m_0+2}, \cdots, t_{2m_0+2}) = \mathscr{S}_{m_0+1,m_0+1}(t_1, \cdots, t_{2m_0+2}). \tag{97}$$

The procedure outlined above made use of the kernels $w_2, \cdots, w_{2m_0+1}$ which form part of the matrices $\mathscr{S}_{m_0,m_0}, \mathscr{S}_{m_0,m_0+1}, \mathscr{S}_{m_0+1,m_0}$. As a result one obtains the matrix $\mathscr{S}_{m_0+1,m_0+1}$ and this, as shown by (90), written for $k = h = m_0 + 1$, contains the additional element w_{2m_0+2}; this additional element is the very kernel it was desired to construct. Since this procedure can be iterated, the proof is complete.

The validity of this theorem does not rely (see the remarks at the end of § 3) upon the assumption that $F(t)$, $G(t)$ and $H(t)$ are functions with proper rational Laplace transforms. If this is the case, it is possible to reduce further the amount of information needed to specify the sequence $\{w_i(t_1, \cdots, t_i)\}_2^\infty$ (see, for example, [4]).

8. Conclusions. We end this paper with some concluding remarks. First of all we would emphasize the strict analogy between the theory developed here and that of the linear systems, whose well-known results can easily be obtained merely by putting $N = 0$. Another advantage of this theory is the systematic use of linear algebra tools.

Results similar to the ones presented here in § 6 are given in [5] for discrete-time bilinear systems. A complete analysis including both discrete-time and continuous-time systems is also presented in [4].

The present theory can be extended to cover the cases in which the initial state is an arbitrary equilibrium state [7]. This makes it possible to handle bilinear systems homogeneous in the state (i.e., $B = 0$ in equation (1)).

Finally, for the sake of completeness, we also note that another approach is possible to the problem of constructing minimal bilinear realizations from a given nonlinear input/output map. This approach reduces the realization problem to that of matching an infinite sequence of input/output parameters; on this basis it is possible to develop, for bilinear systems, a realization theory [6] analogous to the one originated after a well-known paper by B. L. Ho and R. E. Kalman.

REFERENCES

[1] R. W. Brockett, *On the algebraic structure of bilinear systems*, Theory and Applications of Variable Structure Systems, R. R. Mohler and A. Ruberti, eds., Academic Press, New York, 1972, pp. 153–168.

[2] C. Bruni, G. Di Pillo and G. Koch, *On the mathematical models of bilinear systems*, Ricerche di Automatica, 2 (1971), pp. 11–26.

[3] P. d'Alessandro, A. Isidori and A. Ruberti, *A new approach to the theory of canonical decomposition of linear dynamical systems*, this Journal, 11 (1973), pp. 148–158.

[4] ———, *Lectures on Bilinear System Theory*, Notes for a Course held at C.I.S.M., Udine, Italy, Springer-Verlag, Wien, 1972.

[5] P. d'Alessandro, *Structure properties, invariance and insensitivity of bilinear discrete-time systems*, Ricerche di Automatica, 3 (1972), pp. 158–169.

[6] A. Isidori, *Direct construction of minimal bilinear realizations from nonlinear input/output maps*, IEEE Trans. Automatic Control. AC-18 (1973), to appear.

[7] A. Isidori and A. Ruberti, *Realization theory of bilinear systems*, Geometric Methods in System Theory, D. Q. Mayne and R. W. Brockett, eds., D. Reidel, Dordrecht, 1973.

[8] R. R. Mohler, *Natural bilinear control processes*, IEEE Trans. SSC, SSC-6 (1970), pp. 192–197.

[9] D. C. Youla, *The synthesis of linear dynamical systems from prescribed weighting patterns*, SIAM J. Appl. Math., 14 (1966), pp. 527–549.

SIAM J. Control
Vol. 12, No. 3, August 1974

CAUSALITY, STRICT CAUSALITY AND INVERTIBILITY FOR SYSTEMS IN HILBERT RESOLUTION SPACES*

ROMANO M. De SANTIS†

Abstract. The Hilbert resolution space setting is used to obtain some new sufficient conditions for the existence and causality of an inverse system. These conditions are applicable when the system under consideration is characterized by a behavior which is special, in some sense, with respect to the notion of causality. Relevant conceptual connections with recent problems in system sensitivity and stability theory are pointed out.

1. Introduction. Given an operator T defined on the Hilbert space H, and denoting by I the identity operator on H, the main objective of this article is to establish some sufficient conditions on T which insure the invertibility of $(I + T)$ and the causality of $(I + T)^{-1}$.

The question of invertibility for $(I + T)$ has already received considerable attention in the technical literature. In particular, Browder [2], Dolph [10] and Minty [17], among others, have given sufficient invertibility conditions for the case in which T is a monotone operator. Similar results have been obtained by Petryshyn [19] and Shinbrot [28], who considered operators T with special compactness properties. These and other developments are summarized by Damborg [5] and will not be detailed here.

In regard to the causality of $(I + T)^{-1}$, the most familiar results are perhaps those offered by Foures and Segal [12] and Youla, Castriota and Carlin [29]. These results are essentially based on the Paley and Wiener theorem [18] and their application is confined to the case of linear and time-invariant systems. More recent developments concerning systems of a more general type are also available. Sandberg [27], for example, has considered nonlinear time-variant systems and has given an interesting connection between causality and energy related concepts. Damborg [3], [4] has established a sufficient condition for the causality of $(I + T)^{-1}$ in terms of an expression involving "incremental truncated" gain and phase shift concepts. Saeks [25] has considered linear systems in Hilbert space and has established the causality of $(I + T)^{-1}$ when T is causal and satisfies an inner product type condition. In a similar context Porter [22] has shown that $(I + T)^{-1}$ is causal whenever T is causal and dissipative.

A distinctive feature of the present development is that the invertibility of $(I + T)$ and the causality of $(I + T)^{-1}$ are investigated by focusing attention on those causal systems for which the future of the output is determined by the strict past of the input; such systems are said to be strictly causal. This approach leads in a natural way to the utilization of the Hilbert resolution space framework proposed in [22] and [26], and to the exploitation of the strict causality treatment developed in [25].

* Received by the editors February 15, 1972, and in revised form November 13, 1972.

† Ecole Polytechnique, Montréal, Québec, Canada. This work was done in part while the author was associated with the Electrical Engineering Department of the University of Michigan. This work was supported in part by the United States Air Force Office of Scientific Research under Grant 732427 and by the Canadian National Research Council under Grant A8244.

The interrelation of strict causality with causal invertibility has already been emphasized in the technical literature. Relevant efforts in this direction are illustrated by the work of Zames [34] and Willems [31] (see, in particular, [31, § 5.2, pp. 93–101]). More recently it has been shown that some results of Gohberg and Krein [14], [15] on the abstract theory of Volterra operators can be interpreted [7] as showing that if T is linear, completely continuous and strictly causal, then $(I + T)$ is invertible and $(I + T)^{-1}$ is causal. The scope and applicability of these results are hampered, however, on the one hand by the requirements of linearity and complete continuity and on the other hand by the adoption of a strict causality concept which is more restrictive than that usually considered in the technical literature.

The present development will show that linearity and complete continuity requirements are not at all essential to establish results of the Gohberg and Krein type. In particular, we will find, for example, that a sufficient condition for $(I + T)$ to be invertible and for its inverse to be causal is that T be given by the composition of a weakly additive causal and Lipschitz continuous system with a linear bounded and strictly causal system. This type of result will be shown to be applicable to a large class of systems which are not necessarily strictly causal.

The paper is organized as follows. Section 2 establishes the mathematical framework in which the study is embedded. Section 3 summarizes the definitions and properties associated with the notions of causality and strict causality. The aforementioned invertibility and causality results are to be found in § 4 and § 5. In particular, § 4 gives a number of sufficient conditions which simultaneously insure the invertibility of $(I + T)$ and the causality of $(I + T)^{-1}$. In § 5, $(I + T)^{-1}$ is assumed to exist and we offer sufficient conditions for its causality. Section 6 illustrates some connections between present results and problems of sensitivity and stability theory, and § 7 contains some concluding remarks about the overall development.

2. Mathematical preliminaries. The reader is assumed to be familiar with the notions of metric, linear, normed, Banach, inner product, and Hilbert spaces. The notions of linear and nonlinear mappings between such spaces are also assumed to be familiar. A unified treatment of such concepts is available, for example, in [20].

If x is an element of a Banach space B, the norm of x is indicated by the symbol $|x|$. If T is an operator on B, then T is said to be *bounded* if

$$|T| = \sup_{0 \neq x \in B} |Tx|/|x| < \infty.$$

The number $|T|$ is called the *norm* of T. T is *continuous* if for any $x \in B$ and any real $\varepsilon > 0$, there exists a real $\delta > 0$ such that $|Tx - Ty| \leqq \varepsilon$ when $|x - y| \leqq \delta$. T is *Lipschitz continuous* if

$$\|T\| = \sup_{0 \neq x - y \in B} |Tx - Ty|/|x - y| \leqq \infty.$$

$\|T\|$ is called the *Lipschitz norm* of T. Observe that when T is linear, then the concepts of boundedness, continuity and Lipschitz continuity are equivalent. T is called *compact* if $\overline{T(S)}$, the closure under T of a bounded set S, is a compact set. When T is compact and continuous, then it is said to be *completely continuous*.

Next the structure of Hilbert resolution spaces will be presented [22], [24], [25]. Suppose that H is a Hilbert space, and v a linearly ordered set with t_0 and t_∞, respectively, minimum and maximum elements. A family $\mathbb{R} = \{P^t\}, t \in v$, of orthogonal projections on H is a *resolution of the identity* if it enjoys the following two properties:

(Ri) $P^{t_0}H = 0$, $P^{t_\infty}H = H$, and $P^kH \supseteq P^lH$ whenever $k > l$;

(Rii) if $\{P^i\}$ is a sequence of orthogonal projections in $\mathbb{R}$ and there exists an orthogonal projection P such that $\{P^ix\} \to Px$, for every $x \in H$, then $P \in \mathbb{R}$.

A Hilbert space, H, equipped with a resolution of the identity, $\mathbb{R} = \{P^t\}$, is called a *Hilbert resolution space* (in short: HRS) and is denoted by the symbol $[H, P^t]$.

Example 2.1. Suppose that H is given by $L_2[0, \infty)$, the Hilbert space of Lebesgue square integrable real functions. In $L_2[0, \infty)$ a family of orthogonal projections $\mathbb{R} = \{P^t\}$, $t \in [0, \infty]$, can be defined as follows: if $x, y \in L_2[0, \infty)$ and $y = P^tx$, then $y(s) = x(s)$ on $[0, t]$ and $y(s) = 0$ in $[t, \infty)$. When $t = \infty$ then $P^\infty x = x$. The family $\mathbb{R} = \{P^t\}$, as it enjoys properties (Ri) and (Rii), is a resolution of the identity. It is then possible to view $[L_2[0, \infty), P^t]$ as a HRS in the following sections.[1]

The notion of integral on HRS will play a major role in this development. Suppose that $T(s)$, $s \in v$, is a family of operators on a HRS indexed by $s \in v$, and consider the following operations:

(i) Choose a partition Ω of v, $\Omega = \{\xi_0, \xi_1, \cdots, \xi_N\}$, where $\xi_0 = t_0$, $\xi_N = t_\infty$ and $\xi_j < \xi_{j+1}$, $j = 1, 2, \cdots, N-1$.

(ii) Consider the partial sum

$$I^\Omega = \sum_{k=1}^{N} \Delta P(\xi_k)T(s_k), \tag{1}$$

where $\Delta P(\xi_k) = P^{\xi_k} - P^{\xi_{k-1}}$ and $\xi_{k-1} \leqq s_k \leqq \xi_k$.

(iii) On the set of all partitions Ω of v, define a partial order as follows: $\Omega_1 \geqq \Omega_2$ if every element of Ω_2 is contained in Ω_1.

(iv) Suppose that there exists an operator T such that for any $\varepsilon > 0$ there is a partition Ω_ε of v such that the operator norm $|T - I^\Omega|$ is less than ε if $\Omega \leqq \Omega_\varepsilon$.

The operator T obtained through operations (i)–(iv) is called the *integral* of the family $T(s)$ with respect to $\mathbb{R}$ and is denoted by $T = \int dPT(s)$.

It is useful to consider slightly different variations of the above concept of integral. To this purpose, the notations $m\!\!\!\int dPT(s)$ and $M\!\!\!\!\int dPT(s)$ will be used to indicate the integrals which are obtained by choosing s_k in operation (ii) respectively as follows: $s_k = \xi_{k-1}$ or $s_k = \xi_k$. Similarly, the operator $\int dPT(s)\,dP$ will denote the integral which is obtained by replacing equation (1) in operation (ii) by the following:

$$I^\Omega = \sum_{k=1}^{N} \Delta P(\xi_k)T(s_k)\Delta P(\xi_k).$$

In the course of the development it will be natural to associate with an operator T on $[H, P^t]$, the family TP^s. This family will lead to integrals such as $\int dPTP^s$, $m\!\!\!\int dPTP^s$, $M\!\!\!\!\int dPTP^s$, and $\int dPT\,dP$.

[1] The Hilbert space whose elements are square summable sequences, l_2, can be viewed as a HRS in a similar way, and this is true also for the cross-product space $L_2^n \times l_2^m$. For other important examples of HRS, the reader is refered to [11] where reproducing kernel Hilbert spaces are considered.

Example 2.2. In $[L_2[0, \infty), P^t]$, the HRS described in Example 2.1, consider the operator T defined according to the following rule: if $y = Tx$, then $y(t) = g(t)x(t) + \int_0^\infty k(t, s)x(s)\, ds$, where $g \in L_2[0, \infty)$ and $k(t, s)$ is a Lebesgue square integrable kernel. In this case, it is not difficult to recognize that the operators $\int dPTdP$, $\int\!\!\!\!\!\mathrm{m}\, dPTP^s$, and $\int\!\!\!\!\!\mathrm{M}\, dPTP^s$ are well-defined. In particular, they are described as follows:

$$\left(\int dPT\, dPx\right)(t) = g(t)x(t),$$

$$\left(\int\!\!\!\!\!\mathrm{m}\, dPTP^s x\right)(t) = \int_0^t k(t, s)x(s)\, ds,$$

$$\left(\int\!\!\!\!\!\mathrm{M}\, dPTP^s x\right)(t) = g(t)x(t) + \int_0^t k(t, s)x(s)\, ds.$$

3. Some causality properties. In the sequel it will be supposed that T is a bounded operator defined on $[H, P^t]$. As proposed by Porter [22] and Saeks [25], T will be called *causal* (*anticausal*) if $P^tTy_1 = P^tTy_2$, whenever $P^ty_1 = P^ty_2$ ($(I - P^t)Ty_1 = (I - P^t)Ty_2$, whenever $(I - P^t)y_1 = (I - P^t)y_2$) for all $y_1, y_2 \in [H, P^t]$ and $t \in v$. T is *memoryless* if it is simultaneously causal and anticausal. T is *strictly causal* if $T = \int\!\!\!\!\!\mathrm{m}\, dPTP^s$.[2]

For later use and to gain some familiarity with these concepts, some causality properties are presented.

LEMMA 3.1 [25]. *The following statements are equivalent*: T *is causal*; $T = \int\!\!\!\!\!\mathrm{M}\, dPTP^s$; $P^tT = P^tTP^t$.

LEMMA 3.2 [9]. *A necessary and sufficient condition for a linear and causal T to be strictly causal is that* $\int dPTdP = 0$.

The proofs of Lemmas 3.1 and 3.2 are a direct consequence of the definitions of causality and integral, and will be omitted for brevity. A partial illustration of the techniques involved in these proofs is given by the proof of the following result.

THEOREM 3.1. *If T_1 is linear and strictly causal and T_2 is causal, then T_2T_1 and T_1T_2 are also strictly causal.*

Proof. By the definition of strict causality and Lemma 3.1, the following relations hold:

$$T_1 = \int\!\!\!\!\!\mathrm{m}\, dPT_1P^s \quad \text{and} \quad T_2 = \int\!\!\!\!\!\mathrm{M}\, dPT_2P^s.$$

From Lemma 3.2 and the definition of integral, these relations imply that, given any $\varepsilon > 0$, it is possible to find two partitions, Ω_ε' and Ω_ε'' of v, with the property that for all partitions Ω such that $\Omega = \{\xi_0, \xi_1, \cdots, \xi_N\} \geqq \Omega_\varepsilon = \Omega_\varepsilon' \cup \Omega_\varepsilon''$, the following holds:

$$\left|\sum_{k=1}^{N} \Delta P(\xi_k)T_1\Delta P(\xi_k)\right| < \varepsilon, \qquad |T_1 - \tilde{T}_1| < \varepsilon, \quad |T_2 - \tilde{T}_2| < \varepsilon,$$

[2] This definition is conceptually identical to that proposed in [25] and is more restrictive than that considered in [9] or the definition of strong causality used in [31]. In our terminology, a delay time in $L_2[a, b]$ is strictly causal only if a and b are finite.

where

$$\tilde{T}_1 = \sum_{k=1}^{N} \Delta P(\xi_k) T_1 P^{\xi_{k-1}} \quad \text{and} \quad \tilde{T}_2 = \sum_{k=1}^{N} \Delta P(\xi_k) T_2 P^{\xi_k}.$$

It follows that

$$|T_1T_2 - \tilde{T}_1T_2| \leqq |T_1 - \tilde{T}_1||T_2| \leqq \varepsilon|T_2|,$$

$$|\tilde{T}_1T_2 - \tilde{T}_1\tilde{T}_2| \leqq |\tilde{T}_1|\varepsilon \leqq (|T_1| + \varepsilon)\varepsilon.$$

Hence

$$|T_1T_2 - \tilde{T}_1\tilde{T}_2| \leqq |T_1T_2 - \tilde{T}_1T_2| + |\tilde{T}_1T_2 - \tilde{T}_1\tilde{T}_2| \leqq (|T_2| + |T_1| + \varepsilon)\varepsilon,$$

where (by virtue of the linearity of T_1)

$$\left| \tilde{T}_1\tilde{T}_2 - \sum_{k=1}^{N} \Delta P(\xi_k) T_1 T_2 P^{\xi_{k-1}} \right| < \varepsilon |T_2|.$$

These last two equations imply that $T_1T_2 = \mathrm{m}\!\!\!\int dPT_1T_2P^s$. Hence T_1T_2 is strictly causal. A similar argument applies for T_2T_1.

COROLLARY 3.1. *If $T_1, T_2, \cdots, T_n$ are linear causal operators, and T_{i_0} is strictly causal for some $i_0 \in \{1, 2, \cdots, n\}$, then the operator $T_1T_2 \cdots T_n$ is also strictly causal.*

Example 3.1. Consider in $[L_2[0, \infty), P^t]$ the operators T_1, T_2, T_3, T_4 defined as follows: if $y_i = T_ix$, $i = 1, 2, 3, 4$, then:

$$y_1(t) = N_l(x(t)),$$

where $N_l(\cdot)$ is a bounded Lipschitz continuous real function;

$$y_2(t) = \sum_{n=0}^{\infty} g_n x(t - \Delta t_n),$$

where $\Delta t_0 > 0$, $\Delta t_n > \Delta t_{n-1}$, and $\sum_{n=1}^{\infty} g_n < \infty$;

$$y_3(t) = \int_0^t k(t - \tau)x(t)\, d\tau,$$

where $\int_0^\infty |k(t)|\, dt < \infty$;

$$y_4(t) = h(t)x(t),$$

where $h \in L_\infty[0, \infty)$ and $\lim_{t\to\infty} \text{ess sup} |h(t)| = 0$.

It is easy to verify that for each $x \in L_2[0, \infty)$ and $t \in [0, \infty)$ the following relations hold:

$$P^tT_1x = P^tT_1P^tx \quad \text{and} \quad (I - P^t)T_1x = (I - P^t)T_1(I - P^t)x;$$

$$P^tT_2x = P^tT_2P^tx; \qquad P^tT_3x = P^tT_3P^tx;$$

$$P^tT_4x = P^tT_4P^tx_4 \quad \text{and} \quad (I - P^t)T_4x = (I - P^t)T_4(I - P^t)x.$$

It follows then that T_1 and T_4 are simultaneously causal and anticausal, hence they are memoryless; T_2 and T_3 are simply causal.

Observe that while none of the operators T_1, T_2, T_3 and T_4 is strictly causal, the composition T_4T_2 gives an example of a strictly causal operator. To prove this fact, it is sufficient to show (see Lemma 3.2) that given any $\varepsilon > 0$, a partition Ω_ε of $v = [0, \infty]$ can be found with the property that, for any other partition $\Omega' = \{\xi'_0, \xi'_1, \cdots, \xi'_{n'}\} \geqq \Omega_\varepsilon$, the following holds:

$$\left| \sum_{i=1}^{n} \Delta P(\xi'_i) T_4 T_2 \Delta P(\xi'_i) \right| \leqq \varepsilon .$$

This is easily done by choosing an element $\bar{\xi} \in [0, \infty)$ such that

$$|(I - P^{\bar{\xi}})T_4(I - P^{\bar{\xi}})| \leqq \varepsilon/(2|T_2|)$$

and by verifying that every partition

$$\Omega_\varepsilon = \{0, \xi_1, \xi_2, \cdots, \xi_n = \bar{\xi}, \infty\}$$

such that $|\xi_i - \xi_{i-1}| < \Delta t_0$ for all $i = 1, 2, \cdots, n$, enjoys the desired property.

Arguments similar to the above would also show that T_2T_4, T_3T_4 and T_4T_3 are all strictly causal operators. Moreover, it is interesting to note that, for every permutation (i_1, i_2, i_3, i_4) of $(1, 2, 3, 4)$, the operator $T_{i_1}T_{i_2}T_{i_3}T_{i_4}$ is also strictly causal.

4. Existence and causality of $(I + T)^{-1}$. The concepts of causality, strict causality and memorylessness will now be supplemented with the concept of weak additivity.[3] This is done via the following definition and the two subsequent lemmas.

DEFINITION. T is called *weakly additive* if $Tx = TP^tx + T(I - P^t)x$ holds for every $x \in H$, and all $t \in v$.

LEMMA 4.1 [6, p. 114]. *The following operators are weakly additive: every linear operator; every memoryless operator; the linear combination of weakly additive operators; the composition* T_2T_1, *where* T_1 *is weakly additive and* T_2 *is linear; the composition* T_2T_1, *where* T_1 *is memoryless and* T_2 *is weakly additive.*

LEMMA 4.2. *If* T *is weakly additive and causal, then for every pair* $P^1, P^2 \in \{P^t\}$ *one has the following operator identity:*

$$(P^2 - P^1)T = (P^2 - P^1)TP^1 + (P^2 - P^1)T(P^2 - P^1).$$

We are now in a position to state and prove the first fundamental result of the paper.

THEOREM 4.1. *Suppose that* $T = T_1N_lT_2$, *where:*

(i) N_l *is a causal, weakly additive, bounded and Lipschitz continuous operator;*

(ii) T_1 *and* T_2 *are linear, bounded and causal operators;*

(iii) *either* T_1, *or* T_2, *or* T_2T_1 *is strictly causal. Then* $(I + T)$ *is invertible and its inverse is causal, bounded and continuous.*

Proof. First, it will be shown that it is sufficient to prove the theorem in the special case where T_2 coincides with the identity and T_1 is strictly causal. Indeed, if the theorem is valid in this special case, then from Theorem 3.1, one would

[3] In a system theory context, the concept of weak additivity has been explicitly exploited by earlier authors. In particular, it has been adopted by Zadeh [32], and, more recently, by Gersho [13].

obtain the following: if T_1 and T_2 satisfy (ii) and (iii), then the operator $(I + T_2T_1N_l)$ is invertible and $(I + T_2T_1N_l)^{-1}$ is bounded, causal and continuous. This latter result implies the validity of the theorem in the general case. To see this, note that the invertibility of $(I + T)$ is equivalent to the property that for every element $y \in H$ there exists one and only one $x \in H$ such that

$$y = x + T_1N_lT_2x. \tag{2}$$

This, in turn, is equivalent to the existence of a unique element $T_2x \in H$ such that

$$T_2y = T_2x + T_2T_1N_lT_2x. \tag{3}$$

But, from the invertibility of $(I + T_2T_1N_l)$, one has that the desired T_2x exists and is given by

$$T_2x = (I + T_2T_1N_l)^{-1}T_2y. \tag{4}$$

It follows that the solution of equation (2) also exists, is unique and has the property that

$$x = y - T_1N_l(I + T_2T_1N_l)^{-1}T_2y. \tag{5}$$

The operator $(I + T_1N_lT_2)$ is then invertible and its inverse is given by

$$(I + T_1N_lT_2)^{-1} = I - T_1N_l(I + T_2T_1N_l)^{-1}T_2. \tag{6}$$

This last equation shows also that $(I + T_1N_lT_2)^{-1}$ can be expressed by the sum and composition of causal, bounded, and continuous operators, and therefore $(I + T_1N_lT_2)^{-1}$ is itself causal, bounded, and continuous.

It remains then to show that the theorem is valid in the case where T_1 is strictly causal and $T_2 = I$. To verify the invertibility of $(I + T)$ in this special case, one has again to show that for every $y \in H$ there exists a unique $x \in H$ such that

$$y = x + Tx. \tag{7}$$

From Lemma 3.2 one has $\int dPT_1\, dP = 0$. This implies that there exists a partition $\Omega = \{t_0 = \xi_0, \xi_1, \xi_2, \cdots, \xi_N = t_\infty\} \in v$ such that

$$\|\sum \Delta P(\xi_i)T_1\Delta P(\xi_i)\| < 1/\|N_l\|,$$

that is,

$$\|\Delta P(\xi_i)T_1\Delta P(\xi_i)\| < 1/\|N_l\|$$

for each $i = 1, 2, \cdots, N$. Using Lemma 4.2, and the fact that N_l is memoryless, it follows that

$$\|\Delta P(\xi_i)T_1N_l\Delta P(\xi_i)\| = \|\Delta P(\xi_i)T\Delta P(\xi_i)\| < 1 \tag{8}$$

for each $i = 1, 2, \cdots, N$.

Observe now that solving equation (7) is equivalent to finding an $x \in H$ such that

$$\Delta P(\xi_i)y = \Delta P(\xi_i)x + \Delta P(\xi_i)TP^{\xi_i}x, \tag{9}$$

where $i = 1, 2, \cdots, N$. Indeed, if $x \in H$ satisfies (7), then, from the causality of T, $\Delta P(\xi_i)x$ must clearly satisfy (9). Conversely, suppose that the element $x \in H$ is such that $\Delta P(\xi_i)x$ does satisfy (9). Then it would follow that

$$\sum_{i=1}^{N} \Delta P(\xi_i)y = \sum_{i=1}^{N} \Delta P(\xi_i)x + \sum_{i=1}^{N} \Delta P(\xi_i)TP^{\xi_i}x,$$

and this equation coincides with (7).

For $i = 1$, equation (9) becomes

$$\Delta P(\xi_1)y = \Delta P(\xi_1)x + \Delta P(\xi_1)T\Delta P(\xi_1)x. \tag{10}$$

But from (8) one has that $\Delta P(\xi_1)T\Delta P(\xi_1)$, the restriction of T to the Hilbert space $\Delta P(\xi_1)H$, is Lipschitz continuous and has a Lipschitz norm smaller than 1. Applying Lemma 4.3 (given below), there exists then a causal operator K_1 such that the element

$$\Delta P(\xi_1)x = K_1\Delta P(\xi_1)y \tag{11}$$

is the unique solution of equation (10).

For $i = 2$, equation (9) becomes

$$\Delta P(\xi_2)y = \Delta P(\xi_2)x + \Delta P(\xi_2)TP^{\xi_2}x. \tag{12}$$

Noting that T is weakly additive (see Lemma 4.1), and using Lemma 4.2, this last equation can be rewritten as follows:

$$\Delta P(\xi_2)y - \Delta P(\xi_2)T\Delta P(\xi_1)x = \Delta P(\xi_2)x + \Delta P(\xi_2)T\Delta P(\xi_2)x,$$

where, once again, $\Delta P(\xi_2)T\Delta P(\xi_2)$ is Lipschitz continuous and has a Lipschitz norm smaller than 1. Lemma 4.3 can then be applied again and there exists a causal operator K_2 which provides the following unique solution to (12):

$$\Delta P(\xi_2)x = K_2[\Delta P(\xi_2)(y - T\Delta P(\xi_1)x)],$$

where $\Delta P(\xi_1)x$ is defined by (11). By induction, having computed $\Delta P(\xi_1)x, \Delta P(\xi_2)x, \cdots, \Delta P(\xi_{i-1})x$, the unique solution to (9), $\Delta P(\xi_i)x$, can be computed as follows:

$$\Delta P(\xi_i)x = K_i\left[\Delta P(\xi_i)\left(y - T\sum_{j=1}^{i-1} \Delta P(\xi_j)x\right)\right], \tag{13}$$

where K_i is a causal operator.

The above recursive relations define the element $x = \sum_{i=1}^{N} \Delta P(\xi_i)x$, and this element is the unique solution to (7). It can then be concluded that $(I + T)$ is indeed invertible.

With regard to the causality of $(I + T)^{-1}$, it is sufficient to show that for every $t \in v$ and all $y \in H$, the following holds (see Lemma 3.1):

$$P^t(I + T)^{-1}y = P^t(I + T)^{-1}P^ty. \tag{14}$$

Consider the partition Ω' of v given by

$$\Omega' = \{t_0 = \xi_0, \xi_1, \xi_2, \cdots, \xi_{i-1}, t, \xi_i, \cdots, \xi_N = t_\infty\},$$

where $\{\xi_0, \xi_1, \cdots, ,\xi_N\}$ is the partition Ω considered in the first part of the proof, and it has been assumed, without any loss of generality, that $\xi_{i-1} < t < \xi_i$. Use the following notations:

$$x^1 = (I + T)^{-1}y^1, \qquad x^2 = (I + T)^{-1}y^2,$$

where $y^1 = y$ and $y^2 = P^t y$. From equation (13) the following relation holds:

$$\begin{aligned} x^q = (I + T)^{-1}y^q = \sum_{\substack{j=1 \\ j \neq i}}^{N} & K_j[\Delta P(\xi_j)(y^q - TP^{\xi_{j-1}}x^q)] \\ & + K_t[(P^t - P^{\xi_{i-1}})(y^q - TP^{\xi_{i-1}}x^q)] \\ & + K_i[(P^{\xi_i} - P^t)(y^q - TP^{\xi_t}x^q)], \end{aligned}$$

where $q = 1, 2$. By inspection, from this equation it follows that

$$P^{\xi_1}x^1 = P^{\xi_1}x^2, \quad P^{\xi_2}x^1 = P^{\xi_2}x^2, \cdots, P^{\xi_{i-1}}x^1 = P^{\xi_{i-1}}x^2, \quad P^t x^1 = P^t x^2.$$

This implies the validity of equation (14). The proof of the boundedness and continuity of $(I + T)^{-1}$ can be obtained in a similar way (the missing details can be found in [6]).

LEMMA 4.3 [6, p. 131]. *If T is a causal, bounded and Lipschitz continuous operator with Lipschitz norm less than* 1, *then $(I + T)$ is invertible and its inverse is causal, bounded, and Lipschitz continuous.*

Example 4.1. Consider on $[L_2[0, \infty), P^t]$ the operator T given by the composition $T_4T_1(T_3 + T_2)$, where T_1, T_2, T_3 and T_4 are the operators defined in Example 3.1. Note that T_3T_4 and T_2T_4 are linear and strictly causal, hence $T_3T_4 + T_2T_4$ is also linear and strictly causal. Moreover, T_1 is memoryless and Lipschitz continuous. It follows then that T satisfies the hypotheses of Theorem 4.1, and consequently $(I + T)$ is invertible and its inverse $(I + T)^{-1}$ is causal, bounded and continuous.

It is of interest to observe that the statement of Theorem 4.1 is automatically valid when T is linear and strictly causal. In this special case, however, that result can be further strengthened as follows.

THEOREM 4.2 [6, p. 141]. *If T is a linear bounded and strictly causal operator, then $(I + T)$ is invertible and its inverse is causal and bounded. Moreover, $(I + T)^{-1}$ can be computed by the following Neumann series*:

$$(I + T)^{-1} = I + \sum_{n=1}^{\infty} (-1)^n T^n,$$

where $T^1 = T$ and T^{n+1} is given by the composition of T^n with T.[4]

This theorem can be proved by applying techniques similar to those already used in the case of Theorem 4.1. The proof is based on the following well-known specialization of Lemma 4.3 and is omitted for brevity.

LEMMA 4.4 [6, p. 140]. *If T is a linear causal bounded operator with norm less than* 1, *then $(I + T)$ is invertible and its inverse is causal and continuous. Moreover,*

[4] In the case where T is also completely continuous, this theorem provides the Gohberg and Krein result remarked on in the Introduction.

$(I + T)^{-1}$ *can be computed by the following Neumann series*:

$$(I + T)^{-1} = I + \sum_{n=1}^{\infty} (-1)^n T^n.$$

When T is not strictly causal, Theorems 4.1 and 4.2 cannot be directly applied. A number of variations on these theorems are, however, available. This is illustrated, for example, by the following two theorems.

THEOREM 4.3. *Suppose that the operator K is given by the sum $K = \tilde{T} + T$, where $\tilde{T}$ satisfies the hypotheses of Theorem* 4.1, *and T is a linear bounded causal operator with the property that $(I + T)^{-1}$ exists and is causal. Then $(I + K)$ is invertible and its inverse is causal, bounded and continuous.*[5]

THEOREM 4.4 [6, p. 141]. *Suppose that the operator K is given by the sum $K = \tilde{T} + T$, where $\tilde{T}$ satisfies the hypotheses of Theorem* 4.2 *and T is a linear bounded causal operator with norm less than* 1. *Then $(I + K)$ is invertible and its inverse is causal, bounded and continuous. Moreover, $(I + K)^{-1}$ can be computed by the following Neumann series*:

$$(I + K)^{-1} = I + \sum_{n=1}^{\infty} (-1)^n K^n.$$

5. Additional results on the causality of $(I + T)^{-1}$. When an operator T on a Hilbert resolution space $[H, P^t]$ satisfies the hypotheses stated in Theorems 4.1 or 4.3, then those hypotheses are also satisfied for every operator of the family P^sTP^s, where $t_\infty \neq s \in v$. Unfortunately, the converse of this statement is not necessarily true. Indeed, in many system problems of interest it occurs that the operators P^sTP^s, $t_\infty \neq s \in v$, satisfy the hypotheses of Theorems 4.1, 4.3, while the operator T does not. Under this assumption, nothing can in general be said about the invertibility of $(I + T)$. The next result establishes, however, that if $(I + T)$ is invertible, then its inverse is causal.[6]

THEOREM 5.1. *Suppose that T is a causal operator on $[H, P^t]$ and that for each $s \in v$, $s \neq t_\infty$, P^sTP^s satisfies the hypotheses of Theorem* 4.1 (*or Theorem* 4.3). *If $(I + T)$ is invertible, then its inverse is causal.*

Proof. Given any $y \in H$, consider the element $x = (I + T)^{-1}y$. Clearly $y = x + Tx$ and for each $s \in v$, one has that

$$P^s y = P^s x + P^s T x.$$

Using the causality of T, one finds that

$$P^s y = P^s x + P^s T P^s x.$$

[5] The proof of this result is contained in the proof of Theorem 6.2.

[6] It is noted that the causality of T is not sufficient to establish the causality of $(I + T)^{-1}$. In this regard, some illustrative counterexamples can be found in [5] and [30].

But by hypothesis, P^sTP^s satisfies the conditions of Theorem 4.1 (or Theorem 4.3). This implies that $(I + P^sTP^s)$ is invertible in P^sH, and therefore

$$P^sx = [P^s + P^sTP^s]^{-1}P^sy.$$

From this last equation one obtains that

$$P^s(I + T)^{-1} = [P^s + P^sTP^s]^{-1}P^sy = P^s(I + T)^{-1}P^sy,$$

and, by Lemma 3.1, it follows that $(I + T)^{-1}$ is causal.

Example 5.1. Consider in $[L_2[0, \infty), P^t]$ the operator T given by $T = T_3T_1$, or $T = T_1T_3$, where T_1 and T_3 are the operators defined in Example 3.1. The operator T is not necessarily strictly causal and in general neither Theorem 4.1 nor 4.3 can be applied. Indeed, many instances can be mentioned where the operator $(I + T)$ is not invertible. Theorem 5.1 says, however, that if $(I + T)$ is invertible, then $(I + T)^{-1}$ is causal. This result can be applied because, as it is easy to verify, the restriction of the operator T_3 to $[L_2[0, s), P^t]$ is strictly causal for every $s \in [0, \infty)$, and as a consequence the restriction of T to $P^s[L_2[0, \infty), P^t]$ satisfies the conditions of Theorem 4.1.

In a number of other interesting situations, it may happen that while T is strictly causal, neither T nor P^sTP^s satisfies the hypotheses of Theorem 4.1 or 4.3. As in the previous case, the invertibility of $(I + T)$ cannot in general be ascertained. However, if $(I + T)^{-1}$ is known to exist, then its causality can be established.

THEOREM 5.2. *Suppose that T is a strictly causal bounded operator and that $(I + T)^{-1}$ exists and is Lipschitz continuous. Then $(I + T)^{-1}$ is causal.*

Instead of proving this theorem, a slight generalization of it will be proved. This also will provide an opportunity to illustrate the techniques necessary to extend Theorems 4.1 and 4.2 respectively into Theorems 4.3 and 4.4. The generalization in question is the following.

THEOREM 5.3. *Suppose that T satisfies the following conditions*:

(a) *$T = T_{\underline{C}} + T_C$, where $T_{\underline{C}}$ and T_C are respectively strictly causal and causal operators.*

(b) *$T_{\underline{C}}$ is bounded and T_C is Lipschitz continuous.*

(c) *T_C is weakly additive and has Lipschitz norm less than* 1.

(d) *$(I + T)$ is invertible and $(I + T)^{-1}$ is Lipschitz continuous.*

Then $(I + T)^{-1}$ is causal.

Proof. Note first that, since $T_{\underline{C}}$ is strictly causal, there exists a sequence of operators

$$I^{\Omega_i} = \sum_{j=1}^{N} \Delta P(\xi_j)T_{\underline{C}}P^{\xi_{j-1}}$$

such that $\{I^{\Omega_i}\} \to T_{\underline{C}}$, and $\{I^{\Omega_i} + T_C\} \to T$. Moreover, for each operator $(I^{\Omega_i} + T_C)$ and any element $y \in [H, P^t]$, the equation

$$y = x + I^{\Omega_i}x + T_Cx$$

has the unique solution x defined by the following recursive relations:[7]

$$\begin{aligned}
\Delta P(\xi_1)x &= \Delta P(\xi_1)(I + T_C)^{-1}\Delta P(\xi_1)y,\\
\Delta P(\xi_2)x &= \Delta P(\xi_2)(I + T_C)^{-1}\Delta P(\xi_2)y - \Delta P(\xi_2)T_C\Delta P(\xi_1)x - \Delta P(\xi_2)I^{\Omega_i}\Delta P(\xi_1)x,\\
&\ \ \vdots\\
\Delta P(\xi_j)x &= \Delta P(\xi_j)(I + T_C)^{-1}\Delta P(\xi_j)y - \Delta P(\xi_j)T_C P^{\xi_{j-1}}x - \Delta P(\xi_j)I^{\Omega_i}P^{\xi_{j-1}}x,\\
&\ \ \vdots\\
\Delta P(\xi_N)x &= \Delta P(\xi_N)(I + T_C)^{-1}\Delta P(\xi_N)y - \Delta P(\xi_N)T_C P^{\xi_{N-1}}x - \Delta P(\xi_N)I^{\Omega_i}P^{\xi_{N-1}}x.
\end{aligned}$$

These equations clearly imply that $(I + I^{\Omega_i} + T_C)$ is invertible and its inverse $(I + I^{\Omega_i} + T_C)^{-1}$ is causal. In view of Lemma 3.1, to prove that $(I + T)^{-1}$ is causal, it is then sufficient to show that if y is any element of $[H, P^t]$, then the following relation holds:

$$|(I + I^{\Omega_i} + T_C)^{-1}y - (I + T)^{-1}y| \to 0. \tag{15}$$

Indeed, if this relation holds, then it would follow that

$$\begin{gathered}
\{P^t(I + I^{\Omega_i} + T_C)^{-1}y\} \to P^t(I + T)^{-1}y,\\
\{P^t(I + I^{\Omega_i} + T_C)^{-1}P^t y\} \to P^t(I + T)^{-1}P^t y;
\end{gathered}$$

and, since from the causality of $(I + I^{\Omega_i} + T_C)^{-1}$ one has that

$$P^t(I + I^{\Omega_i} + T_C)^{-1}y = P^t(I + I^{\Omega_i} + T_C)^{-1}P^t y,$$

one would also have that

$$P^t(I + T)^{-1}y = P^t(I + T)^{-1}P^t y.$$

Suppose that equation (15) is not true. Then there would exist a positive real ε and $\{T_n\}$, a subsequence of $\{I^{\Omega_i} + T_C\}$, such that

$$|(I + T_n)^{-1}y - (I + T)^{-1}y| > \varepsilon. \tag{16}$$

To see that this is impossible, denote $\{(I + T_n)^{-1}y\}$ and $(I + T)^{-1}y$ respectively by $\{x_n\}$ and x, and observe that

$$y = x_n - T_n x_n = x - Tx. \tag{17}$$

Note also that from the boundedness of $(I + T)^{-1}$, it follows that $(I + T_n)^{-1}$ is uniformly bounded, and consequently the sequence $\{x_n\}$ is also uniformly bounded,

[7] These relations follow from the hypothesis that the Lipschitz norm of T_C is less than 1 and T_C is weakly additive, plus the application of Lemma 4.3.

that is,

$$|x_n| = |(I + T_n)^{-1}y| \leqq My, \tag{18}$$

where M is a positive number conveniently chosen.[8]

Consider now the sequence $\{y_n\}$ given by $y_n = (I + T)x_n$ and observe that from (17) and (18) one has that

$$\begin{aligned} |y_n - y| &= |(I + T)x_n - (I + T_n)x_n| \\ &= |(T - T_n)x_n| \leqq M|T - T_n|y. \end{aligned} \tag{19}$$

Moreover, clearly one has also that

$$(I + T)^{-1}y - (I + T)^{-1}y_n = (I + T)^{-1}y - (I + T_n)^{-1}y.$$

From this equation and (16) it follows that

$$|(I + T)^{-1}y - (I + T)^{-1}y_n| > \varepsilon.$$

Applying (19), one obtains

$$\frac{|(I + T)^{-1}y - (I + T)^{-1}y_n|}{|y - y_n|} \geqq \frac{\varepsilon}{M|T - T_n|y}.$$

But, as the sequence $\{T_n\}$ converges to T, $|T - T_n|$ can be as small as desired. As a consequence, the last equation implies that $(I + T)^{-1}$ is not Lipschitz continuous, and one obtains a contradiction to hypothesis (d).

6. Applications. Potential applications for the ideas and results developed in the previous sections can be envisioned in a number of engineering areas, such as system sensitivity, stability, game theory, optimal control, communication theory and others. In this section we shall briefly illustrate how some of these applications can be realized in a system sensitivity and stability theory context.

Consider the systems represented in Figs. 1 and 2, where the blocks G, $\mathbb{P}$, M are described by linear bounded and causal operators on the Hilbert resolution

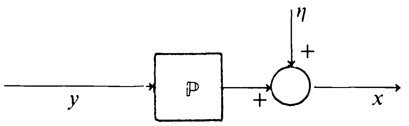

FIG. 1. *The system considered in the sensitivity problem*

[8] By the boundedness of $(I + T)^{-1}$, there exists a positive real m such that for every $x \in [H, P^t]$ the following relation holds:

$$\frac{|(I + T)x|}{|x|} \geqq 2m.$$

On the other hand, one can always choose an integer N such that $n \geqq N$ implies $|T - T_n| \leqq m$. For $n \geqq N$ one has then that

$$\frac{|(I + T_n)x|}{|x|} \geqq \frac{|(I + T)x|}{|x|} - \frac{|(T - T_n)x|}{|x|}.$$

It follows that $|(I + T_n)x| \geqq |x|$ and therefore $|(I + T_n)^{-1}| \leqq 1/m = M$.

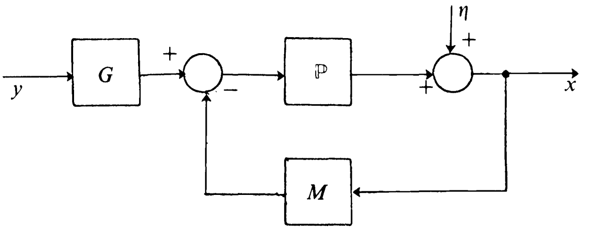

FIG. 2. *A feedback compensation scheme for system sensitivity improvement*

space $[L_2^n[0, \infty), P^t]$ and η is an element of $[L_2^n[0, \infty), P^t]$. In a typical sensitivity problem, $\mathbb{P}$ represents a given physical plant, and G and M are two compensators with the property that the compensated system (Fig. 2) is input-output equivalent to the original system (Fig. 1). This input-output equivalence can be obtained, for example, by choosing $G = I + M\mathbb{P}$. The question of interest is then to determine M in such a way that the compensated system has a sensitivity with respect to the "perturbation" η which is better than that of the original system.

While a proper review of philosophy, motivations and results related to this sensitivity problem is well beyond the scope of the present discussion, here it will be sufficient to consider the following widely adopted sensitivity criteria (see, for example, Anderson and Newcomb [1], Porter [21], [23], and Zahm [33]).

Sensitivity reduction criteria (Zahm [33, p. 51]). The sensitivity of the compensated system is better than the sensitivity of the original system if the following conditions are satisfied:

(i) $\mathscr{S} = (I + \mathbb{P}M)^{-1}$ is a well-defined operator in $L_2^n[0, \infty)$;

(ii) $\mathscr{S}$ is a causal operator;

(iii) $\langle x, x - \mathscr{S}^*\mathscr{S}x\rangle \geqq 0$ for every $x \in L_2^n[0, \infty)$.

The above system sensitivity criteria reduce a good portion of the sensitivity problem to an invertibility and causality problem. The development of the previous sections can then be used to gain insight into the structure of sensitivity reduction. The following theorem illustrates the types of results which are obtainable in this regard.

THEOREM 6.1. *If either the plant $\mathbb{P}$, or the compensator M, or $\mathbb{P}M$, is strictly causal, then the first two conditions for sensitivity reduction are satisfied and the third condition becomes*:

$$\frac{\operatorname{Re}\langle x, (I + \mathbb{P}M)^{-1}\mathbb{P}Mx\rangle}{|(I + \mathbb{P}M)^{-1}\mathbb{P}Mx|^2} \geqq \frac{1}{2},$$

where x is any element in $L_2^n[0, \infty)$.

Proof. Applying Theorem 3.1, the operator $\mathbb{P}M$ is strictly causal, and therefore, from Theorem 4.2, $(I + \mathbb{P}M)$ is invertible and $(I + \mathbb{P}M)^{-1}$ is causal. Moreover, $(I + \mathbb{P}M)^{-1}$ is given by the following expression:

$$(I + \mathbb{P}M)^{-1} = I + \sum_{n=1}^{\infty}(-1)^n(\mathbb{P}M)^n. \tag{20}$$

The third condition for sensitivity improvement then becomes

$$\langle x, x - \mathcal{S}^*\mathcal{S}x\rangle = \langle x, x\rangle - \langle (I + \mathbb{P}M)^{-1}x, (I + \mathbb{P}M)^{-1}x\rangle \geqq 0,$$

and from equation (20),

$$\langle x, x\rangle - \left\langle x + \sum_{n=1}^{\infty} (-1)^n(\mathbb{P}M)^n x, x + \sum_{n=1}^{\infty} (-1)^n(\mathbb{P}M)^n x\right\rangle \geqq 0.$$

This leads to

$$-2\,\mathrm{Re}\left\langle \sum_{n=1}^{\infty} (-1)^n(\mathbb{P}M)^n x, x\right\rangle - \sum_{n=1}^{\infty} (-1)^n(\mathbb{P}M)^n x^2 \geqq 0. \tag{21}$$

Observe now that

$$\sum_{n=1}^{\infty} (-1)^n(\mathbb{P}M)^n = -(I + \mathbb{P}M)^{-1}\mathbb{P}M,$$

and therefore equation (21) becomes

$$\mathrm{Re}\,\langle (I + \mathbb{P}M)^{-1}\mathbb{P}Mx, x\rangle \geqq \tfrac{1}{2}|(I + \mathbb{P}M)^{-1}\mathbb{P}Mx|^2.$$

If, with Damborg [5, p. 32], we interpret $\mathrm{Re}\,\langle x, Tx\rangle/|Tx||x|$ and $|Tx|/|x|$ as the phase and gain respectively of T, then the above result can be paraphrased as follows.

THEOREM 6.1. *If either* $\mathbb{P}$, *or* $\mathbb{P}M$ *is strictly causal, then a necessary and sufficient condition for sensitivity improvement is that the ratio between the phase and the gain of* $(I + \mathbb{P}M)^{-1}\mathbb{P}M$ *be bigger than or equal to* $\frac{1}{2}$.

Let us now turn our attention to the feedback system of Fig. 3 where K is a bounded, continuous and causal operator on $[H, P^t]$. We will say that this feedback system is (bounded input-bounded output) *stable* if it has the property that in correspondence to any input $y \in H$, the output x is a well-defined element of H and the input-output mapping is causal bounded and continuous. This definition of stability is consistent with that used in the normed space stability approach developed by Damborg [5] and Willems [30]; a discussion of its connections with the more classical definition used in the extended space stability approach (see, for example, Zames [35]) goes beyond the scope of the present development and can be found in [3] and [31]. Note that these latter references also consider the case where K is unbounded and not necessarily defined in all of H. For our purpose it is sufficient to recall the following result.

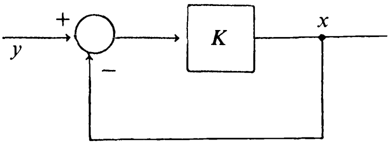

FIG. 3. *The feedback system considered in the stability problem*

LEMMA 6.1. *Necessary and sufficient conditions for the basic feedback system to be stable are that* $(I + K)$ *be invertible and* $(I + K)^{-1}$ *be causal, bounded and continuous.*[9]

This lemma allows us to view most of the development of the previous sections in a stability context. The full exploration of this connection is beyond the intended scope of this paper. However, we shall state a single theorem and corollary which foster conceptual insight.

THEOREM 6.2. *Suppose that* T *is a linear operator and that the basic feedback system is stable for* $K = T$. *Then the basic feedback system is also stable for* $K = (T + \tilde{T})$, *where* $\tilde{T}$ *denotes an operator satisfying the hypotheses of Theorem* 3.1.

Proof. By Lemma 6.1, it will be sufficient to show that $(I + T + \tilde{T})$ is invertible and that $(I + T + \tilde{T})^{-1}$ is causal, bounded and continuous. To this purpose observe first that if the operator $[I + (I + T)^{-1}\tilde{T}]$ is invertible, then $(I + T + \tilde{T})$ is also invertible and

$$(I + T + \tilde{T})^{-1} = [I + (I + T)^{-1}\tilde{T}]^{-1}(I + T)^{-1}. \tag{22}$$

Indeed, suppose that $[I + (I + T)^{-1}\tilde{T}]$ is invertible. Then for every $y \in [H, P^t]$ there would exist an x such that

$$x = [I + (I + T)^{-1}\tilde{T}]^{-1}(I + T)^{-1}y.$$

It would then follow that

$$(I + T)^{-1}y = x + (I + T)^{-1}\tilde{T}y$$

and therefore

$$y = x + Tx + \tilde{T}x.$$

This last equation would imply the invertibility of $(I + T + \tilde{T})$ and the validity of equation (22).

At this point the proof can be completed by showing that $[I + (I + T)^{-1}\tilde{T}]$ is in fact invertible and that $[I + (I + T)^{-1}\tilde{T}]^{-1}$ is causal, bounded and continuous. This is readily done by observing that $(I + T)^{-1}$ is (by hypothesis and Lemma 6.1) causal, bounded and continuous and by verifying that $(I + T)^{-1}\tilde{T}$ satisfies the conditions of Theorem 4.1.

COROLLARY 6.1. *If* K *is given by the composition of a weakly additive, bounded and Lipschitz continuous operator with a linear, bounded and strictly causal operator, then the basic feedback system is stable.*

7. Conclusions. The primary results of this paper are embedded in Theorems 4.1 and 5.2, and emphasize the importance of the concept of strict causality in connection with questions of existence and causality of an inverse system. In particular, Theorem 4.1 states that the strict causality of a system T plus some other reasonable conditions are sufficient to insure existence and causality of $(I + T)^{-1}$. Theorem 5.2 considers more relaxed conditions and establishes the causality of $(I + T)^{-1}$ when this system exists and is Lipschitz continuous. These results can be extended in various directions, and some examples of these extensions are

[9] In a framework slightly different from that adopted here, this result can be found either in [5] or in [31]. In [8] it is shown that the proof used by these two authors is also applicable to the present HRS context.

provided by Theorems 4.3, 5.1 and 5.3. Useful specializations are also possible. For instance, in the case of linear systems, Theorem 4.1 leads to Theorems 4.2 and 4.4 which offer substantial generalizations of a well-known Gohberg and Krein result.

Finally, some connections of the theory to system sensitivity and stability problems have been pointed out. In this regard, Theorem 6.1 illustrates the relevance of the concepts of gain and phase shift in connection with sensitivity theory. Theorem 6.2 provides formalization to the conceptual connections between causality, strict causality and weak additivity on the one hand, and the stability of a basic feedback system on the other.

Acknowledgment. This paper has benefited from a number of suggestions and comments by Professor William A. Porter of the University of Michigan.

REFERENCES

[1] B. D. O. ANDERSON AND R. W. NEWCOMB, *An approach to the time-varying sensitivity problem*, TR 6560, Stanford Electronics Laboratory, Stanford, Calif., 1966.

[2] F. E. BROWDER, *The solvability of nonlinear functional equations*, Duke Math. J., 30 (1962), pp. 557–566.

[3] M. DAMBORG AND A. W. NAYLOR, *Stability structure for feedback systems having unstable open loops*, IEEE Trans. Automatic Control, AC-18 (1973), pp. 318–319.

[4] ———, *The fundamental structure of input-output stability for feedback systems*, IEEE Trans. Systems Science and Cybernetics, 1970, pp. 92–96.

[5] M. DAMBORG, *Stability of the basic nonlinear operator feedback system*, Tech. Rep. 37, Systems Engineering Laboratory, University of Michigan, Ann Arbor, 1969.

[6] R. M. DE SANTIS, *Causality structure of engineering systems*, Ph. D. thesis, University of Michigan, Ann Arbor, 1971.

[7] ———, *On some connections between causality and stability*, Preprints 14th Midwest Symposium on Circuit Theory, Denver, 1971.

[8] ———, *Espaces de resolution Hilbertienne et theorie de la stabilité*, Tech. Rep. EP73-R-1, Ecole Polytechnique de Montréal, 1973.

[9] ———, *Causality for nonlinear systems in Hilbert resolution spaces*, Math. Systems Theory, vol. 17, no. 4, to appear.

[10] C. L. DOLPH AND G. J. MINTY, *On nonlinear integral equations of the Hammerstein type*, Nonlinear Integral Equations, P. M. Anselone, ed., University of Wisconsin Press, Madison, 1964, pp. 99–154.

[11] D. L. DUTTWEILER, *Reproducing kernel Hilbert space techniques for detection and estimation problems*, Tech. Rep. 7050-18, Information Systems Laboratory, Stanford University, Stanford, Calif., 1970.

[12] Y. FOURES AND I. SEGAL, *Causality and analicity*, Amer. Math. Soc. 78 (1955), pp. 385–405.

[13] A. GERSHO, *Nonlinear systems with a restricted additivity property*, IEEE Trans. Circuit Theory, CT-16 (1969), pp. 150–154.

[14] I. Z. GOHBERG AND M. G. KREIN, *Introduction to the Theory of Linear Nonselfadjoint Operators*, vol. 18, American Mathematical Society, Providence, R.I., 1969.

[15] ———, *Theory of Volterra Operators in Hilbert Space and Applications*, vol. 24, American Mathematical Society, Providence, R.I., 1970.

[16] T. KAILATH AND D. L. DUTTWEILER, *Generalized innovation processes and some applications*, 14th Midwest Symposium on Circuit Theory, Denver, 1971.

[17] G. M. MINTY, *Monotone nonlinear operators in Hilbert space*, Duke Math. J., 29 (1962), pp. 341–346.

[18] R. E. A. C. PALEY AND N. WIENER, *Fourier Transform in the Complex Domain*, Colloquium Publications, vol. 19, American Mathematical Society, Providence, R.I., 1934.

[19] W. V. PETRYSHYN, *On a fixed point theorem for nonlinear P-compact operators in Banach space*, Bull. Amer. Math. Soc., 72 (1966), pp. 329–333.

[20] W. A. PORTER, *Modern Foundations of Systems Engineering*, Macmillan, New York, 1965.

[21] ———, *On the reduction of sensitivity in multivariate systems*, Internat. J. Control, 5 (1967), pp. 1–9.

[22] W. A. PORTER AND C. L. ZAHM, *Basic concepts in systems theory*, Tech. Rep. 44, Systems Engineering Laboratory, University of Michigan, Ann Arbor, 1969.

[23] W. A. PORTER, *Sensitivity problems in linear systems*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 20–24.

[24] R. SAEKS, *State in Hilbert space*, SIAM Rev., 15 (1973), pp. 283–308.

[25] ———, *Causality in Hilbert space*, Ibid., 12 (1970), pp. 357–383.

[26] ———, *Resolution space–A function analytic setting for control theory*, 9th Allerton Conference on Circuit and Systems Theory, 1971.

[27] I. W. SANDBERG, *Conditions for the causality of nonlinear operators defined on a function space*, Quart. Appl. Math., 23 (1965), pp. 87–91.

[28] M. SHINBROT, *A fixed point theorem and some applications*, Arch. Rational Mech. Anal., 17 (1964), pp. 255–271.

[29] D. C. YOULA, L. J. CASTRIOTA AND H. L. CARLIN, *Bounded real scattering matrices and the foundations of linear passive network theory*, IRE Trans. Circuit Theory, CT-6 (1959), pp. 102–124.

[30] J. C. WILLEMS, *Stability, instability, invertibility and causality*, this Journal, 7 (1969), pp. 645–671.

[31] ———, *The Analysis of Feedback Systems*, MIT Press, Cambridge, 1971.

[32] L. A. ZADEH, *Optimum nonlinear filters*, J. Appl. Phys., 24 (1953), pp. 396–404.

[33] C. L. ZAHM, *Structure of sensitivity reduction*, Tech. Rep. 33, Systems Engineering Laboratory, University of Michigan, Ann Arbor, 1968.

[34] G. ZAMES, *Realizability conditions for nonlinear feedback systems*, IEEE Trans. Circuit Theory CT-11 (1964), pp. 186–194.

[35] ———, *On the input-output stability of time-varying nonlinear feedback systems*, IEEE Trans. Automatic Control AC-11 (1966), pp. 228–238, pp. 466–476.

SIAM J. CONTROL
Vol. 12, No. 3, August 1974

STABILITY ANALYSIS OF INTERCONNECTED SYSTEMS*

ANTHONY N. MICHEL†

Abstract. Sufficient conditions for the Lyapunov stability of several classes of (a) continuous-time composite systems described by ordinary differential equations, (b) discrete-time composite systems described by difference equations, (c) sampled-data composite systems, and (d) composite systems described by functional differential equations are established. In all cases the objective is the same: to analyze the stability of large-scale composite systems in terms of their lower order subsystems and in terms of their interconnecting structure.

In order to demonstrate the usefulness of the present approach, several specific examples are considered.

1. Introduction.

Despite its elegance and generality, the usefulness of the direct method of Lyapunov is severely limited when applied to problems of high dimension. For this reason it may be advantageous to view high order systems as being composed of several lower order subsystems, which, when interconnected in an appropriate fashion, yield the original composite or interconnected system. The stability analysis of such systems can then often be accomplished in terms of the simpler subsystems and in terms of the interconnecting structure of such composite systems. In this way, complications which usually arise when the direct method is applied to high order systems may be circumvented. Utilizing vector Lyapunov functions, Bailey [1] and others adopted this point of view. An excellent survey of work done in this area is contained in the survey paper by Šiljak [2].

In the present paper some new stability results for several classes of interconnected systems are established. Systems considered include (a) continuous-time systems described by ordinary differential equations, (b) discrete-time systems described by ordinary difference equations, (c) sampled-data systems, and (d) systems described by functional differential equations. The results obtained for (d) are generalizations for most of those established for (a)–(c), and, as such, constitute a unifying setting for a large class of problems.

In order to demonstrate the usefulness of the present approach, several specific examples are considered.

2. Notation and preliminaries.

Let $V \times W$ denote the Cartesian product of arbitrary sets V and W. Let $\in$ denote set membership and let $\subset$ denote set inclusion. Let R^n denote Euclidean n-space and let $|\cdot|$ denote the Euclidean norm. Let $x' = (x_1, \cdots, x_n)$ denote the transpose of $x \in R^n$. If $x, y \in R^n$, then $x \leqq y$ denotes $x_i \leqq y_i$, $i = 1, \cdots, n$.

* Received by the editors January 29, 1973, and in revised form July 16, 1973.

† Electrical Engineering Department, Iowa State University, Ames, Iowa 50010. This work was supported in part by a Faculty Improvement Leave at the Technische Hochschule, Mathematisches Institut II, Graz, Austria granted by Iowa State University and by the National Science Foundation under Grant GK-33739.

For a real rectangular matrix $A = ((a_{ij}))$, $A > 0$ indicates that $a_{ij} > 0$ for each pair (i, j), $A = 0$ denotes the null matrix, and $A \neq 0$ indicates that for some pair (i, j), $a_{ij} \neq 0$.

The eigenvalues of a square matrix A are denoted by $\lambda(A)$. If all eigenvalues of A are real, then the largest and smallest eigenvalues are denoted by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. A square matrix A is said to be stable if the real parts of all its eigenvalues are negative.

The norm of a rectangular matrix D induced by the Euclidean norm is given as $\|D\| = \min \{\alpha : \alpha|x| \geqq |Dx|, x \in R^n\} = \sqrt{\lambda_{\max}(D'D)}$, where D' denotes the transpose of D.

Let $J = [t_0, \infty)$, $t_0 \geqq 0$, and let I denote the sequence $\{t_0 + k\}$, $k = 0, 1, 2, \cdots$.

A real-valued function $\varphi(r)$ is said to belong to *class K* if it is defined, continuous, and strictly increasing over $0 \leqq r < \infty$, and if it vanishes at $r = 0$.

Systems are considered which may be described by ordinary differential equations of the form

$$\dot{x} = g(x, t), \tag{1}$$

where $\dot{x} = dx/dt$ and where $g : R^n \times J \to R^n$.

A function $g : R^n \times J \to R^n$ is said to belong to *class E* if (a) for every $x_0 \in R^n$ and for every $t_0 \geqq 0$, equation (1) possesses one and only one solution $x(t; x_0, t_0)$ for all $t \in J$, where $x_0 = x(t_0; x_0, t_0)$, and (b) $f(x, t) = 0$ for all $t \in J$ if and only if $x = 0$.

Results which yield conditions for the Lyapunov stability of the equilibrium $x = 0$ of (1) involve the existence of mappings $V : R^n \times J \to R^1$. Henceforth it is assumed that such mappings are continuous on $R^n \times J$ and that they satisfy locally a Lipschitz condition with respect to x. The upper right-hand derivative of V with respect to time along solutions of (1) is given as

$$DV_{(1)} = \limsup_{\Delta t \to 0^+} \left(\frac{1}{\Delta t}\right)\{V[x(t + \Delta t; x_0, t_0), t + \Delta t] - V[x(t; x_0, t_0), t]\}.$$

If $V \in C^1$, i.e., if V is continuously differentiable on $R^n \times J$, then the total derivative of V with respect to time along solutions of (1) is given as

$$DV_{(1)} = \nabla V(x, t)' g(x, t) + \frac{\partial V(x, t)}{\partial t},$$

where $\nabla V(x, t)$ denotes the gradient of V.

Also considered are discrete-time systems described by difference equations of the form

$$x(\tau + 1) = g[x(\tau), \tau], \tag{2}$$

where $g : R^n \times I \to R^n$. If for every $x_0 \in R^n$ and for every $t_0 \geqq 0$, equation (2) possesses a unique solution $x(\tau; x_0, t_0)$, with $x_0 = x(t_0; x_0, t_0)$, which is defined for all $\tau \in I$ and if $g(x, \tau) = 0$ for all $\tau \in I$ if and only if $x = 0$, then g is said to belong to *class E*.

Conditions for the Lyapunov stability of the equilibrium $x = 0$ of (2) involve mappings $V:R^n \times I \to R^1$ and the first difference $\Delta V(x, \tau)$ along solutions of (2) expressed by

$$\Delta V_{(2)}(x, \tau) = V[g(x, \tau), \tau + 1] - V(x, \tau).$$

Also considered are systems which are most appropriately described by functional differential equations. The following additional notation is required.

For given $h > 0$, let C denote the space of all continuous mappings of the interval $[-h, 0]$ into R^n. With $\varphi \in C$, define the norm on C by $\|\varphi\| = \sup |\varphi(\theta)|$, $-h \leqq \theta < 0$. Let C_H denote the set of all $\varphi \in C$ such that $\|\varphi\| < H$. If $x(u)$ denotes a continuous n-vector defined on $-h \leqq u$, let $x_t(\cdot)$ denote the restriction of $x(u)$ to the interval $[t - h, t]$, i.e., $x_t(\cdot) \in C$ is defined by $x_t(\cdot) = x(t + \theta)$, $-h \leqq \theta \leqq 0$. In this case t is to be considered as a parameter. Let $\dot{x}(t)$ denote the right-hand derivative of $x(u)$ at $t = u$. Finally, let $f(\varphi, t)$ be an n-vector-valued function defined for all $\varphi \in C_H$ and for all $0 \leqq t < \infty$.

Functional differential equations of the form

$$\dot{x}(t) = f(x_t, t) \tag{3}$$

are considered. A function $x(t; \varphi, t_0)$ is said to be a solution of (3) with initial vector $\varphi \in C_H$ and initial time $t_0 \geqq 0$ if (a) for each $t > t_0$, $x_t(\cdot\,; \varphi, t_0) \in C_H$, (b) $x_{t_0}(\cdot\,; \varphi, t_0) = \varphi(t_0 + \theta) \triangleq \varphi$, $-h \leqq \theta \leqq 0$, and (c) equation (3) is valid for all $t \geqq t_0$, i.e.,

$$\lim_{\Delta t \to 0^+} \frac{\{x_{t+\Delta t}(\cdot\,; \varphi, t_0) - x_t(\cdot\,; \varphi, t_0)\}}{\Delta t} = f[x_t(\cdot\,; \varphi, t_0), t].$$

The function f is said to belong to *class E* if for every $\varphi \in C_H$ and for every $t_0 \geqq 0$, equation (3) possesses a unique solution $x_t(\cdot\,; \varphi, t_0)$ for all $t \geqq t_0$, with $x_{t_0}(\cdot\,; \varphi, t_0) = \varphi$, and if in addition, $f(x_t, t) = 0$ for all $t \geqq t_0$ if and only if $x_t = 0$.

Conditions for the Lyapunov stability of the equilibrium $x = 0$ of (3) involve the existence of functionals $V(\varphi, t)$ defined for all $\varphi \in C_H$ and for all $t \geqq t_0$. Henceforth it is assumed that $V(\varphi, t)$ is continuous for all $\varphi \in C_H$, $t \geqq t_0$, and that it satisfies a local Lipschitz condition in φ. Along solutions of (3), the upper right-hand derivative of V with respect to t is defined as

$$DV_{(3)} = \limsup_{\Delta t \to 0^+} \left(\frac{1}{\Delta t}\right)\{V[x_{t+\Delta t}(\cdot\,; \varphi, t_0), t + \Delta t] - V[x_t(\cdot\,; \varphi, t_0), t]\}.$$

In the remainder of this paper it is assumed that all systems considered belong to class E. The usual definitions and theorems for Lyapunov stability of systems (1), (2) and (3) can be found in any standard text (e.g., [3]).

3. Composite systems considered. Systems are considered which may be represented by the set of *ordinary differential equations*

$$\dot{z}_i = f_i(z_i, t) + g_i(z_1, \cdots, z_m, t), \qquad i = 1, \cdots, m, \tag{4}$$

where $z_i \in R^{n_i}$, $f_i:R^{n_i} \times J \to R^{n_i}$, and $g_i:R^{n_1} \times \cdots \times R^{n_m} \times J \to R^{n_i}$. Letting $\sum_{j=1}^m n_j = n$, letting $x' = (z'_1, \cdots, z'_m) \in R^n$, letting $f(x, t) = [f_1(z_1, t)', \cdots, f_m(z_m, t)']$, and letting $g(x,t)' = [g_1(z_1, \cdots, z_m, t)', \cdots, g_m(z_1, \cdots, z_m, t)'] = [g_1(x,t)', \cdots, g_m(x,t)']$, one can represent (4) equivalently as

$$\dot{x} = f(x,t) + g(x,t) \triangleq h(x,t), \tag{5}$$

where $f: R^n \times J \to R^n$ and where $g: R^n \times J \to R^n$. System (5), which is of the form (1), is called a *composite system*. It may be viewed as a nonlinear, time-varying interconnection of m *isolated subsystems* S_i which are of the form

$$\dot{z}_i = f_i(z_i, t). \tag{6}$$

The unique solutions of (6) are denoted by $z_i(t; z_{i_0}, t_0)$ with $z_{i_0} = z_i(t_0; z_{i_0}, t_0)$.

Subsequently, several special cases of (4) are considered. In particular, let

$$g_i(x,t) = \sum_{\substack{j=1\\ i\neq j}}^{m} C_{ij} z_j,$$

where C_{ij} is a constant $n_i \times n_j$ matrix. Then (4) assumes the form

$$\dot{z}_i = f_i(z_i, t) + \sum_{\substack{j=1\\ i\neq j}}^{m} C_{ij} z_j, \qquad i = 1, \cdots, m. \tag{7}$$

Equation (7) represents a system consisting of m isolated subsystems (6) which are linearly interconnected.

Now let

$$g_i(x,t) = \sum_{\substack{j=1\\ i\neq j}}^{m} C_{ij}(z_1, \cdots, z_m, t) z_j = \sum_{\substack{j=1\\ i\neq j}}^{m} C_{ij}(x,t) z_j,$$

where $C_{ij}(x,t)$ is an $n_i \times n_j$ array. Then (4) assumes the form

$$\dot{z}_i = f_i(z_i, t) + \sum_{\substack{j=1\\ i\neq j}}^{m} C_{ij}(x,t) z_j, \qquad i = 1, \cdots, m. \tag{8}$$

Next, let

$$g_i(x,t) = \sum_{\substack{j=1\\ i\neq j}}^{m} A_{ij} f_j(z_j, t),$$

where A_{ij} is a constant $n_i \times n_j$ matrix and where f_j is determined by the jth isolated subsystem. Then (4) assumes the form

$$\dot{z}_i = \sum_{j=1}^{m} A_{ij} f_j(z_j, t), \qquad i = 1, \cdots, m, \tag{9}$$

where $A_{ii} = I$, $i = 1, \cdots, m$.

Lastly, let

$$g_i(x,t) = \sum_{\substack{j=1\\ i\neq j}}^{m} g_{ij}(z_j, t),$$

where $g_{ij}:R^{n_j} \times J \to R^{n_i}$. Then (4) assumes the form

$$\dot{z}_i = f_i(z_i, t) = \sum_{\substack{j=1 \\ i \neq j}}^{m} g_{ij}(z_j, t), \qquad i = 1, \cdots, m. \tag{10}$$

Also considered are *discrete-time composite systems* described by the set of difference equations

$$z_i(\tau + 1) = f_i[z_i(\tau), \tau] + g_i[z_1(\tau), \cdots, z_m(\tau), \tau], \qquad i = 1, \cdots, m, \tag{11}$$

where $z_i \in R^{n_i}$, $f_i:R^{n_i} \times I \to R^{n_i}$, and $g_i:R^{n_1} \times \cdots \times R^{n_m} \to R^{n_i}$. Defining f, g and h as in (5), equations (11) can equivalently be represented as

$$x(\tau + 1) = f[x(\tau), \tau] + g[x(\tau), \tau] \triangleq h[x(\tau), \tau], \tag{12}$$

where $f:R^n \times I \to R^n$ and $g:R^n \times I \to R^n$. Clearly, (12) is of the same form as (2). Composite system (12) with decomposition (11) can be viewed as a nonlinear, time-varying interconnection of m isolated subsystems which are of the form

$$z_i(\tau + 1) = f_i[z_i(\tau), \tau]. \tag{13}$$

The unique solutions of (13) are denoted by $z_i(\tau; z_{i_0}, t_0)$ with $z_{i_0} = z_i(t_0; z_{i_0}, t_0)$.

Let

$$g_i[x(\tau), \tau] = \sum_{\substack{j=1 \\ i \neq j}}^{m} C_{ij} z_j,$$

where C_{ij} is as defined in (7). Then (11) assumes the form

$$z_i(\tau + 1) = f_i[z_i(\tau), \tau] + \sum_{\substack{j=1 \\ i \neq j}}^{m} C_{ij} z_j(\tau), \qquad i = 1, \cdots, m. \tag{14}$$

Next, let

$$g_i[x(\tau), \tau] = \sum_{\substack{j=1 \\ i \neq j}}^{m} A_{ij} f_j[z_j(\tau), \tau],$$

where A_{ij} is as defined in (9) and where f_j is determined by the jth isolated subsystem. Then (11) assumes the form

$$z_i(\tau + 1) = \sum_{j=1}^{m} A_{ij} f_j[z_j(\tau), \tau], \qquad i = 1, \cdots, m, \tag{15}$$

where $A_{ii} = I$, $i = 1, \cdots, m$.

Finally, let

$$g_i[x(\tau), \tau] = \sum_{\substack{j=1 \\ i \neq j}}^{m} g_{ij}[z_j(\tau), \tau],$$

where g_{ij} is defined similarly as in (10). Then (11) assumes the form

$$z_i(\tau + 1) = f_i[z_i(\tau), \tau] + \sum_{\substack{j=1 \\ i \neq j}}^{m} g_{ij}[z_j(\tau), \tau], \qquad i = 1, \cdots, m. \tag{16}$$

Also considered are *composite sampled-data systems* described by the set of equations

$$\dot{z}_i(t) = f_i[z_i(t), t] + \sum_{\substack{j=1\\ i\neq j}}^{m} C_{ij}z_j(kT), \tag{17}$$
$$kT \leqq t < (k+1)T, \quad k = 0, 1, 2, \cdots, \quad i = 1, \cdots, m,$$

where $\dot{z}_i(t)$ denotes the right-hand derivative of $z(t)$ with respect to t, $z_i \in R^{n_i}$, $f_i : R^{n_i} \times [0, \infty) \to R^{n_i}$, C_{ij} is a constant $n_i \times n_j$ matrix, and T denotes the *sampling time.*

Composite system (17) may be viewed as a linear interconnection of m subsystems described by (6).

Another class of composite sampled-data systems considered are those described by the set of equations

$$\dot{z}_i(t) = f_i[z_i(t), t] + \sum_{\substack{j=1\\ i\neq j}}^{m} \delta(t)C_{ij}z_j(kT), \tag{18}$$
$$kT \leqq t < (k+1)T, \quad k = 0, 1, 2, \cdots, \quad i = 1, \cdots, m,$$

where $\dot{z}_i(t)$, $z_i(t)$, f_i and C_{ij} are defined as in (17) and where

$$\delta(t) = \begin{cases} 1 & \text{if } t = kT, k = 0, 1, 2, \cdots, \\ 0 & \text{if } t \neq kT. \end{cases}$$

Finally, consider *composite systems* described by the set of *functional differential equations*

$$z^i(t) = f_i(z_t^i, t) + \sum_{\substack{j=1\\ i\neq j}}^{m} g_{ij}(z_t^j, t), \qquad i = 1, \cdots, m, \tag{19}$$

where z^i is an n_i-dimensional vector, $f_i(\varphi^i, t)$ is an n_i-vector-valued function defined for all $\varphi^i \in C_{H_i}$ and for all $t \geqq t_0$, and where $g_{ij}(\varphi^j, t)$ is an n_i-vector-valued function defined for all $\varphi^j \in C_{H_j}$ and for all $t \geqq t_0$.

Composite system (19) may be viewed as a nonlinear, time-varying interconnection of m isolated subsystems described by functional differential equations of the form

$$\dot{z}^i(t) = f_i(z_t^i, t). \tag{20}$$

Now let $\sum_{j=1}^m n_j = n$, and let

$$x_t = \begin{bmatrix} z_t^1 \\ \vdots \\ z_t^m \end{bmatrix}, \quad \varphi = \begin{bmatrix} \varphi^1 \\ \vdots \\ \varphi^m \end{bmatrix}, \quad f(x_t, t) = \begin{bmatrix} f_1(z_t^1, t) \\ \vdots \\ f_m(z_t^m, t) \end{bmatrix}, \quad g(x_t, t) = \begin{bmatrix} \sum\limits_{\substack{j=1\\ i\neq j}}^{m} g_{1j}(z_t^j, t) \\ \vdots \\ \sum\limits_{\substack{j=1\\ i\neq j}}^{m} g_{mj}(z_t^j, t) \end{bmatrix}.$$

Then (19) can be written equivalently as

$$(21) \qquad \dot{x}(t) = f(x_t, t) + g(x_t, t) \triangleq h(x_t, t),$$

where x is an n-dimensional vector and where f and g are n-vector-valued functions. Equation (21) is clearly of the same form as (3).

The unique solutions of isolated subsystem (20) are denoted by $z_t^i(\cdot\,; \varphi^i, t_0)$. The initial function vector φ of the solution of (21) is defined over the domain $\mathscr{D} \times E_{t_0}$, where E_{t_0} is the set of values $t - h_i$, $t \geqq t_0$, $i = 1, \cdots, m$, which are smaller or equal to t_0, and where $\mathscr{D} = \{\varphi : \varphi^i \in C_{H_i}, i = 1, \cdots, m\} \subset C_H$.

Remark 1. Systems (7), (8), (9), (10), (14), (15), (16), (17) and (18) are all special cases of system (19).

4. Main results: Continuous-time systems. Subsequently, the following conventions will be employed.

Definition 1. Isolated subsystem (6) possesses *Property* A if there exist a continuously differentiable function $V_i : R^{n_i} \times J \to R^1$, two radially unbounded functions $\varphi_{i1}, \varphi_{i2} \in K$, and a function $\varphi_{i3} \in K$, such that the conditions

(i) $\varphi_{i1}(|z_i|) \leqq V_i(z_i, t) \leqq \varphi_{i2}(|z_i|)$,

(ii) $DV_{i(6)} \leqq -\varphi_{i3}(|z_i|)$

hold for all $z_i \in R^{n_i}$ and for all $t \in J$.

Definition 2. Isolated subsystem (6) possesses *Property* B if there exist a continuously differentiable function $V_i : R^{n_i} \times J \to R^1$ and four positive constants $c_{i1}, c_{i2}, c_{i3}, c_{i4}$, such that the conditions

(i) $c_{i1}|z_i|^2 \leqq V_i(z_i, t) \leqq c_{i2}|z_i|^2$,

(ii) $DV_{i(6)} \leqq -c_{i3}|z_i|^2$,

(iii) $|\nabla V_i(z_i, t)| \leqq c_{i4}|z_i|$

hold for all $z_i \in R^{n_i}$ and for all $t \in J$.

Remark 2. If subsystem (6) possesses Property A, then its equilibrium $z_i = 0$ is asymptotically stable in the large. If it possesses Property B, then its equilibrium is exponentially stable in the large (see, e.g., [3]).

The proofs of the following results are given in §9.

Theorem 1. *The equilibrium $x = 0$ of composite system* (4) *is asymptotically stable in the large if the following conditions are satisfied*:

(i) *each isolated subsystem* (6) *possesses Property* A;

(ii) *for each scalar product $\nabla V_i(z_i, t)' g_i(x, t)$, $i = 1, \cdots, m$, an inequality of the form*

$$\nabla V_i(z_i, t)' g_i(x, t) \leqq [\varphi_{i3}(|z_i|)]^{1/2} \sum_{j=1}^{m} a_{ij}(x, t)[\varphi_{j3}(|z_j|)]^{1/2}$$

can be found; and

(iii) *there exist an m-vector $\alpha' = (\alpha_1, \cdots, \alpha_m)$, $\alpha_i > 0$, $i = 1, \cdots, m$, and $\varepsilon > 0$ such that for each $x \in R^n$ and each $t \in J$, the array $(S + \varepsilon I)$ is negative definite, where $S = ((s_{ij}))$ is defined by*

$$s_{ij} = \begin{cases} -\alpha_i + \alpha_i a_{ii}(x, t) & \text{if } i = j, \\ (\alpha_i a_{ij}(x, t) + \alpha_j a_{ji}(x, t))/2 & \text{if } i \neq j, \end{cases}$$

and where I denotes the identity matrix.

COROLLARY 1. *The equilibrium $x = 0$ of composite system* (8) *is exponentially stable in the large if the following conditions are satisfied*:

(i) *each isolated subsystem* (6) *possesses Property* B;

(ii) *there exists an m-vector* $\alpha' = (\alpha_1, \cdots, \alpha_m)$, $\alpha_i > 0$, $i = 1, \cdots, m$, *and* $\varepsilon > 0$ *such that for each* $x \in R^n$ *and for all* $t \in J$, *the array* $(S + \varepsilon I)$ *is negative definite, where* $S = ((s_{ij}))$ *is defined by*

$$s_{ij} = \begin{cases} -\alpha_i c_{i3} + \alpha_i c_{i4}\|C_{ij}(x, t)\| & \text{if } i = j, \\ (\alpha_i c_{i4}\|C_{ij}(x, t)\| + \alpha_j c_{j4}\|C_{ji}(x, t)\|)/2 & \text{if } i \neq j. \end{cases}$$

Here $\|C_{ij}(x, t)\| = \{\lambda_{\max}[C_{ij}(x, t)'C_{ij}(x, t)]\}^{1/2}$.

THEOREM 2. *The equilibrium $x = 0$ of composite system* (9) *is asymptotically stable in the large if the following conditions are satisfied*:

(i) *each isolated subsystem* (6) *possesses Property* A;

(ii) *for every* $z_i \in R^{n_i}$, $z_i \neq 0$, *there exists a* $\delta > 0$ *such that* $|f_i(z_i, t)| \geqq \delta$ *for all* $t \in J$;

(iii) *for each isolated subsystem* (6) *there exist positive constants* c_{i1} *and* c_{i2} *such that the conditions*

(a) $DV_{i(6)} \leqq -c_{i1}|f_i(z_i, t)|^2$,

(b) $|\nabla V_i(z_i, t)| \leqq c_{i2}|f_i(z_i, t)|$

hold for all $z_i \in R^{n_i}$ *and for all* $t \in J$; *and*

(iv) *there exists an m-vector* $\alpha' = (\alpha_1, \cdots, \alpha_m)$, $\alpha_i > 0$, $i = 1, \cdots, m$, *such that the matrix* $S = ((s_{ij}))$ *defined by*

$$s_{ij} = \begin{cases} -\alpha_i c_{i1} & \text{if } i = j, \\ (\alpha_i c_{i2}\|A_{ij}\| + \alpha_j c_{j2}\|A_{ji}\|)/2 & \text{if } i \neq j, \end{cases}$$

is negative definite. Here,

$$\|A_{ij}\| = \sup_{f_j(z_j, t) \neq 0} \frac{|A_{ij} f_j(z_j, t)|}{|f_j(z_j, t)|}.$$

Remark 3. For autonomous systems, Theorem 2 finds applications when the isolated subsystems

$$\dot{z}_i = f_i(z_i) \tag{22}$$

satisfy the following result due to Krasovskii [3]: The equilibrium $x = 0$ of (22) is asymptotically stable in the large if the Jacobian $J_i(z_i) = \partial f_i(z_i)/\partial z_i$ is such that $\lambda_{\max}[J_i(z_i)' + J_i(z_i)] \leqq -\delta < 0$.

A Lyapunov function which satisfies Krasovskii's result is $V_i(z_i) = f_i(z_i)'f_i(z_i) = |f_i(z_i)|^2$. Now $DV_{i(22)} \leqq -\delta|f_i(z_i)|^2$ and $|\nabla V_i(z_i)| \leqq \{2 \sup_{z_i} |J_i(z_i)|\}|f_i(z_i)|$. In cases where $\sup_{z_i} |J_i(z_i)| < \infty$, let $c_{i2} = 2 \sup_{z_i} |J_i(z_i)|$, and let $c_{i1} = \delta$. Then it follows that

$$DV_{i(22)} \leqq -c_{i1}|f_i(z_i)|^2 \quad \text{and} \quad |\nabla V_i(z_i)| \leqq c_{i2}|f_i(z_i)|,$$

which are precisely the requirements of hypothesis (iii) of Theorem 2.

THEOREM 3. *The equilibrium $x = 0$ of composite system* (10) *is exponentially stable in the large if the following conditions are satisfied*:

(i) *each isolated subsystem* (6) *possesses Property* B;

(ii) *for each* $i, j = 1, \cdots, m$, $i \neq j$, *there exists a positive constant* k_{ij} *such that*

$$|g_{ij}(z_j, t)| \leqq k_{ij}|z_j|$$

for all $z_j \in R^{n_j}$ *and for all* $t \in J$; *and*

(iii) *there exists an m-vector* $\alpha' = (\alpha_1, \cdots, \alpha_m)$, $\alpha_i > 0$, $i = 1, \cdots, m$, *such that the matrix* $S = ((s_{ij}))$ *defined by*

$$s_{ij} = \begin{cases} -\alpha_i c_{i3} & \text{if } i = j, \\ (\alpha_i c_{i4} k_{ij} + \alpha_j c_{j4} k_{ji})/2 & \text{if } i \neq j, \end{cases}$$

is negative definite.

Remark 4. Conditions for the exponential stability of composite system (7) follow immediately from the last result by replacing k_{ij} by $\|C_{ij}\|$, k_{ji} by $\|C_{ji}\|$, and by deleting hypothesis (ii).

Remark 5. Following the proof of Theorem 3, another set of conditions for the exponential stability in the large of composite system (8) can be obtained. In this case hypotheses (ii) and (iii) of Theorem 3 are changed as follows:

(ii) *for every* $x \in R^n$, $x \neq 0$, *there exists a* $\delta > 0$ *such that* $\|C_{ij}(x, t)\| \geqq \delta$ *for all* $t \in J$;

(iii) *there exists an m-vector* $\alpha' = (\alpha_1, \cdots, \alpha_m)$, $\alpha_i > 0$, $i = 1, \cdots, m$, *such that the matrix* $S = ((s_{ij}))$ *defined by*

$$s_{ij} = \begin{cases} -\alpha_i c_{i3} & \text{if } i = j, \\ (\alpha_i c_{i4}\|C_{ij}(x, t)\| + \alpha_j c_{j4}\|C_{ji}(x, t)\|)/2 & \text{if } i \neq j, \end{cases}$$

is negative definite for all $x \in R^n$ *and for all* $t \in J$. *Here* $\|C_{ij}(x, t)\|$ *is as defined in Corollary* 1.

THEOREM 4. *The equilibrium* $x = 0$ *of composite system* (10) *is asymptotically stable in the large if the following conditions are satisfied*:

(i) *for each isolated subsystem* (6) *there exist a continuous function* $V_i: R^{n_i} \times J \to R^1$, *two radially unbounded functions* $\varphi_{i1}, \varphi_{i2} \in K$, *a function* $\varphi_{i3} \in K$, *a continuous function* $c_i(t) \geqq \delta > 0$ *for all* $t \in J$, *and a positive constant* L_i, *such that the conditions*

(a) $\varphi_{i1}(|z_i|) \leqq V_i(z_i, t) \leqq \varphi_{i2}(|z_i|)$,

(b) $DV_{i(6)} \leqq -c_i(t)\varphi_{i3}(|z_i|)$,

(c) $|V_i(z_i', t) - V_i(z_i'', t)| \leqq L_i|z_i' - z_i''|$

hold for all $z_i, z_i', z_i'' \in R^{n_i}$ *and for all* $t \in J$;

(ii) *for each* $i, j = 1, \cdots, m$, $i \neq j$, *there exists a continuous function* $k_{ij}(t) \geqq \varepsilon > 0$ *for all* $t \in J$, *such that the condition*

$$|g_{ij}(z_j, t)| \leqq k_{ij}(t)\varphi_{j3}(|z_j|)$$

is satisfied for all $z_j \in R^{n_j}$ *and for all* $t \in J$; *and*

(iii) *all successive principal minors of the matrix* $S(t) = ((s_{ij}(t)))$ *defined by*

$$s_{ij}(t) = \begin{cases} c_i(t) & \text{if } i = j, \\ -L_i k_{ij}(t) & \text{if } i \neq j, \end{cases}$$

are bounded from below by some $\gamma > 0$ *for all* $t \in J$.

Remark 6. It is strongly emphasized that in Theorem 4, V_i need not be continuously differentiable.

Remark 7. If the conditions of Theorem 4 are relaxed to allow V_i to be continuously differentiable, then subsystem (6) possesses Property A and it is possible to prove the following version of Theorem 4: (a) delete hypothesis (ic); and (b) replace L_i in hypothesis (iii) by c_{i4}, where $|\nabla V_i(z_i, t)| \leqq c_{i4}|z_i|$ for all $z_i \in R^{n_i}$ and for all $t \in J$.

5. Main results: Discrete-time systems. In the next results the following notation is employed.

DEFINITION 3. Isolated subsystem (13) possesses *Property* C if there exist a function $V_i : R^{n_i} \times I \to R^1$, two radially unbounded functions $\varphi_{i1}, \varphi_{i2} \in K$, a function $\varphi_{i3} \in K$, and two positive constants c_i, L_i, such that the conditions

(i) $\varphi_{i1}(|z_i|) \leqq V_i(z_i, \tau) \leqq \varphi_{i2}(|z_i|)$,

(ii) $\Delta V_{i(13)} \leqq -c_i \varphi_{i3}(|z_i|)$,

(iii) $|V_i(z_i', \tau) - V_i(z_i'', \tau)| \leqq L_i|z_i' - z_i''|$

hold for all z_i, z_i', $z_i'' \in R^{n_i}$ *and for all* $\tau \in I$.

Remark 8. If subsystem (13) possesses Property C, then its equilibrium $z_i = 0$ is asymptotically stable in the large (see, e.g., [3]).

The proofs of the subsequent results are given in §9.

THEOREM 5. *The equilibrium* $x = 0$ *of composite system* (16) *is asymptotically stable in the large if the following conditions are satisfied*:

(i) *each isolated subsystem* (13) *possesses Property* C, *with* $\varphi_{i3}(|z_i|) = |z_i|$;

(ii) *for each* $i, j = 1, \cdots, m$, $i \neq j$, *there exists a positive constant* k_{ij} *such that*

$$|g_{ij}(z_j, \tau)| \leqq k_{ij}|z_j|$$

for all $z_j \in R^{n_j}$ *and for all* $\tau \in I$;

(iii) *the matrix* $S = ((s_{ij}))$ *defined by*

$$s_{ij} = \begin{cases} c_i & \text{if } i = j, \\ -L_i k_{ij} & \text{if } i \neq j, \end{cases}$$

has positive successive principal minors.

Remark 9. Conditions for the asymptotic stability in the large of composite system (14) follow from the last result by replacing k_{ij} by $\|C_{ij}\|$, k_{ij} by $\|C_{ji}\|$, and by deleting hypothesis (ii).

THEOREM 6. *The equilibrium* $x = 0$ *of composite system* (15) *is asymptotically stable in the large if the following conditions are satisfied*:

(i) *each isolated subsystem* (13) *possesses Property* C, *with* $\varphi_{i3}(|z_i|) \geqq |f_i(z_i, \tau)|$ *for all* $\tau \in I$ (*i.e.*, $\Delta V_{i(13)} \leqq -c_i|f_i(z_i, \tau)|$, $z_i \in R^{n_i}$, $\tau \in I$);

(ii) *for every* $z_i \in R^{n_i}$, $z_i \neq 0$, *there exists a* $\delta > 0$ *such that* $|f_i(z_i, \tau)| \geqq \delta$ *for all* $\tau \in I$; *and*

(iii) *the matrix* $S = ((s_{ij}))$ *defined by*

$$s_{ij} = \begin{cases} c_i & \text{if } i = j, \\ -L_i\|A_{ij}\| & \text{if } i \neq j, \end{cases}$$

has positive successive principal minors. Here

$$\|A_{ij}\| = \sup_{f_j(z_j, \tau) \neq 0} \frac{|A_{ij} f_j(z_j, \tau)|}{|f_j(z_j, \tau)|}.$$

THEOREM 7. *The equilibrium $x = 0$ of composite system* (11) *is asymptotically stable in the large if the following conditions are satisfied*:

(i) *each isolated subsystem* (13) *possesses Property* C;

(ii) *for each $i, j = 1, \cdots, m$, there exists a function $a_{ij}(x, \tau)$ such that the properties*

(a) $a_{ij}(x, \tau) \geqq \delta > 0, i, j = 1, \cdots, m,$

(b) $c_i - L_i a_{ii}(x, \tau) > 0, i = 1, \cdots, m,$

hold for all $x \in R^n$ and for all $\tau \in I$ (c_i and L_i are given in Definition 3);

(iii) *for each $i = 1, \cdots, m$, an inequality of the form*

$$|g_i(z_1, \cdots, z_m, \tau)| \leqq \sum_{j=1}^{m} a_{ij}(x, \tau)\varphi_{j3}(|z_j|)$$

holds for all $x \in R^n$ and for all $\tau \in I$; and

(iv) *for every $x \in R^n$, each successive principal minor of $S = ((s_{ij}))$ determined by*

$$s_{ij} = \begin{cases} c_i - L_i a_{ii}(x, \tau) & \text{if } i = j, \\ -L_i a_{ij}(x, \tau) & \text{if } i \neq j, \end{cases}$$

is bounded from below by some $\delta > 0$ for all $\tau \in I$.

6. Main results: Sampled-data systems. Presently, the following convention is utilized.

DEFINITION 4. Isolated subsystem (6) possesses *Property* D if there exist a continuous function $V_i : R^{n_i} \times [0, \infty) \to R^1$ and three positive constants c_{i1}, c_{i2} and L_i, such that the conditions

(i) $|z_i| \leqq V_i(z_i, t) \leqq c_{i1}|z_i|$,

(ii) $DV_{i(6)}^{(-)} = \liminf\limits_{\Delta t \to 0^+} \dfrac{\{V_i[z_i + \Delta t \cdot f_i(z_i, t), t + \Delta t] - V_i(z_i, t)\}}{\Delta t} \geqq -c_{i2}|z_i|$,

(iii) $|V_i(z_i', t) - V_i(z_i'', t)| \leqq L_i|z_i' - z_i''|$

hold for all $z_i, z_i', z_i'' \in R^{n_i}$ and for all $t \in [0, \infty)$.

The next results are proved in § 9.

THEOREM 8. *The equilibrium $x = 0$ of composite system* (17) *is asymptotically stable in the large if the following conditions are satisfied*:

(i) *each isolated subsystem* (6) *possesses Property* D;

(ii) *for each isolated subsystem* (6) *there exists a positive constant c_{i3} such that $DV_{i(6)} \leqq -c_{i3}|z_i|$ for all $z_i \in R^{n_i}$ and for all $t \in [0, \infty)$*;

(iii) *the matrix $S = ((s_{ij}))$ with components*

$$s_{ij} = \begin{cases} \dfrac{c_{i3}}{c_{i1}} e^{-c_{i2}T} & \text{if } i = j, \\ -\left\{\dfrac{c_{i3}L_i}{c_{i1}c_{i2}}[1 - e^{-c_{i2}T}] + L_i\right\}\|C_{ij}\| & \text{if } i \neq j, \end{cases}$$

has positive successive principal minors.

THEOREM 9. *The equilibrium $x = 0$ of composite system* (18) *is asymptotically stable in the large if the following conditions are satisfied*:

(i) *for each isolated subsystem* (6), *hypotheses* (i) *and* (ii) *of Theorem* 8 *hold*;

(ii) *the matrix* $S = ((s_{ij}))$ *with components*

$$s_{ij} = \begin{cases} c_{i3} & \text{if } i = j, \\ -L_i\|C_{ij}\| & \text{if } i \neq j, \end{cases}$$

has positive successive principal minors.

7. Main results: Systems described by functional differential equations. It will be convenient to use the following additional nomenclature.

DEFINITION 5. Isolated subsystem (20) possesses *Property* E if there exist a continuous functional $V_i(\varphi^i, t)$ defined for every $\varphi^i \in C_{H_i}$ and for every $t \in J$, three functions $\psi_{i1}, \psi_{i2}, \psi_{i3} \in K$, and two positive constants c_i, L_i, such that the conditions

(i) $\psi_{i1}(\|\varphi^i\|) \leqq V_i(\varphi^i, t) \leqq \psi_{i2}(\|\varphi^i\|)$,
(ii) $DV_{i(20)} \leqq -c_i\psi_{i3}(\|\varphi^i\|)$, and
(iii) $|V_i(\varphi^i, t) - V_i(\eta^i, t)| \leqq L_i\|\varphi^i - \eta^i\|$

hold for all $\varphi^i, \eta^i \in C_{H_i}$ and for all $t \in J$.

DEFINITION 6. For isolated subsystem (20), let $V_i(\varphi^i, t)$ be a continuous functional satisfying (i) and (iii) of Definition 5. Let U_i denote the subset of C_{H_i} of all elements $y_i(t)$ for which the form $V_i[y_i(t), t]$, considered as a function of t, is monotone decreasing. If there exist $\psi_{i3} \in K$ and a constant $c_i > 0$ such that

$$DV_{i(20)} \leqq -c_i\psi_{i3}(\|y_i\|)$$

for all $y_i \in U_i$, then system (20) is said to possess *Property* F.

DEFINITION 7. Let

$$\|\varphi^i\|_1 = \left(\int_{h_i}^{0} \sum_{j=1}^{n_i} [\varphi_j^i(\theta)]^2 \, d\theta\right)^{1/2}.$$

Isolated subsystem (20) possesses *Property* G if the right-hand side of (20) is bounded and if there exist a continuous functional $V_i(\varphi^i, t)$ defined for every $\varphi^i \in C_{H_i}$ and for every $t \in J$, four functions $\psi_{i1}, \psi_{i2}, \psi_{i3}, \psi_{i4} \in K$, and two positive constants c_i, L_i, such that the conditions

(i) $\psi_{i1}(|\varphi^i(0)|) \leqq V_i(\varphi^i, t) \leqq \psi_{i2}(|\varphi^i(0)|) + \psi_{i3}(\|\varphi^i\|_1)$,
(ii) $DV_{i(20)} \leqq -c_i\varphi_{i4}(|\varphi^i(0)|)$, and
(iii) $|V_i(\varphi^i, t) - V_i(\eta^i, t)| \leqq L_i|\varphi^i(0) - \eta^i(0)|$

hold for all $\varphi^i, \eta^i \in C_{H_i}$ and for all $t \in J$.

Remark 10. If isolated subsystem (20) possesses any one of the above properties, then its equilibrium is asymptotically stable (see, e.g., [3]).

THEOREM 10. *The equilibrium* $x = 0$ *of composite system* (19) *is asymptotically stable if the following conditions are satisfied*:

(i) *each isolated subsystem* (20) *possesses Property* E;

(ii) *for each* $i, j = 1, \cdots, m$, $i \neq j$, *there exists a positive constant* k_{ij} *such that*

$$\|g_{ij}(\varphi^j, t)\| \leqq k_{ij}\psi_{j3}(\|\varphi^j\|)$$

for all $\varphi^j \in C_{H_j}$ *and for all* $t \in J$; *and*

(iii) *the matrix* $S = ((s_{ij}))$ *defined by*

$$s_{ij} = \begin{cases} c_i & \text{if } i = j, \\ -L_i k_{ij} & \text{if } i \neq j, \end{cases}$$

has positive successive principal minors.

THEOREM 11. *The equilibrium* $x = 0$ *of composite system* (19) *is asymptotically stable if the following conditions are satisfied*:

(i) *each isolated subsystem* (20) *possesses Property* F;

(ii) *for each* $i, j = 1, \cdots, m$, $i \neq j$, *there exists a positive constant* k_{ij} *such that*

$$\|g_{ij}(\varphi^j, t)\| \leqq k_{ij}\psi_{j3}(\|\varphi^j\|), \qquad i \neq j,$$

for all $\varphi^j \in U_j$ *and for all* $t \in J$; *and*

(iii) *the matrix* $S = ((s_{ij}))$ *defined by*

$$s_{ij} = \begin{cases} c_i & \text{if } i = j, \\ -L_i k_{ij} & \text{if } i \neq j, \end{cases}$$

has positive successive principal minors.

THEOREM 12. *The equilibrium* $x = 0$ *of composite system* (19) *is asymptotically stable if the following conditions are satisfied*:

(i) *each isolated subsystem* (20) *possesses Property* G;

(ii) *for each* $i, j = 1, \cdots, m, i \neq j$, *there exists a positive constant* k_{ij} *such that*

$$|g_{ij}(\varphi^j(0), t)| \leqq k_{ij}\psi_{j4}(|\varphi^j(0)|), \qquad i \neq j,$$

for all $\varphi^j \in C_{H_j}$, $t \in J$; *and*

(iii) *the matrix* $S = ((s_{ij}))$ *defined by*

$$s_{ij} = \begin{cases} c_i & \text{if } i = j, \\ -L_i k_{ij} & \text{if } i \neq j, \end{cases}$$

has positive successive principal minors.

8. Applications. Several specific examples are now considered to demonstrate applications of the above results.

Example 1. The controlled *longitudinal motion of an aircraft* may be represented by the set of equations [4]

$$\begin{aligned} \dot{x}_k &= -\rho_k x_k + \sigma, \qquad k = 1,2,3,4, \\ \dot{\sigma} &= \sum_{k=1}^{4} \beta_k x_k - rp_2\sigma - f(\sigma), \end{aligned} \tag{23}$$

where $\rho_k > 0$, $r > 0$, $p_2 > 0$ are constants and where $f(\sigma)$ is a real, single-valued function with the following properties: (i) $f(\sigma)$ is continuous for all $-\infty < \sigma < \infty$, (ii) $f(\sigma) = 0$ if and only if $\sigma = 0$, and (iii) $\sigma f(\sigma) > 0$ for all $\sigma \neq 0$. Without loss of generality it is assumed that $\rho_1 \leqq \rho_2 \leqq \rho_3 \leqq \rho_4$. Let $z_1' = (x_1, x_2, x_3, x_4)$ and let $z_2 = \sigma$. System (23) may be viewed as a linear interconnection of two isolated subsystems S_1, S_2,

$$S_1 : \dot{x}_k = -\rho_k x_k, \qquad k = 1,2,3,4, \tag{24}$$

$$S_2 : \dot{\sigma} = -rp_2\sigma - f(\sigma), \tag{25}$$

interconnected by means of the matrices $C'_{12} = (1, 1, 1, 1)$ and $C_{21} = (\beta_1, \beta_2, \beta_3, \beta_4)$. Composite system (23) is clearly a special case of (7).

Choosing $V_1(z_1) = c_1 z'_1 z_1$ and $V_2(z_2) = c_2 z^2$, where c_1 and c_2 are positive constants, one obtains

$$DV_{1(24)} \leqq -2c_1\rho_1|z_1|^2, \quad |\nabla V_1(z_1)| \leqq 2c_2|z_1|, \quad DV_{2(25)} \leqq -2rp_2c_2|z_2|^2$$

and $|\nabla V_2(z_2)| \leqq 2c_2|z_2|$ for all $z_1 \in R^{n_1}$ and $z_2 \in R^{n_2}$. Also, $\|C_{12}\| = 2$ and $\|C_{21}\| = (\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2)^{1/2}$.

Choosing $\alpha_1 = 1/(4c_1)$ and $\alpha_2 = 1/(2c_2\sqrt{\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2})$, matrix S of Theorem 3 (and Remark 4) assumes the form

$$S = \begin{bmatrix} -\rho_1/2 & 1 \\ 1 & \dfrac{-rp_2}{(\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2)^{1/2}} \end{bmatrix}.$$

Matrix S is negative definite provided that $-\rho_1^2 p_2^2 r^2 + 4(\sum_{i=1}^4 \beta_i^2) < 0$. It now follows from Theorem 3 that the equilibrium $x' = (z'_1 z'_2) = 0$ of composite system (23) is exponentially stable in the large provided that

$$\sum_{i=1}^{4} \xi_i^2 < 1, \quad \xi_i = \frac{2\beta_i}{\rho_1 p_2 r}. \tag{26}$$

Example 2. Consider the *indirect control problem*

$$\begin{aligned} \dot{x} &= Ax + bf(\sigma), \\ \dot{\sigma} &= -\rho_n\sigma - rf(\sigma) + a'x, \end{aligned} \tag{27}$$

where $x \in R^n$, A is a stable $n \times n$ matrix, b is an n-vector, $r > 0$ is a constant, a is an n-vector, and $f(\sigma)$ is a real, single-valued function with the following properties: (i) $f(\sigma)$ is continuous for all $-\infty < \sigma < \infty$, (ii) $f(\sigma) = 0$ if and only if $\sigma = 0$, and (iii) $0 < \sigma f(\sigma) < k\sigma^2$ for all $\sigma \neq 0$, where $k > 0$ is a constant. System (27) may be viewed as a nonlinear interconnection of two isolated subsystems S_1, S_2,

$$S_1 : \dot{x} = Ax, \tag{28}$$

$$S_2 : \dot{\sigma} = -\rho_n\sigma - rf(\sigma), \tag{29}$$

interconnected by the equations $g_{12}(\sigma) = f(\sigma)b$ and $g_{21}(x) = a'x$. Composite system (27) is clearly a special case of (10).

For S_1 there exist a function $V_1(x_1) = x'Px$, where P is a symmetric, positive definite matrix, and four positive constants c_{1i}, $i = 1, 2, 3, 4$, such that $c_{11}|x|^2 \leqq V_1(x) \leqq c_{12}|x|^2$, $DV_{1(28)} \leqq -c_{13}|x|^2$, $|\nabla V_1(x)| \leqq c_{14}|x|$ for all $x \in R^n$ (for converse stability theorems see, e.g., [3]). For S_2 choose $V_2(\sigma) = \sigma^2/2$. Then $DV_{2(29)} \leqq -\rho_n|\sigma|^2$ and $|\nabla V_2(\sigma)| = |\sigma|$ for all $\sigma \in R^1$.

The constants k_{12} and k_{21} of hypothesis (ii) of Theorem 3 are $k_{12} = k|b|$ and $k_{21} = |a|$.

Choosing $\alpha_1 = 1/(k|b|)$ and $\alpha_2 = c_{14}/|a|$, matrix S of Theorem 3 assumes the form

$$S = \begin{bmatrix} -c_{13}/(k|b|) & c_{14} \\ c_{14} & -c_{14}\rho_n/|a| \end{bmatrix}.$$

This matrix is negative definite provided that $c_{13}c_{14}\rho_n/(k|a||b|) - c_{14}^2 > 0$. Thus, the equilibrium $(x', \sigma) = (0', 0)$ of composite system (27) is exponentially stable in the large provided that

$$k < \rho_n c_{13}/(|a||b|c_{14}). \tag{30}$$

Example 3. Consider again (27). For S_1 choose $V_1(x) = |x|$ and for S_2 choose $V_2(\sigma) = |\sigma|$. Let $W = (A' + A)/2$. Then $DV_{1(28)} \leqq \lambda_{\max}(W)|x|$ and $DV_{2(29)} \leqq -\rho_n|\sigma|$. V_1 is Lipschitzian in x with $L_1 = 1$ and V_2 is Lipschitzian in σ with $L_2 = 1$. Matrix $S(t) = S$ of Theorem 4 assumes the form

$$S = \begin{bmatrix} -\lambda_{\max}(W) & -k|b| \\ -|a| & \rho_n \end{bmatrix},$$

where $-\lambda_{\max}(W) > 0$. This matrix has positive successive principal minors provided that $|\lambda_{\max}(W)|\rho_n - k|a||b| > 0$. It now follows from Theorem 4 that the equilibrium $(x', \sigma) = (0', 0)$ of composite system (27) is asymptotically stable in the large, and hence, absolutely stable, provided that

$$k < \rho_n|\lambda_{\max}(W)|/(|a||b|). \tag{31}$$

Example 4. Consider the nonlinear, time-invariant system

$$\begin{aligned} \dot{z}_1 &= A_1 z_1 + b_1 f_1(\sigma_1), & \dot{z}_2 &= A_2 z_2 + b_2 f_2(\sigma_2), \\ \sigma_1 &= c_1' z_2, & \sigma_2 &= c_2' z_1, \end{aligned} \tag{32}$$

where $z_1 \in R^{n_1}$, $z_2 \in R^{n_2}$, A_1 is a stable $n_1 \times n_1$ matrix, A_2 is a stable $n_2 \times n_2$ matrix, c_1 is an n_2-vector, c_2 is an n_1-vector, and $f_1(\sigma_1)$, $f_2(\sigma_2)$ are real, single-valued and continuous functions with the properties $f_1(0) = f_2(0) = 0$, $0 < \sigma_1 f_1(\sigma_1) < k_1\sigma^2$ for all $\sigma_1 \neq 0$, and $0 < \sigma_2 f_2(\sigma_2) < k_2\sigma^2$ for all $\sigma_2 \neq 0$.

System (32) can be viewed as a nonlinear interconnection of two linear isolated subsystems S_1, S_2,

$$S_i : \dot{z}_i = A_i z_i, \qquad i = 1, 2, \tag{33}$$

interconnected by the equations $g_{12}(z_1) = b_1 f_1(\sigma_1)$ and $g_{21}(z_1) = b_2 f_2(\sigma_2)$. System (32) is clearly a special case of (10).

Isolated subsystems S_1 and S_2 are exponentially stable in the large and there exist functions $V_i(z_i)$, $i = 1, 2$, and positive constants c_{i1}, c_{i2}, c_{i3}, c_{i4}, $i = 1, 2$, such that $c_{i1}|z_i|^2 \leqq V_i(z_i) \leqq c_{i2}|z_i|^2$, $DV_{i(33)} \leqq -c_{i3}|z_i|^2$, and $|\nabla V_i(z_i)| \leqq c_{i4}|z_i|$ for all $z_i \in R^{n_i}$, $i = 1, 2$ (see [3]).

The constants k_{12} and k_{21} of hypothesis (ii) of Theorem 3 are $k_{12} = k_1|b_1||c_1|$ and $k_{21} = k_2|b_2||c_2|$.

Choosing $\alpha_1 = c_{24}/(k_1|c_1||b_1|)$ and $\alpha_2 = c_{14}/(k_2|c_2||b_2|)$, matrix S of Theorem 3 assumes the form

$$S = \begin{bmatrix} -c_{13}c_{24}/(k_1|b_1||c_1|) & c_{14}c_{24} \\ c_{14}c_{24} & -c_{14}c_{23}/(k_2|b_2||c_2|) \end{bmatrix}.$$

From Theorem 3 it follows that the equilibrium $(x_1', x_2') = (0', 0')$ is exponentially stable in the large if S is negative definite, i.e., if

$$(34) \qquad k_1k_2 < c_{13}c_{23}/(c_{14}c_{24}|b_1||b_2||c_1||c_2|).$$

Example 5. Reconsider system (32). For S_1 choose $V_1(z_1) = |z_1|$ and for S_2 choose $V_2(z_2) = |z_2|$. Let $W_i = (A_i' + A_i)/2$, $i = 1, 2$. Then $DV_{i(33)} \leqq \lambda_{\max}(W_i)|z_i|$, $i = 1, 2$. $V_i(z_i)$ is globally Lipschitzian in z_i with $L_i = 1$. Matrix $S(t) = S$ of Theorem 4 assumes the form

$$S = \begin{bmatrix} -\lambda_{\max}(W_1) & -k_1|b_1||c_1| \\ -k_2|b_2||c_2| & -\lambda_{\max}(W_2) \end{bmatrix},$$

where $-\lambda_{\max}(W_i) > 0$, $i = 1, 2$. From Theorem 4 it follows that the equilibrium $x' = (x_1', x_2') = 0'$ of (32) is asymptotically stable in the large if all successive principal minors of S are positive, i.e., provided that

$$(35) \qquad k_1k_2 < \frac{|\lambda_{\max}(W_1)||\lambda_{\max}(W_2)|}{|b_1||b_2||c_1||c_2|}.$$

Example 6. Consider the nonlinear, time-invariant discrete system

$$(36) \qquad \begin{aligned} z_1(\tau + 1) &= A_1z_1(\tau) + b_1f_1(\sigma_1(\tau)), & z_2(\tau + 1) &= A_2z_2(\tau) + b_2f_2(\sigma_2(\tau)), \\ \sigma_1(\tau) &= c_1'z_2(\tau), & \sigma_2(\tau) &= c_2'z_1(\tau), \end{aligned}$$

where $z_i \in R^{n_i}$, A_i is an $n_i \times n_i$ matrix such that $\|A_i\| < 1$, $i = 1, 2$, where c_1 is an n_2-vector, c_2 is an n_1-vector, and $f_i(\sigma_i)$, $i = 1, 2$, are real, single-valued functions with the properties $f_i(0) = 0$ and $0 < \sigma_if_i(\sigma_i) < k_i\sigma_i^2$ for all $\sigma_i \neq 0$.

System (36) can be viewed as a nonlinear interconnection of two linear isolated subsystems S_1, S_2,

$$(37) \qquad S_i : z_i(\tau + 1) = A_iz_i(\tau), \qquad i = 1, 2,$$

interconnected by the equations $g_{12}(z_2(\tau)) = b_1f_1(\sigma_1(\tau))$ and $g_{21}(z_1(\tau)) = b_2f_2(\sigma_2(\tau))$. System (36) is clearly a special case of (16).

For S_i choose $V_i(z_i) = |z_i|$, $i = 1, 2$. Then $\Delta V_{i(37)} \leqq (\|A_i\| - 1)|z_i(\tau)|$, and V_i is globally Lipschitzian in z_i with $L_i = 1$.

The constants k_{12} and k_{21} of hypothesis (ii) of Theorem 5 are $k_{12} = k_1|b_1||c_1|$ and $k_{21} = k_2|b_2||c_2|$.

Matrix S of Theorem 5 assumes the form

$$S = \begin{bmatrix} 1 - \|A_1\| & -k_1|b_1||c_1| \\ -k_2|b_2||c_2| & 1 - \|A_2\| \end{bmatrix},$$

where $\|A_i\| = \sqrt{\lambda_{\max}(A_i'A_i)}$. From Theorem 5 it now follows that the equilibrium $x' = (z_1', z_2') = 0'$ of composite system (36) is asymptotically stable in the large if all successive principal minors of S are positive, i.e., if

$$k_1k_2 < \frac{|1 - \|A_1\|||1 - \|A_2\||}{|b_1||b_2||c_1||c_2|}. \tag{38}$$

Example 7. A large class of time-varying capacitor-resistor networks can be described by [5]–[7]

$$\dot{x} + \{AD_1(t)\}x = b(t), \tag{39}$$

where $x \in R^n$, $A = ((a_{ij}))$ is a constant $n \times n$ matrix, $D(t)$ is a time-varying diagonal $n \times n$ matrix with continuous elements $d_{ii}(t)$ which are nonnegative for all $t \in J$, and where the forcing function $b(t)$ is an n-vector with bounded, continuous elements in t. The case where $d_{ii}(t)$, $i = 1, \cdots, n$, are uniformly bounded from below by $\delta > 0$ for all $t \in J$ is of particular interest. Subsequently it is assumed that $b(t) = 0$ for all $t \in J$.

System (39) may be viewed as a linear interconnection of n isolated subsystems S_i,

$$S_i: \dot{x}_i = -a_{ii}d_{ii}(t)x_i, \qquad i = 1, \cdots, n, \tag{40}$$

interconnected by the equations $g_{ij}(x_j, t) = -a_{ij}d_{jj}(t)x_j \triangleq -m_{ij}(t)x_j$, $i \neq j$. System (39) is clearly a special case of (10).

For each S_i choose $V_i(x_i) = \lambda_i|x_i|$, where $\lambda_i > 0$ is a constant, and note that V_i is globally Lipschitzian in x_i with $L_i = \lambda_i$. If it is assumed that $a_{ii} > 0$, then

$$DV_{i(40)} = -\lambda_i d_{ii}(t)a_{ii}|x_i| \triangleq -\lambda_i m_{ii}(t)|x_i| \leqq -\lambda_i\delta a_{ii}|x_i| \triangleq c_i|x_i|.$$

Also, for each $i, j = 1, \cdots, n$, $i \neq j$, and for each $t \in J$, one has

$$k_{ij}|x_j| \triangleq \delta|a_{ij}||x_j| \leqq |g_{ij}(x_j, t)| \leqq |m_{ij}(t)||x_j|.$$

Hence, hypotheses (i) and (ii) of Theorem 4 are satisfied. Matrix $[S(t)]'$ of this theorem assumes the form

$$[S(t)]' = \begin{bmatrix} \lambda_1 m_{11}(t) & -\lambda_2|m_{21}(t)| & \cdots & -\lambda_n|m_{n1}(t)| \\ -\lambda_1|m_{12}(t)| & \lambda_2 m_{22}(t) & \cdots & -\lambda_n|m_{n2}(t)| \\ \cdots & \cdots & \cdots & \cdots \\ -\lambda_1|m_{1n}(t)| & -\lambda_2|m_{2n}(t)| & \cdots & \lambda_n m_{nn}(t) \end{bmatrix}. \tag{41}$$

It follows from Theorem 4 that the equilibrium of composite system (39), with $b(t) = 0$ for all $t \in J$, is asymptotically stable in the large if there exist positive scalars λ_j, $j = 1, \cdots, n$, such that all successive principal minors of $S(t)$ are bounded uniformly from below by some $\gamma > 0$ for all $t \in J$.

It can be shown [7] that if there exist positive scalars λ_j, $j = 1, \cdots, n$, and γ_i, such that

$$m_{jj}(t) - \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{\lambda_i}{\lambda_j}|m_{ij}(t)| \geqq \gamma_j > 0 \tag{42}$$

for all $t \in J$, then all successive principal minors of (41) are positive. Hence, the equilibrium $x = 0$ of system (39) is asymptotically stable in the large if (42) is satisfied.

Remark 11. Certain classes of nonlinear transistor networks [7], [8] may be analyzed similarly as in Example 7 by means of Theorem 4.

9. Proof of main results. The results of §§ 4–7 are now proved.

Proof of Theorem 1. Choose

$$V(x,t) = \sum_{i=1}^{m} \alpha_i V_i(z_i,t), \qquad \alpha_i > 0, \quad i = 1, \cdots, m.$$

Then

$$\sum_{i=1}^{m} \alpha_i \varphi_{i1}(|z_i|) \leqq V(x,t) \leqq \sum_{i=1}^{m} \alpha_i \varphi_{i2}(|z_i|) \quad \text{for all } x \in R^n, \quad t \in J.$$

Since each subsystem (6) possesses Property A, it follows that $V(x,t)$ is positive definite, decrescent, and radially unbounded.

From hypotheses (i) and (iii) it follows that

$$\begin{aligned} DV_{(4)} &= \sum_{i=1}^{m} \left\{ \alpha_i \left[\frac{\partial V_i(z_i,t)}{\partial t} + \nabla V_i(z_i,t)' f_i(z_i,t) \right] + \alpha_i [\nabla V_i(z_i,t)' g_i(x,t)] \right\} \\ &= \sum_{i=1}^{m} \{\alpha_i D V_{i(6)} + \alpha_i [\nabla V_i(z_i,t)' g_i(x,t)]\} \\ &\leqq \sum_{i=1}^{m} \alpha_i \left\{ -\varphi_{i3}(|z_i|) + [\varphi_{i3}(|z_i|)]^{1/2} \sum_{j=1}^{m} a_{ij}(x,t) [\varphi_{j3}(|z_j|)]^{1/2} \right\}. \end{aligned}$$

Letting $w' \triangleq ([\varphi_{13}(|z_1|)]^{1/2}, \cdots, [\varphi_{m3}(|z_m|)]^{1/2})$, one obtains from hypothesis (iii) that

$$DV_{(4)} \leqq w'Sw \leqq -\varepsilon w'w = -\varepsilon \sum_{i=1}^{m} \varphi_{i3}(|z_i|),$$

and since $\varphi_{i3} \in K$, it follows that $DV_{(4)} < 0$ for all $x \neq 0, x \in R^n, t \in J$, and $DV_{(4)} = 0$ for $x = 0$, $t \in J$. Thus, $DV_{(4)}$ is negative definite and the equilibrium $x = 0$ of (4) is asymptotically stable in the large.

Proof of Corollary 1. The proof is similar to the proof of Theorem 1.

Proof of Theorem 2. Choose

$$V(x,t) = \sum_{i=1}^{m} \alpha_i V_i(z_i,t), \qquad \alpha_i > 0, \quad i = 1, \cdots, m.$$

From hypothesis (i) it follows that $V(x,t)$ is positive definite, decrescent, and radially unbounded for all $x \in R^n$ and $t \in J$.

Using hypotheses (i)–(iii) one obtains

$$\begin{aligned} DV_{(9)} &= \sum_{i=1}^{m} \left\{ \alpha_i D V_{i(6)} + \alpha_i \left[\nabla V_i(z_i,t)' \sum_{\substack{j=1 \\ i \neq j}}^{m} A_{ij} f_j(z_j,t) \right] \right\} \\ &\leqq \sum_{i=1}^{m} \left\{ -\alpha_i c_{i1} |f_i(z_i,t)|^2 + \alpha_i c_{i2} |f_i(z_i,t)| \sum_{\substack{j=1 \\ i \neq j}}^{m} \|A_{ij}\| \, |f_j(z_j,t)| \right\} \end{aligned}$$

Letting $w' = (|f_1(z_1, t)|, \cdots, |f_m(z_m, t)|)$ and taking hypothesis (iv) into account, one has

$$DV_{(9)} \leqq w'Sw \leqq \lambda_{\max}(S) \sum_{i=1}^{m} |f_i(z_i, t)|^2 .$$

Since S is negative definite and $f_i \in E$, it follows, taking (ii) into account, that $DV_{(9)} < 0$ for all $x \neq 0$ and $t \in J$, and $DV_{(9)} = 0$ for $x = 0$ and $t \in J$. Hence, the equilibrium $x = 0$ of (9) is asymptotically stable in the large.

Proof of Theorem 3. Choose

$$V(x, t) = \sum_{i=1}^{m} \alpha_i V_i(z_i, t), \qquad \alpha_i > 0, \quad i = 1, \cdots, m.$$

Since each subsystem (6) possesses Property B, one has

$$\begin{aligned} \min_i \{\alpha_i c_{i1}\} |x|^2 = \min_i \{\alpha_i c_{i1}\} \sum_{i=1}^{m} |z_i|^2 \leqq V(x, t) \leqq \max_i \{\alpha_i c_{i2}\} \sum_{i=1}^{m} |z_i|^2 \\ = \max_i \{\alpha_i c_{i2}\} |x|^2 \end{aligned} \tag{43}$$

for all $x \in R^n$ and $t \in J$. Also, from hypotheses (i) and (ii) one obtains

$$\begin{aligned} DV_{(10)} &= \sum_{i=1}^{m} \left\{ \alpha_i DV_{i(6)} + \alpha_i \nabla V_i(z_i, t)' \sum_{j=1}^{m} g_{ij}(z_j, t) \right\} \\ &\leqq \sum_{i=1}^{m} \left\{ -\alpha_i c_{i3} |z_i|^2 + \alpha_i c_{i4} |z_i| \sum_{\substack{i=1 \\ i \neq j}}^{m} k_{ij} |z_j| \right\} . \end{aligned}$$

Letting $w' = (|z_1|, \cdots, |z_m|)$ and letting $R = ((r_{ij}))$ denote the $m \times m$ matrix determined by

$$r_{ij} = \begin{cases} -\alpha_i c_{i3} & \text{if } i = j, \\ \alpha_i c_{i4} k_{ij} & \text{if } i \neq j, \end{cases}$$

one has, in view of hypothesis (iii),

$$DV_{(10)} \leqq w'Rw = w' \frac{(R + R')}{2} w = w'Sw \leqq \lambda_{\max}(S) |w|^2 = \lambda_{\max}(S) |x|^2 \tag{44}$$

for all $x \in R^n$ and for all $t \in J$. Since S is negative definite, it follows from (43) and (44) that the equilibrium $x = 0$ of (10) is exponentially stable in the large.

Remark. In the remaining proofs the following preliminary result is required.

THEOREM A ([9, p. 305] and [10, p. 71]). *If $A = ((a_{ij}))$ is a real $m \times m$ matrix such that $a_{ij} \leqq 0$, $i \neq j$, and if all successive principal minors of A are positive, then A^{-1} exists and all elements of A^{-1} are nonnegative.*

Proof of Theorem 5. Let $\alpha' = (\alpha_1, \cdots, \alpha_m)$, $\alpha_i > 0$, $i = 1, \cdots, m$, be an *arbitrary* vector, and choose $V(x, \tau) = \sum_{i=1}^{m} \alpha_i V_i(z_i, \tau)$. Then

$$\sum_{i=1}^{m} \alpha_i \varphi_i(|z_i|) \leqq V(x, \tau) \leqq \sum_{i=1}^{m} \alpha_i \varphi_{i2}(|z_i|) \quad \text{for all } x \in R^n, \quad \tau \in I.$$

Since each subsystem (13) possesses Property C, it follows that $V(x, \tau)$ is positive definite and radially unbounded for all $x \in R^n$, $\tau \in I$.

From hypotheses (i) and (ii) it follows that

$$\begin{aligned}\Delta V_{(16)} &= \sum_{i=1}^{m} \alpha_i V_i \left[f_i(z_i, \tau) + \sum_{\substack{j=1\\ i\neq j}}^{m} g_{ij}(z_j, \tau), \tau + 1 \right] - \sum_{i=1}^{m} \alpha_i V_i(z_i, \tau) \\ &= \sum_{i=1}^{m} \alpha_i \left\{ V_i \left[f_i(z_i, \tau) + \sum_{\substack{j=1\\ i\neq j}}^{m} g_{ij}(z_j, \tau), \tau + 1 \right] - V_i[f_i(z_i, \tau), \tau + 1] \right\} \\ &\quad + \sum_{i=1}^{m} \alpha_i \, \Delta V_{i(13)} \leqq \sum_{i=1}^{m} \left\{ \alpha_i L_i \left| \sum_{\substack{j=1\\ i\neq j}}^{m} g_{ij}(z_j, \tau) \right| \right\} + \sum_{j=1}^{m} \{ -\alpha_i c_i |z_i| \} \\ &\leqq \sum_{i=1}^{m} -\alpha_i \left\{ c_i |z_i| + (-L_i) \sum_{\substack{j=1\\ i\neq j}}^{m} k_{ij} |z_j| \right\}.\end{aligned}$$

Letting $y' = \alpha' S$, where S is defined in hypothesis (iii), and letting $w' = (|z_1|, \cdots, |z_m|)$, one has

$$\Delta V_{(16)} \leqq (-\alpha' S) w = (-y') w.$$

Since S has positive successive principal minors and since $s_{ij} \leqq 0$ for all $i \neq j$, it follows from Theorem A that S^{-1} exists and that all elements of S^{-1} are nonnegative. Thus, $\alpha = (S^{-1})' y$. Also, since all elements of S^{-1} are nonnegative and since in each row and column of S^{-1} there is at least one nonzero element, one can choose for *any* α, with strictly positive elements, a vector y in such a fashion that $y_i > 0$, $i = 1, \cdots, m$. In doing so one obtains

$$\Delta V_{(16)} \leqq (-y') w < 0, \qquad x \neq 0,$$

and $\Delta V_{(16)} = 0$, $x = 0$, for all $\tau \in I$ (note that $x = 0$ if and only if $w = 0$). Hence, $\Delta V_{(16)}$ is negative definite and the equilibrium $x = 0$ of (16) is asymptotically stable in the large.

Proof of Theorem 6. Let $\alpha' = (\alpha_1, \cdots, \alpha_m)$, $\alpha_i > 0$, $i = 1, \cdots, m$, be an arbitrary vector, and choose $V(x, \tau) = \sum_{i=1}^{m} \alpha_i V_i(z_i, \tau)$. From hypothesis (i) it follows that $V(x, \tau)$ is positive definite and radially unbounded. Also, in view of hypotheses (i) and (ii) one obtains

$$\begin{aligned}\Delta V_{(15)} &= \sum_{i=1}^{m} \alpha_i \{\Delta V_{i(13)}\} + \sum_{i=1}^{m} \alpha_i \left\{ V_i \left[f_i(z_i, \tau) \right.\right. \\ &\qquad \left.\left. + \sum_{\substack{j=1\\ i\neq j}}^{m} A_{ij} f_j(z_j, \tau), \tau + 1 \right] - V_i[f_i(z_i, \tau), \tau + 1] \right\} \\ &\leqq \sum_{i=1}^{m} -\alpha_i \left\{ c_i |f_i(z_i, \tau)| + \sum_{\substack{j=1\\ i\neq j}}^{m} -L_i \|A_{ij}\| \, |f_j(z_j, \tau)| \right\}.\end{aligned}$$

Let $y' = \alpha' S$, where S is defined in hypothesis (iii), and let $w' = (|f_1(z_1, \tau)|, \cdots, |f_m(z_m, \tau)|)$. Then

$$\Delta V_{(11)} \leqq (-\alpha' S) w = (-y') w.$$

Noting that $x = 0$ if and only if $w = 0$ (recall that $f_i \in E$), noting that S satisfies the condition of Theorem A, and taking hypothesis (ii) into account, the negative definiteness of $\Delta V_{(11)}$ is established as in the proof of Theorem 5.

Proof of Theorem 7. Let $\alpha' = (\alpha_1, \cdots, \alpha_m)$, $\alpha_i > 0$, $i = 1, \cdots, m$, be an arbitrary vector, and choose $V(x, \tau) = \sum_{i=1}^m \alpha_i V_i(z_i, \tau)$. From hypothesis (i) it follows that $V(x, \tau)$ is positive definite and radially unbounded for all $x \in R^n$ and $\tau \in I$. Also, in view of hypotheses (i), (ii) and (iii), one obtains

$$\Delta V_{(11)} = \sum_{i=1}^m \alpha_i\{V_i[f_i(z_i, \tau) + g_i(x, \tau), \tau + 1]\} - \sum_{i=1}^m \alpha_i V_i(z_i, \tau)$$

$$\leqq \sum_{i=1}^m - \alpha_i\left\{c_i\varphi_{i3}(|z_i|) - L_i \sum_{j=1}^m a_{ij}(x, \tau)\varphi_j(|z_j|)\right\}.$$

Letting $y' = \alpha' S$, where S is defined in hypothesis (iv), and letting

$$w' = (\varphi_{13}(|z_1|), \cdots, \varphi_{m3}(|z_m|)),$$

one has

$$\Delta V_{(11)} \leqq (-\alpha' S)w = (-y')w.$$

Noting that $\varphi_{i3} \in K$, $i = 1, \cdots, m$, noting that S satisfies the conditions of Theorem A, and taking hypothesis (ii) into account, the desired result follows as in the proof of Theorem 5.

Proof of Theorem 10. Let $\alpha' = (\alpha_1, \cdots, \alpha_m)$, $\alpha_i > 0$, $i = 1, \cdots, m$, be an arbitrary vector, and choose the functional $V(\varphi, t) = \sum_{i=1}^m \alpha_i V_i(\varphi^i, t)$. From hypothesis (i) it follows that

$$\sum_{i=1}^m \alpha_i \psi_{i1}(\|\varphi^i\|) \leqq V(\varphi, t) \leqq \sum_{i=1}^m \alpha_i \psi_{i2}(\|\varphi^i\|).$$

Thus, $V(\varphi, t)$ is positive definite and decrescent whenever $t \in J$ and $\varphi^i \in C_{H_i}$, $i = 1, \cdots, m$.

Along solutions of (19) one has

$$\sum_{i=1}^m \alpha_i\{V_i[z_{t+\Delta t}^i(\cdot\,; \varphi^i, t_0), t + \Delta t] - V_i[z_t^i(\cdot\,; \varphi^i, t_0), t]\}$$

$$= \sum_{i=1}^m \alpha_i\{V_i[z_t^i(\cdot\,; \varphi^i, t_0) + \Delta t \cdot f_i(z_t^i(\cdot\,; \varphi^i, t_0), t) + o(t), t + \Delta t]$$

$$- V_i[z_t^i(\cdot\,; \varphi^i, t_0), t]\}$$

$$+ \sum_{i=1}^m \alpha_i\{V_i[z_t^i(\cdot\,; \varphi^i, t_0) + \Delta t \cdot f_i(z_t^i(\cdot\,; \varphi^i, t_0), t)$$

$$+ \Delta t \cdot \sum_{\substack{j=1 \\ i \neq j}}^m g_{ij}(z_t^j(\cdot\,; \varphi^j, t_0), t) + o(t), t + \Delta t]$$

$$- V_i[z_t^i(\cdot\,; \varphi^i, t_0) + \Delta t \cdot f_i(z_t^i(\cdot\,; \varphi^i, t_0), t) + o(t), t + \Delta t]\}.$$

In view of hypotheses (i) and (ii) one has

$$\begin{aligned} DV_{(19)} &= \limsup_{\Delta t \to 0^+} \left(\frac{1}{\Delta t}\right) \sum_{i=1}^{m} \alpha_i \{V_i[z^i_{t+\Delta t}(\cdot\,;\varphi^i, t_0), t + \Delta t] - V_i[z^i_t(\cdot\,;\varphi^i, t_0), t]\} \\ &\leqq \sum_{i=1}^{m} \alpha_i DV_{i(20)} + \sum_{i=1}^{m} \alpha_i \left\{ L_i \sum_{\substack{j=1 \\ i \neq j}}^{m} \|g_{ij}(z^j_t(\cdot\,;\varphi^j, t_0), t)\| \right\} \\ &\leqq \sum_{i=1}^{m} \alpha_i \left\{ -c_i \psi_{i3}(\|\varphi^i\|) + L_i \sum_{\substack{j=1 \\ i \neq j}}^{m} k_{ij} \psi_{j3}(\|\varphi^j\|) \right\}. \end{aligned}$$

Now let $w' = (\psi_{13}(\|\varphi^1\|), \cdots, \psi_{m3}(\|\varphi^m\|))$. Since $\psi_{i3} \in K$, $i = 1, \cdots, m$, it follows that $w = 0$ if and only if $x_t(\cdot\,;\varphi, t_0) = 0$ for all $t \in J$. Letting $\alpha' S = y'$, where S is defined in hypothesis (iii), one obtains

$$DV_{(19)} \leqq (-\alpha' S)w = (-y')w.$$

Since S satisfies the conditions of Theorem A, the negative definiteness of $DV_{(19)}$, for all $t \in J$ and for all $\varphi^i \in C_{H_i}$, $i = 1, \cdots, m$, follows from the identical argument used in Theorem 5. Hence, the equilibrium $x = 0$ of (19) is asymptotically stable.

Proofs of Theorems 11 *and* 12. The proofs are similar to the proof of Theorem 10.

Proof of Theorem 4. Let $\alpha' = (\alpha_1, \cdots, \alpha_m)$, $\alpha_i > 0$, $i = 1, \cdots, m$, be an arbitrary vector, and choose $V(x, t) = \sum_{i=1}^m \alpha_i V_i(z_i, t)$. From hypothesis (i) it follows that V is positive definite, decrescent, and radially unbounded for all $x \in R^n$ and for all $t \in J$.

Following the same procedure as in Theorem 10, and taking hypotheses (i) and (ii) into account, one obtains

$$\begin{aligned} DV_{(10)} &= \limsup_{\Delta t \to 0^+} \left(\frac{1}{\Delta t}\right) \left\{ \sum_{i=1}^{m} \alpha_i V_i[z_i(t + \Delta t; z_{i_0}, t_0), t + \Delta t] - V_i[z_i(t; z_{i_0}, t_0), t] \right\} \\ &\leqq \sum_{i=1}^{m} \alpha_i DV_{i(6)} + \sum_{i=1}^{m} \alpha_i \left\{ L_i \sum_{\substack{j=1 \\ i \neq j}}^{m} |g_{ij}[z_j(t; z_{j_0}, t_0), t]| \right\} \\ &\leqq \sum_{i=1}^{m} \alpha_i \left\{ -c_i(t) \psi_{i3}[|z_i(t; z_{i_0}, t_0)|] + L_i \sum_{\substack{j=1 \\ i \neq j}}^{m} k_{ij}(t) \psi_{j3}[|z_j(t; z_{j_0}, t_0)|] \right\}. \end{aligned}$$

Now let $w' = (\psi_{13}[|z_1(t; z_{1_0}, t_0)|], \cdots, \psi_{m3}[|z_m(t; z_{m_0}, t_0)|])$ and let $\alpha' S(t) = y(t)'$, where $S(t)$ is defined in hypothesis (iii). Then

$$DV_{(10)} \leqq (-\alpha' S(t))w = (-y(t)')w.$$

Since $\psi_{i3} \in K$, $i = 1, \cdots, m$, it follows that $w = 0$ if and only if $x = 0$. Noting that $S(t)$ has successive principal minors bounded uniformly from below and noting that $S(t)$ satisfies all conditions of Theorem A, the negative definiteness of $DV_{(10)}$ for all $t \in J$ and for all $x \in R^n$ is established using the same argument as in Theorem 5. Hence, the equilibrium $x = 0$ of (10) is asymptotically stable in the large.

Remark. In the proof of Theorem 8, the following preliminary result is employed.

LEMMA 1. *Assume that each subsystem* (6) *of composite system* (17) *possesses Property* D. *Let* $z_i(t; z_i(kT), kT) \triangleq z_i(t)$ *denote the solution of*

$$\dot{z}_i(t) = f_i(z_i(t), t) + \sum_{\substack{j=1\\ i\neq j}}^{m} C_{ij}z_j(kT) \tag{45}$$

over the interval $[kT, (k + 1)T)$. *Then the solution* $z_i(t)$ *of* (45) *satisfies, for all* $t \in [kT, (k + 1)T)$, *the estimate*

$$|z_i(t)| \geqq \frac{1}{c_{iI}}[e^{-(c_{i2})T}]|z_i(kT)| - \left\{\frac{1}{c_{i1}}\left(\frac{L_i}{c_{i2}}\right)[1 - e^{-(c_{i2})T}]\left\{\sum_{\substack{j=1\\ i\neq j}}^{m} \|C_{ij}\|\,|z_j(kT)|\right\}\right\}. \tag{46}$$

Proof. Let

$$u_i(t) \triangleq \sum_{\substack{j=1\\ i\neq j}}^{m} C_{ij}z_j(kT) = u_i(kT), \qquad t \in [kT, (k + 1)T).$$

Along solutions of (45) one has, since (6) possesses Property D,

$$\begin{aligned} DV^{(-)}_{i(45)} &= \liminf_{\Delta t\to 0^+}\left(\frac{1}{\Delta t}\right)\{V_i[z_i(t + \Delta t), t + \Delta t] - V_i[z_i(t), t]\} \\ &\geqq DV^{(-)}_{i(6)} - L_i\left|\sum_{\substack{j=1\\ i\neq j}}^{m} C_{ij}z_j(kT)\right| \\ &\geqq -c_{i2}|z_i(t)| - L_i|u_i(kT)| \end{aligned}$$

for all $t \in [kT, (k + 1)T)$. Now let

$$k_{i1} \triangleq L_i|u_i(kT)| \leqq L_i \sum_{\substack{j=1\\ i\neq j}}^{m} \|C_{ij}\|\,|z_j(kT)|,$$

and let

$$v_i(t) \triangleq V_i[z_i(t), t] \quad \text{for all } t \in [kT, (k + 1)T).$$

Then

$$Dv^{(-)}_{i(45)} \geqq -c_{i2}v_i(t) - k_{i1}, \quad t \in [kT, (k + 1)T),$$

$$v_i(t) \geqq e^{c_{i2}(t-kT)}v_i(kT) - k_{i1}\int_{kT}^{t} e^{-c_{i2}(t-\tau)}d\tau$$

and

$$V_i(z_i(t), t) \geqq e^{c_{i2}(kT)}\,e^{-c_{i2}t}V_i[z_i(kT), kT] - \frac{k_{i1}}{c_{i2}}[1 - e^{c_{i2}(kT)}\,e^{-c_{i2}t}]$$

for all $t \in [kT, (k + 1)T)$. Since (6) possesses Property D, one has

$$c_{i1}|z_i(t)| \geqq e^{c_{i2}(kT)}\,e^{-c_{i2}t}|z_i(kT)| - \frac{k_{i1}}{c_{i2}}[1 - e^{c_{i2}(kT)}\,e^{-c_{i2}t}]$$

or

$$
\begin{aligned}
|z_i(t)| &\geqq \frac{1}{c_{i1}}[e^{c_{i2}(kT)}][e^{-(c_{i2})t}]|z_i(kT)| \\
&\quad - \left\{\frac{1}{c_{i1}}\left(\frac{L_i}{c_{i2}}\right)[1 - e^{c_{i2}(kT)}][e^{-(c_{i2})t}]\right\}\left\{\sum_{\substack{j=1\\ i\neq j}}^{m} \|C_{ij}\|\,|z_j(kT)|\right\} \\
&\geqq \frac{1}{c_{i1}}[e^{-(c_{i2})T}]|z_i(kT)| - \left\{\frac{1}{c_{i1}}\left(\frac{L_i}{c_{i2}}\right)[1 - e^{-(c_{i2})T}]\right\}\left\{\sum_{\substack{j=1\\ i\neq j}}^{m} \|C_{ij}\|\,|z_j(kT)|\right\}
\end{aligned}
$$

for all $t \in [kT, (k + 1)T)$. This concludes the proof.

Proof of Theorem 8. Let $\alpha' = (\alpha_1, \cdots, \alpha_m)$, $\alpha_i > 0$, $i = 1, \cdots, m$, be an arbitrary vector, and choose $V: R^n \times [0, \infty) \to R^1$ as $V(x, t) = \sum_{i=1}^m \alpha_i V_i(z_i, t)$. Since (6) possesses Property D, it follows that V is positive definite, decrescent, and radially unbounded for all $x \in R^n$ and for all $t \in [0, \infty)$.

Let $z_i(t) \triangleq z_i(t; z_i(kT), kT)$ for all $t \in [kT, (k + 1)T)$. In view of hypotheses (i) and (ii), one has

$$
\begin{aligned}
DV_{(17)} &= \limsup_{\Delta t \to 0^+} \left(\frac{1}{\Delta t}\right)\left\{\sum_{i=1}^{m} \alpha_i V_i[z_i(t + \Delta t; z_i(0), 0), t + \Delta t] - \alpha_i V_i[z_i(t; z_i(0), 0), t]\right\} \\
&\leqq \sum_{i=1}^{m} \alpha_i DV_{i(6)} + \sum_{i=1}^{m} \alpha_i L_i \left\{\sum_{\substack{j=1\\ i\neq j}}^{m} \|C_{ij}\|\,|z_j(kT)|\right\} \\
&\leqq \sum_{i=1}^{m} (-\alpha_i c_{i3})|z_i(t)| + \sum_{i=1}^{m} (\alpha_i L_i) \left\{\sum_{\substack{j=1\\ i\neq j}}^{m} \|C_{ij}\|\,|z_j(kT)|\right\}
\end{aligned}
$$

for all $t \in [kT, (k + 1)T)$, $k = 0, 1, 2, \cdots$. Utilzing (46) obtained in Lemma 1, one has

$$
\begin{aligned}
DV_{(17)} &\leqq \sum_{i=1}^{m} \alpha_i \left\{-\frac{c_{i3}}{c_{i1}}[e^{-c_{i2}T}]|z_i(kT)|\right\} \\
&\quad + \sum_{i=1}^{m} \alpha_i \left\{\frac{c_{i3}L_i}{c_{i1}c_{i2}}[1 - e^{-c_{i2}T}] + L_i\right\} \sum_{\substack{j=1\\ i\neq j}}^{m} \|C_{ij}\|\,|z_j(kT)|
\end{aligned}
$$

for all $t \in [kT, (k + 1)T)$, $k = 0, 1, 2, \cdots$.

Now let $w(kT)' = (|z_1(kT)|, \cdots, |z_m(kT)|)$, and let $\alpha' = y'S$, where S is defined in hypothesis (iii). Then

$$
DV_{(17)} \leqq (-\alpha' S)w(kT) = (-y')w(kT)
$$

for all $t \in [kT, (k + 1)T)$, $k = 0, 1, 2, \cdots$.

Noting that $x = 0$ if and only if $w = 0$, and noting that S satisfies the conditions of Theorem A, the negative definiteness of $DV_{(17)}$ for all $x \in R^n$ and for all $t \in [0, \infty)$ is established using the argument employed in Theorem 5. Hence, the equilibrium $x = 0$ of (17) is asymptotically stable in the large.

Proof of Theorem 9. The proof is similar to the proof of Theorem 8.

10. Concluding remarks. At this point several comments are in order.

1. Most of the results of §§4, 5 and 6 (dealing with interconnected systems described by ordinary differential equations, difference equations, and sampled-data systems) are clearly special cases of the results of § 7 (concerned with systems described by functional differential equations). Thus, the results of § 7 constitute a unifying setting for a relatively large class of stability problems. The stability analysis of other specific classes of composite systems described by functional differential equations (e.g., systems with time lags) is being considered by the author.

2. In the present approach, converse Lyapunov theorems (see, e.g., [3]) which guarantee the existence of suitable Lyapunov functions for stable subsystems play a crucial role.

3. The standard Lyapunov theorems yield actually more information than is usually stated; they yield *uniform* stability, *uniform* asymptotic stability, etc. (see, e.g., [3]). This is also true in the present case, i.e., the hypotheses of all theorems in § 4 through § 7 and of all examples in § 8 are strong enough to insure *uniform asymptotic stability or uniform exponential stability.*

4. The role played by Theorem A in the proofs of the present results suggests that this theorem and related ones may find wide application in the stability analysis of a variety of other problems. The type of argument presented using this theorem appears to be new.

5. It should be noted that Theorem 4 allows the use of V-functions which are not necessarily continuously differentiable.

6. Theorems 8 and 9 can be extended to sampled-data systems with various types of nonlinear interconnections in an obvious way. Also, the method of § 7 can be modified to accommodate systems described by functional differential equations with interconnecting structures not considered here.

7. A good survey of the work dealing with stability analysis of large-scale systems is given in [2]. Related problems are also considered in [11]. The results presented in [12] can be obtained using Theorem 6. The present methods differ significantly from those employed in [12]. Conditions (26), (30) and (34) of Examples 1, 2 and 4 are less conservative than corresponding results reported in [4] and are obtained in a manner far more straightforward than was done previously. Condition (42) of Example 7 was obtained in [7] by methods significantly different from the present procedure.

Acknowledgment. The author is indebted to Prof. Dr. Wolfgang Hahn from the Technical University of Graz, Austria, for numerous suggestions. He also wishes to thank the reviewer and Dr. D. W. Porter for their comments and discussions.

REFERENCES

[1] F. N. Bailey, *The application of Lyapunov's second method to interconnected systems*, this Journal, 3 (1965), pp. 443–462.

[2] D. D. Šiljak, *Stability of large-scale systems*, Proc. Fifth World Congress of IFAC (Session 9,) Nonlinear Systems), Paris, 1972.

[3] W. Hahn, *Stability of Motion*, Springer-Verlag, New York, 1967.

[4] A. A. Piontkovskii and L. D. Rutkovskaya, *Investigation of stability-theory problems by the vector Lyapunov function method*, Automation and Remote Control, 10 (1967), pp. 1422–1429.

[5] H. H. ROSENBROCK, *A Lyapunov function with applications to some nonlinear physical systems*, Automatica, 1 (1963), pp. 31–53.

[6] ———, *A Lyapunov function for some naturally occurring linear homogeneous time dependent equations*, Ibid., 1 (1963), pp. 97–109.

[7] D. MITRA AND H. C. SO, *Existence conditions for L_1 Lyapunov Functions for a class of nonautonomous systems*, IEEE Trans. Cirtuit Theory, CT-19 (1972), pp. 594–598.

[8] I. W. SANDBERG, *Some theorems on the dynamic response of nonlinear transitor networks*, Bell System Tech. J., 48 (1969), pp. 35–54.

[9] R. BELLMAN, *Introduction to Matrix Analysis*, 2nd ed., McGraw-Hill, New York, 1964.

[10] F. R. GANTMACHER, *The Theory of Matrices*, vol. II, Chelsea, New York, 1964.

[11] A. N. MICHEL AND D. W. PORTER, *Stability of composite systems*, Fourth Asilomar Conference on Circuits and Systems, Monterey, Calif., 1970.

[12] M. ARAKI, K. ANDO AND B. KONDO, *Stability of sampled-data composite systems with many nonlinearities*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 22–27.

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

LION AND MAN: THE GENERAL CASE*

JAMES FLYNN†

Abstract. A lion L and a man M, confined to a circular arena, move with speeds bounded by 1 and $w > 1$, respectively. L tries to get as close as possible to M who wants to maintain the greatest possible distance between himself and L. Using the approach of Ryll-Nardzewski, Varaiya and Lin, we formulate this conflict as a differential game. Then we show that there exists a $d^* > 0$ such that the following holds: For each starting position, L has a strategy which brings him as close as d^* to M in a finite time. Furthermore, there exist starting positions such that for each $\varepsilon > 0$, M has a strategy which keeps the distance $|LM| \geqq d^* - \varepsilon$. We also show how to construct L's strategy. The fact that L can always achieve d^* in a *finite time* contrasts our results for the case $w > 1$ with Besicovitch's well-known result for the case $w = 1$ (Rado's pursuit problem).

Introduction. A lion L and a man M, confined to a circular arena, move with speeds bounded by 1 and $w > 1$, respectively. L and M, continually aware of each other's positions, have contrary objectives. L wants to get as close as possible to M while M wants to maintain the greatest possible distance between himself and L. How close L can get to M depends, of course, on the starting position. We will show that there exists a $d^* > 0$ such that the following holds: First, for each starting position, L has a strategy which brings him as close as d^* to M in a finite time. And, second, there exist starting positions such that for each $\varepsilon > 0$, M has a strategy which keeps L as far away as $d^* - \varepsilon$. (This d^* is the same as the one which we define by equation (61) of [4]. Numerical bounds on d^* are given in § 5 of that paper.)

The fact that L can be sure of achieving d^* in a finite time contrasts this case ($w > 1$) with the case of equal speeds ($w = 1$). In the latter case, d^* is equal to 0, since L can get arbitrarily close to M by going to the center and then moving out along the radius OM. However, in that case, L does not have a strategy which always achieves d^* in a finite time. This is a consequence of the following result which Littlewood [6] attributes to A. S. Besicovitch: By moving along an appropriate polygonal path, M can prevent L from ever actually achieving pointwise capture (see [6, pp. 135–136] for details). More recently, Croft [1] has shown that L can guarantee pointwise capture if we require that the curvature of M's path exist and be uniformly bounded.

This paper is a continuation of [3], which dealt with the case where M is restricted to the circumference, and of [4], which developed a characterization of d^*. As in [3] we choose to view the conflict between L and M as a differential game where the payoff is from L to M and is equal to the "smallest" distance between the two. In the first section we formulate the appropriate differential game using the approach developed by Ryll-Nardzewski [7] and Varaiya and Lin [9]. Then, in the second, we describe our results and outline our arguments. The details supporting these arguments are given in the third and fourth sections.

We attack the problem by separating it into a min-max pursuit part and a max-min evasion part. In [4] we consider the evasion part and show that if the

* Received by the editors February 6, 1973.

† Graduate School of Business, University of Chicago, Chicago, Illinois 60637.

starting position is one where M is at a distance $wd^*/(w-1)$ from the center O and the distance $|LM|$ is equal to d^*, then for any given $\varepsilon > 0$, M can keep $|LM| \geqq d^* - \varepsilon$ by using a strategy which keeps $|LM|$ constant whenever L moves at maximum speed along a polygonal path. When L follows an arbitrary trajectory, M's strategy prevents the distance $|LM|$ from deviating from its original value by more than ε (see [4, Remark 2]).

In this paper, we focus on the pursuit part. We will show that given any starting position, L can get as close as d^* to M in a finite time by using a two-stage strategy. In the first stage, L goes to the center O and then moves out along the radius OM until he reaches a position where $|LM| \leqq d^*$ or one where $|OL| \geqq d^*/(w-1)$ and $|OM| > wd^*/(w-1)$. We construct a strategy under which this happens by time $2/\ln(w)$. Now, given any position where L lies on the radius OM, $|OL| \geqq d^*/(w-1)$ and $|OM| > wd^*/(w-1)$, L can force the distance $|LM|$ strictly below d^* (see Theorem 3). Hence, if the position at the end of the first stage is one where $|LM| > d^*$, then L can switch to a second-stage strategy which brings him strictly closer than d^* to M. Hence, L can always achieve d^* in a finite time. We also have as a corollary the result that staying on the boundary of the arena can be a bad strategy for the man.

We wish to thank L. Dubins and D. Blackwell for introducing us to the problem of *pursuit in the circle*. This problem is a generalization of R. Rado's *Lion and Man* problem [6] and Isaac's game of *pursuit in the half-plane* [5, pp. 261–265]. Apparently Isaacs was the first one to formulate it [5, pp. 265, 270] and Gerald J. Smith [8] was the first one to make a serious attempt at solving it. We have benefited greatly from Smith's unpublished results [8].

1. Formulation. The formulation which we use for this problem is essentially the same as the one we developed in [3]. Like the latter, it relies heavily on Ryll-Nardzewski [7] and Varaiya and Lin [9]. For detailed references, arguments and explanations, we refer the reader to Flynn [3, § 2].

Assume that the arena has radius 1. Let $\mathbf{R}$ represent the real numbers and let $\mathbf{R}^2$ represent the space $\mathbf{R} \times \mathbf{R}$ with the metric determined by the norm $\|\cdot\|$, where $\|(r_1, r_2)\| = (r_1^2 + r_2^2)^{1/2}$. Denote the unit circle in $\mathbf{R}^2$ by $\mathbf{C}$, the closed unit disc in $\mathbf{R}^2$ by $\mathbf{D}$ and the time axis $[0, \infty)$ by $\mathbf{T}$. Define:

$$\mathbf{L}(l) = \{\mathbf{l} \mid \mathbf{l}:\mathbf{T} \to \mathbf{D}, \mathbf{l}(0) = l \text{ and } \|\mathbf{l}(t') - \mathbf{l}(t'')\| \leqq \|t' - t''\| \text{ for all } t', t'' \in \mathbf{T}\} \quad \text{for } l \in \mathbf{D};$$

$$\mathbf{M}(\mathscr{m}) = \{\mathscr{m} \mid \mathscr{m}:\mathbf{T} \to \mathbf{D}, \mathscr{m}(0) = \mathscr{m} \text{ and } \|\mathscr{m}(t') - \mathscr{m}(t'')\| \leqq w\|t' - t''\| \text{ for all } t', t'' \in \mathbf{T}\} \quad \text{for } \mathscr{m} \in \mathbf{D};$$

$$\mathbf{L} = \bigcup_{l \in \mathbf{D}} \mathbf{L}(l), \quad \mathbf{M} = \bigcup_{\mathscr{m} \in \mathbf{D}} \mathbf{M}(\mathscr{m}) \quad \text{and} \quad P(\mathbf{l}, \mathscr{m}) = \inf_{t \in \mathbf{T}} \|\mathbf{l}(t) - \mathscr{m}(t)\| \quad \text{for } (\mathbf{l}, \mathscr{m}) \in \mathbf{L} \times \mathbf{M}.$$

$\mathbf{L}(l)$ and $\mathbf{M}(\mathscr{m})$ are sets of trajectories for L and M originating from the respective positions l and $\mathscr{m}$, while $P(\mathbf{l}, \mathscr{m})$ is the payoff from L to M when L uses $\mathbf{l}$ and M uses $\mathscr{m}$. We consider $\mathbf{L}$ and $\mathbf{M}$ as compact metric spaces with respect to the topology of uniform convergence on compact subsets of $\mathbf{T}$. Clearly, P is a continuous function on the compact space $\mathbf{L} \times \mathbf{M}$.

Now we define strategies. Let $(l, \mathscr{m}) \in \mathbf{D} \times \mathbf{D}$. We call a mapping $\pi : \mathbf{M}(\mathscr{m}) \to \mathbf{L}(l)$ a pursuit strategy at $(l, \mathscr{m})$ if it satisfies the *information constraint*: for any $\mathscr{m}', \mathscr{m}'' \in \mathbf{M}(\mathscr{m})$, $\mathscr{m}'(t) = \mathscr{m}''(t)$ for $0 \leqq t \leqq t'$ implies that $\pi(\mathscr{m}')(t) = \pi(\mathscr{m}'')(t)$ for $0 \leqq t \leqq t'$. We can define an evasion strategy η in a similar manner. Let $\Pi(l, \mathscr{m})$ and $H(l, \mathscr{m})$, respectively, denote the sets of pursuit and evasion strategies at $(l, \mathscr{m})$.

We would like to define the outcome which results when L chooses $\pi \in \Pi(l, \mathscr{m})$ and M chooses $\eta \in H(l, \mathscr{m})$ as any pair $(\mathbf{l}, \mathscr{m}) \in \mathbf{L}(l) \times \mathbf{M}(\mathscr{m})$ satisfying

$$\pi(\mathscr{m}) = \mathbf{l}, \quad \eta(\mathbf{l}) = \mathscr{m}.$$

Unfortunately that system might not have a solution. (See Remark 1 of Flynn [3].) The following definition is due to Varaiya and Lin [9].

DEFINITION 1. $(\mathbf{l}, \mathscr{m}) \in \mathbf{L}(l) \times \mathbf{M}(\mathscr{m})$ is an *outcome* of $(\pi, \eta) \in \Pi(l, \mathscr{m}) \times H(l, \mathscr{m})$ if there exist sequences $\langle \mathbf{l}_n \rangle_{n=1}^{\infty} \subset \mathbf{L}(l)$ and $\langle \mathscr{m}_n \rangle_{n=1}^{\infty} \subset \mathbf{M}(\mathscr{m})$ such that

$$\lim_n \mathbf{l}_n = \lim_n \pi(\mathscr{m}_n) = \mathbf{l}, \qquad \lim_n \mathscr{m}_n = \lim_n \eta(\mathbf{l}_n) = \mathscr{m}.$$

Let $O(\pi, \eta)$ denote the set of outcomes of (π, η). One can show that $O(\pi, \eta)$ is a nonempty compact subset of $\mathbf{L}(l) \times \mathbf{M}(\mathscr{m})$. Also, using the continuity of P, one can show that

$$\sup_{\mathscr{m} \in \mathbf{M}(\mathscr{m})} P(\pi(\mathscr{m}), \mathscr{m}) \geqq P(\mathbf{l}, \mathscr{m}) \geqq \inf_{\mathbf{l} \in \mathbf{L}(l)} P(\mathbf{l}, \eta(\mathbf{l})) \tag{1}$$

holds for every $\pi \in \Pi(l, \mathscr{m})$, $\eta \in H(l, \mathscr{m})$, and $(\mathbf{l}, \mathscr{m}) \in O(\pi, \eta)$. (See relationship (7) of [3].)

We define the game $\Gamma(l, \mathscr{m})$ as follows. L selects a $\pi \in \Pi(l, \mathscr{m})$ while M independently selects an $\eta \in H(l, \mathscr{m})$. The payoff from L to M is $P(\mathbf{l}, \mathscr{m})$, where $(\mathbf{l}, \mathscr{m})$ is an arbitrary point in $O(\pi, \eta)$. If

$$\inf_{\pi \in \Pi(l,\mathscr{m})} \sup_{\eta \in H(l,\mathscr{m})} \max_{(\mathbf{l},\mathscr{m}) \in O(\pi,\eta)} P(\mathbf{l}, \mathscr{m}) = \sup_{\eta \in H(l,\mathscr{m})} \inf_{\pi \in \Pi(l,\mathscr{m})} \min_{(\mathbf{l},\mathscr{m}) \in O(\pi,\eta)} P(\mathbf{l}, \mathscr{m}), \tag{2}$$

then we say that the game has a value equal to the common value of the RHS and LHS of (2). We denote that value by $V(l, \mathscr{m})$. If for some $\varepsilon > 0$, $\pi^* \in \Pi(l, \mathscr{m})$ and $\eta^* \in H(l, \mathscr{m})$ satisfy

$$\sup_{\eta \in H(l,\mathscr{m})} \max_{(\mathbf{l},\mathscr{m}) \in O(\pi^*,\eta)} P(\mathbf{l}, \mathscr{m}) - \varepsilon \leqq V(l, \mathscr{m}) \tag{3}$$

and

$$\inf_{\pi \in \Pi(l,\mathscr{m})} \min_{(\mathbf{l},\mathscr{m}) \in O(\pi,\eta^*)} P(\mathbf{l}, \mathscr{m}) + \varepsilon \geqq V(l, \mathscr{m}), \tag{4}$$

respectively, then we say that π^* and η^* are ε-optimal pursuit and evasion strategies. We call a strategy optimal if it is ε-optimal for $\varepsilon = 0$. We have the following theorem.

THEOREM 1. *If*

$$\inf_{\pi \in \Pi(l,\mathscr{m})} \sup_{\mathscr{m} \in \mathbf{M}(\mathscr{m})} P(\pi(\mathscr{m}), \mathscr{m}) = \sup_{\eta \in H(l,\mathscr{m})} \inf_{\mathbf{l} \in \mathbf{L}(l)} P(\mathbf{l}, \eta(\mathbf{l})), \tag{2'}$$

then the game has a value which is equal to the common value of the RHS *and* LHS *of* (2′). *Furthermore, if for some* $\varepsilon \geqq 0$, $\pi^* \in \Pi(l, \mathscr{m})$ *and* $\eta^* \in H(l, \mathscr{m})$ *satisfy*

$$P(\pi^*(\mathscr{m}), \mathscr{m}) - \varepsilon \leqq V(l, \mathscr{m}), \qquad \mathscr{m} \in \mathbf{M}(\mathscr{m}), \tag{3′}$$

and

$$P(\mathbf{l}, \eta^*(\mathbf{l})) + \varepsilon \geqq V(l, \mathscr{m}), \qquad \mathbf{l} \in \mathbf{L}(l), \tag{4′}$$

respectively, then π^* *and* η^* *are* ε*-optimal pursuit and evasion strategies.*

Proof. Because of (1), (2′) implies (2), (3′) implies (3), and (4′) implies (4). (Compare with [3, Thm. 1].)

2. The results. This section begins with a statement of our results in Theorems 2, 3 and 4, and Corollaries 1 and 2. It ends with a proof of these theorems based on Lemmas 2, 3, 5 and 6. We prove the latter three in the next section. Then, in § 4, we prove Lemma 2. The arguments which we use to establish that lemma depend on the results of Flynn [4].

Before proceeding, we introduce a definition. We say that the game is in a *radial position* if L lies on the line segment OM. We will restrict our attention to starting positions which are radial. Clearly, L can bring the game into such a position by moving to the center O.

THEOREM 2. *There exists a* $d^* > 0$ *satisfying the following: First, for each radial* $(l, \mathscr{m})$, *there is a strategy* $\pi \in \Pi(l, \mathscr{m})$ *satisfying*

$$P(\pi(\mathscr{m}), \mathscr{m}) \leqq d^* \quad \textit{for all} \ \ \mathscr{m} \in \mathbf{M}(\mathscr{m}). \tag{5}$$

And, second, there exist radial $(l_0, \mathscr{m}_0)$ *such that for every* $\varepsilon > 0$ *there is a strategy* $\eta \in H(l_0, \mathscr{m}_0)$ *satisfying*

$$P(\mathbf{l}, \eta(\mathbf{l})) \geqq d^* - \varepsilon \quad \textit{for every}\ \mathbf{l} \in \mathbf{L}(l_0). \tag{6}$$

Remark 1. Because of Theorem 1, we can identify d^* with the value $V(l_0, \mathscr{m}_0)$. In addition, Theorem 1 implies that any $\pi \in \Pi(l_0, \mathscr{m}_0)$ satisfying (5) is an optimal pursuit strategy while any $\eta \in H(l_0, \mathscr{m}_0)$ satisfying (6) is an ε-optimal evasion strategy. Finally, we note that any radial $(l_0, \mathscr{m}_0)$ satisfying

$$\|l_0\| = d^*/(w - 1), \qquad \|\mathscr{m}_0\| = wd^*/(w - 1) \tag{7}$$

also satisfies (6) (see [4, § 5]).

THEOREM 3. *For every radial* $(l, \mathscr{m})$ *satisfying*

$$\|l\| \geqq d^*/(w - 1), \qquad \|\mathscr{m}\| > wd^*/(w - 1), \tag{8}$$

we have

$$\inf_{\pi \in \Pi(l, \mathscr{m})} \ \sup_{\mathscr{m} \in \mathbf{M}(\mathscr{m})} P(\pi(\mathscr{m}), \mathscr{m}) < d^*. \tag{9}$$

Theorem 3 has as a corollary a result which, in view of Flynn [3], is essential to the justification of this paper. Its proof is left to the reader.

COROLLARY 1. *If the man stays on the boundary, then the lion can force the distance* $|LM| < d^*$.

Hence, staying on the boundary can be a bad strategy for the man. This illustrates a striking difference between *pursuit in the half-plane* and *pursuit in the* *circle.* In the former game, both players can ignore trajectories which lead to an increase in their distance from the boundary (see Flynn [3, § 7] for details).

THEOREM 4. *For every radial* $(l, \mathscr{m})$, *there exists a* $\pi \in \Pi(l, \mathscr{m})$ *such that, for each* $\mathscr{m} \in \mathbf{M}(\mathscr{m})$, *either*

$$\min_{0 \leqq t \leqq 2/\ln(w)} \|\pi(\mathscr{m})(t) - \mathscr{m}(t)\| \leqq d^* \tag{10}$$

or

$$P(\pi(\mathscr{m}), \mathscr{m}) < d^* \tag{11}$$

holds.

The following corollary illustrates an essential difference between the case $w = 1$ and our case ($w > 1$) (see paragraph 2 of the Introduction). Its proof is left to the reader.

COROLLARY 2. *The lion can achieve* d^* *in a finite time.*

Remark 2. Corollary 2 does not assert that the lion has a strategy under which the time required to achieve d^* is bounded.

We postpone the proof of these theorems until the end of this section. Now we introduce an important type of radial position. Call $(l, \mathscr{m}) \in \mathbf{D} \times \mathbf{D}$ a *stable position* if $l = (1/w)\mathscr{m}$. The reason for the term stable is that once such a position occurs, L can keep the game in a stable position by employing the strategy π described by

$$\pi(\mathscr{m})(t) = (1/w)\mathscr{m}(t), \qquad \mathscr{m} \in \mathbf{M}(\mathscr{m}). \tag{12}$$

We group stable positions into equivalent classes. For each ρ satisfying $0 \leqq \rho \leqq (w - 1)/w$, we define $\Sigma(\rho)$ as the set of all stable $(l, \mathscr{m})$ for which $\|l - \mathscr{m}\| = \rho$.

We want to analyze the situation where the game starts in a stable position. This requires more notation. Let ρ satisfy $0 \leqq \rho \leqq (w - 1)/w$ and let $(l, \mathscr{m}) \in \Sigma(\rho)$. Define

$$\begin{aligned} S(\rho) &= \inf_{\pi \in \Pi(l, \mathscr{m})} \sup_{\mathscr{m} \in \mathbf{M}(\mathscr{m})} P(\pi(\mathscr{m}), \mathscr{m}), \\ I(\rho) &= \sup_{\eta \in H(l, \mathscr{m})} \inf_{\mathbf{l} \in \mathbf{L}(l)} P(\mathbf{l}, \eta(\mathbf{l})), \\ \bar{d} &= \sup \{\rho | S(\rho) = \rho\}, \quad \underline{d} = \sup \{\rho | I(\rho) = \rho\}. \end{aligned} \tag{13}$$

The following lemma characterizes the above quantities.

LEMMA 1.

(a) *If* $S(\rho_0) = \rho_0$, *then* $S(\rho) = \rho$ *for* $0 \leqq \rho \leqq \rho_0$. *Similarly, if* $I(\rho) = \rho_0$, *then* $I(\rho) = \rho$ *for* $0 \leqq \rho \leqq \rho_0$.

(b) *If* $S(\rho_2) = \rho_1$ *for* $\rho_1 \neq \rho_2$, *then* $\rho_2 > \rho_1$ *and* $S(\rho_1) = \rho_1$. *Similarly, if* $I(\rho_2) = \rho_1$ *for* $\rho_1 \neq \rho_2$, *then* $\rho_2 > \rho_1$ *and* $I(\rho_1) = \rho_1$.

(c) *Both* S *and* I *are continuous functions.*

(d) S *is a nonincreasing function on* $[\bar{d}, (w - 1)/w]$, *while* I *is a nonincreasing function on* $[\underline{d}, (w - 1)/w]$.

(e) $S \geqq I$ *and* $\bar{d} \geqq \underline{d}$.

Proof. For (a), take any position $(l, m) \in \Sigma(\rho)$, where $\rho < \rho_0$. Let M travel outward at maximum speed along a radial line. Clearly L's best trajectory follows M along this line. Observe that the distance $|LM|$ remains greater than ρ until a position $(l^0, m^0) \in \Sigma(\rho_0)$ is reached. Since by hypothesis $I(\rho_0) = \rho_0$, we have $I(\rho) = \rho$. The proof for S is similar. Parts (b) and (c) follow from analogous arguments while (d) follows directly from (b). Finally, (e) follows from (1).

In § 4, we will prove Lemma 2 which strengthens part (e) of Lemma 1.

LEMMA 2. *The following applies to stable starting positions: if ρ satisfies $\underline{d} < \rho \leqq (w-1)/w$, then $S(\rho) < \underline{d}$. Hence $\underline{d} = \bar{d}$.*

According to Lemma 2, L can do better given a stable starting position where $|LM| > \underline{d}$ than he can at a stable starting position where $|LM| = \underline{d}$. This seems surprising in view of the fact that the original separation is greater in the former case.

DEFINITION 2. Let d^* equal the common value of $\underline{d}$ and $\bar{d}$.

Remark 3. We will show that d^* satisfies the conditions of Theorems 2, 3 and 4, and Corollary 2. Note that the fact that $S(\rho)$ is strictly less than $\underline{d}$ for $\rho > \underline{d}$ is essential to Theorems 3 and 4, and Corollaries 1 and 2.

As we will soon see, L can always force the game into a position which is approximately stable. To express this formally, we need some additional notation. Given $\varepsilon \geqq 0$ and $0 \leqq \rho \leqq (w-1)/w$, define $\Sigma(\rho, \varepsilon)$ as the set of radial positions satisfying

$$\|l\| = \rho/(w-1) \quad \text{and} \quad \big|w\|l\| - \|\mathscr{m}\|\big| \leqq \varepsilon. \tag{14}$$

Certainly, $\Sigma(\rho, 0) = \Sigma(\rho)$. We leave the proof of the following lemma to the reader.

LEMMA 3. *If $(l, \mathscr{m}) \in \Sigma(\rho, \varepsilon)$, then*

$$\begin{aligned} \Big|S(\rho) - \inf_{\pi \in \Pi(l, \mathscr{m})} \sup_{\mathscr{m} \in \mathbf{M}(\mathscr{m})} P(\pi(\mathscr{m}), \mathscr{m})\Big| &\leqq \varepsilon, \\ \Big|I(\rho) - \sup_{\eta \in H(l, \mathscr{m})} \inf_{\mathbf{l} \in \mathbf{L}(l)} P(\mathbf{l}, \eta(\mathbf{l}))\Big| &\leqq \varepsilon. \end{aligned} \tag{15}$$

We think of positions in $\Sigma(\rho, \varepsilon)$ as being *approximately stable*. As shown in Lemma 4, L can force the game into an approximately stable position by staying on the radius OM. This leads to the notion of *radial strategy*. Given any radial $(l, \mathscr{m})$, we call $\pi \in \Pi(l, \mathscr{m})$ a *radial strategy* if $(\pi(\mathscr{m})(t), \mathscr{m}(t))$ is a radial position for all $t \in \mathbf{T}$ and $\mathscr{m} \in \mathbf{M}(\mathscr{m})$. The proof of the next lemma appears in § 3.

LEMMA 4. *Let $(l, \mathscr{m})$ be any radial position satisfying*

$$\|l\| = r_0 < (1/w)\|\mathscr{m}\|, \tag{16}$$

where $0 < r_0 < 1/w$. For each $\varepsilon > 0$, there exists a radial $\pi \in \Pi(l, \mathscr{m})$ such that for every $\mathscr{m} \in \mathbf{M}(\mathscr{m})$,

$$\|\pi(\mathscr{m})(t)\| \geqq r_0 \quad \text{and} \quad \big|w\|\pi(\mathscr{m})(t)\| - \|\mathscr{m}(t)\|\big| \leqq \varepsilon \tag{17}$$

holds for some $t \leqq 2/\ln(1+\varepsilon)$.

In order to prove Theorems 2, 3 and 4 we need two more lemmas. Their proofs appear in the next section.

LEMMA 5. *For any $0 < r \leqq 1/w$ and any radial $(l, \mathscr{m})$, there exists a radial $\pi \in \Pi(l, \mathscr{m})$ such that for each $\mathscr{m} \in \mathbf{M}(\mathscr{m})$ either*

$$\|\pi(\mathscr{m})(t) - \mathscr{m}(t)\| \leqq (w-1)r \tag{18}$$

or

$$(19)\qquad \|\pi(\mathscr{m})(t)\| = r \quad \textit{and} \quad \|\mathscr{m}(t)\| > wr$$

hold for some $t \leqq 2r/\ln w < \infty$.

LEMMA 6. *Let* $0 < r \leqq 1/w$ *and* $0 < \Delta \leqq 1/3$. *If* $(l, \mathscr{m})$ *is any radial position satisfying*

$$(20)\qquad \|l\| = r \quad \textit{and} \quad (1/w)\|\mathscr{m}\| > r + 3\Delta,$$

then there exists a radial $\pi \in \Pi(l, \mathscr{m})$ *satisfying*

$$(21)\qquad \|\pi(\mathscr{m})(\Delta)\| = r + \Delta^2 \quad \textit{and} \quad (1/w)\|\mathscr{m}(\Delta)\| > r + \Delta^2$$

for every $\mathscr{m} \in \mathbf{M}(\mathscr{m})$.

Proof of Theorem 3. Let $(l, \mathscr{m})$ be any radial position satisfying (8). Let

$$\Delta = (1/3)[(1/w)\|\mathscr{m}\| - d^*/(w-1)] > 0.$$

By Lemma 6, there exists a radial $\pi \in \Pi(l, \mathscr{m})$ such that for each $\mathscr{m} \in \mathbf{M}(\mathscr{m})$,

$$\|\pi(\mathscr{m})(\Delta)\| = (d^* + (w-1)\Delta^2)/(w-1) < (1/w)\|\mathscr{m}(\Delta)\|.$$

Define ε by

$$(22)\qquad \varepsilon = (1/3)[d^* - S(d^* + (w-1)\Delta^2)].$$

Lemma 2 and Definition 2 imply that $\varepsilon > 0$. Because of Lemma 4 we can choose our radial π so that

$$(23)\qquad (\pi(\mathscr{m})(t_1), \mathscr{m}(t_1)) \in \Sigma(\rho_1, \varepsilon)$$

for some $t_1 \leqq \Delta + 2/\ln(1+\varepsilon)$ and some $\rho_1 \geqq d^* + (w-1)\Delta^2$.

By (23) and Lemma 3 there exists a policy $\pi^* \in \Pi(l, \mathscr{m})$ such that

$$(24)\qquad P(\pi^*(\mathscr{m}), \mathscr{m}) \leqq S(\rho_1) + 2\varepsilon$$

for every $\mathscr{m} \in \mathbf{M}(\mathscr{m})$. But Lemma 1 implies that

$$(25)\qquad S(\rho_1) \leqq S(d^* + (w-1)\Delta^2).$$

Hence, (22), (24) and (25) give us

$$P(\pi^*(\mathscr{m}), \mathscr{m}) \leqq d^* - \varepsilon < d^*.$$

Theorem 3 follows.

Proof of Theorem 4. Let $(l, \mathscr{m})$ be any radial position. Applying Lemma 5 with $r = d^*/(w-1)$, we find that there exists a radial $\pi \in \Pi(l, \mathscr{m})$ such that, for each $\mathscr{m} \in \mathbf{M}(\mathscr{m})$, either

$$(10)\qquad \min_{0 \leqq t \leqq 2/\ln(w)} \|\pi(\mathscr{m})(t) - \mathscr{m}(t)\| \leqq d^*$$

or

$$(26)\qquad \|\pi(\mathscr{m})(t_0)\| = d^*/(w-1) < (1/w)\|\mathscr{m}(t_0)\|$$

holds for some $t_0 \leqq 2/\ln(w)$. The rest follows from Theorem 3.

Proof of Theorem 2. The first part follows directly from Theorem 4. The second part follows from Lemma 2 and the definition of $I(\rho)$.

3. Radial pursuit strategies. The purpose of this section is to establish Lemmas 4, 5 and 6. Those lemmas deal with a class of strategies which we call radial pursuit strategies; namely, strategies under which L stays on the line segment OM (see the paragraph preceding Lemma 4 for a formal definition). After a few brief remarks, we prove our key result on radial strategies, Lemma 7. We need the latter to prove Lemmas 4 and 5.

Although radial strategies were used by Littlewood [6] and Croft [1] in their work on the case $w = 1$, Gerald J. Smith was the first to apply such strategies to the present problem ($w > 1$), an idea which he attributed to Lester Dubins. Smith showed that if L starts at O, then for each $\varepsilon > 0$ he has a strategy which takes him to a point Q on the line segment OM which satisfies $|OQ| \geqq (1/w)|OM| - \varepsilon$.

Lemma 7, which we prove below, is somewhat stronger. To establish it we make use of the following result which appears in Flynn [3, § 5]. If L starts at O and M is restricted to trajectories $\mathscr{m}$ which satisfy

$$\|\mathscr{m}(t)\| = \rho, \qquad t \in T, \tag{27}$$

for some arbitrary $\rho > 0$, then under the strategy π which satisfies

$$\pi(\mathscr{m})(t) = \begin{cases} \sin(t/\rho)(1/w)\mathscr{m}(t), & 0 \leqq t \leqq \rho\pi/2, \\ (1/w)\mathscr{m}(t), & t \geqq \rho\pi/2, \end{cases} \tag{28}$$

we have $|OL| = (1/w)|OM|$ for $t \geqq \rho\pi/2$ (compare with (12)).

LEMMA 7. *Let $0 < r \leqq 1/w$ and $v > 1$. Given any radial position $(l, \mathscr{m})$, there exists a radial strategy $\pi \in \Pi(l, \mathscr{m})$ such that for each $\mathscr{m} \in \mathbf{M}(\mathscr{m})$, either*

$$\frac{r - \|\pi(\mathscr{m})(t)\|}{r - \|(1/w)\mathscr{m}(t)\|} \leqq v, \quad \|l\| \leqq \|\pi(\mathscr{m})(t)\| < r, \quad \|(1/w)\mathscr{m}(t)\| < r \tag{29}$$

or

$$\|\pi(\mathscr{m})(t)\| = r \leqq \|(1/w)\mathscr{m}(t)\| \tag{30}$$

holds for some $t \leqq 2r/\ln(v)$.

Remark 4. For each $\mathscr{m} \in \mathbf{M}(\mathscr{m})$, let $\hat{\mathscr{m}}$ satisfy

$$\hat{\mathscr{m}}(t) = (1/w)\mathscr{m}(t), \qquad t \in \mathbf{T}. \tag{31}$$

Observe that Lemma 7 is really a statement about the $\hat{\mathscr{m}}$'s which are the trajectories available to M in the case of equal speeds.

Proof of Lemma 7. We assume $\|l\| < r$, since the problem is trivial otherwise. Denote by k the largest integer satisfying

$$\|l\| \geqq r(1 - v^{-k}).$$

Let

$$r_j = r(1 - v^{-(k+j)}), \qquad j = 0, 1, \cdots.$$

Clearly, if $(1/w)\|\mathscr{m}\| \leqq r_1$, then (29) holds for $t = 0$. Hence we can assume $(1/w)\|\mathscr{m}\| > r_1$. We are going to use induction to construct a sequence

$$0 < t_1 \leqq t_2 \leqq \cdots < 2r/\ln(v)$$

and a radial strategy π such that for each $\mathscr{m} \in \mathbf{M}(\mathscr{m})$, either (29) holds for some t_n or (30) holds for $t_\infty \equiv \lim_n t_n \leqq 2r/\ln(v)$.

We will construct π so that if M plays to avoid condition (29) during the time interval $[0, t_n]$, then $|OL| = r_n$ and $|O\hat{M}| > r_{n+1}$ at time t_n. We do this by having L play according to the following rule during the time interval $[t_{n-1}, t_n]$ whenever $|OL| = r_{n-1}$ and $|O\hat{M}| > r_n$ at time t_{n-1}: Move out along the radius OM toward the circle with center O and radius r_n. (We use the mapping determined by relationship (28) to construct the specific trajectory.) Whenever L follows our strategy and M plays to avoid (29) during $[t_{n-1}, t_n]$, we have $|OL| = r_n$ and $|O\hat{M}| > r_{n+1}$ at time t_n. Note that if M always plays to avoid (29), then $|OL| = r$ and $|O\hat{M}| \geqq r$ at time t_∞ (hence (30) holds for t_∞). The details follow.

Select an arbitrary $\mathscr{m} \in \mathbf{M}(\mathscr{m})$. Define

$$\pi(\mathscr{m})(t) = (r_1 \mathscr{m}(t)/\|\mathscr{m}(t)\|)\sin[(t + u_1)/r_1], \qquad 0 \leqq t \leqq t_1,$$

where

$$t_1 = \min\{t |\, \|\pi(\mathscr{m})(t)\| = r_1 \text{ or } \|\hat{\mathscr{m}}(t)\| = r_1\}$$

and

$$u_1 = r_1 \arcsin(\|l\|/r_1).$$

Evidently,

(32) $$t_1 \leqq (\pi/2)r_1 - u_1 \leqq r_1 \arccos(r_0/r_1).$$

Observe that (29) holds for $t = t_1$ unless

$$\|\pi(\mathscr{m})(t_1)\| = r_1 \quad \text{and} \quad \|\hat{\mathscr{m}}(t_1)\| > r_2.$$

Suppose that we have defined $\pi(\mathscr{m})(t)$ on $[0, t_n]$ for some t_n, where n is a positive integer. If π, $\mathscr{m}$ and t_n satisfy (29), let $t_{n+1} = t_n$. On the other hand, if

$$\|\pi(\mathscr{m})(t_n)\| = r_n \quad \text{and} \quad \|\hat{\mathscr{m}}(t_n)\| > r_{n+1},$$

define

$$\pi(\mathscr{m})(t) = (r_{n+1}\mathscr{m}(t)/\|\mathscr{m}(t)\|)\sin[(t + u_{n+1})/r_{n+1}], \quad t_n \leqq t \leqq t_{n+1},$$

where

$$t_{n+1} = \min\{t \geqq t_n |\, \|\pi(\mathscr{m})(t)\| = r_{n+1} \text{ or } \|\hat{\mathscr{m}}(t)\| = r_{n+1}\}$$

and

$$u_{n+1} = r_{n+1} \arcsin(r_n/r_{n+1}) - t_n.$$

Evidently,

$$t_{n+1} \leqq (\pi/2)r_{n+1} - r_{n+1}\arccos(r_n/r_{n+1}) + t_n.$$

Hence

$$t_{n+1} - t_n \leqq r_{n+1} \arccos(r_n/r_{n+1}). \tag{33}$$

Observe that (29) holds for $t = t_{n+1}$ unless

$$\|\pi(\mathscr{m})(t_{n+1})\| = r_{n+1} \quad \text{and} \quad \|\hat{\mathscr{m}}(t_{n+1})\| > r_{n+2}. \tag{34}$$

We want an upper bound on the t_n's. From (32) and (33), we have

$$t_n \leqq \sum_{j=1}^{n+1} r_j \arccos(r_{j-1}/r_j). \tag{35}$$

But the RHS of (35) is less than

$$\sum_{j=1}^{n+1} r \arccos(1 - v^{-(k+j-1)}) < r \int_0^\infty \arccos(1 - v^{-x})\, dx.$$

Now one can show that

$$\int_0^\infty \arccos(1 - v^{-x})\, dx = K/\ln(v),$$

where

$$K = \int_0^{\pi/2} \frac{y \sin y}{1 - \cos y}\, dy < 2.$$

Hence,

$$t_n < 2r/\ln(v), \qquad n = 1, 2, \cdots.$$

We can define

$$t_\infty \equiv \lim_n t_n \leqq 2r/\ln(v).$$

A continuity argument implies that π must satisfy

$$\pi(\mathscr{m})(t_\infty) = \lim_n \pi(\mathscr{m})(t_n).$$

To finish the argument, observe that either (29) holds for some t_n or (34) holds for every t_n. But in the latter case, (30) holds for t_∞. The result follows.

Lemmas 4 and 5 follow easily from Lemma 7.

Proof of Lemma 4. Applying Lemma 7 with $\|l\| = r_0$, $v = 1 + \varepsilon$ and $r = 1/w$, we see there exists a radial $\pi \in \Pi(l, \mathscr{m})$ such that

$$r_0 \leqq \|\pi(\mathscr{m})(t)\| \quad \text{and} \quad 1/w - \|\pi(\mathscr{m})(t)\| \leqq (1 + \varepsilon)(1/w - \|\hat{\mathscr{m}}(t)\|) \tag{36}$$

or

$$\|\pi(\mathscr{m})(t)\| = 1/w = \|\hat{\mathscr{m}}(t)\| \tag{37}$$

holds for some $t \leqq 2/\ln(1 + \varepsilon)$. Evidently (37) implies (17). Suppose (36) holds. Then

$$\begin{aligned} \|w\pi(\mathscr{m})(t) - \mathscr{m}(t)\| &= w(\|\hat{\mathscr{m}}(t)\| - \|\pi(\mathscr{m})(t)\|) \\ &= w[(1/w - \|\pi(\mathscr{m})(t)\|) - (1/w - \|\hat{\mathscr{m}}(t)\|)] \\ &\leqq w\varepsilon(1/w - \|\hat{\mathscr{m}}(t)\|) \leqq \varepsilon. \end{aligned}$$

Hence (36) also implies (17).

Proof of Lemma 5. Applying Lemma 7 with $v = w$, we see that there exists a radial $\pi \in \Pi(l, \mathscr{m})$ and a $t_0 \leqq 2r/\ln(w)$, where (29) or (30) holds for $t = t_0$. Suppose (29) holds. Evidently

$$\begin{aligned}\|\pi(\mathscr{m})(t_0) - \mathscr{m}(t_0)\| &= w\|\hat{\mathscr{m}}(t_0)\| - \|\pi(\mathscr{m})(t_0)\| \\ &= (w-1)r - w(r - \|\hat{\mathscr{m}}(t_0)\|) + (r - \|\pi(\mathscr{m})(t_0)\|) \\ &\leqq (w-1)r.\end{aligned}$$

Hence (31) implies (18). Clearly, if (18) does not hold, then (30) implies (19). Hence (18) or (19) must hold for $t = t_0$.

Proof of Lemma 6. For any $\mathbf{l} \in \mathbf{L}(l)$ and $\mathscr{m} \in \mathbf{M}(\mathscr{m})$,

$$\|\hat{\mathscr{m}}(t)\| \geqq r + 2\Delta > r + \Delta^2 \quad \text{and} \quad \|\mathbf{l}(t)\| \leqq r + \Delta, \qquad 0 \leqq t \leqq \Delta.$$

Consequently,

$$\|\mathbf{l}(t)\|/\|\hat{\mathscr{m}}(t)\| \leqq (r + \Delta)/(r + 2\Delta) \leqq 1 - \Delta, \qquad 0 \leqq t \leqq \Delta.$$

Hence, during the time interval $[0, \Delta]$, L can move out along the radius OM with speed

$$\Delta < \sqrt{1 - (1 - \Delta)^2} \leqq \sqrt{1 - (\|\mathbf{l}(t)\|/\|\hat{\mathscr{m}}(t)\|)^2}.$$

The strategy which allows him to do this satisfies

$$\pi(\mathscr{m})(t) = (r + t\Delta)\mathscr{m}(t)/\|\mathscr{m}(t)\|, \qquad 0 \leqq t \leqq \Delta.$$

Observe that $\pi(\mathscr{m})(\Delta)$ satisfies (21) for every $\mathscr{m} \in \mathbf{M}(\mathscr{m})$.

4. Lemma 2. In this section, we prove Lemma 2. Many of our arguments depend on results which we obtained in [4, § 4]. However, with the exception of the proofs of Lemmas 10 and 11, the material in this section is self-contained.

We begin with some notation and terminology. As in [4], an object's time derivative means its forward or right-hand time derivative. Also, an object's speed refers to the norm of that derivative. Whenever that time derivative exists, we represent it by the usual "dot" notation. Define

$$\mathbf{L}^*(l) = \{\mathbf{l} \in \mathbf{L}(l) | \mathbf{l} \text{ is piecewise linear and } \|\dot{\mathbf{l}}(t)\| = 1, t \in \mathbf{T}\}, \quad l \in \mathbf{D},$$

and

$$\mathbf{L}^* = \bigcup_{l \in \mathbf{D}} \mathbf{L}^*(l).$$

$\mathbf{L}^*$ consists of all trajectories under which L travels at maximum speed along a polygonal path.

We need a definition of convexity for trajectories. If L selects a trajectory $\mathbf{l} \in \mathbf{L}$ for which $\|\mathbf{l}(\cdot)\|$ is bounded away from 0, then as the vector $\mathbf{l}(t)$ rotates about the center O it sweeps out a directed angle. (Following the usual convention, we let positive angles correspond to counterclockwise displacements.) We can represent this angle by a continuous real-valued function $\theta(\cdot)$ on $\mathbf{T}$ (the value $\theta(t)$, of course, represents the directed angle generated between time 0 and time t). We say that the pursuit trajectory $\mathbf{l}$ is *positively oriented* on an interval $\mathbf{T}_0 \subset \mathbf{T}$ if $\theta(\cdot)$ is a nondecreasing function on $\mathbf{T}_0$. We say that $\mathbf{l}$ is *convex* on $\mathbf{T}_0$ if it is positively oriented on $\mathbf{T}_0$ and the set

$$\bigcup_{t_0 \leqq t \leqq t_1} \{l \in \mathbf{D} | l = K\mathbf{l}(t) \text{ for some } K \in [0, 1]\}$$

is convex whenever $\theta(t_1) - \theta(t_0) \leqq \pi$ and $[t_0, t_1] \subset \mathbf{T}_0$. Similar definitions hold for evasion trajectories. (We will omit any specific reference to $\mathbf{T}_0$ whenever its meaning is unambiguous.)

We say that a trajectory $\mathbf{l} \in \mathbf{L}$ is *radially increasing* if $\|\mathbf{l}(\cdot)\|$ is a strictly increasing function on $\mathbf{T}$ (see the paragraph preceding Lemma 8 of [4]). Let $\mathbf{CL}^*(l)$ denote the set of all radially increasing, convex trajectories in $\mathbf{L}^*(l)$.

Given any position where M is not at the center O, denote by s $(-\pi < s \leqq \pi)$ the angle that the directed line segment ML makes away from the direction MO. s is positive when it is measured in a counterclockwise arc. The following lemma is the key result of this section.

LEMMA 8. *Let* $(l^0, m^0) \in \Sigma(\rho_0)$, *where* $\underline{d} < \rho_0 \leqq (w - 1)/w$. *For every* $\delta_0 > 0$, *there exist an* $\varepsilon_0 > 0$, *a* (*finite*) $t_0 \in \mathbf{T}$ *and a pursuit trajectory* $\mathbf{l}^0 \in \mathbf{CL}^*(l^0)$ *such that the set*

$$\{m \in \mathbf{M}(m^0) | \|\mathbf{l}^0(t) - m(t)\| \geqq \underline{d} - \varepsilon_0 \text{ and } s(t) \geqq 0, t \in \mathbf{T}\}$$

is empty and

$$\|\mathbf{l}^0(t_0)\| \geqq 1 - \delta_0. \tag{38}$$

We postpone the proof of Lemma 8 to the end of this section. Our immediate objective is to use it to prove Lemma 2. Our proof of the latter involves constructing a strategy with certain properties. As an aid to that construction, we introduce the concept of a *look-ahead strategy*.

Let $(l, m) \in \mathbf{D} \times \mathbf{D}$ and $\varepsilon > 0$. We call a mapping $\pi: \mathbf{M}(m) \to \mathbf{L}(l)$ an *ε-look-ahead* (pursuit) *strategy* at (l, m) if it satisfies the following: for any $m', m'' \in \mathbf{M}(m)$, $m'(t) = m''(t)$ for $0 \leqq t \leqq t' + \varepsilon$ implies that $\pi(m')(t) = \pi(m'')(t)$ for $0 \leqq t \leqq t'$.

Remark 5. Under an ε-look-ahead pursuit strategy, L can look ε time units in the future. Hence look-ahead strategies are not valid strategies since they violate the information constraint. However, because of the next lemma we can use them to construct strategies with desired properties.

For each $m \in \mathbf{D}$ and $\varepsilon > 0$, define a mapping β_ε on $\mathbf{M}(m)$ by

$$\beta_\varepsilon(m)(t) = \begin{cases} m & \text{for } 0 \leqq t \leqq \varepsilon, \\ m(t - \varepsilon) & \text{for } t \geqq \varepsilon. \end{cases}$$

LEMMA 9. *If* π *is an ε-look-ahead pursuit strategy at* $(l, m) \in \mathbf{D} \times \mathbf{D}$, *then the composition* $\pi \circ \beta_\varepsilon$ *defines a pursuit strategy at* (l, m). *Furthermore,*

$$\left| \sup_{m \in \mathbf{M}(m)} P(\pi(m), m) - \sup_{m \in \mathbf{M}(m)} P(\pi(\beta_\varepsilon(m)), m) \right| \leqq w\varepsilon. \tag{39}$$

Proof. That $\pi \circ \beta_\varepsilon$ defines a strategy follows from the fact that it satisfies the information constraint (see § 1, paragraph 3). Relation (39) holds because of the Lipschitz condition on M's trajectories.

Proof of Lemma 2. Suppose that $(l^0, m^0) \in \Sigma(\rho_0)$, where ρ_0 satisfies

$$\underline{d} < \rho_0 \leqq (w - 1)/w.$$

Let $\delta_0 > 0$, $\varepsilon_0 > 0$, $t_0 \in \mathbf{T}$ and $l^0 \in \mathbf{CL}^*(l^0)$ satisfy the conditions of Lemma 8 and

$$(40) \qquad (1 - \delta_0)^2 + (\underline{d} - \varepsilon_0/2)^2 \geqq 1.$$

To establish the lemma, it is enough to construct an $(\varepsilon_0/3w)$-look-ahead strategy π_0 at $(l^0, \mathscr{m}^0)$ which satisfies

$$(41) \qquad P(\pi_0(\mathscr{m}), \mathscr{m}) \leqq \underline{d} - \varepsilon_0/2$$

for every $\mathscr{m} \in \mathbf{M}(\mathscr{m}^0)$. By Lemma 9, the existence of π_0 implies

$$S(\rho_0) < \underline{d}.$$

The rest of Lemma 2 follows from that inequality and from parts (d) and (e) of Lemma 1. We turn now to the problem of constructing π_0.

Lemma 8 implies that if L follows $\mathbf{l}^0$ while M selects a trajectory keeping $s(t) \geqq 0$, then the distance $|LM| \leqq \underline{d} - \varepsilon_0$ by time t_0. There is, of course, no reason for M to select such a trajectory when L follows $\mathbf{l}^0$, unless L can move so that M derives no advantage from selecting any other kind of trajectory. Let k denote the smallest integer greater than $3wt_0/\varepsilon_0$. We are going to construct an $(\varepsilon_0/3w)$-look-ahead strategy π_0 and a $t_k \in \mathbf{T}$ satisfying

$$(42) \qquad t_0 \leqq t_k < k(\varepsilon_0/3w + 2/\ln(1 + \varepsilon_0/3w)),$$

such that for each positively oriented (see § 4, paragraph 3) $\mathscr{m} \in \mathbf{M}(\mathscr{m}^0)$,

$$(43) \qquad \|\pi_0(\mathscr{m})(t) - \mathscr{m}(t)\| \leqq \underline{d} - \varepsilon_0/2$$

holds for some $t \leqq t_k$. We define π_0 only for positively oriented $\mathscr{m}$. However, this involves no loss of generality since by using symmetry arguments one can extend the domain of π_0 to all of $\mathbf{M}(\mathscr{m}^0)$.

We digress to introduce some additional terminology. The fact that $\mathbf{l}^0$ satisfies the conditions of Lemma 8 implies the following. For each $l \in \mathbf{D}$ satisfying

$$(44) \qquad \|l^0\| < \|l\| < 1 - \delta_0,$$

there is a unique $t_l \in \mathbf{T}$ satisfying

$$\|\mathbf{l}^0(t_l)\| = \|l\|.$$

Let $\mathbf{l}_l^0$ denote the trajectory which one gets by rotating $\mathbf{l}_0$ about O so that the point $\mathbf{l}^0(t_l)$ coincides with l. We define the *continuation* of $\mathbf{l}^0$ at l as the trajectory $\mathbf{C}_l \in \mathbf{L}^*(l)$ which satisfies

$$(45) \qquad \mathbf{C}_l(t) = \mathbf{l}_l^0(t_l + t), \quad t \in \mathbf{T}.$$

Let $(l, \mathscr{m}) \in \mathbf{D} \times \mathbf{D}$ be any position where $s \geqq 0$ and (44) holds. We say that (l, m) is an *ε-inferior position* if the set

$$\{\mathscr{m} \in \mathbf{M}(\mathscr{m}) | \, \|\mathbf{C}_l(t) - \mathscr{m}(t)\| \geqq \underline{d} - \varepsilon \text{ and } s(t) \geqq 0, t \in \mathbf{T}\}$$

is empty.

An important class of radial positions (see paragraph 2 of § 2) are ε_0-inferior. This class includes any stable position $(l, \mathscr{m})$ which satisfies

$$(46) \qquad \|l\| > \|l^0\|.$$

To see this, examine the situation where starting at $(l^0, \mathscr{m}^0)$, L follows the trajectory $\mathbf{l}^0$, while M uses the strategy η_w defined by

$$(47) \qquad \eta_w(\mathbf{l}) = w\mathbf{l}(t), \qquad t \in \mathbf{T}, \quad \mathbf{l} \in \mathbf{L}(l^0).$$

By applying Lemma 8, one can show that any stable position satisfying (46) is ε_0-inferior. A similar argument can be used to show that any radial $(l, \mathscr{m})$ satisfying $\|\mathscr{m}\| \leqq w\|l\|$ and (46) is ε_0-inferior. It follows that any radial $(l, \mathscr{m})$ satisfying

$$(48) \qquad \|l^0\| < \|l\| \leqq \|\mathscr{m}\| \leqq w\|l\| + \varepsilon_0/3w$$

is $(\varepsilon_0/2)$-inferior. We will use this fact.

In defining π_0, we assume that M follows a positively oriented trajectory $\mathscr{m} \in \mathbf{M}(\mathscr{m}^0)$ (see paragraph 3 of this section). We will use induction to define a sequence $\langle t_j \rangle_{j=1}^k$ satisfying

$$(49) \qquad 0 < t_1 < \cdots < t_j < j(\varepsilon_0/3w + 2/\ln(1 + \varepsilon_0/3w)),$$

and to define $\pi_0(\mathscr{m})(\cdot)$ successively on the intervals $[0, t_j]$, $j = 1, \cdots, k$, so that if

$$(50) \qquad \|\pi_0(\mathscr{m})(t) - \mathscr{m}(t)\| \geqq \underline{d} - \varepsilon_0/2$$

holds for all $t \in [0, t_j]$, the position $(\pi_0(\mathscr{m})(t_j), \mathscr{m}(t_j))$ is $(\varepsilon_0/2)$-inferior and satisfies

$$(51) \qquad \|\pi_0(\mathscr{m})(t_j)\| \geqq \|\mathbf{l}^0(j\varepsilon_0/3w)\|.$$

The construction of π_0 finishes the proof, since we can show that (50) cannot hold for all $t \in [0, t_k]$ when π_0 has the above properties. For suppose (50) does hold for all $t \in [0, t_k]$. Then we certainly have

$$(52) \qquad \|\pi_0(\mathscr{m})(t_k) - \mathscr{m}(t_k)\| \geqq \underline{d} - \varepsilon_0/2.$$

By (51) and (38), we also have

$$(53) \qquad \|\pi_0(\mathscr{m})(t_k)\| \geqq 1 - \delta_0.$$

Using a geometric argument, one can show that (40), (52) and (53) imply that

$$(54) \qquad s(t_k) \geqq \pi/2.$$

But using (52), (54) and the fact that $\mathbf{l}_0$ is radially increasing, one can show that the position $(\pi_0(\mathscr{m})(t_k), \mathscr{m}(t_k))$ is not $(\varepsilon_0/2)$-inferior. The lemma then follows by contradiction.

We turn now to the construction of π_0. At time 0 let L compare $\mathscr{m}(\varepsilon_0/3w)$ (M's position at time $\varepsilon_0/3w$) with the position that L would occupy at that time if he followed $\mathbf{l}^0$ (see Remark 5). If the value of s corresponding to these positions is nonnegative, let π_0 direct L along the trajectory $\mathbf{l}^0$ during the time interval $[O, \varepsilon_0/3w]$. In this case, let

$$(55) \qquad t_1 = \varepsilon_0/3w.$$

On the other hand, if the value of s is negative, let π_0 direct L at full speed to a point on the radius through $\mathscr{m}(\varepsilon_0/3w)$ whose distance from L is $\varepsilon_0/3w$ (of the two possible choices he should pick the one farthest from O). Clearly L lies on the radius, through M at time $\varepsilon_0/3w$. The reader should verify that

$$\|\pi(\mathscr{m})(\varepsilon_0/3w)\| \geqq \|\mathbf{l}^0(\varepsilon_0/3w)\|.$$

Evidently if (50) holds for all $t \in [0, \varepsilon_0/3w]$, then the position at time $\varepsilon_0/3w$ is a radial position which is either $(\varepsilon_0/2)$-inferior or satisfies

$$\|\mathbf{l}^0(\varepsilon_0/3)\| > w\|\pi(\mathscr{m})(\varepsilon_0/3)\| + \varepsilon_0/3w$$

(see (48)). If the position at time $\varepsilon_0/3w$ is $(\varepsilon_0/2)$-inferior, let t_1 satisfy (55); otherwise, let L switch to the radial strategy of Lemma 4 at time $\varepsilon_0/3w$. In the latter case, define t_1 as the first time that an $(\varepsilon_0/2)$-inferior position is reached. By (48), (55) and Lemma 4, t_1 satisfies (49). The reader should verify that if (50) holds for all $t \in [0, t_1]$, then $(\pi_0(\mathscr{m})(t_1), \mathscr{m}(t_1))$ is $(\varepsilon_0/2)$-inferior and (51) holds for $j = 1$.

Assume that for some $t_j, j = 1, \cdots, k - 1$, satisfying (49), we define $\pi_0(\mathscr{m})(\cdot)$ on the interval $[0, t_j]$ so that if (50) holds for all $t \in [0, t_j]$, the position $(\pi_0(\mathscr{m})(t_j), \mathscr{m}(t_j))$ is $(\varepsilon_0/2)$-inferior and satisfies (51). We define t_{j+1} and $\pi_0(\mathscr{m})(\cdot)$ on $[t_j, t_{j+1}]$ as follows. At time t_j, let L compare $\mathscr{m}(t_j + \varepsilon_0/3w)$ (M's position at time $t_j + \varepsilon_0/3w$) with the position that L would occupy at that time if he followed the continuation of $\mathbf{l}^0$ at $\pi_0(\mathscr{m})(t_j)$. If the value of s corresponding to these positions is nonnegative, let π_0 direct L along the continuation of $\mathbf{l}^0$ during the time interval $[t_j, t_j + \varepsilon_0/3w]$. In this case, let

$$t_{j+1} = t_j + \varepsilon_0/3w. \tag{56}$$

If the value of s is negative, let π_0 direct L at full speed to the point farthest from O on the radius through $\mathscr{m}(t_j + \varepsilon_0/3w)$ whose distance from L is $\varepsilon_0/3w$. As before, if (50) holds for all $t \in [t_j, t_j + \varepsilon_0/3w]$, then the position at time $t_j + \varepsilon_0/3w$ is a radial position which is either $(\varepsilon_0/2)$-inferior or satisfies

$$\|l^0(t_j + \varepsilon_0/3w)\| > w\|\pi_0(\mathscr{m})(t_j + \varepsilon_0/3w)\| + \varepsilon_0/3w.$$

If the position is $(\varepsilon_0/2)$-inferior, let t_{j+1} satisfy (56); otherwise, let L switch to the radial strategy of Lemma 4. In the latter case, let t_{j+1} denote the first time that an $(\varepsilon_0/2)$-inferior position is reached. By (48), (56), Lemma 4 and the induction hypotheses, t_{j+1} satisfies (49). The reader should verify that if (50) holds for all $t \in [0, t_{j+1}]$, then $(\pi_0(\mathscr{m})(t_{j+1}), \mathscr{m}(t_{j+1}))$ is $(\varepsilon_0/2)$-inferior and (51) holds. The proof of Lemma 2 is complete.

Our remaining objective is Lemma 8. Suppose $(l, \mathscr{m}) \in \Sigma(\rho)$, where $0 \leqq \rho \leqq (w - 1)/w$. For each $\mathbf{l} \in \mathbf{CL}^*(l)$, define

$$\Lambda(\mathbf{l}) = \{\mathscr{m} \in \mathbf{M}(\mathscr{m}) | \|\mathbf{l}(t) - \mathscr{m}(t)\| \geqq \rho \text{ and } s(t) \geqq 0, t \in T\}. \tag{57}$$

For each $\mathscr{m} \in \Lambda(\mathbf{l})$ define $\tau(\cdot, \cdot)$ by

$$\tau(\mathbf{l}, \mathscr{m}) = \inf\{t \in \mathbf{T} | s(t) = \beta\}, \tag{58}$$

where β satisfies

$$\beta = \arcsin(1/w). \tag{59}$$

(See [4, paragraph 4, § 4].) The next two lemmas follow from results of [4].

LEMMA 10. *Given $(l, \mathscr{m}) \in \Sigma(\rho)$ where $0 \leqq \rho \leqq (w - 1)/w$, there exists an $\mathbf{l} \in \mathbf{CL}^*(l)$ for which $\Lambda(\mathbf{l})$ is empty if and only if $\rho > \underline{d}$.*

Proof. The result follows directly from Lemma 10 and Theorem 2 of [4].

LEMMA 11. *Let* $(l, \mathscr{m}) \in \Sigma(\rho)$, *where* $0 \leqq \rho \leqq (w-1)/w$. *If* $\rho \leqq d$, *then for each* $\mathbf{l} \in \mathbf{CL}^*(l)$ *there exists a trajectory* $\mathscr{m}_* \in \Lambda(\mathbf{l})$ *such that the following conditions hold on the interval* $[0, \tau(\mathbf{l}, \mathscr{m}_*)]$:

(i) $\|\mathbf{l}(t) - \mathscr{m}_*(t)\| = \rho$,

(ii) $\|\dot{\mathscr{m}}_*(t)\| = w$,

(iii) $\mathscr{m}_*$ *is convex.*

Proof. Evidently $\rho \leqq \underline{d}$ implies that condition (I) of [4] holds. Hence, by Theorem 2 of [4], M can keep the distance $|LM| \geqq \|l - \mathscr{m}\|$ by following the *isometric rule* (see the paragraph preceding Lemma 7 of [4]). Let $\mathscr{m}_*$ denote the resulting evasion trajectory. Then conditions (i) and (ii) follow directly from the definition of isometric rule, while condition (iii) follows from Lemma 11 of [4].

Suppose $(l, \mathscr{m}) \in \Sigma(\rho)$, where $0 \leqq \rho \leqq \underline{d}$. Let $\mathbf{l} \in \mathbf{CL}^*(l)$ and $\mathscr{m}_* \Lambda(\mathbf{l})$ satisfy the conditions of Lemma 11. (The reader should verify that this implies $\tau(\mathbf{l}, \mathscr{m}_*) < \infty$.) Define $r(\cdot)$ by

$$r(\mathbf{l}) = \|\mathscr{m}_*(\tau(\mathbf{l}, \mathscr{m}_*))\|, \tag{60}$$

where $\tau(\cdot, \cdot)$ satisfies (58). Note that whenever L selects a trajectory $\mathbf{l}$ for which $r(\mathbf{l}) = 1$, he forces $\mathscr{m}_*$ to the boundary at time $\tau(\mathbf{l}, \mathscr{m}_*)$. We have the following lemma.

LEMMA 12. *There exist sequences* $\langle \rho_n \rangle_{n=1}^{\infty} \langle (l^n, \mathscr{m}^n) \rangle_{n=1}^{\infty}$ *and* $\langle \mathbf{l}^n \rangle_{n=1}^{\infty}$ *such that*

$$0 < \rho_n < \rho_{n+1} < \underline{d}, \quad (l^n, \mathscr{m}^n) \in \Sigma(\rho_n), \quad \mathbf{l}^n \in \mathbf{CL}^*(l^n),$$

and

$$r(\mathbf{l}^n) \geqq 1 - 1/n, \qquad n = 1, 2, \cdots. \tag{61}$$

Remark 6. In this paper, we assume that the radius of the arena is equal to 1 (see § 1, paragraph 2). However, other values cause no problems. Let $\underline{d}(k)$ and $\bar{d}(k)$ denote the respective values of $\underline{d}$ and $\bar{d}$ corresponding to a radius of k. One can show that

$$\underline{d}(k) = k\underline{d}(1), \quad \bar{d}(k) = k\bar{d}(1), \qquad 0 \leqq k < \infty. \tag{62}$$

We will use (62) to prove Lemma 12.

Proof of Lemma 12. Let $(l, \mathscr{m}) \in \Sigma(\rho)$, where $0 < \rho \leqq (w-1)/w$. Using (62) and Lemma 10, one can show that

$$r(\mathbf{l}) \leqq 1 - 1/n, \qquad \mathbf{l} \in \mathbf{CL}^*(l),$$

implies that $\underline{d} \geqq \rho/(1 - 1/n)$. We leave the remaining details for the reader.

Proof of Lemma 8. Let $\langle \rho_n \rangle_{n=1}^{\infty}$, $\langle (l^n, \mathscr{m}^n) \rangle_{n=1}^{\infty}$ and $\langle \mathbf{l}^n \rangle$ satisfy the conditions of Lemma 12. Also let n_0 be any integer which satisfies

$$r(\mathbf{l}^{n_0}) \geqq 1 - \delta_0.$$

The reader should verify that to establish Lemma 8, it is enough to show that for some positive integer $k \geqq n_0$, the set

$$\{\mathscr{m} \in \Lambda(\mathbf{l}^k) | \mathscr{m}(t) = w \cdot \mathbf{l}^k(t) \text{ until play reaches a position in } \Sigma(\rho_0)\}$$

is empty.

Suppose that for each positive integer n there exists an $\mathscr{m}^n \in \Lambda(\mathbf{l}^n)$ for which

$$\mathscr{m}^n(t) = w \cdot \mathbf{l}^n(t) \tag{63}$$

holds until play reaches a position in $\Sigma(\rho_0)$. Let $\mathscr{m}_*^n \in \Lambda(\mathbf{l}^n)$ satisfy the conditions of Lemma 11. Let $\mathscr{R}_n(t)$ denote the region bounded by the arc of a circle of radius ρ_n and center $\mathbf{l}^n(t)$, the circumference of $\mathbf{D}$ and the radii through $\mathbf{l}^n(t)$ and $\mathscr{m}_*^n(t)$ (see Fig. 1). Using the convexity of $\mathscr{m}_*^n$, one can establish

$$\mathscr{m}^n(t) \in \mathscr{R}_n(t), \qquad 0 \leqq \tau \leqq \tau(\mathbf{l}^n, \mathscr{m}_*^n). \tag{64}$$

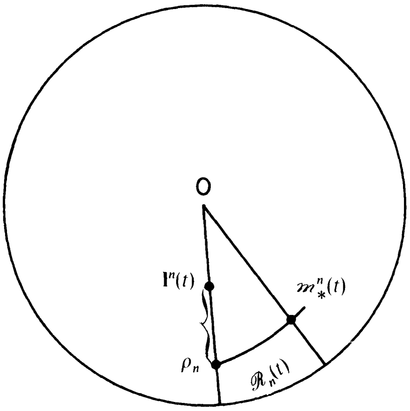

FIG. 1

Let

$$\Delta_n = \sup\{\|l' - l''\| \,|\, l', l'' \in \mathscr{R}_n(\tau(\mathbf{l}^n, \mathscr{m}_*^n))\}.$$

One can use (61) to show that

$$\lim_n \Delta_n = 0. \tag{65}$$

Finally, one can use (63) and the convexity of $\mathscr{m}_*^n$ to show that there exists a constant $K > 0$ such that

$$\|\mathscr{m}_*^n(\tau(\mathbf{l}^n, \mathscr{m}_*^n)) - \mathscr{m}^n(\tau(\mathbf{l}^n, \mathscr{m}_*^n))\| \geqq K, \qquad n = 1, 2, \cdots. \tag{66}$$

The lemma follows from (64), (65), (66) and the triangle inequality.

REFERENCES

[1] H. CROFT, *Lion and man: A postscript*, J. London Math. Soc., 39 (1964), pp. 385–390.

[2] J. FLYNN, *Optimal pursuit strategies for the lion*, Tech. Rep. 174, Statistics Department, Stanford University, Stanford, Calif., 1971.

[3] ———, *Lion and man: The boundary constraint*, this Journal, 11 (1973), pp. 397–411.

[4] ———, *Some results on max-min pursuit*, this Journal, 12 (1974), pp. 53–69.

[5] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

[6] J. LITTLEWOOD, *Lion and Man. A Mathematician's Miscellany*, Methuen, London, 1953.

[7] C. RYLL-NARDZEWSKI, *A theory of pursuit and evasion*, Advances in Game Theory, Princeton University Press, Princeton, 1964, pp. 113–126.

[8] GERALD J. SMITH, *A pursuit evasion game*, unpublished manuscript, 1968.

[9] P. VARAIYA AND J. LIN, *Existence of saddle points in differential games*, this Journal, 7 (1969), pp. 141–157.

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

A GENERAL THEORY OF CONVERGENCE FOR CONSTRAINED OPTIMIZATION ALGORITHMS THAT USE ANTIZIGZAGGING PROVISIONS*

R. KLESSIG†

Abstract. A general convergence theory for algorithms that use an antizigzagging parameter is presented. Certain relationships between this new theory and convergence theories of Polak and Zangwill are established. Applications of the theory are briefly discussed and these applications indicate that the new theory has a wider range of applicability than either the Polak or Zangwill theories.

1. Introduction. Suppose we wish to solve the problem min $\{c(x)|x \in \Omega\}$, where c is a real-valued function and Ω is a constraint set in a normed linear space. Many algorithms for solving such a problem can be analyzed by (1.3.10) of Polak [2] or the theory of Zangwill [7]. Their basic approach is to view such an algorithm as implicitly defining a point-to-set map, call it $\bar{A}$. Thus a sequence $\{x_i\}$, constructed by such an algorithm, satisfies $x_{i+1} \in \bar{A}(x_i)$. In other words, x_{i+1} depends on only x_i. These two theories then specify properties to be satisfied by c, Ω, and $\bar{A}$ that will guarantee that $\{x_i\}$ converge (in some sense) to a solution (or a point satisfying an optimality condition) of the problem. Thus, to analyze an algorithm, c, Ω, and $\bar{A}$ must be identified and then it must be shown that these satisfy the appropriate properties.

Although the Polak and Zangwill theories have been used for analyzing many algorithms, to the best of the author's knowledge they cannot be applied to an important family of algorithms, namely algorithms that use an antizigzagging parameter. As an example of such an algorithm and to facilitate the discussion, consider a variation of an algorithm due to Zoutendijk [8]. The problem to be solved by this particular algorithm is

$$\min \{f^0(x)|f^j(x) \leqq 0, \quad j = 1, \cdots, m\}, \tag{1.1}$$

where $f^j: R^n \to R^1, j = 0, 1, \cdots, m$, are continuously differentiable. We now give the Zoutendijk algorithm modified to use the Armijo [1] step size and applied to (1.1).

1.2. ALGORITHM.

Step 0. Choose any x_0 feasible for (1.1). Select parameters α, β, $\tau \in (0, 1)$ and $\varepsilon_0 > 0$. Select S, a compact neighborhood of the origin in R^n. Set $i = 0$.

Step 1. Compute

$$\theta(\varepsilon_i, x_i) = \min_{h \in S} \{\max \{\langle \nabla f^j(x_i), h\rangle | j \text{ satisfies } f^j(x_i) \geqq -\varepsilon_i \text{ or } j = 0\}\}, \tag{1.3}$$

and let h_i be any solution to this minimax subproblem.

Step 2. If $\theta(\varepsilon_i, x_i) \leqq -\varepsilon_i$, go to Step 3; else set $x_{i+1} = x_i$, set $\varepsilon_{i+1} = \alpha\varepsilon_i$ and go to Step 4.

Step 3. Compute k_i, the smallest nonnegative integer satisfying both

$$f^0(x_i + \beta^{k_i}h_i) - f^0(x_i) - \tau\beta^{k_i}\langle \nabla f^0(x_i), h_i\rangle \leqq 0, \tag{1.4}$$

* Received by the editors July 18, 1972, and in revised form March 29, 1973.

† Bell Laboratories, Holmdel, New Jersey 07733.

$$(1.5)\qquad f^j(x_i + \beta^{k_i}h_i) \leqq 0, \qquad j = 1, \cdots, m.$$

Set $x_{i+1} = x_i + \beta^{k_i}h_i$.

Step 4. Set $i = i + 1$ and go to Step 1.

The parameter ε_i is called the antizigzagging parameter. Its purpose is to allow for consideration of constraints that are almost active (sometimes called tight) in the computation of the search direction h_i. Notice that in order to compute x_{i+1}, both x_i and ε_i must be known. Consequently, a point-to-set map $\bar{A}$, such that $x_{i+1} \in \bar{A}(x_i)$, does not exist for characterizing this algorithm. This means that the Polak and Zangwill theories do not apply in a straightforward manner. In fact, we are not aware of any way of successfully applying these theories.[1]

The theory we present in this paper fills this gap in the existing general convergence theories. In our approach, the antizigzagging parameter is incorporated into the general convergence theory. Thus we have a "search map" A such that $x_{i+1} \in A(\varepsilon_i, x_i)$. In other words, our theory is meant for algorithms where the computation of x_{i+1} requires knowledge of both x_i and ε_i (such as in Algorithm 1.2).

The present work is a refinement of a theory developed by the author and Professor Polak (see (1.3.41) of Polak [2]). The present theory is an improvement in that it is less restrictive. For example, the present theory does not require a one-dimensional minimization and it does not require a compactness assumption. We also believe that it should be much less cumbersome to apply in view of its greater simplicity. Our convergence theory is presented in § 2.

Because we are considering that algorithms are characterized by a map A with $x_{i+1} \in A(\varepsilon_i, x_i)$, it is impossible to directly compare our convergence theory to the theories of Polak and Zangwill. However, certain relationships do exist and we discuss them in § 3.

Finally, in § 4, we consider some algorithms that can be analyzed by our theory. We find that for some algorithms which use an antizigzagging parameter, the antizigzagging parameter need not be included in our convergence theory. That is, a simplified form of our theory can be used to analyze these algorithms.

2. Algorithm prototypes and convergence results. In this section, we present our algorithm prototype for algorithms with an antizigzagging parameter. Our major convergence result is established in Theorem 2.22.

Consider the problem $\inf\{c(x)|x \in \Omega\}$, where c is a real-valued function and Ω is a constraint set in a normed linear space. Usually we can represent a necessary condition of optimality for this problem by a function $\psi: R^+ \times \Omega \to R^-$. That is, if $\hat{x}$ solves the problem, then $\psi(0, \hat{x}) = 0$. (For problem (1.1), $\psi = \theta$.) In proving such a condition, we usually assume that $\psi(0, \hat{x}) < 0$ and then use this fact to construct a feasible point x^* such that $c(x^*) < c(\hat{x})$ (a contradiction). Such a proof is very suggestive of a computational method and this idea is the basis of our general convergence theory.

2.1. *Notation.* Throughout the remainder of our presentation, the following notation will be employed. For z in a normed linear space and $\rho > 0$, we define

[1] It is possible to define the map $\bar{A}$ such that $(\varepsilon_i, x_i) \in \bar{A}(\varepsilon_i, x_i)$ but $\bar{A}$ cannot be shown to have the properties required by the two theories.

$B(z, \rho) \triangleq \{z' \mid \|z' - z\| \leqq \rho\}$, where the particular normed linear space should be clear from the context.

2.2. Algorithm Prototype. Let $(\mathscr{X}, \|\cdot\|)$ be a normed linear space with a closed set Ω. Let ψ, A, b, and c be maps; $\psi : R^+ \times \Omega \to R^-$, $A : R^+ \times \Omega \to 2^{\Omega}$, $b : (0, \infty) \to (0, \infty)$, and $c : \Omega \to R$. The algorithm prototype which finds a point $\hat{x}$ satisfying an optimality condition, $\psi(0, \hat{x}) = 0$, is as follows.

Step 0. Choose any $x_0 \in \Omega$ and choose parameters $\alpha \in (0, 1)$ and $\varepsilon_0 > 0$. Set $i = 0$.

Step 1. If $\psi(\varepsilon_i, x_i) \leqq -b(\varepsilon_i)$, go to Step 2; else set $x_{i+1} = x_i$, set $\varepsilon_{i+1} = \alpha\varepsilon_i$ and go to Step 3.

Step 2. Compute any $y \in A(\varepsilon_i, x_i)$. Set $x_{i+1} = y$ and set $\varepsilon_{i+1} = \varepsilon_i$.

Step 3. Set $i = i + 1$ and go to Step 1.

2.3. *Remark.* We have given (2.2) no provision for stopping in order to simplify the statements of the following lemmas and theorems. Since we are seeking points x that satisfy $\psi(0, x) = 0$, an obvious stopping rule is to stop when $\psi(0, x_i) = 0$. If this stopping rule is added to (2.2), the statements of Lemma 2.8 and Theorem 2.22 must be altered in an obvious way to account for the possibility that (2.2) generates a finite sequence.

2.4. *Assumptions.* In order to establish the major results of this section, we need the following assumptions:

(i) c is either continuous or bounded from below on Ω, $A(\varepsilon, x) \neq \emptyset$ for all $(\varepsilon, x) \in R^+ \times \Omega$, and $c(x') - c(x) \leqq 0$ for all $x' \in A(\varepsilon, x)$, for all $(\varepsilon, x) \in R^+ \times \Omega$.

(ii) $\lim_{\varepsilon \to 0} b(\varepsilon) = 0$.

(iii) ψ is upper semicontinuous on $\{0\} \times \Omega$.

(iv) Given any $\mu > 0$ and any $x \in \Omega$ satisfying $\psi(0, x) < 0$, there exist constants $\delta(\mu, x) > 0$ and $\gamma(\mu, x) > 0$ (possibly depending on μ and x) with the following property. If $\varepsilon' > 0$ and $x' \in B(x, \gamma(\mu, x)) \cap \Omega$ are such that $\psi(\varepsilon', x') \leqq -\mu$, then $c(x'') - c(x') \leqq -\delta(\mu, x)\|x'' - x'\|$ for all $x'' \in A(\varepsilon', x')$.

(v) Given any $\varepsilon > 0$ and any $x \in \Omega$ satisfying $\psi(0, x) < 0$, there exist constants $\sigma(\varepsilon, x) > 0$ and $\kappa(\varepsilon, x) > 0$ (possibly depending on ε and x) with the following property. If $x' \in B(x, \kappa(\varepsilon, x)) \cap \Omega$ is such that $\psi(\varepsilon, x') \leqq -b(\varepsilon)$, then $\|x'' - x'\| \geqq \sigma(\varepsilon, x)$ for all $x'' \in A(\varepsilon, x')$.

Assumption 2.4 (iv) says that if x_i is in a neighborhood of a point x that does not satisfy the optimality condition, the decrease in the cost function is at least proportional to $\|x_{i+1} - x_i\|$. In terms of Algorithm 1.2, this assumption means that h_i must be a direction of descent and β^{k_i} cannot be too large (see (1.4)). Assumption 2.4 (v) says that $\|x_{i+1} - x_i\|$ will be no smaller than $\sigma(\varepsilon_i, x)$. In terms of Algorithm 1.2, this assumption means that the step size β^{k_i} cannot be arbitrarily small so long as ε_i is bounded away from zero.

The following theorem shows that ψ must represent a necessary condition of optimality for the problem $\inf\{c(x) | x \in \Omega\}$.

2.5. Theorem. *Suppose that Assumptions* 2.4 (i) *to* 2.4 (v) *hold.* If $\hat{x}$ *solves* $\inf\{c(x) | x \in \Omega\}$, *then* $\psi(0, \hat{x}) = 0$.

Proof. It will suffice to show that $\psi(0, \hat{x}) < 0$ implies that $\hat{x}$ does not solve the problem $\inf\{c(x) | x \in \Omega\}$. If $\psi(0, \hat{x}) < 0$, by (2.4)(ii) and (2.4)(iii) there exists $\bar{\varepsilon} > 0$ such that $\psi(\bar{\varepsilon}, \hat{x}) \leqq -b(\bar{\varepsilon})$. Let $\delta(b(\bar{\varepsilon}), \hat{x}) > 0$ and $\sigma(\bar{\varepsilon}, \hat{x}) > 0$ be as in (2.4)(iv) and (2.4)(v) respectively. Then applying (2.4)(i), (2.4)(iv), and (2.4)(v) we find that there exists $\tilde{x} \in A(\bar{\varepsilon}, \hat{x}) \subset \Omega$ such that

$$\begin{aligned} c(\tilde{x}) - c(\hat{x}) &\leqq -\delta(b(\bar{\varepsilon}), \hat{x}) \|\tilde{x} - \hat{x}\| \\ &\leqq -\delta(b(\bar{\varepsilon}), \hat{x})\sigma(\bar{\varepsilon}, \hat{x}) \\ &< 0. \end{aligned} \tag{2.6}$$

Inequality (2.6) implies that $\hat{x}$ does not solve the problem $\inf\{c(x) | x \in \Omega\}$ and the proof is complete.

In the sequel, we make use of the following, easy-to-establish lemma.

2.7. LEMMA. *Let* $(x_i\}$ *be a sequence constructed by Algorithm Prototype* 2.2 *and suppose that Assumption* 2.4(i) *holds. If* $\{x_i\}$ *has an accumulation point, then there exists* $T > -\infty$ *such that* $c(x_i) \geqq T$ *for* $i = 0, 1, \cdots$.

We establish our major convergence result by making use of the following lemma.

2.8. LEMMA. *Let* $\{x_i\}$ *and* $\{\varepsilon_i\}$ *be sequences constructed by Algorithm Prototype* 2.2. *Suppose that Assumptions* 2.4(i) *to* 2.4(iv) *hold and suppose that* $\varepsilon_i \to 0$. *If* x^* *is an accumulation point of* $\{x_i\}$, *then it satisfies* $\psi(0, x^*) = 0$.

Proof. Let x^* be an accumulation point of $\{x_i\}$. We consider two cases.

Case 1. $x_i \to x^*$. Since $\varepsilon_i \to 0$, it is clear from Step 1 of (2.2) that there exists a subsequence $\{x_i\}_{i \in K}$ where $K \subset \{0, 1, \cdots\}$ is an infinite index set such that $\psi(\varepsilon_i, x_i) > -b(\varepsilon_i)$ for all $i \in K$. Thus from (2.4)(ii) we have that

$$0 = \overline{\lim_{i \in K}} - b(\varepsilon_i) \leqq \overline{\lim_{i \in K}} \psi(\varepsilon_i, x_i). \tag{2.9}$$

Since $x_i \to x^*$, we must have that $x_i \underset{K}{\to} x^*$. Thus, since (2.4)(iii) holds,

$$0 \leqq \overline{\lim_{i \in K}} \psi(\varepsilon_i, x_i) \leqq \psi(0, x^*), \tag{2.10}$$

and since ψ is nonpositive-valued, (2.10) implies that $\psi(0, x^*) = 0$.

Case 2. $x_i \not\to x^*$. We assume that $\psi(0, x^*) = -2\mu < 0$ and establish a contradiction to Lemma 2.7. By (2.4)(ii) and (2.4)(iii) there exist $\bar{\varepsilon} > 0$ and $\bar{\gamma} > 0$ such that

$$\psi(\varepsilon', x') \leqq -\mu \leqq -b(\varepsilon') \tag{2.11}$$

for all $\varepsilon' \in [0, \bar{\varepsilon}]$, and for all $x' \in B(x^*, \bar{\gamma}) \cap \Omega$. Let $\delta(\mu, x^*) > 0$ and $\gamma(\mu, x^*) > 0$ be as in (2.4)(iv). Since $x_i \not\to x^*$, there exist a $\rho > 0$ and a subsequence indexed by K_1 such that

$$\rho \leqq \tfrac{1}{2} \min\{\bar{\gamma}, \gamma(\mu, x^*)\}, \tag{2.12}$$

$$x_i \notin B(x^*, 2\rho) \quad \text{if and only if} \quad i \in K_1. \tag{2.13}$$

On the other hand, since x^* is an accumulation point of $\{x_i\}$, there must exist a subsequence indexed by K_2 such that

$$x_i \in B(x^*, \rho) \quad \text{if and only if} \quad i \in K_2. \tag{2.14}$$

Now we define two maps; $n: K_2 \to K_1$ and $m: K_2 \to K_2$ by

$$n(i) \triangleq \min\{j \in K_1 | j \geqq i + 1\}, \tag{2.15}$$

$$m(i) \triangleq \min\{j \in K_2 | j \geqq n(i) + 1\}. \tag{2.16}$$

Thus if $x_q \in B(x^*, \rho)$, then $x_{n(q)}$ is the first element following x_q in the sequence $\{x_i\}$ that is not in $B(x^*, 2\rho)$. Also, if $x_q \in B(x^*, \rho)$, $x_{m(q)}$ is the first element following $x_{n(q)}$ in the sequence $\{x_i\}$ that is in $B(x^*, \rho)$. Now since $\varepsilon_i \to 0$, there exists $p \in K_2$ such that $\varepsilon_j \leqq \bar{\varepsilon}$ for $j = p, p + 1, \cdots$. Consider the subsequence $\bar{K} \triangleq \{p, m(p), m(m(p)), \cdots\} \subset K_2$. Since $x_i \in B(x^*, 2\rho)$ for $i = k, k + 1, \cdots, n(k) - 1$, for any $k \in \bar{K}$, we obtain, from (2.11), (2.12), and the choice of p, that

$$\psi(\varepsilon_i, x_i) \leqq -\mu \leqq -b(\varepsilon_i), \qquad i = k, k + 1, \cdots, n(k) - 1, \tag{2.17}$$

for all $k \in \bar{K}$. Consequently, Step 2 is reached for $i = k, k + 1, \cdots, n(k) - 1$, for all $k \in \bar{K}$. But then, applying (2.12), (2.17), and (2.4)(iv) we have that

$$c(x_{i+1}) - c(x_i) \leqq -\delta(\mu, x^*) \|x_{i+1} - x_i\|, \qquad i = k, k + 1, \cdots, n(k) - 1, \tag{2.18}$$

for all $k \in \bar{K}$. By (2.4)(i), $\{c(x_i)\}$ is a nonincreasing sequence and thus (2.18) yields

$$\begin{aligned} c(x_{m(k)}) - c(x_k) &\leqq \sum_{q=k}^{n(k)-1} (c(x_{q+1}) - c(x_q)) \\ &\leqq -\delta(\mu, x^*) \sum_{q=k}^{n(k)-1} \|x_{q+1} - x_q\| \end{aligned} \tag{2.19}$$

for all $k \in \bar{K}$. But by (2.15), $x_k \in B(x^*, \rho)$ and $x_{n(k)} \notin B(x^*, 2\rho)$ for all $k \in \bar{K}$. Hence

$$\rho < \|x_{n(k)} - x_k\| \leqq \sum_{q=k}^{n(k)-1} \|x_{q+1} - x_q\| \tag{2.20}$$

for all $k \in \bar{K}$. Thus (2.19) and (2.20) yield

$$c(x_{m(k)}) - c(x_k) \leqq -\delta(\mu, x^*)\rho < 0 \tag{2.21}$$

for all $k \in \bar{K}$. Hence $c(x_k) \underset{\bar{K}}{\to} -\infty$ which contradicts Lemma 2.7; the proof is complete.

Lemma 2.8 says the following. If we use an algorithm of the form of (2.2) which satisfies (2.4)(i) to (2.4)(iv) to generate sequences $\{x_i\}$ and $\{\varepsilon_i\}$, we know that if $\varepsilon_i \to 0$, then any accumulation point x^* satisfies the optimality condition, $\psi(0, x^*) = 0$. Unfortunately, this knowledge is of dubious value since it does not give us any a priori guarantee that the algorithm will "work". All that we know is that if we try to use the algorithm, we need only monitor $\{\varepsilon_i\}$ in order to determine if the accumulation points of $\{x_i\}$ satisfy the optimality condition. The following theorem indicates how this difficulty may be overcome.

2.22. THEOREM. *Let $\{x_i\}$ be a sequence constructed by Algorithm Prototype* 2.2 *and suppose that Assumptions* 2.4(i) to 2.4(v) *hold. Then if x^* is an accumulation point of $\{x_i\}$, it satisfies* $\psi(0, x^*) = 0$.

Proof. Let x^* be an accumulation point of $\{x_i\}$. We consider two cases.

Case 1. $\varepsilon_i \to 0$. The theorem follows immediately from Lemma 2.8 for this case.

Case 2. $\varepsilon_i \not\to 0$. We assume that $\psi(0, x^*) < 0$ and establish a contradiction to Lemma 2.7. By the way $\{\varepsilon_i\}$ is constructed, there must exist $\varepsilon^* > 0$ and an integer

$N_1 \geqq 0$ such that $\varepsilon_i = \varepsilon^* > 0$ for all $i \geqq N_1$. But then because of Step 1 of (2.2),

$$\psi(\varepsilon_i, x_i) = \psi(\varepsilon^*, x_i) \leqq -b(\varepsilon_i) = -b(\varepsilon^*) \tag{2.23}$$

for all $i \geqq N_1$. Now let $\delta(b(\varepsilon^*), x^*) > 0$ and $\gamma(b(\varepsilon^*), x^*) > 0$ be as in (2.4)(iv) and let $\sigma(\varepsilon^*, x^*) > 0$ and $\kappa(\varepsilon^*, x^*) > 0$ be as in (2.4)(v). If K indexes a subsequence such that $x_i \xrightarrow[K]{} x^*$, then from (2.23) there must exist an integer $N_2 \geqq N_1$ such that

$$x_i \in B(x^*, \gamma(b(\varepsilon^*), x^*)) \cap \Omega, \qquad \psi(\varepsilon_i, x_i) \leqq -b(\varepsilon^*), \tag{2.24}$$

$$x_i \in B(x^*, \kappa(\varepsilon^*, x^*)) \cap \Omega, \qquad \psi(\varepsilon^*, x_i) \leqq -b(\varepsilon^*), \tag{2.25}$$

for all $i \in K$ and $i \geqq N_2$. But this means that (2.4)(iv) and (2.4)(v) can be applied to obtain

$$\begin{aligned} c(x_{i+1}) - c(x_i) &\leqq -\delta(b(\varepsilon^*), x^*)\|x_{i+1} - x_i\| \\ &\leqq -\delta(b(\varepsilon^*), x^*)\sigma(\varepsilon^*, x^*) < 0 \end{aligned} \tag{2.26}$$

for all $i \in K$ and $i \geqq N_2$. Since $\{c(x_i)\}$ is nonincreasing, (2.26) implies that $c(x_i) \to -\infty$ which contradicts Lemma 2.7. The proof is complete.

In many algorithms (for example, (1.2)), the computation of the quantity corresponding to $\psi(\varepsilon_i, x_i)$ is the major part of each iteration. Thus, even though $\psi(\varepsilon_i, x_i) > -b(\varepsilon_i)$, one may not want to "throw away" the information gained in computing the optimality condition. Consequently, these algorithms try to compute a point, x_{i+1}, of lower cost even when $\psi(\varepsilon_i, x_i) > -b(\varepsilon_i)$. For example, compare (1.2) with the ε-perturbation algorithm of Zangwill [7]. This idea can be easily incorporated into Algorithm Prototype 2.2.

2.27. *Remark.* Algorithm Prototype 2.2 can be slightly generalized by replacing Step 2 with Step 2′.

Step 2′. If $\psi(\varepsilon_i, x_i) \leqq -b(\varepsilon_i)$, go to Step 2; else choose any $y \in \Omega$ such that $c(y) \leqq c(x_i)$, set $x_{i+1} = y$, set $\varepsilon_{i+1} = \alpha\varepsilon_i$, and go to Step 3.

Theorem 2.22 and its proof remain unchanged under this modification.

3. Relationships to the Polak and Zangwill theories. As mentioned in § 1, our new convergence theory cannot be directly compared to the Polak or Zangwill theories. However, certain relationships exist and we examine these relationships in this section.

Polak has modified Algorithm 1.2 by changing the way in which the antizigzagging parameter is manipulated. (See (4.3.26) in Polak [2].) Given x_i, his algorithm tries values $\varepsilon_0, \alpha\varepsilon_0, \alpha^2\varepsilon_0, \cdots$ for the antizigzagging parameter until $\theta(\alpha^k\varepsilon_0, x_i) \leqq -\alpha^k\varepsilon_0$ at which time x_{i+1} is computed. This has the effect of making the computation of x_{i+1} dependent only on x_i. With this modification, Polak can apply his convergence theory. If we change the manipulation of ε in Algorithm Prototype 2.2 in the same way, a new prototype results.

3.1. ALGORITHM PROTOTYPE. Let $(\mathscr{X}, \|\cdot\|)$, Ω, ψ, A, b, and c be as in (2.2).

Step 0. Choose any $x_0 \in \Omega$ and choose parameters $\alpha \in (0, 1)$ and $\varepsilon_0 > 0$. Set $i = 0$.

Step 1. Set $\varepsilon = \varepsilon_0$.

Step 2. If $\psi(\varepsilon, x_i) \leqq -b(\varepsilon)$, go to Step 4; else go to Step 3.

Step 3. Set $\varepsilon = \alpha\varepsilon$ and go to Step 2.

Step 4. Compute any $y \in A(\varepsilon, x_i)$. Set $x_{i+1} = y$.

Step 5. Set $i = i + 1$ and go to Step 1.

We establish a convergence result very similar to Theorem 2.22. To do this we apply (1.3.10) of Polak [2].

3.2. THEOREM. *Let* $\{x_i\}$ *be the sequence constructed by Algorithm Prototype* 3.1 *and suppose that* (2.4)(i) *to* (2.4)(v) *hold. If* $\{x_i\}$ *is finite with the last element* x_s, *because* (3.1) *cycled indefinitely between Step* 1 *and Step* 2, *then* x_s *satisfies* $\psi(0, x_s) = 0$. *If* $\{x_i\}$ *is infinite with accumulation point* x^*, *then* x^* *satisfies* $\psi(0, x^*) = 0$.

Proof. In the case of $\{x_i\}$ finite, it is clear that

$$\psi(\alpha^k\varepsilon_0, x_s) > -b(\alpha^k\varepsilon_0), \qquad k = 0, 1, \cdots. \tag{3.3}$$

Because of (2.4)(ii) and (2.4)(iii),

$$\psi(0, x_s) \geqq \overline{\lim_{k\to\infty}}\, \psi(\alpha^k\varepsilon_0, x_s) \geqq \overline{\lim_{k\to\infty}}\, -b(\alpha^k\varepsilon_0) = 0, \tag{3.4}$$

and thus $\psi(0, x_s) = 0$.

For the case of $\{x_i\}$ infinite, we apply (1.3.10) of Polak [2]. Let $\bar{A}: \Omega \to 2^{\Omega}$ denote the point-to-set map defined by Step 1 through Step 4 in (3.1). We must show that given any $x \in \Omega$ such that $\psi(0, x) < 0$, there exist $\bar{\gamma} > 0$ and $\bar{\rho} > 0$ such that

$$c(x'') - c(x') \leqq -\bar{\rho} \quad \text{for all } x'' \in \bar{A}(x'), \quad x' \in B(x, \bar{\gamma}) \cap \Omega. \tag{3.5}$$

Let $x \in \Omega$ be such that $\psi(0, x) < 0$. By (2.4)(ii) and (2.4)(iii) there exist $\gamma_1 > 0$ and $\varepsilon_1 > 0$ such that

$$\psi(\varepsilon', x') \leqq -b(\varepsilon') \quad \text{for all } \varepsilon' \in [0, \varepsilon_1], \quad x' \in B(x, \gamma_1) \cap \Omega. \tag{3.6}$$

Let k be the smallest nonnegative integer such that $\alpha^k\varepsilon_0 < \varepsilon_1$. Let $\gamma(b(\alpha^j\varepsilon_0), x)$ and $\delta(b(\alpha^j\varepsilon_0), x)$, $j = 0, 1, \cdots, k$, be as in (2.4)(iv). Let $\kappa(\alpha^j\varepsilon_0, x)$ and $\sigma(\alpha^j\varepsilon_0, x)$, $j = 0, 1, \cdots, k$, be as in (2.4)(v). Define

$$\begin{aligned} \bar{\gamma} \triangleq \min \{\gamma(b(\alpha^j\varepsilon_0), x), \quad & j = 0, 1, \cdots, k; \\ \kappa(\alpha^j\varepsilon_0, x), \quad & j = 0, 1, \cdots, k\} > 0; \end{aligned} \tag{3.7}$$

$$\bar{\delta} \triangleq \min \{\delta(b(\alpha^j\varepsilon_0), x), \quad j = 0, 1, \cdots, k\} > 0; \tag{3.8}$$

$$\bar{\sigma} \triangleq \min \{\sigma(\alpha^j\varepsilon_0, x), \quad j = 0, 1, \cdots, k\} > 0. \tag{3.9}$$

Now, let x' be any point in $B(x, \bar{\gamma}) \cap \Omega$. Because of (3.6), $\bar{A}(x') = A(\alpha^{j'}\varepsilon_0, x')$ with $j' \in \{0, 1, \cdots, k\}$. Thus it follows from (2.4)(iv), (3.7), and (3.8) that

$$\begin{aligned} c(x'') - c(x') &\leqq -\delta(b(\alpha^{j'}\varepsilon_0), x)\|x'' - x'\| \\ &\leqq -\bar{\delta}\|x'' - x'\| \quad \text{for all } x'' \in \bar{A}(x'). \end{aligned} \tag{3.10}$$

Furthermore, it follows from (2.4)(v), (3.7), and (3.9) that

$$\|x'' - x'\| \geqq \sigma(\alpha^{j'}\varepsilon_0, x) \geqq \bar{\sigma} \quad \text{for all } x'' \in \bar{A}(x'). \tag{3.11}$$

Since $x' \in B(x, \bar{\gamma}) \cap \Omega$ was arbitrary, (3.10) and (3.11) combine to yield (3.5) with $\bar{\rho} = \bar{\delta}\bar{\sigma}$. This completes the proof.

Since we can use (1.3.10) of Polak [2] to prove convergence of Algorithm Prototype 3.1, this prototype, Assumptions 2.4, and Theorem 3.2 form a convergence theory that is a special case of Polak's theory. Furthermore, since (3.5) is sufficient to establish the hypotheses of Convergence Theorem D in Zangwill [7], the convergence theory of this section is a special case of Zangwill's Theorem D.

There is an important consequence of Algorithm Prototype 3.1 and Theorem 3.2. If we have an algorithm that can be analyzed by means of the convergence theory of § 2, then we have identified Ω, ψ, A, c, and b and we have shown that these satisfy Assumptions 2.4. If we rearrange this algorithm so that it fits the structure of Algorithm Prototype 3.1, a new algorithm results with the convergence properties specified by Theorem 3.2. For example, we can analyze Algorithm 1.2 by use of the convergence theory of § 2 and the rearrangement implied by Algorithm Prototype 3.1 yields (4.3.26) in Polak [2]. In a sense, we get two algorithms for the price of one.

Algorithm Prototype 2.2 is specifically constructed to analyze algorithms that use an antizigzagging parameter. In other words, it is meant to apply to algorithms where the computation of x_{i+1} depends on both ε_i and x_i. However, it will also apply to algorithms where x_{i+1} depends on only x_i in which case the ε_i in (2.2) is superfluous. That is, $\psi:\Omega \rightarrow R^-$ and $A:\Omega \rightarrow 2^{\Omega}$, and it is easy to see that (2.2) and Assumptions 2.4 reduce to the following.

3.12. ALGORITHM PROTOTYPE. Let $(\mathscr{X}, \|\cdot\|)$ be a normed linear space with a closed set Ω. Let ψ, A, and c be maps $\psi:\Omega \rightarrow R^-$, $A:\Omega \rightarrow 2^{\Omega}$, and $c:\Omega \rightarrow R$.

Step 0. Choose any $x_0 \in \Omega$. Set $i = 0$.

Step 1. If $\psi(x_i) = 0$, set $x_{i+1} = x_i$ and go to Step 3.[2]

Step 2. Compute any $y \in A(x_i)$. Set $x_{i+1} = y$.

Step 3. Set $i = i + 1$ and go to Step 1.

3.13. ASSUMPTIONS.

(i) c is either continuous or bounded from below on Ω, $A(x) \neq \emptyset$ for all $x \in \Omega$, and $c(x') - c(x) \leqq 0$ for all $x' \in A(x)$ for all $x \in \Omega$.

(ii) ψ is upper semicontinuous on Ω.

(iii) Given any $\mu > 0$ and any $x \in \Omega$ satisfying $\psi(x) < 0$, there exist constants $\delta(\mu, x) > 0$ and $\gamma(\mu, x) > 0$ (possibly depending on μ and x) with the following property. If $x' \in B(x, \gamma(\mu, x)) \cap \Omega$ is such that $\psi(x') \leqq -\mu$, then $c(x'') - c(x') \leqq -\delta(\mu, x)\|x'' - x'\|$ for all $x'' \in A(x')$.

(iv) Given any $x \in \Omega$ satisfying $\psi(0, x) < 0$, there exist constants $\sigma(x) > 0$ and $\kappa(x) > 0$ (possibly depending on x) with the following property. If $x' \in B(x, \kappa(x)) \cap \Omega$ is such that $\psi(x') < 0$, then $\|x'' - x'\| > \sigma(x)$ for all $x'' \in A(x')$.

The following lemma shows that the convergence properties of (3.12) and (3.13) can be established using the Polak theory.

3.14. LEMMA. *Suppose* (3.13)(ii), (3.13)(iii), *and* (3.13)(iv) *hold. Given any* $x \in \Omega$ *such that* $\psi(x) < 0$, *there exist constants* $\bar{\gamma}(x) > 0$ *and* $\bar{\rho}(x) > 0$ (*possibly depending on* x) *such that*

$$(3.15) \qquad c(x'') - c(x') \leqq -\bar{\rho}(x) \quad \textit{for all} \quad x'' \in A(x'), \quad x' \in B(x, \bar{\gamma}(x)) \cap \Omega.$$

Proof. Let $x \in \Omega$ be such that $\psi(x) = -2\mu < 0$. From (3.13)(ii), there exists $\gamma_1 > 0$ such that

$$(3.16) \qquad \psi(x') \leqq -\mu \quad \text{for all } x' \in B(x, \gamma_1) \cap \Omega.$$

[2] Obviously, if $\psi(x_i) = 0$, we should stop. We have put Step 1 in this present form so that we will not have to consider the trivial finite sequence case in the sequel.

Let $\delta(\mu, x) > 0$ and $\gamma(\mu, x) > 0$ be as in (3.13)(iii). Let $\sigma(x) > 0$ and $\kappa(x) > 0$ be as in (3.13)(iv). Define

$$\bar{\gamma}(x) \triangleq \min\{\gamma_1, \gamma(\mu, x), \kappa(x)\} > 0, \tag{3.17}$$

$$\bar{\rho}(x) \triangleq \delta(\mu, x)\sigma(x). \tag{3.18}$$

It follows easily from (3.13)(iii) and (3.13)(iv) that (3.15) holds for $\bar{\gamma}(x)$ and $\bar{\rho}(x)$ as defined in (3.17) and (3.18). This completes the proof.

In view of Lemma 3.14, it is easy to see that $\psi(x) = 0$ is a necessary condition of optimality for $\min\{c(x)|x \in \Omega\}$. Also, because of Lemma 3.14, we can apply (1.3.10) of Polak [2] to obtain the following result.

3.19. Theorem. *Let $\{x_i\}$ be the sequence constructed by Algorithm Prototype* 3.12 *and suppose that Assumptions* 3.13(i) *to* 3.13(iv) *hold. Then if x^* is an accumulation point of $\{x_i\}$, it satisfies $\psi(x^*) = 0$.*

Because Theorem 3.19 can be proved using (1.3.10) of Polak [2], the general convergence theory composed of (3.12), (3.13), and (3.19) is a special case of the Polak theory. Consequently, it also is a special case of Theorem D of Zangwill [7].

Although we cannot directly compare the convergence theory of § 2 with the Polak and Zangwill theories, the results of this section indicate that special cases of the theory in § 2 are special cases of the Polak and Zangwill theories.

4. Applications. Even though the convergence theory of § 2 was motivated by the desire to analyze algorithms which use an antizigzagging parameter, this theory is quite general. At this time we are not aware of any algorithm that can be analyzed by either the Polak theory or by the Zangwill theory but not by the special case of our theory composed of (3.12), (3.13), and (3.19). This is not surprising in view of the similarity of this special case and (1.3.10) of Polak [2]. However, we are mainly concerned with algorithms that use an antizigzagging parameter and thus we shall limit our discussion to these algorithms in this section.

First, we consider those algorithms that can be analyzed by our theory but seem to defy analysis by both the Polak and Zangwill theories. Because of space considerations, we will not carry out the analyses which are usually straightforward.

Method of feasible directions (Zoutendijk [8]). This algorithm is essentially Algorithm 1.2. The analysis is carried out by letting $\psi(\varepsilon, x) = \theta(\varepsilon, x)$.

ε-Perturbation algorithm (Zangwill [7]). The analysis of this algorithm is precisely the same as that for Zoutendijk's algorithm. However, Algorithm Prototype'2.2 must be modified as in Remark 2.27.

Convex programming algorithm ((6.2.2) Zukhovitskiy *et al.* [9]). By use of our convergence theory, we can show that this algorithm can be applied to (1.1) which is not necessarily convex. In this case, $\psi(\varepsilon, x) = \min\{\langle \nabla f^0(x), h\rangle | \langle \nabla f^j(x), h\rangle \leqq -\varepsilon, j \in J(\varepsilon, x)\}$, where $J(\varepsilon, x) = \{j | f^j(x) \geqq -\varepsilon, j \neq 0\}$.

Gradient projection algorithms (Rosen [4], [5]). The well-known gradient projection method of Rosen [5] can easily be modified to use an antizigzagging parameter. Essentially the algorithm should be modified so that the gradient is projected onto the orthogonal complement of the space spanned by the gradients of the ε-active constraints. The process is essentially the same for the related algorithm of Rosen [4]. (Also see Schultz [10].)

Strong variation algorithm for optimal control (Polak and Mayne [3]). Recently Polak and Mayne have used our convergence theory to construct a very complex algorithm for optimal control problems. In the same paper, other complex algorithms were constructed based upon (1.3.10) of Polak [2]. The Polak and Mayne paper serves as an excellent example of the usefulness of the concept of general convergence theory.

Finally, it should be noted that there are algorithms that use an antizigzagging parameter and yet, if the convergence theory of § 2 is used to analyze them, the ε in (2.2) is superfluous. In other words, (3.12), (3.13), and (3.19) can be used to prove convergence for these algorithms. One such algorithm is (4.3.26) of Polak [2] that we have previously cited. Another example is a modification of an algorithm of Topkis and Veinott [6]. This algorithm is meant to solve (1.1) and is obtained by replacing Step 1 and Step 2 of (1.2) by the following.

Step 1′. Compute $\tilde{\theta}(x_i) = \min_{h\in S}\{\max\{\langle\nabla f^0(x_i), h\rangle;\ f^j(x_i) + \langle\nabla f^j(x_i), h\rangle, j \in J(x_i)\}\}$, where $J(x_i) = \{j | f^j(x_i) \geqq -\varepsilon_0, j \neq 0\}$. Let h_i be any solution to this subproblem.

Step 2′. Go to Step 3.

The $\varepsilon_0 > 0$ in this algorithm is required to guarantee convergence and hence can be considered an antizigzagging parameter (Topkis and Veinott take $\varepsilon_0 = +\infty$). However, the antizigzagging parameter is fixed for all iterations and the computation of x_{i+1} only depends upon x_i. As a result, we can use (3.12), (3.13), and (3.19) with $\psi(x) = \tilde{\theta}(x)$ to establish convergence for this algorithm.

5. Conclusion. We have presented a general convergence theory that is applicable to a class of algorithms that cannot be analyzed by either the Polak or Zangwill convergence theories. Some of the relationships between the three general convergence theories have also been examined. Although our theory cannot be directly compared with the Polak or Zangwill theories, our theory seems to be applicable to any algorithm to which the other two theories are applicable.

REFERENCES

[1] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[2] E. Polak, *Computational Methods in Optimization*, Academic Press, New York, 1971.

[3] E. Polak and D. Q. Mayne, *First order, strong variation algorithms for optimal control problems with terminal inequality constraints*, Dept. of Computing and Control, Imperial College of Science and Technology, London, 1972.

[4] J. B. Rosen, *Gradient projection as a least squares solution of Kuhn–Tucker conditions*, Computer Sciences Dept., University of Wisconsin, Madison, 1965.

[5] ———, *The gradient projection method for nonlinear programming, Part I, Linear constraints*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 181–217.

[6] D. M. Topkis and A. Veinott, Jr., *On the convergence of some feasible directions algorithms for nonlinear programming*, this Journal, 5 (1967), pp. 268–279.

[7] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1969.

[8] G. Zoutendijk, *Methods of Feasible Directions*, Elsevier, New York, 1960.

[9] S. I. Zukhovitskiy and T. I. Andeyeva, *Linear and Convex Programming*, W. B. Saunders, Philadelphia, Pa., 1966.

[10] H. Schultz, *A Kuhn–Tucker algorithm*, this Journal, 11 (1973), pp. 438–445.

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

THE EUCLIDEAN SPACE CONTROLLABILITY OF CONTROL SYSTEMS WITH DELAY*

RONALD B. ZMOOD†

Abstract. Algebraic criteria for complete Euclidean space controllability and for Euclidean space null controllability are obtained for control systems with delay. This is achieved by the use of a new form for the fundamental solution of differential-difference equations. In addition, it is shown that the new results reduce to known results for special classes of differential-difference equations.

1. Introduction. In recent years considerable interest has been shown in time lag control systems (see [15] for an extensive bibliography). Systems having transportation delay have been studied extensively, but it is only within the last two decades that a real interest has been shown in the optimal control of such systems. McClamroch [13] and Banks [1] have recently given necessary conditions for the optimality of the solutions of systems with delay. In this work the class of systems with delay that will be considered will be those that can be described by linear differential-difference equations.

In this paper we will consider the controllability of linear differential-difference equations. For these equations it is necessary to differentiate between the notion of function space controllability and controllability in Euclidean space; the reason being that although their solutions are trajectories in Euclidean space, the natural "state space" is a function space. It is hoped that the problem of function space controllability will be treated in a later work, so here we will restrict our attention to controllability in Euclidean space. For Euclidean space controllability we also need to distinguish between the notions of complete controllability and null controllability, in contradistinction with the case for ordinary differential equations.

Complete Euclidean space controllability was first considered by Chyung and Lee in [5], where they generalized the integral criterion of Kalman [10] for the case of differential-difference equations. Kirollova and Curakova [12] in 1967 presented, for the first time, algebraic criteria for Euclidean space null controllability of linear autonomous differential-difference equations. More recently, Gabasov and Curakova [6] have shown that the conditions obtained in 1967 are both necessary and sufficient for complete controllability. A similar algebraic condition to that obtained by Kirillova and Curakova has also been obtained by Johnson [9] in a recent thesis. Weiss [17] has also obtained an algebraic sufficient condition for controllability of time-varying differential-difference equations which includes the results of Buckalo [3] as a special case. Finally, Choudhury [4] has recently published results that are closely related to some of those presented by Gabasov and Curakova [6].

* Received by the editors December 14, 1971, and in revised form May 24, 1973.

† Department of Electrical Engineering, University of Queensland, St. Lucia, Queensland, 4067, Australia. This work was supported in part by the United States Air Force under Grant AF-AFOSR-69-1767 and in part by the National Science Foundation under Grant GK 5798.

In the next section the basic system to be considered is presented, and an important result on the representation of the solution of a differential-difference equation is given. We next present a function space criterion for complete Euclidean space controllability. Then a new form for the fundamental solution of these equations is presented. From these results we obtain an algebraic criterion for complete Euclidean space controllability, and also show that this result reduces to the known results for some special cases. In § 6 the notion of pointwise completeness is introduced, and we then obtain a criterion for Euclidean space null controllability.

2. The basic system. In this section we introduce the basic system that will be considered throughout this paper, including appropriate assumptions.

The system to be considered in this paper is the following differential-difference equation:

$$\dot{x}(t) = A(t)x(t) + B(t)x(t-1) + Cu(t) \quad \text{for } t \in (0, T], \tag{1}$$

$$x(t) = \phi(t) \quad \text{for } t \in [-1, 0]. \tag{2}$$

In what follows we shall use matrix notation, where A^T denotes the transpose of A, and $\|A\|$ denotes the Euclidean norm of A. In (1) and (2) above, $x(t)$ is an n-vector, $u(t)$ is an r-vector, and $u(\cdot)$ is an admissible control (that is, it is contained in the space of square integrable functions L^2 on every finite interval). $A(t)$ and $B(t)$ are $n \times n$ matrix functions, measurable in t and satisfying $\|A(t)\| < m(t)$, $\|B(t)\| < m(t)$, where $m \in L^2(0, T)$. Also $C(t)$ is an $n \times r$ matrix, and $C(\cdot)$ is contained in L^2. We will assume that the initial function $\phi(t)$ is continuous on the interval $t \in [-1, 0]$, that is, $\phi \in C([-1, 0]; R^n)$.

THEOREM 1. *With the above assumptions and for each admissible control* $u(t)$, $t \in [0, T]$, *there exists a unique solution to* (1), (2), *where the solution is continuous on the interval* $t \in [-1, 0]$ *and is absolutely continuous on the interval* $t \in (0, T]$. *Further, the solution at time T is*

$$x(T) = x(T, \phi) + \int_0^T X(T, s)C(s)u(s)\, ds, \tag{3}$$

and

$$x(T, \phi) = X(T, 0)\phi(0) + \int_{-1}^{0} X(T, s+1)B(s+1)\phi(s)\, ds. \tag{4}$$

$X(t, s)$ *is a unique* $n \times n$ *matrix solution, defined on* $[-1, T] \times [0, T]$, *of*

$$\frac{\partial}{\partial t}X(t, s) = A(t)X(t, s) + B(t)X(t-1, s), \tag{5}$$

for $(t, s) \in [s, T] \times [0, T]$, *and*

$$X(t, s) = \begin{cases} I & (n \times n \text{ identity matrix}) \quad \textit{for } t = s, \\ 0 & \text{for } (t, s) \in [-1, s) \times [0, T]. \end{cases} \tag{6}$$

The proof of the existence of the solution is given in [1], while the above variation of parameters formula may be obtained by appropriate substitutions in the results of [2] or [13].

The matrix function $X(t, s)$ will be called the *fundamental solution* of (1). For the case where (1) is an autonomous differential-difference equation, we may write the fundamental solution $X(t, s)$ as $X(t - s)$ without any loss of generality.

3. A function space criterion for complete Euclidean space controllability. In this section we define the concept of complete controllability, and derive a necessary and sufficient condition for differential-difference equations to have this property.

DEFINITION 1. The control system (1), (2) is said to be *completely Euclidean space controllable* at time $T > 0$ if for every $\phi \in C([-1, 0]; R^n)$ and for every $x_1 \in R^n$, there exists an admissible control $u(t)$, $t \in [0, T]$, such that $x(T) = x_1$.

DEFINITION 2. Consider the control system (1), (2). The *reachable set* $\mathscr{R}(T)$ is defined as

$$\mathscr{R}(T) = \{x \in R^n | x = x(T), u(\cdot) \in L^2, \phi = 0\}.$$

It is relatively easy to see that the reachable set $\mathscr{R}(T)$ is a linear subspace of R^n. From (3),

$$x(T) - x(T, \phi) = \int_0^T X(T, s)C(s)u(s)\,ds. \tag{7}$$

We see that for each admissible control, u, the right-hand side of (7) is an element of the reachable set $\mathscr{R}(T)$. Now suppose that (1), (2) is completely Euclidean space controllable. Then for each $x_1 \in R^n$, there exists an admissible control $\tilde{u}$ such that

$$x_1 - x(T, \phi) = \int_0^T X(T, s)C(s)\tilde{u}(s)\,ds \in \mathscr{R}(T).$$

Therefore the system (1), (2) is completely Euclidean space controllable at time T if and only if $\mathscr{R}(T) = R^n$.

THEOREM 2. *The control system* (1), (2) *is completely Euclidean space controllable at time T if and only if $\eta^T X(T, s)C(s) = 0$ for a.e. $s \in [0, T]$ implies $\eta = 0$, where $\eta \in R^n$.*

Proof. From above, the system (1), (2) is completely Euclidean space controllable at time T if and only if $\mathscr{R}(T) = R^n$.

Now $\mathscr{R}(T) \neq R^n$ if and only if there exists a nonzero $\eta \in R^n$ such that $\eta^T x = 0$ for every $x \in \mathscr{R}(T)$. This in turn is equivalent to

$$\int_0^T \eta^T X(T, s)C(s)u(s)\,ds = 0, \tag{8}$$

for all controls $u(\cdot) \in L^2$.

If there exists a nonzero $\eta \in R^n$ such that $\eta^T X(T, s)C(s) = 0$ for a.e. $s \in [0, T]$, then (8) is true for all $u(\cdot) \in L^2$. On the other hand, if there exists a nonzero $\eta \in R^n$ such that (8) is true for all $u(\cdot) \in L^2$, then (8) is true for

$$u(s) = [\eta^T X(T, s)C(s)]^T. \tag{9}$$

Substituting (9) into (8), we obtain

$$\int_0^T \eta^T X(T, s)C(s)C^T(s)X^T(T, s)\eta \, ds = 0, \tag{10}$$

and therefore conclude $\eta^T X(T, s)C(s) = 0$ for a.e. $s \in [0, T]$.

COROLLARY 1. *The control system* (1), (2) *is completely Euclidean space controllable at time T if and only if*

$$W(T, 0) = \int_0^T X(T, s)C(s)C^T(s)X(T, s)\, ds$$

is positive definite.

Proof. It is apparent that $W(T, 0)$ is positive semidefinite. The rest of the proof is then quite straightforward.

Corollary 1 was originally obtained by Chyung and Lee [5] and McClamroch [14].

4. The representation of the fundamental solution of a differential-difference equation. In this section we present a form of the fundamental solution $X(t)$ which will prove to be quite useful in later sections.

Let us consider the matrix differential equation (5) and (6), where we assume the coefficient matrices $A(t)$ and $B(t)$ are constant and equal to A and B, respectively. We introduce the following notation, by defining $X_k(\tau) = X(\tau + k)$ for $\tau \in [0, 1]$ and $k = 0, 1, \cdots$. By direct substitution in (5) and assuming $s = 0$, we obtain

$$\begin{aligned} &\frac{d}{d\tau}X_0(\tau) = AX_0(\tau), && X_0(0) = I, \\ &\frac{d}{d\tau}X_1(\tau) = BX_0(\tau) + AX_1(\tau), && X_1(0) = X_0(1), \\ &\qquad\vdots && \qquad\vdots \\ &\frac{d}{d\tau}X_k(\tau) = BX_{k-1}(\tau) + AX_k(\tau), && X_k(0) = X_{k-1}(1), \end{aligned} \tag{11}$$

so that the solution of (5) and (6) over the interval $t \in [k, k + 1]$ is given by $X(t) = X_k(t - k)$.

If $Z_k(\tau) = [X_0^T(\tau), \cdots, X_k^T(\tau)]^T$, (11) becomes

$$\frac{d}{d\tau}Z_k(\tau) = A_k Z_k(\tau) \quad \text{for } \tau \in [0, 1], \tag{12}$$

$$X_k(\tau) = E_k Z_k(\tau), \tag{13}$$

where $Z_k(\tau)$ is an $n(k+1) \times n$ matrix, and

$$A_k = \begin{bmatrix} A & 0 & 0 & \cdots & 0 & 0 \\ B & A & 0 & \cdots & 0 & 0 \\ 0 & B & A & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & B & A \end{bmatrix}, \qquad E_k = [0, \cdots, 0, I].$$

A_k and E_k are $n(k+1) \times n(k+1)$ and $n \times n(k+1)$ matrices, respectively.

As is well known, the unique solution of (12) is given by

$$Z_k(\tau) = e^{A_k\tau}Z_k(0), \tag{14}$$

and so

$$X_k(\tau) = E_k\, e^{A_k\tau}Z_k(0). \tag{15}$$

It is clear from (11) and the definition of $Z_k(\tau)$ that

$$Z_0(0) = I. \tag{16}$$

Further, by induction,

$$Z_k(0) = \left[\begin{array}{c} I \\ \hdashline e^{A_{k-1}}Z_{k-1}(0) \end{array}\right] \quad \text{for } k = 1, 2, \cdots. \tag{17}$$

From (15) and the definition of $X_k(\tau)$, we obtain

$$X(t) = X_k(t-k) = E_k\, e^{A_k(t-k)}Z_k(0), \tag{18}$$

for $t \in [k, k+1]$. An important observation is the fact that the fundamental matrix $X(t)$ may be singular for some values of $t \in [0, \infty)$. It is this fact that makes it necessary to distinguish between complete Euclidean space controllability as defined above and the notion of Euclidean space null controllability to be defined in a later section.

5. An algebraic criterion for complete Euclidean space controllability of autonomous differential-difference equations. We now develop an algebraic necessary and sufficient condition for complete Euclidean space controllability. From this result we deduce the well-known Kalman [11] condition for the controllability of linear ordinary differential equations. In addition, we obtain the condition of Kirillova and Curakova [12], and also show that the general scalar differential-difference equation is controllable. To complete this section, we present an algebraic characterization of the reachable set $\mathscr{R}(T)$.

THEOREM 3. *A necessary and sufficient condition for the system* (1), (2), *with* $A(t)$, $B(t)$ *and* $C(t)$ *equal to constant matrices* A, B *and* C, *respectively, to be completely Euclidean space controllable at time* $T \in (k, k+1]$, $k = 0, 1, \cdots$, *is that the matrix*

$$Q(T) = [E_0C_0, \cdots, E_0A_0^{n-1}C_0, \cdots, E_kC_k, \cdots, E_kA_k^{n(k+1)-1}C_k]$$

has rank n, *where* $C_i = Z_i(0)C$ *for* $i = 0, 1, 2, \cdots$.

Proof. From Theorem 2, a necessary and sufficient condition for complete Euclidean space controllability is that $\eta^T X(T-s)C = 0$ for $s \in [0, T]$ imply $\eta = 0$.

This statement is equivalent to:

$$\eta^T X(T-s)C = 0 \quad \text{for } s \in (T-1, T], \tag{19}$$

$$\eta^T X(T-s)C = 0 \quad \text{for } s \in (T-2, T-1], \tag{20}$$

$$\vdots$$

$$\eta^T X(T-s)C = 0 \quad \text{for } s \in (0, T-k] \tag{21}$$

implies $\eta = 0$.

From (18), we obtain by direct substitution that

$$\eta^T E_0 \, e^{A_0(T-s)} C_0 = 0 \quad \text{for } s \in (T-1, T], \tag{22}$$

$$\eta^T E_1 \, e^{A_1(T-s-1)} C_1 = 0 \quad \text{for } s \in (T-2, T-1], \tag{23}$$

$$\vdots$$

$$\eta^T E_k \, e^{A_k(T-s-k)} C_k = 0 \quad \text{for } s \in (0, T-k] \tag{24}$$

implies $\eta = 0$.

We now only need to show that (22), (23), (24) implies $\eta = 0$ if and only if the rank of $Q(T)$ equals n. Suppose the rank of $Q(T) < n$; then there exists a nonzero $n \in R^n$ such that

$$\eta^T E_i C_i = \cdots = \eta^T E_i A_i^{n(i+1)-1} C_i = 0, \tag{25}$$

for $i = 0, 1, \cdots, k$. From the Cayley–Hamilton theorem and (25), we find that

$$\eta^T E_i A_i^{n(i+1)} C_i = 0, \tag{26}$$

for $i = 0, 1, \cdots, k$. By induction, it can then be shown that

$$\eta^T E_i A_i^{n(i+1)+l} C_i = 0, \tag{27}$$

for $l = 0, 1, 2, \cdots$ and $i = 0, 1, \cdots, k$. Using the power series expansion of the exponential matrix, we find that

$$\eta^T E_0 \, e^{A_0(T-s)} C_0 = 0 \qquad \text{for } s \in (T-1, T],$$

$$\eta^T E_1 \, e^{A_1(T-s-1)} C_1 = 0 \quad \text{for } s \in (T-2, T-1],$$

$$\vdots$$

$$\eta^T E_k \, e^{A_k(T-s-k)} C_k = 0 \quad \text{for } s \in (0, T-k].$$

This is a contradiction of the statement: (22), (23) and (24) imply $\eta = 0$.

Suppose, on the other hand, there exists a nonzero $\eta \in R^n$ such that (22), (23) and (24) are true. Then by successive differentiation of (22) and letting $s = T$, we obtain

$$\eta^T E_0 C_0 = \cdots = \eta^T E_0 A_0^{n-1} C_0 = 0. \tag{28}$$

Treating (23) and (24) in a similar fashion, we obtain

$$\eta^T E_i A_i^j C_i = 0,$$

for $i = 0, 1, \cdots, k$ and $j = 0, 1, \cdots, n(i + 1) - 1$. It is thus obvious that $\eta^T Q(T) = 0$, which implies the rank of $Q(T)$ is less than n.

It will be observed from Theorem 3 that an autonomous differential-difference equation may become controllable after some nonzero time interval has elapsed. In the following example we present a system which is not completely controllable for $T \leqq 1$, but is completely controllable for all $T > 1$.

Example 1. Consider the following differential-difference equation:

$$(29) \qquad \begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix} + \begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}x_1(t-1)\\ x_2(t-1)\end{bmatrix} + \begin{bmatrix}1\\ 1\end{bmatrix}u(t).$$

Suppose $T \in (0, 1]$; then we find that

$$Q(T) = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix},$$

and the rank of $Q(T) = 1 < 2$. Now suppose $T > 1$; then

$$Q(T) = \begin{bmatrix}1 & 1 & e & 1 + e^1 \cdots \\ 1 & 1 & e & e^1 \cdots \end{bmatrix},$$

and we note that the rank of $Q(T) = 2$. Hence the system (29) is not completely controllable for $T \in (0, 1]$, but is completely controllable for all time $T > 1$.

COROLLARY 2. *If the system* (1), (2), *with* $A(t)$, $B(t)$ *and* $C(t)$ *equal to constant matrices* A, B *and* C, *respectively, is completely Euclidean space controllable at time* T_1, *then it is completely Euclidean space controllable for all time* $T \geqq T_1$.

Proof. Suppose $T \geqq T_1$; then from Theorem 3 we can partition $Q(T)$ as

$$Q(T) = [Q(T_1), \tilde{Q}].$$

It can now be seen that if $Q(T_1)$ has rank n, then $Q(T)$ has rank n.

The following result is the usual algebraic criterion for complete Euclidean space controllability of ordinary differential equations first proved by Kalman [11].

COROLLARY 3. *Suppose that in* (1), (2) *the matrices* $A(t)$ *and* $C(t)$ *equal the constant matrices* A *and* C, *respectively, and* $B(t) \equiv 0$ *for all* $t \in [0, \infty)$. *Then a necessary and sufficient condition for complete Euclidean space controllability at time* T *is that the matrix*

$$Q_A(T) = [C, AC, \cdots, A^{n-1}C]$$

have rank n.

Proof. Suppose $Q_A(T)$ has rank n; then $Q(T)$ has rank n for any T. This follows from the fact that

$$(30) \qquad E_0C_0 = C, \cdots, E_0A_0^{n-1}C_0 = A^{n-1}C,$$

and hence $Q(T)$ has n linearly independent columns for any $T > 0$.

Suppose the rank of $Q_A(T) < n$; then there exists a nonzero $\eta \in R^n$ such that

$$(31) \qquad \eta^T C = \eta^T AC = \cdots = \eta^T A^{n-1}C.$$

Applying the Cayley–Hamilton theorem, we can show that

$$\eta^T A^n C = 0.$$

It can then be shown by induction that

$$\eta^T A^{n+l} C = 0, \tag{32}$$

for $l = 0, 1, \cdots$. The general term in $\eta^T Q(T)$ is given by

$$\eta^T E_i A_i^l C_i,$$

for $i = 0, 1, \cdots, k$ and $l = 0, 1, \cdots, n(i + 1) - 1$. Upon substitution for E_i, A_i and C_i, the general term becomes

$$\eta^T E_i A_i^l C_i = \eta^T[0, \cdots, 0, I] \begin{bmatrix} A^l & 0 & \cdots & 0 \\ 0 & A^l & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & A^l \end{bmatrix} \begin{bmatrix} I \\ e^A \\ \vdots \\ e^{iA} \end{bmatrix} C \tag{33}$$

$$= \eta^T A^l e^{iA} C.$$

Expanding e^{iA} in a power series, and applying (31) and (32) to the general term, we obtain

$$\eta^T E_i A_i^l C_i = \eta^T A^l [I + iA + \cdots] C = 0.$$

Hence $\eta^T Q(T) = 0$ and the rank of $Q(T) < n$.

We next prove the result, originally given by Kirillova and Curakova [12], for the system (1), (2), where $A(t) \equiv 0$ for all $t \in [0, \infty)$, $B(t) \equiv B$ and $C(t) \equiv C$, where B and C are constant matrices.

COROLLARY 4. *Suppose in the system* (1), (2) *that the matrix* $A(t) \equiv 0$ *for all* $t \in [0, \infty)$, $B(t) \equiv B$, *and* $C(t) \equiv C$, *where B and C are constant matrices. Then a necessary and sufficient condition for complete Euclidean space controllability at time* $T \in (k, k + 1]$ *is that the matrix*

$$Q_B(T) = [C, \cdots, B^k C]$$

have rank n.

If $T > n - 1$, *then a necessary and sufficient condition for complete Euclidean space controllability at time T is that the matrix*

$$Q_C(T) = [C, \cdots, B^{n-1} C]$$

have rank n.

Proof. Suppose the rank of $Q_B(T)$ is less than n; then there exists a nonzero $\eta \in R^n$ such that

$$\eta^T Q_B(T) = \eta^T [C, \cdots, B^k C] = 0. \tag{34}$$

The general term in $\eta^T Q(T)$ is given by

$$\eta^T E_i A_i^l C_i, \tag{35}$$

for $i = 0, 1, \cdots, k$ and $l = 0, 1, \cdots, n(i + 1) - 1$. Upon substitution for E_i, A_i and C_i, the general term becomes

$$\eta^T E_i A_i^l C_i = [0, \cdots, 0, I] \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ B & 0 & 0 & \cdots & 0 & 0 \\ 0 & B & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & B & 0 \end{bmatrix}^l \tag{36}$$

$$\times \begin{bmatrix} I \\ I \\ I + B \\ \vdots \\ I + (i-1)B + \cdots + \dfrac{B^{i-1}}{(i-1)!} \end{bmatrix} C.$$

For $l = 0$, (36) becomes

$$\eta^T\left[I + (i-1)B + \cdots + \frac{1}{(i-1)!}B^{i-1}\right]C, \tag{37}$$

and from (34) we conclude that this term equals zero. For $l = 1$, (36) becomes

$$\eta^T\left[I + (i-2)B + \cdots + \frac{1}{(i-2)!}B^{i-2}\right]BC, \tag{38}$$

which is again zero by (34). Similarly, $\eta^T E_i A_i^l C_i = 0$ for $1 < l < i$. For $l = i$, expansion of (36) leads to $\eta^T B^i C$ which is zero from (34). For $l > i$, we see that $A_i^l = 0$; hence $\eta^T E_i A_i^l C_i = 0$ for $l > i$. From the above discussion we may finally conclude that $\eta^T E_i A_i^l C_i = 0$ for $l = 0, 1, \cdots, n(i + 1) - 1$. Hence we have $\eta^T Q(T) = 0$, and the rank of $Q(T)$ is less than n.

For sufficiency, suppose $Q_B(T)$ has rank n. We recognize from the general term (36) and the discussion following it that $E_i A_i^l C_i = B^i C$. Hence, indicating only the important terms, the matrix $Q(T)$ has the form

$$[C, \cdots, BC, \cdots, B^2C, \cdots, B^kC],$$

and so it must also have rank n.

Let us suppose $T > n - 1$. If $Q_C(T)$ has rank n, then since it constitutes the first n submatrices of $Q_B(T)$, we see that $Q_B(T)$ has rank n.

Now suppose Q_C has rank less than n; then there exists a nonzero $\eta \in R^n$ such that

$$\eta^T[C, \cdots, B^{n-1}C] = 0. \tag{39}$$

Applying the Cayley–Hamilton theorem and an inductive argument, we find that

$$\eta^T B^{n+l} C = 0,$$

for $l = 0, 1, \cdots$. Hence the rank of $Q_B(T)$ is also less than n.

Let us now consider the scalar differential-difference equation

$$x^{(n)}(t) + \sum_{i=0}^{n-1} a_i x^{(i)}(t) + \sum_{i=0}^{n-1} b_i x^{(i)}(t-1) = cu(t), \tag{40}$$

where $x(t)$ and $u(t)$ are scalar functions of time t, and the parameters a_i, b_i and c are constant. We denote the ith derivative of $x(t)$ by $x^{(i)}(t)$. This class of system has been discussed by Kirillova and Curakova [12] and Halanay [8], where it has been shown that (40) is null controllable. Here we transform (40) into the form of (1) and thus show that it is completely controllable.

Defining $x_1 = x^{(0)}$, $x_2 = x^{(1)}, \cdots, x_n = x^{(n-1)}$, we can write (40) as

$$\begin{aligned} \dot{x}_1(t) &= x_2(t), \\ \dot{x}_2(t) &= x_3(t), \\ &\vdots \\ \dot{x}_n(t) &= -\sum_{i=0}^{n-1} a_i x_{i+1}(t) - \sum_{i=0}^{n-1} b_i x_{i+1}(t-1) + cu(t). \end{aligned} \tag{41}$$

This equation can be written in the matrix form given in (1), where

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ -b_0 & -b_1 & \cdots & -b_{n-1} \end{bmatrix}, \tag{42}$$

$$C = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

COROLLARY 5. *The scalar control system* (40) *is completely Euclidean space controllable for every time* $T > 0$.

Proof. The proof is an immediate consequence of (42) and Theorem 3.

We now present an algebraic characterization of the reachable set $\mathscr{R}(T)$.

THEOREM 4. *The reachable set* $\mathscr{R}(T)$ *equals the range of the matrix* $Q(T)Q^T(T)$, *where* $Q(T)$ *is the matrix defined in Theorem* 3.

Proof. By examining the proof of Theorem 3, we see that the orthogonal complement, $\mathscr{R}^\perp(T)$, of $\mathscr{R}(T)$ is given by

$$\mathscr{R}^\perp(T) = \{\eta \in R^n \mid \eta^T Q(T) = 0\}. \tag{43}$$

It is clear that (43) can be rewritten as

$$\mathscr{R}^{\perp}(T) = \{\eta \in R^n \mid Q^T(T)\eta = 0\} = \{\eta \in R^n \mid Q(T)Q^T(T)\eta = 0\}$$
$$= \operatorname{null}(Q(T)Q^T(T)) = \operatorname{range}(Q(T)Q^T(T))^{\perp}.$$

The last two equalities follow from the basic properties of linear transformations, where null $(Q(T)Q^T(T))$ denotes the null space of $Q(T)Q^T(T)$, and range $(Q(T)Q^T(T))$ denotes the range of $Q(T)Q^T(T)$. Hence $\mathscr{R}(T) = \operatorname{range}(Q(T)Q^T(T))$.

6. Pointwise completeness. In order to be able to discuss the question of null controllability of differential-difference equations, it is necessary to introduce the concept of pointwise completeness. This notion has been explored more fully in [16] and [18], so we will content ourselves with a definition of the concept and the presentation of the main result.

Consider the system (1), (2) with the control $u(t) \equiv 0$,

$$\dot{x}(t) = A(t)x(t) + B(t)x(t-1), \qquad t \in (0, T], \tag{44}$$

$$x(t) = \phi(t), \qquad t \in [-1, 0]. \tag{45}$$

DEFINITION 3. The system (44), (45) is said to be *pointwise complete* at time T if for every $x_1 \in R^n$, there exists a $\phi \in C([-1,0]; R^n)$ such that $x(T) = x_1$.

The solution of (44), (45) is given by the variation of parameters formula (4), which it will be noted is a linear operator mapping from $C([-1,0]; R^n)$ into R^n.

DEFINITION 4. Consider the system (44), (45). The *pointwise reachable set* $\mathscr{P}(T)$ is defined as

$$\mathscr{P}(T) = \{x \in R^n \mid x = x(T, \phi), \phi \in C([-1,0]; R^n)\}.$$

The term $x(T, \phi)$ is as defined in (4). It can easily be seen that $\mathscr{P}(T)$ is a linear subspace of R^n. We now present the following results without proof.

THEOREM 5. *A necessary and sufficient condition for* (44) *to be pointwise complete for $t = T$ is that for every nonzero $\eta \in R^n$,*

(i) *there exists a set $S \subset [0, 1]$, of nonzero measure, such that $\eta^T X(T, \alpha)B(\alpha) \neq 0$ for $\alpha \in S$*

or

(ii) $\eta^T X(T, 0) \neq 0$.

Proof. This is presented in [18].

COROLLARY 6. *In* (44), *let us suppose that $A(t)$ and $B(t)$ equal the constant matrices A and B, respectively. Then a necessary and sufficient condition for* (44), (45) *to be pointwise complete at time $T \in [k, k+1)$, $k = 0, 1, \cdots$, is that the matrix*

$$M(T) = [E_{k-1}F_{k-1}, \cdots, E_{k-1}A_{k-1}^{nk-1}F_{k-1}, E_{k-1}Z_k(0)]$$

have rank n, where $F_k = Z_k(0)B$.

Proof. This is presented in [18].

Finally, we present an algebraic characterization of the pointwise reachable set $\mathscr{P}(T)$.

THEOREM 6. *In* (44), *we suppose $A(t)$ and $B(t)$ equal the constant matrices A and B, respectively. The pointwise reachable set $\mathscr{P}(T)$ equals the range of the matrix $M(T)M^T(T)$.*

Proof. This result is a consequence of Corollary 6.

7. The null controllability of autonomous differential-difference equations. In this section we define the notion of Euclidean space null controllability, and obtain algebraic necessary and sufficient conditions under which differential-difference equations are null controllable. Gabasov and Kirillova [7, pp. 70–75] have also considered the problem of null controllability (they term it relative controllability) of these equations. Upon examination of their proof, however, it is apparent that they have implicitly assumed the system is pointwise complete. Hence pointwise completeness is a necessary hypothesis of their result. It will be subsequently shown that for systems of order $n = 1, 2$ and 3, a necessary and sufficient condition for null controllability is that the system be completely controllable. It is intended to give a more complete discussion of this question in a future paper.

DEFINITION 5. The system (1), (2) is said to be *Euclidean space null controllable* at time $T > 0$ if for every $\phi \in C([-1, 0]; R^n)$, there exists a control $u(t), 0 \leqq t \leqq T$, such that $x(T) = 0$.

From (3), which we assume is autonomous, we obtain

$$x(T) - x(T, \phi) = \int_0^T X(T - s)Cu(s)\, ds. \tag{46}$$

For null controllability, $x(T) = 0$ for any initial function $\phi \in C([-1,0]; R^n)$; hence (46) becomes

$$-x(T, \phi) = \int_0^T X(T - s)Cu(s)\, ds, \tag{47}$$

for some admissible control $u(t)$, $0 \leqq t \leqq T$. From Definitions 2 and 4, it immediately follows that the control system (1), (2) is Euclidean space null controllable if and only if $\mathscr{P}(T) \subset \mathscr{R}(T)$.

THEOREM 7. *Suppose that in the control system* (1), (2), *the matrices* $A(t)$, $B(t)$ *and* $C(t)$ *are equal to the constant matrices* A, B *and* C, *respectively. Then a necessary and sufficient condition for it to be Euclidean space null controllable at time* $T \in [k, k + 1), k = 0, 1, \cdots$, *is that the rank of* $[M(T)M^T(T), Q(T)Q^T(T)]$ *equal the rank of* $Q(T)Q^T(T)$, *where* $Q(T)$ *and* $M(T)$ *are as defined above.*

Proof. From the discussion preceding the theorem and the results of Theorem 4 and Theorem 6, we see that (1), (2) is null controllable if and only if range $(M(T)M^T(T)) \subset$ range $(Q(T)Q^T(T))$. The result then follows quite simply from the elementary properties of matrices.

THEOREM 8. *Suppose that in the control system* (1), (2), *the matrices* $A(t)$, $B(t)$ *and* $C(t)$ *are equal to the constant matrices* A, B *and* C, *respectively. For* $n = 1$, 2 *or* 3, *a necessary and sufficient condition for it to be Euclidean space null controllable at time* $T \in [k, k + 1), k = 0, 1, \cdots$, *is that it be completely controllable at time* T.

Proof. For $n = 1$ or 2, it has been shown [16] that the system (1), (2) is pointwise complete. From the discussion preceding Theorem 7, it is clear that the proof is complete.

For the case where $n = 3$, we note that if the system is completely controllable, then $\mathscr{R}(T) = R^3$, and so $\mathscr{R}(T) \supset \mathscr{P}(T)$, that is, it is null controllable.

We now prove necessity by contradiction. Suppose the system is null controllable; then $\mathscr{R}(T) \supset \mathscr{P}(T)$. We need to consider two cases: the first is where $\mathscr{P}(T) = R^3$, and the second is where $\mathscr{P}(T) \subsetneq R^3$. In the first case, we see that $\mathscr{R}(T) = R^3$, which implies that the system is completely controllable.

Let us now consider the case where $\mathscr{P}(T) \subsetneq R^3$. Assuming $\mathscr{R}(T) \subsetneq R^3$, it is clear from Theorem 2 that $\mathscr{R}(t) \subsetneq R^3$ for $t \in (0, T]$. This implies that the matrices A and C form an uncontrollable pair. Without loss of generality, we will assume that these matrices are partitioned as below:

$$A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \tag{48}$$

and

$$C = \begin{bmatrix} 0 \\ C_1 \end{bmatrix}, \tag{49}$$

where A_{11}, A_{21}, A_{22} and C_1 are submatrices of A and C, respectively.

Since $\mathscr{R}(T) \subsetneq R^3$, there exists a nonzero vector $q \in R^3$ such that q is orthogonal to every element $x \in \mathscr{R}(T)$. From (48) and (49), it is clear that q has the partitioned form

$$q = \begin{bmatrix} q_1 \\ 0 \end{bmatrix}. \tag{50}$$

Now Popov [16] has shown that for any system for which $n = 3$, the pointwise reachable set $\mathscr{P}(T) \subsetneq R^3$ for $T > 2$ if and only if (44) can be written as

$$\dot{x}(t) = Ax(t) + (AZ - ZA)x(t - 1), \tag{51}$$

where $Z = rq^T e^A$, and q and r are nonzero vectors satisfying

$$\begin{bmatrix} q^T \\ q^T e^A \\ q^T e^A A \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \tag{52}$$

Substituting (48) and (50) into the matrix of (52), we obtain the matrix

$$\left[\begin{array}{l|c} q_1^T & 0 \\ q_1^T e^{A_{11}} & 0 \\ q_1^T e^{A_{11}} A_{11} & 0 \end{array}\right]. \tag{53}$$

It is necessary to consider first the case where q_1 is a 1-vector. The term $q_1^T e^{A_{11}}$ is nonzero if and only if q_1 is nonzero. As a consequence, there exists no vector r such that equation (52) is satisfied. This implies $\mathscr{P}(T) = R^3$.

For the second case, q_1 is a 2-vector. It can be seen that for (52) to be satisfied, it is required that

$$q_1^T \underset{\sim}{r} = 1 \tag{54}$$

and

$$(55)\qquad \begin{bmatrix} q_1^T e^{A_{11}} \\ q_1^T e^{A_{11}}A_{11} \end{bmatrix} \underset{\sim}{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

where $\underset{\sim}{r}^T = [r_1 \;\; r_2]$. Now (55) holds if and only if $q_1^T e^{A_{11}}$ and $q_1^T e^{A_{11}}A_{11}$ are linearly dependent. Letting $\eta^T = q_1^T e^{A_{11}}$, we see that

$$(56)\qquad \eta^T A_{11} = \alpha\eta^T.$$

Also $q_1^T = \eta^T e^{-A_{11}}$, and by the use of the Cayley–Hamilton theorem and (56), we have

$$(57)\qquad q_1^T = \beta\eta^T, \qquad \beta \neq 0.$$

Hence if (55) is satisfied, we see that $\eta^T\underset{\sim}{r} = 0$, which from (57) implies that $q_1^T\underset{\sim}{r} = 0$. It is clear that there exists no r such that (52) is satisfied. Consequently $\mathscr{P}(T) = R^3$.

In both cases we have a contradiction, so that $\mathscr{R}(T) = R^3$, and hence the system is completely controllable.

8. Conclusions.

Necessary and sufficient conditions for the controllability of autonomous differential-difference equations have been presented. It has also been shown that these results reduce to the known results of Kalman et al. [11], and those of Kirillova and Curakova [12], for the particular equations they treated. A general condition for null controllability has also been obtained, and it has been shown that for first, second and third order systems, null controllability is equivalent to complete controllability. The crux of the approach presented here is, of course, the realization of the particularly simple form for the fundamental solution $X(t)$, which enabled us to exploit the analyticity properties of the exponential matrix.

Several extensions are possible, of which the case of multiple delays and the case of delay in the control are the most obvious.

Acknowledgments. The author would like to express his thanks to Professor N. Harris McClamroch of the University of Michigan, under whose guidance this work was performed. The author would also like to thank the reviewer for bringing to his attention some recent references on this problem.

REFERENCES

[1] H. T. BANKS, *Necessary conditions for control problems with variable time lags*, this Journal, 6 (1968), pp. 9–47.

[2] ———, *Representations for solutions of linear functional differential equations*, J. Differential Equations, 5 (1969), pp. 399–409.

[3] A. F. BUCKALO, *Explicit conditions for controllability of linear systems with time lag*, IEEE Trans. Automatic Control, AC-2 (1968), pp. 193–195.

[4] A. K. CHOUDHURY, *Algebraic and transfer-function criteria of fixed-time controllability of delay-differential systems*, Internat. J. Control, 6 (1972), pp. 1073–1082.

[5] D. H. CHYUNG AND E. B. LEE, *Linear optimal systems with time delay*, this Journal, 4 (1966), pp. 548–574.

[6] R. GABASOV AND S. V. CURAKOVA, *The theory of controllability of linear systems with delay lags*, Engrg. Cybernetics, 4 (1969), pp. 16–27.

[7] R. GABASOV AND F. M. KIRILLOVA, *Qualitative Theory of Optimum Control*, Nauk Publ. Moscow, 1971.

[8] A. HALANAY, *On the controllability of linear difference-differential systems*, Lecture Notes on Operations Research and Mathematical Economics 12, Springer, New York, 1969, pp. 329–336.

[9] J. E. JOHNSON, *On the controllability of systems with time delays*, Ph.D. thesis, Univ. of Michigan, Ann Arbor, 1969.

[10] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[11] R. E. KALMAN, Y. C. HO AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1963), pp. 189–213.

[12] F. M. KIRILLOVA AND S. V. CURAKOVA, *On the problem of controllability of linear systems with after effects*, Differensial'nye Uravnenija, 3 (1967), pp. 436–445. (In Russian.)

[13] N. H. MCCLAMROCH, *Theory of extremal controls for linear hereditary processes*, to appear.

[14] ———, *A general adjoint relation for functional differential equations and integral equations with application to control*, J. Optimization Theory Appl., 7 (1971), pp. 346–356.

[15] M. N. OGUZTORELI, *Time-Lag Control Systems*, Academic Press, New York, 1966.

[16] V. M. POPOV, *Pointwise degeneracy of linear, time-invariant, delay-differential equations*, J. Differential Equations, 11 (1972), pp. 541–561.

[17] L. WEISS, *On the controllability of delay-differential equations*, this Journal, 5 (1967), pp. 575–587.

[18] R. B. ZMOOD AND N. H. MCCLAMROCH, *On the pointwise completeness of differential-difference equations*, J. Differential Equations, 12 (1972), pp. 474–486.

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

THE CLOSED-LOOP TIME-OPTIMAL CONTROL. I: OPTIMALITY*

PAVOL BRUNOVSKÝ†

Abstract. The problem of optimality of the trajectories of a linear system with a closed-loop time-optimal control is studied and solved in dimension 2.

1. Introduction. Consider the standard time-optimal control problem for the system

$$\dot{x} = Ax + u \tag{1}$$

($x \in R^n$, A constant), with control constraints $u \in U$, where U is a convex compact polytope in R^n (of possibly lower dimension), containing the origin in its relative interior[1]).

Throughout this paper, we shall assume that the problem is normal, i.e., for any two vertices w_1, w_2 of U, if $\langle \psi, w_1 - w_2 \rangle = 0$, then ψ does not belong to any proper invariant subspace of A' (A' standing for the transpose of A, $\langle \cdot, \cdot \rangle$ for the scalar product), and that U contains the origin in its relative interior.

It is well known from the standard textbooks on optimal control theory (cf., e.g., [1], [7], [8]) that under our assumptions, for any point x which can be steered to the origin, there exists a unique optimal control $u_x(t)$, $t \in [0, T(x)]$, which is piecewise constant with values only at the vertices of U, under the agreement $u_x(t) = \lim_{s \downarrow t} u_x(s)$ for $t \in [0, T(x))$, and $u_x(T(x)) = \lim_{t \uparrow T(x)} u_x(t)$ (which we shall obey throughout this paper).

However, it is frequently desirable to synthesize the optimal control into a closed-loop optimal control (or, optimal feedback), i.e., to express the value of the optimal control as a function of the instant state of the system, independently of the initial state. The possibility of such synthesis follows from the unicity of the open-loop optimal control; it can be verified immediately that the function $v(x) = u_x(0)$ is such a closed-loop optimal control (cf. [1]).

For the desirability of implementing a closed-loop optimal control the following reasons are usually given:

1. There is no need to compute the optimal control for every new initial state separately.

2. The controller is sensitive to instantaneous perturbations: if at any instant of the process the system is deviated from its optimal trajectory, the rest of the process will again lead to the desired final state and will be optimal with respect to this new initial state.

Turning from one instantaneous perturbation to more or less permanently acting perturbations suggests a certain stability of the performance of a system

[1] For the relation of this representation of the time-optimal control problem to the more common $\dot{x} = Ax + Bu, u \in U \subset R^m, m \leqq n$, where U contains 0 in its interior, see [1].

with a time-optimal feedback under perturbations, which, without having been precisely defined, has commonly been intuitively felt.

The attempt to formulate and prove rigorously this stability in some form has motivated this research; another motivation came from linear differential games, where any realistic approach cannot avoid the use of closed-loop strategies.

Working with the mathematical model of the system under the closed-loop optimal control

$$\dot{x} = Ax + v(x), \tag{2}$$

one has first to answer the question, whether all the solutions of (2) are actually optimal trajectories. The problem here is that v is in general discontinuous and thus for (2), neither existence nor unicity of solutions follows from the standard theory.

The answer to the converse question, whether the optimal trajectories of (1) are solutions of (2), seems to be trivial and is answered affirmatively in [6]. This is actually so as far as the solutions in the standard (i.e., Carathéodory) sense are meant. However, it has been demonstrated by several authors that for differential equations with discontinuities in the dependent (state) variable, the Carathéodory concept of solution is inappropriate. Namely, it does not characterize all the movements that can occur in systems modeled by such equations (the so-called chattering or sliding regimes, in particular) and, conversely, not all Carathéodory solutions have a physical significance.

For discontinuous differential equations, apparently the most perfect definition of solution is that of Filippov [4]: One associates with the differential equation in R^n

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

where $F(x) = \bigcap_{\delta > 0} \bigcap_{\mu(N) = 0} \operatorname{co} \operatorname{cl} f(B(x, \delta) - N)$, where $B(x, \delta) = \{y \mid |y - x| < \delta\}$, set of D, a multivalued differential equation

$$\dot{x} \in F(x) \tag{4}$$

where $F(x) = \bigcap_{\delta > 0} \bigcap_{\mu(N) = 0} \operatorname{co} \operatorname{cl} f(B(x, \delta) - N)$, where $B(x, \delta) = \{y \mid |y - x| < \delta\}$, $|\cdot|$ is the Euclidean norm in R^n and μ is the Lebesgue measure in R^n. A function $\varphi: I \to R^n$, I an interval of R, is called a (Filippov) solution of (3) if it is a solution of (4) on I in the usual sense (i.e., it is absolutely continuous and $\dot{\varphi}(t) \in F(\varphi(t))$ for a.e. $t \in I$). Let us note that it can be proven that F is upper semicontinuous with convex compact values and therefore satisfies the usual existence conditions (cf. [2]). For the motivation of the definition, the reader is referred to [4], where further references to sources leading to this definition can also be found.

This definition of solution is not easy to deal with. However, in this paper, we shall need it only in special local situations:

(i) f is continuous in some neighborhood B of x_0.

Then, $F(x_0) = \{f(x_0)\}$ for $x \in B$, and the Filippov solutions through x_0 locally coincide with the Carathéodory ones.

(ii) There is a neighborhood B of x_0 and a smooth submanifold of codimension one $\Sigma = \{x \mid s(x) = 0\}$ of B $((\partial s/\partial x)(x) \neq 0$ for $x \in B)$ such that $B = S^+ \cup \Sigma$

$\cup S^-$, where $S^+ = \{x|s(x) > 0\}$, $S^- = \{x|s(x) < 0\}$ and

$$f(x) = \begin{cases} f^+(x) & \text{for } x \in S^+, \\ f^-(x) & \text{for } x \in S^-, \end{cases} \tag{5}$$

f^+ and f^- being continuous on $S^+ \cup \Sigma$ and $S^- \cup \Sigma$, respectively, and satisfying

$$\left\langle \frac{\partial s}{\partial x}(x), f^+(x) \right\rangle < 0, \quad \left\langle \frac{\partial s}{\partial x}(x), f^-(x) \right\rangle \leqq 0 \tag{6}$$

for $x \in \Sigma$ (note that the definition of f on Σ is irrelevant since by the Filippov theory, directions on any set of measure zero cannot be sensed by a system and therefore are not counted). Geometrically this means that x_0 lies on a surface of discontinuity of f and the limit direction vectors from both sides of this surface point to one side of the surface. In this case, the solutions of (4) through x_0 coincide locally with the solutions of $\dot{x} = f^-(x)$(see Fig. 1).

(iii) This case is like (ii) with (6) replaced by

$$\left\langle \frac{\partial s}{\partial x}(x), f^+(x) \right\rangle < 0, \quad \left\langle \frac{\partial s}{\partial x}(x), f^-(x) \right\rangle > 0 \quad \text{for } x \in \Sigma.$$

In this case, a continuous vector field σ on Σ is defined as follows: $\sigma(x)$ is the intersection of the segment joining $f^+(x)$ and $f^-(x)$ with the tangent hyperplane to Σ at x; the solution of (4) through x_0 is locally that of σ. This trajectory is called the *sliding* or *chattering* trajectory (see Fig. 2).

The problem we are going to study in this paper can now be formulated as follows:

Are all the (Filippov) solutions of the discontinuous differential equation

$$\dot{x} = Ax + v(x), \tag{2}$$

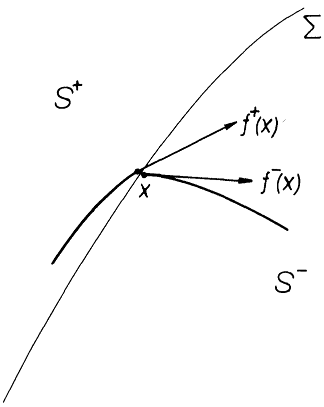

FIG. 1

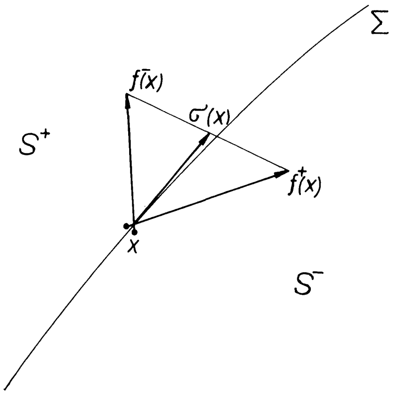

FIG. 2

or, in other words, are all the solutions of the multivalued differential equation

$$\dot{x} \in Ax + V(x),$$

where $V(x) = \bigcap_{\delta > 0} \bigcap_{\mu(N) = 0} \text{co cl } v(B(x, \delta) - N)$, optimal trajectories of (1) and, conversely, are all the optimal trajectories of (1) solutions of (2)? The problem of stability of the system (2) under perturbations will be the subject of part II of this paper.

However basic the question formulated in the last paragraph is, the answer to it is not easy. This paper presents a complete answer in dimension 2. The intricacy of the details of the analysis as well as the fact that the results in dimension 2 show that the systems for which the answer is negative are by no means pathological or extraordinary make the results appear rather pessimistic. Moreover, it seems that in higher dimensions the systems with nonoptimal sliding regimes are even more common, if not typical.

2. Main results. Let C be a convex subset of R^n. For $x \in C$ denote $H(x, C) = \{\psi \in R^n \mid \langle \psi, x \rangle = \max_{y \in c} \langle \psi, y \rangle\}$, $H_0(x, C) = \{\psi \in H(x, C) \mid |\psi| = 1\}$. For $\psi \in R^n$, denote $Q(\psi, C) = \{x \in C \mid \psi \in H(x, C)\}$. We have $H_0(x, C) \neq \emptyset$ if and only if x is a boundary point of C: x is called a regular boundary point of C if $H_0(x, C)$ is a one-point set, a singular boundary point otherwise. For a given convex set C, H, H_0 and Q are upper semicontinuous set-valued functions on C, C and R^n, respectively.

The following theorem gives an answer to the question raised in §1 for two-dimensional systems.

THEOREM. *Let $n = 2$. Then every trajectory of (2) is optimal and every optimal trajectory is a solution of (2) if and only if there does not exist a vertex w of U such that $H(w, U)$ contains the eigenvector of $-A'$ corresponding to its larger eigenvalue but does not contain the other eigenvector of $-A'$.*

So (2) can have nonoptimal trajectories only if A has two distinct real eigenvalues (since, otherwise, either there is no eigenvector or no larger eigenvalue). Further, (2) can have no nonoptimal solutions if $\dim U = 1$. For, in this case, U has two vertices w_1, and w_2, both $H(w_1, U)$ and $H(w_2, U)$ being half-spaces and therefore containing both eigenvectors of A' (provided these exist).

The rest of this section is devoted to the proof of the theorem. Note that Lemmas 1–7 are valid in any dimension, while the remaining three are valid only in dimension 2.

Denote by $\mathscr{R}(t)$, $t \geqq 0$, the set of points which can be steered to the origin in time t, i.e. $\mathscr{R}(t) = \{\int_0^t - e^{-sA}u(s)ds \mid u \text{ measurable with values in } U\}$. It is well known that the sets $\mathscr{R}(t)$ have the following properties.

Property 1. $\mathscr{R}(t)$ are compact, strictly convex, depend continuously on t (in the Hausdorff set topology) and have a nonempty interior for $t > 0$.

Property 2. $x \in \partial\mathscr{R}(t)$ (the boundary of $\mathscr{R}(t)$) if and only if $t = T(x)$, where $T(x)$ is the minimal time for steering x to 0.

Property 3. Denote $E(x) = -H(x, \mathscr{R}(T(x)))$. The open-loop optimal control $u_x(t)$ satisfies the maximum principle

$$\langle \psi(t), u_x(t) \rangle = \max_{u \in U} \langle \psi(t), u \rangle$$

for any solution $\psi(t)$ of the adjoint equation with initial value $\psi(0) = \psi \in E(x)$. In other words, if

$$x = -\int_0^{T(x)} e^{-tA}u_x(t)\,dt, \tag{7}$$

and $\xi_x(t)$ is the optimal trajectory from x, then

$$\langle e^{-tA'}\psi, u_x(t) \rangle = \max_{u \in U} \langle e^{-tA'}\psi, u \rangle \quad \text{for } t \in [0, T(x)]$$

and $\langle e^{-tA'}\psi, A\xi_x(t) + u_x(t) \rangle \geqq 0$ is constant for all $\psi \in E(x)$, $\psi \neq 0$.

The proofs of Properties 1–3 can be found, e.g., in [6].

Denote $\mathscr{R} = \bigcup_{t \geqq 0} \mathscr{R}(t)$. $\mathscr{R}$ is open in R^n. The following lemma is an immediate consequence of Property 3.

LEMMA 1. *For all* $x \in \mathscr{R}$, $E(x) = -\bigcap_{t \in [0, T(x)]} e^{tA'}H(u_x(t), U)$.

COROLLARY 1. *Going backwards in time along an optimal trajectory, E is nonincreasing.*

We shall call a boundary point of a convex set C a *corner* if $H(x, C)$ has a nonempty interior. Note that in dimension 2, every singular boundary point is a corner. A point $x \in \mathscr{R}$ will be called a *corner point* if it is a corner of $\mathscr{R}(T(x))$.

LEMMA 2. *Let C be convex and* $x, y \in C$, $x \neq y$. *Then* $H(x, C) \cap H(y, C) = \partial H(x, C) \cap \partial H(y, C)$.

Proof. Assume that there exists a $\psi \in \operatorname{int} H(x, C) \cap H(y, C)$. Then, for sufficiently small $\kappa > 0$,

$$\psi + \kappa(y - x) \in H(x, C). \tag{8}$$

On the other hand, we have $\langle \psi + \kappa(y - x), y \rangle - \langle \psi + \kappa(y - x), x \rangle = \kappa\langle y - x, y - x \rangle > 0$, which contradicts (9).

LEMMA 3. $x \neq 0$ *can be a corner point in* $\mathscr{R}$ *only if* $u_x(t)$ *is constant on* $[0, T(x)]$.

Proof. If $u_x(t)$ is not constant, then there exist $0 \leqq t_1 < t^* < t_2 \leqq T(x)$ such that $u_x(t) = w_1$ on $[t_1, t^*)$ and $u_x(t) = w_2$ on $[t^*, t_2)$. We have, by Lemma 1,

$$E(x) \subset \bigcap_{t \in [t_1, t^*)} e^{tA'}H(w_1, U) \cap \bigcap_{t \in [t^*, t_2)} e^{tA'}H(w_2, U). \tag{9}$$

Obviously, $e^{tA'}H_0(w, U)$ depends continuously on t, from which it follows that

$$\bigcap_{t\in[t_1,t)} e^{tA'}H(w_1, U) = \bigcap_{t\in[t_1,t^*]} H(w_1, U).$$

From this and (9) we obtain $E(x) \subset e^{t^*A}[H(w_1, U) \cap H(w_2, U)]$. By Lemma 2, $H(w_1, U) \cap H(w_2, U) = \partial H(w_1, U) \cap \partial H(w_2, U)$ and, therefore, $E(x)$ cannot contain an interior point.

COROLLARY 2. *Every corner point belongs to one of the curves*

$$\Gamma(w) = \left\{ - \int_0^t e^{-sA} w \, ds \,\middle|\, 0 \leqq t < \tau(w) \right\},$$

where w is a vertex of U and $\tau(w) = \max\{t \mid \bigcap_{0\leqq s \leqq t} e^{sA'}H(w, U) \neq 0\}$.

LEMMA 4. *If $\psi \in \operatorname{int} E(x)$ or $\psi \in E(x), \psi \neq 0$ and* $\dim U = n$, *then* $\langle \psi, Ax + v(x)\rangle > 0$.

Proof. By Property 3, $\langle \psi, Ax + v(x)\rangle \geqq 0$ for every $\psi \in E(x)$. Assume first that $Ax + v(x) = 0$. Then, $\langle \chi, Ax + v(x)\rangle = 0$ and, by Property 3, also $\langle e^{-T(x)A'}\chi, u_x(T(x))\rangle = 0$ for all $\chi \in E(x)$ or equivalently, $\langle \varphi, u_x(T(x))\rangle = 0$ for all $\varphi \in e^{-T(x)A'} \cdot E(x)$. Since, by assumption, $E(x)$ has a nonempty interior, so does $e^{-T(x)A'}E(x)$, which implies $u_x(T(x)) = 0$. This, however, contradicts the bang-bang principle.

For $Ax + v(x) \neq 0$, there exists an $\varepsilon > 0$ such that $\psi \pm \varepsilon(Ax + v(x)) \in E(x)$. If $\langle \psi, Ax + v(x)\rangle = 0$, then $\langle \psi \pm \varepsilon(Ax + v(x)), Ax + v(x)\rangle = \pm|Ax + v(x)|^2$. Since $Ax + v(x) \neq 0$, one of these numbers would be negative, which contradicts Property 3.

If $\dim U = n$ and $\psi \neq 0$, $\psi \in E(x)$, then by Property 3,

$$\langle \psi, Ax + v(x)\rangle = \max_{u\in U} \langle \psi, Ax + u\rangle = \max_{u\in U} \langle e^{-T(x)A'}\psi, u\rangle > 0.$$

LEMMA 5. *E is upper semicontinuous.*

Proof. See [6, Lemma 8] for the proof.

Denote $E_0(x) = \{\psi \in E(x) \mid |\psi| = 1\}$, $W(x) = \operatorname{co} \bigcap_{\psi\in E_0(x)} Q(\psi, U)$.

LEMMA 6. *W is upper semicontinuous. $W(x)$ is closed and $V(x) \subset W(x)$ for every $x \in \mathscr{R}$.*

Proof. Assume $x_k \to x$, $u_k \in W(x_k)$, $u_k \to u_0$. Then there exist $\psi_{k1}, \cdots, \psi_{k,n+1} \in E_0(x_k)$, $y_{k1}, \cdots, y_{k,n+1} \in Q(\psi_{ki}, U)$ and $\lambda_{k1}, \cdots, \lambda_{k,n+1} \in [0, 1]$ such that $u_k = \sum_i \lambda_{ki} y_{ki}$, $\sum_i \lambda_{ki} = 1$. Extracting a subsequence if necessary, we may assume $\psi_{ki} \to \psi_i$, $y_{ki} \to y_i$, $\lambda_{ki} \to \lambda_i \in [0, 1]$, $\sum_i \lambda_i = 1$. From Lemma 5, it follows that $\psi_i \in E_0(x)$ and, from the upper semicontinuity of Q, $y_i \in Q(\psi_i, U)$. Thus $u_0 = \sum_i \lambda_i y_i \in W(x)$, which proves the upper semicontinuity of W.

From the upper semicontinuity and boundedness of Q, it follows that $\bigcup_{\psi\in E_0(x)} Q(\psi, U)$ is compact and, consequently, that $W(x)$ is closed.

To prove the inclusion $V(x) \subset W(x)$, we note that $u \in V(x)$ means that there exist sequences $x_{ki} \to x$, $\lambda_{ki} \in [0, 1]$, $i = 1, \cdots, n + 1$, such that $\sum_i \lambda_{ki} = 1$ and $u = \lim_k \sum_i \lambda_{ki} v(x_{ki})$. By Property 3 and the definition of v, $v(x_{ki}) \in Q(\psi, U)$ for all $\psi \in E(x_{ki})$. Turning to a subsequence if necessary, we can choose from $E_0(x_{ki})$ convergent subsequences $\{\psi_{ki}\}$ whose limits ψ_i, due to Lemma 5, belong to $E_0(x)$.

Also, we may assume that $\lambda_{ki} \to \lambda_i \in [0,1]$, $\sum_i \lambda_i = 1$ and $v(x_{ki}) \to v_i$. By upper semicontinuity of Q, $v_i \in Q(\psi_i, U)$. Thus, $u = \sum_i \lambda_i v_i \in \operatorname{co}(\bigcup_i Q(\psi_i, U)) \subset W(x)$.

LEMMA 7. *If* $E_0(x_0) = \{\psi_0\}$ (in dimension 2, this means that x_0 is not a corner point) *and* $\max_{u\in U}\langle e^{-T(x)A'}\psi_0, u\rangle$ $(= \max_{u\in U}\langle \psi_0, Ax_0 + u\rangle) > 0$, *then* T *is differentiable at* x_0 *and* $(\partial T/\partial x)(x_0) = -[\max_{u\in U}\langle \psi_0, Ax_0 + u\rangle]^{-1}\psi_0$.

Proof. This lemma is a consequence of [9, Thm. 3.2].

Thanks to normality, for a given $\psi_0 \neq 0$, one obtains a unique trajectory $\Delta(\psi_0) = \{\delta(t) \mid t \leqq 0\}$ by choosing $u(t)$ $(t \leqq 0)$ so as to satisfy the maximum principle with the solution of the adjoint equation $\psi(t) = e^{-tA'}\psi_0$ $(\psi(0) = \psi_0)$.

We have $\xi_{\delta(t)}(s) = \delta(t+s)$, $s \in [0, -t]$, and $\delta(t) = -\int_t^0 e^{(t-s)A'}u(s)\,\mathrm{d}s$, $u(s) \in Q(e^{-sA'}\psi_0, U)$.

Denote by $U^\perp = \{\psi_0 \mid \max_{u\in U}\langle \psi_0, u\rangle = 0\}$, $\Gamma = \bigcup\{\Gamma(w) \mid w$ is a vertex of $U\}$ and $\Delta = \bigcup\{\Delta(\psi_0) \mid \psi_0 \in U^\perp, \psi_0 \neq 0\}$. Note that Δ is empty if $\dim U = n$.

In the rest of the paper we shall assume $n = 2$ without further notice.

LEMMA 8. *The set of points of* $\mathscr{R}$ *at which* T *is not differentiable is contained in* $\Gamma \cup \Delta$.

Proof. If $E_0(x)$ contains more than a point, then according to Corollary 2, $x \in \Gamma$; $\max_{u\in U}\langle e^{-T(x)A'}\psi_0, u\rangle = 0$ for $\{\psi_0\} = E_0(x)$ means $x \in \Delta(e^{-T(x)A'}\psi_0)$, $\psi_0 \in U^\perp$. Lemma 7 completes the proof.

For the next lemma, let us choose a coordinate system in R^2, and for a given point $x \in R^2$, denote its coordinates by x^1, x^2.

LEMMA 9. *Let* w *be a vertex of* U, $x_0 \in \Gamma(w)$ *and* $e_1 = (1,0) \in \operatorname{int} E(x_0) \subset H(w, U)$. *Denote by* ψ_1, ψ_2 *the unit boundary vectors of* $E(x_0)$ *in such a way that* $\psi_1^2 > 0 > \psi_2^2$. *Then there exist a neighborhood* B *of* x_0 *and a smooth function* γ *defined in some neighborhood of* x_0^1 *such that*

(i) $B \cap \Gamma(w^*) = \emptyset$ *for any vertex* $w^* \neq w$ *of* U,

(ii) $B \cap \Gamma(w) = \{(x^1, x^2) \in B \mid x^2 = \gamma(x^1)\}$,

(iii) $\Gamma(w) \cap B$ *divides* $B - \Gamma(w)$ *into disjoint parts*

$$B^+ = \{(x^1, x^2) \in B \mid x^2 > \gamma(x^1)\}, \quad B^- = \{(x^1, x^2) \mid x^2 < \gamma(x^1)\},$$

(iv) *for* $x \in B^+$ (*resp.* B^-), $\psi \in E_0(x)$, $\psi^2 < 0$ (*resp.* $\psi^2 > 0$),

(v) $y(x_0) = \kappa_1\psi_1 - \kappa_2\psi_2$, *where* $\kappa_1 \geqq 0, \kappa_2 \geqq 0$ *and* $y(x) = (1, (d\gamma/dx_1)(x^1))$ *is the normal to* Γ *at* x; $\kappa_1 > 0$ *if* $\dim U = 2$.

Proof. (i) is obvious. (ii) and (iii) follow from the fact that $Ax_0 + w$ is tangent to $\Gamma(w)$ and by Lemma 4, $(Ax_0 + w)^1 = \langle e_1, Ax_0 + w\rangle > 0$.

If B is chosen to be a sufficiently small ball, then from $x \in B^+$ it follows that $\langle x^1, \gamma(x^1)\rangle \in \mathscr{R}(T(x))$. If $\psi \in E(x)$, from this and the strict convexity of $\mathscr{R}(T(x))$, we obtain $0 > \langle \psi, x - (x^1, \gamma(x^1))\rangle = \psi^2(x^2 - \gamma(x^1))$, which implies $\psi^2 < 0$ and thus proves (iv) for $x \in B^+$. The proof of the B^- part of (iv) is similar.

Since $y(x)$ is orthogonal to Γ at x, we have $\langle y(x), Ax + w\rangle = 0$, which by Lemma 4 implies that neither y nor $-y$ can belong to $\operatorname{int} E(x)$ and, in case $\dim U = 2$, to $E(x)$; this proves (v).

For $x \in R - \Gamma$, $E_0(x)$ consists of one point, which we denote by $\Psi(x)$.

LEMMA 10. *Let* $x \in \mathscr{R} - \Gamma$. *Then* $0 \notin Ax + W(x)$.

Proof. Either $\Psi(x) \in \operatorname{int} H(w, U)$ for some vertex w of U or $\Psi(x) = H_0(w_1, U) \cap H_0(w_2, U)$ for two distinct vertices w_1, w_2 of U. In the first case, $Ax + W(x) = \{Ax + w\} = \{Ax + v(x)\}$, which cannot be zero by Lemma 4. In the second case, $Ax + W(x) = \operatorname{co}\{Ax + w_1, Ax + w_2\}$, and

$$\langle \Psi(x), w_1 - w_2 \rangle = 0. \tag{10}$$

$0 \in \operatorname{co}\{Ax + w_1, Ax + w_2\}$ would mean that there exists a $\lambda \in [0, 1]$ such that $Ax + \lambda w_1 + (1 - \lambda)w_2 = 0$, from which and (10) we obtain $\langle \Psi(x), Ax + w_1 \rangle = \langle \Psi(x), (1 - \lambda)(w_2 - w_1) \rangle = 0$ and, similarly, $\langle \Psi(x), Ax + w_2 \rangle = 0$. Since $v(x) =$ either w_1 or w_2, $\langle \Psi(x), Ax + v(x) = 0$. From Lemma 4 it follows that in this case, $\dim U = 1$. Choosing proper coordinates, we may assume that

$$U = \operatorname{co}\{w_1, w_2\}, \quad w_1 = \alpha e_1, \quad w_2 = -\beta e_1, \qquad \alpha, \beta > 0.$$

By Property 3 we have

$$\langle e^{-T(x)A'}\Psi(x), u_x(T(x)) \rangle = \langle \Psi(x), Ax + v(x) \rangle = 0. \tag{11}$$

Since $u_x(T(x))$ is w_1 or w_2, from (9) and (10) we obtain that $\Psi(x)$ as well as $e^{-T(x)A'}\Psi(x)$ are nonzero multiples of $e_2 = (0, 1)$. From the normality it follows that both half-lines $\{\kappa\Psi(x) | \kappa > 0\}$ and $\{\kappa\Psi(x) | \kappa < 0\}$ have to be crossed transversally by the solutions of the adjoint equation, each of them in one direction. Therefore, if t_1 is the first switching time of u_x, i.e., the first zero of $\max_{u \in U}\langle e^{-tA'}\Psi(x), u \rangle$ (which is a zero of $\langle e^{-tA'}\Psi(x), e_1 \rangle$) on $(0, T(x)]$, $e^{-t_1A'}\Psi(x)$ has to be a negative multiple of $\Psi(x)$. Note that t_1 exists, because $\langle e^{-T(x)A'}\Psi(x), e_1 \rangle = 0$.

Going through the phase portraits of linear two-dimensional equations (see, e.g., [3]) one finds that this is possible only if A has a pair of complex eigenvalues and, therefore, the switching times are multiples of t_1.

Let $0 < t_1 < t_2 < \cdots < t_p < T(x)$ be the switching times of $u_x(t)$. Assume without loss of generality that $u_x(t) = w_1$ on $[t_p, T(x)]$, $u_x(t) = w_2$ on $[t_{p-1}, t_p)$, etc. On every interval $[t_i, t_{i+1})$, $\xi_x(t)$ satisfies the linear differential equation $\dot\eta = A\eta$. Therefore, from $u_x(T(x)) = \dot\xi_x(T(x)) = w_1$, we obtain $A\xi_x(t_p) + v(\xi_x(t_p)) = A\xi_x(t_p) + w_1 = \dot\xi_x^+(t_p) = -\nu_p w_1 = -\mu_p v(\xi_x(t_p))$, where $\nu_p > 0$, $\mu_p > 0$ and $\dot\xi_x^+$ stands for the right derivative of ξ_x. Therefore, the left derivative $\dot\xi_x^-(t_p) = \lim_{t \uparrow t_p}\dot\xi_x(t) = A\xi_x(t_p) + w_2 = -\mu_p w_1 - w_1 + w_2 = \sigma_p w, \sigma_p > 0$. Replacing $T(x)$ by t_p and t_p by t_{p-1}, we obtain, as above, $A\xi_x(t_{p-1}) + v(\xi_x(t_{p-1})) = -\mu_{p-1}v(\xi_x(t_{p-1}))$, $\mu_{p-1} > 0$. Repeating this argument p times, we finally obtain $Ax + v(x) = -\mu v(x), \mu > 0$. Without loss of generality, assume $v(x) = w_1$. We then have by (10) that for any $0 < \lambda < 1$, $\lambda(Ax + w_1) + (1 - \lambda)(Ax + w_2) = Ax + w_1 + (1 - \lambda) \cdot (w_2 - w_1) = -\mu w_1 + (1 - \lambda)(w_2 - w_1) = [-\mu - (1 - \lambda)(\alpha + \beta)]e_1 < 0$, so $0 \in \operatorname{co}\{Ax + w_1, Ax + w_2\}$.

Proof of the theorem. We prove two statements.

1°. Any solution $\varphi(t)$ of (2) starting at a point $x_0 \in \mathscr{R} - (\bar\Gamma \cup \Delta)$ coincides with $\xi_{x_0}(t)$ for sufficiently small t, and the same is true if $x_0 \in \Delta(\psi_0) - \Gamma$ for $\psi_0 \in U^\perp$, $\psi_0 \neq 0$ and $e^{T(x)}\psi_0 = \psi(x_0)$ is interior to some $H(w, U)$.

2°. For $x_0 \in \Gamma(w)$, w a vertex of U, we show that the trajectory of (2) from x_0 coincides with $\xi_{x_0}(t)$ for sufficiently small t if and only if w does not satisfy the condition of the theorem.

The set P of the points x_0 which are not taken care of in 1° and 2° consists of the points of $\bar\Gamma - \Gamma$ (which is discrete in R) and of the points x_0 of $\Delta(\psi_0) - \Gamma$ for

$\psi_0 \neq 0$, $\psi \in U^\perp$ with $\Psi(x_0) = e^{T(x_0)A'}\psi_0 \in U^\perp$ (which, due to the normality, is also discrete in $\mathscr{R}$). Therefore, by Lemmas 8 and 10, for any trajectory $\varphi(t)$ of (2), $\varphi^{-1}(P)$ cannot contain an interval. If no vertex of U satisfies the conditions of the theorem, it follows from 1° and 2° that $T(\varphi(t))$ has to be strictly decreasing in the open intervals separating the points of $\varphi^{-1}(P)$ and, therefore, no trajectory of (2) can return to one point of P twice. As a consequence, we obtain that $\varphi^{-1}(P)$ is discrete, from which, in addition to 1° and 2°, the theorem follows immediately. So it remains to prove 1° and 2°.

Proof of 1°. Since $\mathscr{R} - (\bar{\Gamma} \cup \Delta)$ is open, if $x_0 \in \mathscr{R} - (\bar{\Gamma} \cup \Delta)$, then some neighborhood B of x_0 is contained in $\mathscr{R} - (\bar{\Gamma} \cup \Delta)$. By Lemma 8, T is differentiable in B. For sufficiently small t, we have $\varphi(t) \in B$ and, therefore, $dT(\varphi(t))/dt = (\partial T/\partial x)(\varphi(t))(A\varphi(t) + w(t))$ for almost all sufficiently small t, where $w(t)$ is measurable with values in $W(\varphi(t))$. Since $(\partial T/\partial x)(\varphi(t)) \in E(\varphi(t))$ and $E(\varphi(t))$ is a half-line for $\varphi(t) \in B$, we have $(\partial T/\partial x)(\varphi(t))(A\varphi(t) + w(t)) = \max_{u \in U} (\partial T/\partial x)(\varphi(t))(A\varphi(t) + u) = 1$. This implies $T(\varphi(t)) - T(x_0) = t$ for small t, which means that $\varphi(t)$ is optimal.

If $x_0 \in \Delta(\psi_0) - \Gamma$, $\psi_0 \neq 0$, and $\Psi(x) = e^{T(x_0)A'}\psi_0$ is interior to $H(w, U)$, then by Lemma 5, $\Psi(x)$ is contained in the interior of $H(w, U)$ for x from some neighborhood B of x_0, and therefore $v(x)$ is constant in B and equal to w. Thus, by (i) in § 1, $V(x) = \{w\}$ for $x \in B$, and the solution of (2) is a solution of $\dot{x} = Ax + w$ and, consequently, is optimal.

Proof of 2°. Let $x_0 \in \Gamma(w)$. Choosing suitable coordinates, we may assume that the conditions of Lemma 9 are satisfied. Denote by χ_1, χ_2 the unit boundary vectors of $H(w, U)$ in such a way that $\chi_1^2 \geqq \psi_1^2 > 0 > \psi_2^2 \geqq \chi_2^2$. Since, due to normality, the vectors χ_1, χ_2 cannot be eigenvectors of A', the trajectories of the adjoint equation cross them transversally. According to whether they are crossed inwards or outwards with respect to $H(w, U)$, the following three cases have to be distinguished.

Case 1. Both vectors χ_1, χ_2 are outside $e^{tA'}H(w, U)$ for all $t > 0$ sufficiently small.

Case 2. One of the vectors χ_1, χ_2 is contained and the other is not contained in $e^{tA'}H(w, U)$ for small $t > 0$.

Case 3. Both χ_1 and χ_2 are contained in $e^{tA'}H(w, U)$ for small $t > 0$.

We prove that in Cases 1 and 2, the only trajectory of (2) from x_0 coincides locally with the optimal trajectory, while in Case 3, the solution of (2) from x_0 (which is unique in this case) always differs from the optimal trajectory.

In Case 1, $E_0(x)$ is contained in the interior of $H(w, U)$ for $x \in \Gamma(w)$ and, by Lemma 5, also for x from some neighborhood B of x_0. Thus, by Lemma 6, $V(x) = \{w\} = \{v(x)\}$ for $x \in B$. Thus, the trajectory of (2) from x_0 must coincide with the optimal trajectory in B.

To analyze Case 2, we assume that χ_1 is outside and χ_2 inside $E(x)$ (the other case is similar). Then, by Lemma 9, for $x \in B^-$, $\psi \in E_0(x)$ implies $\psi_2 > 0$. Using, in addition, Lemma 5, we obtain that $\psi \in H(w, U)$, provided B is suitably restricted, which implies $v(x) = w$ for $x \in B^-$. For $x \in B^+$, $\psi \in E_0(x)$, we have according to Lemma 9, $\psi^2 < 0$. Since $\mathscr{R}(t)$ are strictly convex, $\psi \notin E(x^*)$ where $\{x^*\} = \Gamma \cap \mathscr{R}(T(x))$. By choosing B sufficiently small, x^* will be arbitrarily close to x_0 for $x \in B$, and therefore $E(x^*)$ will contain e_1 and also χ_2. From this, we obtain $\psi \notin H(w, U)$ and therefore $\psi \in H(w_1, U)$, where w_1 is the vertex adjacent to w in the clockwise direction. However, $\psi \in H(w_1, U)$ implies $v(x) = w_1$. Consequently,

$$V(x) = \begin{cases} \{w\} & \text{for } x \in B^-, \\ \{w_1\} & \text{for } x \in B^+ \\ \operatorname{co}\{w, w_1\} & \text{for } x \in \Gamma(w) \cap B. \end{cases}$$

Since $\psi_2 = \chi_2$ is common to $H(w, U)$ and $H(w_1, U)$ and $\psi_1 \in \operatorname{int} H(w, U)$, we have by Lemma 9

$$\langle y(x_0), w_1 - w\rangle = \kappa_1\langle \psi_1, w_1 - w\rangle = -q\kappa_1, \tag{12}$$

where q is positive.

We prove $\kappa_1 > 0$. Assume $\kappa_1 = 0$. Then, $y(x_0) = -\kappa_2\psi_2, \kappa_2 \neq 0$, which implies $\langle \psi_2, Ax + w\rangle = 0$. Property 3 gives $\langle e^{-tA'}\psi_2, A\xi_{x_0}(t) + w\rangle = 0$. However, $e^{-tA'}\psi_2$ is interior to $E(x(t))$ for $t > 0$ sufficiently small, which contradicts Lemma 4.

We now prove that the local situation at x_0 is of type (ii) in § 1, from which we obtain that the solution of (2) from x_0 will be, for small t, identical with the trajectory of $\dot{x} = Ax + w$, i.e., with the optimal trajectory.

Using (11) and $\kappa_1 > 0$, from $\langle y(x), Ax + w\rangle = 0$, we obtain

$$\langle y(x), Ax + w_1\rangle < 0 \tag{13}$$

for $x = x_0$. By continuity of y, (13) will remain true for $x \in B$ for B suitably restricted. Since $y(x) = (\partial s/\partial x)(x)$, (13) means $\langle (\partial s/\partial x)(x), Ax + w\rangle < 0$.

In Case 3, the situation in B^- is analogous to that in B^- which is the same as in Case 2. Therefore, we obtain

$$V(x) = \begin{cases} \{w_2\} & \text{for } x \in B^-, \\ \{w_1\} & \text{for } x \in B^+, \\ \operatorname{co}\{w_1, w_2\} & \text{for } x \in \Gamma(w) \cap B, \end{cases}$$

where w_2 is the vertex adjacent in the counterclockwise direction. As in Case 2, we obtain $\langle y(x), Ax + w_2\rangle > 0$, $\langle y(x), Ax + w_1\rangle < 0$ for $x \in B$. Thus, the situation at x_0 is that of (iii) in § 1 and the solution of (2) will move along $\Gamma(w)$ with a speed that is obtained as the intersection of co $\{Ax + w_1, Ax + w_2\}$ with $Ax + w$, which is certainly smaller than $Ax + w$ (otherwise, $Ax + w_1$, $Ax + w, Ax + w_2$ and, consequently, w_1, w, w_2 would have to lie on a line, which is not true).

To finish the proof of the theorem, we have only to find out when Case 3 happens. Going through the phase pictures of two-dimensional linear differential equations (see, e.g., [3]) one finds that Case 3 occurs precisely if w satisfies the conditions of the theorem.

Remark. It follows from the proof of the theorem that if (2) has sliding trajectories, these are never optimal (these sliding trajectories should not be confused with the "optimal sliding regimes" of [5], which refer to open-loop optimal controls).

3. Examples. The following two simple examples illustrate the phenomena of nonoptimal sliding. Further examples can be found among the examples of [1].

Example 1.

$$\dot{x}^1 = -x^1 + u^1,$$
$$\dot{x}^2 = x^2 + u^2,$$

$U = \{(u^1, u^2) \mid |u^1| + |u^2| \leqq 1\}$. Denote $w_1 = (1, 0)$, $w_2 = (0, 1)$, $w_3 = (-1, 0)$, $w_4 = (0, -1)$. One finds that

$$v(x) = \begin{cases} w_1 & \text{if } x^1 < 0, \quad x^2 = 0, \\ w^2 & \text{if } x^2 < 0, \\ w_3 & \text{if } x^1 > 0, \quad x^2 = 0, \\ w_4 & \text{if } x^2 > 0. \end{cases} \tag{14}$$

The sliding trajectories of (2) appear on the x^1-axis: we have $V(x^1, 0) = \operatorname{co}\{(-x^1, 0) + w_2, (-x^1, 0) + w_4\}$. Thus, the sliding trajectories are solutions of $\dot{x}^1 = -x^1$ instead of $\dot{x}^1 = -x^1 \pm 1$. Note that the sliding trajectories in this case need an infinite time to reach the origin.

Example 2. The system is as in Example 1 with $U = \operatorname{co}\{w_1, w_2, w_3, w_4\}$, where $w_1 = (1, 0)$, $w_2 = (\frac{1}{2}, 1)$, $w_3 = (-1, 0)$, $w_4 = (\frac{1}{2}, -1)$. Also in this case, v is defined by (10). However, in this case the sliding trajectories satisfy the differential equation $\dot{x}^1 = \frac{1}{2} - x^1$, which means, that on the interval $[0, \frac{1}{2}]$ of the x^1-axis the system, instead of moving towards the origin, will slide away from it, towards the point $(\frac{1}{2}, 0)$.

REFERENCES

[1] V. G. Bolťanskij, *The Mathematical Theory of Optimal Processes*, Nauka, Moscow, 1966.

[2] P. Brunovský, *On the best stabilizing control under a given class of perturbations*, Czech. Math. J., 15 (1965), pp. 329–369.

[3] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[4] A. F. Filippov, *Differential equations with discontinuous right-hand sides*, Mat. Sb., 51 (1960), pp. 99–128.

[5] R. V. Gamkrelidze, *On sliding optimal states*, Dokl. Akad. Nauk SSSR, 143 (1962), pp. 1243–1245.

[6] O. Hájek, *Geometric theory of time-optimal control*, this Journal, 9 (1971), pp. 339–350.

[7] H. Hermes and J. P. LaSalle, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[8] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[9] B. N. Pšeničnyj, *Linejnyje differenciaľnyje igry*, Avtomat. i Telemeh., 1968, no. 1, pp. 65–78.

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

METHODES DE DECOMPOSITION POUR LA MINIMISATION D'UNE FONCTION SUR UN ESPACE PRODUIT*

B. MARTINET† ET A. AUSLENDER‡

Abstract. Methods are studied for minimizing a convex functional on a closed convex set in a product space, which can be considered as extensions of the Gauss–Seidel method. In this well-known method, a partial optimization problem must be solved at each step. This paper shows that the first step of an optimization method as gradient method, penalization method, etc., is in fact sufficient.

Soit H_i, $i = 1, 2, \cdots, n$, des espaces de Hilbert réels. On note le produit scalaire en H_i par $(\cdot, \cdot)_i$ et la norme correspondante par $\|\cdot\|_i$. On considère l'espace produit $H = \prod_{i=1}^n H_i$ sur lequel on définit un produit scalaire $(\cdot, \cdot)$ et une norme $\|\cdot\|$ tels que:

$$(x, y) = \sum_{i=1}^{n} (x_i, y_i)_i, \qquad \|x\|^2 = \sum_{i=1}^{n} \|x_i\|_i^2$$

avec $x_i, y_i \in H_i$ et $x = (x_1, x_2, \cdots, x_n)$ et $y = (y_1, y_2, \cdots, y_n)$. Soit des ensembles C_i qui sont des convexes fermés de H_i et $C = \prod_{i=1}^n C_i$. Etant donnée une fonction f à valeurs réelles définie sur H, on se propose de résoudre le problème d'optimisation suivant:

$$\text{(P)} \qquad \text{Trouver} \quad u^* \in C \quad \text{tel que} \quad f(u^*) = \inf(f(u), u \in C) = f^*.$$

On suppose que:

(i) La fonction f est fortement convexe, c'est-à-dire qu'il existe $K > 0$ tel que pour tout $x_1, x_2 \in H$ et pour tout $\lambda \in [0, 1]$ on ait:

$$f(\lambda x_1 + (1 - \lambda)x_2) \leqq \lambda f(x_1) + (1 - \lambda)f(x_2) - K\lambda(1 - \lambda)\|x_1 - x_2\|^2.$$

(ii) La dérivée f' de f existe et vérifie une condition de Lipschitz sur tout borné de H, c'est-à-dire que pour tout borné B de H il existe une constant $M > 0$ telle que:

$$\|f'(u) - f'(v)\| \leqq M\|u - v\| \quad \text{pour tout} \quad u, v \in B.$$

Ces hypothéses permettent de montrer que (P) admet une solution unique notée u^*.

En fait, on peut faire des hypothèses moins sévères, en particulier, supposer que la fonction f est fortement convexe séparément par rapport à chaque variable (cf. Martinet et Auslender [11]).

On emploie très souvent pour résoudre de façon approchée le problème (P) des méthodes de décomposition qui ont été étudiées entre autres par Auslender [1], Cea [2], Martinet [8], Miellou [12], Rheinboldt et Ortega [12],

* Received by the editors December 30, 1971, and in revised form December 12, 1972.

† U.E.R. de Mathématiques, Université de Provence, 13-Marseille, France.

‡ Département de Mathématiques Appliquées, Clermont-Ferrand, Boite Postale n° 45, 63-Aubiere, France.

Schechter [14]. Certaines de ces méthodes généralisent le procédé bien connu de Gauss-Seidel [4] (méthodes de relaxation).

Les procédés originaux que nous présentons dans cet article permettent de construire à partir d'un élément u^0 de H une suite $\{u^m\}$ d'éléments de H qui convergent vers la solution unique u^* du problème (P). On note:

$$u^m = (u_1^m, u_2^m, \cdots, u_n^m), \quad \text{où} \quad u_i^m \in H_i,$$

et pour tout $i = 1, 2, \cdots, n$:

$$\varphi_i^m(v_i) = f(u_1^{m+1}, u_2^{m+1}, \cdots, u_{i-1}^{m+1}, v_i, u_{i+1}^m, \cdots, u_n^m)$$

($\varphi_i'^m$ désigne alors la dérivée de φ_i^m).

Dans ces procédés,. on associe à la méthode de relaxation une certaine méthode d'optimisation M de la façon suivante.

On passe de u^m à u^{m+1} en déterminant successivement pour $i = 1, 2, \cdots, n$ l'élément u_i^{m+1} à partir de u_i^m en effectuant une seule itération de la méthode M appliquée au problème:

$$(\mathrm{P}_{im}) \qquad \min\,(\varphi_i^m(u_i)|u_i \in C_i).$$

Ceci présente souvent un grand avantage pratique par rapport à la méthode de relaxation habituelle où les problèmes (P_{im}) sont résolus exactement.

Au lieu de la relaxation, on pourrait de la même façon envisager des méthodes de décomposition en parallèle (cf. Auslender [1]). On peut montrer (cf. Martinet et Auslender [11]) qu'on peut ainsi associer à la relaxation la plupart des méthodes classiques d'optimisation. Pour alléger ce texte, nous n'étudions ici que deux exemples de tels procédés où on associe à la relaxation respectivement une méthode de gradient dans le cas sans contrainte et une méthode de pénalité extérieure.

1. 1ère méthode: Méthode de gradient associée à la relaxation dans le cas sans contrainte. On suppose ici que $C_i = H_i, i = 1, \cdots, n$.

Au cours des itérations, on passe de u^m à u^{m+1} en faisant successivement pour $i = 1, 2, \cdots, n$:

$$u_i^{m+1} = \begin{cases} u_i^m & \text{si} \quad \varphi_i'^m(u_i^m) = 0, \\ u_i^m - \rho_i^m \varphi_i'^m(u_i^m) & \text{sinon}, \end{cases}$$

ou si l'on note $\psi_i^m(\rho) = \varphi_i^m(u_i^m - \rho\varphi_i'^m(u_i^m))$ le nombre ρ_i^m est choisi de sorte que $\rho_i^m > 0$ et $\psi_i^m(\rho_i^m) = \inf(\psi_i^m(\rho)|\rho > 0)$. On peut alors montrer le résultat suivant qui justifie l'intérêt du procédé.

THÉORÈME. *La suite* $\{f(u^m)\}$ *est décroissante et la suite* $\{u^m\}$ *converge fortement vers la solution* u^* *de* (P).

Démonstration. On a, par construction, pour tout m et tout $i = 1, \cdots, n$:

$$f(u^{m+1}) \leqq \varphi_i^m(u_i^{m+1}) \leqq \varphi_i^m(u_i^m) \leqq f(u^m).$$

Donc, puisque la suite $\{f(u^m)\}$ est bornée inférieurement, on a pour tout $i = 1, \cdots, n$:

$$(1.1) \qquad \lim_{m \to \infty} \varphi_i^m(u_i^{m+1}) - \varphi_i^m(u_i^m) = 0.$$

D'autre part, puisque la fonction f est fortement convexe, l'ensemble $S_0 = \{u \in H | f(u) \leqq f(u_0)\}$ est borné. Comme $u^m \in S_0$ pour tout m, la suite $\{u^m\}$ est bornée. On sait de plus, puisque f fortement convexe, que pour tout i, on a:

$$(1.2) \qquad \varphi_i^m(u_i^m) - \varphi_i^m(u_i^{m+1}) \geqq (\varphi_i'^m(u_i^{m+1}), u_i^m - u_i^{m+1})_i + K\|u_i^{m+1} - u_i^m\|_i^2.$$

Par définition, de u_i^{m+1}, on a pour tout i:

$$(1.3) \qquad (\varphi_i'^m(u_i^{m+1}), u_i^m - u_i^{m+1})_i = 0.$$

Les relations (1.1)–(1.3) entraînent que:

$$(1.4) \qquad \lim_{m \to \infty} \|u_i^{m+1} - u_i^m\|_i = 0 \quad \text{pour tout } i.$$

Comme f' vérifie une condition de Lipschitz sur S_0 on tire de (1.3) et (1.4) que:

$$(1.5) \qquad \lim_{m \to \infty} \rho_i^m \|\varphi_i'^m(u_i^m)\|^2 = 0.$$

Or:

$$\|\varphi_i'^m(u_i^m) - \varphi_i'^m(u_i^{m+1})\| \leqq M\rho_i^m.$$

De plus, par définition de u_i^{m+1}:

$$(\varphi_i'^m(u_i^m), \varphi_i'^m(u_i^{m+1}))_i = 0;$$

il vient alors:

$$\|\varphi_i'^m(u_i^m)\| \leqq M\rho_i^m.$$

D'où on tire de (1.5) que pour tout i:

$$(1.6) \qquad \lim_{m \to \infty} \varphi_i'^m(u_i^m) = 0.$$

Et grâce à (1.4):

$$(1.7) \qquad \lim_{m \to \infty} \|f'(u^m)\| = 0$$

on a, grâce à la convexité de f:

$$f(u^*) \geqq f(u^m) + (f'(u^m), u^m - u^*).$$

D'où on tire de (1.7) et du fait que la suite $\{u^m\}$ est bornée que:

$$(1.8) \qquad \lim_{m \to \infty} f(u^m) = f(u^*).$$

Comme f est fortement convexe, on a:

$$f(u^*) \leqq f\left(\frac{u^* + u^m}{2}\right) \leqq \frac{1}{2}(f(u^*) + f(u^m)) - \frac{K}{4}\|u^m - u^*\|^2,$$

donc:

$$\|u^m - u^*\|^2 \leqq \frac{2}{K}(f(u^m) - f(u^*)).$$

D'où il vient de (1.8) que:

$$\lim_{m \to \infty} \|u^m - u^*\| = 0.$$

Remarque. D'autres choix de directions et de longueur de déplacement pour le passage de u_i^m à u_i^{m+1} peuvent être choisis. En particulier de formalisme général des méthodes de descente adopté par Cea en [2] peut être associé à la méthode de relaxation (cf. [11]). Signalons par exemple le cas où $H_i = \mathbb{R}$ et où on détermine la longueur de déplacement par un procédé de tabulation par pas (voir Cea [2, pp. 83–90]). On montre alors que la méthode de relaxation converge encore si on remplace la minimisation sur chaque composante par un processus très grossier d'approximation de ce minimum.

2. 2ème méthode: Méthode de pénalisation extérieure associée à la relaxation. Les ensembles C_i sont définis par:

$$C_i = \{u_i \in D_i | g_i(u_i) \leqq 0\}, \qquad i = 1, 2, \cdots, n,$$

où D_i est une partie convexe fermée bornée de H_i, g_i une fonctionnellé définie sur H_i convexe et continue, l'ensemble C_i est "défini correctement" c'est-à-dire que pour toute suite $\{u_i^m\}$ dans D_i telle que

$$\lim_{m\to\infty} \sup g_i(u_i^m) \leqq 0,$$

alors on a, si $d(u_i^m, C_i)$ désigne la distance de u_i^m à C_i:

$$\lim_{m\to\infty} d(u_i^m, C_i) = 0.$$

C'est le cas, si compte-tenu des hypothèses précédentes, on suppose de plus que D_i est compact ou bien qu'il existe $z_i \in D_i$ tel que:

$$g_i(z_i) < 0.$$[1]

On suppose de plus que $\alpha = \sup(f(x) | x \in D)$ est fini.

On se donne n suites de réels positifs K_i^m, $i = 1, \cdots, n$, tels que $\lim_{m\to\infty} K_i^m = +\infty$ et des applications l_i de $\mathbb{R}$ dans $\mathbb{R}^+$ décroissantes convexes dérivables et telles que:

$$l_i(\tau) = l_i'(\tau) = 0 \quad \text{si} \quad \tau \leqq 0,$$
$$l_i(\tau) > 0 \quad \text{si} \quad \tau > 0$$

(exemple: $l_i(\tau) = (\tau^+)^2$).

Posons $\psi_i^m(v_i) = \varphi_i^m(v_i) + K_i^m l_i(g_i(v_i))$.

Algorithme. On construit la suite $\{u^m\}$ de la façon suivante: à l'itération m, on obtient successivement pour $i = 1, 2, \cdots, n$ l'élément u_i^{m+1} tel que:

$$u_i^{m+1} \in D_i \cdot \psi_i^m(u_i^{m+1}) = \min(\psi_i^m(u_i) | u_i \in D_i).$$

On a alors le théorème suivant.

THÉORÈME: *La suite $\{u^m\}$ converge fortement vers la solution u^* de* (P).

Démonstration.

1. Montrons d'abord que la suite $\{u^m\}$ ainsi construite vérifie les deux conditions suivantes:

(i)

$$\lim_{m\to\infty} d(u_i^m, C_i) = 0, \tag{2.1}$$

[1] Dans le cas où D_i n'est pas compact (pratiquement si H_i n'est pas de dimension finie), on retrouve la condition standard de qualification.

(ii) Pour toute suite $\{v_i^m\}$, $v_i^m \in C_i$ on a:

$$\lim_{m\to\infty} \inf (\varphi_i'^m(u_i^{m+1}), v_i^m - u_i^{m+1})_i \geqq 0. \tag{2.2}$$

(a) Par construction, on a:

$$\psi_i^m(u_i^{m+1}) \leqq \varphi_i^m(v_i) \quad \text{pour tout} \quad v_i \in C_i$$

car:

$$l_i(g_i(v_i)) = 0 \quad \text{si} \quad v_i \in C_i.$$

Donc:

$$0 \leqq K_i^m l_i(g_i(u_i^{m+1})) \leqq \varphi_i^m(v_i) - \varphi_i^m(u_i^{m+1}) \quad \text{pour tout} \quad v_i \in C_i,$$

et si on pose: $\beta = \inf(f(x) | x \in D) > -\infty$, $D = \prod_{i=1}^n D_i$, il vient:

$$0 \leqq K_i^m l_i(g_i(u_i^{m+1})) \leqq \alpha - \beta.$$

Les propriétés de K_i^m et de l_i entraînent que:

$$\lim_{m\to\infty} \sup g_i(u_i^{m+1}) \leqq 0$$

et comme C_i est défini correctement:

$$\lim_{m\to\infty} d(u_i^m, C_i) = 0$$

et la relation (2.1) est démontrée.

(b) D'après la proposition 2 (Lescarret [6]) l'application $v_i \to l_i(g(v_i))$ est convexe et sous dérivable et si $\partial g_i(u_i)$ désigne le sous différentiel de g_i au point u_i, on a, d'après cette proposition, la formule:

$$\partial(l_i(g_i(v_i)) = l_i'(g_i(v_i))\partial(g_i(v_i)).$$

Par définition de u_i^{m+1}, on a alors:

$$(\varphi_i'^m(u_i^{m+1}), v_i^m - u_i^{m+1})_i + K_i^m l_i'(g_i(u_i^{m+1}))(d_i^m, v_i^m - u_i^{m+1})_i \geqq 0 \tag{2.3}$$
$$\text{pour tout} \quad v_i^m \in D_i,$$

où $d_i^m \in \partial g_i(u_i^{m+1})$ (sous différentiel de g_i au point u_i^{m+1} non vide puisque g_i est continue).

Or, pour tout $v_i^m \in C_i$, pour tout $\bar{d}_i^m \in \partial g_i(v_i^m)$ et pour tout $d_i^m \in \partial g_i(u^{m+1})$, on a, grâce à la convexité de la fonction $l_i(g_i(\cdot))$:

$$0 = l_i'(g_i(v_i^m))(\bar{d}_i^m, v_i^m - u_i^{m+1})_i \geqq l_i'(g_i(u_i^{m+1}))(d_i^m, v_i^m - u_i^{m+1})_i.$$

Ce qui, compte-tenu de (2.3) montre (2.2).

2. Montrons alors le lemme suivant qui a son intérêt propre et qui permet de conclure.

LEMME. *Si C est borné, toute suite $\{u^m\}$ vérifiant les relations* (2.1) *et* (2.2) *est minimisante.*

Démonstration. Remarquons d'abord que grâce à (2.2) et à la convexité de φ_i^m on a si l'on note $P_{C_i}(u_i)$ la projection sur C_i de u_i:

$$\lim_{m\to\infty} \sup \varphi_i^m(u_i^{m+1}) - \varphi_i^m(P_{C_i}(u_i^m)) \leqq 0 \quad \text{pour tout } i. \tag{2.1'}$$

De la relation (2.1) il vient que la suite $\{u^m\}$ est bornée. On déduit de (2.1), de (2.1′) et de l'uniforme continuité de φ_i^m que:

"Il existe une suite $\{\eta_m\}$ de réels positifs tels que $\lim_{m\to\infty}\eta_m = 0$ et que:

$$\varphi_i^m(u_i^{m+1}) \leqq \varphi_i^m(u_i^m) + \eta_m \quad \text{pour tout } m, \qquad i = 1, \cdots, n."$$

Posons:

$$\psi_i^m = \varphi_i^m(u_i^{m+1}) - \varphi_i^m(u_i^m), \qquad A_m = \{i|\psi_i^m > 0\}, \quad B_m = \{i|\psi_i^m \leqq 0\}. \tag{2.2′}$$

Par construction et d'après (2.2′) on a:

$$f(u^{m+1}) - f(u^m) = \sum_{i\in A_m} |\psi_i^m| - \sum_{i\in B_m} |\psi_i^m|, \tag{2.3′}$$

$$|\psi_i^m| \leqq \eta_m \quad \text{pour tout} \quad i \in A_m, \tag{2.4′}$$

$$|\psi_i^m| \leqq (n-1)\eta_m + f(u^m) - f(u^{m+1}) \quad \text{pour tout} \quad i \in B_m. \tag{2.5′}$$

Soit $S = \{m|f(u^m) \leqq f(u^{m+1})\}$. Nous distinguerons successivement le cas où le cardinal de S est infini et celui où il est fini.

Le cardinal de S est infini.

(a) Pour tout $m \in S$, on a d'après (2.5′) et (2.4′):

$$|\psi_i^m| \leqq (n-1)\eta_m. \tag{2.6′}$$

Dans ce cas, l'ensemble A_m ne peut être vide alors que B_m peut l'être. De (2.6′) on tire que:

$$\lim_{\substack{m\to\infty\\ m\in S}} \psi_i^m = 0 \quad \text{pour tout } i. \tag{2.7′}$$

Comme f est fortement convexe, on a:

$$\varphi_i^m(u_i^m) \geqq \varphi_i^m(u_i^{m+1}) + (\varphi_i'^m(u_i^{m+1}), u_i^m - u_i^{m+1})_i + K(\|u_i^{m+1} - u_i^m\|)^2, \quad K > 0,$$

donc, si on pose: $C_i^m = (\varphi_i'^m(u_i^{m+1}), u_i^m - P_{C_i}(u_i^m))_i$, on a, compte-tenu de (2.1) et de (2.2):

$$-\psi_i^m \geqq \varepsilon_m + C_i^m + K(\|u_i^{m+1} - u_i^m\|),$$

la suite $\{\varepsilon_m\}$ vérifiant $\liminf \varepsilon_m \geqq 0$. D'où il vient, grâce à (2.1) et (2.7′) que:

$$\lim_{\substack{m\to\infty\\ m\in S}} \|u_i^{m+1} - u_i^m\| = 0.$$

On en déduit, grâce à (2.2) et au fait la fonctionnelle $\varphi_i'^m$ est uniformément continue que:

$$\liminf_{\substack{m\to\infty\\ m\in S}} (f'(u^m), u - u^m) \geqq 0 \quad \text{pour tout} \quad u \in C.$$

Par convexité de f, cela entraîne que:

$$f(u^*) \geqq \limsup_{\substack{m\to\infty\\ m\in S}} f(u^m). \tag{2.8′}$$

Soit $\tilde{u}$ une valeur d'adhérence faible de la suite $\{u^m\}$, $m \in S$, il en existe au moins une. D'après (2.1) $\tilde{u} \in C$ et comme f est faiblement semi-continue inférieurement, on a:

$$f(u^*) \leqq f(\tilde{u}) \leqq \liminf_{\substack{m \to \infty \\ m \in S}} f(u^m).$$

D'où on tire, grâce à (2.8′) que:

$$\lim_{\substack{m \to \infty \\ m \in S}} f(u^m) = f(u^*). \tag{2.9′}$$

(b) Si le complémentaire de S est vide ou de cardinal fini le théorème est démontré. Dans le cas contraire, à tout $m \notin S$, on fait correspondre les nombres $k(m)$ et $j(m)$ comme suit:

$$k(m) \in S, \quad k(m) = \min(l | l > m, l \in S),$$

$$j(m) \in S, \quad j(m) = \max(l | l < m, l \in S).$$

Par définition, on aura:

$$\begin{cases} f(u^m) \geqq f(u^{k(m)}), \\ f(u^m) \leqq f(u^{j(m)}) + n \cdot \eta_m. \end{cases}$$

Comme, grâce à (2.9′) on a:

$$\lim_{m \to \infty} f(u^{k(m)}) = \lim_{m \to \infty} f(u^{j(m)}) = f^*$$

on en déduit que:

$$\lim_{m \to \infty} f(u^m) = f^*.$$

Le cardinal de S est fini. Dans ce cas, il existe m_0 tel que pour tout $m \geqq m_0$, on ait:

$$f(u^{m+1}) < f(u^m).$$

Cela entraîne, puisque la suite $f(u^m)$ est bornée inférieurement (grâce à (2.1)) que:

$$\lim_{m \to \infty} (f(u^{m+1}) - f(u^m)) = 0.$$

Les relations (2.4′) et (2.5′) entraînent que:

$$\lim_{m \to \infty} \psi_i^m = 0.$$

La fin de la démonstration se fait comme dans la partie (a) de le cas où le cardinal de S est infini.

Remarque. Le cas où les ensembles C_i sont définis par plusieurs contraintes est analogue. Nous ne l'avons pas introduit ici car il n'apporte rien de plus à l'idée de ces procédés. Dans ce cas, il est possible aussi de ne faire qu'une pénalisation partielle des contraintes (cf. Martinet [9]).

Remarque. Comme nous l'avons signalé dans l'introduction, on peut justifier sous des hypothèses convenables des procédés où on associe à la relaxation la plupart des méthodes d'optimisation courantes:

(a) Méthodes de descente:

—de type Frank et Wolfe (cf. Cea [2]),

—de type gradient projeté (cf. Levitin et Polyak [7]),

—de type linéarisation (cf. Cea [2] et Zoutendijk [17]).

(b) Autres méthodes:

—méthodes de coupe ("cut-off methods") (cf. Levitin et Polyak [7]),

—méthodes de régularisation (cf. Tikhonov [15a, b], Martinet [10]).

BIBLIOGRAPHIE

[1] A. Auslender, *Méthodes numériques pour la décomposition et la minimisation de fonctions non différentiables*, Numer. Math., 18 (1971), pp. 213–223.

[2] J. Cea, *Optimisation, Théorie et Algorithmes*, Dunod, Paris, 1971.

[3] Frank Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[4] N. Gastinel, *Analyse Numérique Linéaire*, Hermann, Paris, 1966.

[5] J. E. Kelley, *The cutting-plane methodfor nonlinear programming*, J. SIAM, 8 (1960), pp. 703–712.

[6] C. Lescarret, *Sous différentiabilité de fonctions composées*, Faculté des Sciences de Montpellier, Séminaire d'Analyse unilatérale, 1968, Exposé n° 2.

[7] 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.

[8] B. Martinet, *Convergence de certaines méthodes de relaxation en programmation convexe*, C.R. Acad. Sci. Paris, 265 (1967), pp. 210–212.

[9] ———, Thèse, Université Scientifique et Médicale de Grenoble, Avril, 1972.

[10] ———, *Minimisation d'une fonctionnelle dans un espace produit par une méthode de relaxation*, R.I.R.O., 1971.

[11] B. Martinet et A. Auslender, *Méthodes de relaxation pour la minimisation d'une fonction dans un espace produit*, Séminaire d'Analyse Numérique, Université Scientifique et Médicale de Grenoble, Octobre–Décembre, 1971.

[12] J. C. Miellou, *Sur une notion de monotonie conduisant à une extension de l'application de la méthode variationnelle dans l'étude des systèmes d'équations et d'inéquations aux dérivées partielles—Opérateurs para-monotones*, Thèse, Grenoble, Octobre 1970.

[13] J. M. Ortega and W. C. Rheinboldt, *Iteration Solution of Nonlinear Equations in Several Variables*, Academic Press, New York—London, 1970.

[14] S. Schechter, *Relaxation methods for convex problems*, SIAM J. Numer. Anal., 5 (1968), pp. 601–612.

[15a] A. N. Tikhonov, *On the solution of ill-posed problems and the method of regularization*, Dokl. Akad. Nauk SSSR, 151 (1963), pp. 501–504.

[15b] ———; *On the regularization of ill-posed problems*, Ibid., 153 (1963), pp. 49–52.

[16] G. Zoutendijk, *Methods of Feasible Directions: A Study in Linear and Non-Linear Programming*, Elsevier, Amsterdam, 1960.

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

AN INTERIOR PENALTY METHOD FOR MINIMAX PROBLEMS WITH CONSTRAINTS*

HITOSHI SASAI†

Abstract. In this paper the interior penalty function method is applied to the minimax problems. Fiacco and McCormick's penalty technique is used in order to solve the minimax problems with side constraints.

As a result, the constrained minimax problem is reduced to the one solving a sequence of unconstrained approximation problems and a new computational algorithm is established.

1. Introduction. The application of penalty methods to mathematical programming problems subject to side constraints is currently of much interest.

Carroll [1] proposed an interior penalty method which transforms a constrained minimization problem into a sequence of unconstrained minimization problems.

Further development of this study has been made by A. V. Fiacco and G. P. McCormick [2], [3], and their results are called Sequential Unconstrained Minimization Technique (SUMT).

These are all concerned with minimization problems and do not deal with minimax problems with side constraints.

A. Auslender [4] treated an exterior penalty method for minimax problems in Hilbert space.

In this paper we apply Fiacco and McCormick's interior penalty technique to minimax problems in Banach space. As a result, a minimax problem with constraints is replaced by a sequence of unconstrained minimax problems and a new computational algorithm is established.

2. Preliminaries. Let $\mathscr{X}$ and $\mathscr{U}$ be real reflexive Banach spaces, and $\mathscr{X}'$ and $\mathscr{U}'$ be their dual spaces. We denote the canonical bilinear forms on $\mathscr{X} \times \mathscr{X}'$ and $\mathscr{U} \times \mathscr{U}'$ by $\langle x, x' \rangle$ and $\langle u, u' \rangle$, respectively, i.e., the values of the linear functionals $x' \in \mathscr{X}'$ and $u' \in \mathscr{U}'$ on the elements $x \in \mathscr{X}$ and $u \in \mathscr{U}$, respectively.

We denote by X and U arbitrary nonempty sets of $\mathscr{X}$ and $\mathscr{U}$, respectively. Now let $f(x, u)$ be a function from $X \times U$ to $R^1 \cup \{-\infty\} \cup \{+\infty\}$, where R^1 denotes the real line.

DEFINITION 1. If $\alpha = \sup_U \inf_X f(x, u) = \inf_X \sup_U f(x, u)$, α is called *the saddle value* of $f(x, u)$ with respect to maximizing over U and minimizing over X.

DEFINITION 2. If $(\hat{x}, \hat{u}) \in X \times U$ and

$$f(\hat{x}, u) \leqq f(\hat{x}, \hat{u}) \leqq f(x, \hat{u}) \quad \forall\, x \in X \text{ and } \forall\, u \in U,$$

$(\hat{x}, \hat{u})$ is called *the saddle point.*

DEFINITION 3. If $f(x, u)$ is a convex function of $x \in X$ for each $u \in U$ and a concave function of $u \in U$ for each $x \in X$, we say that $f(x, u)$ is a *convex-concave* function.

Remark. If $(\hat{x}, \hat{u})$ is a saddle point, $f(\hat{x}, \hat{u})$ is the saddle value.

* Received by the editors November 17, 1971, and in revised form July 19, 1973.

† Faculty of Business Administration, Yokohama National University, Yokohama, Japan.

DEFINITION 4. For $f(x, u)$ defined on $\mathscr{X} \times \mathscr{U}$ and taking values in $R^1 \cup \{-\infty) \cup \{+\infty\}$, we write

$$\operatorname{dom} f = \{x, u : |f(x, u)| < \infty\}.$$

dom f is called the effective domain of $f(x, u)$ [5]. We denote the projections of dom f on $\mathscr{X}$ and $\mathscr{U}$ by $\operatorname{dom}_{\mathscr{X}} f$ and $\operatorname{dom}_{\mathscr{U}} f$, respectively.

DEFINITION 5. Let $f(x)$ be a function from $\mathscr{X}$ to $R^1 \cup \{-\infty\} \cup \{+\infty\}$. If $f(x)$ is weakly (strongly) lower semicontinuous, we say that $f(x)$ is weakly (strongly) closed.

3. Statement of problem. $\mathscr{X}$ and $\mathscr{U}$ are real reflexive Banach spaces. Let X_1 and U_1 be convex subsets of $\mathscr{X}$ and $\mathscr{U}$, which are strongly closed and contain interior points.

The following conditions are imposed in our discussions.

(C_1): $f(x, u)$, $X_1 \times U_1 \to R^1$, is a convex-concave function and strongly continuous in each variable.

(C_2): $g_i(x)$, $X_1 \to R^1$ and $h_j(u)$, $U_1 \to R^1$ are concave functions and strongly continuous, where $i = 1, 2, \cdots, m$ and $j = 1, 2 \cdots, n$.

(C_3): X and U are bounded sets, and $X^0 \times U^0 = \phi$, where

$$X = \{x \in X_1 : g_i(x) \geqq 0, i = 1, \cdots, m\},$$

$$X^0 = \{x \in X_1 : g_i(x) > 0, i = 1, \cdots, m\},$$

$$U = \{u \in U_1 : h_j(u) \geqq 0, j = 1, \cdots, n\},$$

$$U^0 = \{u \in U_1 : h_j(u) > 0, j = 1, \cdots, n\}.$$

Remark. If (C_1) and (C_2) are satisfied, it follows that $\overline{X^0} = X$ and $\overline{U^0} = U$, where $\overline{X^0}$ and $\overline{U^0}$ denote the strong closures of X^0 and U^0, respectively.

Remark. The following fact should be noted: if a real-valued convex function on a Banach space is weakly lower semicontinuous, then it is strongly continuous [6].

The minimax problem under consideration is the following one.

Original problem (O). Find a saddle point of $f(x, u)$ with respect to $X \times U$, i.e., a point $(\hat{x}, \hat{u})$ such that:

$$f(\hat{x}, u) \leqq f(\hat{x}, \hat{u}) \leqq f(x, \hat{u}), \tag{1}$$

for any $x \in X$ and $u \in U$.

An attempt to solve Problem (O) by applying a penalty method results in the following sequence of problems.

Approximation problem (A). Find a saddle point of $p(x, u, r_k)$ with respect to $X^0 \times U^0$ for each $r_k > 0$, i.e., a point $(x_k, u_k) \in X^0 \times U^0$ such that

$$p(x_k, u, r_k) \leqq p(x_k, u_k, r_k) \leqq p(x, u_k, r_k) \tag{2}$$

for every $x \in X^0$ and $u \in U^0$, where

$$p(x, u, r_k) = f(x, u) + \sum_1^m \frac{r_k}{g_i(x)} - \sum_1^n \frac{r_k}{h_j(u)},$$

and r_k is a strictly monotonic decreasing sequence and $r_k \to 0$ as $k \to \infty$.

Problem (A) is an unconstrained minimax problem for each r_k. By generating a sequence of solutions of Problem (A), consisting of interior points of the constrained set, which converges to a solution of Problem (O), we can obtain a new computational algorithm for solving a minimax problem subject to side constraints.

In the following we discuss the properties of Problem (A) and the convergency conditions of Problem (A) relative to Problem (O).

4. Existence of solutions of Problem (A). We modify $p(x, u, r_k)$ as follows:

$$p^*(x,u,r_k) = \begin{cases} p(x,u,r_k) & \text{for } x \in X^0 \text{ and } u \in U^0, \\ +\infty & \text{for } x \notin X^0 \text{ and } u \in U^0, \\ -\infty & \text{for } u \notin U^0. \end{cases} \tag{3}$$

This modification is called a convex-concave extension [7]. We should note that $\operatorname{dom} p^* = X^0 \times U^0$, $\operatorname{dom}_{\mathscr{X}} p^* = X^0$ and $\operatorname{dom}_{\mathscr{U}} p^* = U^0$.

We first prove the following lemmas.

LEMMA 1. *Let the conditions* (C_1) *and* (C_2) *be satisfied. Then* $p^*(x,u,r_k)$ *is convex-concave. Moreover* $p^*(x,u,r_k)$ *is weakly closed for every fixed* $u \in \operatorname{dom}_{\mathscr{U}} p^*$ *and* $-p^*(x,u,r_k)$ *is weakly closed for every fixed* $x \in \operatorname{dom}_{\mathscr{X}} p^*$.

Proof. By the definition (3), it is clear that p^* is convex-concave.

Now let us show the second part of the lemma. We may demonstrate that the sets $M = \{x \in \mathscr{X} : p^*(x, u_0, r_k) \leqq \alpha\}$ and $N = \{u \in \mathscr{U} : -p^*(x_0, u, r_k) \leqq \alpha\}$ are weakly closed for any fixed $x_0 \in \operatorname{dom}_{\mathscr{X}} p^*$, $u_0 \in \operatorname{dom}_{\mathscr{U}} p^*$ and $\alpha \in R^1$.

$M \subseteq X^0$ and $N \subseteq U^0$, because $\operatorname{dom}_{\mathscr{X}} p^* = X^0$ and $\operatorname{dom}_{\mathscr{U}} p^* = U^0$. On the other hand, $p^*(x, u_0, r_k)$ and $-p^*(x_0, u, r_k)$ are strongly continuous on X^0 and U^0 by (C_1) and (C_2). Hence, M and N are strongly closed, because X_1 and U_1 are strongly closed.

Since M and N are convex sets by (C_1) and (C_2), we can conclude that M and N are weakly closed.

LEMMA 2. Let the conditions (C_1), (C_2) and (C_3) be satisfied. If $(\hat{x}, \hat{u})$ is a saddle point of $p^*(x,u,r_k)$ with respect to $\mathscr{X} \times \mathscr{U}$, it is also a saddle point of $p(x,u,r_k)$ with respect to $X^0 \times U^0$.

Proof. By the definition of p^* and (C_3),

$$\inf_{\mathscr{X}} p^*(x,u,r_k) = \inf_{X^0} p^*(x,u,r_k) < +\infty \quad \text{for any } u \in \mathscr{U}, \tag{4}$$

$$\sup_{\mathscr{U}} p^*(x,u,r_k) = \sup_{U^0} p^*(x,u,r_k) > -\infty \quad \text{for any } x \in \mathscr{X}. \tag{5}$$

The assumption means that

$$\sup_{\mathscr{U}} p^*(\hat{x},u,r_k) \leqq p^*(\hat{x},\hat{u},r_k) \leqq \inf_{\mathscr{X}} p^*(x,\hat{u},r_k). \tag{6}$$

By (4), (5) and (6), we have

$$-\infty < \sup_{U^0} p^*(\hat{x},u,r_k) \leqq p^*(\hat{x},\hat{u},r_k) \leqq \inf_{X^0} p^*(x,\hat{u},r_k) < +\infty. \tag{7}$$

Now we can prove $(\hat{x}, \hat{u}) \in X^0 \times U^0$. If $\hat{x} \notin X^0$, $\sup_{U^0} p^*(\hat{x}, u, r_k) = +\infty$ and hence $\inf_{X^0} p^*(x, \hat{u}, r_k) \geqq +\infty$. If $\hat{u} \notin U^0$, $\inf_{X^0} p^*(x, \hat{u}, r_k) = -\infty$ and hence $\sup_{U^0} p^*(\hat{x}, u, r_k) \leqq -\infty$. These contradict (7) and hence $(\hat{x}, \hat{u}) \in X^0 \times U^0$.

Since $p^*(x, u, r_k) = p(x, u, r_k)$ on $X^0 \times U^0$, we have

$$\sup_{U^0} p(\hat{x}, u, r_k) \leqq p(\hat{x}, \hat{u}, r_k) \leqq \inf_{X^0} p(x, \hat{u}, r_k).$$

This completes the proof.

We use Moreau's theorem [5] to prove the existence of solutions of Problem (A). Concerning other theorems on the existence of saddle points, see [7], [8].

THEOREM (Moreau). *Suppose that the following conditions are satisfied*:

(i) $f(x, u)$ *is a convex-concave function from* $\mathscr{X} \times \mathscr{U}$ *to* $R^1 \cup \{-\infty\} \cup \{+\infty\}$ *and* $f(x, u)$ *is weakly closed for every fixed* $u \in \operatorname{dom}_{\mathscr{U}} f$ *and* $-f(x, u)$ *is weakly closed for every fixed* $x \in \operatorname{dom}_{\mathscr{X}} f$;

(ii) *there exist* $x_0 \in \operatorname{dom}_{\mathscr{X}} f$ *and a number* $a_0 < \sup_{\mathscr{U}} f(x_0, u)$ *such that* $\{u : f(x_0, u) \geqq a_0\}$ *is weakly compact*;

(iii) *there exist* $u_0 \in \operatorname{dom}_{\mathscr{U}} f$ *and a number* $b_0 > \inf_{\mathscr{X}} f(x, u_0)$ *such that* $\{x : f(x, u_0) \leqq b_0\}$ *is weakly compact. Then* $f(x, u)$ *has a saddle point with respect to* $\mathscr{X} \times \mathscr{U}$.

If we apply Moreau's theorem to $p^*(x, u, r_k)$, we have the following theorem.

THEOREM 1. *Let the conditions* (C_1), (C_2) *and* (C_3) *be satisfied. Then there exists a saddle point of* $p(x, u, r_k)$ *with respect to* $X^0 \times U^0$ *for every* r_k.

Proof. By Lemma 2, we may show that there exists a saddle point of $p^*(x, u, r_k)$ with respect to $\mathscr{X} \times \mathscr{U}$.

Now we apply Moreau's theorem to $p^*(x, u, r_k)$. By Lemma 1, condition (i) is satisfied.

Let us show that conditions (ii) and (iii) are satisfied. By (C_3), there exists $(x_0, u_0) \in X^0 \times U^0$. We put

$$a_0 = f(x_0, u_0) - \sum_1^n \frac{r_k}{h_j(u_0)}, \qquad b_0 = f(x_0, u_0) + \sum_1^m \frac{r_k}{g_i(x_0)}.$$

Then we have

$$a_0 = f(x_0, u_0) - \sum_1^n \frac{r_k}{h_j(u_0)} < p^*(x_0, u_0, r_k) \leqq \sup_{\mathscr{U}} p^*(x_0, u, r_k),$$

$$b_0 = f(x_0, u_0) + \sum_1^m \frac{r_k}{g_i(x_0)} > p^*(x_0, u_0, r_k) \geqq \inf_{\mathscr{X}} p^*(x, u_0, r_k).$$

We define M and N as follows:

$$M = \{u : p^*(x_0, u, r_k) \geqq a_0\}, \qquad N = \{x : p^*(x, u_0, r_k) \leqq b_0\}.$$

Since $M \subseteq U^0$, $N \subseteq X^0$ and X^0, U^0 are bounded by (C_3), M and N are bounded. On the other hand we can conclude that M and N are convex and strongly closed by the same technique as that used in the proof of Lemma 1. Thus it follows that M and N are weakly compact, and hence (ii) and (iii) are satisfied.

Consequently there exists a saddle point of $p(x, u, r_k)$ with respect to $X^0 \times U^0$ for each r_k.

In the following we denote the saddle point of Problem (A) for each r_k by (x_k, u_k).

COROLLARY. *If $f(x, u)$, $g_i(x)$ and $h_j(u)$ are Fréchet differentiable in each variable, respectively, then*

$$\nabla_x p(x, u_k, r_k)|_{x=x_k} = 0, \qquad \nabla_u p(x_k, u, r_k)|_{u=u_k} = 0,$$

where ∇_x and ∇_u are the gradients of $p(x, u, r_k)$ in x and u, respectively.

Proof. This corollary is clear because $(x_k, u_k) \in X^0 \times U^0$ and X^0, U^0 are interior sets of X, U, respectively.

5. Condition of convergence. Since X and U are weakly compact, there exists a saddle point (and hence the saddle value) for Problem (O) from Moreau's theorem by extending f as in (3), but with X and U instead of X^0 and U^0.

Now we put as follows:

$$\inf_X \sup_U f(x, u) = \sup_U \inf_X f(x, u) = \delta^*. \tag{8}$$

THEOREM 2. *Let the conditions* (C_1), (C_2) *and* (C_3) *be satisfied, and* (x_k, u_k) *be a saddle point of Problem* (A) *for each* r_k. *Then,*

$$p(x_k, u_k, r_k) \to \delta^* \quad \textit{and} \quad f(x_k, u_k) \to \delta^*.$$

Proof. By (8), $\overline{X^0} = X$ and $\overline{U^0} = U$, there exist $x_0 \in X^0$ and $u_0 \in U^0$ such that for any $\varepsilon > 0$,

$$\delta^* + \varepsilon \geqq \sup_U f(x_0, u), \qquad \delta^* - \varepsilon \leqq \inf_X f(x, u_0). \tag{9}$$

Since (x_k, u_k) is a saddle point,

$$\begin{aligned} f(x_k, u) + \sum_1^m \frac{r_k}{g_i(x_k)} - \sum_1^n \frac{r_k}{h_j(u)} \leqq p(x_k, u_k, r_k) &= f(x_k, u_k) + \sum_1^m \frac{r_k}{g_i(x_k)} - \sum_1^n \frac{r_k}{h_j(u_k)} \\ &\leqq f(x, u_k) + \sum_1^m \frac{r_k}{g_i(x)} - \sum_1^n \frac{r_k}{h_j(u_k)}, \end{aligned} \tag{10}$$

for any $x \in X^0$ and $u \in U^0$. Hence,

$$\begin{aligned} &\inf_X f(x, u_0) - \sum_1^n \frac{r_k}{h_j(u_0)} \\ &\leqq f(x_k, u_0) - \sum_1^n \frac{r_k}{h_j(u_0)} \\ &\leqq f(x_k, u_0) + \sum_1^m \frac{r_k}{g_i(x_k)} - \sum_1^n \frac{r_k}{h_j(u_0)} \\ &\leqq p(x_k, u_k, r_k) \\ &\leqq f(x_0, u_k) + \sum_1^m \frac{r_k}{g_i(x_0)} - \sum_1^n \frac{r_k}{h_j(u_k)} \\ &\leqq f(x_0, u_k) + \sum_1^m \frac{r_k}{g_i(x_0)} \\ &\leqq \sup_U f(x_0, u) + \sum_1^m \frac{r_k}{g_i(x_0)}. \end{aligned} \tag{11}$$

Thus by (9) and $(x_0, u_0) \in X^0 \times U^0$, for sufficiently large k,

$$\delta^* - \varepsilon - \varepsilon \leqq p(x_k, u_k, r_k) \leqq \delta^* + \varepsilon + \varepsilon .$$

Consequently we have

$$p(x_k, u_k, r_k) \to \delta^* .$$

Again by (10) the following inequalities are clear:

$$f(x_k, u_0) - \sum_1^n \frac{r_k}{h_j(u_0)} \leqq f(x_k, u_k) - \sum_1^n \frac{r_k}{h_j(u_k)} \leqq f(x_k, u_k), \tag{12}$$

$$f(x_0, u_k) + \sum_1^m \frac{r_k}{g_i(x_0)} \geqq f(x_k, u_k) + \sum_1^m \frac{r_k}{g_i(x_k)} \geqq f(x_k, u_k). \tag{13}$$

The left-hand side of (12) is greater than $\delta^* - 2\varepsilon$ and the left-hand side of (13) is smaller than $\delta^* + 2\varepsilon$ by the inequality (11). Therefore we have

$$f(x_k, u_k) \to \delta^* .$$

Remark. It should be easy to see the following results using (12) and (13):

$$\sum_1^m \frac{r_k}{g_i(x_k)} \to 0 \quad \text{and} \quad \sum_1^n \frac{r_k}{h_j(u_k)} \to 0. \tag{14}$$

THEOREM 3. *Let the conditions* (C_1), (C_2) *and* (C_3) *be satisfied and* (x_k, u_k) *be a saddle point of Problem* (A) *for each* r_k. *Then the sequence* $\{(x_k, u_k)\}$ *contains at least one cluster point* (*in the sequence of weak topology*) *and every cluster point* $(\hat{x}, \hat{u})$ *of* $\{(x_k, u_k)\}$ *is a saddle point of Problem* (O).

Proof. X^0 and U^0 are bounded sets of reflexive Banach space by (C_3) and since $(x_k, u_k) \in X_0$, there exist subsequences $\{x_{k_j}\}$ and $\{u_{k_j}\}$ converging weakly to some $\hat{x} \in X$ and $\hat{u} \in U$. We write k for k_j.

We first prove that

$$f(x, u') \leqq \delta^* \leqq f(x', u) \tag{15}$$

for any $u' \in U^0$ and $x' \in X^0$.

We have the following inequalities because $-f(x, u)$ is weakly lower semicontinuous in u, $\sum r_k/g_i(x') \to 0$ and $\sum r_k/h_j(u_k) \to 0$ (by (14)):

$$\begin{aligned} \delta^* &= \lim_k p(x_k, u_k, r_k) \leqq \overline{\lim}\, p(x', u_k, r_k) \\ &\leqq \overline{\lim}\, f(x', u_k) + \overline{\lim} \sum \frac{r_k}{g_i(x')} + \overline{\lim} \left\{ -\sum \frac{r_k}{h_j(u_k)} \right\} \\ &\leqq f(x', \hat{u}). \end{aligned}$$

On the other hand, by a similar technique,

$$\delta^* = \lim p(x_k, u_k, r_k) \geqq f(\hat{x}, u'),$$

because $f(x, u)$ is weakly lower semicontinuous in x.

We prove secondly that

$$f(\hat{x}, \hat{u}) = \delta^* .$$

Since $\overline{X^0} = X$ and $\overline{U^0} = U$, there exist sequences $\{x_n\} \subseteq X^0$ and $\{u_n\} \subseteq U^0$ which converge strongly to $\hat{x}$ and $\hat{u}$, respectively. Hence by (15) we have

$$f(\hat{x}, u_n) \leqq \delta^* \leqq f(x_n, \hat{u}).$$

Since $f(x, u)$ is strongly continuous in each variable, it follows that

$$\delta^* = f(\hat{x}, \hat{u}). \tag{16}$$

Last we prove $(\hat{x}, \hat{u})$ is a saddle point of Problem (O). By (15) and (16) we have

$$f(\hat{x}, u') \leqq f(\hat{x}, \hat{u}) \leqq f(x', \hat{u})$$

for any $x' \in X^0$ and $u' \in U^0$. But again by strong continuity of $f(x, u)$, $\overline{X^0} = X$ and $\overline{U^0} = U$, we can conclude that

$$f(\hat{x}, u) \leqq f(\hat{x}, \hat{u}) \leqq f(x, \hat{u})$$

for any $x \in X$ and $u \in U$.

COROLLARY. *Let the conditions stated in Theorem* 3 *be satisfied. If a saddle point of Problem* (O) *denoted by* $(\hat{x}, \hat{u})$ *is unique,* $\{x_k\}$ *and* $\{u_k\}$ *converge weakly to* $\hat{x}$ *and* $\hat{u}$, *respectively.*

Proof. This corollary is clear because any subsequence of $\{(x_k, u_k)\}$ has also its subsequence converging weakly to $(\hat{x}, \hat{u})$.

Acknowledgment. The author wishes to thank Prof. H. Ishigaki, the School of Education, Waseda University, for helpful suggestions and discussions.

REFERENCES

[1] C. W. CARROLL, *The created response surface technique for optimizing nonlinear restrained systems*, Operations Res., 9 (1961), pp. 169–184.

[2] A. V. FIACCO AND G. P. MCCORMICK, *The sequential unconstrained minimization technique for nonlinear programming, a primal-dual method*, Management Sci., 10 (1964), pp. 360–366.

[3] A. V. FIACCO AND A. P. JONES, *Generalized penalty method in topological spaces*, SIAM J. Appl. Math., 5 (1969), pp. 996–1000.

[4] A. AUSLENDER, *Problèmes de minimax via l'analyse convexe et les inégalités variationelles: Théorie et algorithmes*, No. 77, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1972.

[5] A. D. IOFFE AND V. M. TIKHOMIROV, *Duality of convex functions and extremum problems*, Uspehi Mat. Nauk, XXIII (1968), pp. 51–116.

[6] R. T. ROCKAFELLAR, *Level sets and continuity of conjugate convex functions*, Trans. Amer. Math. Soc., 123 (1966), pp. 46–63.

[7] ———, *Convex Analysis*, Princeton Univ. Press, Princeton, N.J., 1969.

[8] K. FAN, *Minimax theorems*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 42–47.

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

AN EXTENSION OF AN OPTIMAL CONTROL SUFFICIENCY THEORY*

GERALD M. ARMSTRONG†

Abstract. An extension is given for a general sufficiency theory in time-optimal control which was developed by L. C. Young. The extension enables sufficient conditions to be applied to a wider class of trajectories than in the original theory, including a class of trajectories with initial Lagrange multiplier unity.

L. C. Young has developed a very general sufficiency theory for time-optimal control problems [1], [2]. The purpose of this paper is to extend this theory to include a wider class of trajectories than in his original work.

1. We consider an autonomous, time-optimal control problem with a given set as target. The letter t will denote time, x will be a point in n-dimensional space, and u will be the control value. Here u may be a point in some elementary figure, a chattering control, or even some sort of label distinguishing one function of (t, x) from another.

The function $x(t)$ will be assumed absolutely continuous, and $u(t)$ will be arbitrary, with points or labels as values. We suppose that $g(x, u)$ is a smooth function of x, with values in n-space. We call the pair $x(t)$, $u(t)$, which are defined for $t \leqq 0$, an admissible trajectory if they satisfy almost everywhere in t the differential equation

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

where we suppose that the terminal value $x(0)$ lies on the target. If t_1 is the initial time on a trajectory, $t_1 < 0$, our problem is to determine for a given value of $x(t_1)$ the smallest value of $-t_1$ in the class of admissible trajectories.

2. We restrict the class of admissible trajectories by requiring them to satisfy a strengthened form of Pontryagin's maximum principle; namely, that along a trajectory $x(t)$ with control $u(t)$, there exists a conjugate vector function $y(t)$, absolutely continuous in t, which never vanishes, and satisfies

A. $\dot{y}(t) = -y(t)g_x(x(t), u(t))$ for almost all t,
B. $y(t)g(x(t), u) \leqq 1$ for all u, with equality when $u = u(t)$,
C. $y(0)$ is normal to the target at the point $x(0)$.

This differs from the maximum principle in that for inequality B, the right-hand side is unity instead of some nonnegative quantity which is constant in t. This is equivalent to requiring that the initial Lagrange multiplier be unity, and is an unpleasant restriction. Our main result is to show how the theory can be extended to include a class of trajectories not satisfying this restriction.

Trajectories satisfying A, B and C are called lines of flight. The corresponding trios $x(t)$, $y(t)$, $u(t)$ will be called canonical lines of flight. Arcs of these will be called arcs, or canonical arcs, of flight. We also assume the synchronization condition for lines of flight: if two lines, or arcs of flight meet at a point x, they do so at the same time t.

* Received by the editors April 12, 1973.

† Department of Mathematics, Brigham Young University, Provo, Utah 84601. This paper contains results from the author's thesis, University of Wisconsin, 1971.

3. We next suppose that the arcs of flights are grouped into families depending on some Euclidean parameter σ. Similarly, the conjugate variable $y(t)$ depends on an additional parameter ρ. The hypotheses concerning the sets of σ and ρ are given elsewhere (see [1, p. 267]).

The families of arcs using these parameters are defined by functions

$$x(t,\sigma), \quad y(t,\sigma,\rho), \quad u(t,\sigma), \quad t^{-}(\sigma) \leqq t \leqq t^{+}(\sigma),$$

where $t \neq -\infty$. The functions t^{-}, t^{+} satisfy $-\infty \leqq t^{-}(\sigma) < t^{+}(\sigma) \leqq 0$ as well as other hypotheses (see [1, p. 267]).

We denote by S^{-}, S, S^{+} and by S^{*-}, S^{*}, S^{*+} the sets of (t,σ) and (t,σ,ρ) for which t is subject to the corresponding conditions $-\infty < t^{-}(\sigma) = t$, or $t^{-}(\sigma) < t < t^{+}(\sigma)$, or $t = t^{+}(\sigma)$. The images under the function $x(t,\sigma)$ of S^{-}, S, S^{+} will be written E^{-}, E, E^{+} and the images under the pair $x(t,\sigma)$, $y(t,\sigma,\rho)$ of S^{*-}, S^{*}, S^{*+} will be written E^{*-}, E^{*}, E^{*+}.

Subject to certain assumptions, [1, p. 268], the family Σ of functions of t given by $x(t,\sigma)$, $u(t,\sigma)$, $t^{-}(\sigma) \leqq t \leqq t^{+}(\sigma)$, will be termed a spray of flights. With the inclusion of $y(t,\sigma,\rho)$, we obtain a canonical spray of flights. The sets E^{-}, E, E^{+}, or E^{*-}, E^{*}, E^{*+} will be termed its source, its flight corridor, its destination. The arcs themselves will be taken to be open arcs in E or E^{*}, but possessing endpoints situated in the source or destination. In the following we omit further references to canonical sprays.

4. A finite or countable sequence $\Sigma_1, \Sigma_2, \cdots$ of sprays of flight will be called a chain of flights if, for $k = 1, 2, \cdots$, they fit together in inverse order so that the source of Σ_k contains the destination of Σ_{k+1}. When the destination of Σ_1 is the target, we speak of a chain of flights to the target. The sources and flight corridors of the individual sprays are called the constituent sets of the chain. We call any collection of chains of flight to the target a concourse of flights if the collection satisfies conditions given below.

The concept of unimpaired union allows us to pass from sprays of flight, the local theory, to a concourse of flights, the global results. A set R, with a class of subsets P, is the unimpaired union of the P if it is their union, and if further, there exists a repairable decomposition of R into at most countably many sets R_k, such that each P is the union of those R_k which contain it. A set R which is covered by the lines of flight is termed a concourse of flights, if R is the unimpaired union of the constituent sets of the chains.

The decomposition of R into the R_k is called repairable if every rectifiable curve in R can be obtained from the rectifiable curves in the R_k by finite fusion and cutting, and countable embellishment and trimming. Fusion means adding two curves having a common endpoint; cutting is the reverse operation. Embellishment means adding to a curve a closed curve which intersects it, and trimming is the reverse operation (see [1, p. 277]).

5. Suppose that R is a concourse of flights. The function $T(x)$, for x in R, called the flight-time from x to the target, is defined to be the length of the time-interval for a line of flight issuing from the point x. A set or curve will be termed "of bounded flight-time" if $T(x)$ is bounded on it.

The basic existence theorem of Young under these hypotheses is as follows (see [1, p 281]).

THEOREM. *Let $x \in R$. Then the flight-time $T(x)$ is the least time for transferring the point x to the target along an admissible trajectory in R.*

The fact that the admissible trajectories considered must all be in R restricts applications of the theorem in some problems. For example, a problem seen in many books is the "controlled slowing of an oscillator" [1, p. 233]. (Optimal trajectories are determined for this problem in [3, p. 568], [4, pp. 52, 80], [5, p. 11], and [6, pp. 27, 53].) In this problem, each spray of flights in the plane is bounded by a source, a destination, and two trajectories which do not satisfy property B of the maximum principle, and therefore are not in R. We call these trajectories not in R weak lines of flight. (Illustrations in [1, p. 240] and [3, p. 579] show the sprays of flight as well as the weak lines of flight bounding the sprays.)

To complete the proof of optimality for the oscillator problem, it is necessary to deal with the weak lines of flight. A proof is given by Young [1, p. 239] by showing that a trajectory can only cross a weak line of flight outwardly, away from the origin. Thus the weak lines of flight constitute one way barriers, and so a trajectory crossing a weak line of flight cannot be optimal.

The extension of this theory which we give below allows us to consider trajectories in a set R^* which includes both R and a certain class of weak lines of flight. Our extension, in particular, provides an easy sufficiency verification for the controlled slowing of an oscillator problem.

6. We first give two lemmas which enable us to define a set R^* of lines of flight. The proofs of these lemmas may be given by sequential continuity arguments, and are omitted here.

LEMMA 1. *Let X, Y, Z be Euclidean spaces, and suppose X is bounded. Let $f: X \to Y$, $g: Y \to Z$, and $h: X \to Z$ be the composition $g(f)$ of f and g. If h and f are continuous, so is g.*

LEMMA 2. *Let X and Y be Euclidean spaces. Let $f: X \to Y$ be a function which is continuous on each rectifiable curve in X. Then f is continuous on X.*

We now prove a theorem based on these lemmas.

THEOREM 1. *Assume R, the set covered by lines of flight, is the unimpaired union of the constituent sets. Let $T(x)$ be the flight-time from $x \in R$. Then T is continuous on R.*

Proof. We first show T is continuous on each rectifiable curve $x(t)$ of bounded flight-time in R. Let $X = \{x : x = x(t)\}$. Let I denote a time interval. Consider the commutative diagram in Fig. 1 where $s(t)$ represents the arc length along $x(t)$.

$$\begin{tikzcd}
I \arrow[r, "x(t)"] \arrow[dr, "s(t)"'] & X \arrow[r, "T(x)"] & I \\
& I \arrow[u, "x(s)"'] \arrow[ur, "T(s)"'] &
\end{tikzcd}$$

FIG. 1

$T(s)$ is continuous [1, p. 280], $x(s)$ is $x(t)$ represented in terms of its arc length, and so is continuous. Therefore $T(x)$ is continuous by Lemma 1.

If we know that the curve $x(t)$ is of bounded flight-time, the theorem follows by Lemma 2. By [1, p. 272], $x(t)$ is of bounded flight-time if it lies in E or E^-, a constituent set of R. Since R is the unimpaired union of the constituent sets, $x(t)$ is of bounded flight-time in R.

THEOREM 2. *Assume $x(t)$ is a trajectory with corresponding control $u(t)$ which satisfies the maximum principle and has no self-intersections. Then T is continuous along $x(t)$.*

Proof. If T were not continuous at $x_0 = x(t_0)$, then $T(x_0^-) < T(x_0^+)$, where the right- and left-hand limits are taken along the trajectory $x(t)$. Let $t_1 = T(x_0^-)$, $t_2 = T(x_0^+)$. Then either $x([t_1, t_2]) = x_0$, or $x(t)$ has a self-intersection at x_0. The equation $x([t_1, t_2]) = x_0$ implies $x(t) = 0$ for each t in (t_1, t_2), and so $g(x(t), u(t)) = 0$ and $g_x(x(t), u(t)) = 0$ for t in (t_1, t_2). By equation A of the strengthened maximum principle, $y(t)$ must be constant over (t_1, t_2), so $y(t)g(x(t), u(t)) = 0$ on this interval. But $y(t)g(x, u)$ is constant on trajectories, and so 0 over the entire t-interval. Since $y(t) \neq 0$, $g(x, u) = 0$, so $x(t) = 0$ and $x(t)$ is a constant for all t. But then $x(t)$ is a degenerate trajectory. This completes the proof of Theorem 2.

We now specify the additional trajectories which we include with R to form the larger set R^*. First, the additional trajectories must be synchronized with those in R: if two trajectories from R^* meet at a point x, they must do so in the same time t. Thus $T(x)$, which is continuous in R by Theorem 1, is well-defined in R^*. Note that the synchronization condition implies that no trajectory in R^* can have self-intersections.

We also require our trajectories in R^* to satisfy the maximum principle. We call the additional trajectories in R^* weak lines of flight. We shall see that because the function T is continuous on R, and continuous on each weak line of flight by Theorem 2, that it is continuous in all of R^*.

Our method of proof requires three additional hypotheses. We assume that each point on a weak line of flight has a neighborhood intersecting at most finitely many other weak lines of flight. We assume each point on a weak line of flight is contained in the closure of R. Finally, we assume that each point on a weak line of flight is contained in a neighborhood on which T is bounded. This completes the description of the set R^*.

THEOREM 3. *Let R^* be as above. Then T is continuous on R^*.*

Proof. Let x_0 be a point on $x_0(t)$, a weak line of flight in R^*. Suppose x_0 is not in R. Let $\{x_i\}$ be a sequence in R^* such that $x_i \to x_0$. We must show that $T(x_i) \to T(x_0)$.

We assume without loss of generality that $\{x_i\}$ is contained in R. In fact, if infinitely many of the x_i lie on weak lines of flight, our finiteness condition implies infinitely many lie on a particular line of flight, which must intersect $x_0(t)$ at x_0. But T is continuous along any line of flight, so $T(x_i) \to T(x_0)$.

For $i = 0, 1, 2, \cdots$, let $-t_i$ be selected so that there is a line of flight $x_i(t)$ with $x_i(t_i) = x_i$, $i = 1, 2, \cdots$, and $x_0(t_0) = x_0$. By the synchronization condition, $T(x_i) = -t_i$ for all i. Since $\{t_i\}$ contains a bounded subsequence, there exists a subsequence of $\{t_i\}$ which we denote $\{t_k\}$, and a point t' such that $t_k \to t'$. But then $x_i(t_k) \to x_i(t')$ for each i. Since also $x_i(t') \to x_0(t')$ as $i \to \infty$, we obtain $x_k(t_k) \to x_0(t')$ as $k \to \infty$. But $\{x_k(t_k)\}$ is a subsequence of $\{x_i\}$, so $x_0(t') = x_0$. But then, by synchronization, $t' = t_0$, and the proof is complete.

Our basic sufficiency theorem will now be given.

THEOREM 4. *Let R^* be as in Theorem 3. Let x_0 be a point on the weak line of flight $x_0(t)$, where x_0 is not in R. Then $T(x_0)$ is the least time to transfer x_0 to the target along a trajectory in R^*.*

Proof. Let $x_1(t)$ be any trajectory in R^* from x_0 to the target. If $x_1(t)$ is a weak line of flight, the proof follows by synchronization. Otherwise, let $S = \{t : x_1(t) \in R^* - R\}$. Then S contains at most countably many nondegenerate closed intervals $[t', t'']$. Let $x' = x_1(t'), x'' = x_1(t'')$. Then $T(x') - T(x'') \leqq -(t' - t'')$. In fact, let $\{t_i'\}$ and $\{t_i''\}$ be sequences such that $t_i' \to t'$ and $t_i'' \to t''$ as $i \to \infty$, and $x_1(t_i')$ and $x_1(t_i'')$ are points of R for all i. Since these points lie in R,

$$T(x_1(t_i')) - T(x_1(t_i'')) \leqq -(t_i' - t_i'').$$

Since $T(x)$ is continuous in R^*, $T(x') - T(x'') \leqq -(t' - t'')$.

Now let a be the greatest lower bound of the set of t in S for which $t > t''$ and let there exist a, b with $[a, b]$ contained in S. If there is no such point, let $a = 0$. Then in the same way as above, we see that

$$T(x'') - T(x_1(a)) \leqq -(t'' - a),$$

where $T(x) = 0$ if x is on the target. Then by addition the time $T(x_0)$ is not greater than the time along $x_1(t)$ from x_0 to the target. This completes the proof.

REFERENCES

[1] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, Pa., 1969.

[2] ———, *Strengthening Caratheodory's method to apply in control problems*, Control Theory and the Calculus of Variations, A. V. Balakrishnan, ed., Academic Press, New York, 1969.

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

[4] H. HERMES AND J. LA SALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[5] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1968.

[6] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, L. W. Neustadt, ed., Interscience, New York, 1962.

SIAM J. Control
Vol. 12, No. 4, November 1974

ACCELERATED FRANK-WOLFE ALGORITHMS*

GERARD G. L. MEYER†

Abstract. This paper presents a class of iterative procedures, called accelerated Frank–Wolfe algorithms. It shows that a subclass, namely the nontrivial proper algorithms, is of special interest. An algorithm parametrized by an integer q is exhibited and it is then seen that for every $q > 0$, the algorithm is a nontrivial proper accelerated Frank–Wolfe algorithm.

Introduction. Since its presentation in 1956 (see [9]), the Frank–Wolfe algorithm, also referred to as the conditional gradient method (see [6], [12]), has been used to solve many types of optimization problems [1], [2], [5], [8], [10], [11], [15]. However, there is a drawback to the method, namely its slow convergence [4], [9]. This paper presents an acceleration procedure which differs from existing ones [2], [3] in that it does not require that new types of subproblems be solved at each iteration. The approach consists of "weakening" the inherent memory of the Frank–Wolfe algorithm.

The paper's first section is devoted to a rapid review of the Frank–Wolfe algorithm and its properties in finite-dimensional spaces. Section 2 defines a class of algorithms which are called accelerated Frank–Wolfe algorithms. It is shown that the class is far from being empty and four algorithms in it are given. This leads to a discussion of the various families of accelerated Frank–Wolfe algorithms, from which one concludes that it would be of interest to exhibit nontrivial proper accelerated Frank–Wolfe algorithms. Section 3 contains a scalar parametrized algorithm, i.e., an algorithm depending on a scalar parameter q. One sees that for $q = 0$, the algorithm is equivalent to the Frank–Wolfe algorithm and that for every $q > 0$, the algorithm is a nontrivial, proper accelerated Frank–Wolfe algorithm. Finally, § 4 gives some indications of the computational behavior of the scalar parametrized algorithm.

1. The Frank–Wolfe algorithm. This paper is restricted to the application of the Frank–Wolfe algorithm to finite-dimensional problems. For a discussion of the infinite-dimensional case, one may consult [6] or [12].

Problem 1. Given a convex, continuously differentiable map $f(\cdot)$ from R^n into R^1 and a convex, compact subset T of R^n, find a point z^* in T such that $f(z^*) \leqq f(z)$, for all z in T.

DEFINITION 1. An *algorithm is convergent* for Problem 1 if every sequence of points $\{z_i\}$, generated by it, satisfies one of the following conditions:

(i) If the sequence $\{z_i\}$ is finite, i.e., $\{z_i\} = \{z_1, z_2, \cdots, z_{k+1}\}$, then z_k is a solution of Problem 1.

(ii) If the sequence $\{z_i\}$ is infinite, then each of its cluster points is a solution of Problem 1.

* Received by the editors November 23, 1971, and in final revised form June 4, 1973.

† Department of Electrical Engineering, Johns Hopkins University, Baltimore, Maryland 21218.
This research was supported in part by the National Science Foundation under Grant GK-36221.

The Frank–Wolfe algorithm requires that an initial point y in the set T be given and it is characterized by the type of subproblems which must be solved at each iteration. These subproblems consist of the minimization of a linear map on the set T and the minimization of the map $f(\cdot)$ on a segment.

ALGORITHM 1 (Frank–Wolfe). Let y be a point in T.

Step 0. Set $z_1 = y$ and set $i = 1$.

Step 1. Compute a point v_i in T such that $\langle \nabla f(z_i), v_i \rangle \leqq \langle \nabla f(z_i), v \rangle$, for all v in T.

Step 2. Compute a scalar ξ_i in $[0, 1]$ such that $f((1 - \xi_i)z_i + \xi_i v_i) \leqq f((1 - \xi)z_i + \xi v_i)$, for all ξ in $[0, 1]$.

Step 3. Set $z_{i+1} = (1 - \xi_i)z_i + \xi_i v_i$.

Step 4. If $f(z_{i+1}) < f(z_i)$, set $i = i + 1$ and go to Step 1; else, stop.

For proof of the following theorem, see Theorem 3 and Remark 1 in § 2.

THEOREM 1. *Algorithm* 1 *is convergent for Problem* 1.

Theorem 1 shows that Algorithm 1 may be used to obtain the solution of Problem 1 but does not give any indications on the computational efficiency of the algorithm. Experiments with the Frank–Wolfe algorithm point to the fact that the behavior of the algorithm may be rather poor. This observation coincides with the analytical result obtained by Canon and Cullum [4]. These authors have studied the behavior of Algorithm 1 when applied to a special case of Problem 1. Let $f(\cdot)$ be defined by $f(z) = \langle z, Qz \rangle + \langle d, z \rangle$, let T be the set $\{z | Bz \leqq c\}$ and assume that T is bounded. Here Q is a positive definite symmetric $n \times n$ matrix, and B, d and c are $s \times n$, $n \times 1$ and $s \times 1$ matrices respectively. For this special problem the following result holds.

THEOREM 2. *Let z^* be the solution of the special problem and let $\{z_i\}$ be an infinite sequence generated by Algorithm* 1 *when applied to the special problem. Suppose that*:

(i) $z^* \in \partial T$ (*the boundary of* T);

(ii) $z_i \in \mathring{T}$ (*the interior of* T) *for infinitely many* $i \in \{1, 2, \cdots\}$.

Then, for every constant $\alpha > 0$ and for every $\varepsilon > 0$, $f(z_i) - f(z^*) \geqq \alpha/i^{1+\varepsilon}$, for infinitely many $i \in \{1, 2, \cdots\}$.

Theorem 2 does not imply that Algorithm 1 behaves inefficiently for all problems of the form of Problem 1. In fact, Levitin and Polyak in [12] have shown that Algorithm 1 is convergent at the rate of a geometric progression for Problem 1 if T is strictly convex and $\|\nabla f(z)\| \geqq \varepsilon$ on T, for some strictly positive scalar ε. It must be noted that the results obtained in [12] are given in general reflexive Banach spaces.

2. Accelerated Frank–Wolfe algorithms.

The Frank–Wolfe algorithm is applied to a wide variety of practical problems, as witnessed by the considerable amount of literature existing on the subject [2], [5], [8], [10], [11], [12]. In view of § 1, it is desirable to find improved versions of the Frank–Wolfe algorithm, i.e., algorithms which retain the desirable features of the Frank–Wolfe algorithm, without having its slow convergence.

Algorithms of the following form will be considered.

ALGORITHM 2. Let y be a point in T.

Step 0. Set $z_1 = y$, set $v_0 = y$ and set $i = 1$.

Step 1. Compute a point v_i in T, such that

$$\langle \nabla f(z_i), v_i \rangle \leqq \langle \nabla f(z_i), v \rangle \quad \text{for all } v \text{ in } T.$$

Step 2. Compute a scalar ξ_i in $[0, 1]$, such that

$$f((1 - \xi_i)z_i + \xi_i v_i) \leqq f((1 - \xi)z_i + \xi v_i) \quad \text{for all } \xi \text{ in } [0, 1].$$

Step 3. Set $y_{i0} = (1 - \xi_i)z_i + \xi_i v_i$.

Step 4. Compute a point z_{i+1} in T, such that

$$f(z_{i+1}) \leqq f(y_{i0}).$$

Step 5. If $f(z_{i+1}) < f(z_i)$, set $i = i + 1$ and go to Step 1; else, stop.

Before attempting to prove the convergence of Algorithm 2, two simple lemmas are given.

LEMMA 1. *A point z in T is a solution of Problem* 1 *if and only if* $\langle \nabla f(z), v - z \rangle \geqq 0$, *for all v in T.*

LEMMA 2. *Let z and v be such that* $f((1 - \xi)z + \xi v) \geqq f(z)$, *for all* ξ *in* $[0, 1]$. *Then* $\langle \nabla f(z), v - z \rangle \geqq 0$.

THEOREM 3. *Algorithm* 2 *is convergent for Problem* 1.

Proof. Suppose that Algorithm 2 generates only a finite sequence $\{z_1, z_2, \cdots, z_k, z_{k+1}\}$. Then $f(z_{k+1}) = f(z_k)$; otherwise the algorithm would have generated the point z_{k+2}, and $f(z_{k+1}) = f(z_k) = f(y_{k0}) \leqq f((1 - \xi)z_k + \xi v_k)$, for all ξ in $[0, 1]$. Lemma 2 implies that $\langle \nabla f(z_k), v_k - z_k \rangle \geqq 0$, and by construction $\langle \nabla f(z_k), v - v_k \rangle \geqq 0$, for all v in T. It follows that $\langle \nabla f(z_k), v - z_k \rangle \geqq 0$, for all v in T, and therefore in view of Lemma 1, z_k is a solution of Problem 1.

Suppose now that Algorithm 2 generates an infinite sequence $\{z_i\}$ and let z^* be a cluster point of this sequence. Then there exists K_1, an infinite subset of the integers, such that the subsequence $\{z_i\}_{K_1}$ converges to z^*. The set T is compact by assumption and therefore there exists K, an infinite subset of K_1, such that the subsequences $\{\xi_i\}_K$, $\{v_i\}_K$, $\{y_{i0}\}_K$ and $\{z_{i+1}\}_K$ converge to some points ξ^*, v^*, y_0^* and z^{**} respectively. The property of convergent sequences ensures that the sequence $\{z_i\}_K$ converges to z^*. The continuity of the map $\nabla f(\cdot)$ and the continuity of the scalar product with respect to its arguments imply the following relations:

$$(1) \qquad \langle \nabla f(z^*), v^* \rangle \leqq \langle \nabla f(z^*), v \rangle \quad \text{for all } v \text{ in } T,$$

$$(2) \qquad f((1 - \xi^*)z^* + \xi^* v^*) \leqq f((1 - \xi)z^* + \xi v^*) \quad \text{for all } \xi \text{ in } [0, 1],$$

$$(3) \qquad y_0^* = (1 - \xi^*)z^* + \xi^* v^*,$$

$$(4) \qquad f(z^{**}) \leqq f(y_0^*).$$

Now, assume that $f(z^{**}) \leqq f(z^*) - \delta$, for some $\delta > 0$. Then there exists k, such that $f(z_{i+1}) \leqq f(z_i) - \delta/2$, for all $i \geqq k$, i in K, i.e., the monotonically decreasing sequence $\{f(z_i)\}$ is unbounded from below. But this is impossible because $f(\cdot)$ is a continuous map and the set T is compact. It follows that there does not exist any $\delta > 0$, such that $f(z^{**}) \leqq f(z^*) - \delta$, i.e.,

$$(5) \qquad f(z^{**}) \geqq f(z^*).$$

Equations (2), (3), (4) and (5) imply that $f(z^{**}) = f(z^*) = f(y_0^*) \leqq f((1 - \xi)z^* + \xi v^*)$, for all ξ in $[0, 1]$, and Lemma 2 implies that $\langle \nabla f(z^*), v^* - z^* \rangle \geqq 0$. It follows from equation (1) that $\langle \nabla f(z^*), v - z^* \rangle \geqq 0$, for all v in T, and therefore z^* is a solution of Problem 1.

DEFINITION 2. An algorithm is an *accelerated Frank–Wolfe algorithm* if it is of the form of Algorithm 2.

There exist many algorithms of the form of Algorithm 2. To illustrate this point some of them are given now.

ALGORITHM 3. Algorithm 2 with Step 4 being defined as follows:

Step 4. Set $z_{i+1} = y_{i0}$.

ALGORITHM 4. Algorithm 2 with Step 4 being defined as follows:

Step 4. Compute a point z_{i+1} in T, such that

$$f(z_{i+1}) \leqq f(z) \quad \text{for all } z \text{ in } T.$$

ALGORITHM 5. Algorithm 2 with Step 4 being defined as follows:

Step 4. Compute a point z_{i+1} in the convex hull of $v_0, v_1, \cdots, v_i$, such that

$$f(z_{i+1}) \leqq f(z) \quad \text{for all } z \text{ in the convex hull of } v_0, v_1, \cdots, v_i.$$

ALGORITHM 6. Algorithm 2 with Step 4 being defined as follows:

Step 4. Compute a point w_i in T, such that

$$\langle \nabla f(y_{i0}), w_i \rangle \leqq \langle \nabla f(y_{i0}), w \rangle \quad \text{for all } w \text{ in } T,$$

and compute a point z_{i+1} in $[y_{i0}, w_i]$, such that $f(z_{i+1}) \leqq f(z)$, for all z in $[y_{i0}, w_i]$.

Remarks.

1. Algorithm 3 is equivalent to Algorithm 1, and therefore Theorem 3 implies Theorem 1.

2. In Algorithm 4 and Algorithm 5, the computation of z_{i+1} does not depend on y_{i0}. This implies that Step 2 and Step 3 can be deleted in these algorithms.

3. Algorithm 4 generates at most three points, namely $\{z_1, z_2, z_3\}$, where z_1 is a given point in T and z_2 is the solution of Problem 1.

Remark 1 shows that an accelerated Frank–Wolfe algorithm does not necessarily behave in a better way than Algorithm 1. The important question, which has not yet been raised, is the following: "How does one choose Step 4 in Algorithm 2, so that Algorithm 2 is preferable, computationally speaking, to Algorithm 1?" An easy and rather trivial answer is obtained by looking at Algorithm 4. Here is an algorithm which gives a solution of Problem 1 in one iteration. But it is obvious that this is not the type of algorithm one is looking for, because Step 4 in this case is as complicated as the original problem. Therefore, one is led to define a class of algorithms which does not contain Algorithm 4.

DEFINITION 3. A *proper accelerated Frank–Wolfe algorithm* is an algorithm of the form of Algorithm 2, in which the computation of z_{i+1} in Step 4 does not involve the resolution of subproblems other than:

(i) minimization of a linear functional on T;

(ii) minimization of $f(\cdot)$ on a given segment in T.

Immediately, it is apparent that Algorithm 4 and Algorithm 5 are not proper accelerated Frank–Wolfe algorithms. Now, an examination of Algorithm 3 and Algorithm 6 reveals that not much has been achieved by exhibiting these algorithms.

DEFINITION 4. An algorithm of the form of Algorithm 2 is *trivial* if every sequence $\{z_i\}$, generated by it, is a subsequence of some sequence generated by Algorithm 1.

Using Definitions 3 and 4, one can define with precision the class of algorithms one is interested in here, namely, the class of nontrivial, proper accelerated Frank–Wolfe algorithms. It can be remarked that all the algorithms given so far, i.e., Algorithms 3, 4, 5 and 6, are either nonproper or trivial.

3. Scalar parametrized algorithm. An accelerated Frank–Wolfe algorithm depending on a nonnegative integer q is now presented.

ALGORITHM 7. Let y be a point in T and let q be a scalar.

Step 0. Set $z_1 = y$, set $v_0 = y$, set $\mu_1^0 = 1$ and set $i = 1$.

Step 1. Compute a point v_i in T, such that

$$\langle \nabla f(z_i), v_i \rangle \leqq \langle \nabla f(z_i), v \rangle \quad \text{for all } v \text{ in } T.$$

Step 2. Compute a scalar ξ_i in $[0, 1]$, such that

$$f((1 - \xi_i)z_i + \xi_i v_i) \leqq f((1 - \xi)z_i + \xi v_i) \quad \text{for all } \xi \text{ in } [0, 1].$$

Step 3. Set $j = 0$, set $\lambda_{ij}^k = (1 - \xi_i)\mu_i^k$ for $k = 0, 1, \cdots, i - 1$, set $\lambda_{ij}^i = \xi_i$ and set $y_{ij} = (1 - \xi_i)z_i + \xi_i v_i$.

Step 4a. If $j \geqq q$, go to Step 4e; else, go to Step 4b.

Step 4b. Set $I_{ij} = \{k \in \{0, 1, \cdots, i\} | \langle \nabla f(y_{ij}), v_k - y_{ij} \rangle \leqq 0\}$, set $\psi_{ij} = 1/(\sum_{k \in I_{ij}} \lambda_{ij}^k)$, and set $x_{ij} = \psi_{ij} \sum_{k \in I_{ij}} \lambda_{ij}^k v_k$.

Step 4c. Compute a scalar χ_{ij} in $[0, 1]$, such that

$$f((1 - \chi_{ij})y_{ij} + \chi_{ij}x_{ij}) \leqq f((1 - \chi)y_{ij} + \chi x_{ij}) \quad \text{for all } \chi \text{ in } [0, 1].$$

Step 4d. Set $\lambda_{i(j+1)}^k = (1 - \chi_{ij} + \psi_{ij}\chi_{ij})\lambda_{ij}^k$, for $k \in I_{ij}$; set $\lambda_{i(j+1)}^k = (1 - \chi_{ij})\lambda_{ij}^k$, for $k \notin I_{ij}$; set $y_{i(j+1)} = (1 - \chi_{ij})y_{ij} + \chi_{ij}x_{ij}$; set $j = j + 1$ and go to Step 4a.

Step 4e. Set $z_{i+1} = y_{iq}$ and set $\mu_{i+1}^k = \lambda_{iq}^k$, $k = 0, 1, \cdots, i$.

Step 5. If $f(z_{i+1}) < f(z_i)$, set $i = i + 1$ and go to Step 1; else, stop.

It is clear, in view of Definition 3, that for any scalar q, Algorithm 7 is a proper accelerated Frank–Wolfe algorithm. This immediately implies that, for any scalar q, Algorithm 7 is convergent for Problem 1.

One may characterize an algorithm by the class of problems on which it is finitely convergent. Consider the following special case of Problem 1.

Problem 1′. Given a convex, continuously differentiable map $f(\cdot)$ from R^2 into R^1 and a finite number of points $x_1, x_2, \cdots, x_p$ in R^2, let $T = [x_1, x_2, \cdots, x_p]$ and suppose that $\nabla f(z) \neq 0$ for every z in T. Find a point z^* in T such that $f(z^*) \leqq f(z)$ for every z in T.

THEOREM 4. *When applied to Problem* 1′, Algorithm 7 with $q > 0$ gives the solution of the problem in a finite number of steps, i.e., Algorithm 7 is finitely convergent.

The proof of the theorem is somewhat lengthy but not very complicated and has been deleted. A simple example proposed by one of the reviewers illuminates the difference in behavior between Algorithm 1 and Algorithm 7 with $q > 0$.

Example 1. Let $x_1 = (-1, 0.5)$, $x_2 = (1, 0.5)$, $x_3 = (0, 20)$, $y = (-0.5, 1)$ and $f(x) = \|x\|_2$. In this case Algorithm 1 will generate an infinite sequence of points

$\{z_i\}$ converging to $z^* = (0, 0.5)$. If Algorithm 7 with $q = 1$ is used, then a finite sequence of points is generated, namely:

$$z_1 = (-0.5, 1);$$

$$z_2 = (0.25, 0.75);$$

$$z_3 = (0.06, 0.5);$$

$$z_4 = (0, 0.5) = z^*;$$

$$z_5 = (0, 0.5) = z^*.$$

Example 1 shows that Algorithm 1 may not converge in a finite number of steps on Problem 1. This remark coupled with Theorem 4 shows that Algorithm 7 is a nontrivial proper accelerated Frank–Wolfe algorithm.

Now a short heuristic argument is given to justify the form of Algorithm 7. It is well known that Algorithm 7, with $q = 0$, is not a very efficient algorithm when the μ_i^k coefficients do not die off fast enough. In other words, the algorithm shows its limitations when, at iteration i, the coefficients μ_i^k are different from 0, for k much smaller than i. The idea behind the form of Algorithm 7 is to use a procedure which has a tendency to decrease the dependence of the point z_i on the points v_k for k much smaller than i. This effect is obtained by constructing the auxiliary points x_{ij} which have the property that $\langle x_{ij} - y_{ij}, \nabla f(y_{ij})\rangle \leqq 0$. The procedure adopted to construct the points x_{ij} does not require that any new subproblem be solved.

It does not seem possible to show analytically that, for some value of $q > 0$, Algorithm 7 behaves in a better way than Algorithm 1. The only recourse is to experiment with the algorithm and try to determine if there exists a value of $q > 0$, for which the algorithm does not have the slow convergence of Algorithm 1.

4. Computational results. In order to obtain an idea of the computational behavior of Algorithm 7, the following problem was considered.

Problem 2. Let T be the subset of R^3 defined by

$$T = \{x | x^1 \geqq v + (x^2)^2/2\beta_2 + (x^3)^2/2\beta_3, x^1 \leqq 10\}.$$

Find a point z^* in T, such that $\|z^*\|_2 \leqq \|z\|_2$, for all z in T.

Remark 4. This type of problem was used by Barr in [2]. The notation $\|\cdot\|_2$ is used to denote the Euclidean norm, i.e., $\|z\|_2 = ((z^1)^2 + (z^2)^2 + (z^3)^2)^{1/2}$. For any $\beta_2 > 0$ and any $\beta_3 > 0$, the set T is convex and the solution of Problem 2 is $z^* = (v, 0, 0)$.

Problem 2 was solved for various values of β_2, β_3 and v. Algorithm 7 was used with $y = (6, 2, 2)$ and two values of q, namely $q = 0$ and $q = 2$. The results of the experiments are contained in Tables 1, 2, 3 and 4. These tables indicate the number of iterations N_0 and N_2 necessary to obtain a point z in T satisfying $\|z\|_2 \leqq \|z^*\|_2 + (\|y\|_2 - \|z^*\|_2)/c$ when Algorithm 7 is used with $q = 0$ and $q = 2$ respectively. The reader may note that it is preferable to use $q = 2$. However in cases in which the Frank–Wolfe method is efficient, i.e., N_0 small, there is no marked advantage in using $q = 2$. This tends to indicate that the acceleration procedure takes effect only when really needed.

TABLE 1

	$v = 1$ $\beta_2 = 10$ $\beta_3 = 10$		$v = 1$ $\beta_2 = 100$ $\beta_3 = 100$		$v = 1$ $\beta_2 = 500$ $\beta_3 = 500$	
c	N_0	N_2	N_0	N_2	N_0	N_2
5	3	3	6	6	15	8
10	6	3	6	10	15	12
50	9	3	7	12	16	12
100	10	3	11	12	16	14
500	20	7	27	22	76	16
1,000	21	7	28	22	76	38
5,000	27	9	51	23	76	43
10,000	29	11	57	23	84	43
50,000	31	15	71	23	92	52
100,000	31	17	90	23	100	52

TABLE 2

	$v = 1$ $\beta_2 = 10$ $\beta_3 = 1$		$v = 1$ $\beta_2 = 100$ $\beta_3 = 1$		$v = 1$ $\beta_2 = 500$ $\beta_3 = 1$	
c	N_0	N_2	N_0	N_2	N_0	N_2
5	3	3	2	2	23	11
10	7	3	2	2	34	15
50	10	4	6	6	42	15
100	14	5	24	11	42	16
500	15	9	24	19	42	20
1,000	15	9	24	19	50	21
5.000	20	11	37	27	98	35
10,000	22	13	37	31	98	54
50,000	25	16	45	40	106	55
100,000	25	18	66	40	107	55

TABLE 3

	$v = 0.5$ $\beta_2 = 10$ $\beta_3 = 10$		$v = 0.5$ $\beta_2 = 100$ $\beta_3 = 100$		$v = 0.5$ $\beta_2 = 500$ $\beta_3 = 500$	
c	N_0	N_2	N_0	N_2	N_0	N_2
5	3	3	6	6	28	16
10	6	3	6	7	61	22
50	8	7	6	7	62	23
100	8	9	11	15	62	23
500	12	14	31	28	117	35
1,000	12	14	51	35	150	40
5,000	16	18	79	36	181	51
10,000	16	18	98	38	183	52
50,000	20	23	98	38	185	69
100,000	20	23	98	39	233	69

TABLE 4

	$\nu = 0.5$ $\beta_2 = 10$ $\beta_3 = 1$		$\nu = 0.5$ $\beta_2 = 100$ $\beta_3 = 1$		$\nu = 0.5$ $\beta_2 = 500$ $\beta_3 = 1$	
c	N_0	N_2	N_0	N_2	N_0	N_2
5	4	3	3	2	30	11
10	5	4	3	2	41	16
50	13	8	17	14	83	40
100	13	11	25	14	124	50
500	13	13	84	20	177	95
1,000	13	15	104	26	177	95
5,000	13	21	160	28	193	99
10,000	14	21	162	32	226	107
50,000	17	24	178	36	251	132
100,000	18	26	191	41	251	143

Remark 5. The stop rule used in the numerical experiments is unconventional. It involves the initial guess y, the optimal point z^* and the cost function $\|\cdot\|_2$. If by chance, the initial guess y satisfies $\|y\|_2 - \|z^*\|_2 = 1$, then one gets the usual stop rule $\|z\|_2 \leqq \|z^*\|_2 + 1/c$. In the case at hand, $\|y\|_2 = \sqrt{44}$ and therefore the quantity $\|y\|_2 - \|z^*\|_2$ can be interpreted as a scaling factor.

One iteration of Algorithm 7 with $q = 0$ requires one Definition 3(i) minimization and one Definition 3(ii) minimization; one iteration of Algorithm 7 with $q = 2$ requires one Definition 3(i) minimization and three Definition 3(ii) minimizations. Since Definition 3(i) minimizations are the more difficult in applications [2], [8], [11], it is meaningful to use N_0 and N_2 as indicators of efficiency.

One drawback of Algorithm 7 as written is the ever growing data base. It is possible to limit this growth by specifying the maximum length of the string of v_i and restarting the algorithm when the allowable length has been reached. To be more specific one may add a Step 4f in Algorithm 7 between Step 4e and Step 5 defined as follows.

Step 4f. If the allowable length of the string of v_i has been reached, set $y = z_{i+1}$ and go to Step 0.

It is clear that Algorithm 7 with $q > 0$ and Step 4f included is still a nontrivial proper accelerated Frank–Wolfe algorithm. In fact this modified algorithm converges in a finite number of steps on Problem 1′ if the allowable length of the string of v_i is at least two.

5. Conclusion. It has been shown that many accelerated Frank–Wolfe algorithms exist. The discussion of the different types of such algorithms has indicated that it is not possible to give a realistic answer to the question: find the best accelerated algorithm? This is due to the existence of a nonproper algorithm which solves Problem 1 in one iteration. This has lead to the definition of a restricted class of algorithms, namely the class of nontrivial proper accelerated Frank–Wolfe algorithms. The main purpose of this paper has been to show that such algorithms exist and are efficient. At this point, one would like to emphasize that nonproper, nontrivial algorithms may be of great interest, as witnessed by the results obtained by Barr (see [2] and [3]), but this paper has been restricted to proper accelerated algorithms.

REFERENCES

[1] A. Auslender and F. Brodeau, *Convergence d'un algorithme de Frank et Wolfe appliqué à un problème de contrôle*, Rev. Française Informat. Recherche Opérationnelle, no. 7, 1968. pp. 3–12.

[2] R. O. Barr, *Computation of optimal controls on convex reachable sets*, Mathematical Theory of Control, Academic Press, New York, 1967, pp. 63–70.

[3] ———, *An efficient computational procedure for a generalized quadratic programming problem*, this Journal, 7 (1969), pp. 415–429.

[4] M. D. Canon and C. D. Cullum, *A tight upper bound on the rate of convergence of the Frank–Wolfe algorithm*, this Journal, 6 (1968), pp. 509–516.

[5] V. F. Dem'yanov, *Determination of the optimum program in a linear system*, Avtomat. i Telemeh., 25 (1964), no. 1, pp. 3–11.

[6] V. F. Dem'yanov and A. M. Rubinov, *The minimization of a smooth convex functional on a convex set*, this Journal, 5 (1967), pp. 280–294.

[7] J. H. Eaton, *An iterative solution to time-optimal control*, J. Math. Anal. Appl., 5 (1962), pp. 329–344.

[8] P. S. Fancher, *Iterative computation procedures for an optimum control problem*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 346–348.

[9] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.

[10] T. Fujisawa and Y. Yasuda, *An iterative procedure for solving the time-optimal regulator problem*, this Journal, 5 (1967), pp. 501–512.

[11] E. G. Gilbert, *An iterative procedure for computing the minimum of quadratic form on a convex set*, this Journal, 4 (1966), pp. 61–80.

[12] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, Ž. Vyčisl. Mat. i Mat. Fiz., 6 (1969), no. 5, pp. 787–823 = U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), no. 5, pp. 1–50.

[13] E. Polak, *On the convergence of optimization algorithms*, Rev. Française Informat. Recherche Opérationnelle, Serie Rouge, 1969, no. 16, pp. 17–34.

[14] B. T. Polyak, *Gradient methods for the minimization of functionals*, Ž. Vyčisl. Mat. i Mat. Fiz., 3 (1963), no. 4, pp. 643–653 = U.S.S.R. Comput. Math. and Math. Phys., 3 (1963), no. 4, pp. 864–878.

[15] M. Valadier, *Extension d'un algorithm de Frank–Wolfe*, Rev. Française Recherche Opérationnelle, 1965, no. 36, pp. 251–253.

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

INSTABILITY OF NONLINEAR INFINITE-DIMENSIONAL FEEDBACK SYSTEMS USING LYAPUNOV FUNCTIONALS*

THOMAS L. STEDING AND ARTHUR R. BERGEN†

Abstract. Sufficient conditions are given for the L_2 instability of a broad class of nonlinear, time-varying feedback systems. The system under consideration is assumed to be decomposed into two subsystems: one passive and nonlinear, not necessarily memoryless, and the other unstable and linear, not necessarily finite-dimensional. The main results essentially state that the feedback system is unstable if the linear subsystem is strictly passive and bounded on a proper subset of L_2. The results apply both to instability in the input-output sense and to instability of unforced systems.

The principal conceptual tools of the analysis are a Lyapunov function and a state, both of which are defined on the linear subsystem in a manner not depending on the dimensionality of the system. As an application of these results, an instability counterpart to the circle criterion is presented which applies to a class of systems more general than those of previous results. The conditions of this counterpart imply, in addition to L_2 instability, global state space instability. Conditions are also given under which the Lyapunov function defined in the analysis may be used to establish asymptotic stability in the large.

1. Introduction. We present herein certain theorems on the L_2 instability of a broad class of nonlinear, time-varying feedback systems. The system under investigation is illustrated in Fig. 1, where the linear time-invariant operator G is assumed to be the sum of an infinite-dimensional, stable component, G_1, and a finite-dimensional, unstable component, G_2; the nonlinear time-varying operator N is assumed to be passive on L_2. The key results of the paper, given in § 3, essentially state that the feedback system is unstable if G is strictly passive and bounded on a certain proper subset of L_2. The analysis involves a state and a Lyapunov function, both of which are defined on the linear subsystem G using only its impulse response, and not requiring a finite-dimensional vector differential representation. In contrast with recent efforts in instability theory, our results apply both to instability in the input-output sense *and* to instability of unforced systems (where the zero-input response of G is constrained to arise from previous inputs).

In 1967 Brockett and Lee [3] published an instability counterpart to the circle criterion where the linear subsystem was required to have a rational transfer function. Instability results in a functional analysis setting were derived by J. C. Willems [9]. Departing from the state instability formulation of [3], these results deal with input-output unboundedness in the L_2 sense. Conditions derived in [9] are similar to those in [3], but admit infinite-dimensional linear subsystems.

We give, as an application of the results of § 3, an instability counterpart to the circle criterion which is of greater generality than the above results. In particular, we extend the results of [9] to nonlinear systems while simultaneously reformulating the analysis of L_2 instability using such concepts as passivity, energy, and state. In fact, the pattern of our analysis is closer to that used in [3], and includes their state space instability result as a special case.

* Received by the editors May 13, 1971, and in final revised form August 20, 1973.

† Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720. This research was supported by the Joint Services Electronics Program under Grant AFOSR-68-1488.

Recently, in a separate effort, Bergen and Takeda [2] also generalized the results in [3], and paralleled some of the work of this paper. Their analysis, however, depending on notions of causality, is based on a different approach than the one used here.

The outline of the paper is as follows. After preliminaries in § 2, we give the main results of the paper in § 3. The analysis begins in § 4 with a definition of a function which will serve as a state for the infinite-dimensional, linear subsystem; at each instant of time, this state is a real-valued function defined on the interval $[0, \infty)$. The L_2-boundedness of the system solutions is then related to the behavior of the state for large values of time. Following this, a Lyapunov function is defined on the linear subsystem G by considering certain energy-like quantities associated with G and the state. After establishing the existence and certain further properties of this Lyapunov function, the proof of the results follows directly. The instability counterpart of § 5 is established from these results by transforming the system under consideration to one satisfying the hypotheses of § 3. Finally, we make some concluding remarks in § 6.

The Lyapunov function and state used in the analysis are analogous to the concepts developed by Baker and Bergen in [1]. Contrary to their results, however, our Lyapunov function is not shown to possess all of the usual properties (e.g., radial unboundedness) of Lyapunov functions, but only those necessary for the proof of the main result. Thus, "quasi-Lyapunov function" may be a formally more appropriate name for this function; for brevity, however, we have carried over the terminology of [1].

2. Preliminary notations and definitions. The framework used here is an adaptation of the functional analytic approach to stability analyses (see, e.g., [11]). Let R denote the real number and let Ω be the set of all measurable functions mapping R into R. For real t and for f belonging to Ω the *truncation operator*, P_t, is defined by

$$(1) \qquad P_t f(\tau) = \begin{cases} f(\tau) & \text{for } \tau \leqq t, \\ 0 & \text{for } \tau > t; \end{cases}$$

and the *shift operator*, S_t, is defined by

$$(2) \qquad S_t f(\tau) = f(t + \tau)$$

for all real τ. For simplicity of notation, we write P in place of P_0.

We use the notation $f_{[t_1,t_2]}$ to mean the function which coincides with f over the interval $[t_1, t_2]$ and is zero elsewhere. We write f^+ for $f_{[0,\infty)}$, and f^- for $f_{(-\infty,0]}$.

The analysis is carried out on the Hilbert spaces[1]

$$(3) \qquad L_2(R) = \left\{ f \in \Omega \middle| \, \|f\|_2^2 = \int_{-\infty}^{\infty} |f(t)|^2 \, dt < \infty \right\}$$

and

$$(4) \qquad L_{2_\sigma}(R) = \left\{ f \in \Omega \middle| \, \|f\|_\sigma^2 = \int_{-\infty}^{\infty} |f(t)|^2 e^{-\sigma t} \, dt < \infty \right\},$$

[1] Since we never explicitly consider the case $\sigma = 2$, the notation is unambiguous.

for some real σ, where the inner products on $L_2(R)$ and $L_{2_\sigma}(R)$ are given by, respectively,

$$(5) \qquad \langle f_1, f_2 \rangle_2 = \int_{-\infty}^{\infty} f_1(t) f_2(t)\, dt$$

and

$$(6) \qquad \langle f_1, f_2 \rangle_\sigma = \int_{-\infty}^{\infty} f_1(t) f_2(t) e^{-\sigma t}\, dt.$$

We write L_2 and L_{2_σ} in place of, respectively, $L_2(R)$ and $L_{2_\sigma}(R)$. If F is an operator on L_2 (L_{2_σ}), its operator norm is denoted by $\|F\|_2$ ($\|F\|_\sigma$).

The *extended space* of L_2, denoted by L_{2_e}, is defined by

$$(7) \qquad L_{2_e} = \{f \in \Omega | P_t f \in L_2 \text{ for all } t \text{ in } R\}.$$

We let L_2^+ denote that subspace of L_2 given by

$$(8) \qquad L_2^+ = \{f \in L_2 | f(t) = 0 \text{ for almost all } t < 0\},$$

with analogous definitions for L_2^-, $L_{2_\sigma}^+$, $L_{2_\sigma}^-$, and $L_{2_e}^+$.

3. Problem formulation and principal results. We consider the feedback system equations (Fig. 1):

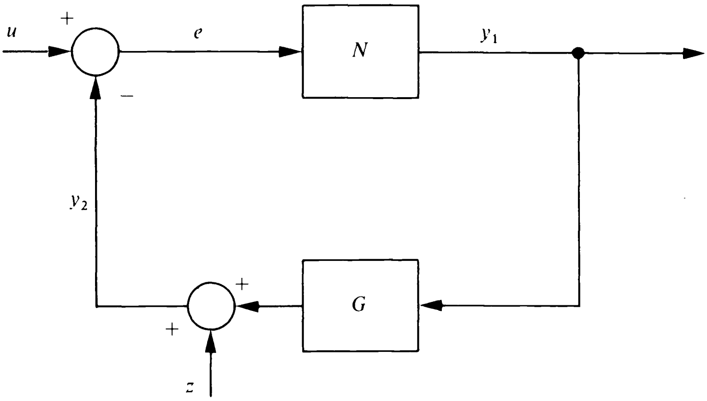

Fig. 1. *The system S*

$$(9) \qquad (I + GN)e = u - z,$$

$$(10) \qquad y_1 = Ne,$$

$$(11) \qquad y_2 = Gy_1 + z,$$

where u is the input; z is the zero-input response of the subsystem G arising from initial conditions; and e, y_1, or y_2 can be viewed as outputs. We make the following assumptions.

A1. There exist linear, time-invariant, nonanticipative operators G_1 and G_2 such that on the intersection of their domains,

$$(12) \qquad G = G_1 + G_2,$$

where

(i) G_1 maps $L_{2_e}^+$ into $L_{2_e}^+$ and L_2 into L_2, and is a bounded operator on L_2;

(ii) for all $t \geqq 0$, and $y_1 \in L_{2_e}$,

$$G_2 y_1(t) = \int_{-\infty}^{t} g_2(t - \tau) y_1(\tau)\, d\tau, \tag{13}$$

where $g_2(t) = 0$ for all $t < 0$ and g_2 is not the zero function. The Laplace transform of g_2, $G_2(s)$, exists and satisfies

$$G_2(s) = l(s)/m(s), \tag{14}$$

where $l(s)$ and $m(s)$ are real polynomials, the degree of $m(s)$ exceeds that of $l(s)$, and for some real $\sigma > 0$ the zeros of $m(s)$ lie in the open strip $0 < \operatorname{Re} s < \sigma/2$ in the complex plane. Zeros on the $j\omega$-axis are excluded.

A2. Let D_N be a set in $L_{2_e}^+$. The nonanticipative relation N is a subset of $D_N \times L_{2_e}^+$ and its restriction to L_2^+ is a subset of $(D_N \cap L_2^+) \times L_2^+$.

A3. u, e, y_1, y_2, and z are real-valued functions defined on the real line and equal to zero over the interval $(-\infty, 0)$; furthermore,

(i) u and $z \in L_{2_\sigma}^+$;

(ii) e, y_1, and $y_2 \in L_{2_e}^+$;

(iii) when e solves (9), then $e \in D_N$.

Standard arguments show that assumption A1 (ii) implies that G_2 maps $L_{2_e}^+$ into $L_{2_e}^+$ and is a bounded operator on L_{2_σ}. This together with A1 (i) ensures that the operator $G = G_1 + G_2$ is defined on the domains L_{2_σ} and $L_{2_e}^+$. Note that from A1 (ii), G_2 is not the zero operator.

The instability problem is to find conditions on G and N under which there is an input u and/or a zero-input response z such that e belongs to[2] $L_{2_e}^+ \sim L_2^+$.

We shall be investigating two separate cases: (i) the system S is unforced ($u \equiv 0$); (ii) u is an (arbitrary) unspecified element of $L_{2_\sigma}^+$ and the initial conditions are considered to be fixed. In the unforced case, to avoid physically meaningless results, we restrict our consideration to those initial conditions belonging to the set

$$Z = \{z \in L_{2_\sigma}^+ \mid z = G^+\zeta \text{ for some } \zeta \in L_{2_\sigma}^-\}, \tag{15}$$

where $G^+ = (I - P)G$. We thus require that z arise from inputs to the subsystem G occurring previous to time $t = 0$, a condition which we accept as guaranteeing that the initial condition z corresponds to those occurring in the physical system.

Since, as shown in § 4, G^+ is a bounded map from $L_{2_\sigma}^-$ into $L_{2_\sigma}^+$, Z is indeed a subset of $L_{2_\sigma}^+$. Also when G has a rational transfer function the class Z coincides with the usual set of zero-input responses generated by choosing all possible initial states in the corresponding finite-dimensional space.

Finally, we define the subspace $M \subset L_{2_\sigma}^-$ by

$$M = \mathscr{N}(G_2^+) \cap L_{2_\sigma}^-, \tag{16}$$

where $\mathscr{N}(G_2^+)$ is the null space of the operator $G_2^+ (= (I - P)G_2)$ defined on $L_{2_\sigma}^-$. We may now state the main result of this paper.

[2] The symbol $\sim$ denotes set complement.

THEOREM 1. *Under the assumptions* A1, A2 *and* A3, if

(a) *for all* $e \in D_N \cap L_2^+$,

$$\langle Ne, e\rangle_2 \geqq 0, \tag{17}$$

and

(b) *there are constants* $\delta > 0$ *and* $\lambda > 0$ *such that for all* $\eta \in M$,

$$\langle G\eta, \eta\rangle_2 \geqq \delta\|\eta\|_2^2 \tag{18}$$

and

$$\|G\eta\|_2 \leqq \lambda\|\eta\|_2, \tag{19}$$

then:

(i) *if* $u \equiv 0$, *there is a* $z \in Z$ *such that if* e *solves* (9), *then* $e \in L_{2_e}^+ \sim L_2^+$;

(ii) *otherwise, for* any $z \in Z$ *there is a* $u \in L_{2_\sigma}^+$ *such that if* e *solves* (9), *then* $e \in L_{2_e}^+ \sim L_2^+$.

Comments. Although G represents an unstable subsystem, M is one class of inputs over which G produces L_2-*bounded* outputs; in fact, M is precisely that class of inputs in $L_{2_\sigma}^-$ for which the exponentially growing terms of the output (due to G_2) do not appear over the interval $[0, \infty)$. Theorem 1 states, then, that if G has a *nontrivial* unstable component, but is well-behaved over M, the feedback system is unstable in the sense described.

4. Analysis. For purposes of the analysis, we assume a solution e to (9) exists in $L_{2_e}^+$ in D_N. Since part (ii) of Theorem 1 follows directly from part (i), we assume hereafter that $u \equiv 0$, and hence restrict z to belong to Z.

Before proceeding to the definition of the state, we note that G^+ is a bounded map from $L_{2_\sigma}^-$ into $L_{2_\sigma}^+$ since if ζ belongs to $L_{2_\sigma}^-$,

$$\begin{aligned}\|G^+\zeta\|_\sigma &= \|((G_1 + G_2)\zeta)^+\|_\sigma\\ &\leqq \|(G_1\zeta)^+\|_\sigma + \|(G_2\zeta)^+\|_\sigma\\ &\leqq \|(G_1\zeta)^+\|_2 + \|(G_2\zeta)^+\|_\sigma\\ &\leqq \|G_1\|_2\|\zeta\|_2 + \|G_2\|_\sigma\|\zeta\|_\sigma\\ &\leqq \|G_1\|_2\|\zeta\|_\sigma + \|G_2\|_\sigma\|\zeta\|_\sigma.\end{aligned} \tag{20}$$

4.1. Definition and properties of the state of the subsystem G. Consider the system S. For each $t \geqq 0$ the *state of the subsystem* G, $x_t(\cdot)$, is a real-valued function defined on $[0, \infty)$ given by

$$x_t \triangleq (I - P)S_t[Gy_{1[0,t]} + z], \qquad t \geqq 0. \tag{21}$$

Therefore x_t is simply the backward shift of t units of the zero-input response of the subsystem G over $[t, \infty)$ with input $y_{1[0,t]}$. Hence x_t completely summarizes the effect of past inputs to G. The shift ensures that a state is independent of the time it is reached. Note that $x_0 = z$.

It is interesting to compare this state with the usual state of a finite-dimensional system. Thus, suppose G has a rational transfer function, and let $\{A, b, c, d\}$ be a minimal representation of the linear subsystem G with state

$\mathbf{x} \in R^n$. Then it is easily shown that for each (fixed) $t \geqq 0$,

$$x_t(\tau) = ce^{A\tau}\mathbf{x}(t), \qquad \tau \geqq 0. \tag{22}$$

In addition, since the above representation is completely observable, there is a one-to-one correspondence between the states $\mathbf{x}(t)$ and x_t.

As defined, the state x_t does not necessarily correspond to the internal state of the subsystem G, but is only determined by its input-output structure. Indeed, in the above example, if the representation $\{A, b, c, d\}$ is not minimal, then x_t contains no information about the unobservable portion of $\mathbf{x}(t)$.

The *trajectory* of the subsystem G is the set $\{x_t | t \geqq 0\}$. A particular state x_t will often be referred to as a *point* on the trajectory.

The following relationships are easily established: between two points on a trajectory,

$$x_t = (I - P)S_t[Gy_{1[t_1,t]} + S_{-t_1}x_{t_1}], \qquad 0 < t_1 < t; \tag{23}$$

between the output over $[t, \infty)$, the input over $[t, \infty)$, and the state at time t,

$$y_{2[t,\infty)} = Gy_{1[t,\infty)} + S_{-t}x_t, \qquad 0 < t. \tag{24}$$

These relationships can be used to show that the subsystem G along with the state x_t and the appropriate spaces from § 3 form a dynamical system [4, p. 49].

The following lemma establishes some additional properties of x_t.

LEMMA 1. *For all x_t on a trajectory of the system S,*

(i) *$x_t \in L_{2_\sigma}^+$ for all $t > 0$;*

(ii) *when $x_0 \in Z$, x_t can be written for all $t \geqq 0$ as*

$$x_t = x_{1t} + x_{2t}, \tag{25}$$

where $x_{1t} \in L_2^+$ and x_{2t} is either the zero function or belongs to $L_{2_\sigma}^+ \sim L_2^+$;

(iii) *when e belongs to L_2^+, $\|x_{1t}\|_2$, $\|x_{2t}\|_\sigma$, and $\|x_t\|_\sigma$ all $\to 0$ as $t \to \infty$.*

Proof. Part (i) follows immediately from the definition of x_t. To show (ii), define

$$x_{1t} = G_1^+ S_t y_{1[0,t]} + (I - P)S_t G_1^+ \zeta, \tag{26}$$

$$x_{2t} = G_2^+ S_t y_{1[0,t]} + (I - P)S_t G_2^+ \zeta, \tag{27}$$

where $G_i^+ = (I - P)G_i$, $i = 1$ or 2, and ζ is an element of $L_{2_\sigma}^-$ such that $G^+\zeta = x_0$. Then $x_t = x_{1t} + x_{2t}$, and $x_{1t} \in L_2^+$. Also, x_{2t} is the sum of exponentially growing terms whose coefficients are possibly all zero, and whose exponents λ_i satisfy $0 < \operatorname{Re} \lambda_i < \sigma/2$. Thus either $x_{2t} = 0$ or $x_{2t} \in L_{2_\sigma}^+ \sim L_2^+$.

To show (iii), assume $e \in L_2^+$. Then from (21) and (11),

$$\begin{aligned} x_t &= (I - P)S_t G y_1 + (I - P)S_t x_0 - (I - P)S_t G y_{1[t,\infty)} \\ &= (I - P)S_t y_2 - (I - P)S_t G y_{1[t,\infty)} \\ &= (I - P)S_t y_2 - S_t G y_{1[t,\infty)}. \end{aligned} \tag{28}$$

Using the triangle inequality, the properties of the L_2- and L_{2_σ}-norms, and A1 give

$$
\begin{aligned}
\|x_t\|_\sigma &\leqq \|(I - P)S_t y_2\|_\sigma + \|S_t G y_{1[t,\infty)}\|_\sigma \\
&\leqq \|(I - P)S_t y_2\|_2 + \|S_t G_1 y_{1[t,\infty)}\|_\sigma + \|S_t G_2 y_{1[t,\infty)}\|_\sigma \\
&\leqq \|y_{2[t,\infty)}\|_2 + \|G_1 S_t y_{1[t,\infty)}\|_2 + \|G_2 S_t y_{1[t,\infty)}\|_\sigma \\
&\leqq \|y_{2[t,\infty)}\|_2 + \|G_1\|_2 \|S_t y_{1[t,\infty)}\|_2 + \|G_2\|_\sigma \|S_t y_{1[t,\infty)}\|_\sigma \\
&\leqq \|y_{2[t,\infty)}\|_2 + (\|G_1\|_2 + \|G_2\|_\sigma)\|S_t y_{1[t,\infty)}\|_2 \\
&\to 0 \quad \text{as } t \to \infty,
\end{aligned}
\tag{29}
$$

since y_1 and y_2 are in L_2.

Using an analogous argument, $\|x_{1t}\|_2 \to 0$ as $t \to \infty$. Finally,

$$
\begin{aligned}
\|x_{2t}\|_\sigma &= \|x_t - x_{1t}\|_\sigma \\
&\leqq \|x_t\|_\sigma + \|x_{1t}\|_2 \\
&\to 0 \quad \text{as } t \to \infty.
\end{aligned}
\tag{30}
$$

4.2. Definition and properties of the Lyapunov function. Having defined a state appropriate for infinite-dimensional systems, we are in a position to construct a Lyapunov function related to the system S. Let D_L be a domain in $L_{2_\sigma}^+$. We define the map $L: D_L \to R$ by

$$L(x_t) = \inf\{\langle G\zeta, \zeta\rangle_2 | \zeta \in L_{2_\sigma}^- \text{ and } G^+\zeta = x_t\}. \tag{31}$$

Now for $\zeta \in L_{2_\sigma}^-$, $G\zeta \in L_{2_\sigma}$, and hence $PG\zeta \in L_2$. Also, since $L_{2_\sigma}^- \subset L_2$, $\zeta \in L_2$ and thus the quantity $\langle G\zeta, \zeta\rangle_2 = \langle PG\zeta, \zeta\rangle_2$ is well-defined. We will need the following inequality:

$$
\begin{aligned}
\|PG\zeta\|_2 &= \|PG_1\zeta + PG_2\zeta\|_2 \\
&\leqq \|PG_1\zeta\|_2 + \|PG_2\zeta\|_2 \\
&\leqq \|\mathrm{PG}_1\zeta\|_2 + \|PG_2\zeta\|_\sigma \\
&\leqq \|G_1\zeta\|_2 + \|G_2\zeta\|_\sigma \\
&\leqq \|G_1\|_2\|\zeta\|_2 + \|G_2\|_\sigma\|\zeta\|_\sigma \\
&\leqq (\|G_1\|_2 + \|G_2\|_\sigma)\|\zeta\|_\sigma \\
&\triangleq k_1\|\zeta\|_\sigma.
\end{aligned}
\tag{32}
$$

Suppose for the moment that the indicated infimum is finite. If G is the input impedance of a 1-port network, L has the following physical interpretation (see Fig. 2). Let ζ be an input current source and $G\zeta$ be the corresponding terminal voltage. Then $L(x_t)$ is the infimum of the energy pumped into the terminals 1–1′ using $L_{2_\sigma}^-$ inputs which drive from the zero state at time $\tau = -\infty$ to the state x_t at time $\tau = 0$. Equivalently, $L(x_t)$ is the negative of the supremum of the energy *extracted* from 1–1′ in driving to x_t. Thus if $L(x_t)$ is negative, at least one function exists in L_2^- that in driving to x_t generates a net outward flow of energy over the interval $(-\infty, 0]$.

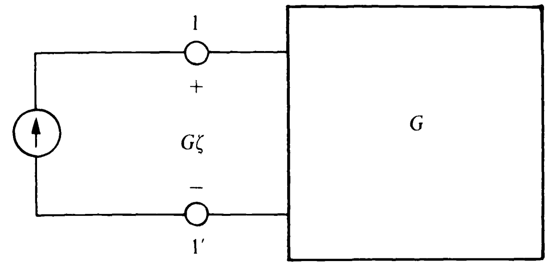

FIG. 2. *Network interpretation of* $L(x_t)$

For the function L to be meaningful in our analysis, we first must establish that $L(x_t) > -\infty$ along trajectories of interest. It is not immediately evident, however, that this property holds for a given state x_t. Indeed, the restriction of the map G^+ to $L_{2_\sigma}^-$ is in general not onto $L_{2_\sigma}^+$, and the set

$$C(x_t) = \{\zeta \in L_{2_\sigma}^- | G^+\zeta = x_t\}, \tag{33}$$

over which the infimum is taken, will therefore be void for certain states. Furthermore, the unconstrained minimization $\inf\langle G\zeta, \zeta\rangle$ is not finite, since (Lemma 5) there exists at least one $\eta \in L_{2_\sigma}^-$ such that $\langle G\eta, \eta\rangle < 0$, and hence $\langle Gk\eta, k\eta\rangle \to -\infty$ as $K \to \infty$. This minimization thus provides no useful lower bound for $L(x_t)$.

LEMMA 2. *Assume* $x_0 \in Z$. *Then* $L(x_t) > -\infty$ *for all* x_t *on the resulting system trajectory.*

Proof. Since $x_0 \in Z$, there exists an element μ in $L_{2_\sigma}^-$ satisfying $G^+\mu = x_0$. Then, for all $t \geqq 0$, the function $\eta = S_t(\mu + y_{1[0,t]})$ belongs to $C(x_t)$. Hence $C(x_0) \neq \varnothing$ implies $C(x_t) \neq \varnothing$ for all $t \geqq 0$.

Since the restriction $G_2^+ : L_{2_\sigma}^- \to L_{2_\sigma}^+$ is bounded, $\mathcal{N}(G_2^+)$ is a closed subspace of L_{2_σ}, and M is a closed subspace of the Hilbert space $L_{2_\sigma}^-$. Thus [5, p. 257],

$$L_{2_\sigma}^- = M \oplus M^\perp, \tag{34}$$

where $M^\perp$ denotes the orthogonal complement of M in $L_{2_\sigma}^-$. Hence, $\zeta \in C(x_t)$ implies $\zeta = \zeta_1 + \zeta_2$ *uniquely*, where $\zeta_1 \in M$, $\zeta_2 \in M^\perp$; and

$$x_{1t} = G_1^+(\zeta_1 + \zeta_2), \tag{35}$$

$$x_{2t} = G_2^+\zeta_2. \tag{36}$$

Now if ζ' is any other element of $C(x_t)$, then $\zeta_2' = \zeta_2$ since $G_2^+\zeta_2' = G_2^+\zeta_2$, implying $(\zeta_2' - \zeta_2) \in M \cap M^\perp = \{0\}$. Denoting this unique (over $C(x_t)$) second component by ξ, we have from (31),

$$\begin{aligned} L(x_t) &= \inf_{\substack{\theta \in M \\ G^+(\theta+\xi) = x_t}} \langle G(\theta + \xi), (\theta + \xi)\rangle_2 \\ &\geqq \inf_{\theta \in M} \langle G(\theta + \xi), (\theta + \xi)\rangle_2. \end{aligned} \tag{37}$$

Note that the term on the right-hand side is *not* the unconstrained minimization mentioned above. Now it follows from the Schwarz inequality and condition (b) of Theorem 1, that for all θ in M,

$$
\begin{aligned}
\langle G(\theta + \xi), (\theta + \xi)\rangle_2 &= \langle G\theta, \theta\rangle_2 + \langle G\theta, \xi\rangle_2 + \langle G\xi, \theta\rangle_2 + \langle G\xi, \xi\rangle_2 \\
&\geqq \delta\|\theta\|_2^2 - k_2\|\theta\|_2 + k_3 \qquad (38)\\
&\geqq k_3 - k_2^2/2\delta,
\end{aligned}
$$

where using (32),

$$
\begin{aligned}
k_2 &\triangleq \lambda\|\xi\|_2 + \|PG\xi\|_2 \qquad (39)\\
&\leqq (\lambda + k_1)\|\xi\|_\sigma
\end{aligned}
$$

and

$$
\begin{aligned}
k_3 &\triangleq \langle G\xi, \xi\rangle_2 \qquad (40)\\
&\leqq \|PG\xi\|_2\|\xi\|_2 \leqq k_1\|\xi\|_\sigma^2.
\end{aligned}
$$

Thus $L(x_t) > -\infty$ whenever $C(x_t) \neq \varnothing$.

We now prove that the Lyapunov function has the following important property.

LEMMA 3. *Let $x_0 \in Z$. Then $L(x_t)$ is monotonically nonincreasing along the ensuing trajectory.*

Proof. Let $0 < t_1 < t$ and $\hat{t} = t - t_1 > 0$. Define the set

$$D(x_t) = \{S_{t-t_1}\eta + S_t y_{1[t_1,t]} | \eta \in C(x_{t_1})\}. \qquad (41)$$

Functions in $D(x_t)$ drive the subsystem G to x_{t_1} at time $\tau = -\hat{t}$, and then along the system trajectory to x_t at time $\tau = 0$ (Fig. 3). Clearly, $D(x_t) \subset C(x_t)$. Hence

$$
\begin{aligned}
L(x_t) &= \inf_{\zeta \in C(x_t)} \langle G\zeta, \zeta\rangle \\
&\leqq \inf_{\mu \in D(x_t)} \langle G\mu, \mu\rangle \qquad (42)\\
&= \inf_{\eta \in C(x_{t_1})} \{\langle GS_{\hat{t}}\eta, \eta\rangle + \langle G(S_{\hat{t}}\eta + S_t y_{1[t_1,t]}), S_t y_{1[t_1,t]}\rangle\}.
\end{aligned}
$$

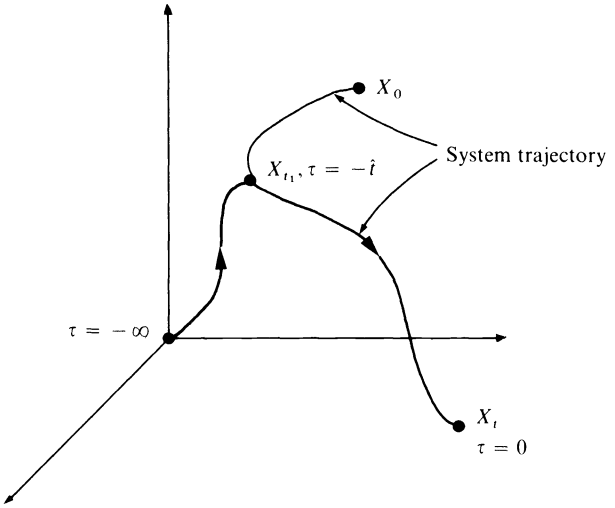

FIG. 3. *Path (heavy line) in state space in proof of Lemma* 4

Now for $t_1 \leqq \tau \leqq t$, using equation (23),

$$G(S_t\eta + y_{1[t_1,t]})(\tau) = -e(\tau). \tag{43}$$

Thus, from (42),

$$L(x_t) \leqq L(x_{t_1}) - \int_{t_1}^{t} e(\tau)y_1(\tau)\,d\tau. \tag{44}$$

Since N is passive, the second term on the right-hand side is nonpositive, and the lemma is proved.

LEMMA 4. *Let $x_0 \in Z$ and e solve* (9). *If $e \in L_2^+$, then $L(x_t) \geqq 0$ for all $t \geqq 0$.*

Proof. Consider the function

$$L_r(x_{2t}) = \inf_{\zeta \in C_r(x_{2t})} \langle G\zeta, \zeta\rangle_2, \tag{45}$$

where $C_r(x_{2t}) = \{\zeta \in L_{2_\sigma}^- | G_2^+\zeta = x_{2t}\}$. Since $C_r(x_{2t}) \supset C(x_t)$, $L_t(x_{2t}) \leqq L(x_t)$ for all $t \geqq 0$. Now clearly

$$C_r(x_{2t}) = \{\theta + \xi | \theta \in M\}, \tag{46}$$

where, recalling the proof of Lemma 2, ξ is the unique element in $M^\perp$ satisfying $G_2^+\xi = x_{2t}$. Thus

$$L_r(x_{2t}) = \inf_{\theta \in M} \langle G(\theta + \xi), (\theta + \xi)\rangle_2, \tag{47}$$

and from (38), $L_r(x_{2t})$ exists for all x_{2t} along the system trajectory.

Now if $x_{2t} = 0$, we have, since G is strictly passive on M,

$$\begin{aligned} L_r(0) &= \inf_{\theta \in M} \langle G\theta, \theta\rangle_2 \\ &\geqq \inf_{\theta \in M} \delta\|\theta\|_2^2. \end{aligned} \tag{48}$$

The infimum is reached with $\theta = 0$, and hence $L_r(0) = 0$. We now show L_r is continuous at the origin in its domain.

First consider the restriction of the map G_2^+ to $M^\perp$. This is a one-to-one map onto its finite-dimensional range. Since $M^\perp$ is thus also finite-dimensional, the inverse map is bounded and there exists a constant $k_4 > 0$ such that

$$\|\xi\|_\sigma \leqq k_4\|x_{2t}\|_\sigma. \tag{49}$$

We now show e belonging to L_2^+ implies $L_r(x_{2t}) \to 0$ as $t \to \infty$. Recalling from Lemma 1 (iii) that $e \in L_2^+$ implies $\|x_{2t}\|_\sigma \to 0$ as $t \to 0$, it is only necessary to show that $L_r(x_{2t}) \to 0$ when $\|x_{2t}\|_\sigma \to 0$.

Consider the case where for a given x_{2t} the infimum in (47) is reached for some θ, say θ^*, i.e.,

$$L_r(x_{2t}) = \langle G(\theta^* + \xi), \theta^* + \xi\rangle_2. \tag{50}$$

By the definition (45) and using (40),

$$L_r(x_{2t}) \leqq \langle G\xi, \xi\rangle_2 \leqq k_1\|\xi\|_\sigma^2, \tag{51}$$

where $k_1 = \|G_1\|_2 + \|G_2\|_\sigma$. On the other hand, using (38), (39) and (40),

$$\begin{aligned} L_r(x_{2t}) &\geqq k_3 - \frac{k_2^2}{2\delta} \geqq -k_1\|\xi\|^2 - \frac{(\lambda + k_1)^2}{2\delta}\|\xi\|_\sigma^2 \\ &\triangleq -k_5\|\xi\|_\sigma^2, \end{aligned} \tag{52}$$

which bounds $L_r(x_{2t})$ from below. Hence

$$|L_r(x_{2t})| \leqq k_6\|\xi\|_\sigma^2, \tag{53}$$

where $k_6 = \max\{k_1, k_5\}$. Then (49) establishes the continuity of $L_r(x_{2t})$ in the special case where the infimum in (48) is reached. The extension to the case where the infimum is not reached is treated by conventional techniques and may be found in [8]. Thus we conclude that $e \in L_2^+$ implies $L_r(x_{2t}) \to 0$ as $t \to \infty$.

Suppose now $e \in L_2^+$ and $L(x_{t_1}) = \mu < 0$ for some $t_1 > 0$. Since $L_r(x_{2t}) \to 0$ as $t \to \infty$, for $t_2 > t_1$ large enough,

$$\mu < L_r(x_{2t_2}); \tag{54}$$

and since $L(x_t)$ is nonincreasing,

$$L(x_{t_2}) \leqq L(x_{t_1}) = \mu. \tag{55}$$

Combining (54) and (55) gives

$$L(x_{t_2}) < L_r(x_{2t_2}), \tag{56}$$

which contradicts the fact that $L(x_t) \geqq L_r(x_{2t})$ for all $t \geqq 0$. Hence, $e \in L_2^+$ implies $L(x_t) \geqq 0$ for all $t \geqq 0$, and the lemma is proved.

4.3. Proof of Theorem 1. The presence of the right-half-plane poles of $G_2(s)$ can be shown to imply G is *not* passive over $L_{2_\sigma}^-$. We make use of this fact in proving the next lemma.

LEMMA 5. *There exists at least one initial state x_0 belonging to Z for which $L(x_0) < 0$.*

Proof. Consider the functions $v_1(t) = 1$ for $t \in [0, 2T]$ and zero elsewhere; and $v_2(t) = 1$ for $t \in [0, T]$, $v_2(t) = -1$ for $t \in [T, 2T]$, and zero elsewhere. Direct calculations show that there always exists a $T > 0$ and a choice of $u = v_1$ or $u = v_2$ such that $\langle Gu, u\rangle_2 < 0$. Then, setting $x_0 = G^+S_{2T}u$, x_0 belongs to Z, and

$$\begin{aligned} L(x_0) &= \inf_{\zeta \in C(x_t)} \langle G\zeta, \zeta\rangle_2 \\ &\leqq \langle GS_{2T}u, S_{2T}u\rangle_2 \\ &= \langle Gu, u\rangle_2 \\ &< 0. \end{aligned} \tag{57}$$

By selecting $x_0 \in Z$ as in Lemma 5, we have from the contraposition of Lemma 4 that $e \notin L_2^+$. Since it was assumed $e \in L_{2_e}^+$, it is concluded that $e \in L_{2_e}^+ \sim L_2^+$. The proof of Theorem 1 is complete.

5. Application: an instability counterpart to the circle criterion. We now use the results obtained above to develop an instability counterpart to the circle criterion.

5.1. System description. Consider now the nonlinear, time-varying feedback system $\tilde{S}$ governed by the equations

$$(I + \tilde{G}\tilde{N})\tilde{e} = \tilde{u} - \tilde{z}, \tag{58}$$

$$\tilde{y}_2 = \tilde{G}\tilde{y}_1 + \tilde{z}, \tag{59}$$

$$\tilde{y}_1 = \tilde{N}\tilde{e}. \tag{60}$$

We make the following assumptions on $\tilde{S}$.

Ã1. $\tilde{G}$ is given by

$$\tilde{G}\tilde{y}_1(t) = \int_0^t \tilde{g}(t - \tau)\tilde{y}_1(\tau)\,d\tau, \tag{61}$$

where $\tilde{g} \in L_1^+$. Standard theory shows that $\tilde{G}$ is a bounded nonanticipative, time-invariant operator on L_2 and L_{2_σ}, and, additionally, maps $L_{2_e}^+$ into $L_{2_e}^+$.

Ã2. $\tilde{N}$ is a memoryless, (possibly) time-varying nonlinearity given by

$$\tilde{N}\tilde{e}(t) = \tilde{n}(\tilde{e}(t), t) \quad \text{for all } t \geqq 0, \tag{62}$$

where $\tilde{n}: R \times [0, \infty) \to R$ satisfies

(i) $\tilde{n}(0, t) = 0$ for all $t \geqq 0$;

(ii) there exist real constants α and β such that for some arbitrarily small $\varepsilon > 0$,

$$0 < \alpha \leqq \tilde{n}(r, t)/r \leqq \beta - \varepsilon \tag{63}$$

for all real $r \neq 0$ and all $t \geqq 0$; and

(iii) $n(\tilde{e}(\cdot), \cdot)$ is a measurable function whenever $e(\cdot)$ is measurable.

Ã3. $\tilde{u}$, $\tilde{e}$, $\tilde{y}_1$, $\tilde{y}_2$, and $\tilde{z}$ are real-valued functions defined on the real line and equal to zero over $(-\infty, 0)$; furthermore,

(i) $\tilde{u}$ and $\tilde{z} \in L_2^+$, and

(ii) $\tilde{e}$, $\tilde{y}_1$, and $\tilde{y}_2 \in L_{2_e}^+$.

By analogy with § 3, we define the map $\tilde{G}^+ = (I - P)\tilde{G}$ which, by Ã1, is a bounded map from L_2^- into L_2^+. Also, the set of permissible zero-input responses, $\tilde{Z}$, is given by

$$\tilde{Z} = \{\tilde{z} \in L_2^+ \mid \tilde{z} = \tilde{G}^+\zeta, \text{for some } \zeta \in L_2^-\}. \tag{64}$$

$D[\alpha, \beta]$ denotes the closed disc in the complex plane with center at $(-(\alpha + \beta)/2\alpha\beta, 0)$ and radius $(\beta - \alpha)/2\alpha\beta$, and Γ denotes the Nyquist locus (the set $\{\tilde{G}(j\omega) | \omega \in R\} \cup \{0\}$).

THEOREM 2. *Under the above hypotheses if* $\tilde{\Gamma}$ *does not intersect* $D[\alpha, \beta]$ *and encircles it* $\rho > 0$ *times in the clockwise direction, then*

(i) *if* $\tilde{u} \equiv 0$, *there exists a* $\tilde{z} \in \tilde{Z}$ *such that if* $\tilde{e}$ *solves* (59), *then* $\tilde{e} \in L_{2_e} \sim L_2^+$;

(ii) *otherwise, for any* $\tilde{\zeta} \in \tilde{Z}$ *there exists a* $\tilde{u} \in L_2^+$ *such that if* $\tilde{e}$ *solves* (59), *then* $\tilde{e} \in L_{2_e}^+ \sim L_2^+$.

5.2. Definition of the transformed system S_1. We may assume throughout the analysis that $\tilde{u} \equiv 0$.

Since the point $(-1/\alpha, 0)$ belongs to the disc $D[\alpha, \beta]$, we conclude from the principle of the argument [7, p. 252] that $(1 + \alpha\tilde{G}(s))$ has ρ zeros in the open right-half-plane. Choose $\sigma > \operatorname{Re} s_r$, where s_r is the right-most zero of $(1 + \alpha\tilde{G}(s))$. Then $(I + \alpha G)$ has a time-invariant, nonanticipative, bounded inverse on L_{2_σ} [8, p. 53]. Let $\hat{\gamma} = (\beta - \alpha)/2$. The transformed system S_1 is defined in the form of Fig. 1 with

$$G = (2/\hat{\gamma})(I + \beta\tilde{G})(I + \alpha\tilde{G})^{-1}, \tag{65}$$

$$e = (1/\hat{\gamma})(\tilde{N} - \beta I)\tilde{e}, \tag{66}$$

$$y_1 = \tfrac{1}{2}(\tilde{N} - \alpha I)\tilde{e}, \tag{67}$$

$$z = 2(I + \alpha\tilde{G})^{-1}\tilde{z}. \tag{68}$$

It can be verified that $(I + GN)e = -z$ if and only if $(I + \tilde{G}\tilde{N})\tilde{e} = -\tilde{z}$, and the systems S_1 and $\tilde{S}$ are therefore equivalent.

By analogy with the state x_t, we define the *state of the subsystem* $\tilde{G}$ by

$$\tilde{x}_t = (I - P)S_t[\tilde{G}\tilde{y}_{1[0,t]} + \tilde{z}]. \tag{69}$$

Since G maps L_2 into L_2, $\tilde{x}_t$ belongs to L_2 for all $t \geqq 0$.

We now show that S_1 satisfies assumptions A1 through A3 of § 3. Let $\tilde{G}(I + \alpha\tilde{G})^{-1} = G_l + G_r$, where G_r is the map associated with the right-half-plane zeros of $(1 + \alpha\tilde{G}(s))$, and $G_l : L_{2_e}^{+} \to L_{2_e}^{+}$ and is a bounded map on L_2. Defining $G_1 = (2/\hat{\gamma})I + 4G_r$ and $G_2 = 4G_r$, $G = G_1 + G_2$ and G_1, G_2 satisfy A1. Also, N satisfies A2, where N is the relation defined by $Ne = y_1$. By equations (66) through (68), A3 is satisfied.

We now show that conditions (a) and (b) of Theorem 1 are satisfied. Condition (a) follows from the definition of N, assumption $\tilde{\text{A}}$2 (ii), and equations (66) and (67). To establish condition (b), we observe that by definition $\zeta \in M$ implies $G_2\zeta \in L_2$, and hence the limit-in-the-mean Fourier transform of $G_2\zeta$, denoted by $\mathscr{F}(G_2\zeta)$, exists. It is claimed that

$$\mathscr{F}(G_2\zeta)(j\omega) = G_2(j\omega)Z(j\omega) \tag{70}$$

for all $\omega \in R$. To show this we define the map G_2' by

$$G_2'\zeta(t) = \underset{\Delta \to \infty}{\text{l.i.m.}} \frac{1}{2\pi} \int_{-\Delta}^{\Delta} G_2(j\omega)Z(j\omega)\exp(j\omega t)\,d\omega \tag{71}$$

for all real t and for all $\zeta \in L_2$. Since $\max_{\omega \in R} |G_2(j\omega)| < \infty$, G_2' is a bounded map on L_2. We now show that $G_2 - G_2' = 0$ on M, and the conclusion then follows from Theorem 13J of [6].

We first define the subset M_D of M consisting of those functions in M whose Laplace transforms are rational with simple poles; then M_D is *dense* in M [10, p. 467]. For functions ζ in M_D, it can be verified by direct computation that $(G_2 - G_2')\zeta = 0$ [8, p. 61]. If ζ is an arbitrary element of M, there exists a sequence $\{\zeta_n\}_{n=1}^{\infty} \subset M_D$ such that $\zeta = \lim_{n\to\infty} \zeta_n$. Then using the continuity of G_2 and G_2',

$$\begin{aligned}(G_2 - G_2')\zeta &= (G_2 - G_2') \lim_{n\to\infty} \zeta_n \\ &= \lim_{n\to\infty} (G_2 - G_2')\zeta_n \qquad (72)\\ &= 0,\end{aligned}$$

and the claim is established.

Well-known manipulations [3, p. 607] show that there exists a $\delta > 0$ such that Re $G(j\omega) \geqq \delta$ for all real ω. Similarly there exists a $\lambda > 0$ such that $|G(j\omega)| \leqq \lambda$ for all real ω. Hence, using Parseval's theorem we now have that for all $\zeta \in M$,

$$\begin{aligned}\langle G\zeta, \zeta\rangle_2 &= \langle (G_1 + G_2)\zeta, \zeta\rangle_2 \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} (G_1(j\omega) + G_2(j\omega))|Z(j\omega)|^2\, d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \operatorname{Re} G(j\omega)|Z(j\omega)|^2\, d\omega \qquad (73)\\ &\geqq \frac{\delta}{2\pi}\int_{-\infty}^{\infty} |Z(j\omega)|^2\, d\omega \\ &= \delta\|\zeta\|_2^2,\end{aligned}$$

$$\begin{aligned}\|G\zeta\|_2^2 &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |G(j\omega)|^2|Z(j\omega)|^2\, d\omega \\ &\leqq \frac{\lambda^2}{2\pi}\int_{-\infty}^{\infty} |Z(j\omega)|^2\, d\omega \qquad (74)\\ &= \lambda^2\|\zeta\|_2^2.\end{aligned}$$

Hence condition (b) of Theorem 1 is satisfied.

Now suppose $x_0 \in Z$ and let $\zeta \in L_{2_\sigma}^-$ solve $G^+\zeta = x_0$. There exists an element $\tilde{x} \in L_{2_\sigma}^+$ such that $x_0 = 2(I + \alpha\tilde{G})^{-1}\tilde{x}_0$, and a simple calculation [8, p. 64] shows that

$$\begin{aligned}\tilde{x}_0 &= \tfrac{1}{2}(I + \alpha\tilde{G})G^+\zeta \\ &= \tilde{G}^+\tilde{\zeta},\end{aligned} \qquad (75)$$

where $\tilde{\zeta} = 2\zeta - 2\alpha P\tilde{G}(I + \alpha\tilde{G})^{-1}\zeta \in L_{2_\sigma}^- \subset L_2^-$. Hence $\tilde{x}_0 \in \tilde{Z}$.

Now let $\tilde{e}$ solve $(I + \tilde{G}\tilde{N})\tilde{e} = -\tilde{x}_0$. From Theorem 1 there exists an initial state $x_0 \in Z$ (and, therefore, an initial state $\tilde{x}_0 \in \tilde{Z}$) such that $e \in L_{2_e}^+ \sim L_2^+$. Since $e \in L_2^+$ if and only if $\tilde{e} \in L_2^+$, we have that there exists an $\tilde{x}_0 \in \tilde{Z}$ such that $\tilde{e} \in L_{2_e}^- \sim L_2^+$, and the proof of Theorem 2 is complete.

6. Concluding remarks.

1. An examination of the proof of Theorem 2 shows that this result also holds for the case $\alpha < \beta < 0$. Instability results were obtained in [3] for the case

$\alpha < 0 < \beta$ by requiring that the open-loop system be unstable; without using loop-shifting transformations, this is precluded here by the assumption that $\tilde{g}$ belongs to L_1^+.

2. It appears possible to construct other instability criteria using the methods developed here. An example is the instability counterpart to Popov's theorem [3] generalized to the infinite-dimensional case. A plausible procedure would be to transform the original system to one having the form considered in Theorem 1, and then conclude instability by using arguments analogous to those of § 5.

3. Under the conditions of Theorem 2, in addition to L_2 instability, it can also be concluded that $\|\tilde{x}_t\|_2 \to \infty$ as $t \to \infty$ [8, p. 67]. In the finite-dimensional case, this reduces to the state space instability of [3].

4. The Lyapunov function defined in § 4 may be used, at least in some restricted cases, to establish asymptotic stability in the large. Specifically, it is shown in [8] that for the case $g_2(t) = 0$ for all real t (and hence $\sigma = 0$) and $u \equiv 0$, the origin of the system S is asymptotically stable in the large whenever $\mathscr{R}(G^+|L_2^-)$, the range of G^+ restricted to L_2^-, is a closed subspace of L_2^+. This includes, for example, the case where the subsystem of G of the system S has a rational transfer function. It is noted that, in view of the relationship $x_t = 2(I + \alpha G)^{-1}\tilde{x}_t$, the stability of the origin of S implies the stability of the origin of $\tilde{S}$.

Acknowledgment. The authors would like to thank Dr. John J. Anton for some suggested improvements to the paper. They also highly appreciate the thorough and scholarly review given by the final reviewer.

REFERENCES

[1] R. A. Baker and A. R. Bergen, *Lyapunov stability and Lyapunov functions of infinite dimensional feedback systems*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 325–334.

[2] A. R. Bergen and S. Takeda, *On instability of feedback systems with a single nonlinear time-varying gain*, Ibid., AC-16 (1971), pp. 462–464.

[3] R. W. Brockett and H. B. Lee, *Frequency domain instability criteria for time-varying and nonlinear systems*, IEEE Proc., 55 (1967), pp. 604–619.

[4] C. A. Desoer, *Notes for a Second Course in Linear Systems*, Van Nostrand-Reinhold, New York, 1970.

[5] N. Dunford and J. T. Schwartz, *Linear Operators, I*, Interscience, New York, 1967.

[6] R. R. Goldberg, *Fourier Transform*, Cambridge University Press, New York, 1961.

[7] E. Hille, *Analytic Function Theory*, vol. 1, Ginn, Boston, 1959.

[8] T. L. Steding, *Instability of nonlinear infinite dimensional feedback systems using Lyapunov functions*, Doctoral dissertation, University of California, Berkeley, 1971.

[9] J. C. Willems, *Stability, instability, invertibility, and causality*, this Journal, 7 (1969), pp. 645–671.

[10] L. A. Zadeh and C. A. Desoer, *Linear System Theory*, McGraw-Hill, New York, 1963.

[11] G. Zames, *Input-output stability of time-varying, nonlinear feedback systems. Parts I and II*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 228–238.

SIAM J. Control
Vol. 12, No. 4, November 1974

A CLASS OF EXPONENTIAL PENALTY FUNCTIONS*

FREDERIC H. MURPHY†

Abstract. A class of penalty functions where the trial solutions may be interior or exterior to the feasible region of a nonlinear program is developed. Conditions under which the trial solutions become feasible are presented and a convergence rate is established. Also, trial values for the Lagrange multipliers where convergent subsequences of the trial multipliers converge to optimal Lagrange multipliers can be constructed from the trial solutions to the nonlinear program.

1. Consider the nonlinear programming problem (NLP)

$$\underset{x \in E^n}{\text{maximize}}\, f(x) \tag{1}$$

subject to

$$g_i(x) \leqq 0 \quad \text{for } i = 1, \cdots, m, \tag{2}$$

where f and g_i for $i = 1, 2, \cdots, m$ are real-valued functions defined on E^n.

Penalty function algorithms for solving NLP translate this difficult constrained problem into a sequence of easier maximizations, where with each iteration the penalty for infeasibility increases in the exterior algorithms [3], or the penalty for being near the boundary while feasible decreases in the interior algorithms [3]. The penalty functions are constructed so that all convergent subsequences of solutions to the easier problems converge to optimal solutions of NLP, either finitely or in the limit; and the value of the objective function either increases or decreases to the value of an optimal solution depending on the choice of penalty function.

We propose below a class of differentiable penalty functions in which with each iteration the advantage of being interior to the feasible region improves and the penalty for being exterior increases. The trial solutions, unlike with other penalty function algorithms, can be interior or exterior to the feasible region, and the value of the objective function is not necessarily monotonically increasing or decreasing at each iteration.

Separately and independently similar approaches were developed by Evans and Gould [2] and Allran and Johnsen [1]. Evans and Gould have made the most general statement of this new class of algorithms and therefore have the least detailed results.

Allran and Johnsen develop the most restrictive form of the penalty function, and their conditions for convergence must be clarified. The more general results herein reduce to statements of their results under appropriate restrictions.

Our class of functions is

$$F_k(x) = f(x) - \sum_{i=1}^{m} \frac{1}{s(k)} e^{r(k)g_i(x)}, \tag{3}$$

where $r(k) \geqq s(k) \geqq 1$ and $r(k) \to \infty$.

* Received by the editors March 29, 1972, and in revised form November 6, 1973.

† Graduate School of Management, Northwestern University, Evanston, Illinois 60201. This work was supported in part by the National Science Foundation under Grant GS-3032.

Members of this class of penalty functions are

$$f(x) - \sum_{i=1}^{m} e^{kg_i(x)}, \tag{4}$$

$$f(x) - \sum_{i=1}^{m} \frac{1}{a^k} e^{a^k g_i(x)}, \quad \text{where } a > 1, \tag{5}$$

$$f(x) - \sum_{i=1}^{m} \frac{1}{a^k} e^{b^k g_i(x)}, \quad \text{where } 1 < a < b. \tag{6}$$

The class of functions treated by Allran and Johnsen [1] is (3) with $s(k) = 1$, and (3) is a specific example of the class developed by Evans and Gould [2].

A disadvantage with interior penalty functions is that it can take as long to find an interior starting solution as it takes to solve the nonlinear program [4, p. 213]. Because the trial solution to an exponential penalty function can be feasible or infeasible in NLP, we need not look for an initial feasible solution. Also, by an appropriate choice of parameters, in the convex case we can guarantee that all trial solutions are feasible in NLP.

Exponential penalty functions have the property that they are uniformly bounded over any compact feasible region, and again by an appropriate choice of parameters the gradient remains uniformly bounded over the feasible region. In the limit as $r(k) \to \infty$, the penalty function (3) has the value $f(x)$ if x well satisfies all the constraints. If $s(k) \to \infty$, this is true even for the points at the boundary of the feasible region. As with other penalty functions, it can be shown that if $f(x)$ is concave and the constraints are convex, (3) forms a concave function, ensuring that a local maximum is a global maximum. A convergence rate, trial values for the Lagrange multipliers and upper and lower bounds on the value of an optimal solution are provided.

2. Convergence results. Let

$$S = \{x | g_i(x) \leqq 0 \text{ for } i = 1, \cdots, m\}, \tag{7}$$

$$T = \{x | g_i(x) < 0 \text{ for } i = 1, \cdots, m\}, \tag{8}$$

$$x^k \in X \text{ maximize } F_k(x) \text{ over } X \text{ for } k = 1, 2, \cdots, \tag{9}$$

where X is a compact set in E^n containing the feasible region S and x^* is an optimal solution to NLP. We use $|y|$ to represent the Euclidean norm of y for y of any dimension, and $\bar{T}$ to represent the closure of the set T.

THEOREM 1. *Assume that*

(a) *S is a nonempty compact set,*

(b) $\bar{T} = S$,

(c) *the functions $f(x), g_1(x), \cdots, g_m(x)$ are continuous.*

(d) $r(k) \geqq s(k) \geqq 1$.

Then any convergent subsequence of x^k converges to an optimal solution of NLP.

Instead of restricting x^k to be in X, Evans and Gould [2] provide a growth rate condition on the objective function and constraints to guarantee the existence of x^k and ensure that x^k is in some compact set after a finite number of iterations.

This restriction of x^k is motivated by the fact that nonlinear programming algorithms are designed to be used on a digital computer, which automatically restricts x^k to a compact set. Condition (c) along with (9) guarantees the existence of x^k since we are maximizing a continuous function (3) over a compact set. In the convex case, x^k is automatically contained in a compact set without any formal restriction.

Note that condition (b) implies $T \neq \varnothing$. Allran and Johnsen [1] use the weaker assumption that the interior of S is nonempty. This assumption is not sufficient for Theorem 1 to be valid in the nonconvex case. To see the need for condition (b), consider the example:

$$\text{(10)} \qquad \underset{x \in E^1}{\text{maximize}}\ \tfrac{1}{2}x$$

subject to

$$\text{(11)} \qquad g(x) \leqq 0,$$

where

$$\text{(12)} \qquad g(x) = \begin{cases} 4|x| - 1 & \text{for } |x| \leqq \tfrac{1}{4}, \\ 0 & \text{for } \tfrac{1}{4} < |x| \leqq 1, \\ |x| - 1 & \text{elsewhere.} \end{cases}$$

Letting $s(k) = 1$ and $r(k) = k$, we can determine $F_k(x)$:

$$\text{(13)} \qquad F_k(x) = \begin{cases} \tfrac{1}{2}x - e^{k(4|x|-1)} & \text{for } |x| \leqq \tfrac{1}{4}, \\ \tfrac{1}{2}x - 1 & \text{for } \tfrac{1}{4} < |x| \leqq 1, \\ \tfrac{1}{2}x - e^{k(|x|-1)} & \text{elsewhere.} \end{cases}$$

Hence,

$$\text{(14)} \qquad F_\infty(x) = \begin{cases} \tfrac{1}{2}x & \text{for } |x| < \tfrac{1}{4}, \\ \tfrac{1}{2}x - 1 & \text{for } \tfrac{1}{4} \leqq |x| \leqq 1, \\ -\infty & \text{elsewhere.} \end{cases}$$

In this case with $F_\infty(x)$ the discontinuous limit of a sequence of continuous functions, the maximum of $F_\infty(x)$ does not exist. The supremum is $\tfrac{1}{8}$, the limit of any sequence x_h, $h = 1, 2, \cdots$, with $x_h \uparrow \tfrac{1}{4}$, whereas the maximum of the original problem (10) and (11) is $\tfrac{1}{2}$ at $x = 1$. If $s(k) = k$, then

$$\text{(15)} \qquad F_k(x) = \begin{cases} \tfrac{1}{2}x - \dfrac{1}{k}e^{k(4|x|-1)} & \text{for } |x| \leqq \tfrac{1}{4}, \\ \tfrac{1}{2}x - \dfrac{1}{k} & \text{for } \tfrac{1}{4} < |x| \leqq 1, \\ \tfrac{1}{2}x - \dfrac{1}{k}e^{k(|x|-1)} & \text{elsewhere,} \end{cases}$$

with the result that

$$F_\infty(x) = \begin{cases} \frac{1}{2}x & \text{for } -1 \leqq x \leqq 1, \\ -\infty & \text{elsewhere.} \end{cases} \tag{16}$$

From this we may infer that condition (b) is unnecessary if $s(k) \uparrow \infty$, which will be stated in Theorem 2 below.

LEMMA 1. *Under conditions* (a), (c) *and* (d), *there exists a convergent subsequence of* x^k; *and for any such subsequence indexed, say, by* k_u, *we have* $x^{k_u} \to \bar{x} \in S$.

Proof. Assume $\bar{x} \notin S$. This means that for some constraint $g_h(x)$ and for some k_u sufficiently large, we have

$$g_h(x^{k_u}) \geqq \delta > 0. \tag{17}$$

This implies $F_{k_u}(x^{k_u}) \to -\infty$ as $k_u \to \infty$ since $(1/s(k_u))\, e^{r(k_u)\delta} \to +\infty$ when $r(k) \geqq s(k)$ (use L'Hospital's rule on $x^{-1}e^x$ as $x \to \infty$ to see this). However,

$$F_{k_u}(x) \geqq f(x) - m \tag{18}$$

for x feasible in NLP since e raised to a negative number is less than one and $s(k) \geqq 1$. That is, we have a uniform lower bound on the maximum of $F_{k_u}(x)$ and a contradiction. Therefore, $\bar{x} \in S$.

Proof of Theorem 1. For any $x_0 \in T$ we have

$$f(x^{k_u}) \geqq F_{k_u}(x^{k_u}) \geqq F_{k_u}(x_0). \tag{19}$$

Because $F_{k_u}(x_0) \to f(x_0)$, and $f(x^{k_u}) \to f(\bar{x})$ by the continuity of $f(x)$, taking limits in (19) we have

$$f(\bar{x}) \geqq f(x_0). \tag{20}$$

Since $\bar{T} = S$, we can choose x_0 arbitrarily close to x^*, and

$$f(\bar{x}) \geqq f(x^*). \tag{21}$$

By Lemma 1, $\bar{x}$ is feasible; therefore, $\bar{x}$ is optimal in NLP.

If we require $s(k) \uparrow \infty$, we may drop assumption (b) in Theorem 1; that is, we no longer require that $\bar{T} = S$.

THEOREM 2. *If conditions* (a), (c) *and* (d) *hold, and if* $s(k) \uparrow \infty$, *then any convergent subsequence of* x^k *converges to an optimal solution of NLP.*

Proof. Let $x^{k_u} \to \bar{x}$. Then by Lemma 1, $\bar{x}$ is feasible. Since $s(k) \to \infty$, $F_k(x^*) \to f(x^*)$ as $k \to \infty$. Now

$$f(x^{k_u}) \geqq F_{k_u}(x^{k_u}) \geqq F_{k_u}(x^*). \tag{22}$$

By taking limits as $k_u \to \infty$ and noting that $f(x)$ is continuous we have

$$f(\bar{x}) \geqq f(x^*). \tag{23}$$

That is, (23) is an equality and $\bar{x}$ is optimal in NLP.

By noting (19) and (22) we have the following.

COROLLARY 1. *If the conditions of either Theorem* 1 *or Theorem* 2 *are satisfied, then*

$$F_k(x^k) \to f(x^*) \quad \textit{as } k \to \infty. \tag{24}$$

All exterior algorithms have the disadvantage of producing trial solutions that are infeasible at each iteration and are feasible only in the limit. Algorithms like the cutting-plane methods of Kelley [5] and Veinott [8] have this difficulty as well as does the differentiable exterior penalty function discussed in [9].

Allran and Johnsen [1] show that with their penalty function, after a finite number of iterations there exists an $x \in T$ that is a local maximum of $F_k(x)$. This is not true for the more general function (3). For example, let $f(x) = x$, $g_1(x) = (x - 1)^3$ and $g_2(x) = -x$ in NLP, $x \in E^1$. Using (5),

$$\nabla F_k(x) = 1 - e^{a^k(x-1)^3}3(x - 1)^2 + e^{-a^k x} \tag{25}$$

For k large, $\nabla F_k(x) \approx 1$, or greater than one, for $x \in S = [0, 1]$, which means any local maximum of $F_k(x)$ is infeasible. Also, it is not true that any convergent subsequence of local maxima of $F_k(x)$ converges to a feasible point of NLP for all possible choices of the $g_i(x)$, even with the penalty function of Allran and Johnsen [1]. However, under certain conditions all of the trial solutions will be feasible after a finite number of iterations. It is *not* necessarily true, however, that once a trial point is feasible, all the subsequent x^k are feasible. The trial solutions may be feasible and then infeasible a *finite* number of times.

THEOREM 3. *If $f(x)$ is concave, $g_1(x), \cdots, g_m(x)$ are convex on E^n, T is nonempty, S is compact, $r(k)/s(k) \to \infty$, and $\nabla f(x)$, $\nabla g_1(x), \cdots, \nabla g_m(x)$ are continuous, then there exists an integer K' such that every penalty function maximizer over E^n is feasible for all $k \geqq K'$.*

Proof. We show that for a fixed point $x_0 \in T$ and any x_B on the boundary of the feasible region, the directional derivative of $F_k(x)$ at x_B in the direction $(x_B - x_0)$ is negative for k sufficiently large. Although the value of k is dependent on the choice of x_B, the continuity of gradients allows us to find a K when the directional derivative is negative for all x_B for $k \geqq K$. Hence $F_k(x)$ is decreasing as x is translated from x_B out of the feasible region in the direction $(x_B - x_0)$. Now, for any $x \notin S$, there is an x_B on the line connecting x_0 and x. Because the directional derivative at x_B in the direction $(x_B - x_0)$, which is the same as $(x - x_B)$, is negative, $F_k(x_B) > F_k(x)$ for k sufficiently large by the concavity of $F_k(x)$. Thus, there is a boundary point x_B with $F_k(x_B) > F_k(x)$ corresponding to each infeasible x for k sufficiently large, which means x^k is feasible for k sufficiently large.

Note that

$$\nabla F_k(x) = \nabla f(x) - \sum_{i=1}^{m} e^{r(k)g_i(x)}\frac{r(k)}{s(k)}\nabla g_i(x). \tag{26}$$

Let x_0 be a point in T and x_B be a boundary point of S, that is, for at least one $h \in \{1, \cdots, m\}$, $g_h(x_B) = 0$. Let

$$2\delta = \max\{g_i(x_0) : i = 1, \cdots, m\}. \tag{27}$$

Since each $g_i(x)$ is convex,

$$0 > 2\delta \geqq g_i(x_0) \geqq g_i(x) + \nabla g_i(x)(x_0 - x). \tag{28}$$

If $g_i(x) > \delta$, we have

$$0 > \delta > \nabla g_i(x)(x_0 - x). \tag{29}$$

At iteration k,

$$\begin{aligned}\nabla F_k(x_B) = \nabla f(x_B) &- \sum_{\{i:g_i(x_B)>\delta\}} e^{r(k)g_i(x_B)}\frac{r(k)}{s(k)}\nabla g_i(x_B)\\ &- \sum_{\{i:g_i(x_B)\leqq\delta\}} e^{r(k)g_i(x_B)}\frac{r(k)}{s(k)}\nabla g_i(x_B),\end{aligned} \tag{30}$$

$$\begin{aligned}\nabla F_k(x_B)\cdot(x_0 - x_B) &= \nabla f(x_B)\cdot(x_0 - x_B)\\ &- \sum_{\{i:g_i(x_B)>\delta\}} e^{r(k)g_i(x_B)}\frac{r(k)}{s(k)}\nabla g_i(x_B)\cdot(x_0 - x_B)\\ &- \sum_{\{i:g_i(x_B)\leqq\delta\}} e^{r(k)g_i(x_B)}\frac{r(k)}{s(k)}\nabla g_i(x_B)\cdot(x_0 - x_B).\end{aligned} \tag{31}$$

Observe that for h such that $g_h(x) \leqq \delta$, $x \in S$,

$$e^{r(k)g_h(x)}\frac{r(k)}{s(k)}|\nabla g_h(x)\cdot(x_0 - x)| \leqq e^{r(k)\delta}\frac{r(k)}{s(k)}|\nabla g_h(x)\cdot(x_0 - x)| \to 0 \tag{32}$$

as $k \to \infty$, since $e^{r(k)\delta}r(k) \to 0$ as $k \to \infty$ and $s(k) \geqq 1$. Note that (32) is uniform for $x \in S$ with $g_h(x) \leqq \delta$ since $\nabla g_h(x)\cdot(x_0 - x)$ is uniformly bounded for $x \in S$ by our assumption on the continuity of the gradients of $g_1(x), \cdots, g_m(x)$.

Therefore, there is a K such that, for $k \geqq K$ and $x \in S$ with $g_i(x) \leqq \delta$,

$$|e^{r(k)g_i(x)}\frac{r(k)}{s(k)}\nabla g_i(x)\cdot(x_0 - x)| < \varepsilon. \tag{33}$$

For $x \in S$ with $0 > g_h(x) > \delta$ for some $h \in \{1, \cdots, m\}$ by (29),

$$\frac{r(k)}{s(k)}e^{r(k)g_h(x)}\nabla g_h(x)\cdot(x_0 - x) \leqq \frac{r(k)}{s(k)}e^{r(k)g_h(x)}\delta < 0. \tag{34}$$

For $x \in S$ with $g_h(x) = 0$ for some $h \in \{1, \cdots, m\}$ by (29),

$$\frac{r(k)}{s(k)}\nabla g_h(x)\cdot(x_0 - x) \leqq \frac{r(k)}{s(k)}\delta \downarrow -\infty. \tag{35}$$

Note that (35) like (33) is uniform for $x \in S$.

By our assumptions that $\nabla f(x)$ is continuous and S is compact, we know that there is an $M > 0$ with

$$|\nabla f(x)\cdot(x_0 - x)| \leqq M. \tag{36}$$

Using (29), (33) and (34) for $k \geqq K$ with $x = x_B$,

$$\begin{aligned}\nabla F_k(x_B)\cdot(x_0 - x_B) &\geqq -M - \frac{r(k)}{s(k)} \sum_{\{i:g_i(x_B)=0\}} \nabla g_i(x_B)\cdot(x_0 - x_B) - \sum_{\{i:g_i(x_B)\leqq\delta\}} \varepsilon \\ &\geqq -M - \frac{r(k)}{s(k)} \sum_{\{i:g_i(x_B)=0\}} g_i(x_0) - m\varepsilon \qquad (37) \\ &\geqq -M - \frac{r(k)}{s(k)}\delta - m\varepsilon.\end{aligned}$$

Consequently, by (35) there is a K' such that $\nabla F_k(x_B)\cdot(x_0 - x_B) > 0$ for all x_B on the boundary of S for $k \geqq K'$.

Since the directional derivative of $F_k(x)$ at x_B in the direction $(x_0 - x_B)$ (towards x_0) is positive, it is negative in the direction $(x_B - x_0)$ (from x_B away from the feasible region). Therefore, $F_k(x) < F_k(x_B)$ for all $x - x_B + q(x_0 - x_B)$, $q < 0$, because $F_k(x)$ is concave. We know that for each $x \notin S$, there is a corresponding boundary point x_B such that

$$x = x_B + q(x_0 - x_B) \quad \text{with } q < 0. \tag{38}$$

Therefore, for each $x \notin S$, there is an associated point $x_B \in S$, where

$$F_k(x) < F_k(x_B) \quad \text{for } k \geqq K'. \tag{39}$$

Since $F_k(x)$ is continuous and S is compact, the maximum of $F_k(x)$ over S exists and by (39) is the global maximum of $F_k(x)$ over E^n. That is, x^k exists and is in S for $k \geqq K'$.

We next establish convergence rates for our class of penalty functions $F_k(x)$. Since we have shown that the trial solutions are feasible after a finite number of iterations, we need only establish rates of convergence for feasible trial solutions.

THEOREM 4. *If $x^k \in S$, then*

$$\frac{m}{s(k)} \geqq f(x^*) - f(x^k) \geqq 0. \tag{40}$$

Proof. Since $x^k \in S, f(x^*) - f(x^k) \geqq 0$. Also, $g_i(x^*) \leqq 0$; that is, $e^{r(k)g_i(x^*)} \leqq 1$, and

$$\begin{aligned} f(x^*) - \frac{m}{s(k)} &\leqq F_k(x^*) \\ &\leqq F_k(x^k) \qquad (41) \\ &\leqq f(x^k), \end{aligned}$$

or

$$0 \leqq f(x^*) - f(x^k) \leqq \frac{m}{s(k)}. \tag{42}$$

We see that from the same set of inequalities (41), we can express the convergence rate in terms of the value of the penalty function at the trial solution. That is, from (41),

$$\frac{m}{s(k)} \geqq f(x^*) - F_k(x^k) \geqq 0. \tag{43}$$

Note that the upper bound of (42) and (43) does not depend on x^k being feasible.

Under the assumption of uniform concavity of $f(x)$, we may establish the rate of convergence of x^k to an optimal solution.

DEFINITION 1. A real-valued function $f(x)$ is uniformly concave [6] on a convex set T if there exists a nondecreasing function $\delta(v) > 0$ on $(0, \infty)$ such that for $x, y \in T$,

$$f(\tfrac{1}{2}(x + y)) \geqq \tfrac{1}{2}f(x) + \tfrac{1}{2}f(y) + \delta(|x - y|), \tag{44}$$

where $|x - y|$ is the Euclidean norm.

An example of a uniformly concave function is any strictly concave function over T with T compact [6]. Since uniformly concave is stronger than strictly concave, x^* is the unique solution to NLP and $x^k \to x^*$.

THEOREM 5. *If $f(x)$ is uniformly concave on S and $\delta(v)$ is strictly increasing in v, then for $x^k \in S$,*

$$|x^k - x^*| \leqq \delta^{-1}\left(\frac{m}{2s(k)}\right). \tag{45}$$

Proof. Since x^k is feasible, by Theorem 4,

$$|f(x^*) - f(x^k)| \leqq \frac{m}{s(k)}. \tag{46}$$

And since x^* is optimal and $\tfrac{1}{2}(x^* + x^k)$ is feasible,

$$\begin{aligned} f(x^*) &\geqq f(\tfrac{1}{2}(x^* + x^k)) \\ &\geqq \tfrac{1}{2}f(x^*) + \tfrac{1}{2}f(x^k) + \delta(|x^* - x^k|) \end{aligned} \tag{47}$$

by uniform concavity. Thus using (47),

$$\tfrac{1}{2}f(x^*) - \tfrac{1}{2}f(x^k) \geqq \delta(|x^* - x^k|). \tag{48}$$

Using (46),

$$\frac{1}{2}\,\frac{m}{s(k)} \geqq \delta(|x^* - x^k|), \tag{49}$$

which means

$$\delta^{-1}\left(\frac{m}{2s(k)}\right) \geqq |x^* - x^k|. \tag{50}$$

As an example, if $\delta(r) = r^2$, then

$$\sqrt{\frac{m}{2s(k)}} \geqq |x^* - x^k|. \tag{51}$$

As with the penalty functions, we can generate trial Lagrange multipliers where the limit of any convergent subsequence is an optimal set of Lagrange multipliers. Here setting

$$u_i^k = \frac{r(k)}{s(k)} e^{r(k)g_i(x^k)} \quad \text{for } i = 1, \cdots, m, \tag{52}$$

we have trial Lagrange multipliers. With $s(k) = 1$, these are the trial multipliers of Allran and Johnsen [1]. The proof that convergent subsequences of u_i^k converge to the Lagrange multipliers of NLP is omitted as it is routine. It can be found in [7].

Acknowledgment. I wish to thank Professor Robert Mifflin of Yale University for his helpful comments.

REFERENCES

[1] R. R. Allran and S. E. V. Johnsen, *An algorithm for solving nonlinear programming problems subject to nonlinear inequality constraints*, Comput. J., 13 (1970), pp. 171–177.

[2] V. R. Evans and F. J. Gould, *Stability and exponential penalty function techniques in nonlinear programming*, Institute of Statistics Mimeo Series no. 723, University of North Carolina, Chapel Hill, 1970.

[3] A. Fiacco and G. McCormick, *Nonlinear Programming, Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[4] R. Fletcher, ed., *Optimization*, Academic Press, London, 1969.

[5] J. E. Kelley, Jr., *The cutting-plane method for solving convex programs*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 703–712.

[6] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), pp. 1–50.

[7] F. H. Murphy, *Topics in nonlinear programming: Penalty function and column generation algorithms*, Ph.D. thesis, Yale University, New Haven, Conn., 1971.

[8] A. F. Veinott, Jr., *The supporting hyperplane method for unimodal programming*, Operations Res., 15 (1967), pp. 147–152.

[9] W. Zangwill, *Nonlinear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

GENERIC SOLVABILITY OF THE DECOUPLING PROBLEM*

E. FABIAN AND W. M. WONHAM†

Abstract. For the linear multivariable system $\dot{x} = Ax + Bu$, $z_i = D_i x$ ($i \in \mathbf{k}$), decoupled (noninteracting) control is shown to be achievable for "almost all" sets of real matrices $\{A\,(n \times n), B\,(n \times m); D_i\,(q_i \times n), i \in \mathbf{k}\}$, having the dimensions shown, if and only if

$$\sum_{i=1}^{k} q_i \leqq \min(n, m - 1 + \min_{1 \leqq i \leqq k} q_i).$$

With $(n, m, q_1, \cdots, q_k)$ subject to the inequality, the exceptional matrix sets for which decoupling is impossible belong to a proper algebraic variety in $\mathbb{R}^N$, $N = n^2 + nm + (q_1 + \cdots + q_k)n$.

1. Introduction. Given any formally stated problem of control law synthesis, it is important to know whether the problem is solvable for "most" or "typical" sets of parameter values which are likely to be met in practice, or solvable only for special or highly restricted parameter sets, the occurrence of which is rare. In the former case the problem may be called "generically solvable", or simply "generic". In this paper we formalize the property of genericity for a wide class of synthesis problems of linear multivariable control, and obtain necessary and sufficient conditions for genericity of the extended decoupling problem in the version studied in [1], [2]. The main result is stated in § 3.

2. Genericity and well-posedness. Let $A, B, \cdots$ be matrices with elements in $\mathbb{R}$ and suppose $\Pi(A, B, \cdots)$ is some property which may be asserted about them. Most properties of interest to us will turn out to hold true for all points (parameter sets) except some which lie on an algebraic hypersurface, and which are thus, in an intuitive sense, atypical. For the sake of precision we borrow some terminology from algebraic geometry (cf. [3]). Let $\mathbf{p} = (p_1, \cdots, p_N) \in \mathbb{R}^N$ and consider polynomials $\psi \in \mathbb{R}[\lambda_1, \cdots, \lambda_N]$, where the λ_i are indeterminates. A *variety* $\mathbf{V} \subset \mathbb{R}^N$ is the set of common zeros of a finite number of polynomials $\psi_1, \cdots, \psi_k$:

$$\mathbf{V} = \{\mathbf{p} : \psi_i(p_1, \cdots, p_N) = 0, i \in \mathbf{k}\}.$$

$\mathbf{V}$ is *proper* if $\mathbf{V} \neq \mathbb{R}^N$ and *nontrivial* if $\mathbf{V} \neq \emptyset$. A point $\mathbf{p} \in \mathbb{R}^N$ is *generic relative to* $\mathbf{V}$ if $\mathbf{p} \notin \mathbf{V}$. A *property* Π is a function from $\mathbb{R}^N$ to $\{0, 1\}$, where $\Pi(\mathbf{p}) = 0$ (or 1) means Π fails (or holds) at $\mathbf{p}$. Let $\mathbf{V}$ be a proper variety. Π is *generic relative to* $\mathbf{V}$ if

$$\ker \Pi \equiv \{\mathbf{p} : \Pi(\mathbf{p}) = 0\} \subset \mathbf{V};$$

and Π is *generic* if such a $\mathbf{V}$ exists.

Assign to $\mathbb{R}^N$ the usual topology. If $\mathbf{V}$ is a variety, it is clear from the continuity of its defining polynomials that $\mathbf{V}$ is closed. Thus if Π is generic relative to $\mathbf{V}$ and if $\mathbf{p} \in \mathbf{V}^c$, Π holds throughout a sufficiently small neighborhood of $\mathbf{p}$: in this sense Π is *well-posed* at points $\mathbf{p} \in \mathbf{V}^c$.

* Received by the editors February 1, 1973.

† Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada M5S 1A4. This research was supported in part by the National Research Council of Canada under Grant A-7399.

Let $\mathbf{p}_0 \in \mathbf{V}$, with $\mathbf{V}$ nontrivial and proper. It is clear that every (open) neighborhood of $\mathbf{p}_0$ contains points $\mathbf{p} \notin \mathbf{V}$: otherwise, each defining polynomial of $\mathbf{V}$ vanishes identically in some neighborhood of $\mathbf{p}_0$, hence vanishes on $\mathbb{R}^N$, in contradiction to the assumption that $\mathbf{V}$ is proper. Thus if Π is generic relative to $\mathbf{V}$ and if Π fails at $\mathbf{p}_0$, Π can be made to hold if $\mathbf{p}_0$ is subject to a suitable perturbation, arbitrarily small. We conclude that the set of points $\mathbf{p}$ where a generic property is well-posed is both open and dense in $\mathbb{R}^N$. However, from a practical viewpoint not too much can be made of this fact alone, as there exist in $\mathbb{R}^N$ open dense subsets with (positive) Lebesgue measure arbitrarily small.

As a simple example of genericity, let A $(n \times n)$, B $(n \times m)$ be real matrices, write $(A, B) = \mathbf{p} \in \mathbb{R}^N$ $(N = n^2 + nm)$, and let $\Pi(\mathbf{p}) = 1$ if and only if $\mathbf{p}$ (i.e., the pair (A, B)) is controllable. It is easy to see that Π is generic. Indeed, (A, B) is controllable if and only if the $n \times nm$ matrix

$$X = [B, AB, \cdots, A^{n-1}B]$$

has rank n. Write x_i, $i = 1, \cdots, nm$, for the columns of X. Controllability fails if and only if every determinant formed by selecting n columns x_i vanishes: that is, $\psi(\mathbf{p}) = 0$, where

$$\psi(\mathbf{p}) \equiv \sum_i (\det [x_{i_1} x_{i_2} \cdots x_{i_n}])^2,$$

and i ranges over all sets of indices $(i_1, \cdots, i_n)$ with $1 \leqq i_1 < i_2 < \cdots < i_n \leqq nm$. Let $\mathbf{V}$ be the locus of zeros of ψ. Clearly $\mathbf{V}$ is a variety in $\mathbb{R}^N$ and, since controllable pairs exist, $\mathbf{V}$ is proper. Of course, the fact that the controllable pairs are open and dense in parameter space is well known [4, p. 100], but the proof of denseness here is a little simpler than the one just cited.

Finally, we remark that if Π is generic, the set

$$\mathbf{V}_\pi = \bigcap_{\mathbf{V} \supset \ker \Pi} \mathbf{V}$$

is again an algebraic variety [3, Chap. 1], and Π is well-posed at every point in $\mathbf{V}_\pi^c$. In general, $\ker \Pi$ is a *proper* subset of $\mathbf{V}_\pi$, that is, there exist points $\mathbf{p}$ (in $\mathbf{V}_\pi$) where Π holds and may or may not be well-posed.

3. Extended decoupling problem.

Consider the system

$$\dot{x} = Ax + Bu, \tag{1}$$

$$z_i = D_i x, \qquad i \in \mathbf{k}. \tag{2}$$

Here A, B and the D_i are real matrices of dimension respectively $n \times n$, $n \times m$ and $q_i \times n$ $(i \in \mathbf{k})$, with $n \geqq 1$, $1 \leqq m \leqq n$, and $1 \leqq q_i \leqq n$ $(i \in \mathbf{k})$. In the geometric decoupling theory [1], [2], one considers the corresponding linear transformations $A: \mathscr{X} \to \mathscr{X}$, $B: \mathscr{U} \to \mathscr{X}$, $D_i: \mathscr{X} \to \mathscr{Z}_i$ on $\mathbb{R}$-vector spaces $\mathscr{X}, \mathscr{U}, \mathscr{Z}_i$ with dimension $d(\mathscr{X}) = n$, $d(\mathscr{U}) = m$, $d(\mathscr{Z}_i) = q_i$. It is known that the extended decoupling problem (EDP) is solvable, possibly by use of dynamic compensation, if and only if

$$\mathscr{R}_i^* + \ker D_i = \mathscr{X}, \qquad i \in \mathbf{k}, \tag{3}$$

where $\mathscr{R}_i^*$ is the largest controllability subspace contained in the subspace

$$\mathscr{K}_i = \bigcap_{j \neq i} \ker D_j. \tag{4}$$

Solvability of EDP is thus a property Π of the parameter set $(A, B, D_1, \cdots, D_k)$, regarded as a point $\mathbf{p} \in \mathbb{R}^N$, with $N = n^2 + nm + (q_1 + \cdots + q_k)n$. Our main result is the following.

THEOREM. *Π is generic if and only if*

$$\sum_{i=1}^{k} q_i \leqq n \tag{5}$$

and

$$m \geqq 1 + \sum_{i=1}^{k} q_i - \min_{1 \leqq i \leqq k} q_i. \tag{6}$$

Roughly speaking, the theorem means that EDP is generically solvable if and only if the row spaces of the D_i are (generically) independent, and the number of (generically independent) control vectors (i.e., columns of B) is sufficiently large. Since $q_i \geqq 1$ we always need $m \geqq k$; and if, for instance, $m = k$, we must have $q_i = 1$ for all i.

The following notation will be used. A prime denotes matrix transpose, dual linear transformation, or dual space. If n, m are integers,

$$n \vee m = \max(n, m), \qquad n \wedge m = \min(n, m).$$

If for each $\mathbf{p} \in \mathbb{R}^N$, $\mathscr{R}(\mathbf{p}) \subset \mathscr{X}$ is a linear subspace, we write

$$d(\mathscr{R}) = r\,(\mathrm{g})$$

to mean that the generic dimension of $\mathscr{R}$ is r, i.e., that $d(\mathscr{R}(\mathbf{p})) \neq r$ only for $\mathbf{p}$ in some fixed proper variety $\mathbf{V} \subset \mathbb{R}^N$ depending on the function $\mathscr{R}(\cdot)$. Subspace inclusions written $\mathscr{R} \subset \mathscr{S}$ (g) are to be interpreted in the same fashion. We observe that a finite union of proper varieties is a proper variety; hence if a finite set of propositions each holds (g), the entire set holds simultaneously (g).

Proof of theorem.

(i) *Preliminaries.* It is clear that

$$d(\ker D_i) = n - q_i\,(\mathrm{g}), \qquad i \in \mathbf{k}, \tag{7}$$

and

$$d(\operatorname{Im} D_i') = q_i\,(\mathrm{g}), \qquad i \in \mathbf{k}, \tag{8}$$

since the dimensional evaluations fail at $\mathbf{p}$ only if all $q_i \times q_i$ minors of the $q_i \times n$ matrix D_i vanish. Similarly,

$$d\left(\sum_{j \neq i} \operatorname{Im} D_j'\right) = n \wedge \sum_{j \neq i} q_j\,(\mathrm{g}), \qquad i \in \mathbf{k}, \tag{9}$$

and

$$\begin{aligned} d(\mathscr{K}_i) &= n - d(\mathscr{K}_i^{\perp}) \\ &= n - n \wedge \sum_{j \neq i} q_j\,(\mathrm{g}), \qquad i \in \mathbf{k}. \end{aligned} \tag{10}$$

By the same reasoning,

$$d(\operatorname{Im} B) = m\,(\mathrm{g}) \tag{11}$$

and

$$\begin{aligned} d(\operatorname{Im} B \cap \mathscr{K}_i) &= d(\operatorname{Im} B) + d(\mathscr{K}_i) - d(\operatorname{Im} B + \mathscr{K}_i) \\ &= m + \left(n - n \wedge \sum_{j \neq i} q_j\right) - n \wedge \left[m + \left(n - n \wedge \sum_{j \neq i} q_j\right)\right] (\mathrm{g}) \\ &= 0 \vee \left(m - n \wedge \sum_{j \neq i} q_j\right) (\mathrm{g}), \qquad i \in \mathbf{k}. \end{aligned} \tag{12}$$

(ii) *Necessity.* Suppose EDP is solvable at **p**. By (3) and (4),

$$\mathscr{K}_i + \ker D_i = \mathscr{X}, \qquad i \in \mathbf{k},$$

or

$$\left(\sum_{j \neq i} \operatorname{Im} D_j'\right) \cap \operatorname{Im} D_i' = 0, \qquad i \in \mathbf{k},$$

that is, the subspaces $\operatorname{Im} D_i' \subset \mathscr{X}'$ are independent. It follows from this and (8) that EDP is generically solvable-only if

$$\sum_{i=1}^{k} q_i \leqq n, \tag{13}$$

as claimed in (5). By (10) and (13),

$$d(\mathscr{K}_i) = n - \sum_{j \neq i} q_j\,(\mathrm{g}), \qquad i \in \mathbf{k}; \tag{14}$$

and by (12) and (13),

$$d(\operatorname{Im} B \cap \mathscr{K}_i) = 0 \vee \left(m - \sum_{j \neq i} q_j\right) (\mathrm{g}). \tag{15}$$

Now if $D_i \neq 0$, (3) implies $\mathscr{R}_i^* \neq 0$ and therefore (see [1]) $\operatorname{Im} B \cap \mathscr{K}_i \neq 0$. Thus by (15) generic solvability of EDP implies

$$m - \sum_{j \neq i} q_j \geqq 1, \qquad i \in \mathbf{k},$$

which is equivalent to (6).

(iii) *Sufficiency.* Suppose (5) and (6) hold. Write $\sum_{j \neq i} q_j = q_i'$. By (5) and (10),

$$d(\mathscr{K}_i) = n - q_i'\,(\mathrm{g}). \tag{16}$$

Write $\operatorname{Im} B = \mathscr{B}$. Using (10), (11) and (16) we have

$$\begin{aligned} d(\mathscr{K}_i + \mathscr{B}) &= n \wedge (n - q_i' + m) \\ &\geqq n \wedge (n + 1) \qquad (\text{by } (6)) \\ &= n\,(\mathrm{g}). \end{aligned}$$

It follows that $\mathscr{K}_i + \mathscr{B} = \mathscr{X}\,(\mathrm{g})$ and thus

$$A\mathscr{K}_i \subset \mathscr{K}_i + \mathscr{B}\,(\mathrm{g}), \qquad i \in \mathbf{k}.$$

By [2, Lemma 4.3, part (i) of proof] we have

$$\mathscr{R}_i^* = \mathscr{R}_i^n\,(\mathrm{g}),$$

where

$$\mathscr{R}_i^{\mu+1} = \mathscr{K}_i \cap (A\mathscr{R}_i^\mu + \mathscr{B}), \qquad \mu = 0, 1, \cdots, n, \tag{17a}$$

and

$$\mathscr{R}_i^0 = 0. \tag{17b}$$

It will be shown that $\mathscr{R}_i^n = \mathscr{K}_i(\mathrm{g})$. For this it is convenient to use a more refined method than heretofore. Replace $\mathbf{p} \in \mathbb{R}^N$ by the indeterminate $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_N)$; i.e., $\boldsymbol{\lambda}$ is simply a list representing the N entries of the matrices $A, B, D_1, \cdots, D_k$ regarded as literal variables. We shall consider $A, \cdots, D_k$ as matrices over the integral domain $\mathbb{R}[\boldsymbol{\lambda}]$ or its fraction field $\mathbb{R}(\boldsymbol{\lambda})$. We then regard the $\mathscr{R}_i^\mu$ defined by (17) as subspaces of the vector space $\mathbb{R}^n(\boldsymbol{\lambda})$. Let $r_{i\mu}$ (resp. $s_{i\mu}$) be the dimension of $\mathscr{R}_i^\mu$ (resp. $S_i^\mu = A\mathscr{R}_i^\mu + \mathscr{B}$) over $\mathbb{R}(\boldsymbol{\lambda})$. We now compute the $r_{i\mu}$ and $s_{i\mu}$, dropping the subscript i for convenience.

LEMMA. *In the setup just described, let*

$$\begin{aligned} \mathscr{S}^\mu &= A\mathscr{R}^\mu + \mathscr{B}, \\ \mathscr{R}^{\mu+1} &= \mathscr{K} \cap \mathscr{S}^\mu, \qquad \mu = 0, 1, \cdots, n, \\ \mathscr{R}^0 &= 0. \end{aligned} \tag{18}$$

Then

$$\begin{aligned} s_\mu &= n \wedge (r_\mu + m), \\ r_{\mu+1} &= n - n \wedge (n - s_\mu + q'), \qquad \mu = 0, 1, \cdots, n, \\ r_0 &= 0. \end{aligned} \tag{19}$$

Proof. Let R_μ be an $n \times r_\mu$ matrix over $\mathbb{R}(\boldsymbol{\lambda})$ whose columns are a basis of $\mathscr{R}^\mu$, introduce the $n \times (r_\mu + m)$ matrix

$$\hat{S}_\mu = [AR_\mu, B], \tag{20}$$

let S_μ be an $n \times s_\mu$ matrix such that $\operatorname{Im} S_\mu = \operatorname{Im} \hat{S}_\mu$, and let $S_\mu^\perp$ be an $(n - s_\mu) \times n$ matrix such that $\ker S_\mu^\perp = \operatorname{Im} S_\mu$. Let D be a $q' \times n$ matrix such that $\ker D = \mathscr{K}$, and write

$$\begin{bmatrix} S_\mu^\perp \\ D \end{bmatrix} = T_\mu. \tag{21}$$

Then by (18),

$$\mathscr{R}^{\mu+1} = \ker T_\mu. \tag{22}$$

We define the *rank polynomial* $\psi(M)$ of a matrix M over $\mathbb{R}(\lambda)$ to be the sum of squares of the minors of M having maximal dimension. Thus M is of full (i.e., maximal) rank if and only if $\psi(M) \neq 0$. By (20),

$$s_\mu = n \wedge (r_\mu + m) \Leftrightarrow \psi(\hat{S}_\mu) \neq 0. \tag{23}$$

By (21),

$$\operatorname{rank} T_\mu = n \wedge (n - s_\mu + q') \Leftrightarrow \psi(T_\mu) \neq 0; \tag{24}$$

hence by (22) and (24),

$$r_{\mu+1} = n - n \wedge (n - s_\mu + q') \Leftrightarrow \psi(T_\mu) \neq 0.$$

Since $r_0 = 0$ trivially, it follows that (19) is true if and only if

$$\psi(T_{n-1})\psi(\hat{S}_{n-1}) \cdots \psi(T_0)\psi(\hat{S}_0) \neq 0. \tag{25}$$

A simple inductive argument on μ shows that the rational matrices $\hat{S}_\mu(\lambda)$ and $T_\mu(\lambda)$ have the property that $\hat{S}_\mu(\mathbf{p})$ and $T_\mu(\mathbf{p})$ are defined (g) for $\mathbf{p} \in \mathbb{R}^N$, and that, if $\mathscr{R}^\mu(\mathbf{p})$, $\mathscr{S}^\mu(\mathbf{p})$ are computed by (18) with $\lambda = \mathbf{p}$, we shall have

$$\mathscr{S}^\mu(\mathbf{p}) = \operatorname{Im} \hat{S}_\mu(\mathbf{p})\,(\mathrm{g})$$

and

$$\mathscr{R}^{\mu+1}(\mathbf{p}) = \ker T_\mu(\mathbf{p})\,(\mathrm{g}).$$

Therefore, to prove (25) it is enough to replace λ by some $\mathbf{p} \in \mathbb{R}^N$ for which the asserted dimensional evaluations can be verified. For this let $e_1, \cdots, e_n$ be the unit vectors in $\mathbb{R}^N$, put $e_v = 0$ if $v > n$, and define

$$\begin{aligned} Ae_j &= 0, & j &= 1, \cdots, q', \\ Ae_{q'+r} &= e_{m+r}, & r &= 1, \cdots, n - q', \\ D &= [I_{q'} \quad 0], & B &= \begin{bmatrix} I_m \\ 0 \end{bmatrix}. \end{aligned} \tag{26}$$

Then easy computations verify that the $\mathscr{R}^\mu$ and $\mathscr{S}^\mu \subset \mathscr{X}$ generated by (18) with (26) indeed have dimensions r_μ and s_μ given by (19). Hence the polynomial in (25) cannot vanish, and the lemma follows.

From the lemma it results at once that

$$d(\mathscr{R}_i^\mu) = r_{i\mu}\,(\mathrm{g})$$

for the subspaces $\mathscr{R}_i^\mu \subset \mathscr{X}$ defined by (17). We claim that

$$r_{in} = n - q_i', \qquad i \in \mathbf{k}, \tag{27}$$

and hence that

$$\mathscr{R}_i^n = \mathscr{K}_i\,(\mathrm{g}), \qquad i \in \mathbf{k}, \tag{28}$$

as asserted. Dropping the subscript i we have from (19):

$$\begin{aligned} r_{\mu+1} &= n - n \wedge [n - n \wedge (r_\mu + m) + q'] \\ &= 0 \vee [n \wedge (r_\mu + m) - q'] \\ &\geqq n \wedge (r_\mu + m) - q' \\ &\geqq n \wedge (r_\mu + 1 + q') - q' \qquad \text{(by (6))}. \end{aligned}$$

If $r_\mu < n - q'$, then $r_\mu + 1 + q' \leqq n$, so

$$\begin{aligned} r_{\mu+1} &\geqq (r_\mu + 1 + q') - q' \\ &= r_\mu + 1. \end{aligned}$$

Since [1] the $\mathscr{R}^\mu$ defined by (18) are nondecreasing, and since $d(\mathscr{R}^\mu) \leqq d(\mathscr{K}) = n - q'$, it follows that $r_\mu \uparrow n - q'$ with convergence in at most $n - q' < n$ steps, and the claim (27) is proved.

Finally, (5) implies that

$$\begin{aligned} \operatorname{rank} \begin{bmatrix} D_1 \\ \vdots \\ D_k \end{bmatrix} &= \sum_{i=1}^{k} q_i \,(\mathrm{g}) \\ &= \sum_{i=1}^{k} \operatorname{rank} D_i \,(\mathrm{g}); \end{aligned}$$

hence that

$$\left(\sum_{j \neq i} \operatorname{Im} D'_j \right) \cap \operatorname{Im} D'_i = 0 \,(\mathrm{g}), \qquad i \in \mathbf{k};$$

and so

$$\mathscr{K}_i + \ker D_i = \mathscr{X} \,(\mathrm{g}), \qquad i \in \mathbf{k}.$$

This combined with (28) shows that (3) is true generically.

4. Concluding remark. The main result may be a useful practical guide in identifying situations where dynamic decoupling is likely, in principle, to be feasible. The notions of "generic solvability" and of "well-posedness" introduced here are nevertheless rather primitive. They imply nothing about the "well-conditionedness" of the computations leading to a solution at a well-posed parameter set, or about the sensitivity of the solution in the neighborhood of such a set. It seems likely, however, that little could be hoped for in those respects if genericity were absent.

Acknowledgment. The authors are indebted to B. Francis for several helpful discussions.

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] W. Fulton, *Algebraic Curves*, W. A. Benjamin, New York, 1969.

[4] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

SIAM J. Control
Vol. 12, No. 4, November 1974

GLOBAL CONTROLLABILITY OF DISTURBED NONLINEAR EQUATIONS*

D. L. LUKES†

Abstract. In [1] the author dealt with the global controllability of nonlinear ordinary differential equations. The present paper treats a more general problem wherein the controller is required to do its steering in the face of measurable but unpredictable disturbances entering the system equation. Appropriate definitions of strong and weak controllability are made and lead to what the author calls a compensator. Sufficient conditions for controllability and existence of a compensator are derived. The results are of some interest in differential games as well as in control.

1. Statement of the weak and strong controllability problems. The control theory of this article is based upon the equations

$$\dot{x}(t) = F(x(t), u(t), v(t), t), \tag{1.1}$$

$$x(t_0) = x_0, \quad x(t_1) = x_1, \tag{1.2}$$

$$u(t) = \Phi(x_0, x_1, t, v), \tag{1.3}$$

with $t_0 \leqq t \leqq t_1$ and $v \in \mathscr{D}$.

Equation (1.1) is a differential equation with variables $x \in R^a$, $u \in R^b$, $v \in R^c$ and $t \in [t_0, t_1]$. The map $F: R^a \times R^b \times R^c \times [t_0, t_1] \to R^a$ and the real parameters t_0, t_1 are assumed to be given. $\mathscr{D}$ denotes a dense linear subspace of the Banach space $L_1^c[t_0, t_1]$ of equivalence classes of Lebesgue integrable mappings of $[t_0, t_1]$ into R^c and $v(\cdot)$ denotes an element of $v \in \mathscr{D}$. To simplify the presentation $\mathscr{D}$ will always be selected so that each $v \in \mathscr{D}$ can be represented by a continuous $v(\cdot)$.

In the context of a discussion in which x_0, x_1 of (1.2) are treated as fixed, the notation $\{x, u\}_v$ will be employed to denote the collection of pairs of maps $x:[t_0, t_1] \to R^a$ absolutely continuous and $u:[t_0, t_1] \to R^b$ continuous such that (1.1) is satisfied at a.e. $t \in [t_0, t_1]$ and both equalities in (1.2) hold. The $v \in \mathscr{D}$ is called a *disturbance* and x, u compatible *state* and *control responses*, respectively. The following definitions are based upon equations (1.1)–(1.2).

DEFINITION 1.1. Call F *weakly controllable* if for some dense $\mathscr{D}$, as described above, the following conditions are satisfied for each x_0 and x_1 in R^a:

(a) $\{x, u\}_v \neq \emptyset$ for each $v \in \mathscr{D}$.

(b) There exists a map $v \to x_v, u_v \in \{x, u\}_v$ such that for any $\tilde{v} \in \mathscr{D}$ satisfying $\tilde{v}|_{[t_0,t]} = v|_{[t_0,t]}$ for some $t \in [t_0, t]$ there exist $\tilde{x}_v, \tilde{u}_v \in \{x, u\}_{\tilde{v}}$ such that $\tilde{x}_v, \tilde{u}_v|_{[t_0,t]} = x_v, u_v|_{[t_0,t]}$.

Remark 1.1. Part (a) of Definitions 1.1 and 1.2 ask that there be control functions doing the steering and (b) asks that the state-control function pair $x(t)$, $u(t)$ resemble a response to v depending upon only the past history of v.

* Received by the editors March 15, 1973, and in revised form July 26, 1973.

† Department of Applied Mathematics and Computer Science, University of Virginia, Charlottesville, Virginia 22901. This work was supported by the United States Office of Naval Research under Contract N00014-69-A-0060-0010.

DEFINITION 1.2. Call F *strongly controllable* or simply *controllable* if for some dense $\mathscr{D}$ the following conditions are satisfied for each x_0, x_1 in R^a:

(a) $\{x, u\}_v \neq \emptyset$ for each $v \in \mathscr{D}$.

(b) There exists a map $v \to x_v, u_v \in \{x, u\}_v$ such that for any $\tilde{v} \in \mathscr{D}$ satisfying $\tilde{v}|_{[t_0,t]} = v|_{[t_0,t]}$ for some $t \in [t_0, t_1]$, $x_{\tilde{v}}, u_{\tilde{v}}|_{[t_0,t]} = x_v, u_v|_{[t_0,t]}$.

In other words a controllable F is a weakly controllable F for which it suffices to use $\tilde{x}_v, \tilde{u}_v = x_{\tilde{v}}, u_{\tilde{v}}$ in the definition of weak controllability.

The next definition leads to a useful alternative definition of strong controllability.

DEFINITION 1.3. A *compensator for F* is a map $\Phi: R^a \times R^a \times [t_0, t_1] \times \mathscr{D} \to R^b$ with the properties:

(a) $\Phi(x_0, x_1, t, v)$ is continuous in t for each fixed $x_0, x_1 \in R^a$ and $v \in \mathscr{D}$.

(b) $\Phi(x_0, x_1, \cdot, \tilde{v})|_{[t_0,t]} = \Phi(x_0, x_1, \cdot, v)|_{[t_0,t]}$ for all $x_0, x_1 \in R^a$ whenever $\tilde{v}$, $v \in \mathscr{D}$ satisfy $\tilde{v}|_{[t_0,t]} = v|_{[t_0,t]}$ for some $t \in [t_0, t_1]$.

(c) Equations (1.1)–(1.3) have at least one solution $x(\cdot)$ for each $x_0, x_1 \in R^a$.

Remark 1.2. Comparison of Definitions 1.2, 1.3 shows that F has a compensator if and only if it is strongly controllable. If F has a compensator Φ then use $u_v = \Phi(x_0, x_1, \cdot, v)$ and the axiom of choice to select a compatible $x(\cdot)$ (if the solutions of (1.1) are not unique) to obtain the required mapping $v \to x_v, u_v$. When F is strongly controllable the composition $v \to x_v, u_v \to u_v$ induces the required compensator Φ.

The primary aim of this article is to establish the weak or strong controllability of a substantial class of nonlinear (and linear) differential equations.

2. Relationships with other control problems. The success of Kalman's original geometric definition of controllability together with its algebraic characterization is widely recognized in linear control theory. The controllability hypothesis occurs frequently in the theory and in particular has led to elegant results illuminating the ultimate capabilities of feedback in performance and stabilization. The situation for nonlinear control is understandably in a much more primitive state. Although the geometric definition readily extends to nonlinear systems the detection of controllability there is to a considerable extent an unsolved problem. In [3] the author did achieve some success in dealing with nonlinear systems possessing a controllable linear part.

The basic difference between this earlier work and the problem to be studied here is the occurrence of the disturbance term v in (1.1). In problems where v is absent from F the definitions of controllability given clearly reduce to the earlier notion, i.e., the concept has been appropriately generalized.

The need for the generalization arises in applications where an unpredictable external disturbance enters the system to be controlled. In some cases it might be in the nature of a stochastic disturbance and in other problems simply be the output of some adjacent equipment. The theory of differential games frequently uses such a model with the v being interpreted as the control variables of other players who are trying to influence the state response.

In some control problems the disturbance $v(\tau)$, $t_0 \leqq \tau \leqq t_1$, is unavoidable but known at the initial time. Those problems require no generalization of the controllability concept since they can be treated as one of the original undisturbed type, $\tilde{F}(x, u, t) = F(x, u, v(t), t)$. However it is frequently unrealistic, in problems involving disturbances, to assume that $u(\tau)$, $\tau \leqq t$, is in some sense dependent upon $v(\tau)$, $t \leqq \tau$. The definitions of weak and strong controllability state this notion of independence in a technically precise manner. To a considerable extent potential future independence is already present in a weakly controllable system. But it is the compensator function Φ of a strongly controllable F which an engineer requires in order to build the hardware for doing the steering.

In the theory developed in the following sections continuity properties of F contribute to concluding weak controllability while differentiability strengthens the conclusion to strong controllability.

3. The main results. Attention will be restricted to those differential equations (1.1) for which F can be represented in the form

$$F(x, u, v, t) = Ax + Bu + f(x, u, v, t) \tag{3.1}$$

in which A and B are real matrices and f is continuous.

Define the *Kalman index* $k(A, B)$ to be the rank of the matrix $[B, AB, \cdots, A^{a-1}B]$ of size $a \times ab$ whose columns are those of the indicated matrix products of A and B.

THEOREM 3.1. *F is (weakly) controllable if the following conditions are satisfied*:

(a) $k(A, B) = a$;

(b) $f(x, u, v_0, t) = 0$ *for some* $v_0 \in R^c$ *and all* $x \in R^a$, $u \in R^b$ *and t in a neighborhood of* t_1;

(c) $\sup |f(x, u, v, t)| < \infty$ *for each* $\gamma < \infty$, *the* sup *being taken over the set* $|x| < \infty$, $|u| < \infty$, $|v| < \gamma$, $t \in [t_0, t_1]$.

COROLLARY 3.1. *The system* $F = Ax + Bu + C(x, u, v, t)v$ *is (weakly) controllable if* $k(A, B) = a$, *the matrix function C is continuous and for each* $\gamma < \infty$ *the* $\sup |C|$ *taken over the set* $|x| < \infty$, $|u| < \infty$, $|v| < \gamma$, $t \in [t_0, t_1]$ *is finite.*

Remark 3.1. The word weakly has been enclosed in parentheses in the conclusions of Theorem 3.1 and Corollary 3.1 since it can be dropped, i.e., strong controllability can be concluded, by making the additional assumption that F satisfies a Lipschitz condition. The required condition for Theorem 3.1 would state that for each bounded subset $K \subset R^a \times R^b \times R^c \times [t_0, t_1]$ there exists a number L_K such that

$$|f(x, u, v, t) - f(\tilde{x}, \tilde{u}, v, t)| \leqq L_K[|x - \tilde{x}| + |u - \tilde{u}|] \tag{3.2}$$

for all (x, u, v, t) and $(\tilde{x}, \tilde{u}, v, t)$ in K. In particular it is easy to check that (3.2) would hold if, for example, f has continuous first order partial derivatives in the coordinates of x and u. Corollary 3.1 can be dealt with by inequality (3.2) applied to $C(x, u, v, t)$ with continuity of the first order partial derivatives of C being sufficient.

THEOREM 3.2. *The system* $F = Ax + Bu + f(v, t)$ *is controllable if* $k(A, B) = a$, *f is continuous and* $f(v_0, t) = 0$ *for some* $v_0 \in R^c$ *and all t in a neighborhood of* t_1. *If f is continuous then* $k(A, B) = a$ *is necessary for F to be controllable.*

COROLLARY 3.2. *The system* $F = Ax + Bu + C(v, t)v$ *with the matrix function C continuous is controllable if and only if* $k(A, B) = a$.

Remark 3.2. Important special cases of Corollary 3.2 are those in which $C(v, t)$ is independent of one or both of the variables v, t and more particularly where C is the identity matrix.

Remark 3.3. The proof of Theorem 3.1 provides a synthesized formula for a compensator for F.

THEOREM 3.3. *The members of the one-parameter family of control systems $F_\varepsilon = Ax + Bu + \varepsilon f(x, u, v, t)$ corresponding to the values of the parameter ε in some neighborhood of zero are controllable if*:

(a) $k(A, B) = a$;

(b) $f(x, u, v_0, t) = 0$ *for some $v_0 \in R^c$ and all $x \in R^a$, $u \in R^b$ and t in a neighborhood of t_1*;

(c) $f(x, u, v, t)$ *is continuous and satisfies inequality* (3.2) *for some $L = L_K$ independent of K.*

4. The proof of Theorem 3.1. The proof is developed by means of a sequence of lemmas. To simplify the notation a preliminary translation of the origin in t-space allows us to let $t_0 = 0$ and we denote t_1 as T, and x_1 as x_T. In view of the earlier observation that when v is absent from F the definition of a controllable system reduces to the standard one there is no reason to present the proof of the following lemma which is proven in most texts on linear control theory (see [2, p. 99]). (Ordinary matrix transposition is denoted by *.)

LEMMA 4.1. *The following statements are equivalent*:

(a) $F = Ax + Bu$ *is controllable*;

(b) $k(A, B) = a$;

(c) $S_\varepsilon = \int_0^\varepsilon e^{A\sigma}BB^* e^{A^*\sigma}\, d\sigma$ *is symmetric positive definite and hence nonsingular for each $\varepsilon > 0$.*

It will prove convenient to introduce the following notations:

$$r_1(t) = e^{At}x_0 + S_t e^{A^*(T-t)} S_T^{-1}(x_T - e^{AT}x_0), \tag{4.1}$$

$$r_2(t) = B^* e^{A^*(T-t)} S_T^{-1}(x_T - e^{AT}x_0), \tag{4.2}$$

$$K_1(t, \omega) = e^{A(t-\omega)} - S_{t-\omega} e^{A^*(T-t)} S_{T-\omega}^{-1} e^{A(T-\omega)}, \tag{4.3}$$

$$K_2(t, \omega) = -B^* e^{A^*(T-t)} S_{T-\omega}^{-1} e^{A(T-\omega)}, \tag{4.4}$$

$$F_1(x, u)(t) = \int_0^t K_1(t, \omega) f(x(\omega), u(\omega), v(\omega), \omega)\, d\omega, \tag{4.5}$$

$$F_2(x, u)(t) = \int_0^t K_2(t, \omega) f(x(\omega), u(\omega), v(\omega), \omega)\, d\omega. \tag{4.6}$$

LEMMA 4.2. *Any continuous solution $x(\cdot)$, $u(\cdot)$ to the system of equations*

$$x(t) = F_1(x, u)(t) + r_1(t), \tag{4.7}$$

$$u(t) = F_2(x, u)(t) + r_2(t), \tag{4.8}$$

$0 \leqq t \leqq T$, *provides a solution to the boundary value problem* (1.1)–(1.2) *for F given by* (3.1).

Proof. Since Theorem 3.1 assumes that $k(A, B) = a$ it will be sufficient to deal with the situation wherein S_ε as defined by (c) of Lemma 4.1 is nonsingular for each $\varepsilon > 0$. Clearly (4.1)–(4.2) then define continuously differentiable functions $r_i(t)$ $(i = 1, 2)$ on $0 \leqq t \leqq T$. Moreover the $K_i(t, \omega)$ $(i = 1, 2)$ determined by (4.3)–(4.4) have continuous derivatives in the triangular region $0 \leqq \omega \leqq t \leqq T$ with the vertex $t = \omega = T$ deleted. The singularity of the K_i at this one point will not prove to be an insurmountable difficulty.

Assuming that there exists a solution $x(\cdot)$, $u(\cdot)$ to (4.7)–(4.8) it is a trivial matter to check that $x(0) = x_0$ since both S_t and $F_1(x, u)(t)$ are zero at $t = 0$. By noting that $K_1(T, \omega) = 0$ and $r_1(T) = x_T$ it is apparent that $F_1(x, u)(T) = 0$ and consequently (4.7) computes $x(T) = x_T$ as required.

The conclusion of the proof requires the verification of (1.1). The formal differentiation of (4.7) together with employment of (4.5), (4.6), (4.8) and the observation that $K_1(t, t) = I$ results in the calculation

$$\begin{aligned} &\dot{x}(t) - [Ax(t) + Bu(t) + f(x(t), u(t), v(t), t)] \\ (4.9)\quad &= \int_0^t \left\{ \frac{\partial}{\partial t} K_1(t, \omega) - AK_1(t, \omega) - BK_2(t, \omega) \right\} f(x(\omega), u(\omega), v(\omega), \omega)\, d\omega \\ &\quad + \dot{r}_1(t) - Ar_1(t) - Br_2(t). \end{aligned}$$

Using the defining equations (4.1)–(4.4) one can easily show that

$$(4.10)\qquad \frac{\partial}{\partial t} K_1(t, \omega) - AK_1(t, \omega) - BK_2(t, \omega) = 0,$$

$$(4.11)\qquad \dot{r}_1(t) - Ar_1(t) - Br_2(t) = 0.$$

The immediate conclusion drawn from (4.9)–(4.11) is the required equation

$$(4.12)\qquad \dot{x}(t) = Ax(t) + Bu(t) + f(x(t), u(t), v(t), t).$$

Remark 4.1. The formal differentiation in the last part of the proof of Lemma 4.2 requires a word of justification. First note that by restricting $0 \leqq t < T$ the singularity in K_1 at $\omega = t = T$ produces no problem. The formal differentiation of the integral F_1 is legitimate as long as $f(x(\omega), u(\omega), v(\omega), \omega)$ is continuous in ω. By the continuity assumptions on f, $x(\cdot)$ and $u(\cdot)$, that condition is met if $v(\cdot)$ is continuous. If $v(\cdot)$ is only bounded and measurable then equation (4.12) will in general hold almost everywhere (Lebesgue measure). For the purpose of this article it will be satisfactory to deal primarily with continuous $v(\cdot)$.

LEMMA 4.3. *If F satisfies conditions* (a)–(c) *of Theorem* 3.1 *then there is a dense linear subspace of continuous functions $\mathscr{D} \subset L_1^c[0, T]$ such that* (4.7)–(4.8) *have a continuous solution $x(\cdot)$, $u(\cdot)$ for each $v \in \mathscr{D}$.*

Preliminary simplification of the problem. Notice that without any loss in generality v_0 can be taken equal to zero (replace $f(x, u, v, t)$ by $f(x, u, v - v_0, t)$). Furthermore the r_i $(i = 1, 2)$ in (4.7)–(4.8) can be assumed to be identically zero (since the transformation $x \to x - r_1$, $u \to u - r_2$ preserves properties (a)–(c) of f).

The candidate for $\mathscr{D}$ is the collection of continuous maps $v(\cdot)$ from $[0, T]$ into R^c having the property that $v(\omega) = 0$ for some neighborhood (depending upon v) of $\omega = T$. Select arbitrary $v \in \mathscr{D}$ and consider the new function $\tilde{f}(x, u, \omega) = f(x, u, v(\omega), \omega)$. This function inherits the properties of f assumed in (b)–(c) of

Theorem 3.1. Consequently it is clearly adequate to establish the existence of a continuous solution $x(\cdot)$, $u(\cdot)$ to the simplified equations

$$x(t) = F_1(x, u)(t), \tag{4.13}$$

$$u(t) = F_2(x, u)(t), \tag{4.14}$$

where in (4.5)–(4.6), $f = f(x, u, \omega)$ is continuous on $R^a \times R^b \times [0, T]$ and

$$f(x, u, \omega) = 0 \quad \text{for all } x \in R^a, \tag{4.15}$$

$u \in R^b$ and $\omega \in [T - \varepsilon, T]$ for some $\varepsilon > 0$ (depending upon v). In addition there exists a bound

$$|f(x, u, \omega)| \leqq \gamma_1 \quad \text{for all } x \in R^a,\ u \in R^b \text{ and } \omega \in [0, T]. \tag{4.16}$$

Completion of the proof of Lemma 4.3. The proof is based upon an application of the Schauder fixed-point theorem (see [1, p. 456]).

Consider the vector space $E = E^a \oplus E^b$, where E^a is the linear space of continuous maps from $[0, T]$ into R^a and E^b is similarly defined. E is to be normed using $|(x, u)| = \max |x(t)| + \max |u(t)|$ in which the maximum is taken over $t \in [0, T]$. This makes E a Banach space with convergence in the norm topology being uniform convergence.

Consider the formula

$$\tilde{F}(x, u)(t) = [F_1(x, u)(t), F_2(x, u)(t)]. \tag{4.17}$$

The following related statements will be verified:

(a) $\tilde{F}$ defines a continuous map of E into itself.

(b) $|\tilde{F}(x, u)|$ is bounded on E.

(c) For all ρ sufficiently large, $\tilde{F}$ maps E into the subset of E,

$$E_\rho = \{(x, u) : |(x, u)(t + \Delta) - (x, u)(t)| \leqq \rho|\Delta|, |(x, u)| \leqq \rho, \text{ all } t, \Delta\}.$$

(d) E_ρ is a compact and convex subset of the Banach space E.

To aid in seeing that $\tilde{F}(x, u) \in E$ whenever $(x, u) \in E$, utilize (4.15) to write

$$F_i(x, u)(t) = \int_0^{\min(t, T-\varepsilon)} K_i(t, \omega) f(x(\omega), u(\omega), \omega)\, d\omega, \tag{4.18}$$

and notice that the integrand is continuous for $0 \leqq \omega \leqq t \leqq T - \varepsilon$ and that the upper limit of integration is continuous in t. Hence (4.18) defines continuous functions of t ($i = 1, 2$).

By continuity on compact sets it follows that there exists a bound

$$|K_i(t, \omega)| \leqq \gamma_2 \quad \text{for } 0 \leqq \omega \leqq t \leqq T - \varepsilon \tag{4.19}$$

(i = 1, 2). Let $(x_\infty, u_\infty) \in E$ and $(x_k, u_k) \in E$ with (x_k, u_k) converging to (x_∞, u_∞) as $k \to \infty$. Estimation with (4.19) gives

$$\begin{aligned} &|F_i(x_k, u_k)(t) - F_i(x_\infty, u_\infty)(t)| \\ &\leqq \gamma_2 \int_0^T |f(x_k(\omega), u_k(\omega), \omega) - f(x_\infty(\omega), u_\infty(\omega), \omega)|\, d\omega, \end{aligned} \tag{4.20}$$

from which it follows by the continuity of f and hence uniform continuity of f on compact subsets of $R^a \times R^b \times [0, T]$ that the right-hand side of (4.20) converges to zero as $k \to \infty$. But this implies that the left-hand side converges uniformly to zero and finally then that $\tilde{F}(x_k, u_k) \to \tilde{F}(x_\infty, u_\infty)$ as $k \to \infty$. This proves (a). Part (b) follows readily from (4.18) and (4.16) by continuity.

The first step in establishing (c) is to notice that from formulas (4.3) and (4.4) one can show there is a constant γ_3 such that

$$|K_i(t + \Delta, \omega) - K_i(t, \omega)| \leqq \gamma_3 |\Delta| \tag{4.21}$$

for all $0 \leqq \omega \leqq t \leqq T - \varepsilon$, $(i = 1, 2)$. Then with this inequality estimate,

$$\begin{aligned} &|F_i(x, u)(t + \Delta) - F_i(x, u)(t)| \\ &\leqq \left| \int_0^{\min(t+\Delta, T-\varepsilon)} [K_i(t + \Delta, \omega) - K_i(t, \omega)] f(x(\omega), u(\omega), \omega)\, d\omega \right. \\ &\quad + \left| \int_{\min(t, T-\varepsilon)}^{\min(t+\Delta, T-\varepsilon)} K_i(t, \omega) f(x(\omega), u(\omega), \omega)\, d\omega \right| \\ &\geqq T\gamma_3 |\Delta| \gamma_1 + |\Delta| \gamma_2 \gamma_1 = T\gamma_1(\gamma_2 + \gamma_3)|\Delta| \end{aligned} \tag{4.22}$$

for all $(x, u) \in E$. Conclusion (c) follows immediately from (4.22).

It is a simple exercise to show E_ρ is a closed and convex subset of E. Since E_ρ is a bounded subset of E, once its closed character has been established the observation that it is an equi-uniformly-continuous family allows application of the Arzela–Ascoli theorem to conclude its compactness.

A direct consequence of (a)–(d) is that for all sufficiently large ρ the restriction of $\tilde{F}$ to E_ρ provides a continuous map of a compact and convex subset of a Banach space into itself. Hence the Schauder theorem applies to produce the conclusion that $\tilde{F}$ has a fixed point in E_ρ. In other words (4.13)–(4.14) has a continuous solution $x(\cdot)$, $u(\cdot)$ in E and the proof of Lemma 4.3 is complete.

Conclusion of the proof of Theorem 3.1. Lemmas 4.1–4.3 show that under the hypothesis of Theorem 3.1, for the dense linear subspace $\mathscr{D} \subset L_1^c[0, T]$ of continuous functions selected, there exists for each fixed x_0, $x_T \in R^a$ a solution $x(\cdot)$, $u(\cdot)$ to (4.7)–(4.8) on $[0, T]$ for each $v \in \mathscr{D}$. This says that the first condition in Definition 1.1, $\{x, u\}_v \neq \emptyset$ for each $v \in \mathscr{D}$, is met. Although the fixed point is not necessarily unique the axiom of choice provides the means for selecting a unique solution x_v, $u_v \in \{x, u\}_v$ and thus a candidate map $v \to x_v$, u_v is available for testing against Definition 1.1.

To show that condition (b) of the latter definition is met by the candidate map, consider v, $\tilde{v} \in \mathscr{D}$ with $\tilde{v}|_{[0,t]} = v|_{[0,t]}$ for some fixed $t \in [0, T]$. What is required is showing that $x_v, u_v|_{[0,t]}$ can be extended to $[0, T]$ as a solution to equations (4.7)–(4.8) corresponding to replacement of v by $\tilde{v}$ in (4.5)–(4.6). Consider the equations

$$\tilde{x}(\tau) = \int_t^\tau K_1(\tau, \omega) f(\tilde{x}(\omega), \tilde{u}(\omega), \tilde{v}(\omega), \omega)\, d\omega + \tilde{r}_1(\tau), \tag{4.23}$$

$$\tilde{u}(\tau) = \int_t^\tau K_2(\tau, \omega) f(\tilde{x}(\omega), \tilde{u}(\omega), \tilde{v}(\omega), \omega)\, d\omega + \tilde{r}_2(\tau), \tag{4.24}$$

$t \leqq \tau \leqq T$, where

$$\tilde{r}_i(\tau) = \int_0^t K_i(\tau, \omega) f(x_v(\omega), u_v(\omega), v(\omega), \omega)\, d\omega + r_i(\tau) \tag{4.25}$$

$(i = 1, 2)$.

If (4.23)–(4.24) can be shown to have a solution those functions could be extended backward to $[0, T]$ by setting $\tilde{x}(\tau), \tilde{u}(\tau) = x_v(\tau), u_v(\tau)$ for $0 \leqq \tau \leqq t$. Then using the fact that $\tilde{v}|_{[0,t]} = v|_{[0,t]}$, the substitution of (4.25) back into (4.23)–(4.24) and addition of the integrals would show that the extended maps $\tilde{x}$, $\tilde{u}$ satisfy the equations

$$\tilde{x}(\tau) = \int_0^\tau K_1(\tau, \omega) f(\tilde{x}(\omega), \tilde{u}(\omega), \tilde{v}(\omega), \omega)\, d\omega + r_1(\tau), \tag{4.26}$$

$$\tilde{u}(\tau) = \int_0^\tau K_2(\tau, \omega) f(\tilde{x}(\omega), \tilde{u}(\omega), \tilde{v}(\omega), \omega)\, d\omega + r_2(\tau) \tag{4.27}$$

for $0 \leqq \tau \leqq T$. Consequently by Lemma 4.2, condition (b) of Definition 1.1, namely $\tilde{x}, \tilde{u} \in \{x, u\}_{\tilde{v}}$ with $\tilde{x}, \tilde{u}|_{[0,t]} = x_v, u_v|_{[0,t]}$, would be verified.

The existence of a solution to (4.23)–(4.24) however has already been taken care of by the remarks about the preliminary simplification preceding the proof of Lemma 4.3 and the proof of that lemma. This completes the proof of Theorem 3.1.

Proof of Remark 3.1. This is an appropriate point to ask "why, without additional assumptions about f, does the above argument not conclude strong controllability?" If (4.7)–(4.8) fail to have a unique solution for each v there would be some reason to expect that although F is weakly controllable it might not be strongly controllable. (Of course those equations might not have solutions but F could still be strongly controllable since the equations are obviously not equivalent to (1.1).) However, any additional condition on f which guarantees unique solutions would give the strengthened conclusion. One example is condition (3.2).

Suppose f satisfies (3.2) and for some $v, \tilde{v} \in \mathscr{D}$, $\tilde{v}|_{[0,t]} = v|_{[0,t]}$ for some $t \in [0, T)$. Since $|x_v|$, $|u_v|$, $|v|$ and $|x_{\tilde{v}}|$, $|u_{\tilde{v}}|$, $|\tilde{v}|$ are continuous they are bounded on $[0, t]$ and (3.2) can be applied to estimate the differences via (4.5)–(4.8):

$$\begin{aligned} |x_v(\tau) - x_{\tilde{v}}(\tau)| &\leqq \int_0^\tau |K_1(\tau, \omega)| L[|x_v(\omega) - x_{\tilde{v}}(\omega)| + |u_v(\omega) - u_{\tilde{v}}(\omega)|]\, d\omega \\ &\leqq \gamma_2 L \int_0^\tau [|x_v(\omega) - x_{\tilde{v}}(\omega)| + |u_v(\omega) - u_{\tilde{v}}(\omega)|]\, d\omega \end{aligned} \tag{4.28}$$

and similarly

$$|u_v(\omega) - u_{\tilde{v}}(\omega)| \leqq \gamma_2 L \int_0^\tau [|x_v(\omega) - x_{\tilde{v}}(\omega)| + |u_v(\omega) - u_{\tilde{v}}(\omega)|]\, d\omega \tag{4.29}$$

for $0 \leqq \tau \leqq t < T$ with the constant γ_2 possible dependent upon t but not τ.

Addition of (4.28)–(4.29) gives

$$\frac{dp(\tau)}{d\tau} \leqq 2\gamma_2 L p(\tau), \tag{4.30}$$

where $p(\tau)$ is the integral appearing in (4.29). Multiplication of (4.30) by $\exp(-2\gamma_2 L\tau)$ and integration shows that

$$p(\tau)\exp(-2\gamma_2 L\tau) \leqq 0. \tag{4.31}$$

But $p(\tau) \geqq 0$ and it is clear $p(\tau)$ must be zero for $0 \leqq \tau \leqq t$. Hence $x_{\tilde{v}}, u_{\tilde{v}}|_{[0,t]} = x_v$, $u_v|_{[0,t]}$ and the proof of the strong controllability is complete.

5. Remark about the proof of Theorem 3.3. The details of the proof of Theorem 3.3 are not presented here since the approach is very much like the one used on Theorem 3.1. Since the parameter ε can be freely restricted to keep $|\varepsilon|$ small, a contractive mapping argument replaces the Schauder theorem to obtain weak controllability. The strong controllability is a result of uniqueness exactly as before.

6. Other remarks about proofs. Most of the corollaries follow from the proven theorems without further explanation. In Corollary 3.1 the required boundedness results from that of $|C|$ and the boundedness of $|v(t)|$ which comes from the assumed continuity of v. The appropriate choice $v_0 = 0$ is clear.

To show the necessity of the condition $k(A, B) = a$ in Theorem 3.2, suppose that $k(A, B) < a$. From the linear theory (see [2, p. 99]) there exists a nonsingular change of state variables which transforms F into the form $\tilde{F} = \tilde{A}\tilde{x} + \tilde{B}u + \tilde{f}(v, t)$ in which $\tilde{A}$ and $\tilde{B}$ have the form

$$\tilde{A} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}.$$

It is then clear that there are states which cannot be connected by a response for each choice of v. The argument is concluded by noting that this property is preserved by nonsingular linear transformations on R^a.

7. Discussion of the results. Although the proofs were frequently nonconstructive, there are problems for which the developed theory presents a computational procedure. For example if the control variable u is absent from f so F is of the form $F(x, u, v, t) = Ax + Bu + f(x, v, t)$ with (3.2) holding, then the steering compensator can be computed from (4.8) as

$$u(t) = \Phi(x_0, x_1, t, v) = \int_0^t K_2(t, \omega) f(x(\omega), v(\omega), \omega)\, d\omega + r_2(t). \tag{7.1}$$

The instantaneous measurement of the state $x(t)$ and the disturbance $v(t)$ would be required in the implementation of (7.1). This synthesized compensator applies as well to problems where u enters f.

Simple examples show that in general there will not exist a compensator that will do the required steering if $\mathscr{D} = L_1^c$, at least not if the steering signals are to contain finite energy in the sense that $\int_0^\infty |u(t)|\, dt \leqq \infty$. That is why the definition requires that a compensator need only be able to steer against a linear subspace $\mathscr{D}$ which is dense in L_1^c. In view of the remarks made in § 2 and the results of [3] this sacrifice results from not allowing the controller future information about v. Of course in some special cases a compensator exists with $\mathscr{D} = L_1^c$. For example consider $F = Ax + Bu + Cv$ and suppose there exists a matrix D such that $BD = -C$. Let $u = Dv + \tilde{u}$ to transform F into $F = Ax + B\tilde{u}$. Then a compensator for $\mathscr{D} = L_1^c$ is obtained by solving the resulting controllability problem for the undisturbed system.

One can easily show that the compensators shown to exist in § 3 can handle a disturbance if $|v(t) - v_0| \to 0$ sufficiently fast as $t \to t_1$. In the differential game setting in which v is interpreted as the control variables of opposing players this says the v-players get weak toward the end. In capture problems the state splits into $x = (x_1, x_2)$ with $x_1, x_2 \in R^\alpha$ and $2\alpha = a$. The capture condition then can be taken as $x_1 = x_2$. By selecting the final point on the diagonal in $R^\alpha \times R^\alpha$, the player in command of u (the pursuer) can then guarantee capture for all evasive maneuvers $v \in \mathscr{D}$.

As a final remark it should be noted that although the members v of the disturbance class $\mathscr{D}$ selected in the proofs "turned off" prior to $t = t_1$ and thereby killed off the nonlinear term the problem solved could not be treated by simply waiting for the nonlinearity to disappear and then solving the problem of steering an undisturbed linear system. The compensator never knows ahead of time when the disturbance might vanish and in fact might not vanish until the very end time, in which case it would be too late to do any corrective steering.

REFERENCES

[1] N. Dunford and J. T. Schwartz, *Linear Operators, Part 1*, Interscience, New York, 1958.

[2] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[3] D. L. Lukes, *Global controllability of nonlinear systems*, this Journal, 10 (1972), pp. 112–126 (Erratum: 11 (1973), p. 186).

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

CONVEXITY AND PROPERTY (Q) IN OPTIMAL CONTROL THEORY*

L. CESARI† AND M. B. SURYANARAYANA‡

Abstract. In the present paper the authors present a number of results concerning closure, lower closure and lower semicontinuity in problems of optimal control. In these theorems the convexity of the relevant sets plays a major role, together with certain further conditions, usually denoted as "seminormality" conditions, among which are, in particular, properties of upper semicontinuity of variable sets. Particular emphasis is given to recent results where seminormality conditions have been reduced to a minimum, or not used at all. We also show that certain analytical conditions of practical significance, recently used in lieu of seminormality, imply a weak form of seminormality.

Introduction. In the present paper we present a number of results concerning closure, lower closure, and lower semicontinuity in problems of the calculus of variations and optimal control theory. In these theorems the convexity of the relevant sets plays a major role, together with certain further conditions, usually denoted as "seminormality" conditions, among which are, in particular, Kuratowski's property of upper semicontinuity [27] of variable sets, or Property (K) (also denoted as Property (U) in [7]) and its recent variant, Property (Q) [7]–[15]. Property (Q) has been recently used by many authors, such as C. Olech [31], A. LaSota and C. Olech [29], T. S. Angell [1], R. F. Baum [2], D. E. Cowles [21], [22], M. B. Suryanarayana [33], L. Cesari, J. R. LaPalm and D. A. Sanchez [17], M. F. Bidaut [4], C. Çastaing and M. Valadier [6], S. N. Chow and J. D. Schuur [19] and P. Kaiser [26].

In the present paper we summarize results concerning closure theorems (§ 1), lower closure theorems (§ 2), and lower semicontinuity (§ 3), with particular emphasis on recent results where "seminormality" conditions have been reduced to a minimum, or not used at all. In § 1.7 we restate Property (Q) in different forms and we prove, in particular, that certain analytical conditions of practical significance (§ 1.6), recently used in lieu of seminormality, imply weak forms of seminormality.

1. Abstract Mayer problems.

1.1. Notations. Let G be a given measurable subset of the t-space E^ν of finite measure, $t = (t^1, \cdots, t^\nu)$; for every $t \in G$, let $A(t)$ be a given nonempty subset of the x-space E^n, $x = (x^1, \cdots, x^n)$, and let $A = \{(t, x) | t \in G,\ x \in A(t)\}$. For every $(t, x) \in A$, let $Q(t, x)$ be a given subset of the z-space E^r, $z = (z^1, \cdots, z^r)$.

Main Problem 1. Given a sequence of measurable functions $x(t)$, $\xi(t)$, $x_k(t)$, $\xi_k(t)$, $t \in G$, $k = 1, 2, \cdots$, $x, x_k : G \to E^n$, $\xi, \xi_k : G \to E^r$ with

$$(1)\qquad x_k(t) \in A(t), \qquad \xi_k(t) \in Q(t, x_k(t)), \qquad t \in G \text{ a.e.}, \quad k = 1, 2, \cdots,$$

find modes of convergence $x_k \to x$, $\xi_k \to \xi$ and conditions on $A(t)$, $Q(t, x)$, such that

$$(2)\qquad x(t) \in A(t), \qquad \xi(t) \in Q(t, x(t)), \qquad t \in G \text{ a.e.}$$

* Received by the editors October 2, 1972, and in revised form August 1973. This research was done in the frame of AFOSR Research Project 71-2122 at the University of Michigan.

† Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48104.

‡ Department of Mathematics, Eastern Michigan University, Ypsilanti, Michigan 48197.

Let us assume now that for every $(t, x) \in A$, a subset $U(t, x)$ of the u-space E^m is given, let $M = \{(t, x, u) | (t, x) \in A, u \in U(t, x)\}$, let $f(t, x, u) = (f_1, \cdots, f_r)$ be defined on M, and let $Q(t, x)$ denote the sets

$$Q(t, x) = f(t, x, U(t, x)) = \{z | z = f(t, x, u), u \in U(t, x)\}.$$

Main Problem 2. Given a sequence of measurable functions $x(t)$, $\xi(t)$, $x_k(t)$, $\xi_k(t)$, $u_k(t)$, $t \in G$, $k = 1, 2, \cdots$, $x, x_k \in G \to E^n$, $\xi, \xi_k : G \to E^r$, $u_k : G \to E^m$ with

$$\begin{aligned} &x_k(t) \in A(t), \quad u_k(t) \in U(t, x_k(t)), \\ &\xi_k(t) = f(t, x_k(t), u_k(t)), \qquad t \in G \text{ a.e.}, \quad k = 1, 2, \cdots, \end{aligned} \tag{3}$$

find "modes of convergence" $x_k \to x$, $\xi_k \to \xi$ and conditions on $A(t)$, $U(t, x)$, M, $Q(t, x)$, $f(t, x, u)$ which imply that there is a measurable $u : G \to E^m$ such that

$$\begin{aligned} &x(t) \in A(t), \quad u(t) \in U(t, x(t)), \\ &\xi(t) = f(t, x(t), u(t)), \qquad t \in G \quad \text{a.e.} \end{aligned} \tag{4}$$

These are the two main forms of the so-called *closure problem* in abstract Mayer problems. Problem 1 is obviously more general than Problem 2. On the other hand, when Q is defined in terms of f and U as in Problem 2, then implicit function theorems (E. J. McShane and R. B. Warfield [30]) guarantee that solutions to Problem 2 can be derived from solutions to Problem 1. Other implicit function theorems have been proven, e.g., by C. Castaing and M. Valadier [6], C. J. Himmelberg and F. S. VanVleck [24] and M. Q. Jacobs [25].

For the sake of brevity we shall report here only results concerning Problem 2 as more immediately applicable. Under the conditions of all statements below, the implicit function theorems in [30] apply.

1.2. Filippov's statements. To simplify the exposition, let $\nu = 1$, $G = [a, b] \subset E^1$, and A be compact. For any $(\bar{t}, \bar{x}) \in A$ we denote by $N_\delta(\bar{t}, \bar{x})$ (closed δ-neighborhood of $(\bar{t}, \bar{x})$ in A) the set of all $(t, x) \in A$ at a distance $\leqq \delta$ from $(\bar{t}, \bar{x})$. For any set $F \subset E^r$ we denote by F_ε the closed ε-neighborhood of F in E^r.

We say that variable sets $Q(t, x)$ are metric upper semicontinuous at $(\bar{t}, \bar{x})$ provided we have the following property.

Property (*). Given $\varepsilon > 0$ there is a $\delta > 0$ such that for all $(t, x) \in N_\delta(\bar{t}, \bar{x})$, we have $Q(t, x) \subset [Q(\bar{t}, \bar{x})]_\varepsilon$.

The sets $Q(t, x)$ are said to be *metric upper semicontinuous on A* provided they have the above property at every point $(\bar{t}, \bar{x}) \in A$. This property is also denoted as *upper semicontinuity by set inclusion.*

Filippov's solution [23] to Problem 2 is well represented by the following simple statement.

STATEMENT 1.2.i. *Let $x, x_k : [a, b] \to E^n$ be absolutely continuous functions, let $\xi(t) = x'(t)$, $\xi_k(t) = x_k'(t)$, $t \in [a, b]$ a.e., $u_k : [a, b] \to E^m$ measurable, for which* (3) *holds, and let $x_k \to x$ uniformly in $[a, b]$ as $k \to \infty$* (*no further requirement on ξ, ξ_k*). *Let us assume that the sets A and M are compact, that f is continuous on M, and the sets $Q(t, x)$ are all convex. Then there is a measurable $u : [a, b] \to E^m$ for which* (4) *holds.*

Under the hypotheses of this theorem, say A and M compact and f continuous on M, then the sets $Q(t, x)$ are necessarily compact, equibounded, and metric upper semicontinuous on A.

1.3. Properties (K) and (Q). The property of "metric upper semicontinuity" mentioned in § 1.2, though quite adequate for equibounded closed variable sets, does not seem to be general enough for closed sets which may be unbounded, or at least not equibounded. A more general concept is the following one. Let S be a metric space, Y a topological vector space, and let 2^Y denote the set of all subsets of Y. Then a set-valued function $Q:S \to 2^Y$ is said to have Property (K) (Kuratowski's upper semicontinuity [27]) at a point $\bar{s} \in S$ provided

$$Q(\bar{s}) = \bigcap_{\delta>0} \operatorname{cl} \bigcup_{s \in N_\delta(\bar{s})} Q(s), \tag{5}$$

where $N_\delta(\bar{s})$ is the closed δ-neighborhood of $\bar{s}$ in S. Analogously, Q is said to have Property (Q) at the point $\bar{s} \in S$ provided

$$Q(\bar{s}) = \bigcap_{\delta>0} \operatorname{cl} \operatorname{co} \bigcup_{s \in N_\delta(\bar{s})} Q(s).$$

Here cl Z denotes the closure of a set Z in the space Y, and cl co Z denotes the closure of the convex hull of Z.

Since $Q(\bar{s})$ is in any case a subset of the second member of (5), then the $\subset$ sign is always valid there, and so we are actually requiring that the $\supset$ sign hold in (5). Also, Q is said to have Property (K) in S if it has property (5) at every point $\bar{s} \in S$. Sets $Q(s)$ having Property (K) are closed as intersections of closed sets.

Note that if $\mathscr{M} = \{(s, z) | s \in S, z \in Q(s)\} \subset A \times Y$, then $\mathscr{M}$ is the *graph* of the set-valued function Q, and it is easy to prove that $\mathscr{M}$ is closed in $A \times Y$ if and only if Q has Property (K) in A (cf. [7]).

For sets $Q(t, x)$, $(t, x) \in A$, $Q(t, x) \subset E^r$, as in § 1.1, we may require Properties (K), (Q) *with respect to* (t, x) *in* A, say for Property (Q):

$$Q(\bar{t}, \bar{x}) = \bigcap_{\delta>0} \operatorname{cl} \operatorname{co} \bigcup_{(t,x) \in N_\delta(\bar{t},\bar{x})} Q(t, x), \tag{6}$$

where $N_\delta(\bar{t}, \bar{x})$ denotes the set of all $(t, x) \in A$ at a distance $\leqq \delta$ from $(\bar{t}, \bar{x})$. Alternately, we may be concerned with the lesser requirement that, for every $\bar{t} \in G$, the sets $Q(\bar{t}, x)$, $x \in A(\bar{t})$, $Q(\bar{t}, x) \subset E^r$, we have Properties (K) or (Q) *with respect to* x *in* $A(t)$, say, for Property (Q):

$$Q(\bar{t}, \bar{x}) = \bigcap_{\delta>0} \operatorname{cl} \operatorname{co} \bigcup_{x \in N_{\delta;\bar{t}}(\bar{x})} Q(\bar{t}, x), \tag{7}$$

where $N_{\delta;\bar{t}}(\bar{x})$ denotes the set of all $x \in A(\bar{t})$ with $|x - \bar{x}| \leqq \delta$.

Concerning the sets $U(t, x)$ and M in § 1.1, we may remark that M is the graph of $U : A \to 2^{E^m}$, and thus, if A is closed, then M is closed in $E^{\nu+n+m}$ if and only if U has Property (K) with respect to (t, x) in A. Analogously, for each $\bar{t} \in G$, the set $M(\bar{t}) = \{(x, u) | x \in A(\bar{t}), u \in U(\bar{t}, x)\}$ is the graph of $U : A(\bar{t}) \to 2^{E^m}$, and thus, if $A(\bar{t})$ is closed, then $M(\bar{t})$ is closed if and only if U has Property (K) with respect to x on $A(\bar{t})$.

There are many interrelations between the properties of the sets $U(t, x)$, M, $Q(t, x)$, in particular concerning Properties (K) and (Q). For the sake of brevity we refer to [7]–[12], [26].

Let us mention here merely that for sets Q which are equibounded, convex, and closed, Property (*) of § 1.2 (upper semicontinuity by set inclusion) implies Property (Q). Also, let us mention here that, if we think of x as an n-vector

$(x^1, \cdots, x^n)$, then Properties (Q_ρ), $0 \leqq \rho \leqq n$, have been proposed [21], [22], which are intermediate between Properties (K) and (Q), and which actually are similar to Property (K) with respect to ρ of the variables x^i and to Property (Q) with respect to the remaining variables. For the sake of brevity we refer to [16], [21], [22], [26].

1.4. A first extension of Filippov's statements. Again we assume $\nu = 1$, $G = [a, b] \subset E^1$, and A compact as in § 1.2. The following statement extends Statement 1.2.i.

STATEMENT 1.4.i. *Let* $x, x_k : [a, b] \rightarrow E^n$ *be absolutely continuous,* $\xi(t) = x'(t)$, $\xi_k(t) = x_k'(t)$, $t \in [a, b]$ *a.e.,* $u_k : [a, b] \rightarrow E^m$ *measurable for which* (3) *holds, and let* $x_k \rightarrow x$ *as* $k \rightarrow \infty$ *uniformly on* $[a, b]$ (*no further requirements on* ξ, ξ_k). *Let* A *and* M *be closed sets,* f *continuous on* M, *and the sets* $Q(t, x)$ *be all convex, closed, and satisfying Property* (Q) *with respect to* (t, x) *in* A. *Then* (4) *holds* [7].

Examples show that (4) may not hold if the sets $Q(t, x)$ do not have the required Property (Q), which, on the other hand, is not in general a consequence of the continuity of f on the closed set M. For a detailed analysis we refer to [7] and for further remarks to [8], [9], [10] and [25].

The question has been asked whether a "seminormality" requirement lesser than "the sets $Q(t, x)$ having Property (Q) with respect to (t, x) in A" would suffice in Statement 1.4.i, and for this "Property (Q) with respect to x only in $A(t)$ (for almost all t)" had been proposed. As long as we take "uniform convergence $x_k \rightarrow x$" as the sole mode of convergence, this reduction in "seminormality" requirements is not possible, as counterexamples have shown [10, p. 313]. In the proof of a number of known existence theorems, "uniform convergence $x_k \rightarrow x$" can indeed be replaced by other modes of convergences (see §§1.5–2.3 below) for which far weaker seminormality requirements, or none at all, suffice. However, in certain existence theorems "with an exceptional set" (see, e.g., [17], [28]), "uniform convergence" in small neighborhoods of the exceptional set seem to be relevant, and the statement of this section still applies.

1.5. Solutions to Problem 2 in connection with weak convergence. Again concerning Problem 2, here are a few statements which parallel Statements 1.2.i and 1.4.i in a more general context. We shall denote by $|H|$ the usual Lebesgue measure of a measurable subset H of E. Let G, $A(t)$, A, $U(t, x)$, f, M, $Q(t, x)$ be as in § 1.1, and $M(t)$ as in § 1.3. Instead of the usual requirements such as A, M closed and f continuous on M, we shall need below only the following Condition (C) (or Carathéodory-type continuity requirement):

Condition (C). For every $\varepsilon > 0$ there is a compact subset K of G such that $|G - K| < \varepsilon$, the sets $A_K = \{(t, x) \in A | t \in K\}$ and $M_K = \{(t, x, u) \in M | t \in K\}$ are closed, and $f(t, x, u)$ is continuous on M_K.

As is well known, this Condition (C) is certainly satisfied whenever $M = G \times E^n \times E^m$ and $f(t, x, u)$, defined on M, is measurable with respect to t for every $(x, u) \in E^n \times E^m$, and continuous with respect to (x, u) for every $t \in G$. Let us mention here that Condition (C) as stated implies that for almost all $\bar{t} \in G$, the sets $A(\bar{t})$ and $M(\bar{t})$ are closed and $f(\bar{t}, x, u)$ is continuous on $M(\bar{t})$.

In the present context no further relation is needed between the functions ξ_k, x_k, u_k, ξ, x besides those indicated below, or (3) of § 1.1.

STATEMENT 1.5.i. *Let $x(t)$, $x_k(t)$, $\xi(t)$, $\xi_k(t)$, $u_k(t)$, $t \in G$, $k = 1, 2, \cdots$, be measurable functions, $x, x_k : G \to E^n$, ξ, $\xi_k \in (L_1(G))^r$, $u_k : G \to E^m$, with $x_k \to x$ in measure, $\xi_k \to \xi$ weakly in $(L_1(G))^r$, for which* (3) *holds. Let us assume that G, $A(t)$, A, $U(t, x)$, f, M, $Q(t, x)$ are as in* § 1.1, *that Condition* (C) *holds, and that for almost all $\bar{t} \in G$, the sets $Q(\bar{t}, x)$, $x \in A(\bar{t})$, are closed, convex and satisfy Property* (Q) *with respect to x in $A(\bar{t})$. Then there is a measurable function $u(t)$, $t \in G$, $u : G \to E^m$ for which* (4) *holds* [13], [14].

STATEMENT 1.5.ii. *This statement is the same as Statement* 1.5.i *with $\xi_k \to \xi$ strongly in $(L_1(G))^r$ as $k \to \infty$, and for almost every $\bar{t} \in G$, the sets $Q(\bar{t}, x)$, $x \in A(\bar{t})$, are only closed and satisfy Property* (K) *in $A(\bar{t})$* [13], [14].

Remark 1. If the sets $U(t)$ depend on t only, then the differences

$$\delta_k(t) = f(t, x_k(t), u_k(t)) - f(t, x(t), u_k(t)), \qquad t \in G, \quad k = 1, 2, \cdots,$$

exist. If we know that $\delta_k \in (L_1(G))^r$, and that $\delta_k \to 0$ weakly in $(L_1(G))^r$, then Statement 1.5.i holds with the sets $Q(\bar{t}, x)$ only closed and convex (for almost all $\bar{t} \in G$) (no Condition (K) or (Q) required) [13]. This remark will be used in § 1.6.i below, and in § 1.6 we will give analytical conditions which guarantee that $\delta_k \to 0$ strongly in L_1.

STATEMENT 1.5.iii. *Let $x(t)$, $x_k(t)$, $\xi(t)$, $\xi_k(t)$, $u_k(t)$, $t \in G$, $k = 1, 2, \cdots$, be measurable functions, $x, x_k : G \to E^n$, ξ, $\xi_k \in (L_1(G))^r$, $u_k : G \to E^m$, with $x_k \to x$ in measure, $\xi_k \to \xi$ weakly in $(L_1(G))^r$, and for which* (3) *holds. Let us assume that G, $A(t)$, A, $U(t, x)$, f, M, $Q(t, x)$, are as in* § 1.1, *that Condition* (C) *holds, and that, for almost all $\bar{t} \in G$, the sets $Q(\bar{t}, x)$ are closed and convex. Also let us assume the following two conditions.*

Condition (α). *There are measurable bounded functions $p(t)$, $t \in G$, $|p(t)| \leqq \sigma$, $p : G \to E^r$, such that, for almost all $\bar{t} \in G$, we have $p(\bar{t}) \in Q(\bar{t}, x)$ for all $x \in A(\bar{t})$.*

Condition (β). *For every $N > \sigma$ and almost all $\bar{t} \in G$ the sets $Q(\bar{t}, x) \cap V(0, N)$, $x \in A(\bar{t})$, have Property* (K) *with respect to x in $A(\bar{t})$. ($V(0, N)$ is the ball of radius N around $0 \in E^r$.)*

Then there is a measurable function $u(t)$, $t \in G$, $u : G \to E^m$, for which (4) *holds* [15].

Remark 2. If for almost all $\bar{t} \in G$, and every $\bar{x} \in A(\bar{t})$, it occurs that $|f(\bar{t}, x, u)| \to +\infty$ as $|u| \to \infty$, $u \in U(\bar{t}, x)$, uniformly in some neighborhood $N_{\delta;\bar{t}}(\bar{x})$ in $A(\bar{t})$, then Condition (β) is a consequence of the remaining hypotheses. Thus, we see that under the conditions of Statement 1.5.iii and $|f(t, x, u)| \to +\infty$ as $|u| \to \infty$ as stated, no "seminormality" requirement is needed. For the simple proof of this remark, see [15], [7] and, in the present specific hypotheses, also [26].

Remark 3. A slight extension of Statement 1.5.iii can be obtained by replacing (α), (β) by the following new requirements.

Condition (α^*). There is a measurable function $p(t)$, $t \in G$, $p \in (L_1(G))^r$ such that, for almost all $\bar{t} \in G$, we have $p(\bar{t}) \in Q(\bar{t}, x)$ for all $x \in A(\bar{t})$.

Condition (β^*). For any $N \geqq 1$ sufficiently large and almost all $t \in G$, the nonempty sets $Q(\bar{t}, x) \cap V(0, q_N(\bar{t}))$, $x \in A(\bar{t})$ satisfy Property (K) with respect to x in $A(\bar{t})$. Here $q_N(t) = \max\{N, |p(t)|\}$, $t \in G$ [15], [34].

If the sets $Q(\bar{t}, x)$ are closed and convex and for almost any $\bar{t}$ and $\bar{x} \in A(\bar{t})$ it happens that $|f(\bar{t}, x, u)| \to \infty$ as $|u| \to +\infty$, $u \in U(\bar{t}, x)$, uniformly (in x) in some neighborhood $N_{\delta;\bar{t}}(\bar{x})$ of $A(\bar{t})$, then Condition (β^*) is a consequence of the remaining hypotheses, as pointed out in Remark 2 above.

1.6. Lipschitz-type and growth-type conditions on f. Recently a number of analytical conditions on f have been proposed, under which closure theorems (as well as lower closure and lower semicontinuity theorems) have been proved without seminormality conditions. In a previous paper [18] we have considered a great many of these conditions, all of the Lipschitz type or of the growth type. We have shown in [18] that all these conditions easily imply a rather general Property $(\mathfrak{D})$ on f of function-theoretical type (see below), on the basis of which the expected closure theorems without seminormality conditions (and lower closure and lower semicontinuity theorems) immediately ensue on the basis of the general statements of § 1.5. We list here a few of these Lipschitz-type and growth-type conditions, say, (F_p), (G_{pq}), (H_q) (see [18] for many others). It is convenient here to express these conditions in terms of the same sequences $x_k(t)$, $u_k(t)$, $t \in G$, $k = 1, 2, \cdots$, of § 1.1. We assume that G, $A(t)$, A, $U(t)$, M, f are as in § 1.1, with $U(t)$ depending on t only.

Condition (F_p) is a wide extension of the uniform Lipschitz condition of $f(t, x, u)$ with respect to u.

Condition (F_p). For $1 \leqq p < \infty, x, x_k \in (L_p(G))^n$, $\|x_k - x\|_p \to 0$, and

$$|f(t, x_k(t), u_k(t)) - f(t, x(t), u_k(t))| \leqq F_k(t)h(|x_k(t) - x(t)|), \quad t \in G, \quad k = 1, 2, \cdots,$$

where $h(\zeta)$, $0 \leqq \zeta < +\infty$, is a given monotone nondecreasing function with $h(0+) = 0, h(\zeta) = c\zeta^\gamma, c \geqq 0, 0 < \gamma \leqq p$, for all $\zeta \geqq \zeta_0 \geqq 0$ (c, γ, ζ_0 given constants), and $F_k(t) \geqq 0$, $t \in G$, $F_k \in L_{p'}(G)$, $k = 1, 2, \cdots$, are given functions with $p' = p/(p - \gamma)(p' = \infty$ if $\gamma = p)$, and $\|F_k\|_{p'} \leqq C$, a given constant.

For analogous conditions, (F_∞) and others, see [18]. For previous work concerning Lipschitz-type conditions in an analogous context, see E. H. Rothe [32] and L. D. Berkowitz [3].

Condition (G_{pq}) partakes of both the Lipschitz-type and growth-type conditions.

Condition (G_{pq}). For $1 \leqq p, q < \infty, x, x_k \in (L_p(G))^n$, u, $u_k \in (L_q(G))^m$, $\|x\|_p$, $\|x_k\|_p \leqq L_0$, $\|u\|_q$, $\|u_k\|_q \leqq L$, (L, L_0 given constants), $x_k \to x$ in measure in G as $k \to \infty$, and there are constants $c, c', \alpha, \beta, 0 < \alpha \leqq p, 0 < \beta \leqq q$, and a function $\psi(t) \geqq 0$, $t \in G$, $\psi \in L_1(G)$, such that for all (t, x, u), $(t, y, u) \in M$, we have

$$|f(t, x, u) - f(t, y, u)| \leqq \psi(t) + c(|x|^{p-\alpha} + |y|^{p-\alpha}) + c'|u|^{q-\beta}.$$

For analogous conditions, $(\mathrm{G}_{\infty q})$ and others, see [18]. For previous work concerning these types of conditions see F. E. Browder [5], and as mentioned, E. H. Rothe [32].

Condition (H_q) is a variant of conditions (G_{pq}).

Condition (H_q). For $1 \leqq q < \infty$, x, x_k measurable, $x_k \to x$ in measure in G as $k \to \infty$, $u, u_k \in (L_q(G))^m$, $\|u\|_q$, $\|u_k\|_q \leqq L$, a constant, and there are other constants c', β, $0 < \beta \leqq q$, and a function ψ as above such that for all (t, x, u), $(t, y, u) \in M$, we have

$$|f(t, x, u) - f(t, y, u)| \leqq \psi(t) + c'|u|^{q-\beta}.$$

For analogous conditions, (H_∞) and others, see [18]. As mentioned, all these conditions imply Property $(\mathfrak{D})$ below.

Again we assume that G, $A(t)$, A, $U(t)$, M, f are as in § 1.1, with $U(t)$ depending on t only, and we consider sequences $x, \xi, x_k, \xi_k, u_k, k = 1, 2, \cdots$, as in Statement

1.5.i satisfying (3) or

$$x(t), x_k(t) \in A(t), \quad u_k(t) \in U(t), \quad \xi_k(t) = f(t, x_k(t), u_k(t)), \qquad t \in G, \quad k = 1, 2, \cdots.$$

Since $U(t)$ depends on t only, also the function $\bar{\xi}_k(t) = f(t, x(t), u_k(t))$, $t \in G$, $k = 1, 2, \cdots$, are defined on G. Under the sole Carathéodory-type Condition (C) all functions ξ_k and $\bar{\xi}_k$ are measurable in G, and therefore so are the differences

$$\delta_k(t) = f(t, x_k(t), u_k(t)) - f(t, x(t), u_k(t)), \qquad t \in G, \quad k = 1, 2, \cdots.$$

We have proved in [18] that all of Conditions (F_p), (G_{pq}), (H_q) (and the others considered in [18]) imply that $\delta_k \in (L_1(G))^r$ (or $\delta_k \in (L_p(G))^r$ for some $p \geqq 1$), and that

$$\lim_{k \to \infty} \|\delta_k\|_1 = 0. \tag{8}$$

Whenever (8) holds, we say that Property ($\mathfrak{D}$) holds for the given sequence. We have now the following further closure theorem without seminormality conditions.

STATEMENT 1.6.i. *Let $x(t), x_k(t), \xi(t), \xi_k(t), u_k(t), t \in G, k = 1, 2, \cdots$, be measurable functions, $x, x_k : G \to E^n$, ξ, $\xi_k \in (L_1(G))^r$, $u_k : G \to E^m$, with $\xi_k \to \xi$ weakly in $(L_1(G))^r$, for which* (3) *holds. Let us assume that G, $A(t)$, A, $U(t)$, f, M, $Q(t, x)$ are as in* § 1.1, *that $U(t)$ depends on t only, that Condition* (C) *holds, and that, for almost all $\bar{t} \in G$, the sets $Q(\bar{t}, x)$, $x \in A(\bar{t})$, are closed and convex for every $x \in A(\bar{t})$. Let one of the following conditions hold*:

(a) *For some p, $1 \leqq p \leqq \infty$, x, $x_k \in (L_p(G))^n$, and Condition* (F_p) *holds.*

(b) *For some p, q, $1 \leqq p, q \leqq \infty$, x, $x_k \in (L_p(G))^n$, $u_k \in (L_q(G))^m$, and Condition* (G_{pq}) *holds.*

(c) *For some q, $1 \leqq q \leqq \infty$, $u_k \in (L_q(G))^m$ and Condition* (H_q) *holds.*

Then there is a measurable function $u(t)$, $t \in G$, $u: G \to E^m$, for which (4) *holds.*

This statement is a corollary of Statement 1.5.i and of Remark 1 of § 1.5, proved in [13]. The same conclusion of Statement 1.6.i of course holds if we only know that $x_k \to x$ in measure and Property ($\mathfrak{D}$) holds.

1.7. Property ($\mathfrak{D}$) implies a weak form of Property (Q).

Property (Q) with respect to x as defined by (7) has been used recently by many authors. Some (see, e.g., M. F. Bidaut [4], C. Castaing [6]) have formulated it in the equivalent form:

(9) For every sequence of points x_s, $s = 1, 2, \cdots$, with $x_s \in A(\bar{t})$, $x_s \to \bar{x}$ as $s \to \infty$, we have

$$Q(\bar{t}, \bar{x}) \supset \bigcap_{h=1}^{\infty} \operatorname{cl} \operatorname{co} \left\{ \bigcup_{s=h}^{\infty} Q(\bar{t}, x_s) \right\}.$$

Another equivalent form is as follows:

(10) For every sequence of points (x_s, z_s), $s = 1, 2, \cdots$, with $x_s \in A(\bar{t})$, $x_s \to \bar{x}$ as $s \to \infty$, $z_s \in Q(\bar{t}, x_s)$, we have

$$Q(\bar{t}, \bar{x}) \supset \bigcap_{h=1}^{\infty} \operatorname{cl} \operatorname{co} \left\{ \bigcup_{s=h}^{\infty} z_s \right\}.$$

Keeping in mind the definition of the sets $Q(t, x)$ we see that an equivalent definition of Property (Q) with respect to x at $(\bar{t}, \bar{x})$ is as follows:

(11) For every sequence of points (x_s, u_s), $s = 1, 2, \cdots$, $x_s \in A(\bar{t})$, $u_s \in U(\bar{t}, x_s)$, $x_s \to \bar{x}$, we have

$$Q(\bar{t}, \bar{x}) \supset \bigcap_{h=1}^{\infty} \operatorname{cl} \operatorname{co} \left\{ \bigcup_{s=h}^{\infty} f(\bar{t}, x_s, u_s) \right\}.$$

In [13] the remark was also made that for the proof of Statement 1.5.i, the full force of Property (Q) with respect to x is not needed. We know already from real analysis that there is a subsequence $k_s, s = 1, 2, \cdots$, such that $x_{k_s}(t) \to x(t)$ pointwise a.e. in G as $s \to \infty$. All we need for the proof of Statement 1.5.i is the following.

(12) We can select a further subsequence, say still k_s for the sake of simplicity, such that, for almost all $t \in G$, we have

$$Q(t, x(t)) \supset \bigcap_{h=1}^{\infty} \operatorname{cl} \operatorname{co} \left\{ \bigcup_{s=h}^{\infty} f(t, x_{k_s}(t), u_{k_s}(t)) \right\}.$$

We are now in a position to prove that Property ($\mathfrak{D}$) implies the weak form of Property (Q) we have mentioned in (12).

STATEMENT 1.7.i. *If the sets $U(t)$ depend on t only, if Condition* (C) *holds, if $x(t)$, $x_k(t)$, $u_k(t)$, $t \in G$, $k = 1, 2, \cdots$, are measurable functions, if the sets $Q(t, x(t))$ are closed and convex for almost all $t \in G$, if $\delta_k \to 0$ strongly in $(L_1(G))^r$, $x_k(t) \to x(t)$ in measure in G as $k \to \infty$, then* (12) *holds, that is, there is a subsequence $k_s, s = 1, 2, \cdots$, such that $x_{k_s}(t) \to x(t)$ pointwise a.e. in G as $s \to \infty$, and, for almost all $t \in G$, we also have*

$$Q(t, x(t)) \supset \bigcap_{h=1}^{\infty} \operatorname{cl} \operatorname{co} \left\{ \bigcup_{s=h}^{\infty} f(t, x_{k_s}(t), u_{k_s}(t)) \right\}.$$

Proof. Here $\|\delta_k\|_1 \to 0$, hence $\delta_k(t) \to 0$ in measure, and there is a subsequence $[k_s]$ such that $\delta_{k_s}(t) \to 0$ pointwise a.e. in G as $s \to \infty$. Let T denote the set of measure zero in G where this does not occur, or the sets $Q(t, x(t))$ are not closed and convex. For every $t \in G - T$ we have, then, $\delta_{k_s}(t) \to 0$ as $t \to \infty$, or

$$f(t, x_{k_s}(t), u_{k_s}(t)) - f(t, x(t), u_{k_s}(t)) \to 0 \quad \text{as } s \to \infty.$$

Let t be any fixed point in $G - T$. Given $\varepsilon > 0$, there is some $\bar{s} = \bar{s}(\varepsilon, t) > 0$, such that for all $s \geqq \bar{s}$ we have

$$|\delta_{k_s}(t)| = |f(t, x_{k_s}(t), u_{k_s}(t)) - f(t, x(t), u_{k_s}(t))| < \varepsilon.$$

We shall take $\bar{s} = \bar{s}(\varepsilon, t) > 1/\varepsilon$. Note that the sets $U(t)$ depend on t only; hence, $u_k(t) \in U(t)$, and $f(t, x(t), u_{k_s}(t)) \in Q(t, x(t))$, $t \in G - T$. First, for the chosen t and all $s \geqq \bar{s}$, we have then

$$f(t, x_{k_s}(t), u_{k_s}(t)) \in [Q(t, x(t))]_\varepsilon,$$

where $[\cdot]_\varepsilon$ denotes the closed ε-neighborhood of $Q(t, x(t))$. Since $t \in G - T$, the set $Q(t, x(t))$ is closed and convex; hence, the set $[Q(t, x(t))]_\varepsilon$ is also closed and convex, or

$$\operatorname{cl} \operatorname{co} [Q(t, x(t))]_\varepsilon = [Q(t, x(t))]_\varepsilon,$$

and, in addition,

$$Q(t, x(t)) = \bigcap_{\varepsilon > 0} [Q(t, x(t))]_\varepsilon.$$

Thus

$$\operatorname{cl} \operatorname{co} \left\{ \bigcup_{s=\bar{s}}^{\infty} f(t, x_{k_s}(t), u_{k_s}(t)) \right\} \subset \operatorname{cl} \operatorname{co} [Q(t, x(t))]_\varepsilon = [Q(t, x(t))]_\varepsilon.$$

As $\varepsilon \to 0$ we certainly have $\bar{s} = \bar{s}(\varepsilon, t) \to \infty$, and

$$\bigcap_{h=1}^{\infty} \operatorname{cl} \operatorname{co} \left\{ \bigcup_{s=h}^{\infty} f(t, x_{k_s}(t), u_{k_s}(t)) \right\} \subset \bigcap_{\varepsilon > 0} [Q(t, x(t))]_\varepsilon = Q(t, x(t)).$$

This relation holds for all $t \in G - T$, that is, almost everywhere in G. Statement 1.7.i is thereby proved.

Remark 4. Since in Statement 1.7.i the only information we have on the behavior of the sets $Q(t, x)$ is the one we can derive from the fact that $\|\delta_k\|_1 \to 0$, it is clear that the form (12) of Property (Q) is the best we can derive. As mentioned, this form (12) is still strong enough so that the argument for original theorem 1.5.i can be applied. Suppose now that there is a set T of measure zero on G such that for every $(\bar{t}, x)$, $\bar{t} \in G - T$, $x \in A(\bar{t})$, the set $Q(\bar{t}, x)$ is closed and convex, and in addition, $\lim_{x \to \bar{x}} f(\bar{t}, x, u) = f(\bar{t}, \bar{x}, u)$ uniformly with respect to u. This is certainly the case if f satisfies a Lipschitz condition in x uniformly with respect to u.

Then Property (Q) holds at $(\bar{t}, \bar{x})$ in the usual form. Indeed, given $\varepsilon > 0$ we can find δ such that $|f(\bar{t}, x, u) - f(\bar{t}, \bar{x}, u)| < \varepsilon$ for all $x \in A(t)$, $|x - \bar{x}| \leqq \delta$, and all $u \in U(t)$. Then, for $x \in N_{\delta;\bar{t}}(\bar{x})$ we have $Q(\bar{t}, x) \subset [Q(\bar{t}, \bar{x})]_\varepsilon$, and, as before,

$$\bigcap_{\varepsilon > 0} \operatorname{cl} \operatorname{co} \bigcup_{x \in N_{\delta;\bar{t}}(\bar{x})} Q(\bar{t}, x) \subset \bigcap_{\varepsilon > 0} \operatorname{cl} \operatorname{co} [Q(t, x)]_\varepsilon = Q(t', x).$$

Remark 5. Statement 1.7.i holds even in situations where $U(t, x)$ depends on both t and x. Here are some cases of interest (see [34] for details).

Case (A). There is a subsequence $[k_\alpha]$ such that $u_{k_\alpha}(t) \in U(t, x(t))$, $t \in G$ a.e., $\alpha = 1, 2, \cdots$. The argument is now as for Statement 1.4.i starting with the sequence $[k_\alpha]$ instead of $[k]$.

Case (B). Let $f(t, x, u)$ be defined in $A \times E_m$, though $U(t, x)$ is still a subset of E_m which may depend on t and x. Let us assume two things: (B1) for every $t \in G - T$ and $\varepsilon > 0$ there is an $\eta > 0$ such that $u, \bar{u} \in E_m$, $|u - \bar{u}| \leqq \eta$ implies $|f(t, x(t), u) - f(t, x(t), \bar{u})| < \varepsilon$; (B2) for every $t \in G - T$ and $\eta > 0$ there is $\sigma > 0$ such that $x \in A(t)$, $|x - x(t)| \leqq \sigma$ implies $U(t, x) \subset [U(t, x(t))]_\eta$.

Indeed, in the proof of Statement 1.7.i we have considered any fixed $t \in G - T$ and any given $\varepsilon > 0$. Let us now determine η as in (B1), and then σ as in (B2). Also, let us take $\bar{s} = \bar{s}(t, \varepsilon, \sigma)$ so that $s \geqq \bar{s}$ implies $|x_{k_s}(t) - x(t)| \leqq \sigma$. Then, $|\delta_{k_s}(t)| \leqq \varepsilon$ and $u_{k_s}(t) \in U(t, x_{k_s}(t)) \subset [U(t, x(t))]_\eta$. Thus, there is some $\bar{u} \in U(t, x(t))$ such that $|\bar{u} - u_{k_s}(t)| \leqq \eta$, and finally

$$|\delta'_{k_s}(t)| = |f(t, x(t), u_{k_s}(t)) - f(t, x(t), \bar{u})| \leqq \varepsilon.$$

Then

$$|f(t, x_{k_s}(t), u_{k_s}(t)) - f(t, x(t), \bar{u})| \leqq |\delta_{k_s}(t)| + |\delta'_{k_s}(t)| \leqq 2\varepsilon,$$

and

$$f(t, x_{k_s}(t), u_{k_s}(t)) \in [Q(t, x(t))]_{2\varepsilon}.$$

The proof of Statement 1.7.i can now be repeated without further changes.

Note that (B1) is the uniform continuity of $f(t, x(t), u)$, $u \in E_m$, as a function of u in E_m for any fixed $t \in G - T$, and that (B2) is the metric upper semicontinuity with respect to x only of the sets $U(t, x)$ at $(t, x(t))$.

Case (C). Let us assume that $f(t, x, u)$ is still defined on $M = \{(t, x, u)|(t, x) \in A, u \in U(t, x)\}$ and that, for every $t \in G - T$ and $\varepsilon > 0$, there is $\delta > 0$ such that for $x \in A(t)$, $|x - x(t)| \leqq \delta$, $u \in U(t, x)$, there is some $\bar{u} \in U(t, x(t))$ such that $|f(t, x, u) - f(t, x(t), \bar{u})| \leqq \varepsilon$. The argument is the same as above.

Remark 6. Note that in Case (B) above the metric upper semicontinuity with respect to x of the sets $U(t, x)$ cannot be replaced by Kuratowski's upper semicontinuity, as the following example shows.

Example. Take $\nu = n = m = r = 1$, $f(t, x, u) = u$, $0 \leqq t \leqq 1$, $-1 \leqq x \leqq 1$, $U(t, x) = \{0\}$ if $x = 0$, $U(t, x) = \{u \geqq x^{-1}\}$ if $0 < x \leqq 1$, $U(t, x) = \{u \leqq x^{-1}\}$ if $-1 \leqq x < 0$. The sets $U(t, x)$ are identical to the sets $Q(t, x)$. Property (Q) with respect to x is not satisfied at the points $(t, 0)$, $0 \leqq t \leqq 1$.

2. Abstract Lagrange problems.

2.1. Notations. Let G be a measurable subset of the t-space E^ν of finite measure, $t = (t^1, \cdots, t^\nu)$, for every $t \in G$ let $A(t)$ be a given nonempty subset of the x-space E^n, $x = (x^1, \cdots, x^n)$, and let $A = \{(t, x)|t \in G, x \in A(t)\}$. For every $(t, x) \in A$ let $\tilde{\tilde{Q}}(t, x)$ be a given subset of the $\tilde{z}$-space E^{r+1}, $\tilde{z} = (z^0, z) = (z^0, z^1, \cdots, z^r)$.

Problem 1$'$. We consider here a sequence of measurable functions $x(t)$, $\xi(t)$, $x_k(t)$, $\xi_k(t)$, $\eta_k(t)$, $t \in G$, $k = 1, 2, \cdots$, $x, x_k : G \to E^n$, $\xi, \xi_k : G \to E^r$, $\eta_k : G \to E^1$, $\eta_k \in L_1(G)$, with

$$x_k(t) \in A(t), \qquad (\eta_k(t), \xi_k(t)) \in \tilde{\tilde{Q}}(t, x_k(t)), \tag{13}$$
$$t \in G \text{ a.e.}, \quad k = 1, 2, \cdots.$$

We define i, $-\infty \leqq i \leqq +\infty$, by taking

$$i = \liminf_{k \to \infty} \int_G \eta_k(t)\, dt, \tag{14}$$

and we denote by $Q(t, x) \subset E^r$ the projection of the sets $\tilde{\tilde{Q}}(t, x)$ on the z-space E^r. We seek "modes of convergence" $x_k \to x$, $\xi_k \to \xi$, and conditions on $A(t)$, $\tilde{\tilde{Q}}(t, x)$, $Q(t, x)$, which imply that i is finite, and there is a measurable function $\eta(t)$, $t \in G$, $\eta \in L_1(G)$ such that

$$x(t) \in A(t), \quad (\eta(t), \xi(t)) \in \tilde{\tilde{Q}}(t, x(t)), \qquad t \in G \text{ a.e.},$$
$$\int_G \eta(t)\, dt \leqq i. \tag{15}$$

We assume here that for every $(t, x) \in A$, a subset $U(t, x)$ of the u-space E^m is given, we denote by M the set $\{(t, x, u)|(t, x) \in A, u \in U(t, x)\}$, we consider functions $f_0(t, x, u)$, $f(t, x, u) = (f_1, \cdots, f_r)$ defined on M, and for every $(t, x) \in A$, we denote by $\tilde{\tilde{Q}}(t, x)$ the sets

$$\tilde{\tilde{Q}}(t, x) = \{(z^0, z)|(z^0 \geqq f_0(t, x, u), z = f(t, x, u), u \in U(t, x)\} \subset E^{r+1}.$$

Then the projection $Q(t, x)$ of $\tilde{Q}(t, x)$ on the z-space E^r is the set $Q(t, x) = f(t, x, U(t, x)) = \{z | z = f(t, x, u), u \in U(t, x)\}$.

Problem 2′. We consider here a sequence of measurable functions $x(t), \xi(t), x_k(t), \xi_k(t), \eta_k(t), u_k(t), t \in G, k = 1, 2, \cdots, x, x_k: G \to E^n, \xi, \xi_k: G \to E^r, \eta_k \in L_1(G), u_k: G \to E^m$, with

$$(16)\qquad \begin{aligned} &x_k(t) \in A(t), \quad u_k(t) \in U(t, x_k(t)), \quad \xi_k(t) = f(t, x_k(t), u_k(t)), \\ &\eta_k(t) = f_0(t, x_k(t), u_k(t)), \qquad\qquad t \in G, \quad k = 1, 2, \cdots, \end{aligned}$$

and we define i as above, $-\infty \leqq i \leqq +\infty$. We seek "modes of convergence" $x_k \to x$, $\xi_k \to \xi$, and conditions on $A(t)$, $U(t, x)$, M, $\tilde{Q}(t, x)$, $Q(t, x)$, f_0 and f, which imply that there is a measurable $u: G \to E^m$, such that

$$(17)\qquad x(t) \in A(t), \quad u(t) \in U(t, x(t)), \quad \xi(t) = f(t, x(t), u(t)), \quad t \in G \text{ a.e.},$$

$$(18)\qquad \eta \in L_1(G), \quad \int_G \eta(t)\, dt \leqq i,$$

where $\eta(t) = f_0(t, x(t), u(t)), t \in G$.

These are the two main forms of the so-called *lower closure* problem in abstract Lagrange problems. The same considerations at the end of § 1.1 hold also in the present situation.

Again, for the sake of simplicity, we shall refer below mainly to Problem 2′.

For $v = 1$, $G = [a, b]$, A compact, a Filippov-type solution to Problem 2′ is well represented by the following statement.

STATEMENT 2.1.i. *Let $x, x_k: [a, b] \to E^n$ be absolutely continuous functions, let $\xi(t) = x'(t)$, $\xi_k(t) = x_k'(t)$, $t \in [a, b]$ a.e., $u_k: [a, b] \to E^m$, measurable, for which* (16) *holds, and $x_k \to x$ as $k \to \infty$ uniformly on $[a, b]$ (no further requirements on ξ, ξ_k). Let us assume that A, M are compact sets, f_0, f are continuous on M, and the sets $\tilde{Q}(t, x)$ are all convex. Then, i is finite, and there is $u: [a, b] \to E^m$ measurable for which* (17), (18) *hold.*

Under the conditions of Statement 2.1.i, say A and M compact and f_0, f continuous on M, it is not restrictive to replace $\tilde{Q}(t, x)$ by $\tilde{Q}_L(t, x) = [(z^0, z) \in \tilde{Q}(t, x) | z^0 \leqq L]$, where L is the maximum of f_0 on M. Then the sets $Q(t, x)$ and $\tilde{Q}_L(t, x)$ are compact, equibounded, and certainly are "metric upper semicontinuous" on A.

2.2. A first extension of Filippov's statements for Problem 2′. Again we assume $v = 1, G = [a, b] \subset E^1$, and A compact as in § 1.2. The following statements extend Statement 2.1.i.

STATEMENT 2.2.i. *Let $x, x_k: [a, b] \to E^n$ be absolutely continuous functions, let $\xi(t) = x'(t)$, $\xi_k(t) = x_k'(t)$, $t \in [a, b]$ a.e., $u_k: [a, b] \to E^m$ measurable, $\eta_k \in L_1([a, b])$, for which* (16) *holds with $i < +\infty$, and $x_k \to x$ as $k \to \infty$ uniformly on $[a, b]$ (no further requirements on ξ, ξ_k, η_k). Let A and M be closed sets, f_0 and f continuous on M, and the sets $\tilde{Q}(t, x)$ all convex, closed, and satisfying Property* (Q) *with respect to (t, x) in A. Let us assume that there is a constant $\gamma \geqq 0$ and a function $\psi(t) \geqq 0$, $t \in G$, $\psi \in L_1[a, b]$, such that $f_0(t, x, u) \geqq -\psi(t) - \gamma |f(t, x, u)|$ for all $(t, x, u) \in M$. Then i is finite, and there is a measurable $u: [a, b] \to E^m$ for which* (17), (18) *hold* [7].

2.3. Solutions of Problem 2′ in connection with weak convergence. We assume here as in § 1.5 that $v \geqq 1$, and that G is a given measurable subset of the t-space E^v of finite measure. The following statements parallel those in § 1.6.

STATEMENT 2.3.i. *Let $x(t)$, $x_k(t)$, $\xi(t)$, $\xi_k(t)$, $\eta_k(t)$, $u_k(t)$, $t \in G$, $k = 1, 2, \cdots$, be measurable functions, $x, x_k : G \to E^n$, $\xi, \xi_k \in (L_1(G))^r$, $\eta_k \in L_1(G)$, $u_k : G \to E^m$ measurable, for which* (16) *holds with $i < +\infty$, and $x_k \to x$ in measure in G, $\xi_k \to \xi$ weakly in $(L_1(G))^r$, and let us assume that there are also functions $\lambda, \lambda_k \in L_1(G)$, $k = 1, 2, \cdots$, with $\eta_k(t) \geqq \lambda_k(t)$, $\lambda_k \to \lambda$ weakly in $L_1(G)$ as $k \to \infty$. Also, let us assume that Condition* (C) *holds* (*for f and f_0*) *and that, for almost all $\bar{t} \in G$, the sets $\tilde{\tilde{Q}}(\bar{t}, x)$, $x \in A(\bar{t})$, are convex and have Property* (Q) *with respect to x in $A(\bar{t})$. Then i is finite, and there are measurable functions $u(t)$, $\eta^*(t)$, $t \in G$, $\eta^* \in L_1(G)$, such that* (17) *holds, $\lambda(t) \leqq \eta^*(t)$, $\eta(t) = f_0(t, x(t), u(t)) \leqq \eta^*(t)$ and $\int_G \eta^*(t)\,dt \leqq i$* [13].

Under a variety of conditions, we can guarantee that, in addition, (18) holds, that is, $\eta \in L_1(G)$, and hence $\int_G \eta(t)\,dt \leqq i$. This is certainly the case if any one of the following hypotheses holds, under each of which there are also natural choices for the functions λ_k above [13].

Hypothesis 1. There is a constant $\gamma \geqq 0$ and a function $\psi(t) \geqq 0$, $t \in G$, $\psi \in L_1(G)$, such that $f_0(t, x, u) \geqq -\psi(t) - \gamma|f(t, x, u)|$ for all $(t, x, u) \in M$.

Under this hypothesis we may well take $\lambda_k(t) = -\psi(t) - \gamma|\xi_k(t)|$, $t \in G$. Since $\xi_k \to \xi$ weakly in $(L_1(G))^r$, the functions ξ_k, as well as $|\xi_k|$, are equiabsolutely integrable, and then there is some $\omega(t) \geqq 0, t \in G, \omega \in L_1(G)$, and a subsequence, say still $\{k\}$, such that $|\xi_k| \to \omega$ (and $\lambda_k \to -\psi - \gamma\omega$) weakly in $L_1(G)$ as $k \to \infty$. Then $\lambda(t) \leqq \eta(t) \leqq \eta^*(t), t \in G, \lambda, \eta, \eta^* \in L_1(G)$.

Hypothesis 2. x, $x_k \in (L_p(G))^n$, $\|x_k - x\|_p \to 0$ for some p, $1 \leqq p \leqq +\infty$, ξ, $\xi_k \in (L_1(G))^r$, $\xi_k \to \xi$ weakly in $(L_1(G))^r$, and there are constants γ, $\gamma' \geqq 0$ and a function $\psi(t) \geqq 0$, $t \in G$, $\psi \in L_1(G)$, such that $f_0(t, x, u) \geqq -\psi(t) - \gamma|f(t, x, u)| - \gamma'|x|^p$ for all $(t, x, u) \in M$.

Under this hypothesis we may well take $\lambda_k(t) = -\psi(t) - \gamma|\xi_k(t)| - \gamma'|x_k(t)|^p$. With the same notations and conventions as above, we have, then, $\lambda(t) = -\psi(t) - \gamma\omega(t) - \gamma'|x(t)|$, $t \in G$, and again $\lambda(t) \leqq \eta(t) \leqq \eta^*(t)$, $t \in G$, λ, η, $\eta^* \in L_1(G)$.

Hypothesis 3. $x_k \in (L_\infty(G))^n$, $\|x_k\|_\infty \leqq L_1$, $\xi_k \in (L_\infty(G))^r$, $\|\xi_k\|_\infty \leqq L_2$ for given constants L_1, L_2, and there are a function $\psi(t) \geqq 0$, $t \in G$, $\psi \in L_1(G)$, and a monotone nondecreasing function $\sigma(\zeta)$, $0 \leqq \zeta < +\infty$, such that $f_0(t, x, u) \geqq -\psi(t)\sigma[|x| + |f(t, x, u)|]$ for all $(t, x, u) \in M$.

Under this hypothesis we may well take in Statement 2.3.i, $\lambda_k(t) = \lambda(t) = -\psi(t)\sigma(L_1 + L_2)$, $t \in G$, and then again $\lambda(t) \leqq \eta(t) \leqq \eta^*(t)$, $t \in G$, $\lambda, \eta, \eta^* \in L_1(G)$.

For other conditions, examples, and remarks we refer to [13].

STATEMENT 2.3.ii. This is the same as Statement 2.3.i with $\xi_k \to \xi$ strongly in $(L_1(G))^r$ as $k \to \infty$, and, for almost every $\bar{t} \in G$, the sets $\tilde{\tilde{Q}}(\bar{t}, x)$, $x \in A(\bar{t})$, are only closed and satisfy Property (K) in $A(\bar{t})$ [13], [14].

Remark 7. Here, as in § 1.5, if the sets $U(t)$ depend on t only, then the differences

$$\delta_k(t) = f(t, x_k(t), u_k(t)) - f(t, x(t), u_k(t)),$$
$$\delta_k^0(t) = f_0(t, x_k(t), u_k(t)) - f_0(t, x(t), u_k(t)), \qquad t \in G, \quad k = 1, 2, \cdots,$$

exist. If we know that $\delta_k \in (L_1(G))^r$, $\delta_k^0 \in L_1(G)$, that $\delta_k \to 0$ weakly in $(L_1(G))^r$, and that $\limsup \int_G \delta_k^0(t)\,dt = 0$ as $k \to \infty$, then Statement 2.3.i holds with the sets $\tilde{\tilde{Q}}(\bar{t}, x)$ only closed and convex (for almost all $\bar{t} \in G$) (no Conditions (K) or (Q) required) [13]. This remark will be used below.

To state the next lower closure theorem, we use the same notations as above. For any $(t, x) \in A$ and corresponding set $\tilde{Q}(t, x) \subset E^{r+1}$, we consider the projection $Q(t, x)$ of $\tilde{Q}(t, x)$ on the z-space E^r, and the real-valued function

$$T(z; t, x) = \inf \{z^0 | (z^0, z) \in \tilde{Q}(t, x)\}, \qquad z \in Q(t, x).$$

In other words, for every $(t, x) \in A$ and $z \in E^r$, $z \in Q(t, x)$, we consider the set $\{u\}$ of all $u \in U(t, x)$ for which $z \in f(t, x, u)$, and then $T(z; t, x) = \inf \{f_0(t, x, u), u \in \{u\}\}$. Note that $-\infty \leqq T(z; t, x) < +\infty$ for every $z \in Q(t, x)$. If $\tilde{Q}(t, x)$ is convex, so is $Q(t, x)$, and in [9] it was proved that either $T(z: t, x) = -\infty$ for all $z \in \operatorname{Rint} Q(t, x)$, or $T(z; t, x) > -\infty$ for all $z \in Q(t, x)$. In the latter case, $T(z; t, x)$ is a real-valued convex function on the convex set $Q(t, x) \subset E^r$. (Here Rint Q denotes the set of all points z which are interior to Q with respect to the hyperspace of minimum dimension containing Q.)

We shall denote by $V(0, R)$ the closed ball in E^r of center the origin and radius R.

STATEMENT 2.3.iii. *Let $x(t)$, $x_k(t)$, $\xi(t)$, $\xi_k(t)$, $\eta_k(t)$, $u_k(t)$, $t \in G$, $k = 1, 2, \cdots$, be measurable functions, $x, x_k: G \to E^n$, ξ, $\xi_k \in (L_1(G))^r$, $\eta_k \in L_1(G)$, $u_k: G \to E^m$ measurable, for which* (16) *holds with $i < +\infty$, and $x_k \to x$ in measure in G, $\xi_k \to \xi$ weakly in $(L_1(G))^r$. Let us assume that Property* (C) *holds* (*for both f_0 and f*), *and that for almost all $\bar{t} \in G$ and all $x \in A(\bar{t})$ the sets $\tilde{Q}(t, x)$ are convex and closed. Let us assume these further conditions.*

Condition (α). *There is a measurable bounded function $p(t)$, $t \in G$, $|p(t)| \leqq \sigma$, $p: G \to E^r$, such that $p(t) \in Q(t, x)$ for $t \in G$ and all $x \in A(t)$.*

Condition (β). *There is a constant $c \geqq 0$ such that $|T(z; t, x)| \leqq c$ for almost all $t \in G$, all $x \in A(t)$, and $z \in Q(t, x) \cap V(0, 2\sigma)$.*

Condition (γ). *For every $N > \sigma$ and almost all $t \in G$, the sets $\tilde{Q}(t, x) \cap [E^1 \times V(0, N)]$ satisfy Property* (K) *with respect to x in $A(t)$. Then there is a measurable function $u: G \to E^m$ for which* (17), (18) *hold.*

Remark 8. Condition (γ) is certainly verified if it happens that for almost every $\bar{t} \in G$ and every $\bar{x} \in A(\bar{t})$, we have $|f(\bar{t}, \bar{x}, u)| \to +\infty$ as $|u| \to \infty$, $u \in U(\bar{t}, \bar{x})$, uniformly in some neighborhood $N_{\delta;\bar{t}}(\bar{x})$ of $\bar{x}$ in $A(\bar{t})$. We see, therefore, that under very mild conditions on f and f_0, no "seminormality" conditions are needed.

Remark 9. A slight extension of Statement 2.3.iii can be obtained by replacing Conditions (α), (β), (γ) by the following new requirements:

(a) There is a measurable function $p(t)$, $t \in G$, $p: G \to E^r$, such that $p(t) \in Q(t, x)$ for almost all $t \in G$ and all $x \in A(t)$.

(b) There is a measurable function $\rho(t) \geqq 0$, $t \in G$, $\rho \in L_1(G)$, such that $|T(z; t, x)| \leqq \rho(t)$ for almost all $t \in G$, all $x \in A(t)$, and all $z \in Q(t, x) \cap V(0, 2\sigma(t))$, where $\sigma(t) = \max \{N_0, |p(t)|\}$ for some constant $N_0 > 0$.

(c) For all $N \geqq N_0$ and almost all $t \in G$ the sets $\tilde{Q}(t, x) \cap [E^1 \times V(0, \sigma_N(t))]$ have Property (K) with respect to x in $A(t)$, where $\sigma_N(t) = \max \{N, |p(t)|\}$.

(d) For some p, q, $1 \leqq p$, $q \leqq \infty$, $1/p + 1/q = 1$, the function $\rho(t)/\sigma(t)$ is of class $L_q(G)$, while ξ, $\xi_k \in (L_p(G))^r$, $k = 1, 2, \cdots$, $\xi_k \to \xi$ weakly in $(L_p(G))^r$ [15], [34].

Finally, as in § 1.7, and by the use of the analytical Conditions (F_p), (G_{pq}), (H_q) on f_0 and f, we can state another lower closure theorem without seminormality conditions, which parallels the closure theorem 1.6.i.

STATEMENT 2.3.iv. *Let $x(t)$, $x_k(t)$, $\xi(t)$, $\xi_k(t)$, $\eta_k(t)$, $u_k(t)$, $t \in G$, $k = 1, 2, \cdots$, be measurable functions, $x, x_k : G \to E^n$, $\xi, \xi_k \in (L_1(G))^r$, $\eta_k \in L_1(G)$, $u_k : G \to E^m$, for which* (16) *holds, $\xi_k \to \xi$ weakly in $(L_1(G))^r$, and $-\infty < i < +\infty$. Let us assume that G, $A(t)$, A, $U(t)$, f_0, f, M, $\tilde{Q}(t, x)$ are as in* § 2.1 (*Problem* 2'), *that $U(t)$ depends on t only, that Condition* (C) *holds* (*for both f_0 and f*), *and that for almost all $\bar{t} \in G$, the sets $\tilde{Q}(\bar{t}, x)$, $x \in A(\bar{t})$, are closed and convex for every $x \in A(\bar{t})$. Let us assume that one of the following conditions holds*:

(a) *For some p, $1 \leqq p \leqq \infty$, x, $x_k \in (L_p(G))^n$, and Condition* (F_p) *holds* (*for both f_0 and f*).

(b) *For some p, q, $1 \leqq p, q \leqq \infty$, $x, x_k \in (L_p(G))^n$, $u_k \in (L_q(G))^m$, and Condition* (G_{pq}) *holds* (*for both f_0 and f*).

(c) *For some q, $1 \leqq q \leqq \infty$, $u_k \in (L_q(G))^m$, and Condition* (H_q) *holds* (*for both f_0 and f*).

Then i is finite and there is a measurable function $u(t)$, $t \in G$, $u : G \to E^m$, such that (17) *and* (18) *hold with $\eta(t) = f_0(t, x(t), u(t))$, $t \in G$. In particular, $\eta \in L_1(G)$, and $\int_G \eta(t)\, dt \leqq i$.*

This statement 2.3.iv is a corollary of Statement 2.3.i and Remark 7 proved in [13].

3. Lower semicontinuity theorems for free problems.

3.1. A necessary and sufficient condition for lower semicontinuity. Here we take $v = 1$, $G = [a, b] \subset E^1$. For every $t \in [a, b]$, then, $A(t)$ is a given subset of the x-space E^n; we take $A = \{(t, x) | a \leqq t \leqq b, x \in A(t)\} \subset E^{n+1}$, and $M = A \times E^n \subset E^{2n+1}$. We assume A (and M) closed, and $f_0(t, x, u)$ is a continuous function on M.

We denote by $\mathscr{T}$ the class of all functions $x(t) = (x^1, \cdots, x^n)$, $a \leqq t \leqq b$, x absolutely continuous, with $x(t) \in A(t)$, $t \in [a, b]$, and $f_0(t, x(t), x'(t))$ of class $L_1([a, b])$. Then for $x \in \mathscr{T}$ the functional $I[x]$ is defined by

$$I[x] = \int_a^b f_0(t, x(t), x'(t))\, dt.$$

For elements $x, x_k \in \mathscr{T}$, $k = 1, 2, \cdots$, we choose the "mode of convergence $x_k \to x$" defined by $x_k \to x$ uniformly on $[a, b]$ and $x'_k \to x'$ weakly in $L_1([a, b])$. As an application of the Dunford–Pettis theorem on weak convergence in L_1, it is easy to prove that, if $x, x_k \in \mathscr{T}$, $k = 1, 2, \cdots$, and $x'_k \to x'$ weakly in $L_1([a, b])$, and $x_k(\bar{t}) \to x(\bar{t})$ pointwise at at least one point $\bar{t} \in [a, b]$, then $x_k(t) \to x(t)$ pointwise at every $t \in [a, b]$, and actually $x_k \to x$ uniformly in $[a, b]$ as $k \to \infty$.

We say that $I[x]$ is lower semicontinuous in $\mathscr{T}$ provided for all $x, x_k \in \mathscr{T}$ with $x_k \to x$ uniformly in $[a, b]$ and $x'_k \to x'$ weakly in $L_1([a, b])$, it occurs that $I[x] \leqq \liminf_{k \to \infty} I[x_k]$.

A more specific hypothesis is needed on A, namely that A be the closure of its interior, $A = \operatorname{cl}(\operatorname{int} A)$.

STATEMENT 3.1.i. *If A is closed, $A = \operatorname{cl}(\operatorname{int} A)$, $M = A \times E^n$, and $f_0(t, x, u)$ is continuous on A, then a necessary and sufficient condition for $I[x]$ to be lower semicontinuous in $\mathscr{T}$ is that $f_0(t, x, u)$ be convex with respect to u in E^n for every $(t, x) \in A$.*

The necessity of the condition was proved by Tonelli in 1912 (Tonelli's argument for uniform convergence applies as well for the mode of convergence chosen above). The sufficiency has been proved by Cesari in [15].

3.2. A lower semicontinuity theorem with weak convergence. Here $v \geqq 1$, G is any measurable subset of E^v of finite measure, for every $t \in G$ a subset $A(t)$ of E^n is assigned; $A = \{(t, x) | t \in G, x \in A(t)\} \subset E^{v+n}$, $M = A \times E^m \subset E^{v+n+m}$, and $f_0(t, x, u)$ is a given function on M. We shall assume that usual Carathéodory Condition (C) holds. Then, for any pair $x(t)$, $u(t)$, $t \in G$, of measurable functions with $x(t) \in A(t)$, $t \in G$ a.e., the function $f_0(t, x(t), u(t))$ is measurable. We shall consider only pairs for which $f_0(t, x(t), u(t))$, $t \in G$, has a Lebesgue integral on G (finite, $+\infty$, or $-\infty$), so that the functional $I[x, u]$ is defined for such pairs

$$I[x, u] = \int_G f_0(t, x(t), u(t))\, dt.$$

STATEMENT 3.2.i. *Let $x(t)$, $x_k(t)$, $u(t)$, $u_k(t)$, $t \in G$, $k = 1, 2, \cdots$, be measurable functions, $x, x_k : G \to E^n$, $u, u_k \in (L_1(G))^m$, with $x_k \to x$ in measure in G and $u_k \to u$ weakly in $(L_1(G))^m$ as $k \to \infty$. Let $M = A \times E^m$, $f_0(t, x, u)$ defined on M, convex with respect to u for every $(t, x) \in A$, let us assume that Condition (C) holds, and that $f_0(t, x, u)$ is bounded on the set $M_1 = \{(t, x, u) \in M | \, |u| \leqq 1\}$. Then*

$$I[x] \leqq \liminf_{k \to \infty} I[x_k],$$

where each of these integrals necessarily is finite or $+\infty$.

REFERENCES

[1] T. S. ANGELL, *Existence theorems for optimal control problems involving functional differential equations*, J. Optimization Theory Appl., 7 (1971), pp. 149–169.

[2] R. F. BAUM, *An existence theorem for optimal control systems with state variable in C and stochastic control problems*, Ibid., 5 (1970), pp. 335–346.

[3] L. D. BERKOWITZ, *Existence and lower closure theorems for abstract control problems*, to appear.

[4] M. F. BIDAUT, *Quelques resultats d'existence pour des problèmes de controle optimal*, C. R. Acad. Sci. Paris Ser A, 274 (1972), pp. 62–65.

[5] F. E. BROWDER, *Remarks on the direct method of the calculus of variations*, Arch. Rational Mech. Anal., 20 (1965), pp. 251–258.

[6] C. CASTAING AND M. VALADIER, *Equations differentielles multivoques dans les espaces vectoriels localment convexes*, Rev. Française Informat. Recherche Opérationnelle, 3^{e}année, no. 16, 1969, pp. 3–16.

[7] 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. 396–412, 413–429.

[8] ———, *Existence theorems for optimal controls of the Mayer type*, this Journal, 6 (1968), pp. 517–552.

[9] ———, *Seminormality and upper semicontinuity in optimal control*, J. Optimization Theory Appl., 6 (1970), pp. 114–137.

[10] ———, *Closure, lower closure, and semicontinuity theorems in optimal control*, this Journal, 9 (1971), pp. 287–315.

[11] ———, *Convexity and seminormality in the calculus of variations*, Atti e Memorie Accad. Sci. Lettere Arti Modena, to appear.

[12] ———, *Lagrange problems of optimal control and convex sets not containing any straight line*, Atti Sem. Mat. Fis. Univ. Modena, to appear.

[13] ———, *Closure theorems for orientor fields and weak convergence*, Arch. Rational Mech. Anal., to appear.

[14] ———, *Closure theorems for orientor fields*, Bull. Amer. Math. Soc., to appear.

[15] ———, *Lower semicontinuity and lower closure theorems without seminormality conditions*, Ann. Mat. Pura Appl., 98 (1974), pp. 381–397.

[16] L. CESARI AND D. E. COWLES, *Existence theorems in multidimensional problems of optimization with distributed and boundary controls*, Arch. Rational Mech. Anal., 46 (1972), pp. 321–355.

[17] L. CESARI, J. R. LAPALM AND D. A. SANCHEZ, *An existence theorem for Lagrange problems with unbounded controls and a slender set of exceptional points*, this Journal, 9 (1971), pp. 590–605.

[18] L. CESARI AND M. B. SURYANARAYANA, *Closure theorems without seminormality conditions*, J. Optimization Theory Appl., to appear.

[19] S. N. CHOW AND J. D. SCHUUR, *An existence theorem for ordinary differential equations in Banach spaces*, Bull. Amer. Math. Soc., 77 (1971), pp. 1018–1020.

[20] ———, *Fundamental theory of contingent differential equations in Banach space*, Trans. Amer. Math. Soc., 178 (1973), pp. 1–12.

[21] D. E. COWLES, *Upper semicontinuity properties of variable sets in optimal control*, J. Optimization Theory Appl., 10 (1972), pp. 222–236.

[22] ———, *Lower closure theorems for Lagrange problems of optimization with distributed and boundary controls*, Ibid., 10 (1972), pp. 300–320.

[23] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, Vestnik Moscov. Univ., Soc. Math. Mech. Astr., 2 (1959), pp. 25–32; English transl., this Journal, (1962), pp. 76–84.

[24] C. J. HIMMELBERG AND F. S. VANVLECK, *Selection and implicit function theorems for multifunctions with Souslin graph*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 19 (1971), pp. 911–916.

[25] M. Q. JACOBS, *Filippov's implicit function lemma*, this Journal, 5 (1967), pp. 622–627.

[26] P. KAISER, *Questions of seminormality in optimal control theory*, to appear.

[27] C. KURATOWSKI, *Les fonctions semi-continues dans l'espace des ensembles fermés*, Fund. Math., 18 (1932), pp. 148–166.

[28] J. R. LAPALM, *Contributions to the existence for Lagrange problems*, to appear.

[29] 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.

[30] E. J. MCSHANE AND R. B. WARFIELD, *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1969), pp. 41–47.

[31] C. OLECH, *Existence theorems for optimal problems with vector valued cost functions*, Trans. Amer. Math. Soc., 136 (1969), pp. 157–180.

[32] E. H. ROTHE, *An existence theorem in the calculus of variations*, Arch. Rational Mech. Anal., 21 (1966), pp. 151–162.

[33] M. B. SURYANARAYANA, *Existence theorems for optimization problems concerning hyperbolic partial differential equations*, J. Optimization Theory Appl., to appear.

[34] ———, *Remarks on lower semicontinuity*, Ibid., to appear.

SIAM J. Control
Vol. 12, No. 4, November 1974

REMARKS ON THE CONTROL OF DISCRETE-TIME DISTRIBUTED PARAMETER SYSTEMS*

J. ZABCZYK†

Abstract. The aim of this paper is to investigate limit properties of the infinite-dimensional control system described by a difference equation with a quadratic cost functional. Some related results of independent interest concerning stability and the Riccati operator difference equation are given also.

1. Introduction. K. Y. Lee, S. N. Chow and R. O. Barr investigated in [7] the infinite-dimensional system described by a difference equation with a quadratic cost functional. It turned out that a theory similar to that of the finite-dimensional case can be developed. The control on the infinite interval and the behavior of the solutions as the length of intervals tends to $+\infty$ were also considered.

In this paper we are mainly interested in limit properties of the solutions. We shall show, among others, that in all theorems of [7], weak convergence can be replaced by strong convergence or even by norm convergence. We shall consider also a more general problem: What must be the conditions to ensure the minimal cost taken on finite intervals $[0, \cdots, N-1]$ to be bounded for $N \to +\infty$? Some related results of independent interest concerning stability and the Riccati equation will be given also.

2. Notation, statement of the problem in the case of finite intervals. We shall use notations similar to those in [7]. So let H and U be Hilbert spaces, $\Phi : H \to H$, $D : U \to H$ be linear and bounded operators. H is called the *state space*, U the *space of controls.*

The control system is given by

$$x_{i+1} = \Phi x_i + Du_i, \qquad x_i \in H, \quad u_i \in U, \quad i = 0, 1, \cdots,$$

and the quadratic cost functional is given by

$$J_N(x_0; u_0, \cdots, u_{N-1}) = \sum_{i=0}^{N-1} [(Qx_i, x_i) + (Ru_i, u_i)], \qquad N = 1, 2, \cdots.$$

Here $Q : H \to H$ is a bounded self-adjoint, positive semidefinite operator, and $R : U \to U$ is a bounded, self-adjoint, positive definite operator. An operator P transforming a Hilbert space X into space X itself is called *positive semidefinite* if for every $x \in X$, $(Px, x) \geqq 0$; in this case we shall write $P \geqq 0$. If, in addition, P is an invertible operator, then it will be called *positive definite*, and in this case we shall write $P > 0$.

* Received by the editors April 23, 1973.

† Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland.

Let $\ell_N^2(H)$, $N = 1, 2, \cdots$, be the Hilbert space consisting of all sequences $(x_i)_{0 \leqq i < N}$ with $x_i \in H$, $i = 0, 1, \cdots$, and with an inner product defined by: for $\mathbf{x}, \mathbf{y} \in \ell_N^2(H)$, $\mathbf{x} = (x_i)$, $\mathbf{y} = (y_i)$,

$$(\mathbf{x}, \mathbf{y}) = \sum_{0 \leqq i < N} (x_i, y_i).$$

Similarly, the Hilbert space $\ell_N^2(U)$ may be defined.

Now we state the problem of the optimal control on the finite interval $[0, 1, 2, \cdots, N-1]$, $N = 1, 2, \cdots$.

Problem 1. For every $x_0 \in H$, determine a sequence $(u_0^*, u_1^*, \cdots, u_{N-1}^*) \in \ell_N^2(U)$, so that for any sequence $(u_0, \cdots, u_{N-1}) \in \ell_N^2(U)$,

$$J_N(x_0; u_0^*, \cdots, u_{N-1}^*) \leqq J_N(x_0; u_0, \cdots, u_{N-1}).$$

3. Solution of the problem. In this section we summarize a part of the results obtained in [7]. We give another proof of these results based on dynamic programming. This proof seems to be simpler than that in [7].

We start with some lemmas.

LEMMA 3.1. *Let A, B be linear, bounded operators $A: H \to H$, $B: H \to H$. Then the operators $I + AB$, $I + BA$ are both invertible or both not invertible; if they are invertible, then*

$$A(I + BA)^{-1} = (I + AB)^{-1}A. \tag{3.1}$$

If $A \geqq 0$, $B \geqq 0$, then $I + AB$ is an invertible operator and

$$A(I + BA)^{-1} \geqq 0. \tag{3.2}$$

Proof. It is well known that the operators AB and BA have the same spectrum (see [4, p. 311]); in particular $I + AB$ is an invertible operator iff $I + BA$ is. If the operators $(I + AB)^{-1}$ and $(I + BA)^{-1}$ exist, then since

$$(I + AB)A = A(I + BA),$$

therefore

$$(I + AB)^{-1}(I + AB)A(I + BA)^{-1}$$
$$= (I + AB)^{-1}A(I + BA)(I + BA)^{-1},$$

and therefore (3.1) holds.

Using a similar argument, we see that the spectrum of the operator AB is equal to the spectrum of $\sqrt{A}\,B\sqrt{A}$ if $A \geqq 0$, $B \geqq 0$. But the operator $\sqrt{A}\,B\sqrt{A}$ is positive semidefinite, so evidently, $(I + \sqrt{A}\,B\sqrt{A})^{-1}$ exists; therefore $(I + AB)^{-1}$ exists also.

To prove (3.2) we may, without loss of generality, suppose that A^{-1} exists. Then

$$A(I + BA)^{-1} = A[(A^{-1} + B)A]^{-1} = (A^{-1} + B)^{-1} \geqq 0.$$

LEMMA 3.2. *The system of the equations*

$$V_{i+1}(x) = \inf_{u \in U} [(Qx, x) + (Ru, u) + V_i(\Phi x + Du)] \tag{3.3}$$

where $V_0(x) = 0, x \in H, i = 0, 1, 2, \cdots, N - 1$, *has the unique solution* $(V_0, V_1, \cdots, V_N)$. *Functions* V_i, $i = 0, 1, \cdots, N$, *have the following forms*: $V_i(x) = (K_i x, x)$, $i = 0, 1, \cdots, N$, *where the operators* $K_i : H \to H$, $i = 0, 1, \cdots, N$, *are defined recursively*

$$(3.4)\quad K_0 = 0, \quad K_{i+1} = \Phi^* K_i (I + DR^{-1} D^* K_i)^{-1} \Phi + Q, \qquad i = 0, \cdots, N - 1.$$

Functions $f_i : U \to (-\infty, +\infty)$, $i = 1, 2, \cdots, N$,

$$f_i(u) = (Qx, x) + (Ru, u) + (K_{i-1}(\Phi x + Du), \Phi x + Du)$$

take on their minima at the unique points

$$u_i = -R^{-1} D^* K_{i-1} (I + DR^{-1} D^* K_{i-1})^{-1} \Phi x, \qquad i = 1, \cdots, N.$$

Proof. Lemma 3.2 is evidently true for $N = 1$. Let $V_N(x) = (K_N x, x)$, $x \in H$, where K_N is the linear, bounded, semidefinite operator given by (3.4). Then

$$V_{N+1}(x) = \min_{u \in U} f_{N+1}(u) = (Qx, x) + (\Phi^* K_N \Phi x, x)$$
$$+ \min_{u \in U} \{([R + D^* K_N D]u, u) + 2(D^* K_N \Phi x, u)\}.$$

But the operator $R + D^* K_N D$ is an invertible operator, so the function f_{N+1} takes on its minimum at a unique point $\bar{u}$ (see [8, p. 7]). But $df_{N+1} u \in U$ is given by $df_{N+1} u = 2\{[R + D^* K_N D]u + D^* K_N \Phi x\}$, so $[R + D^* K_N D]\bar{u} = -D^* K_N \Phi x$ and therefore

$$\bar{u} = -[R + D^* K_N D]^{-1} D^* K_N \Phi x = -(I + R^{-1} D^* K_N D)^{-1} R^{-1} D^* K_N \Phi x.$$

Using Lemma 3.1, we get

$$\bar{u} = -R^{-1} D^* K_N (I + DR^{-1} D^* K_N)^{-1} \Phi x = u_{N+1}.$$

Let us simplify the notation by defining

$$W_i = R^{-1} D^* K_i (I + DR^{-1} D^* K_i)^{-1} \Phi, \qquad i = 0, 1, \cdots.$$

Thus we have

$$V_{N+1}(x) = (Qx, x) + (RW_N x, W_N x) + (K_N[\Phi - DW_N]x, [\Phi - DW_N]x).$$

So the proof of Lemma 3.2 will be completed if we prove that $K_{N+1} = \bar{K}_{N+1}$, where

$$\bar{K}_{N+1} = Q + W_N^* R W_N + (\Phi - DW_N)^* K_N (\Phi - DW_N).$$

But

$$\Phi - DW_N = [I - D(I + R^{-1} D^* K_N D)^{-1} R^{-1} D^* K_N]\Phi,$$

and easy calculations give

$$\Phi - DW_N = (I + DR^{-1} DK_N)^{-1} \Phi, \qquad \text{(see also [4, p. 311])}.$$

Thus

$$\begin{aligned}\bar{K}_{N+1} &= Q + [R^{-1}D^*K_N(I + DR^{-1}D^*K_N)^{-1}\Phi]^*D^*K_N(I + DR^{-1}D^*K_N)^{-1}\Phi \\ &\quad + \Phi^*[(I + DR^{-1}D^*K_N)^{-1}]^*K_N(I + DR^{-1}D^*K_N)^{-1}\Phi \\ &= Q + \Phi^*[(I + K_NDR^{-1}D^*)^{-1}(K_NDR^{-1}D^* + I)]K_N(I + DR^{-1}D^*K_N)^{-1}\Phi \\ &= Q + \Phi^*K_N(I + DR^{-1}D^*K_N)^{-1}\Phi = K_{N+1}.\end{aligned}$$

Let us remark that, by virtue of the Lemma 3.1, the operators K_I, $i = 0, 1, \cdots$, are well-defined and self-adjoint, and $K_i \geqq 0$.

We now proceed to prove the next theorem.

THEOREM 3.1 (see [7]). *For every initial state $x_0 \in H$ and $N = 1, 2, \cdots$, the optimal control $(u_0^*, \cdots, u_{N-1}^*)$ is given by the feedback form*

$$u_i^* = -R^{-1}D^*K_{N-i-1}(I + DR^{-1}D^*K_{N-i-1})^{-1}\Phi x_i = W_{N-i-1}x_i,$$

$$i = 0, 1, \cdots, N - 1.$$

The optimal cost is given by

$$J_N(x_0; u_0^*, \cdots, u_{N-1}^*) = (K_Nx_0, x_0) = V_N(x_0).$$

Proof. The theorem is evidently true for $N = 1$. If it is true for N, then for $N + 1$ we will have

$$\begin{aligned}J_{N+1}(x_0; u_0, \cdots, u_N) &= (Qx_0, x_0) + (Ru_0, u_0) + \sum_{i=1}^{N} [(Qx_i, x_i) + (Ru_i, u_i)] \\ &= (Qx_0, x_0) + (Ru_0, u_0) + J_N(x_1; u_1, \cdots, u_N).\end{aligned}$$

So

$$J_{N+1}(x_0; u_0, \cdots, u_N) \geqq (Qx_0, x_0) + (Ru_0, u_0) + V_N(\Phi x_0 + Du_0)$$

and, by virtue of Lemma 3.2,

$$J_{N+1}(x_0; u_0, \cdots, u_N) \geqq (Qx_0, x_0) + (RW_Nx_0, W_Nx_0) + (K_Nx_1, x_1).$$

Therefore

$$J_{N+1}(x_0; u_0, \cdots, u_N) \geqq J_{N+1}(x_0; -W_Nx_0, \cdots, -W_0x_N).$$

In this way the proof of Theorem 3.1 is completed.

4. Properties of the Riccati operator difference equation. Let $\mathscr{E}$ be the Banach space of all bounded self-adjoint operators $K: H \to H$ with the usual operator norm and $\mathscr{K}$ the cone of all positive semidefinite operators. The cone is normal. For the notions of cone, normal cone and other properties of cones which will be needed in the sequel, we refer to [6].

In this section we shall be interested in the properties of the equations

$$K_{i+1} = \Phi^*K_i(I + SK_i)^{-1}\Phi + Q, \qquad i = 0, 1, 2, \cdots, \tag{4.1}$$

$$K = \Phi^*K(I + SK)^{-1}\Phi + Q, \tag{4.2}$$

where K_0, Q, $S \in \mathscr{K}$. The results of this section will be needed in § 6.

THEOREM 4.1. *Let us define the transformation* $A:\mathscr{K} \to \mathscr{K}$ *by the formula*

$$A(K) = \Phi^*K(I + SK)^{-1}\Phi + Q.$$

The operator is monotonic, and for $K_n = A^n(K_0), L_n = A^n(L_0), K_0, L_0 \in \mathscr{K}$, $n = 1, 2, \cdots$,

$$\begin{aligned}\|K_n - L_n\| \leqq \|K_0 - L_0\| \, &\|(I + SK_0)^{-1}\Phi(I + SK_1)^{-1}\Phi \cdots (I + SK_{n-1})^{-1}\Phi\| \\ &\cdot \|(I + SL_0)^{-1}\Phi(I + SL_1)^{-1}\Phi \cdots (I + SL_{n-1})^{-1}\Phi\|.\end{aligned} \tag{4.3}$$

If, in addition, $Q > 0$, *then* A *is a I-concave operator.*

In the proof we shall need the following lemma.

LEMMA 4.1. *Let* $A, B \in \mathscr{K}$ *and* $A \geqq B > 0$; *then*

$$B^{-1} \geqq A^{-1} > 0.$$

Proof of the lemma. Let $C \geqq 0$ be such that $A = B + C$. It is not difficult to check that

$$B^{-1} - (B + C)^{-1} = B^{-1}C(I + B^{-1}C)^{-1}B^{-1}.$$

But, by virtue of Lemma 3.1,

$$C(I + B^{-1}C)^{-1} \geqq 0.$$

Thus

$$B^{-1}C(I + B^{-1}C)^{-1}B^{-1} \geqq 0.$$

Proof of Theorem 4.1. Let us remark that Lemma 3.1 implies that A really transforms $\mathscr{K}$ into $\mathscr{K}$. To prove the monotonicity, let us suppose $K_1 \geqq K_2 \geqq 0$; then

$$A(K_1) - A(K_2) = \Phi^*[K_1(I + SK_1)^{-1} - K_2(I + SK_2)^{-1}]\Phi.$$

Let, in addition $K_2 > 0$. Then

$$K_1(I + SK_1)^{-1} - K_2(I + SK_2)^{-1} = (K_1^{-1} + S)^{-1} - (K_2^{-1} + S)^{-1}.$$

In view of Lemma 4.1,

$$K_1^{-1} + S \leqq K_2^{-1} + S,$$

so, applying the same lemma, we have

$$(K_1^{-1} + S)^{-1} \geqq (K_2^{-1} + S)^{-1}.$$

Therefore in the case $K_2 > 0$, $A(K_1) - A(K_2) \geqq 0$; we obtain the general case $K_2 \geqq 0$ by the easy limit argument (we can add to operators K_1 and K_2 the operator εI, $\varepsilon > 0$ with $\varepsilon \to 0$).

We recall (see [6]) that an operator $A:\mathscr{K} \to \mathscr{K}$ is called an I-concave operator iff:

(1°) for every $K \in \mathscr{K}$, there exist numbers $\alpha, \beta > 0$ such that $\alpha I \leqq A(K) \leqq \beta I$,

(2°) for every number $t \in (0, 1)$ and operator $K \in \mathscr{K}$, there exists $\varepsilon_0 > 0$ such that $A(tK) \geqq (t + \varepsilon)A(K)$, for every $\varepsilon \in (0, \varepsilon_0)$.

Of course if $Q > 0$, then condition (1°) is satisfied. To prove that condition (2°) is satisfied also, let us observe that

$$\begin{aligned}A(tK) - (t + \varepsilon)A(K) = (1 - t - \varepsilon)Q + \Phi^*\{&-\varepsilon K(I + SK)^{-1} \\ &+ t[K(I + tSK)^{-1} - K(I + SK)^{-1}]\}\Phi.\end{aligned}$$

So it is sufficient to prove that

$$K(I + tSK)^{-1} - K(I + SK)^{-1} \geqq 0. \tag{4.4}$$

But the inequality (4.4) is true if $K > 0$, because then it is equivalent (see Lemma 4.1) to the true inequality

$$K^{-1} + tS \leqq K^{-1} + S.$$

We get the general case by the simple limit argument.

To complete the proof of the Theorem 4.1, we have to show that (4.3) holds.

It is not difficult to check that

$$K(I + SK)^{-1} - L(I + SL)^{-1} = (I + KS)^{-1}(K - L)(I + SL)^{-1}.$$

So

$$\begin{aligned} K_n - L_n &= \Phi^*(I + K_{n-1}S)^{-1}(K_{n-1} - L_{n-1})(I + SL_{n-1})^{-1}\Phi \\ &= \Phi^*(I + K_{n-1}S)^{-1} \cdots \Phi^*(I + K_1S)^{-1}\Phi^*(I + K_0S)^{-1}(K_0 - L_0) \\ &\quad \cdot (I + SL_0)^{-1}\Phi(I + SL_1)^{-1}\Phi \cdots (I + SL_{n-1})^{-1}\Phi, \end{aligned}$$

and therefore the inequality (4.3) is true.

THEOREM 4.2. *Let us suppose* $Q > 0$; *then* (4.2) *has at most one solution* $K \in \mathscr{K}$. *If* $K \in \mathscr{K}$ *is the solution of* (4.2), *then* $\|K_i - K\| \to 0$ *geometrically fast as* $i \to +\infty$ *and* K_i, $i = 0, 1, 2, \cdots$, *are defined by* (4.1).

Proof. Let us suppose $Q > 0$. Using Theorem 4.1, we see that an operator A satisfies all conditions of [6, Thm. 6.3], so the equation $K = A(K)$ has at most one solution. If a solution K exists, then the conditions of [6, Thm. 6.4] are satisfied, so the successive approximations $K_i = A^i(K_0)$ tend to K. Let us define

$$U = H, \quad D^* = D = \sqrt{S}, \quad T_i = (I + SK_i)^{-1}\Phi, \quad T = (I + SK)^{-1}\Phi, \quad R = I.$$

We shall see in §6 (see the proof of Theorem 6.2) that $r(\Phi - \sqrt{S}(I + \sqrt{S}K\sqrt{S})^{-1}\sqrt{S}K\Phi) < 1$ but $\Phi - \sqrt{S}(I + \sqrt{S}K\sqrt{S})^{-1}\sqrt{S}K\Phi = (I - \sqrt{S}(I + \sqrt{S}K\sqrt{S})^{-1}\sqrt{S}K)\Phi = (I + SK)^{-1}\Phi$, so $r(T) < 1$. Therefore there exists a natural number k_0 such that $\|T^{k_0}\| < 1$. Since $\|K_i - K\| \to 0$ as $i \to +\infty$, so $\|T_{i+1}T_{i+2} \cdots T_{i+k_0}\| \to \|T^{k_0}\|, i \to +\infty$. From Theorem 4.1 it follows that

$$\begin{aligned} \|K - K_n\| \leqq \|K - K_0\| \, \|T^{k_0}\|^p \|T^q\| \, \|T_0T_1 \cdots T_{k_0-1}\| \cdots \\ \cdot \|T_{(p-1)k_0} \cdots T_{pk_0-1}\| \, \|T_{pk_0}T_{pk_0+1} \cdots T_n\|, \end{aligned}$$

where $n = pk_0 + q, 0 \leqq q < k_0, p, q$ nonnegative integers, and therefore $\|K_i - K\| \to 0$ geometrically fast.

Remark 4.1. If $K_0 \leqq K$ (for instance if $K_0 = 0$), then the sequence of successive approximations is monotonic.

5. Stability theorem. We now prove a general stability result. A special case of it will be needed in the next section.

THEOREM 5.1. *Let X be a Banach space and $\Phi: X \to X$ be a bounded, linear operator, and let $N:[0, +\infty) \to [0, +\infty)$, $N(0) = 0$ be a continuous, strictly increasing, convex function. If, for every $x \in X$, there exists an $\alpha > 0$ such that*

$$\sum_{n=0}^{+\infty} N(\alpha\|\Phi^n x\|) < +\infty,$$

then the spectral radius $r(\Phi)$ of Φ,

$$r(\Phi) = \lim_n \sqrt[n]{\|\Phi^n\|},$$

is less than 1.

Proof. Let ℓ^N be a set consisting of all sequences $(x_i)_{i=0,1,\cdots}$, $x_i \in X$, for which there exist $\alpha = \alpha(x_i) > 0$ such that $\sum_{i=0}^{+\infty} N(\alpha\|x_i\|) < +\infty$. It is not difficult to check that ℓ^N is a normed space with the norm

$$\|(x_i)\| = \inf\left\{\varepsilon > 0; \sum_{i=0}^{+\infty} N\left(\frac{1}{\varepsilon}\|x_i\|\right) \leqq 1\right\},$$

Let us remark that if $(x_i) \in \ell^N$, then

$$\sum_{i=0}^{+\infty} N\left(\frac{1}{\|(x_j)\|}\|x_i\|\right) \leqq 1.$$

For every $n = 0, 1, 2, \cdots$ and $x \in X$, we define

$$Z_n(x) = (x, \Phi x, \Phi^2 x, \cdots, \Phi^n x, 0, 0, \cdots).$$

Then $Z_n: X \to \ell^N$ and is a bounded linear operator. Since for every $x \in X$, $\sup_n \|Z_n(x)\| < +\infty$, so in view of the Banach–Steinhaus theorem, there exists an $M > 0$ such that $\|Z_n(x)\| \leqq M\|x\|$, $n = 0, 1, \cdots$. Therefore, for $n = 0, 1, \cdots$ and $x \neq 0$, we have

$$1 \geqq \sum_{i=0}^{n} N\left(\frac{1}{\|Z_n(x)\|}\|\Phi^i x\|\right) \geqq \sum_{i=0}^{n} N\left(\frac{\|\Phi^i x\|}{M\|x\|}\right).$$

The same Banach–Steinhaus theorem implies that $\sup_i \|\Phi^i\| \leqq \bar{M} < +\infty$ because for every $x \in X$, $\lim_i \|\Phi^i x\| = 0$.

For some γ, $0 < \gamma < 1/\bar{M}$, we define the function $m: X \to \{0, 1, 2, \cdots\}$ as follows:

$$m(x) = \inf\{i; \|\Phi^i x\| \leqq \gamma\|x\|\}.$$

The function m is bounded.

Since we have

$$m(x)N\left(\frac{\gamma}{M}\right) \leqq \sum_{0 \leqq i \leqq m(x)-1} N\left(\frac{\|\Phi^i x\|}{M\|x\|}\right) \leqq 1,$$

hence $m(x) \leqq (N(\gamma/M))^{-1}$.

For $i_0 \geqq (N(\gamma/M))^{-1}$, we have

$$\|\Phi^{i_0}(x)\| = \|\Phi^{i_0 - m(x)}(\Phi^{m(x)}(x))\| \leqq \bar{M}\gamma\|x\|.$$

So we have proved that

$$\|\Phi^{i_0}\| \leqq \bar{M}\gamma < 1, \quad \text{thus} \quad r(\Phi) < 1.$$

Remark 5.1. The method used in the proof is similar to that used in [1, pp. 190–193] in the case of $N(r) = r^p$ and for the semigroup of operators.

Remark 5.2. It is not difficult to prove by the same method an analogous theorem in the case of continuous-time: *If $T(t)$ is strongly continuous semigroup of operators and for every $x \in X$ there exists $\alpha > 0$ such that*

$$\int_0^{+\infty} N(\alpha\|T_t x\|)\, dt < +\infty,$$

then there exist $M > 0$, $\lambda \in (0, 1)$ such that $\|T_t\| \leqq Me^{-\lambda t}$ for every $t \geqq 0$ (see also [3]).

6. Control on the infinite interval. First case. We shall now consider a control on the infinite interval with the condition: For every $x \in H$,

$$\sup_N \inf_{(u_0, \cdots, u_{N-1})} J_N(x; u_0, \cdots, u_{N-1}) < +\infty. \tag{6.1}$$

This condition is equivalent, by virtue of Theorem 3.1 to the condition: For every $x \in H$,

$$\sup_N (K_N x, x) < +\infty, \tag{6.2}$$

where $K_0 = 0$ and

$$\begin{aligned} K_{N+1} &= \Phi^* K_N (I + DR^{-1}DK_N)^{-1}\Phi + Q, \\ &= A(K_N) \quad \text{with} \quad S = DR^{-1}D^*, \qquad N \geqq 0 \end{aligned}$$

LEMMA 6.1. *Let* (6.1) *hold; then $K_N \to K$ strongly as $N \to +\infty$, where K is a bounded, positive semidefinite operator.*

Proof. Since the sequence (K_N) is monotonic, Lemma 6.1 is an immediate consequence of the well-known theorem (see [5, p. 925, Ex. 21]) which follows.

THEOREM 6.1. *The positive semidefinite solution of the equation $K = A(K)$ exists iff the condition* (6.1) *holds.*

Proof. If $\bar{K} \in \mathscr{K}$ and $\bar{K} = A(\bar{K})$, then $\bar{K} \geqq K_N = A^N(0)$, so $\sup_N (K_N x, x) < +\infty$, for every $x \in H$.

To prove the inclusion in the opposite direction we show, first of all, that if $L_N \xrightarrow[\text{strongly}]{} L$, $L_N, L \in \mathscr{K}$ then $A(L_N) \xrightarrow[\text{strongly}]{} A(L)$ also. In fact, from § 3 (see the proof of Lemma 3.2) it follows that for every $L \in \mathscr{K}$,

$$A(L) = Q + W^*RW + (\Phi - DW)^*L(\Phi - DW),$$

where

$$W = W(L) = (R + D^*LD)^{-1}D^*L\Phi.$$

Using [5, Thm. 2, p. 922] we see that $W(L_N) \xrightarrow[\text{strongly}]{} W(L)$ and $W(L_N) \xrightarrow[\text{strongly}]{} W(L)$. But then $A(L_N) \xrightarrow[\text{strongly}]{} A(L)$ also.

Let now K be the operator defined in Lemma 6.1; then $K_N \xrightarrow[\text{strongly}]{} K$ so $A(K_N) \xrightarrow[\text{strongly}]{} A(K)$ also. But $K_{N+1} = A(K_N)$, so $K = A(K)$.

THEOREM 6.2. *Let $Q > 0$. The equation $K = A(K)$, $K \in \mathscr{K}$, has a solution iff there exists a bounded operator $W: H \to U$ such that*

$$r(\Phi - DW) < 1. \tag{6.3}$$

Here $r(\Phi - DW)$ denotes the spectral radius of $\Phi - DW$.

Proof. If $K = A(K)$, $K \in \mathscr{K}$, then we define $W = (R + D^*KD)^{-1}D^*K\Phi$. To prove (6.2) it is sufficient to show that for every $x \in H$, $\sum_{n=0}^{+\infty} \|(\Phi - DW)^n x\|^2 < +\infty$ and then apply Theorem 5.1, setting $N(r) = r^2$. Let us introduce some notations: $T = \Phi - DW$, $T_i = \Phi - DW_i$, where the operators $W_i = (R + D^*K_iD)^{-1}D^*K_i\Phi$, $i = 0, 1, \cdots$, were introduced before in § 3. Let us remark (see Theorem 3.1) that for every $N = 1, 2, \cdots$, the optimal trajectory $(x_0, x_1, \cdots, x_{N-1})$ starting from x is given by formula

$$x_i = T_{N-i}T_{N-i+1} \cdots T_{N-1}x_0, \qquad i = 1, 2, \cdots, N-1,$$

and the optimal control $(u_0, \cdots, u_{N-1})$ by

$$u_i = -W_{N-i-1}x_i, \qquad i = 0, 1, \cdots, N-1.$$

Therefore

$$\begin{aligned}(K_Nx_0, x_0) &= (Qx_i, x_i) + (Ru_i, u_i) \\ &= \sum_{i=1}^{N-1} ([Q + W^*_{N-i-1}RW_{N-i-1}]T_{N-i} \cdots T_{N-1}x_0, T_{N-i} \cdots T_{N-1}x_0) \\ &\quad + ([Q + W^*_{N-1}RW_{N-1}]x_0, x_0).\end{aligned}$$

If we fix $N_0 < N$ and base on the fact that $W_i \xrightarrow[\text{strongly}]{} W$, $W_i^* \xrightarrow[\text{strongly}]{} W^*$, $T_i \xrightarrow[\text{strongly}]{} T$, then we will get from the above relations:

$$\begin{aligned}(Kx_0, x_0) \geqq (K_Nx_0, x_0)&([Q + W^*RW]x_0, x_0) \\ &+ \sum_{i=1}^{N_0} ([Q + W^*RW]T^ix_0, T^ix_0).\end{aligned}$$

Thus

$$\sum_{i=0}^{+\infty} ([Q + W^*RW]T^ix_0, T^ix_0) \leqq (Kx_0, x_0).$$

Since $Q > 0$, we have $\sum_{i=0}^{+\infty} \|T^ix_0\|^2 < +\infty$ for every $x_0 \in H$, and therefore $r(T) = r(\Phi - DW) < 1$.

If $r(\Phi - D\bar{W}) < 1$ for some operator $\bar{W}: H \to U$, then the feedback control

$$\bar{u}_i = -\bar{W}x_i \qquad i = 0, 1, \cdots,$$

gives

$$J_N(x_0; \bar{u}_0, \cdots, \bar{u}_{N-1}) = \sum_{i=0}^{N-1} ([Q + \bar{W}^*R\bar{W}]\bar{T}^i x_0, \bar{T}^i x_0),$$

where $\bar{T} = \Phi - D\bar{W}$. So

$$\begin{aligned}\lim_N (K_N x_0, x_0) &\leqq \lim_N J_N(x_0; \bar{u}_0, \cdots, \bar{u}_{N-1})\\ &\leqq \sum_{i=0}^{+\infty} ([Q + \bar{W}^*R\bar{W}]\bar{T}^i x_0, \bar{T}^i x_0) < +\infty\end{aligned}$$

because $\|\bar{T}^{*i}\| = \|\bar{T}^i\| \leqq \bar{M}\lambda^i$, $i = 0, 1, \cdots$, for some $\bar{M} > 0$ and λ, $0 < \lambda < 1$.

Remark 6.1. Using the same method as in the proof of Theorem 6.2, we can prove that for every $x \in H$,

$$(Kx, x) = \sum_{i=0}^{+\infty} ([Q + W^*RW]T^i x, T^i x). \tag{6.4}$$

THEOREM 6.3. *Let us define the operators* $P_N : H \to \ell^2_\infty(H)$, $N = 1, 2, \cdots$, *by the formulas*

$$P_N x = (x, T_{N-1}x, T_{N-2}T_{N-1}x, \cdots, T_1 \cdots T_{N-1}x, 0, 0, \cdots), \qquad N = 1, 2, \cdots$$

$$P_\infty x = (x, Tx, T^2x, \cdots, T^{N-1}x, T^N x, \cdots), \qquad x \in H.$$

If $Q > 0$ *and there exists* $K \in \mathcal{K}$ *such that* $K = A(K)$, *then* $\lim_N \|P_N - P_\infty\| = 0$.

This theorem says that the optimal trajectories for the finite control intervals converge to the optimal trajectory for the control problem on the infinite interval and the convergence is in the operator norm.

Proof. Let us remark that for any natural numbers N, N_1, with $N > N_1$, it holds that

$$\begin{aligned}\|P_\infty x - P_N x\|^2 &\leqq \sum_{i=1}^{N_1} \|T^i x - T_{N-i} \cdots T_{N-1}x\|^2\\ &\quad + 2 \sum_{i > N_1} \|T^i x\|^2 + 2 \sum_{i > N_1}^{N-1} \|T_{N-i} \cdots T_{N-1}x\|^2.\end{aligned} \tag{6.5}$$

We suppose now $Q > 0$ and $K = A(K)$ for some $K \in \mathcal{K}$, and let us fix $\varepsilon > 0$.

Let N_1 be a natural number such that

$$2 \sum_{i > N_1} \|T^i x\|^2 < \frac{\varepsilon}{3}\|x\|^2, \tag{6.6}$$

$$\left\|K - \sum_{i=0}^{N_1} T^{*i}(Q + W^*RW)T^i\right\| < \frac{\varepsilon\gamma}{12}, \tag{6.7}$$

where $\gamma = \inf_{\|y\|=1} (Qy, y) > 0$. Such a number N_1 exists. In fact, since $r(T) < 1$, there exists numbers $M > 0$ and λ, $0 < \lambda < 1$, for which

$$(6.8)\qquad \|T^i\| = \|T^{*i}\| \leqq M\lambda^i, \qquad i = 0,1,2,\cdots,$$

so

$$\sum_{i>N_1} \|T^i\|^2 \leqq \frac{M^2\lambda^2}{1-\lambda^2}\lambda^{2N_1}.$$

On the other hand, by virtue of (6.4),

$$(6.9)\qquad K = \sum_{i=0}^{+\infty} T^{*i}(Q + W^*RW)T^i.$$

This together with (6.8) gives (6.7). By standard considerations, relying on the fact that $\|K_i - K\| \to 0$ as $i \to +\infty$, we show that

$$(6.10)\qquad \|T_i - T\| \to 0, \quad \|W_l - W\| \to 0 \quad \text{as } i \to +\infty.$$

In view of (6.9) and (6.7), there exists an $N_2 > N_1$ such that for $N > N_2$, $x \in H$, $\|x\| \leqq 1$,

$$\begin{aligned}(Kx,x) \geqq (K_Nx,x) &\geqq ([Q + W^*_{N-1}RW_{N-1}]x,x)\\ &\quad + \sum_{i=1}^{N_1} (T^*_{N-1}\cdots T^*_{N-i}[Q + W^*_{N-1-i}RW_N]\\ &\qquad \cdot T_{N-i}\cdots T_{N-1}x,x)\\ &\geqq (Kx,x) - \varepsilon\gamma/6.\end{aligned}$$

Thus for $N > N_2$, $x \in H$, $\|x\| \leqq 1$,

$$\begin{aligned}\frac{\varepsilon\gamma}{6} &\geqq \sum_{i>N_1}^{N-1} ([Q + W^*_{N-1-i}RW_{N-1-i}]T_{N-i}\cdots T_{N-1}x, T_{N-i}\cdots T_{N-1}x)\\ &\geqq \gamma \sum_{i>N_1}^{N-1} \|T_{N-i}\cdots T_{N-1}x\|^2.\end{aligned}$$

Let now $N_3 > N_2$ be such that for $N > N_3$,

$$\sum_{i=1}^{N_1} \|T^ix - T_{N-i}\cdots T_{N-1}x\|^2 \leqq \frac{\varepsilon}{3}\|x\|^2;$$

see (6.10).

At last, for $x \in H$, $\|x\| \leqq 1$, and $N > N_3$,

$$\|P_Nx - P_\infty x\|^2 \leqq \varepsilon/3 + \varepsilon/3 + 2\varepsilon/6 = \varepsilon.$$

This completes the proof in the case $Q > 0$.

THEOREM 6.4. *Let us define the operators*: $V_N: H \to \ell^2_\infty(U)$, $N = 1,2,\cdots$, *by the formulas*

$$V_Nx = (-W_{N-1}x, -W_{N-2}T_{N-1}x, \cdots, -W_1T_2\cdots T_{N-1}x, 0, 0, \cdots), \qquad N = 1,2,\cdots,$$

$$V_\infty x = (-Wx, -WTx, -WT^2x, \cdots).$$

Under the same assumptions as in Theorem 6.3,

$$\lim_N \|V_N - V_\infty\| = 0.$$

Proof. The proof of the theorem is analogous to that of Theorem 6.3.

This theorem states that the optimal controls for the control problem on the finite intervals tend to the optimal control for the control problem on the infinite interval and the convergence is in the operator norm.

7. Control on the infinite interval. Second case. In [7] the control problem on the infinite interval was investigated under the following hypothesis.

Hypothesis 7.1. For every sequence $(u_i) \in \ell^2_\infty(U)$ and for every $x_0 \in H$, the sequence (x_i) given by

$$x_{i+1} = \Phi x_i + Du_i, \qquad i = 0, 1, 2, \cdots,$$

belongs to $\ell^2_\infty(H)$.

In [7] it was proved that the condition $\|\Phi\| < 1$ was sufficient for Hypothesis 7.1 to hold.

In this section we strengthen, otherwise than in § 6, some of the results obtained in [7].

Lemma 7.1. *Hypothesis* 7.1 *is true iff* $r(\Phi) < 1$.

Proof. To prove the sufficiency, let us denote $v_i = Du_i$, $i = 0, 1, 2, \cdots$. We have

$$x_{i+1} = v_i + \Phi v_{i-1} + \cdots + \Phi^i v_0 + \Phi^{i+1} x_0, \qquad i = 0, 1, \cdots,$$

so

$$\|x_{i+1}\| \leqq \|v_i\| + \|\Phi\| \, \|v_{i-1}\| + \cdots + \|\Phi^i\| \, \|v_0\| + \|\Phi^{i+1}\| \, \|x_0\|.$$

Therefore we may write

$$(\|x_0\|, \|x_1\|, \cdots) \leqq (1, \|\Phi\|, \|\Phi^2\|, \cdots) * (\|x_0\|, \|v_0\|, \|v_1\|, \cdots)$$

where $*$ denotes the convolution transform.

Using the well-known Young's inequality (see [5, p. 951, (c)]), we get

$$\left(\sum_{i=0}^{+\infty} \|x_i\|^2\right)^{1/2} \leqq \left(\sum_{i=0}^{+\infty} \|\Phi^i\|\right)(\|x_0\|^2 + \sum_{i=0}^{+\infty} \|v_i\|^2)^{1/2}$$

But, of course, $r(\Phi) < 1$ implies $\sum_{i=0}^{+\infty} \|\Phi^i\| < +\infty$.

The necessity follows immediately from Lemma 5.1. It is necessary only to put $N(r) = r^2$ and $u_i = 0$, $i = 0, 1, \cdots$.

When Hypothesis 7.1 holds, then it is possible to formulate the control problem on the infinite interval as follows (see [7]).

Problem 2. For every initial state $x_0 \in H$, find a sequence $\mathbf{u}^* = (u_i^*) \in \ell^2_\infty(U)$ such that for every $(u_i) \in \ell^2_\infty(U)$,

$$J_\infty(x; u_0, u_1, \cdots) \geqq J_\infty(x; u_0^*, u_1^*, \cdots)$$

where

$$J_\infty(x; u_0, u_1, \cdots) = \sum_{i=0}^{+\infty} [(Qx_i, x_i) + (Ru_i, u_i)]$$

and $x_{i+1} = \Phi x_i + Du_i$, $i = 0, 1, \cdots$.

Let us introduce sequence $\mathbf{x}^* = (x_i^*)$, $x_{i+1}^* = \Phi x_i^* + Du_i^*$, $x_0^* = x_0$.

Here $\mathbf{u}^* = (u_0^*, u_1^*, \cdots)$ is the only point for which the minimum is achieved.

As in § 6 we shall use the following notions:

$$W_N = (R + D^*K_ND)^{-1}D^*K_N\Phi, \qquad T_N = \phi - DW_N, \qquad N = 0, 1, \cdots,$$
$$K = \lim_N K_N, \qquad W = (R + D^*KD)^{-1}D^*K\Phi, \qquad T = \phi - DW.$$

The following theorem holds.

Theorem 7.1. *If* $r(\Phi) < 1$ *and* $R > 0$, *then the condition* (6.1) *is satisfied and for every* $x \in H$,

$$\mathbf{x}^* = (x, Tx, T^2x, \cdots), \mathbf{u}^* = (-Wx, -WTx, \cdots),$$

(7.1) $$P_N x \xrightarrow[\text{strongly}]{} \mathbf{x}^* \quad in \quad l^2_\infty(H),$$

(7.2) $$V_N x \xrightarrow[\text{strongly}]{} \mathbf{u}^* \quad in \quad l^2_\infty(U) \quad as\ N \to +\infty.$$

Proof. To prove that (7.1), (7.2) hold we introduce the operators

$$\mathbf{A}: l^2_\infty(U) \to l^2_\infty(H), \qquad \mathbf{B}: \mathrm{H} \to l^2_\infty(H),$$
$$\mathbf{Q}: l^2_\infty(H) \to l^2_\infty(H), \qquad \mathbf{R}: l^2_\infty(U) \to l^2_\infty(U),$$
$$\mathbf{C}_N: l^2_\infty(H) \to l^2_\infty(H), \qquad N = 1, 2, \cdots$$

by the formulas

$$\mathbf{A}(u_i) = (x_i),$$

where

$$x_0 = 0, \qquad x_{i+1} = \Phi x_i + Du_i, \qquad i = 0, 1, \cdots,$$
$$\mathbf{B}x = (\Phi^i x), \qquad \mathbf{Q}(x_i) = (Qx_i),$$
$$\mathbf{R}(u_i) = (Ru_i), \qquad \mathbf{C}_N(x_i) = (x_0, x_1, \cdots, x_{N-1}, 0, \cdots).$$

Since, for every $N > N_0$,

$$(RW_{N-1}x, W_{N-1}x) + \sum_{l=1}^{N_0} (RW_{N-1-i}T_{N-i} \cdot \cdots \cdot T_{N-1}x, W_{N-1-i}T_{N-i} \cdot \cdots \cdot T_{N-1}x)$$
$$\leqq (K_Nx, x) \leqq (Kx, x),$$

and so, for every $N_0 = 1, 2, \cdots$,

$$(RWx, Wx) + \sum_{l=0}^{N_0} (RWT^i x, WT^i x) \leqq (Kx, x)$$

Therefore,

$$\bar{\mathbf{u}} = (-Wx, -WTx, -WT^2x, \cdots) \in l^2_\infty(U).$$

Let us define the sequence

$$\bar{\mathbf{x}} = (x, Tx, T^2x, \cdots);$$

then $\bar{\mathbf{x}} = \mathbf{A}\bar{\mathbf{u}} + \mathbf{B}x$. Consequently $\bar{\mathbf{x}} \in l^2_\infty(H)$ and it is easy to see that

$$(Kx, x) = J_\infty(x; \bar{\mathbf{u}}).$$

So we have proved that

$$\bar{\mathbf{u}} = \mathbf{u}^*, \qquad \bar{\mathbf{x}} = \mathbf{x}^*.$$

Since $J_N(x; \cdot)$ takes on its minimum at $V_N x$, it follows that for any sequence $\mathbf{u}^N = (u_0, u_1, \cdots, u_{N-1}, 0, 0, \cdots)$, which can be identified with $(u_0, u_1, \cdots, u_{N-1})$, we have

$$J_N(x; \mathbf{u}^N) - J_N(x; V_N x) = (\mathbf{R}(\mathbf{u}^N - V_N x), \mathbf{u}^N - V_N x) \\ + (\mathbf{Q}\mathbf{C}_N\mathbf{A}(\mathbf{u}^N - V_N x), \mathbf{C}_N\mathbf{A}(\mathbf{u}^N - V_N x)).$$

Using this equality we obtain that

(7.3) $$J_N(x; \mathbf{u}^{*N}) - J_N(x; V_N x) \geqq \inf_{\|z\|=1} (Rz, z)\|\mathbf{u}^{*N} - V_N x\|^2,$$

where

$$\mathbf{u}^{*N} = (u_0^*, u_1^*, u_2^*, \cdots, u_{N-1}^*, 0, \cdots).$$

But

$$(Kx, x) = J_\infty(x; \mathbf{u}^*) \geqq J_N(x; \mathbf{u}^{*N}) \geqq J_N(x; V_N x) = (K_N x, x);$$

therefore

$$J_N(x; \mathbf{u}^{*N}) - J_N(x; V_N x) \to 0,$$

and consequently (see (7.3))

$$\mathbf{u}^{*N} - V_N x \xrightarrow[\text{strongly}]{} 0.$$

But

$$\mathbf{u}^{*N} \xrightarrow[\text{strongly}]{} \mathbf{u}^*, \quad \text{so} \quad V_N x \xrightarrow[\text{strongly}]{} \mathbf{u}^*.$$

To prove the relation (7.1) let us remark that

$$P_N x = \mathbf{C}_N[\mathbf{A}(V_N x) + \mathbf{B}x],$$
$$\mathbf{x}^* = \mathbf{A}\mathbf{u}^* + \mathbf{B}x.$$

Since

$$\mathbf{A}(V_N x) \xrightarrow[\text{strongly}]{} \mathbf{A}\mathbf{u}^*$$

and

$$\mathbf{C}_N \mathbf{B} x \xrightarrow[\text{strongly}]{} \mathbf{B} x$$

we get that

$$P_N x \xrightarrow[\text{strongly}]{} x^* \quad \text{in } l^2_\infty(H).$$

The proof of Theorem 7.1 is completed.

8. Final remarks. There exist many papers which developed a similar theory for the continuous-time problem (see [2], [9]). But the restriction to the discrete-time case enabled us to prove some stronger results in the sense that for instance strong or weak convergence in the formulation of appropriate theorems were replaced by the convergence in operator norm.

REFERENCES

[1] I. Daleckiĭ and M. Krein, *Stability of Solutions of Differential Equations in Banach Space*, Science, Moscow, 1970. (In Russian.)

[2] R. Datko, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

[3] ———, *Extending a theorem of A. M. Liapunov to Hilbert space*, J. Math. Anal. Appl., 32 (1970), pp. 610–616.

[4] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[5] N. Dunford and T. Schwartz, *Linear Operators, Part II*, Interscience, New York, 1963.

[6] M. A. Krasnosel'skiĭ, *Positive Solutions of Operator Equations*, P. Noordhoff, Groningen, the Netherlands, 1964.

[7] K. Y. Lee, S. Chow and R. Barr, *On the control of discrete-time distributed parameter systems*, this Journal, 10 (1972), pp. 361–376.

[8] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[9] D. L. Lukes and D. L. Russell, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

SIAM J. CONTROL
Vol. 12, No. 4, November, 1974

A NEW ALGORITHM FOR OPTIMAL FILTERING OF DISCRETE-TIME STATIONARY PROCESSES*

ANDERS LINDQUIST†

Abstract. An algorithm (which does not involve the usual Riccati-type equation) for computing the gain matrices of the Kalman filter is presented. If the dimension k of the state space is much larger than that of the observation process, the number of nonlinear equations to be solved in each step is of order k rather than k^2 as by the usual procedure.

1. Introduction. Let $\{x_n\}$ be a k-dimensional and $\{y_n\}$ an m-dimensional (wide sense) stationary stochastic vector process generated by the well-known model:

$$x_{n+1} = Fx_n + v_n, \tag{1.1}$$

$$y_n = Hx_n + w_n, \tag{1.2}$$

where, for convenience, x_0, $\{v_n\}$ and $\{w_n\}$ have zero mean and are pairwise uncorrelated, and

$$E\{x_0x_0'\} = P_0, \tag{1.3}$$

$$E\{v_iv_j'\} = P_1\delta_{ij}, \tag{1.4}$$

$$E\{w_iw_j'\} = P_2\delta_{ij} \tag{1.5}$$

(δ_{ij} is the Kronecker delta and $'$ denotes transpose). The matrices F, H, P_0, P_1 and P_2 are constant and have the appropriate dimensions. To simplify matters we assume that P_2 is positive definite (of course all P_i are nonnegative definite).

Now, it is well known that the linear least squares estimate $\hat{x}_n$ of x_n given $\{y_0, y_1, \cdots, y_{n-1}\}$ can be determined by the Kalman filter [5]:

$$\hat{x}_{n+1} = F\hat{x}_n + K_n(y_n - H\hat{x}_n) \tag{1.6}$$

with initial condition $\hat{x}_0 = 0$ and the gain matrix K_n given by

$$K_n = F\Sigma_nH'(H\Sigma_nH' + P_2)^{-1}. \tag{1.7}$$

Here the error covariance matrix

$$\Sigma_n = E\{(x_n - \hat{x}_n)(x_n - \hat{x}_n)'\} \tag{1.8}$$

can be recursively computed from the equation

$$\Sigma_{n+1} = F[\Sigma_n - \Sigma_nH'(H\Sigma_nH' + P_2)^{-1}H\Sigma_n]F' + P_1 \tag{1.9}$$

with initial condition $\Sigma_0 = P_0$. Therefore, this procedure requires computation of the (symmetric) $k \times k$ matrix Σ_n in each step in order to obtain the gain K_n,

* Received by the editors December 26, 1972. Material added to the original version of this paper is enclosed in square brackets.

† Institute of Optimization and Systems Theory, Royal Institute of Technology, 10044 Stockholm 70, Sweden. Now at Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506.

while actually the $k \times m$ matrix

$$Q_n = \Sigma_n H' \tag{1.10}$$

is needed. When, as is often the case, $m \ll k$, this amounts to computing plenty of unnecessary information.

In this paper we present an algorithm by which Q_n can be computed directly without using the Riccati-type equation (1.9). Instead of the $k(k + 1)/2$ equations of (1.9), we need only solve $2km + m(m + 1)/2$ equations, which is a major reduction when $m \ll k$. In the scalar output case ($m = 1$) we actually only need $2k$ equations.

[Before turning to the derivation of our algorithm we shall present some facts about the classical problem of determining the linear least squares estimate of y_n given $\{y_0, y_1, \cdots, y_{n-1}\}$ when $\{y_n\}$ is an *arbitrary* m-dimensional (wide sense) stationary stochastic sequence. Due to the stationarity it is possible to solve the normal equations recursively for the ($m \times m$ matrix) filter coefficients as n increases. For the scalar case ($m = 1$) such recursions can be found in [8], [9], [13] and also in the theory of orthogonal polynomials [1], [2].[1] The latter is of course no coincidence since the connection between prediction and orthogonal polynomials is well established [3]. When $m > 1$, the situation is somewhat more complicated (which accounts for the fact that the number of equations in our algorithm increases "discontinuously" as m becomes greater than 1). Recursive equations for this case can be found in [7], [12], [14]. However, a relation which is important for our purposes is missing in [7] and although this relation is mentioned in [12], [14], there is no proof for it. Therefore, in presenting a set of such equations we shall supply the reader with a short but complete proof. At the same time we shall be able to relate these equations to certain forward and backward prediction problems.]

In § 2 we shall introduce some notations and recall certain facts from estimation theory, in § 3 the abovementioned recursions for the filter coefficients will be developed, and in §§ 4 and 5 we shall return to what is the basic contribution of this paper, namely, the derivation of an algorithm for Q_n without making use of (1.9).

Independently, Kailath [4] has recently shown that (under certain conditions which are fulfilled in the stationary case) the Riccati equation for the continuous-time Kalman–Bucy filter can be factorized to yield equations similar to ours. Indeed, our method modified to the continuous-time case gives exactly the corresponding equations of Kailath, as we shall demonstrate in [10]. Similar results have also recently been announced by Rissanen [6], who, however, does not consider the model (1.1)–(1.2).

We have presented our algorithm for Q_n in connection with the one-step prediction problem (which is the standard problem in the literature). However, the algorithm can also be used for the pure filtering problem as pointed out in § 4.

[1] The equations in [1], [2] were first made known to us by R. E. Kalman, who suggested the theory of orthogonal polynomials studied in an algebraic context as a possible vehicle in obtaining a more effective algorithm. Our approach, however, is quite elementary in the sense that only facts of linear algebra are used. References [9], [12], [13], [14] were brought to our attention by a referee.

2. The forward and backward prediction problem. Let $y_0, y_1, y_2, y_3, \cdots$ be a wide sense stationary sequence of m-dimensional stochastic vectors with zero mean and covariances

$$C_i = E\{y_{n+i}y_n'\},$$

which, of course, are $m \times m$ matrices such that $C_{-i} = C_i'$. To simplify matters, we assume that this vector process has full rank in the sense that the generalized Toeplitz matrices

$$T_n = \begin{pmatrix} C_0 & C_1' & C_2' & \cdots & C_n' \\ C_1 & C_0 & C_1' & \cdots & C_{n-1}' \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ C_n & C_{n-1} & C_{n-2} & \cdots & C_0 \end{pmatrix},$$

$n = 0, 1, 2, \cdots$, are positive definite.

Now, for each $n \in \{1, 2, 3, \cdots\}$ we can define two problems of estimation, namely, the *forward* (P_n) and *backward* (P_n^*) one-step prediction problem.

Problem P_n. Find the linear least squares estimate $\hat{y}_n$ of y_n given $\{y_0, y_1, \cdots, y_{n-1}\}$.

Problem P_n^*. Find the linear least squares estimate $\hat{y}_n^*$ of y_0 given $\{y_1, y_2, \cdots, y_n\}$.

Clearly these estimates have the following form:

$$\hat{y}_n = -\sum_{i=0}^{n-1} \Phi_{ni}' y_i,$$

$$\hat{y}_n^* = -\sum_{i=1}^{n} \Phi_{ni}^{*\prime} y_i,$$

where Φ_{ni} and Φ_{ni}^* are $m \times m$ matrices. Then, by defining Φ_{nn} and Φ_{n0}^* to be unit matrices, the estimation errors can be written

$$\tilde{y}_n = y_n - \hat{y}_n = \sum_{i=0}^{n} \Phi_{ni}' y_i, \tag{2.1}$$

$$\tilde{y}_n^* = y_0 - \hat{y}_n^* = \sum_{i=0}^{n} \Phi_{ni}^{*\prime} y_i. \tag{2.2}$$

We introduce the following notations for the error covariances:

$$R_n = E\{\tilde{y}_n \tilde{y}_n'\}, \tag{2.3}$$

$$R_n^* = E\{\tilde{y}_n^* \tilde{y}_n^{*\prime}\}. \tag{2.4}$$

Now let E_n be the matrix formed by an infinite number of $m \times m$ matrices arranged in a vertical array with zero matrices in positions $0, 1, \cdots, n-1, n+1, n+2, \cdots$ and a unit matrix in position n:

$$E_n' = (0, \cdots, 0, I, 0, 0, \cdots).$$

Furthermore, given $X = \sum_{k=0}^{p} E_k X_k$ and $Y = \sum_{k=0}^{p} E_k Y_k$, where X_k, $k = 0, 1, \cdots, p < \infty$, and Y_k, $k = 0, 1, \cdots, p$, are $m \times m$ matrices (some of which may be zero), define the $m \times m$ matrix $[X, Y]$:

$$[X, Y] = X'TY,$$

where T is the infinite matrix T_∞. (Of course we could define $[X, Y]$ in terms of T_p, but we prefer an expression which is independent of p.) Then we have

$$[Y, X] = [X, Y]' \quad \text{and} \quad [E_i, E_j] = C_{i-j}. \tag{2.5}$$

Finally, introduce the shift operator σ:

$$\sigma^i X = \sum_{k=0}^{p} E_{k+i} X_k \qquad (i \geqq 0). \tag{2.6}$$

It is then easily seen that

$$[E_k, \sigma^i X] = [E_{k-i}, X] \qquad (k \geqq i) \tag{2.7}$$

and that

$$[\sigma X, \sigma Y] = [X, Y]. \tag{2.8}$$

We are now in a position to express some well-known orthogonality properties in terms of

$$\Phi_n = \sum_{k=0}^{n} E_k \Phi_{nk} \tag{2.9}$$

and

$$\Phi_n^* = \sum_{k=0}^{n} E_k \Phi_{nk}^*. \tag{2.10}$$

LEMMA 2.1.

$$[E_k, \Phi_n] = 0 \quad \textit{for } 0 \leqq k < n, \tag{2.11}$$

$$[E_n, \Phi_n] = R_n, \tag{2.12}$$

$$[\Phi_k, \Phi_l] = R_k \delta_{kl}. \tag{2.13}$$

Proof. Equations (2.11) and (2.12) follow from

$$[E_k, \Phi_n] = \sum_{i=0}^{n} C_{k-i} \Phi_{ni} = \sum_{i=0}^{n} E\{y_k y_i'\} \Phi_{ni} = E\{y_k \tilde{y}_n'\},$$

which by orthogonality is 0 for $k < n$ and R_n for $k = n$. We obtain (2.13) by observing that

$$[\Phi_k, \Phi_l] = \sum_{i=0}^{k} \Phi_{ki}' [E_i, \Phi_l],$$

which, by (2.11) and (2.12), is 0 for $k < l$ and R_k for $k = l$. Then it follows from (2.5) that (2.13) holds for $k > l$ also.

LEMMA 2.2.

$$[E_k, \Phi_n^*] = 0 \quad \textit{for } 0 < k \leqq n, \tag{2.14}$$

$$[E_0, \Phi_n^*] = R_n^*, \tag{2.15}$$

$$[\sigma^{n-k}\Phi_k^*, \sigma^{n-l}\Phi_l^*] = R_k^*\delta_{kl} \qquad (0 \leqq k, l \leqq n). \tag{2.16}$$

Proof. Equations (2.14) and (2.15) are obtained in the same way as (2.11) and (2.12), only replacing $\tilde{y}_n$ by $\tilde{y}_n^*$. To obtain (2.16), also observe (2.6) and (2.7) to see that (for $k \leqq l$)

$$[\sigma^{n-k}\Phi_k^*, \sigma^{n-l}\Phi_l^*] = \sum_{i=0}^{k} \Phi_{ki}^{*\prime}[E_{i+n-k}, \sigma^{n-l}\Phi_l^*] = \sum_{i=0}^{k} \Phi_{ki}^{*\prime}[E_{l-k+i}, \Phi_l^*].$$

Remark 2.1. Observe that Φ_n is *uniquely* determined by the system (2.11) of *normal equations*:

$$\sum_{i=0}^{n-1} C_{k-i}\Phi_{ni} = -C_{k-n}, \qquad k = 0, 1, \cdots, n-1, \tag{2.17}$$

for the coefficient matrix T_{n-1} is nonsingular. Likewise, Φ_n^* is *uniquely* determined by (2.14):

$$\sum_{i=1}^{n} C_{k-i}\Phi_{ni}^* = -C_k, \qquad k = 1, 2, \cdots, n, \tag{2.18}$$

which can also be written

$$\sum_{i=0}^{n-1} C_{k-i}'\Phi_{n,n-i}^* = -C_{k-n}', \qquad k = 0, 1, \cdots, n-1. \tag{2.19}$$

Remark 2.2. In the scalar case ($m = 1$) we have a particularly simple relationship between P_n and P_n^*, namely $\Phi_{ni}^* = \Phi_{n,n-i}$ and $R_n^* = R_n$. In fact, the first relation follows from (2.17) and (2.19) (for $C_i' = C_i$). Then the second relation is obtained by comparing (2.12) and (2.15).

Remark 2.3. Note that R_n and R_n^* are positive definite. In fact, observing that T_n is positive definite, this follows from $R_n = [\Phi_n, \Phi_n]$ and $R_n^* = [\Phi_n^*, \Phi_n^*]$. Clearly, R_n and R_n^* are also symmetric.

3. Difference equations for Φ_n and Φ_n^*.

LEMMA 3.1. *The following equations hold with the initial condition given by* $\Phi_0 = \Phi_0^* = E_0$:

$$\Phi_{n+1} = \sigma\Phi_n - \Phi_n^*\Gamma_n^*, \tag{3.1}$$

$$\Phi_{n+1}^* = \Phi_n^* - \sigma\Phi_n\Gamma_n, \tag{3.2}$$

where Γ_n *and* Γ_n^* *are* $m \times m$ *matrices defined by the following equations*:

$$R_n\Gamma_n = (R_n^*\Gamma_n^*)' = S_n, \tag{3.3}$$

where

$$S_n = [\sigma\Phi_n, \Phi_n^*] \tag{3.4}$$

$$= [\sigma\Phi_n, E_0] \tag{3.5}$$

$$= [E_{n+1}, \Phi_n^*]. \tag{3.6}$$

Proof. Let S_n be defined by (3.4). Then

$$S_n = \sum_{k=0}^{n} \Phi_{nk}'[E_{k+1}, \Phi_n^*],$$

which is equal to (3.6) by Lemma 2.2. Also, due to (2.7),

$$S_n = [\sigma\Phi_n, E_0] + \sum_{k=1}^{n} [\Phi_n, E_{k-1}]\Phi_{nk}^*,$$

which, by Lemma 2.1, is equal to (3.5).

To prove (3.1), first observe that Φ_{n+1} can be represented in the following form:

$$\Phi_{n+1} = E_{n+1} + \sum_{k=0}^{n} \sigma^{n-k}\Phi_k^* B_{nk},$$

where B_{nk} are $m \times m$ matrices. (In fact, this equation is equivalent to the following system:

$$\sum_{i=k}^{n} \Phi_{i,i-k}^* B_{ni} = \Phi_{n+1,n-k}, \qquad k = 0, 1, \cdots, n,$$

which can be solved for the B_{ni} matrices, for $\Phi_{i0}^* = I$.) By Lemma 2.2, we have

$$[\sigma^{n-i}\Phi_i^*, \Phi_{n+1}] = [\sigma^{n-i}\Phi_i^*, E_{n+1}] + R_i^* B_{ni} \qquad (0 \leqq i \leqq n).$$

The left member, being equal to $\sum_{k=0}^{i} \Phi_{ik}^{*\prime}[E_{k+n-i}, \Phi_{n+1}]$, is zero by Lemma 2.1, and due to (2.7) the first term on the right side equals S_i' as given by (3.6). Hence, since R_i^* is nonsingular, by (3.3), $B_{ni} = -\Gamma_i^*$. Then form $\Phi_{n+1} - \sigma\Phi_n$ to see that (3.1) holds.

Similarly (3.2) can be proved by considering the representation

$$\Phi_{n+1}^* = E_0 + \sum_{k=0}^{n} \sigma\Phi_k B_{nk}.$$

Then Lemma 2.1 (together with (2.8)) and (3.5) imply

$$[\sigma\Phi_i, \Phi_{n+1}^*] = S_i + R_i B_{ni},$$

the left member of which is zero (Lemma 2.2), and therefore $B_{ni} = -\Gamma_i$. Thus (3.2) holds. This concludes the proof of the lemma.

In the case $m = 1$ (i.e., the process $\{y_n\}$ is scalar) we have $\Phi_{nk}^* = \Phi_{n,n-k}$, $R_n^* = R_n$ (see Remark 2.2) and, consequently, $\Gamma_n^* = \Gamma_n$. The corresponding versions of (3.1) and (3.2) can be found in the theory of orthogonal polynomials (see [1, p. 183] or [2, p. 155]). In the general case ($m > 1$) similar equations can be found in [7], [12], [13]. However, [7] does not contain relation (3.3), and although this relation is mentioned in [12], [14], there is no proof for it.

LEMMA 3.2. *The error covariances R_n and R_n^* satisfy the following difference equations with $R_0 = R_0^* = C_0$:*

$$R_{n+1} = R_n - \Gamma_n^{*\prime} R_n^* \Gamma_n^*, \tag{3.7}$$

$$R_{n+1}^* = R_n^* - \Gamma_n' R_n \Gamma_n \tag{3.8}$$

[*or the uncoupled equations*

$$R_{n+1} = R_n(I - \Gamma_n\Gamma_n^*), \tag{3.9}$$

$$R_{n+1}^* = R_n^*(I - \Gamma_n^*\Gamma_n). \tag{3.10}$$

Also the following relations hold:

$$R_{n+1}\Gamma_n = (R_{n+1}^*\Gamma_n^*)' \tag{3.11}$$

and

$$(R_{n+1})^{-1} = (R_n)^{-1} + \Gamma_n(R_{n+1}^*)^{-1}\Gamma_n', \tag{3.12}$$

$$(R_{n+1}^*)^{-1} = (R_n^*)^{-1} + \Gamma_n^*(R_{n+1})^{-1}\Gamma_n^{*\prime}.]^2 \tag{3.13}$$

Proof. From (3.1) and (3.6) we have

$$[E_{n+1}, \Phi_{n+1}] = [E_{n+1}, \sigma\Phi_n] - S_n\Gamma_n^*$$

which, by (2.7) and Lemma 2.1, is the same as (3.7) or (3.9), depending on which of the two expressions (3.3) for S_n is used. Likewise (3.2) and (3.5) yield

$$[E_0, \Phi_{n+1}^*] = [E_0, \Phi_n^*] - S_n'\Gamma_n,$$

which is the same as (3.8) or (3.10) (Lemma 2.2). To obtain (3.11), postmultiply (3.7) and (3.8) by Γ_n and Γ_n^* respectively, and use (3.3). Finally, from (3.9) we have

$$(R_n)^{-1} = (I - \Gamma_n\Gamma_n^*)(R_{n+1})^{-1}$$

which, by (3.11), is equal to (3.12). (To see this, transpose (3.11), premultiply by $(R_{n+1}^*)^{-1}$ and postmultiply by $(R_{n+1})^{-1}$.) Equation (3.13) is derived in the same way.

4. An algorithm for the gain matrix. We now return to the problem described in § 1. Thus the innovation process (2.1) will be

$$\tilde{y}_n = H(x_n - \hat{x}_n) + w_n. \tag{4.1}$$

Our object is to determine the gain matrix (1.7):

$$K_n = FQ_n(HQ_n + P_2)^{-1}, \tag{4.2}$$

where Q_n, defined by (1.10) and (1.8), can be written

$$Q_n = E\{x_n\,\tilde{y}_n'\}. \tag{4.3}$$

Here we have first used the orthogonality between $\hat{x}_n$ and $(x_n - \hat{x}_n)$ to obtain $Q_n = E\{x_n(x_n - \hat{x}_n)'\}H'$ and then (4.1) and the fact that x and w are uncorrelated. By a similar argument, we can express the error covariance $R_n = E\{\tilde{y}_n\tilde{y}_n'\}$, defined in § 2, in terms of Q_n:

$$R_n = HQ_n + P_2. \tag{4.4}$$

Of course, (4.2), (4.3) and (4.4) can easily and in a well-known fashion be derived directly, and our reference to the equations in § 1 is merely for the purpose of comparison. Note in particular that we make no use of the Riccati equation (1.9).

[2] The corresponding part of the proof should also be bracketed.

Now, since $E\{v_j x_i'\} = 0$ for $j \geqq i$ and $E\{x_i x_i'\} = P_0$ (stationarity), (1.1) yields

$$E\{x_n x_i'\} = F^{n-i}P_0 \qquad (n \geqq i), \tag{4.5}$$

and therefore, remembering that $y_i = Hx_i + w_i$, we have

$$E\{x_n y_i'\} = F^{n-i}P_0 H' \qquad (n \geqq i) \tag{4.6}$$

(for x and w are uncorrelated) and

$$C_i = HF^i P_0 H' + P_2 \delta_{i0} \qquad (i \geqq 0). \tag{4.7}$$

Inserting (2.1) into (4.3) and observing (4.6), we obtain

$$Q_n = \sum_{i=0}^{n} F^{n-i}P_0 H' \Phi_{ni}. \tag{4.8}$$

Then, if we define

$$Q_n^* = \sum_{i=0}^{n} F^{n+1-i}P_0 H' \Phi_{ni}^*, \tag{4.9}$$

we can exploit Lemma 3.1 to obtain

$$Q_{n+1} = Q_n - Q_n^* \Gamma_n^*, \tag{4.10}$$

$$Q_{n+1}^* = FQ_n^* - FQ_n \Gamma_n, \tag{4.11}$$

where $Q_0 = P_0 H'$ and $Q_0^* = FP_0 H'$.

Furthermore, from (3.6) we have

$$S_n = \sum_{i=0}^{n} C_{n+1-i} \Phi_{ni}^*,$$

which by (4.7) and (4.9) equals

$$S_n = HQ_n^*. \tag{4.12}$$

This enables us to determine Γ_n and Γ_n^* from (3.3):

$$\Gamma_n = R_n^{-1} HQ_n^*, \tag{4.13}$$

$$\Gamma_n^* = (R_n^*)^{-1} Q_n^{*\prime} H'. \tag{4.14}$$

By (4.4), R_n can be expressed in terms of Q_n, while for R_n^* we must employ the recursion (3.8) of Lemma 3.2.

Hence we are now in a position to state our main result.

THEOREM 4.1. *The optimal gain matrix for the filter* (1.6) *can be determined in the following way*:

$$K_n = FQ_n(HQ_n + P_2)^{-1}, \tag{4.2}$$

where

$$Q_{n+1} = Q_n - Q_n^*(R_n^*)^{-1} Q_n^{*\prime} H', \tag{4.15}$$

$$Q_{n+1}^* = FQ_n^* - FQ_n(HQ_n + P_2)^{-1} HQ_n^*, \tag{4.16}$$

$$R_{n+1}^* = R_n^* - Q_n^{*\prime} H'(HQ_n + P_2)^{-1} HQ_n^*, \tag{4.17}$$

with initial conditions $Q_0 = P_0 H'$, $Q_0^* = FP_0 H'$ *and* $R_0^* = HP_0 H' + P_2$.

Remark 4.1. Note that Q_n and Q_n^* are $k \times m$ matrices and R_n^* is a symmetric $m \times m$ matrix. Thus we have $2km + [m(m + 1)]/2$ equations to determine Q_n.

[*Remark* 4.2. Since only the inverse of R_n^* is needed, we may replace (4.17) by

$$(4.18)\quad (R_{n+1}^*)^{-1} = (R_n^*)^{-1} + (R_n^*)^{-1}Q_n^{*\prime}H'(HQ_{n+1} + P_2)^{-1}HQ_n^*(R_n^*)^{-1}.$$

To see this, just insert (4.14) and (4.4) into (3.13). Equation (4.18) can also be obtained directly from (4.17) by applying the matrix inversion lemma[3] and noticing that $R_n = HQ_n + P_2$ is given by

$$(4.19)\qquad R_{n+1} = R_n - HQ_n^*(R_n^*)^{-1}Q_n^{*\prime}H'$$

with initial condition $R_0 = HP_0H' + P_2$.]

Remark 4.3. Equations (4.15), (4.16) and (4.17) can also be used for the pure filtering problem to determine the linear least squares estimate of x_n given $\{y_0, y_1, \cdots, y_n\}$. In fact, it is well known that we now have the following filtering equation (which of course is derived without resort to the Riccati equation):

$$(4.20)\qquad \hat{x}_n = F\hat{x}_{n-1} + L_n(y_n - HF\hat{x}_{n-1})$$

with initial condition $\hat{x}_0 = 0$, where the gain L_n is given by

$$(4.21)\qquad L_n = Q_n(HQ_n + P_2)^{-1}.$$

[*Remark* 4.4. We have made an effort to present our algorithm for Q_n in a compact form using as few equations as possible. Equations (4.15), (4.16) and (4.17) contain all the information needed for determining the gain sequences K_n and L_n. However, as usual, a certain judgment has to be exercised in implementing our algorithm. Computational requirements call for minimizing the number of arithmetic operations (see, e.g., [11] for details), and different considerations have to be made for the one-step predictor and for the pure filter. The reader should convince himself that *in general* Table 1 describes the natural implementation of our algorithm (when $m \ll k$), although the number of equations has increased. For example, instead of computing the quantities $R_n = HQ_n + P_2$ from Q_n in each step, amending the projected equation (4.19) = (3.7) (which of course is contained in (4.15)) usually (but not always) reduces the number of arithmetic operations. Also, which is even more important, there should be a minimum of multiplications by the large matrix F. We have introduced some auxiliary variables in addition to the ones defined in the text, $U_n \equiv (R_n)^{-1}$, $U_n^* \equiv (R_n^*)^{-1}$, $\bar{Q}_n \equiv FQ_n$ and $\bar{Q}_n^* \equiv FQ_n^*$ (the last two used for the one-step predictor only). However, it should be noted that special properties of the system's matrices *may* call for some other implementation of the algorithm. For example, with a sparse F (e.g., a companion matrix) the multiplication by F becomes less critical.]

[3] The author would like to thank Prof. I. H. Rowe (among others) for suggesting this. (This remark was communicated to the editor on March 20, 1973.)

TABLE 1

Pure filtering		One-step prediction
$Q_0 = P_0H'$		$\bar{Q}_0 = FP_0H'$
	$Q_0^* = FP_0H'$	
	$R_0 = HP_0H' + P_2$	
	$U_0 = U_0^* = R_0^{-1}$	
$L_n = Q_nU_n$		$K_n = \bar{Q}_nU_n$
	$S_n = HQ_n^*$	
	$\Gamma_n^* = U_n^*S_n'$	
		$\bar{Q}_n^* = FQ_n^*$
$Q_{n+1} = Q_n - Q_n^*\Gamma_n^*$		$\bar{Q}_{n+1} = \bar{Q}_n - \bar{Q}_n^*\Gamma_n^*$
$Q_{n+1}^* = F(Q_n^* - L_nS_n)$		$Q_{n+1}^* = \bar{Q}_n^* - K_nS_n$
	$R_{n+1} = R_n - S_n\Gamma_n^*$	
	$U_{n+1} = (R_{n+1})^{-1}$	
	$U_{n+1}^* = U_n^* + \Gamma_n^*U_{n+1}\Gamma_n^{*\prime}$	

5. The scalar output case. In the case $m = 1$ we have a somewhat simpler situation. Since $R_n^* = R_n$, which is given by (4.4), equation (4.17) now becomes superfluous, and therefore we end up with $2k$ equations. (Also note that S_n is now a scalar.)

We can also write our equations directly in terms of the gain vector k_n without increasing the number of equations[4]

$$(5.1) \qquad k_{n+1} = [1 - (h'k_n^*)^2]^{-1}[k_n - (h'k_n^*)Fk_n^*],$$

$$(5.2) \qquad k_{n+1}^* = [1 - (h'k_n^*)^2]^{-1}[Fk_n^* - (h'k_n^*)k_n],$$

with initial conditions $k_0 = k_0^* = (h'P_0h + P_2)^{-1}FP_0h$, where we write H as h' to emphasize that it is a vector.

In fact, observe that $k_n = FQ_nR_n^{-1}$. Then define $k_n^* = Q_n^*R_n^{-1}$, from which we have $\Gamma_n = h'k_n^*$. Therefore, (3.8) gives

$$R_{n+1} = [1 - (h'k_n^*)^2]\,R_n,$$

and (4.10) and (4.11) yield the desired result. (Since R_{n+1} and R_n are both positive, so is $[1 - (h'k_n^*)^2]$, and therefore we can safely divide by this quantity.)

[The equations can be simplified at the expense of the "symmetry" by adding $(h'k_n^*)$ times (5.2) to (5.1):

$$(5.3) \qquad k_{n+1} = k_n - (h'k_n^*)k_{n+1}^*,$$

$$(5.4) \qquad k_{n+1}^* = [1 - (h'k_n^*)^2]^{-1}[Fk_n^* - (h'k_n^*)k_n],$$

but we should remember that computational requirements may call for retaining the original algorithm of § 4.]

Similar equations can be obtained for the pure filtering problem.

[4] [We can obtain similar equations for $m > 1$ if we amend the equations for both R_n and $R_n{}^*$.]

REFERENCES

[1] N. I. Akhiezer, *The Classical Moment Problem*, Hafner, New York, 1965.

[2] L. Ya. Geronimus, *Orthogonal Polynomials*, Consultant Bureau, New York, 1961.

[3] U. Grenander and G. Szegö, *Toeplitz Forms and their Applications*, University of California Press, Berkeley and Los Angeles, 1958.

[4] T. Kailath, *Some Chandrasekhar-type algorithms for quadratic regulators*, IEEE Decision and Control Conference, Dec. 1972.

[5] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Trans. ASME, Ser. D, 82 (1960), pp. 35–43.

[6] J. Rissanen, *A fast algorithm for optimum linear predictors*, Abstract, Oct. 1972.

[7] P. Whittle, *On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix*, Biometrica, 50 (1963), pp. 129–134.

[8] J. Durbin, *The fitting of time-series models*, Rev. Internat. Inst. Statist., 28 (1960), pp. 233–244.

[9] N. Levinson, *The Wiener RMS (root mean square) error in filter design and prediction*, Extrapolation, Interpolation and Smoothing of Stationary Time Series, Appendix B, N. Wiener, ed., MIT Press, Cambridge, Mass., 1942.

[10] A. Lindquist, *Optimal filtering of continuous-time stationary processes by means of the backward innovation process*, this Journal, 12 (1974), pp. 747–754.

[11] J. M. Mendel, *Computational requirements for a discrete Kalman filter*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 748–758.

[12] E. A. Robinson, *Multichannel time series analysis with digital computer programs*, San Francisco, 1967.

[13] ———, *Statistical Communication and Detection with Special Reference to Digital Data Processing of Radar and Seismic Signals*, Hafner, New York, 1967.

[14] R. A. Wiggins and E. A. Robinson, *Recursive solutions to the multichannel filtering problem*, J. Geophys. Res., 70 (1965), pp. 1885–1891.

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

OPTIMAL FILTERING OF CONTINUOUS-TIME STATIONARY PROCESSES BY MEANS OF THE BACKWARD INNOVATION PROCESS*

ANDERS LINDQUIST†

Abstract. A new approach to linear least squares estimation of continuous-time (wide sense) stationary stochastic processes is presented. The basic idea is that the relevant estimates can be expressed not only in terms of the usual (forward) innovation process but also in terms of a backward innovation process. The functions determining the optimal filter as well as the error covariance functions are seen to satisfy some differential equations. As an important example the Kalman–Bucy filter is considered. It is demonstrated that the optimal gain matrix can be determined from $2mn$ equations (where n is the dimension of the system and m of the output) rather than $\frac{1}{2}n(n + 1)$ as in the conventional theory. This is an advantage when, as is usually the case, $m \ll n$. These equations were first derived by Kailath, who used a different method. Also they are the continuous-time versions of some equations previously obtained (independently of Kailath) by the author.

1. Introduction. In this paper we consider linear least squares filtering of wide sense stationary stochastic vector processes, where the estimation is based on past observations of the process on an increasing but *finite* time interval. Since therefore the filtering estimate will be a nonstationary process, the weighting function of the filter will be a function of two time variables rather than one as in classical Wiener theory, where observations from the infinite past are assumed to be available. This weighting function satisfies a generalized Wiener–Hopf equation for which no general method of solution is known. Since this is also the case when the process to be estimated is nonstationary, it may seem unnecessarily restrictive to assume stationarity. However, it turns out that this assumption will enable us to give simple differential equations for the weighting function and the error covariance function. These equations are completely characterized by the covariance between the estimation error process and the initial value of the estimated process, and therefore, at least in theory, we have reduced the problem to determining this function of one variable.

The usefulness of our results becomes apparent when applying them to Kalman–Bucy filtering [4] of wide sense stationary processes. It is well known that the computation of the "gain matrix" for such a filter requires the solution of an $n \times n$ matrix Riccati differential equation, where n is the dimension of the system. The number n is usually much larger than the dimension m of the observed process. Our approach will yield $2mn$ nonlinear differential equations instead of the $\frac{1}{2}n(n + 1)$ of the Riccati equation and therefore will supply a more effective algorithm for the gain matrix whenever $m \ll n$. The same equations were recently presented by Kailath [3], who derived them directly from the Riccati equation. However, our approach helps to reveal the fact that the property of the error covariance matrix which makes Kailath's method work holds for wide sense stationary stochastic processes in general, and not only for those realized by a finite-dimensional linear stochastic system.

* Received by the editors February 12, 1973.

† Department of Mathematics, University of Florida, Gainesville, Florida. Now at Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506.

In order to obtain our results, in § 3 we define the *backward innovation processes*. This is in the spirit of a previous paper [6] by the author on filtering of discrete-time processes, and we shall find some similarities in the structures of analogous equations, but also some important differences. So, for example, the equations for the "Kalman gain" are somewhat more complicated in the discrete-time case in that an equation for the "backward error covariance" is required. The reason for this, of course, is that in the continuous-time case the innovation processes can be defined (as we have done) to have constant incremental covariances, while in the discrete-time case they are identical to the error processes, for which no such constancy holds. (Hence the discrete-time counterpart of (4.7) also contains the backward error covariance matrix.)

2. Preliminaries. Let $z(t)$ be the m-dimensional stochastic process

$$z(t) = \int_0^t y(\tau)\,d\tau + w(t) \tag{2.1}$$

defined on $[0, T]$, where $y(t)$ is a zero mean vector process[1] such that

$$\int_0^T E|y(t)|^2\,dt < \infty,$$

and $w(t)$ is a process with zero mean and covariance function

$$E\{w(s)w(t)'\} = I \min(s, t) \tag{2.2}$$

(prime denotes transpose), which implies that $w(t)$ has *orthogonal increments.*

If H is the Hilbert space of all second order stochastic variables (of course, we have tacitly assumed an underlying probability space (Ω, B, P)) with inner product $(\xi, \eta) = E\{\xi\eta\}$, then define $H_t(z)$ to be the closed linear hull in H of the stochastic variables $\{z_i(s);\ 0 \leqq s \leqq t,\ i = 1, 2, \cdots, m\}$. Furthermore, for any $\xi \in H$, let $\hat{E}_t^z\xi$ denote the projection of ξ onto $H_t(z)$, i.e., the wide sense conditional mean of ξ given $\{z(s);\ 0 \leqq s \leqq t\}$. If x is a vector with components $x_i \in H$, we shall take $\hat{E}_t^z x$ to mean the vector with components $\hat{E}_t^z x_i$.

We shall need a few results from linear filtering theory which in the present form are essentially due to Kailath. Denote $\hat{E}_t^z y(t)$ by $\hat{y}(t)$ and define the innovation process

$$v(t) = z(t) - \int_0^t \hat{y}(\tau)\,d\tau, \tag{2.3}$$

for which we have the following lemmas.

LEMMA 2.1. *The process $v(t)$ has zero mean and covariance function* (2.2) *and hence orthogonal increments. Moreover,*

$$H_t(v) = H_t(z), \qquad 0 \leqq t \leqq T. \tag{2.4}$$

LEMMA 2.2. *Let x be a stochastic vector with components in H, and let $v(t)$ be a zero mean vector process with orthogonal increments and covariance function*

[1] In order to take full advantage of integration theory, we assume that all stochastic processes defined are measurable in (t, ω).

(2.2). *Then*

$$\hat{E}_t^v x = \int_0^t \frac{d}{ds} E\{xv(s)'\}\, dv(s). \tag{2.5}$$

Informal versions of these lemmas have appeared in a series of papers by Kailath on the "innovation method". For rigorous proofs see Kailath [2] or Lindquist [5].

3. The backward innovation processes. For the moment assuming that $t \in [0, T]$ is fixed, define the following stochastic processes for $s \in [0, t]$:

$$y_t(s) = y(t - s), \tag{3.1}$$

$$z_t(s) = z(t) - z(t - s), \tag{3.2}$$

$$w_t(s) = w(t) - w(t - s), \tag{3.3}$$

where y, x and w are the processes defined in § 2. Then equation (2.1) gives us

$$z_t(s) = \int_0^s y_t(\tau)\, d\tau + w_t(s), \tag{3.4}$$

which is an equation of the same type as (2.1), for it is immediately clear that

$$E\{w_t(s) w_t(\tau)'\} = I \min(s, \tau). \tag{3.5}$$

Therefore $w_t(s)$ has orthogonal increments for each fixed t. Also it is clear that

$$z(s) = z_t(t) - z_t(t - s) \tag{3.6}$$

and therefore

$$H_t(z_t) = H_t(z). \tag{3.7}$$

Now, introducing the notation $\hat{y}_t(s) = \hat{E}_s^{z_t} y_t(s)$, we consider the innovation process corresponding to (3.4):

$$v_t(s) = z_t(s) - \int_0^s \hat{y}_t(\tau)\, d\tau, \tag{3.8}$$

which we shall call the *backward innovation process* for $\{z(s); 0 \leqq s \leqq t\}$. Clearly we have one such process for each $t \in [0, T]$. The following lemma is then an immediate consequence of Lemma 2.1 and equation (3.7).

LEMMA 3.1. *For each fixed t, $v_t(s)$ has zero mean and covariance function* (3.5), *and hence orthogonal increments. Moreover,*

$$H_s(v_t) = H_s(z_t), \qquad 0 \leqq s \leqq t, \tag{3.9}$$

and, in particular,

$$H_t(v_t) = H_t(z_t) = H_t(z) = H_t(v). \tag{3.10}$$

Therefore, whenever we wish to determine a linear least squares estimate based on the data $\{z(s); 0 \leqq s \leqq t\}$, we can also express it in terms of z_t, v or v_t, whichever we find appropriate.

4. An equation for the error covariance. Let $z(t)$ be the m-dimensional data process defined by (2.1) and $x(t)$ a related n-dimensional zero mean stochastic process such that the compound process $(x(t), y(t))$ is *wide sense stationary* with $E\{x(t)x(t)'\} = P_0$. Also, to simplify matters, assume that (x, y) and w are uncorrelated.

Now, our problem is to determine the linear least squares estimate $\hat{x}(t) = \hat{E}_t^z x(t)$, and in the process of doing so we are interested in the estimation error covariance function

$$P(t) = E\{\tilde{x}(t)\tilde{x}(t)'\}, \tag{4.1}$$

where $\tilde{x}(t) = x(t) - \hat{x}(t)$.

To this end, we recall the well-known fact that (for a fixed t) the filtering estimate $\hat{y}_t(s)$ of $y_t(s)$ given the data $\{z_t(\tau); 0 \leqq \tau \leqq s\}$ is

$$\hat{y}_t(s) = \int_0^s G(s, \tau)\, dz_t(\tau), \tag{4.2}$$

where G is a function (defined through a generalized Wiener–Hopf equation) only of C_t, where $C_t(\tau, s) = E\{y_t(\tau)y_t(s)'\}$. However, due to the stationarity, $C_t(\tau, s) = E\{y(s)y(\tau)'\}$ does not depend on the parameter t, and hence G is good for all $t \in [0, T]$. Therefore we have

$$\begin{aligned} E\{x(t)\hat{y}_t(s)'\} &= E\left\{x(t)\left[\int_0^s G(s, \tau)y(t - \tau)\, d\tau\right]'\right\} \\ &= E\left\{x(s)\left[\int_0^s G(s, \tau)y(s - \tau)\, d\tau\right]'\right\} \\ &= E\{x(s)\hat{y}_s(s)'\}, \end{aligned} \tag{4.3}$$

where we have used the fact that (x, y) is wide sense stationary and x and w are uncorrelated.

We are now in a position to apply Lemmas 2.2 and 3.1 to see that

$$\hat{x}(t) = \int_0^t Q(t, s)\, dv_t(s), \tag{4.4}$$

where

$$\begin{aligned} Q(t, s) &= E\{x(t)[y_t(s) - \hat{y}_t(s)]'\} \\ &= E\{x(s)[y_s(s) - \hat{y}_s(s)]'\} \\ &= E\{\tilde{x}(s)y(0)'\}, \end{aligned} \tag{4.5}$$

where again we have exploited the stationarity and uncorrelatedness properties mentioned above, relation (4.3), and also the fact that $\hat{x}(s)$ and $y_s(s) - \hat{y}_s(s)$ are orthogonal and that the same is true for $\tilde{x}(s)$ and $\hat{y}_s(s)$. Hence $Q(t, s)$ does not depend on t, and we shall therefore call it $Q(s)$:

$$\hat{x}(t) = \int_0^t Q(s)\, dv_t(s). \tag{4.6}$$

Then we have the error covariance

$$\begin{aligned} P(t) &= E\{x(t)x(t)'\} - E\{\hat{x}(t)\hat{x}(t)'\} \\ &= P_0 - \int_0^t Q(s)Q(s)'\,ds, \end{aligned} \tag{4.7}$$

which concludes the proof of the following theorem.

THEOREM 4.1. *With the conditions imposed in the beginning of this section, the error covariance* (4.1) *satisfies the matrix differential equation*

$$\begin{aligned} \dot{P}(t) &= -Q(t)Q(t)', \\ P(0) &= P_0, \end{aligned} \tag{4.8}$$

where

$$Q(t) = E\{\tilde{x}(t)y(0)'\}. \tag{4.9}$$

Here the dynamics of the $n \times m$ matrix function Q is essentially that of the error signal $\tilde{x}$, so to proceed we have to impose further conditions on the process $x(t)$. We choose to illustrate this by applying Theorem 4.1 to the Kalman–Bucy filter.

Assume that the n-dimensional wide sense stationary process $x(t)$ is given by the stochastic differential equation

$$dx = Ax\,dt + B\,dv, \qquad x(0) = x_0, \tag{4.10}$$

and the m-dimensional data process by

$$dz = Hx\,dt + dw, \qquad z(0) = 0,$$

so that $y(t)$ is in fact equal to $Hx(t)$. Here $v(t)$ is a vector process of type (2.2), x_0 is a zero mean stochastic variable, and x_0, v and w are pairwise uncorrelated. The matrices A, B and H are constant.

Now, it is well known [4] that the filtering estimate $\hat{x}(t)$ is generated by

$$d\hat{x} = A\hat{x}\,dt + K(dz - H\hat{x}\,dt), \qquad \hat{x}(0) = 0, \tag{4.11}$$

where the "gain-matrix" function K is given by

$$K(t) = P(t)H'. \tag{4.12}$$

The $n \times n$ matrix function P is usually determined from a matrix Riccati equation, which amounts to solving $\frac{1}{2}n(n + 1)$ nonlinear differential equations, in order to obtain the nm functions in the gain matrix K. Our procedure yields $2mn$ equations, which is a major advantage whenever, as is often the case, $m \ll n$. To see this, first observe that the error process $\tilde{x}(t)$ is given by

$$d\tilde{x} = (A - KH)\tilde{x}\,dt + B\,dv - dw \tag{4.13}$$

with initial condition $\tilde{x}(0) = x_0$. Moreover, $y(0) = Hx_0$, which is uncorrelated with v and w. It is then easy to see that

$$\dot{Q} = (A - KH)Q, \qquad Q(0) = P_0H',$$

and that we therefore have the following $2mn$ equations to determine the optimal gain:

$$\begin{aligned} \dot{K}(t) &= -Q(t)Q(t)'H', \\ \dot{Q}(t) &= (A - K(t)H)Q(t), \end{aligned} \tag{4.14}$$

with initial conditions $K(0) = Q(0) = P_0H'$. Clearly, there exists a *unique* solution of the system (4.14). Indeed, establishing this is a standard exercise in the use of the contraction mapping principle.

Equations (4.14) have also been obtained by Kailath [3] by differentiating the Riccati equation. However, unlike Kailath's method, ours completely avoids the Riccati equation. It also demonstrates the fact that the low rank property (in the interesting case $m \ll n$) of $\dot{P}$ is not only a property of "lumped" stationary processes (4.10) but is one of stationary processes in general.

5. Differential equations for the weighting function. Consider the problem to determine the linear least squares estimate $\hat{y}(t) = \hat{E}_t^z y(t)$, where as before the data process $z(t)$ is given by (2.1):

$$z(t) = \int_0^t y(\tau)\, d\tau + w(t). \tag{5.1}$$

We assume that y is wide sense stationary and that y and w are uncorrelated. Then as we pointed out in § 4, we have

$$\hat{y}(t) = \int_0^t F(t, t - s)\, dz(s), \tag{5.2}$$

where F is the weighting function to be determined. This function is known to satisfy a Fredholm integral equation (a generalized Wiener–Hopf equation), but we shall demonstrate that it also satisfies a system of differential equations.

Now it is easily seen that we can rewrite (5.2) in terms of the backward data process (3.2) to obtain

$$\hat{y}(t) = \int_0^t F(t, s)\, dz_t(s). \tag{5.3}$$

Also we define the backward weighting function F^* by the equation

$$\hat{y}_t(s) = \int_0^s F^*(s, s - \tau)\, dz_t(\tau). \tag{5.4}$$

As we pointed out in § 4, the stationarity insures that F^* is the same for all values of the parameter t. Moreover,

$$\hat{y}_t(t) = \int_0^t F^*(t, \tau)\, dz(\tau). \tag{5.5}$$

We can also express $\hat{y}$ and $\hat{y}_t$ in terms of the innovation processes v and v_t.

In fact, equation (4.6) yields for the case $x = y$:

$$\hat{y}(t) = \int_0^t \Gamma(s)\, dv_t(s), \tag{5.6}$$

where

$$\Gamma(t) = E\{\tilde{y}(t)y(0)'\} \tag{5.7}$$

or, introducing the notation $\tilde{y}_t(s) = y_t(s) - \hat{y}_t(s)$,

$$\Gamma(t) = E\{\tilde{y}(t)\tilde{y}_t(t)'\}. \tag{5.8}$$

Also, a straightforward application of Lemma 2.2 (with $v = \nu$) yields

$$\hat{y}_t(t) = \int_0^t \Gamma(s)'\, dv(s). \tag{5.9}$$

Hence, by (3.8),

$$\begin{aligned}\hat{y}(t) &= \int_0^t \Gamma(s)\left[dz_t(s) - \int_0^s F^*(s, s - \tau)\, dz_t(\tau)\, ds\right]\\ &= \int_0^t \left[\Gamma(s) - \int_s^t \Gamma(\tau)F^*(\tau, \tau - s)\, d\tau\right] dz_t(s)\end{aligned}$$

and, by (2.3),

$$\begin{aligned}\hat{y}_t(t) &= \int_0^t \Gamma(s)'\left[dz(s) - \int_0^s F(s, s - \tau)\, dz(\tau)\, ds\right]\\ &= \int_0^t \left[\Gamma(s)' - \int_s^t \Gamma(\tau)'F(\tau, \tau - s)\, d\tau\right] dz(s).\end{aligned}$$

(The change of the order of integration is permitted due to a Fubini-type theorem for stochastic integrals. See, e.g., [1, p. 197].) It is clear from these expressions that

$$F(t, s) = \Gamma(s) - \int_s^t \Gamma(\tau)F^*(\tau, \tau - s)\, d\tau, \tag{5.10}$$

$$F^*(t, s) = \Gamma(s)' - \int_s^t \Gamma(\tau)'F(\tau, \tau - s)\, d\tau. \tag{5.11}$$

Now, the following theorem is a continuous-time analogue of Lemmas 3.1 and 3.2 in [6].

THEOREM 5.1. *The weighting functions F and F* satisfy the following differential equations*:

$$\frac{\partial F}{\partial t}(t, s) = -\Gamma(t)F^*(t, t - s), \tag{5.12}$$

$$\frac{\partial F^*}{\partial t}(t, s) = -\Gamma(t)'F(t, t - s) \tag{5.13}$$

for $t \geqq s$, *with initial conditions* $F(s,s) = \Gamma(s)$ *and* $F^*(s,s) = \Gamma(s)'$, *where* Γ *is defined by* (5.7) *or* (5.8). *The error covariances* $R(t) = E\{\tilde{y}(t)\tilde{y}(t)'\}$ *and* $R^*(t) = E\ \{\tilde{y}_t(t)\tilde{y}_t(t)'\}$ *satisfy*:

$$\dot{R}(t) = -\Gamma(t)\Gamma(t)', \tag{5.14}$$

$$\dot{R}^*(t) = -\Gamma(t)'\Gamma(t), \tag{5.15}$$

with initial conditions $R(0) = R^*(0) = E\{y(t)y(t)'\}$.

Moreover, if x *is the process of Theorem* 4.1 *and* N *is the weighting function defined by*

$$\hat{x}(t) = \int_0^t N(t, t-s)\,dz(s), \tag{5.16}$$

then N *satisfies*

$$\frac{\partial N}{\partial t}(t,s) = -Q(t)F^*(t, t-s) \tag{5.17}$$

for $t \geqq s$, *with initial condition* $N(s,s) = Q(s)$, *where* Q *is given by* (4.9).

Proof. Equations (5.12) and (5.13) follow from (5.10) and (5.11), and (5.17) is derived in the same way as (5.12) only exchanging (5.6) for (4.6). Finally, (5.14) and (5.15) are consequences of Theorem 4.1. To obtain (5.14), put $x = y$, and to obtain (5.15), put $x = y_t$ and exchange y for y_t.

As an example, we can now apply equations (5.12) and (5.13) to obtain an alternative derivation of equations (4.14). In fact, by (5.2) and (5.5), $K(t) = E\{x(t)\tilde{y}(t)'\}$ and $Q(t) = E\{x(t)\tilde{y}_t(t)'\}$ can be expressed in terms of F and F^* respectively. Also observe that $E\{x(t)y(s)'\} = e^{A(t-s)}P_0H'$ and finally that $\Gamma(t) = HQ(t)$.

REFERENCES

[1] I. I. Gikman and A. V. Skorokhod, *Introduction to the Theory of Random Processes*, Saunders, London, 1969.

[2] T. Kailath, *A note on least squares estimation by the innovations method*, this Journal, 10 (1972), pp. 477–486.

[3] ———, *Some Chandrasekhar-type algorithms for quadratic regulators*, Proc. IEEE Conference on Decision and Control, New Orleans, December 1972.

[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] A. Lindquist, *Optimal control of linear stochastic systems with applications to time lag systems*. Information Sci., 5 (1973), pp. 81–126.

[6] ———, *A new algorithm for optimal filtering of discrete-time stationary processes*, this Journal, 12 (1974), pp. 736–746.

SIAM J. CONTROL
Vol. 12, No. 4, November 1974

HILBERT NETWORKS. I*

VACLAV DOLEZAL†

Abstract. In this paper there is constructed a simple model of a general nonlinear network. The network considered consists of at most countably many lumped elements described by (not necessarily linear) operators from a Hilbert space into itself. Several theorems are proved on the existence and uniqueness of the solution of the network. Also, it is shown that, under certain conditions, the solution of an infinite network is a limit of solutions of finite subnetworks.

Introduction. In the last two decades, considerable attention has been paid to networks consisting of finitely many nonlinear elements. The most profound results in this respect have been obtained by Minty [1]. The networks considered in [1] are assumed to consist of finitely many nonlinear resistors which are described by monotone multivalued functions. On the other hand, an abstract model of a network was constructed by Roth [11].

While the approach developed in [1] is purely "network-theoretic", i.e., it starts from the description of the network by its graph and its elements, many results were also obtained that view the network as a simple interconnection of blocks, each of which is described by some operator on a Hilbert space or possibly on an extended Hilbert space. In this context, let us mention the paper by Zames [2], which considers mostly feedback systems and establishes relations between stability and passivity or monotonicity, and several papers by Sandberg. A nice survey of results dealing with systems described by causal operators on extended Banach or Hilbert spaces is given in Sandberg's paper [3]. There are also scores of papers dealing with systems which have a particular structure; let us mention only the paper by Zames and Falb [4].

On the other hand, very recently efforts have been made to investigate networks consisting of infinitely many elements. Pioneering results in this field were obtained by Flanders [5] and Zemanian [6]. While results obtained in [5] apply mostly to networks with constant linear resistors under a direct current regime, [6] is concerned with infinite networks whose elements are described by certain linear operators on some Hilbert space. However, it is assumed that the network can be partitioned into subnetworks each of which has an R- or L- or C-network-like character.

In the present paper we attempt to unify all the approaches to the network problem indicated above. To be more specific, we assume that the network elements are described by (in general nonlinear) operators from some Hilbert space H into itself; thus, within this framework, currents and voltages are elements in H. Moreover, for the network model discussed it is immaterial whether it consists of finitely or infinitely many elements; also, we do not make any special assumptions

* Received by the editors March 22, 1973, and in revised form February 6, 1974.

† Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, New York 11790. This research was supported by the National Science Foundation under Grant PO33568-X00.

about the network structure except for the requirement that the graph of the network be locally finite.

Naturally, our network model is not "universal", i.e., there are physical systems which cannot be analyzed by our model. This is mainly due to the fact that, for the sake of simplicity, we have assumed that the operators describing the network elements are defined on the entire space. Thus, for example, our model does not apply to the case when the underlying space is $L_2[0, \tau]$, $\tau > 0$, and the network contains differentiators (inductors), unless the system of integro-differential equations governing the regime is replaced by a system of integral equations. Similarly, if we deal with the space $L_2[0, \infty)$, we encounter difficulties when integrators (capacitors) are present, unless we make some additional assumptions about the growth of capacitors as $t \to \infty$. However, as mentioned above, it does not matter whether finitely or infinitely many elements are present.

On the other hand, our model can be immediately applied to any specific R, L, C-network with constant elements in the frequency domain.

Fortunately, the above limitations are of a technical nature; they can be easily removed by constructing a more general (and, consequently, more complicated) model of a network. This will be done in a later paper.

Also, let us point out the following fact: since the analysis takes place in a Hilbert space, we assume a priori that all energies associated with the network are finite. This, however, appears to be a quite natural assumption.

As for the development of our theory, it is convenient to introduce first the concept of an abstract network $\mathcal{N}$ as a pair (Z, a) of certain operators. This is done in the first part of the paper. We define the solution of $\mathcal{N}$ in a fashion which is a straightforward generalization of Kirchhoff's laws, and establish several theorems on the existence and uniqueness of a solution.

In the second part we define a Hilbert network $\hat{\mathcal{N}}$ as a pair $(\hat{Z}, G)$, where G is an oriented graph (finite or infinite) and $\hat{Z}$ is an (impedance) operator describing the relations between currents and voltage drops in branches of $\hat{\mathcal{N}}$. It is shown that the concept of a solution of $\hat{\mathcal{N}}$ can be defined in two different but equivalent ways, i.e., either as a solution of the associated abstract network or as a solution of the classical Kirchhoff's laws. Due to this fact, theorems obtained in the first part apply to a Hilbert network. As a result, we obtain theorems on the existence and uniqueness of a solution of a Hilbert network, and some results which are extensions of certain laws in classical network theory. Finally, it is shown that if the impedance operator $\hat{Z}$ of an infinite Hilbert network $\hat{\mathcal{N}}$ satisfies a certain monotonicity condition, then the solution of $\hat{\mathcal{N}}$ is a limit of solutions of finite subnetworks of $\hat{\mathcal{N}}$.

1. Abstract networks. If N_1, N_2 are nonzero normed linear spaces, $[N_1, N_2]$ will denote the space of all linear continuous operators from N_1 into N_2; moreover, if $\alpha > 0$, then $\mathrm{Lip}_\alpha[N_1, N_2]$ will signify the set of all (not necessarily linear) operators $A : N_1 \to N_2$ such that

$$\|Ax_1 - Ax_2\| \leqq \alpha \|x_1 - x_2\| \tag{1.1}$$

for all $x_1, x_2 \in N_1$.

In the sequel, $\mathscr{H}$ and $\mathscr{H}'$ will be fixed Hilbert spaces, i.e., complex or real, complete (not necessarily separable) inner product spaces.

Let Z be a (not necessarily linear) operator from $\mathscr{H}$ into itself, and let $a \in [\mathscr{H}, \mathscr{H}']$, $a \neq 0$; then the ordered pair $\mathscr{N} = (Z, a)$ will be called an abstract network over $\mathscr{H}$. Also, Z will be called the impedance operator of $\mathscr{N}$, and a the structural operator of $\mathscr{N}$. If, in particular, Z is linear, $\mathscr{N}$ will be called linear.

DEFINITION. Let $\mathscr{N} = (Z, a)$ be an abstract network over $\mathscr{H}$, and let $e \in \mathscr{H}$; an element $i \in \mathscr{H}$ will be called a *solution of $\mathscr{N}$ corresponding to e* if

$$\mathrm{K}_1: \qquad \langle c, Zi - e\rangle = 0$$

for every $c \in \mathscr{H}$ such that $ac = 0$, and

$$\mathrm{K}_2: \qquad ai = 0.$$

A network $\mathscr{N}$ will be called *regular* if for every $e \in \mathscr{H}$ there exists a unique solution $i \in \mathscr{H}$ corresponding to e, i.e., if there exists an operator $A : \mathscr{H} \to \mathscr{H}$, called the admittance operator of $\mathscr{N}$, such that $i = Ae$ for each $e \in \mathscr{H}$.

For further purposes, let us introduce the following notation.

If a is the structural operator of $\mathscr{N}$, let

$$N_a = \{x : x \in \mathscr{H}, ax = 0\}; \tag{1.2}$$

clearly, N_a is a closed linear subspace of $\mathscr{H}$. Furthermore, let $N_a^\perp$ be the orthogonal complement of N_a in $\mathscr{H}$, and let P be the orthogonal projection of $\mathscr{H}$ onto N_a.

If $N_a = \{0\}$, then $\mathscr{N}$ is clearly regular and $i = 0$ is the unique solution of $\mathscr{N}$ corresponding to any chosen $e \in \mathscr{H}$; thus, in the sequel we will assume that $N_a \neq \{0\}$.

Observe that, using this notation, a solution of $\mathscr{N}$ corresponding to $e \in \mathscr{H}$ can be defined as an element $i \in \mathscr{H}$ which satisfies the relations

$$\mathrm{K}_1^*: \qquad Zi - e \in N_a^\perp,$$

$$\mathrm{K}_2^*: \qquad i \in N_a.$$

The following proposition gives necessary and sufficient conditions for a network to be regular.

THEOREM 1.1. *Let $\mathscr{N} = (Z, a)$ be an abstract network over $\mathscr{H}$, and let N_a, P have the meaning defined above; furthermore, let $\tilde{Z}$ be the restriction of Z to N_a. Then $\mathscr{N}$ is regular iff the operator $P\tilde{Z} : N_a \to N_a$ possesses an inverse $(P\tilde{Z})^{-1} : N_a \to N_a$. In this case, the admittance operator A of $\mathscr{N}$ is given by*

$$A = (P\tilde{Z})^{-1}P. \tag{1.3}$$

If, in addition, $\mathscr{N}$ is linear, then A is linear, and $A \in [\mathscr{H}, \mathscr{H}]$ provided $Z \in [\mathscr{H}, \mathscr{H}]$.

Proof. (a) Assume first that $(P\tilde{Z})^{-1}$ exists, and define A by (1.3). Choose $e \in \mathscr{H}$ and let $i = Ae$. Since $A : \mathscr{H} \to N_a$, we have $i \in N_a$ and K_2^* holds. Next, let $c \in N_a$; since $Pc = c$, we have

$$\begin{aligned}\langle c, Zi - e\rangle &= \langle Pc, Z(P\tilde{Z})^{-1}Pe - e\rangle \\ &= \langle Pc, \tilde{Z}(P\tilde{Z})^{-1}Pe - e\rangle \\ &= \langle c, P^*\tilde{Z}(P\tilde{Z})^{-1}Pe - P^*e\rangle \\ &= \langle c, P\tilde{Z}(P\tilde{Z})^{-1}Pe - Pe\rangle = 0.\end{aligned}$$

Hence, $Zi - e \in N_a^\perp$, and consequently, i is a solution of $\mathscr{N}$ corresponding to e.

Suppose now that there exists some $\tilde{\imath} \in \mathscr{H}$ that satisfies K_1^* and K_2^* with the same $e \in \mathscr{H}$. Then, for any $c \in N_a$, $\langle c, Zi - e\rangle = 0$ and $\langle c, Z\tilde{\imath} - e\rangle = 0$, and consequently, $\langle c, Zi - Z\tilde{\imath}\rangle = 0$. Because $i, \tilde{\imath}$ are elements of N_a, we have

$$(1.4) \qquad 0 = \langle c, \tilde{Z}i - \tilde{Z}\tilde{\imath}\rangle = \langle Pc, \tilde{Z}i - \tilde{Z}\tilde{\imath}\rangle = \langle c, P\tilde{Z}i - P\tilde{Z}\tilde{\imath}\rangle.$$

However, since both $P\tilde{Z}i$ and $P\tilde{Z}\tilde{\imath}$ are in N_a, we can set $c = P\tilde{Z}i - P\tilde{Z}\tilde{\imath}$ and obtain from (1.4), $P\tilde{Z}i - P\tilde{Z}\tilde{\imath} = 0$. The existence of $(P\tilde{Z})^{-1}$ then implies that $i = \tilde{\imath}$. Hence, $\mathscr{N}$ is regular and (1.3) holds.

(b) Conversely, suppose that $\mathscr{N}$ is regular, i.e., there exists an operator $\tilde{A}:\mathscr{H} \to \mathscr{H}$ such that, for any $e \in \mathscr{H}$, $i = \tilde{A}e$ is the unique solution of $\mathscr{N}$ corresponding to e. We are going to show that the operator $P\tilde{Z}: N_a \to N_a$ is invertible. Thus, suppose first that there exist $x_1, x_2 \in N_a$ such that $P\tilde{Z}x_1 = P\tilde{Z}x_2 = y$. Then we have for any $c \in N_a$,

$$\langle c, Zx_1 - y\rangle = \langle Pc, Zx_1 - y\rangle = \langle c, P\tilde{Z}x_1 - Py\rangle = \langle c, P\tilde{Z}x_1 - y\rangle = 0;$$

similarly, $\langle c, Zx_2 - y\rangle = 0$. Consequently, x_1 and x_2 are solutions of $\mathscr{N}$ corresponding to y, so that necessarily $x_1 = x_2$; hence, $P\tilde{Z}$ is one-to-one.

Next, arbitrarily choose $y \in N_a$ and put $i = \tilde{A}y$. Then we have by K_2, K_1, $i \in N_a$ and $\langle c, Zi - y\rangle = 0$ for all $c \in N_a$. This equality yields

$$(1.5) \qquad 0 = \langle Pc, Zi - y\rangle = \langle c, P\tilde{Z}i - Py\rangle = \langle c, P\tilde{Z}i - y\rangle.$$

However, since $P\tilde{Z}i \in N_a$ and $y \in N_a$, we can set $c = P\tilde{Z}i - y$; then we get from (1.5), $P\tilde{Z}i - y = 0$, i.e., $P\tilde{Z}$ is onto. Hence, $P\tilde{Z}$ is invertible.

Finally, if $\mathscr{N}$ is linear, (1.3) shows that A is also linear. If $Z \in [\mathscr{H}, \mathscr{H}]$, then $P\tilde{Z}$ is bounded on N_a since $\|P\| = 1$; thus, by the open mapping theorem, $(P\tilde{Z})^{-1}$ is also bounded and (1.3) shows that $A \in [\mathscr{H}, \mathscr{H}]$. This completes the proof.

The following three theorems give various sufficient conditions for regularity of an abstract network.

THEOREM 1.2. *Let $\mathscr{N} = (Z, a)$ be an abstract network over $\mathscr{H}$, and let N_a, P have the meaning defined above. Assume that there exist a number $\mu \neq 0$ and a real λ with $0 < \lambda < 1$ such that*

$$(1.6) \qquad I - \mu PZ \in \mathrm{Lip}_\lambda [N_a, N_a],$$

where I is the identity operator on N_a. Then $\mathscr{N}$ is regular and the admittance operator A of $\mathscr{N}$ is in $\mathrm{Lip}_\beta [\mathscr{H}, \mathscr{H}]$, where $\beta = (1 - \lambda)^{-1}|\mu|$.

In view of Theorem 1.1, the proof of Theorem 1.2 is a straightforward application of the contraction mapping theorem to the operator $Q_y: N_a \to N_a$ defined by $Q_y x = (I - \mu PZ)x + \mu y$, and is omitted.

THEOREM 1.3. *Let $\mathscr{N} = (Z, a)$ be an abstract network over $\mathscr{H}$, and let N_a, P have the meaning defined above. Assume that there exist constants $c > 0$ and $\lambda > 0$ such that*

$$\text{(i)} \qquad \mathrm{Re}\,\langle Zx_1 - Zx_2, x_1 - x_2\rangle \geqq c\|x_1 - x_2\|^2$$

for all $x_1, x_2 \in N_a$,

$$\text{(ii)} \qquad PZ \in \mathrm{Lip}_\lambda\,[N_a, N_a].$$

Then $\mathscr{N}$ *is regular and the admittance operator* A *of* $\mathscr{N}$ *is in* $\mathrm{Lip}_\gamma\,[\mathscr{H}, \mathscr{H}]$ *with* $\gamma = (c\lambda)^{-1}(\lambda + \sqrt{\lambda^2 - c^2})$.

The proof of this theorem is again a routine application of properties of monotonic operators and is omitted. (See, for example, [7, p. 718].)

Remark 1. If $\mathscr{N} = (Z, a)$ is an abstract network and the operator Z is such that $\mathrm{Re}\,\langle Zx_1 - Zx_2, x_1 - x_2\rangle \geqq c\|x_1 - x_2\|^2$ for all $x_1, x_2 \in \mathscr{H}$ and some $c > 0$, and $Z \in \mathrm{Lip}_\lambda\,[\mathscr{H}, \mathscr{H}]$ for some $\lambda > 0$, then conditions (i), (ii) in Theorem 1.3 are trivially satisfied, and consequently, $\mathscr{N}$ is regular. In other words, a network $\mathscr{N}$ with such Z is always regular whatever the structural operator a is.

For the proof of the next theorem we will need the following proposition.

LEMMA 1.1. *Let* $\mathscr{H}_i$ *be a Hilbert space with inner product* $\langle\,\cdot\,,\ \cdot\,\rangle_i$, $i = 1, 2$, *and let* $A \in [\mathscr{H}_1, \mathscr{H}_2]$.

(a) *If*

$$\text{(1.7)} \qquad \inf_{\substack{\|x\|=1 \\ x \in \mathscr{H}_1}} \sup_{\substack{\|y\|=1 \\ y \in \mathscr{H}_2}} |\langle Ax, y\rangle_2| = c_1 > 0$$

and

$$\text{(1.8)} \qquad \inf_{\substack{\|y\|=1 \\ y \in \mathscr{H}_2}} \sup_{\substack{\|x\|=1 \\ x \in \mathscr{H}_1}} |\langle Ax, y\rangle_2| = c_2 > 0,$$

then A *is one-to-one from* $\mathscr{H}_1$ *onto* $\mathscr{H}_2$, $A^{-1} \in [\mathscr{H}_2, \mathscr{H}_1]$ *and*

$$\text{(1.9)} \qquad \|A^{-1}\| \leqq c_1^{-1}.$$

(b) *If* $\mathscr{H}_1 = \mathscr{H}_2 = \mathscr{H}$ *and there exists* $c > 0$ *such that*

$$\text{(1.10)} \qquad |\langle Ax, x\rangle| \geqq c\|x\|^2$$

for all $x \in \mathscr{H}$, *then* (1.7) *and* (1.8) *are satisfied with* $c_1 = c_2 = c$.

Since the proof is similar to that of the Lax–Milgram theorem (see [8, p. 92]), we will indicate only the main steps. It follows from (1.7) that

$$\text{(1.11)} \qquad |\langle Ax, x\rangle_2| \geqq c\|x\|_2^2$$

for all $x \in \mathscr{H}_1$. This shows that A is one-to-one. Moreover, (1.11) and the boundedness of A imply readily that $A\mathscr{H}_1$ is closed in $\mathscr{H}_2$. Next, suppose that $A\mathscr{H}_1 \neq \mathscr{H}_2$; then, by a standard theorem, there exists $y_0 \in \mathscr{H}_2$ with $\|y_0\| = 1$ such that $y_0 \perp A\mathscr{H}_1$. Thus, for all $x \in \mathscr{H}_1$ with $\|x\| = 1$, $\langle Ax, y_0\rangle_2 = 0$ which contradicts (1.8); hence, A is onto $\mathscr{H}_2$. Inequality (1.9) follows then from (1.11). The assertion (b) is obvious.

Remark 2. It is a matter of routine to show that conditions (1.7) and (1.8) are necessary and sufficient conditions for A and the adjoint A^* to have a bounded inverse.

THEOREM 1.4. *Let $\mathcal{N} = (Z, a)$ be an abstract network over $\mathcal{H}$, and let N_a and P have the meaning defined above. Assume that $Z = Z_1 + Z_2$, where $Z_1 \in [\mathcal{H}, \mathcal{H}]$. If*

(i) $$\inf_{\substack{\|x\|=1\\ x\in N_a}} \sup_{\substack{\|y\|=1\\ y\in N_a}} |\langle Z_1x, y\rangle| = c > 0,$$

$$\inf_{\substack{\|y\|=1\\ y\in N_a}} \sup_{\substack{\|x\|=1\\ x\in N_a}} |\langle Z_1x, y\rangle| > 0,$$

(ii) *there exists $\lambda > 0$ such that $PZ_2 \in \mathrm{Lip}_\lambda [N_a, N_a]$,*

(iii) $c^{-1}\lambda < 1$,

then $\mathcal{N}$ is regular and the admittance operator A of $\mathcal{N}$ is in $\mathrm{Lip}_\delta [\mathcal{H}, \mathcal{H}]$, where $\delta = (c - \lambda)^{-1}$.

Proof. Let $\tilde{Z}_1$ and $\tilde{Z}_2$ be the restriction of Z_1 and Z_2 to N_a, respectively. First, consider the operator $P\tilde{Z}_1 : N_a \to N_a$; note that N_a as a closed linear subspace of $\mathcal{H}$ is a Hilbert space in its own right. If $x, y \in N_a$, we have $\langle P\tilde{Z}_1x, y\rangle = \langle \tilde{Z}_1x, Py\rangle = \langle Z_1x, y\rangle$; hence, conditions (i) show by Lemma 1.1 that $(P\tilde{Z}_1)^{-1}$ exists, is bounded and $\|(P\tilde{Z}_1)^{-1}\| \leqq c^{-1}$. To conclude the proof, it suffices to verify that the operator $Q_y : N_a \to N_a$ defined by

$$Q_yx = -(P\tilde{Z}_1)^{-1}P\tilde{Z}_2x + (P\tilde{Z}_1)^{-1}y$$

is a contraction, and consequently, the equation $P\tilde{Z}x = y$ possesses a unique solution for any chosen $y \in N_a$.

From Theorem 1.4 we get immediately the following proposition.

COROLLARY 1.1. *Let $\mathcal{N} = (Z, a)$ be a linear abstract network over $\mathcal{H}$ and let $Z \in [\mathcal{H}, \mathcal{H}]$.*

(a) *If Z satisfies conditions* (i) *in Theorem* 1.4, *then $\mathcal{N}$ is regular.*

(b) *If there exists $c > 0$ such that*

$$|\langle Zx, x\rangle| \geqq c\|x\|^2 \tag{1.12}$$

for all $x \in N_a$, then $\mathcal{N}$ is regular.

In either case, the admittance operator A of $\mathcal{N}$ is in $[\mathcal{H}, \mathcal{H}]$, and $\|A\| \leqq c^{-1}$.

Remark 3. If the network $\mathcal{N} = (Z, a)$ is linear, $Z \in [\mathcal{H}, \mathcal{H}]$ and (1.12) holds for all $x \in \mathcal{H}$ with some $c > 0$, then $\mathcal{N}$ is clearly regular; thus, in this case, regularity of $\mathcal{N}$ is independent of a.

Let us now establish an equivalent formulation of the network solution; to this end, we will need the following lemma.

LEMMA 1.2. *Let $\mathcal{H}, \mathcal{H}'$ be Hilbert spaces, let $a \in [\mathcal{H}, \mathcal{H}']$ and let $a^*\mathcal{H}'$ be closed in $\mathcal{H}$ ($a^* \in [\mathcal{H}', \mathcal{H}]$ signifies the adjoint of a). Furthermore, let $y \in \mathcal{H}$; then there exists $x \in \mathcal{H}'$ such that*

$$a^*x = y \tag{1.13}$$

iff for every $z \in \mathcal{H}$ with $az = 0$ we have $\langle z, y\rangle = 0$.

Proof. If $N_a \subset \mathcal{H}$ is the null space of a, then our lemma claims that $a^*x = y$ has a solution $x \Leftrightarrow y \in N_a^\perp$. However, by a standard theorem (see [3, p. 133]), $N_a = (a^*\mathcal{H}')^\perp$; hence $N_a^\perp = (a^*\mathcal{H}')^{\perp\perp} = \overline{a^*\mathcal{H}'} = a^*\mathcal{H}'$, which proves the lemma.

An obvious application of Lemma 1.2 to conditions K_1, K_2 yields the following equivalent formulation of the network solution.

THEOREM 1.5. *Let $\mathcal{N} = (Z, a)$ be an abstract network over $\mathcal{H}$, and let $a^*\mathcal{H}'$ be a closed subspace of $\mathcal{H}$, where a^* is the adjoint of a. Then the following statements are true*:

(i) *If, for some $e \in \mathcal{H}$, the network $\mathcal{N}$ possesses a* (*not necessarily unique*) *solution $i \in \mathcal{H}$, then there exists a $v \in \mathcal{H}$ such that*

$$Zi + a^*v = e, \tag{1.14}$$

$$ai = 0. \tag{1.15}$$

(ii) *If, for a given $e \in \mathcal{H}$, there exist elements $i \in \mathcal{H}$ and $v \in \mathcal{H}$ such that* (1.14) *and* (1.15) *hold, then i is a solution of $\mathcal{N}$ corresponding to e.*

It is clear that if $N_{a^*} \neq \{0\}$ (N_{a^*} being the nullspace of a^*), then the element v in (1.14) is not determined uniquely, even in the case that i is unique. However, if we impose the requirement that v be in $N_{a^*}^{\perp}$, v will be determined uniquely.

Combining this fact with Theorem 1.5, we get the following result.

THEOREM 1.6. *Let $\mathcal{N} = (Z, a)$ be an abstract network over $\mathcal{H}$ and let $a^*\mathcal{H}'$ be closed in $\mathcal{H}$. Define the operator $K : N_a \times N_{a^*}^{\perp} \to \mathcal{H}$ by*

$$K(x, y) = Zx + a^*y. \tag{1.16}$$

Then $\mathcal{N}$ is regular iff K possesses an inverse K^{-1}.

Proof. Assume first that K^{-1} exists. If $e \in \mathcal{H}$, then there exist unique elements $i \in N_a$ and $v \in N_{a^*}^{\perp}$ such that $K(i, v) = e$, i.e., (1.14) and (1.15) hold. However, by Theorem 1.5, i is a solution of $\mathcal{N}$ corresponding to e; furthermore, i is unique, because if there existed another $i' \in N_a$, we had by Theorem 1.5, $Zi' + a^*v' = e$ for some $v' \in \mathcal{H}'$, which contradicts the uniqueness of (i, v). Hence, $\mathcal{N}$ is regular.

Conversely, assume that $\mathcal{N}$ is regular. Choose $e \in \mathcal{H}$, and let $i \in N_a$ be the unique solution of $\mathcal{N}$ corresponding to e. Then, by Theorem 1.5 there exists a unique $v \in N_{a^*}^{\perp}$ such that (1.14) holds, i.e., we have $K(i, v) = e$. Moreover, if there existed some different pair $(i', v') \in N_a \times N_{a^*}^{\perp}$ with $K(i', v') = e$, then, by Theorem 1.5, i' would be a solution of $\mathcal{N}$ corresponding to e; however, since necessarily $v' = v$, we would have $i' \neq i$, which contradicts the regularity of $\mathcal{N}$. Hence, K^{-1} exists and the proof is complete.

Remark 4. Comparing Theorems 1.1 and 1.6, we obtain the following equivalence:

$$\mathcal{N} \text{ is regular} \Leftrightarrow (P\tilde{Z})^{-1} \text{ exists} \Leftrightarrow K^{-1} \text{ exists.}$$

The following theorem provides a basis for proving an approximation theorem on infinite Hilbert networks.

THEOREM 1.7. *Let $\mathcal{N} = (Z, a)$ be an abstract network over $\mathcal{H}$, let N_a be the nullspace of a and P be the projection of $\mathcal{H}$ onto N_a. Let there exist constants $c, \lambda > 0$ such that*

$$\operatorname{Re}\langle Zx_1 - Zx_2, x_1 - x_2\rangle \geqq c\|x_1 - x_2\|^2 \tag{1.17}$$

for all $x_1, x_2 \in N_a$, and

$$PZ \in \operatorname{Lip}_\lambda[N_a, N_a]. \tag{1.18}$$

Furthermore, assume that the following conditions are satisfied:

(i) $\mathscr{H}_n, n = 1, 2, 3, \cdots$, *is a sequence of closed linear subspaces of* $\mathscr{H}$ *such that* $\mathscr{H}_1 \subset \mathscr{H}_2 \subset \mathscr{H}_3 \subset \cdots \subset \mathscr{H}$ *and* P_n *is the orthogonal projection from* $\mathscr{H}$ *onto* $\mathscr{H}_n$;

(ii) *For every* $n, a_n \in [\mathscr{H}, \mathscr{H}']$ *is such that*

$$N_{a_n} = \{x : x \in \mathscr{H}, a_n x = 0\} \subset \mathscr{H}_n$$

and $N_{a_n} \subset N_{a_{n+1}}$;

(iii) $N = \bigcup_{n=1}^{\infty} N_{a_n}$ *is dense in* N_a, *i.e.*, $\bar{N} = N_a$.

Let $e \in \mathscr{H}$, *let* $i \in \mathscr{H}$ *be the* (*unique*) *solution of* $\mathscr{N}$ *corresponding to e, and let* $i_n \in \mathscr{H}$ *be the* (*unique*) *solution of* $\mathscr{N}_n = (Z, a_n)$ *corresponding to* $P_n e$. *Then* $i_n \to i$ *in* $\mathscr{H}$.

If, in addition, $\mathscr{N}$ *is linear, then conditions* (1.17), (1.18) *may be replaced by requirements*

$$|\langle Zx, x\rangle| \geqq c\|x\|^2, \qquad c > 0, \tag{1.19}$$

for all $x \in N_a$, *and*

$$Z \in [\mathscr{H}, \mathscr{H}]. \tag{1.20}$$

Proof. First, from Theorem 1.3 it follows that $\mathscr{N}$ is regular. Moreover, by (iii), $N_{a_n} \subset N_a$ for every n, and consequently, (1.17) holds for all $x_1, x_2 \in N_{a_n}$. Let Q_n be the orthogonal projection from $\mathscr{H}$ onto N_{a_n}; due to the inclusion $N_{a_n} \subset N_a$ we have $Q_n = Q_n P$. Consequently, (1.18) yields for all $x_1, x_2 \in N_{a_n}$, $\|Q_n(Zx_1 - Zx_2)\| = \|Q_n P(Zx_1 - Zx_2)\| \leqq \|P(Zx_1 - Zx_2)\| \leqq \lambda\|x_1 - x_2\|$, i.e., $Q_n Z \in \mathrm{Lip}_\lambda [N_{a_n}, N_{a_n}]$; hence, $\mathscr{N}_n$ is regular by Theorem 1.3.

Next, choose $e \in \mathscr{H}$ and construct the solution i of $\mathscr{N}$ corresponding to e. If $\varepsilon > 0$, then by (iii) there exists $j \in N$ such that $\|j - i\| < (\varepsilon/2) \min [1, c\lambda^{-1}]$. Thus, due to (ii), there exists an integer $M > 0$ so that $j \in N_{a_n}$ for all $n \geqq M$. Choose some $n \geqq M$ and let i_n be the solution of $\mathscr{N}_n$ corresponding to $P_n e$; thus, we have $i_n \in N_{a_n}$ and

$$\langle c, Zi_n\rangle = \langle c, P_n e\rangle \tag{1.21}$$

for every $c \in N_{a_n}$.

On the other hand, we have $i \in N_a$ and

$$\langle c', Zi\rangle = \langle c', e\rangle \tag{1.22}$$

for all $c' \in N_a$, and consequently, also for all $c \in N_{a_n}$. However, if $c \in N_{a_n} \subset \mathscr{H}_n$, we have $\langle c, P_n e\rangle = \langle P_n c, e\rangle = \langle c, e\rangle$. Hence, (1.21), (1.22) yield $\langle c, Zi_n - Zi\rangle = 0$, and consequently,

$$\langle c, Zi_n - Zj\rangle = \langle c, Zi - Zj\rangle \tag{1.23}$$

for any $c \in N_{a_n}$. Since $i_n, j \in N_{a_n}$, we can set $c = i_n - j$ and get from (1.23),

$$\langle i_n - j, Zi_n - Zj\rangle = \langle i_n - j, Zi - Zj\rangle. \tag{1.24}$$

Using (1.17), (1.18) and the fact that P is an identity on N_{a_n}, we obtain from (1.24),

$$\begin{aligned} c\|i_n - j\|^2 &\leqq \mathrm{Re}\,\langle i_n - j, Zi_n - Zj\rangle = \mathrm{Re}\,\langle i_n - j, Zi - Zj\rangle \\ &\leqq |\langle P(i_n - j), Zi - Zj\rangle| = |\langle i_n - j, P(Zi - Zj)\rangle| \\ &\leqq \|i_n - j\| \cdot \|P(Zi - Zj)\| \leqq \|i_n - j\| \cdot \lambda\|i - j\|. \end{aligned} \tag{1.25}$$

Consequently, $\|i_n - j\| \leqq c^{-1}\lambda\|i - j\| < \varepsilon/2$; thus, we have, finally, $\|i_n - i\| \leqq \|i_n - j\| + \|j - i\| < \varepsilon$, which is what we wanted to show.

As for the last proposition concerning a linear network, observe first that $\mathcal{N}$ and each $\mathcal{N}_n$ is regular by Corollary 1.1 (b). On the other hand, except for the proof of regularity of $\mathcal{N}$ and $\mathcal{N}_n$, (1.17) was used only for establishing the inequality (1.25). However, it is easy to see that (1.25) follows from (1.19), too; hence, the proof.

Concluding this section, let us mention the abstract analogues of the classical "reciprocity law" and "energy conservation principle".

THEOREM 1.8. *Let $\mathcal{N} = (Z, a)$ be an abstract network over $\mathcal{H}$.*

(a) *If $i_k \in \mathcal{H}$ is a (not necessarily unique) solution of $\mathcal{N}$ corresponding to $e_k \in \mathcal{H}, k = 1, 2$, then*

$$\langle i_2, e_1\rangle = \langle i_1, e_2\rangle \qquad \textit{provided } Z \textit{ is symmetric,}$$

i.e.,

$$\langle Zx, y\rangle = \langle Zy, x\rangle \qquad \textit{for all } x, y \in \mathcal{H},$$

and

$$\langle i_2, e_1\rangle = \overline{\langle i_1, e_2\rangle} \qquad \textit{provided } Z \textit{ is self-adjoint.}$$

(b) *If $i \in \mathcal{H}$ is a solution of $\mathcal{N}$ corresponding to $e \in \mathcal{H}$, then*

$$\langle i, e\rangle = \langle i, Zi\rangle.$$

The proof follows immediately from conditions K_1, K_2 and is omitted.

Let us note the fact that (a) applies practically only to the linear case, since self-adjointness of Z clearly implies linearity of Z, and symmetry of Z implies antilinearity of Z.

2. Hilbert networks. In this section we will apply the results obtained in the preceding section to nonlinear Hilbert networks, i.e., networks described by an oriented graph and a (not necessarily linear) impedance operator from some Hilbert space $\mathcal{H}^c$ into itself. Since the graph and the vector of voltages or currents can be either finite or infinite, it will be useful to introduce some convenient notation.

Let c_1, c_2 be cardinal numbers such that $c_1, c_2 \leqq \aleph_0$; a matrix $M = [m_{ik}]$ (the element m_{ik} stands in the ith row and kth column) will be said to be of type $c_1 \times c_2$ if the set of rows (columns) has cardinal $c_1(c_2)$. A $c \times 1$ matrix ξ will be called a c-vector and denoted by $[\xi_k]$ whenever it will be necessary to display its components. The sum of two matrices of the same type is defined elementwise.

If M and N are conformable matrices, the product $M \cdot N$ is defined in the usual way by assuming that each element of $M \cdot N$ is defined. (We will use the dot " $\cdot$ " for denoting the matrix multiplication in order to distinguish it from the product of operators defined by such matrices.)

Furthermore, if $c = \aleph_0$, l^c will stand for the Hilbert space l_2; if $c = n < \aleph_0$, l^c will denote the n-dimensional Euclidean space. For convenience, we will interpret elements in l^c as c-vectors; also, if $\xi, \eta \in l^c$, then $\langle \xi, \eta\rangle_c = \xi^T\bar{\eta}$ and $\|\xi\|_c^2 = \xi^T\bar{\xi}$.

In the sequel, $\mathcal{H}$ will denote a fixed separable Hilbert space. If $c \leqq \aleph_0$, we let

$$\mathcal{H}^c = \{x : x = [x_k] - c\text{-vector}, x_k \in \mathcal{H}, \sum_k \|x_k\|^2 < \infty\} \tag{2.1}$$

and

$$\langle x, y\rangle_c = \sum_k \langle x_k, y_k\rangle \tag{2.2}$$

for every pair $x, y \in \mathcal{H}^c$. Clearly, $\mathcal{H}^c$ is again a separable Hilbert space.

Next, let $\mathscr{B}$ (set of branches) and $\mathscr{V}$ (set of nodes) be nonempty sets having cardinals c_2 and c_1, respectively, $c_1, c_2 \leqq \aleph_0$. Thus, we will assume that all elements in $\mathscr{B}$ are ordered into a fixed (finite or infinite) sequence (b_i), and elements in $\mathscr{V}$ into a fixed sequence (v_k).

A mapping $G:\mathscr{B} \to \mathscr{V} \times \mathscr{V}$ will be called an oriented graph if

(i) for any $v_k \in \mathscr{V}$ the sets

$$\mathscr{B}_k^+ = \{b_j : b_j \in \mathscr{B}, Gb_j = (v_\alpha, v_k)\}$$

and

$$\mathscr{B}_k^- = \{b_j : b_j \in \mathscr{B}, Gb_j = (v_k, v_\alpha)\}$$

are finite and at least one of them is nonempty,

(ii) $$\{b_i : b_i \in \mathscr{B}, Gb_i = (v_\alpha, v_\alpha) \text{ for any } v_\alpha \in \mathscr{V}\} = \varnothing.$$

Thus, our graph does not contain "isolated" nodes and branches beginning and ending at the same node; also, each node $v_k \in \mathscr{V}$ is incident with only a finite number of branches. Observe that we have $c_1 \leqq 2c_2$.

Let G be an oriented graph; define the (incidence) matrix $d = [d_{ik}]$ of type $c_2 \times c_1$ by

$$d_{ik} = \begin{cases} 1 & \text{if } Gb_i = (v_\alpha, v_k), \\ -1 & \text{if } Gb_i = (v_k, v_\alpha), \\ 0 & \text{if } Gb_i = (v_\alpha, v_\beta) \text{ and } \alpha \neq k, \beta \neq k. \end{cases} \tag{2.3}$$

Next, let K be an arbitrary diagonal matrix of type $c_1 \times c_1$, $K = \operatorname{diag}(k_1, k_2, \cdots)$ such that the number $k_i \neq 0$ for every i, and such that the matrix

$$a = K \cdot d^T = [a_{ik}] \tag{2.4}$$

has the property that $\sum_{ik} |a_{ik}|^2 < \infty$. Since each row of d^T contains only finitely many nonzero elements due to the assumption (i), such a matrix K always exists.

Define now the operator $\hat{a}$: $\mathscr{H}^{c_2} \to \mathscr{H}^{c_1}$ by

$$\hat{a}x = a \cdot [x_k], \tag{2.5}$$

where $x = [x_k]$ is any c_2-vector in $\mathscr{H}^{c_2}$. Due to our construction of the matrix a it follows that $\hat{a} \in [\mathscr{H}^{c_2}, \mathscr{H}^{c_1}]$.

Furthermore, let

$$N_{\hat{a}} = \{x : x \in \mathscr{H}^{c_2}, \hat{a}x = 0\}; \tag{2.6}$$

it is easy to see that $N_{\hat{a}}$ is a closed subspace of $\mathscr{H}^{c_2}$ and does not depend on the choice of the matrix K.

Let G be an oriented graph, and let $c_2 \leqq \aleph_0$ be the cardinal of its set of branches $\mathscr{B}$; moreover, let $\hat{Z}$ be a (not necessarily linear) operator from $\mathscr{H}^{c_2}$ into itself. Then the ordered pair $\hat{\mathscr{N}} = (\hat{Z}, G)$ will be called a Hilbert network, and $\hat{Z}$ the impedance operator of $\hat{\mathscr{N}}$.

A network $\hat{\mathscr{N}}$ will be called linear if $\hat{Z}$ is linear, and infinite or finite if $c_2 = \aleph_0$ or $c_2 < \aleph_0$, respectively.

If a vector $x = [x_k] \in \mathscr{H}^{c_2}$ has some relevance to the network $\hat{\mathscr{N}}$, it will be understood that the kth component x_k of x is associated with the branch $b_k \in \mathscr{B}$, where the branches in $\mathscr{B}$ have been enumerated in some fixed manner as mentioned above. Thus, using physical terminology, if $e = [e_k] \in \mathscr{H}^{c_2}$ has the meaning of a vector of voltages in $\hat{\mathscr{N}}$, then $e_k \in \mathscr{H}$ is interpreted as the voltage in the branch b_k; similarly, if $i = [i_k] \in \mathscr{H}^{c_2}$ is the vector of currents in $\hat{\mathscr{N}}$, $i_k \in \mathscr{H}$ is the current in b_k. If we denote $u = [u_k] = \hat{Z}i \in \mathscr{H}^{c_2}$, then the kth component $u_k \in \mathscr{H}$ of u is the voltage drop in b_k caused by currents $i_i, i_2, i_3, \cdots$.

A special case of a Hilbert network is obtained, if the impedance operator $\hat{Z}$ is described by a matrix. Indeed, assume that to each ordered pair $(b_i, b_k) \in \mathscr{B} \times \mathscr{B}$ there is assigned an operator $Z_{ik}: \mathscr{H} \to \mathscr{H}$ such that the mapping $\hat{Z}$ defined on $\mathscr{H}^{c_2}$ by

$$\hat{Z}x = [Z_{ik}] \cdot [x_k] \tag{2.7}$$

maps $\mathscr{H}^{c_2}$ into itself. In this case, the voltage drop u_l in branch b_l is additive in drops caused by currents $i_1, i_2, i_3, \cdots$, i.e., $u_l = \sum_k Z_{lk} i_k$. Of course, if each operator Z_{ik} is linear, then the network $\hat{\mathscr{N}}$ is linear, and vice versa.

DEFINITION 1. Let $\hat{\mathscr{N}} = (\hat{Z}, G)$ be a Hilbert network, and let $e \in \mathscr{H}^{c_2}$; an element $i \in \mathscr{H}^{c_2}$ will be called a *solution of* $\hat{\mathscr{N}}$ *corresponding to* e if i is a solution of the associated abstract network $\mathscr{N} = (\hat{Z}, \hat{a})$ over $\mathscr{H}^{c_2}$ corresponding to e, i.e., if

$$\mathrm{K}_1^*: \qquad \hat{Z}i - e \in N_{\hat{a}}^{\perp},$$

$$\mathrm{K}_2^*: \qquad i \in N_{\hat{a}}.$$

A network $\hat{\mathscr{N}}$ will be called *regular*, if $\mathscr{N}$ is regular, i.e., if for each $e \in \mathscr{H}^{c_2}$ there exists a unique $i \in \mathscr{H}^{c_2}$ such that K_1^* and K_2^* are satisfied. In this case, the admittance operator of $\mathscr{N}$ is called the admittance operator of $\hat{\mathscr{N}}$.

Obviously, we have defined the solution of a Hilbert network $\hat{\mathscr{N}}$ as a solution of the associated abstract network. The question is, however, what justifies such a definition. Having the classical Kirchhoff's laws in mind, it would be perhaps more natural to define the solution concept for $\hat{\mathscr{N}}$ by the following definition.

DEFINITION 2. Let $\hat{\mathscr{N}} = (\hat{Z}, G)$ be a Hilbert network, and let $e \in \mathscr{H}^{c_2}$; an element $i \in \mathscr{H}^{c_2}$ will be called a *classical solution of* $\hat{\mathscr{N}}$, *corresponding to* e if

$$\mathrm{K}_1^+: \qquad \bar{\gamma}^T \cdot (\hat{Z}i - e) = 0$$

for every $\gamma \in l^{c_2}$ satisfying the equation $a \cdot \gamma = 0$, and

$$\mathrm{K}_2^+: \qquad a \cdot i = 0.$$

Observe that this definition is meaningful. Indeed, if $e, i \in \mathscr{H}^{c_2}$, then $\hat{Z}i - e = u$ is also in $\mathscr{H}^{c_2}$, and consequently, $\bar{\gamma}^T \cdot u$ is defined and belongs to $\mathscr{H}$ for any $\gamma \in l^{c_2}$; also, it is easy to verify that the classical solution does not depend on the choice of the matrix K.

Fortunately, it turns out that both solution concepts coincide; to prove this fact, we will have to carry out some auxiliary considerations.

First, let us define the mapping $\tilde{a}: l^{c_2} \to l^{c_1}$ by

$$\tilde{a}\xi = a \cdot \xi, \tag{2.8}$$

and let

$$N_{\tilde{a}} = \{\xi : \xi \in l^{c_2}, \tilde{a}\xi = 0\}. \tag{2.9}$$

Clearly, $N_{\tilde{a}}$ does not depend on K. Moreover, since $\tilde{a} \in [l^{c_2}, l^{c_1}]$, $N_{\tilde{a}}$ is a closed linear subspace of l^{c_2}, and consequently, $N_{\tilde{a}}$ is a separable Hilbert space in its own right.

It is obvious that if $N_{\tilde{a}} = \{0\}$, then $i = 0 \in \mathscr{H}^{c_2}$ is the only classical solution of $\mathscr{N}$ for any $e \in \mathscr{H}^{c_2}$; thus, we will assume in the sequel that $N_{\tilde{a}} \neq \{0\}$.

Choose a fixed orthonormal basis $\{\xi^i\}$ in $N_{\tilde{a}}$, and let $c_0 \leqq \aleph_0$ be the cardinal of $\{\xi^i\}$. Moreover, let X be the $c_2 \times c_0$ matrix, whose columns are all the c_2-vectors ξ^i enumerated in some fixed fashion. It is clear that

$$a \cdot X = 0. \tag{2.10}$$

Define now the mapping $\tilde{X} : l^{c_0} \to l^{c_2}$ by

$$\tilde{X}\eta = X \cdot \eta. \tag{2.11}$$

Then we have the following lemma.

LEMMA 2.1. (a) *$\tilde{X} \in [l^{c_0}, l^{c_2}]$ and $N_{\tilde{a}} = \tilde{X}l^{c_0}$.*

(b) *$\|\tilde{X}\xi\| = \|\xi\|$ for each $\xi \in l^{c_0}$.*

(c) *If $\tilde{X}^* \in [l^{c_2}, l^{c_0}]$ is the adjoint of $\tilde{X}$, then*

$$\tilde{X}^*\tilde{X} = I, \tag{2.12}$$

with I being the identity operator on l^{c_0}, and $\tilde{X}^*\eta = \overline{X}^T \cdot \eta$ for every $\eta \in l^{c_2}$.

Proof. Proposition (a) is an obvious consequence of the definition of X, and (b) is nothing else than Parseval's equality for the basis $\{\xi^i\}$. Proposition (c) follows immediately from the definition of an adjoint operator.

The next lemma is crucial for our further considerations.

LEMMA 2.2. *Let the operator $\hat{X}$ be defined on $\mathscr{H}^{c_0}$ by*

$$\hat{X}z = X \cdot [z_k]; \tag{2.13}$$

then we have:

(i) *$\hat{X} \in [\mathscr{H}^{c_0}, \mathscr{H}^{c_2}]$;*

(ii) *$N_{\hat{a}} = \hat{X}\mathscr{H}^{c_0}$ and $\hat{X}$ is one-to-one;*

(iii) *$\hat{a}\hat{X} = 0$ on $\mathscr{H}^{c_0}$;*

(iv) *$\|\hat{X}z\|_{c_2} = \|z\|_{c_0}$ for every $z \in \mathscr{H}^{c_0}$;*

(v) *if $\hat{X}^* \in [\mathscr{H}^{c_2}, \mathscr{H}^{c_0}]$ is the adjoint of $\hat{X}$, then*

$$\hat{X}^*\hat{X} = I, \tag{2.14}$$

where I is the identity operator on $\mathscr{H}^{c_0}$, and

$$\hat{X}^*v = \overline{X}^T \cdot v \tag{2.15}$$

for all $v \in \mathscr{H}^{c_2}$;

(vi) *$N_{\hat{X}^*} = N_{\hat{a}}^{\perp}$, where $N_{\hat{X}^*} = \{x : x \in \mathscr{H}^{c_2}, \hat{X}^*x = 0\}$;*

(vii) *if P is the orthogonal projection from $\mathscr{H}^{c_2}$ onto $N_{\hat{a}}$, then $P = \hat{X}\hat{X}^*$.*

Proof. First, let us show that the definition (2.13) is meaningful. To this end, choose some fixed orthonormal basis $\{f_i\}$ in $\mathscr{H}$; since $\mathscr{H}$ is separable, the cardinal c of $\{f_i\}$ does not exceed $\aleph_0$. Let the c-vector f be defined by $f = [f_i]$.

Now, arbitrarily choose $z = [z_k] \in \mathscr{H}^{c_0}$; by definition of $\mathscr{H}^{c_0}$,

$$\sum_k \|z_k\|^2 = \|z\|_{c_0}^2 < \infty. \tag{2.16}$$

Since $\{f_i\}$ is a basis in $\mathscr{H}$, for every k there exist numbers $r_{kj}, j = 1, 2, 3, \cdots$, such that

$$z_k = \sum_k r_{kj} f_j \quad \text{and} \quad \|z_k\|^2 = \sum_j |r_{kj}|^2. \tag{2.17}$$

Define the $c_0 \times c$ matrix R by $R = [r_{kj}]$; thus, $z = R \cdot f$. Moreover, from (2.16) it follows that

$$\sum_{kj} |r_{kj}|^2 = \|z\|_{c_0}^2 < \infty. \tag{2.18}$$

Next, consider the $c_2 \times c$ matrix $Y = X \cdot R$; if y^m and ρ^m are the mth column of Y and R, respectively, we have $y^m = X \cdot \rho^m$. However, each column ρ^m is in l^{c_0} due to (2.17); since the columns of X constitute an orthonormal set in l^{c_2}, y^m will be in l^{c_2} and, by Parseval's equality, $\|y^m\| = \|\rho^m\|$. Thus, denoting $y^m = [y_k^m]$, we have for every m, $\sum_k |y_k^m|^2 = \|y^m\|^2 = \sum_k |r_{km}|^2$. Consequently, by (2.18),

$$\sum_{mk} |y_k^m|^2 = \sum_{mk} |r_{km}|^2 = \|z\|_{c_0}^2. \tag{2.19}$$

Hence, for each i, the ith row $y_i = [y_i^1, y_i^2, y_i^3, \cdots]$ of Y is in l^c, and consequently, the product $x = [x_i] = Y \cdot f$ is defined, i.e., $x_i \in \mathscr{H}$ and $\|x_i\|^2 = \sum_l |y_i^l|^2$ by Parseval's equality. Thus, by (2.19),

$$\sum_i \|x_i\|^2 = \|x\|_{c_2}^2 = \sum_{il} |y_i^l|^2 = \|z\|_{c_0}^2, \tag{2.20}$$

and we have $X \cdot z = X \cdot R \cdot f = Y \cdot f = x$; hence, (2.20) proves (i) and (iv).

Next, if $z \in \mathscr{H}^{c_0}$, then $\hat{a}\hat{X}z = a \cdot X \cdot z = 0 \cdot z = 0$ and (iii) holds.

Let us now prove (ii). First of all, (iv) shows that $\hat{X}$ is one-to-one. To prove that the range of $\hat{X}$ is $N_{\hat{a}}$, choose some $x = [x_k] \in N_{\hat{a}}$; then $\hat{a}x = a \cdot [x_k] = 0$. Since $x \in \mathscr{H}^{c_2}$, it follows in the same way as above that there exists a $c_2 \times c$ matrix $S = [s_{ik}]$ such that $x = S \cdot f$ and

$$\sum_{ik} |s_{ik}|^2 = \|x\|_{c_2}^2 < \infty. \tag{2.21}$$

Thus, we have $a \cdot S \cdot f = 0$, which yields $a \cdot S = 0$. Since each column s^k of S is in l^{c_2} due to (2.21) and $a \cdot s^k = 0$, then by Lemma 2.1 (a), (b) there exists a c_0-vector $\eta_k \in l^{c_0}$ such that $s^k = X \cdot \eta_k$, and $\|s^k\| = \|\eta_k\|$. Hence,

$$\sum_i |s_{ik}|^2 = \|\eta_k\|^2 \tag{2.22}$$

for every k. Denote $\eta_k^T = [\eta_{1k}, \eta_{2k}, \eta_{3k}, \cdots]$ and define the $c_0 \times c$ matrix E by $E = [\eta_{ik}]$, (i.e., the η_k's are columns of E). Then it follows from (2.22) and (2.21) that

$$\sum_{ki} |\eta_{ik}|^2 = \sum_{ki} |s_{ik}|^2 < \infty. \tag{2.23}$$

Moreover, $S = X \cdot E$, i.e., $x = X \cdot E \cdot f$, and $E \cdot f \in \mathscr{H}^{c_0}$ due to Parseval's equality and (2.23). Hence, $x \in \hat{X}\mathscr{H}^{c_0}$, and consequently, $N_{\hat{a}} \subset \hat{X}\mathscr{H}^{c_0}$.

Conversely, if $x \in \hat{X}\mathscr{H}^{c_0}$, then $x = \hat{X}w$ for some $w \in \mathscr{H}^{c_0}$, and thus, by (iii), $\hat{a}x = \hat{a}\hat{X}w = 0$, i.e., $x \in N_{\hat{a}}$; hence $\hat{X}\mathscr{H}^{c_0} \subset N_{\hat{a}}$, and (ii) is proved.

To prove (v), denote $X = [\xi_{ik}]$; choosing $x = [x_i] \in \mathscr{H}^{c_0}$ and $u = [u_i] \in \mathscr{H}^{c_2}$, we have, by continuity of the inner product and absolute convergence of the series involved,

$$\begin{aligned}\langle \hat{X}x, u\rangle_{c_2} &= \langle [\sum_{\mu} \xi_{i\mu}x_{\mu}], [u_i]\rangle_{c_2}\\ &= \sum_{i} \langle \sum_{\mu} \xi_{i\mu}x_{\mu}, u_i\rangle = \sum_{i}\sum_{\mu} \xi_{i\mu}\langle x_{\mu}, u_i\rangle\\ &= \sum_{\mu}\sum_{i} \langle x_{\mu}, \bar{\xi}_{i\mu}u_i\rangle = \sum_{\mu} \langle x_{\mu}, \sum_{i} \bar{\xi}_{i\mu}u_i\rangle\\ &= \langle [x_{\mu}], [\sum_{i} \bar{\xi}_{i\mu}u_i]\rangle_{c_0} = \langle x, \bar{X}^T \cdot u\rangle_{c_0}.\end{aligned}$$

This with the identity $\langle \hat{X}x, u\rangle_{c_2} = \langle x, \hat{X}^*u\rangle_{c_0}$ proves (2.15). Choosing now $x \in \mathscr{H}^{c_0}$, we have by (2.15) and the equality $\bar{X}^T \cdot X = I$, $\hat{X}^*\hat{X}x = \bar{X}^T \cdot X \cdot x = x$, which concludes the proof of (v).

Next, by a standard theorem, (see [6, p. 133]), $N_{\hat{X}^*}^{\perp} = \overline{\hat{X}\mathscr{H}^{c_0}}$. However, (ii) yields, due to closedness of $N_{\hat{a}}$, $\overline{\hat{X}\mathscr{H}^{c_0}} = \bar{N}_{\hat{a}} = N_{\hat{a}}$ which proves (vi).

Finally, let us prove (vii). To this end, observe first that $\hat{X}^*$ maps $\mathscr{H}^{c_2}$ onto $\mathscr{H}^{c_0}$. This follows readily from (2.14), which means that $\hat{X}$ is a right inverse of $\hat{X}^*$.

Next, let $Q: \mathscr{H}^{c_2} \to \mathscr{H}^{c_2}$ be defined by $Q = \hat{X}\hat{X}^*$; then we have by (2.14), $Q^2 = \hat{X}\hat{X}^*\hat{X}\hat{X}^* = \hat{X}\hat{X}^* = Q$. Also, $Q^* = (\hat{X}\hat{X}^*)^* = \hat{X}\hat{X}^* = Q$. Hence, Q is an orthogonal projection. Now, we are going to show that $Q\mathscr{H}^{c_2} = N_{\hat{a}}$. Indeed, if $x \in Q\mathscr{H}^{c_2}$, then $x = \hat{X}\hat{X}^*y$ for some $y \in \mathscr{H}^{c_2}$. By (iii), however, $\hat{a}x = \hat{a}\hat{X}(\hat{X}^*y) = 0$, i.e., $x \in N_{\hat{a}}$; hence $Q\mathscr{H}^{c_2} \subset N_{\hat{a}}$. Conversely, let $x \in N_{\hat{a}}$; then $x = \hat{X}y$ for some $y \in \mathscr{H}^{c_0}$ according to (ii). However, since $\hat{X}^*$ is onto $\mathscr{H}^{c_0}$, there exists some $z \in \mathscr{H}^{c_2}$ such that $y = \hat{X}^*z$; consequently, $x = \hat{X}\hat{X}^*z = Qx$, i.e., $N_{\hat{a}} \subset Q\mathscr{H}^{c_2}$. Hence, Q is the orthogonal projection of $\mathscr{H}^{c_2}$ onto $N_{\hat{a}}$, and the proof is complete.

Note that Lemma 2.2 can also be proved by using results on matrix representation of linear operators defined on a separable Hilbert space. (See [10, p. 48].)

Returning to the definition of a classical solution, we can state the following proposition.

LEMMA 2.3. *Let $\hat{\mathscr{N}}$ be a Hilbert network and let $e \in \mathscr{H}^{c_2}$; then $i \in \mathscr{H}^{c_2}$ is a classical solution of $\mathscr{N}$ corresponding to e iff*

$$(2.24) \qquad \hat{Z}i - e \in N_{\hat{X}^*}, \qquad i \in N_{\hat{a}}.$$

Proof. If $i \in \mathscr{H}^{c_2}$ is a classical solution, then, by K_1^+, $\bar{\gamma}^T \cdot u = 0$ for every $\gamma \in N_{\hat{a}}$, where $u = \hat{Z}i - e$. Since each column ξ^i of X is in $N_{\hat{a}}$, we have $\bar{X}^T \cdot u = 0$. Consequently, by Lemma 2.2 (v), $\hat{X}^*u = 0$, i.e., $u \in N_{\hat{X}^*}$ and (2.24) holds.

Conversely, let (2.24) hold; then by Lemma 2.2 (v), $\bar{X}^T \cdot u = 0$. Now, if $\gamma \in N_{\hat{a}}$, i.e., $\gamma \in l^{c_2}$ and $a \cdot \gamma = 0$, then by Lemma 2.1 (a), there exists $\eta \in l^{c_0}$ such that

$\gamma = X\cdot\eta$. Thus, we have $0 = \bar{\eta}^T\cdot(\bar{X}^T\cdot u) = \overline{(X\cdot\eta)}^T\cdot u = \bar{\gamma}^T\cdot u$; hence, K_1^+ holds and our lemma is proved.

Using Lemma 2.2 (vi) and combining Lemma 2.3 with Definition 1, we get immediately the following important result.

THEOREM 2.1. *For any Hilbert network $\hat{\mathcal{N}}$, the solution and classical solution coincide.*

A comment on the significance of this theorem is in order. The relations K_1^+ and K_2^+ defining the classical solution are clearly straightforward extensions of classical Kirchhoff's laws, i.e., in the case of a finite network, K_1^+ and K_2^+ coincide with the loop and node law, respectively. Thus the classical solution appears to be the most natural concept of a solution. Since by Theorem 2.1 the classical solution of a Hilbert network $\hat{\mathcal{N}}$ coincides with the solution of the associated abstract network, the transparent theory developed in the first section applies to Hilbert networks.

Now we are ready to state theorems on regularity of Hilbert networks.

THEOREM 2.2. *Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network and let $\hat{X}$ have the meaning defined above. Then $\hat{\mathcal{N}}$ is regular iff the operator $\hat{X}^*\hat{Z}\hat{X}:\mathcal{H}^{c_0}\to\mathcal{H}^{c_0}$ possesses an inverse. In this case, the admittance operator $A:\mathcal{H}^{c_2}\to\mathcal{H}^{c_2}$ of $\hat{\mathcal{N}}$ is given by*

$$A = \hat{X}(\hat{X}^*\hat{Z}\hat{X})^{-1}\hat{X}^*, \tag{2.25}$$

i.e., $i = Ae$ is the unique solution of $\hat{\mathcal{N}}$ corresponding to $e\in\mathcal{H}^{c_2}$. If, in addition, $\hat{\mathcal{N}}$ is linear, then A is linear, and $A\in[\mathcal{H}^{c_2},\mathcal{H}^{c_2}]$ provided $\hat{Z}\in[\mathcal{H}^{c_2},\mathcal{H}^{c_2}]$.

Proof. Recalling Theorem 1.1, $\hat{\mathcal{N}}$ is regular iff the operator $P\tilde{Z}:N_{\hat{a}}\to N_{\hat{a}}$ possesses an inverse, where $\hat{Z}$ is the restriction of $\hat{Z}$ to $N_{\hat{a}}$ and P is the projection of $\mathcal{H}^{c_2}$ onto $N_{\hat{a}}$.

Assume first that $\hat{X}^*\hat{Z}\hat{X}$ has an inverse. In agreement with Lemma 2.2 (ii), let $Y\in[N_{\hat{a}},\mathcal{H}^{c_0}]$ be the inverse of $\hat{X}\in[\mathcal{H}^{c_0},N_{\hat{a}}]$. Then $\hat{X}^*\tilde{Z} = \hat{X}^*\hat{Z}\hat{X}Y$ is one-to-one from $N_{\hat{a}}$ onto $\mathcal{H}^{c_0}$, and consequently, $\hat{X}\hat{X}^*\tilde{Z}$ is one-to-one from $N_{\hat{a}}$ onto $N_{\hat{a}}$. By Lemma 2.2 (vii), however, $\hat{X}\hat{X}^*\tilde{Z} = P\tilde{Z}$; hence, $(P\tilde{Z})^{-1}$ exists and $\hat{\mathcal{N}}$ is regular.

Conversely, if $\hat{\mathcal{N}}$ is regular, then reversing the above argument we conclude that $\hat{X}^*\hat{Z}\hat{X}$ possesses an inverse.

To show that the admittance operator A of $\hat{\mathcal{N}}$ is given by (2.25), choose some $e\in\mathcal{H}^{c_2}$ and put $i = \hat{X}(\hat{X}^*\hat{Z}\hat{X})^{-1}\hat{X}^*e$. Then from Lemma 2.2 (iii) it follows immediately that $i\in N_{\hat{a}}$. Moreover, $\hat{X}^*(\hat{Z}i - e) = \hat{X}^*\hat{Z}\hat{X}(\hat{X}^*\hat{Z}\hat{X})^{-1}\hat{X}^*e - \hat{X}^*e = 0$, i.e., $\hat{Z}i - e\in N_{\hat{X}^*} = N_{\hat{a}}^{\perp}$. Hence, i is the solution of $\hat{\mathcal{N}}$ corresponding to e and consequently, A is the admittance operator of $\hat{\mathcal{N}}$.

The last assertion concerning a linear network is an immediate consequence of Theorem 1.1.

Remark 5. Let us caution the reader about the following fact. Suppose that the operator $\hat{Z}$ is described by a matrix $[Z_{ik}]$, $Z_{ik}:\mathcal{H}\to\mathcal{H}$, i.e., $\hat{Z}$ is defined by (2.7) and maps $\mathcal{H}^{c_2}$ into itself. If the Z_{ik}'s are nonlinear, then the operator $\hat{X}^*\hat{Z}\hat{X}$ considered in Theorem 2.2 clearly need not be equal to the operator described by the matrix $\bar{X}^T\cdot[Z_{ik}]\cdot X$ even in the case of a finite network. However, if all operators Z_{ik} are linear, then such a representation is true.

THEOREM 2.3. *Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network, let $\hat{X}$ have the meaning defined above and let $W = \hat{X}^*\hat{Z}\hat{X}:\mathcal{H}^{c_0}\to\mathcal{H}^{c_0}$. Assume that there exists a number*

$\mu \neq 0$ and a real λ with $0 < \lambda < 1$ such that

$$I - \mu W \in \mathrm{Lip}_\lambda[\mathscr{H}^{c0}, \mathscr{H}^{c0}]. \tag{2.26}$$

Then $\hat{\mathscr{N}}$ is regular and the admittance operator A of $\hat{\mathscr{N}}$ is in

$$\mathrm{Lip}_\beta[\mathscr{H}^{c2}, \mathscr{H}^{c2}], \quad \textit{where } \beta = (1-\lambda)^{-1}|\mu|.$$

Proof. Referring to Theorem 1.2, we are going to show that condition (1.6) is satisfied for the abstract network $(\hat{Z}, \hat{a})$ associated with $\hat{\mathscr{N}}$. Choose $x_1, x_2 \in N_{\hat{a}}$; by Lemma 2.2 (ii) there exist $y_1, y_2 \in \mathscr{H}^{c0}$ such that $x_1 = \hat{X}y_1$ and $x_2 = \hat{X}y_2$. Then we have by Lemma 2.2 (vii), (iv), (v) and (2.26),

$$\begin{aligned} \|x_1 - x_2 - \mu P(\hat{Z}x_1 - \hat{Z}x_2)\|_{c_2} &= \|P\{x_1 - x_2 - \mu(\hat{Z}x_1 - \hat{Z}x_2)\}\|_{c_2} \\ &= \|\hat{X}\hat{X}^*\{\cdots\}\|_{c_2} \\ &= \|\hat{X}^*\{\hat{X}y_1 - \hat{X}y_2 - \mu(\hat{Z}\hat{X}y_1 - \hat{Z}\hat{X}y_2)\}\|_{c_2} \\ &= \|y_1 - y_2 - \mu(Wy_1 - Wy_2)\|_{c_0} \\ &\leqq \lambda\|y_1 - y_2\|_{c_0}. \end{aligned} \tag{2.27}$$

However, by Lemma 2.2 (iv), $\|x_1 - x_2\|_{c_2} = \|\hat{X}(y_1 - y_2)\|_{c_2} = \|y_1 - y_2\|_{c_0}$, which together with (2.27) shows that $I - \mu P\hat{Z} \in \mathrm{Lip}_\lambda[N_{\hat{a}}, N_{\hat{a}}]$; hence, the proof.

THEOREM 2.4. *Let $\hat{\mathscr{N}} = (\hat{Z}, G)$ be a Hilbert network, let $\hat{X}$ have the meaning defined above and let $W = \hat{X}^*\hat{Z}\hat{X} : \mathscr{H}^{c0} \to \mathscr{H}^{c0}$. Assume that there exist constants $\gamma > 0$ and $\lambda > 0$ such that*

$$\mathrm{Re}\langle Wx_1 - Wx_2, x_1 - x_2\rangle_{c0} \geqq \gamma\|x_1 - x_2\|_{c0}^2 \tag{2.28}$$

for all $x_1, x_2 \in \mathscr{H}^{c0}$, and

$$W \in \mathrm{Lip}_\lambda[\mathscr{H}^{c0}, \mathscr{H}^{c0}]. \tag{2.29}$$

Then $\hat{\mathscr{N}}$ is regular and the admittance operator A of $\hat{\mathscr{N}}$ is in

$$\mathrm{Lip}_\delta[\mathscr{H}^{c2}, \mathscr{H}^{c2}] \quad \textit{with } \delta = (\gamma\lambda)^{-1}(\lambda + \sqrt{\lambda^2 - \gamma^2}).$$

Proof. Referring to Theorem 1.3, choose $x_1, x_2 \in N_{\hat{a}}$; then $x_1 = \hat{X}y_1, x_2 = \hat{X}y_2$ for some $y_1, y_2 \in \mathscr{H}^{c0}$, and we have by (2.28),

$$\begin{aligned} \mathrm{Re}\langle \hat{Z}x_1 - \hat{Z}x_2, x_1 - x_2\rangle_{c_2} &= \mathrm{Re}\langle \hat{Z}\hat{X}y_1 - \hat{Z}\hat{X}y_2, \hat{X}(y_1 - y_2)\rangle_{c_2} \\ &= \mathrm{Re}\langle \hat{X}^*\hat{Z}\hat{X}y_1 - \hat{X}^*\hat{Z}\hat{X}y_2, y_1 - y_2\rangle_{c0} \\ &= \mathrm{Re}\langle Wy_1 - Wy_2, y_1 - y_2\rangle_{c0} \\ &\geqq \gamma\|y_1 - y_2\|_{c0}^2. \end{aligned}$$

However, by Lemma 2.2 (iv), $\|y_1 - y_2\|_{c_0} = \|x_1 - x_2\|_{c_2}$, and consequently, (i) in Theorem 1.3 is satisfied.

Also, as before we have by (2.29),

$$\begin{aligned}\|P(\hat{Z}x_1 - \hat{Z}x_2)\|_{c_2} &= \|\hat{X}\hat{X}^*(\hat{Z}\hat{X}y_1 - \hat{Z}\hat{X}y_2)\|_{c_2} \\ &= \|\hat{X}^*\hat{Z}\hat{X}y_1 - \hat{X}^*\hat{Z}\hat{X}y_2\|_{c_0} \\ &= \|Wy_1 - Wy_2\|_{c_0} \leqq \lambda\|y_1 - y_2\|_{c_0} \\ &= \lambda\|x_1 - x_2\|_{c_2}.\end{aligned}$$

Thus, the assumptions of Theorem 1.3 are satisfied and consequently, $\hat{\mathcal{N}}$ is regular.

THEOREM 2.5. *Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network and let $\hat{X}$ have the meaning defined above. Assume that $\hat{Z} = \hat{Z}_1 + \hat{Z}_2$, where $\hat{Z}_1 \in [\mathscr{H}^{c_2}, \mathscr{H}^{c_2}]$, and denote $W_1 = \hat{X}^*\hat{Z}_1\hat{X}$, $W_2 = \hat{X}^*\hat{Z}_2\hat{X}$. If*

(a) $$\inf_{\substack{\|x\| = 1 \\ \mathscr{H}^{c_0}}} \sup_{\substack{\|y\| = 1 \\ \mathscr{H}^{c_0}}} |\langle W_1 x, y\rangle_{c_0}| = \gamma > 0,$$

$$\inf_{\substack{\|y\| = 1 \\ \mathscr{H}^{c_0}}} \sup_{\substack{\|x\| = 1 \\ \mathscr{H}^{c_0}}} |\langle W_1 x, y\rangle_{c_0}| > 0,$$

(b) *there exists $\lambda > 0$ such that $W_2 \in \mathrm{Lip}_\lambda [\mathscr{H}^{c_0}, \mathscr{H}^{c_0}]$,*

(c) $$\gamma^{-1}\lambda < 1,$$

then $\hat{\mathcal{N}}$ is regular and the admittance operator A of $\hat{\mathcal{N}}$ is in $\mathrm{Lip}_\delta [\mathscr{H}^{c_2}, \mathscr{H}^{c_2}]$, where $\delta = (\gamma - \lambda)^{-1}$.

Proof. Choose $x, y \in N_{\hat{a}}$; then $x = \hat{X}x'$ and $y = \hat{X}y'$ for some $x', y' \in \mathscr{H}^{c_0}$; consequently,

$$\langle \hat{Z}_1 x, y\rangle_{c_2} = \langle \hat{Z}_1\hat{X}x', \hat{X}y'\rangle_{c_2} = \langle \hat{X}^*\hat{Z}_1\hat{X}x', y'\rangle_{c_0} = \langle W_1 x', y'\rangle_{c_0}.$$

Since the spaces $N_{\hat{a}}$ and $\mathscr{H}^{c_0}$ are in a one-to-one correspondence by Lemma 2.2 (ii), conditions (a) imply conditions (i) in Theorem 1.4. Moreover, as in the proof of Theorem 2.4 we confirm easily that (b) implies (ii) in Theorem 1.4; also, (iii) $\Leftrightarrow$ (c), which concludes the proof.

COROLLARY 2.1. *Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network, let $\hat{Z} \in [\mathscr{H}^{c_2}, \mathscr{H}^{c_2}]$ and let $W = \hat{X}^*\hat{Z}\hat{X}$.*

(a) *If W satisfies conditions* (a) *in Theorem* 2.5, *then $\hat{\mathcal{N}}$ is regular.*

(b) *If there exists a $\gamma > 0$ such that*

$$|\langle Wx, x\rangle_{c_0}| \geqq \gamma\|x\|_{c_0}^2$$

for all $x \in \mathscr{H}^{c_0}$, then $\hat{\mathcal{N}}$ is regular. In either case, the admittance operator A of $\hat{\mathcal{N}}$ is in $[\mathscr{H}^{c_2}, \mathscr{H}^{c_2}]$ and $\|A\| \leqq \gamma^{-1}$.

The proof follows immediately from Corollary 1.1.

By applying Theorem 1.6 we can establish another necessary and sufficient condition for regularity of a Hilbert network; since the application is straightforward, we omit the details.

Let us now discuss some specific examples of networks.

Example 1. Let $\mathscr{H}$ be the complex plane C^1 and let G be an oriented graph with the set of branches $\mathscr{B}$ having cardinal $c_2 \leqq \aleph_0$. For each of indices i, k, let $Z_{ik} \in C^1$. Assume that $\hat{Z} \in [l^{c_2}, l^{c_2}]$, where $\hat{Z}$ is defined by $\hat{Z}v = [Z_{ik}] \cdot [v_k]$, and consider the network $\hat{\mathcal{N}} = (\hat{Z}, G)$. Obviously, $\hat{\mathcal{N}}$ can be interpreted as an

R, L, C-network with constant elements in frequency domain when the frequency has some fixed value ω_0. In this case, $\mathscr{H}^c$ is just the space l^c for any $c \leqq \aleph_0$. Consequently, the operators $\hat{a}$ and $\tilde{a}$ coincide, and the same is true for $\hat{X}$ and $\tilde{X}$. Thus, referring to Theorem 2.2 and Remark 6, our network $\mathscr{N}$ will be regular precisely if the matrix $\bar{X}^T \cdot [Z_{ik}] \cdot X$ is invertible. In this case the admittance operator A of $\mathscr{N}$ is bounded and described by the matrix $X \cdot (\bar{X}^T \cdot [Z_{ik}] \cdot X)^{-1} \cdot \bar{X}^T$.

Example 2. Let $\mathscr{H}$ be a fixed (complex or real) Hilbert space, let G be an oriented graph and let Z^i be a (not necessarily linear) operator from $\mathscr{H}$ into itself such that $Z^i 0 = 0$. Let the operator $\hat{Z}$ on $\mathscr{H}^{c_2}$ be defined by

$$(2.30) \qquad \hat{Z}u = \operatorname{diag}(Z^1, Z^2, Z^3, \cdots) \cdot [u_k] = [Z^1 u_1, Z^2 u_2, Z^3 u_3, \cdots]^T,$$

and let $\mathscr{N} = (\hat{Z}, G)$; thus, a Hilbert network without mutual couplings is under consideration. Moreover, assume that there exist constants $a_1, a_2 > 0$ such that

$$(2.31) \qquad \operatorname{Re} \langle Z^i x_1 - Z^i x_2, x_1 - x_2 \rangle \geqq a_1 \|x_1 - x_2\|^2$$

for all $x_1, x_2 \in \mathscr{H}$ and all i, and

$$(2.32) \qquad Z^i \in \operatorname{Lip}_{a_2} [\mathscr{H}, \mathscr{H}]$$

for every i. We are going to show that $\mathscr{N}$ is regular and its admittance operator A is in $\operatorname{Lip}_\alpha [\mathscr{H}^{c_2}, \mathscr{H}^{c_2}]$ with $\alpha = (a_1 a_2)^{-1}(a_2 + \sqrt{a_2^2 - a_1^2})$.

Indeed, if $u, v \in \mathscr{H}^{c_2}$ and $m > 0$ is an integer, we have by (2.32),

$$(2.33) \qquad \sum_{i=1}^{m} \|Z^i u_i - Z^i v_i\|^2 \leqq a_2^2 \|u - v\|_{c_2}^2.$$

However, (2.33) shows that, due to condition $Z^i 0 = 0$, $\hat{Z}u \in \mathscr{H}^{c_2}$ whenever $u \in \mathscr{H}^{c_2}$, i.e., $\hat{Z} : \mathscr{H}^{c_2} \to \mathscr{H}^{c_2}$; also, we have $\hat{Z} \in \operatorname{Lip}_{a_2} [\mathscr{H}^{c_2}, \mathscr{H}^{c_2}]$.

Similarly, (2.31) implies that

$$(2.34) \qquad \operatorname{Re} \langle \hat{Z}u - \hat{Z}v, u - v \rangle_{c_2} \geqq a_1 \|u - v\|_{c_2}^2.$$

Thus, recalling Remark 1, assumptions of Theorem 1.3 are satisfied and consequently, the abstract network $(\hat{Z}, \hat{a})$ is regular; hence, $\mathscr{N}$ is regular, which is what we wanted to show.

Note that if $\mathscr{N}$ is a finite network, i.e., $c_2 < \aleph_0$, then the assumption $Z^i 0 = 0$ is not necessary.

Before we consider the next example, let us make the following observation.

A $c \times c$ matrix $M = [m_{ik}]$, $c \leqq \aleph_0$, will be called a ribbon matrix if there exists an integer $p \geqq 1$ such that $m_{ik} = 0$ whenever $|i - k| \geqq p$.

LEMMA 2.4. *Let $\mathscr{H}$ be a Hilbert space and let $Y = [Y_{ik}]$ be a $c \times c$ matrix ($c \leqq \aleph_0$) whose elements Y_{ik} are operators from $\mathscr{H}$ into itself. If*

(a) *Y is a ribbon matrix,*

(b) *there exists an $a > 0$ such that $Y_{ik} \in \operatorname{Lip}_a [\mathscr{H}, \mathscr{H}]$ for all i, k,*

(c) *$Y_{ik} 0 = 0$ for all i, k,*

then the operator $\hat{Y}$ defined on $\mathscr{H}^c$ by $\hat{Y}x = Y \cdot [x_k]$ is in $\operatorname{Lip}_{a'} [\mathscr{H}^c, \mathscr{H}^c]$ with $a' = (2p + 1)a$.

If, in addition, $c < \aleph_0$, the assumption (c) can be dropped.

Proof. First, observe that due to (a) the sum $\sum_\mu Y_{i\mu} x_\mu$ involves at most $2p + 1$ nonzero terms for any i. Choosing $u, v \in \mathscr{H}^c$ and denoting $\hat{Y}u = [y_k]$, $\hat{Y}v = [z_k]$, we have by Schwartz's inequality and (b),

$$
\begin{aligned}
\sum_{i=1}^{m} \|y_i - z_i\|^2 &= \sum_{i=1}^{m} \|\sum_{\mu} Y_{i\mu}u_\mu - Y_{i\mu}v_\mu\|^2 \\
&\leqq \sum_{i=1}^{m} \left(\sum_{|\mu - i| \leqq p} c\|u_\mu - v_\mu\| \right)^2 \\
&\leqq \sum_{i=1}^{m} c^2(2p+1) \sum_{|\mu - i| \leqq p} \|u_\mu - v_\mu\|^2 \\
&\leqq c^2(2p+1)^2 \|u - v\|_c^2.
\end{aligned} \tag{2.35}
$$

However, (2.35) shows that, due to (c), $[y_k] \in \mathscr{H}^c$, i.e., $\hat{Y}$ maps $\mathscr{H}^c$ into itself; moreover, (2.35) implies that $\hat{Y} \in \mathrm{Lip}_q[\mathscr{H}^c, \mathscr{H}^c]$ with $q = c(2p+1)$, which concludes the proof.

Example 3. Let G be an oriented graph having the set of branches $\mathscr{B}$ with cardinal $c_2 \leqq \aleph_0$, and let $\mathscr{H}$ be the real Hilbert space $L_2[0, \tau]$, $\tau > 0$. For every index i, let $r_i(\sigma)$ be a real function on $(-\infty, \infty)$ satisfying the condition $r_i(0) = 0$, and, for every pair i, k, let $S_{ik}(t)$ be a real function having a bounded derivative $S'_{ik}(t)$ on $[0, \tau]$. Moreover, assume that the following conditions are satisfied:

(i) there exist constants $0 < a \leqq b$ such that

$$a(\sigma_1 - \sigma_2)^2 \leqq (r_i(\sigma_1) - r_i(\sigma_2))(\sigma_1 - \sigma_2) \leqq b(\sigma_1 - \sigma_2)^2 \tag{2.36}$$

for all i and $\sigma_1, \sigma_2 \in (-\infty, \infty)$;

(ii) there exists a constant $d > 0$ such that

$$|S_{ik}(t)| \leqq d \tag{2.37}$$

for all i, k and $t \in [0, \tau]$;

(iii) the $c_2 \times c_2$ matrices $S(t) = [S_{ik}(t)]$ and $-S'(t) = [-S'_{ik}(t)]$ are symmetric positive semidefinite ribbon matrices on $[0, \tau]$. (By saying that a $c \times c$ matrix $A(t)$ is positive semidefinite on $[0, \tau]$, we mean that for any real c-vector $\xi \in l^c$ and $t \in [0, \tau]$ we have $\xi^T \cdot A(t) \cdot \xi \geqq 0$.)

For every pair i, k, let $Z_{ik}: L_2[0, \tau] \to L_2[0, \tau]$ be defined by

$$(Z_{ik}x)(t) = r_{ik}(x(t)) + S_{ik}(t) \int_0^t x(\sigma)\, d\sigma, \tag{2.38}$$

where $r_{ii} = r_i$ for all i and $r_{ik} = 0$ for $i \neq k$. Furthermore, define the operator $\hat{Z}$ on $L_2^{c_2}[0, \tau]$ by

$$\hat{Z}y = [Z_{ik}] \cdot [y_k], \tag{2.39}$$

and consider the network $\hat{\mathscr{N}} = (\hat{Z}, G)$. Obviously, $\hat{\mathscr{N}}$ can be interpreted as an R, C-network with nonlinear resistors and time-varying capacitors. We are going to show that, under the assumptions made, $\hat{\mathscr{N}}$ is regular and its admittance operator A is in $\mathrm{Lip}_\alpha[L_2^{c_2}, L_2^{c_2}]$ for some $\alpha > 0$.

To this end, define operators R_{ik}, U_{ik} on $L_2[0, \tau]$ by

$$(R_{ik}x)(t) = r_{ik}(x(t)), \qquad (U_{ik}x)(t) = S_{ik}(t) \int_0^t x(\sigma)\, d\sigma, \tag{2.40}$$

and operators $\hat{R}, \hat{U}$ on $L_2^{c_2}[0,\tau]$ by

$$\hat{R}y = [R_{ik}] \cdot [y_k], \qquad \hat{U}y = [U_{ik}] \cdot [y_k]. \tag{2.41}$$

Then $\hat{Z} = \hat{R} + \hat{U}$; moreover, it is easy to verify that $R_{ik} \in \mathrm{Lip}_b[L_2, L_2]$, $U_{ik} \in \mathrm{Lip}_{\tau d}[L_2, L_2]$ and $R_{ik}0 = U_{ik}0 = 0$. Hence, by Lemma 2.4, $\hat{R}$, $\hat{U}$, and consequently, $\hat{Z}$ too, are in $\mathrm{Lip}_\alpha[L_2^{c_2}, L_2^{c_2}]$ for some $\alpha > 0$.

On the other hand, (2.36) implies by an easy calculation that

$$a\|x_1 - x_2\|_{c_2}^2 \leqq \langle \hat{R}x_1 - \hat{R}x_2, x_1 - x_2 \rangle_{c_2} \tag{2.42}$$

for all $x_1, x_2 \in L_2^{c_2}$. Moreover, observe that by the dominated convergence theorem we have for any $x_i = [x_k^i(t)] \in L_2^{c_2}[0,\tau], i = 1, 2,$

$$\langle x_1, x_2 \rangle_{c_2} = \sum_k \langle x_k^1, x_k^2 \rangle = \sum_k \int_0^\tau x_k^1(\sigma) x_k^2(\sigma)\, d\sigma$$

$$= \int_0^\tau \left(\sum_k x_k^1(\sigma) x_k^2(\sigma) \right) d\sigma = \int_0^\tau x_1^T(\sigma) \cdot x_2(\sigma)\, d\sigma.$$

Thus, if $x = [x_k(t)] \in L_2^{c_2}$ and $y(t) = \int_0^t x(\sigma)\, d\sigma$, we get by integrating by parts,

$$\langle \hat{U}x, x \rangle_{c_2} = \int_0^\tau x^T(t) \cdot S(t) \cdot \left(\int_0^t x(\sigma)\, d\sigma \right) dt$$

$$= \tfrac{1}{2} y^T(\tau) \cdot S(\tau) \cdot y(\tau) - \int_0^\tau y^T(\sigma) \cdot S'(\sigma) \cdot y(\sigma)\, d\sigma \geqq 0.$$

Combining this with (2.42), it follows that

$$a\|x_1 - x_2\|_{c_2}^2 \leqq \langle \hat{Z}x_1 - \hat{Z}x_2, x_1 - x_2 \rangle_{c_2}$$

for all $x_1, x_2 \in L_2^{c_2}$. Thus, recalling Remark 1, it follows immediately that our network $\hat{\mathcal{N}}$ is regular.

Note also the fact that assumption (iii) may be replaced by the weaker requirement that matrices $X^T \cdot S(t) \cdot X$ and $-X^T \cdot S'(t) \cdot X$ are symmetric positive semidefinite ribbon matrices on $[0, \tau]$.

In order to formulate the approximation theorem for infinite networks, let us introduce some further concepts and notation.

Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be an infinite Hilbert network ($c_2 = \aleph_0$), and let $\mathscr{B} = (b_1, b_2, b_3, \cdots)$ be the sequence of all branches of G. If $n \geqq 1$ is an integer, let G_n be the restriction of the mapping G to the set $\mathscr{B}_n = (b_1, b_2, \cdots, b_n)$; then G_n is again a (finite) oriented graph.

Next, if $x = [x_k]$ is an $\aleph_0$-vector, let $(x)_n$ signify the n-vector $[x_1, x_2, x_3 \cdots, x_n]^T$; similarly, if $y = [y_k]$ is an n-vector, let y' stand for the $\aleph_0$-vector $[y_1, y_2, \cdots, y_n, 0, 0, \cdots]^T$. An $\aleph_0$-vector x will be called simple if all but finitely many of its components are zero.

Finally, let $\hat{Z}_n$ be the operator from $\mathscr{H}^n$ into itself defined by

$$\hat{Z}_n x = (\hat{Z}x')_n, \tag{2.43}$$

and let $\hat{\mathcal{N}}_n = (\hat{Z}_n, G_n)$; thus, $\hat{\mathcal{N}}_n$ is a finite subnetwork of $\hat{\mathcal{N}}$. Of course, if $e_n \in \mathscr{H}^n$, we will look for a solution of $\hat{\mathcal{N}}_n$ corresponding to e_n in the same space $\mathscr{H}^n$.

Note that if in particular the operator $\hat{Z}$ is described by a matrix $[Z_{ik}]$ and defined by (2.7), then $\hat{Z}_n$ is described by the $n \times n$ submatrix of $[Z_{ik}]$ standing in the upper left corner.

THEOREM 2.6. *Let $\hat{\mathscr{N}} = (\hat{Z}, G)$ be an infinite Hilbert network, let $W = \hat{X}^* \hat{Z} \hat{X} : \mathscr{H}^{co} \to \mathscr{H}^{co}$ and let $\hat{\mathscr{N}}_n$ have the meaning defined above. Assume that the following conditions are satisfied*:

(i) *there exist constants $\gamma, \lambda > 0$ such that*

$$\operatorname{Re} \langle Wx_1 - Wx_2, x_1 - x_2 \rangle_{co} \geqq \gamma \|x_1 - x_2\|_{co}^2 \tag{2.44}$$

for all $x_1, x_2 \in \mathscr{H}^{co}$, and

$$\hat{Z} \in \operatorname{Lip}_\lambda [\mathscr{H}^{\aleph_0}, \mathscr{H}^{\aleph_0}]; \tag{2.45}$$

(ii) *the set S of all simple vectors in $N_{\hat{a}}$ is dense in $N_{\hat{a}}$.*

Then

(a) *$\hat{\mathscr{N}}$ and each $\hat{\mathscr{N}}_n$ is regular,*

(b) *if $i \in \mathscr{H}^{\aleph_0}$ is the solution of $\hat{\mathscr{N}}$ corresponding to $e \in \mathscr{H}^{\aleph_0}$ and $i_n \in \mathscr{H}^n$ is the solution of $\hat{\mathscr{N}}_n$ corresponding to $(e)_n \in \mathscr{H}^n$, then $i_n' \to i$ in $\mathscr{H}^{\aleph_0}$.*

If, in addition, $\mathscr{N}$ is linear, then conditions (2.44), (2.45) *may be replaced by the requirements*

$$|\langle Wx, x \rangle_{co}| \geqq \gamma \|x\|_{co}^2$$

for all $x \in \mathscr{H}^{co}$, and $\hat{Z} \in [\mathscr{H}^{\aleph_0}, \mathscr{H}^{\aleph_0}]$.

Proof. For each $n \geqq 1$, let $\mathscr{H}^{n'} = \{x' : x \in \mathscr{H}^n\}$; then $\mathscr{H}^{1'} \subset \mathscr{H}^{2'} \subset \mathscr{H}^{3'} \subset \cdots \subset \mathscr{H}^{\aleph_0}$. Let $P_n : \mathscr{H}^{\aleph_0} \to \mathscr{H}^{n'}$ be defined by $P_n x = \{(x)_n\}'$; then P_n is clearly the orthogonal projection from $\mathscr{H}^{\aleph_0}$ onto $\mathscr{H}^{n'}$.

Next, let d_n be the incidence matrix of G_n, i.e., d_n is defined by (2.3). From the definition of the graph it follows that if $x \in \mathscr{H}^n$ satisfies the equation $d_n^T \cdot x = 0$, then $d_{n+1}^T \cdot (x')_{n+1} = 0$ and $a \cdot x' = 0$. Thus, letting

$$N_{\hat{a}_n^T} = \{x : x \in \mathscr{H}^n, d_n^T \cdot x = 0\} \quad \text{and} \quad N'_{\hat{a}_n^T} = \{x' : x \in N_{\hat{a}_n^T}\},$$

we have

$$N'_{\hat{a}_n^T} \subset \mathscr{H}^{n'}, \quad N'_{\hat{a}_n^T} \subset N'_{\hat{a}_{n+1}^T} \subset N_{\hat{a}}. \tag{2.46}$$

Define now the $\aleph_0 \times \aleph_0$ matrices a_n by

$$a_n = \left[\begin{array}{c|c} d_n^T & 0 \\ \hline 0 & \tilde{d} \end{array}\right], \tag{2.47}$$

where $\tilde{d}$ is the $\aleph_0 \times \aleph_0$ diagonal matrix

$$\tilde{d} = \operatorname{diag}(1, \tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{4}, \cdots), \tag{2.48}$$

and let $\hat{a}_n \in [\mathscr{H}^{\aleph_0}, \mathscr{H}^{\aleph_0}]$ be defined by $\hat{a}_n x = a_n \cdot x$. From (2.47) and (2.48) it follows readily that $N_{\hat{a}_n} = N'_{\hat{a}_n^T}$, where $N_{\hat{a}_n} \subset \mathscr{H}^{\aleph_0}$ is the nullspace of $\hat{a}_n$. Hence, $N_{\hat{a}_n} \subset \mathscr{H}^{n'}$ and $N_{\hat{a}_n} \subset N_{\hat{a}_{n+1}}$, i.e., condition (ii) in Theorem 1.7 is satisfied.

Also, it is clear that assumption (ii) is exactly the condition (iii) in Theorem 1.7.

On the other hand, as we have seen in the proof of Theorem 2.4, condition (2.44) is equivalent to condition

$$\operatorname{Re} \langle \hat{Z}y_1 - \hat{Z}y_2, y_1 - y_2 \rangle_{\aleph_0} \geqq \gamma \|y_1 - y_2\|_{\aleph_0}^2 \tag{2.49}$$

for all $y_1, y_2 \in N_{\hat{a}}$. Moreover, if P is the orthogonal projection of $\mathscr{H}^{\aleph_0}$ onto $N_{\hat{a}}$, then we have by (2.45),

$$(2.50) \qquad P\hat{Z} \in \mathrm{Lip}_\lambda [N_{\hat{a}}, N_{\hat{a}}].$$

Hence, by Theorem 1.3, the abstract network $(\hat{Z}, \hat{a})$ over $\mathscr{H}^{\aleph_0}$ is regular, i.e., our network $\hat{\mathscr{N}}$ is regular.

Next, we are going to show that each network $\hat{\mathscr{N}}_n$ is regular. Indeed, if $u_1, u_2 \in N_{\hat{a}_n^T} \subset \mathscr{H}^n$, we have by (2.49), (2.43) and the inclusion $N'_{\hat{a}_n^T} \subset N_{\hat{a}}$,

$$(2.51) \qquad \begin{aligned} &\mathrm{Re}\,\langle \hat{Z}_n u_1 - \hat{Z}_n u_2, u_1 - u_2\rangle_n \\ &\quad = \mathrm{Re}\,\langle (\hat{Z}u_1')_n - (\hat{Z}u_2')_n, u_1 - u_2\rangle_n = \mathrm{Re}\,\langle \hat{Z}u_1' - \hat{Z}u_2', u_1' - u_2'\rangle_{\aleph_0} \\ &\quad \geqq \gamma\|u_1' - u_2'\|_{\aleph_0}^2 = \gamma\|u_1 - u_2\|_n^2. \end{aligned}$$

Furthermore, if Q_n is the orthogonal projection from $\mathscr{H}^n$ onto $N_{\hat{a}_n^T}$ and $u_1, u_2 \in N_{\hat{a}_n^T}$, then $u_1', u_2' \in N_{\hat{a}}$ and (2.45) yields

$$(2.52) \qquad \begin{aligned} \|Q_n(\hat{Z}_n u_1 - \hat{Z}_n u_2)\|_n &\leqq \|\hat{Z}_n u_1 - \hat{Z}_n u_2\|_n = \|(\hat{Z}u_1')_n - (\hat{Z}u_2')_n\|_n \\ &\leqq \|\hat{Z}u_1' - \hat{Z}u_2'\|_{\aleph_0} \leqq \lambda\|u_1' - u_2'\|_{\aleph_0} \\ &= \lambda\|u_1 - u_2\|_n, \end{aligned}$$

i.e.,

$$(2.52) \qquad Q_n\hat{Z}_n \in \mathrm{Lip}_\lambda [N_{\hat{a}_n^T}, N_{\hat{a}_n^T}].$$

Thus, according to Theorem 1.3, (2.51) and (2.52) show that the abstract network $(\hat{Z}_n, \hat{a}_n^T)$ over $\mathscr{H}^n$ is regular, and consequently, $\hat{\mathscr{N}}_n$ is regular.

Consider now the abstract network $\mathscr{N}_n' = (\hat{Z}, \hat{a}_n)$ over $\mathscr{H}^{\aleph_0}$. Since $N_{\hat{a}_n} = N'_{\hat{a}_n^T} \subset N_{\hat{a}}$, we have readily from (2.49),

$$(2.53) \qquad \mathrm{Re}\,\langle \hat{Z}y_1 - \hat{Z}y_2, y_1 - y_2\rangle_{\aleph_0} \geqq \gamma\|y_1 - y_2\|_{\aleph_0}^2$$

for all $y_1, y_2 \in N_{\hat{a}_n}$. Similarly, if K_n is the orthogonal projection from $\mathscr{H}^{\aleph_0}$ onto $N_{\hat{a}_n}$, (2.45) implies as above that

$$(2.54) \qquad K_n\hat{Z} \in \mathrm{Lip}_\lambda [N_{\hat{a}_n}, N_{\hat{a}_n}].$$

Hence, (2.53), (2.54) show by Theorem 1.3 that $\mathscr{N}_n'$ is regular.

Finally, we are going to show that if $i_n \in \mathscr{H}^n$ is the solution of $\hat{\mathscr{N}}_n$ corresponding to $(e)_n \in \mathscr{H}^n$, then $i_n' \in \mathscr{H}^{\aleph_0}$ is the solution of $\mathscr{N}_n'$ corresponding to $\{(e)_n\}' = P_n e$. Indeed, we have $i_n \in N_{\hat{a}_n^T}$ and

$$(2.55) \qquad \langle \tilde{c}, \hat{Z}_n i_n - (e)_n\rangle_n = 0$$

for each $\tilde{c} \in N_{\hat{a}_n^T}$. Denote $J_n \in \mathscr{H}^{\aleph_0}$ the solution of $\mathscr{N}_n'$ corresponding to $P_n e$; then $J_n \in N_{\hat{a}_n}$ and

$$(2.56) \qquad \langle c, \hat{Z}J_n - \{(e)_n\}'\rangle_{\aleph_0} = 0$$

for all $c \in N_{\hat{a}_n}$. However, due to definition (2.47) of a_n it follows that there exists $I_n \in \mathscr{H}^n$ such that $J_n = I_n'$, and by equality $a_n \cdot J_n = 0$ we necessarily have $I_n \in N_{\hat{a}_n^T}$.

Thus, for any $c^+ \in N_{\hat{a}_n^T}$ we have $(c^+)' \in N_{\hat{a}_n}$ and (2.56) yields

$$0 = \langle (c^+)', \hat{Z}I'_n - \{(e)\}_n\}' \rangle_{\aleph_0} = \langle c^+, (\hat{Z}I'_n)_n - (e)_n \rangle_n$$
$$= \langle c^+, \hat{Z}_n I_n - (e)_n \rangle_n.$$

Comparing this with (2.54), we see that, due to regularity of $\hat{\mathcal{N}}_n$ and $\mathcal{N}_n$, necessarily $i_n = I_n$; consequently, $J_n = i'_n$ as desired.

Now, applying Theorem 1.7 to networks $\mathcal{N}$ and $\mathcal{N}_n$, we conclude that $i'_n \to i$ in $\mathcal{H}^{\aleph_0}$ and the proof is complete.

Recalling Theorem 1.8, we immediately obtain the following result.

THEOREM 2.7. *Let $\hat{\mathcal{N}} = (\hat{Z}, G)$ be a Hilbert network.*

(a) *Let $h \in \mathcal{H}$ be a fixed element, and for every $m \geqq 1$ let $e_m \in \mathcal{H}^{c_2}$ be such that the m-th component of e_m is h and all others are zero; if $\hat{\mathcal{N}}$ has a solution $[i_k^m] \in \mathcal{H}^{c_2}$ corresponding to e_m and a solution $[i_k^n] \in \mathcal{H}^{c_2}$ corresponding to e_n, then*

$$\langle i_m^n, h \rangle = \langle i_n^m, h \rangle \quad \textit{provided } \hat{Z} \textit{ is symmetric,}$$

and

$$\langle i_m^n, h \rangle = \overline{\langle i_n^m, h \rangle} \quad \textit{provided } \hat{Z} \textit{ is self-adjoint.}$$

(b) *If $\hat{Z}$ is a positive operator, i.e., $\langle \hat{Z}x, x \rangle \geqq 0$ for all $x \in \mathcal{H}^{c_2}$, and if $i \in \mathcal{H}^{c_2}$ is a solution of $\hat{\mathcal{N}}$ corresponding to $e \in \mathcal{H}^{c_2}$, then $\langle i, e \rangle \geqq 0$.*

Proof. The proof is obvious.

Concluding the paper, let us mention the fact that the Tellegen's theorem holds for Hilbert networks, too, and is an immediate consequence of Definitions 1, 2 and Theorem 2.1. Indeed, we have the following theorem.

THEOREM 2.8. *Let $\hat{\mathcal{N}}_m = (\hat{Z}_m, G), m = 1, 2,$ be Hilbert networks having the same graph G, and let $i_m = [i_k^m] \in \mathcal{H}^{c_2}$ be a solution of $\hat{\mathcal{N}}_m$ corresponding to $e_m = [e_k^m] \in \mathcal{H}^{c_2}$. Then*

$$\langle u_1, i_2 \rangle_{c_2} = 0, \tag{2.57}$$

i.e., $\sum_k \langle u_k^1, i_k^2 \rangle = 0$, where $u_1 = [u_k^1] = \hat{Z}_1 i_1 - e_1$.

Proof. By K_i^*, $u_1 \in N_{\hat{a}}^{\perp}$, and by K_2^*, $i_1, i_2 \in N_{\hat{a}}$; from this, (2.57) follows immediately.

REFERENCES

[1] G. J. MINTY, *Monotone networks*, Proc. Roy. Soc. London Ser. A, 257 (1960), pp. 94–212.

[2] G. ZAMES, *On the input-output stability of time-varying nonlinear systems*, IEEE Trans. Automatic Control, AC-11 (1964), pp. 228–238, 465–476.

[3] I. W. SANDBERG, *A survey of the theory of systems governed by nonlinear functional equations*, System Theory, L. A. Zadeh and E. Polak, eds., McGraw-Hill, New York, 1969.

[4] G. ZAMES AND P. L. FALB, *Systems with monotone and slope-restricted nonlinearities*, this Journal, 6 (1968), pp. 89–107.

[5] H. FLANDERS, *Infinite networks: I—resistive networks*, IEEE Trans. Circuit Theory, CT-18 (1971), pp. 326–331.

[6] A. H. ZEMANIAN, *Passive operator networks*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 184–193.

[7] I. I. KOLODNER, *Equations of Hammerstein type in Hilbert space*, J. Math. Mech., 13 (1964), pp. 701–750.

[8] K. YOSIDA, *Functional Analysis*, Academic Press, New York, 1965.

[9] S. B. BERBERIAN, *Introduction to Hilbert Space*, Oxford University Press, New York, 1961.

[10] N. I. AKHIEZER AND I. M. GLAZMAN, *Theory of Linear Operators in Hilbert Space*, vol. 1, Frederick Ungar, New York, 1966.

[11] J. P. ROTH, *An application of algebraic topology to a problem in numerical analysis: On the existence of a solution to the network problem*, Proc. Nat. Acad. Sci. U.S.A., 41 (1958), pp. 518–521.